A Churn for the Better

Localizing Censorship using Network-level Path Churn and Network Tomography

Shinyoung Cho
Stony Brook University / SUNY Korea
shicho@cs.stonybrook.edu

Rishab Nithyanand
Data & Society Research Institute
rishab@datasociety.net

Abbas Razaghpanah
Stony Brook University
arazaghpanah@cs.stonybrook.edu

Phillipa Gill
University of Massachusetts, Amherst
phillipa@cs.umass.edu

ABSTRACT

Recent years have seen the Internet become a key vehicle for citizens around the globe to express political opinions and organize protests. This fact has not gone unnoticed, with countries around the world repurposing network management tools (*e.g.,* URL filtering products) and protocols (*e.g.,* BGP, DNS) for censorship. Previous work has focused on identifying how censorship is performed. However, there is no major studies to identify, at a global scale, the networks responsible for performing censorship. Also, repurposing network products for censorship can have unintended *international* impact, which we refer to as "censorship leakage". While there have been anecdotal reports of censorship leakage, there has yet to be a systematic study of censorship leakage at a global scale.

In this paper, we combine a global censorship measurement platform (ICLab) with a general-purpose technique – *boolean network tomography* – to identify which AS on a network path is performing censorship. At a high-level, our approach exploits BGP churn to narrow down the set of potential censoring ASes by 97%. We identify 108 censoring ASes and find that the censorship introduced by 32 of the 108 censoring ASes has an impact on users located outside of the jurisdiction of the censoring AS, resulting in the *leaking* of regional censorship policies.

CCS CONCEPTS

• **Networks** → **Network measurement**; **Network dynamics**;

KEYWORDS

Network Measurement, Internet Censorship, Censorship Leakage, Localization, Boolean Network Tomography, Boolean Satisfiability

ACM Reference Format:
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 *Proceedings of CoNEXT '17*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3143361.3143386

CoNEXT '17, December 12–15, 2017, Incheon, Republic of Korea

ACM ISBN 978-1-4503-5422-6/17/12…$15.00
https://doi.org/10.1145/3143361.3143386

1 INTRODUCTION

The Internet is now regarded as part of the critical infrastructure, with citizens relying on it for dissemination of information and organizing political action. Consequently, governments and network-level entities – *e.g.,* Autonomous Systems (ASes) – are implementing forms of censorship to restrict access to specific content, in many cases, repurposing existing network management tools [10] and protocols (*e.g.,* DNS [1, 21], BGP [5]) to filter Internet content.

Previous work has focused on identifying how governments and network level entities performed such censorship. Most of these studies have conducted measurements focusing on country-specific censorship (e.g., China [3], Iran [2], and Pakistan [1]). However, there are no major studies to identify ASes, at a global scale, responsible for implementing such censorship and there is a lack of understanding such censorship from longitudinal perspectives. By identifying ASes responsible for such censorship, we can also quantify its unintended international impact, or *censorship leakage*. We define censorship leakage as the leakage of regional information controls policies into countries and networks outside of the network implementing the policy. The most well-known non-BGP hijack instance of such leakage is the case of censorship implemented via DNS root-servers located in China impacting international users [3]. It must be noted that in this paper, we do not detect censorship and its leakage caused due to Internet routing anomalies (*e.g.,* the 2008 BGP hijack of YouTube traffic by Pakistan Telecom [5]).

To identify ASes performing censorship on a global scale, we need a technique that is able to identify specific instances of censorship globally. This is a challenge due to the general lack of vantage points that are available for measuring censorship on an ongoing basis. We address this challenge by combining the ICLab measurement platform with the idea of *boolean network tomography* [33]. ICLab is a platform of ∼1,000 globally distributed vantage points that have been performing measurements of censorship on an ongoing basis since November 2015 (more details in §2.1). Our intuition is that we can observe multiple tests from a given vantage point to a given destination and that, if there is sufficient path churn between the vantage point and destination, we can create a set of boolean constraints where the constraint is true if censorship is observed, and false otherwise. The idea of exploiting network-level path churn has also been used to improve the success of de-anonymization attacks on Tor users [30].

In the case where censorship is observed, it must be the case that at least one autonomous system (AS) on the path is performing censorship. We can then input these constraints into an off-the-shelf

SAT solver to identify the AS performing censorship. In this study, we demonstrate the applicability of this intuition by answering the following key questions: (1) Is there enough path churn observed in our measurements to create a solvable set of constraints? We need to validate that we have enough variability in paths, especially in the cases where we observe censorship, to create a solvable set of constraints to narrow down the set of potential censors. (2) Will our constraints generate a small set of potential censoring ASes? We want to make sure that the set of potential censoring ASes is not intractably large, making it impossible to exactly identify ASes responsible for implementing censorship. While answering these questions, we make the following contributions:

Problem reformulation. We demonstrate how measurements gathered by the ICLab platform can be used to formulate a boolean network tomography problem solvable by off-the-shelf SAT solvers. Our approach carries over to other measurement databases such as those generated by the OONI [15] and the M-Lab [22] platforms.

Measuring and exploiting network-level churn. We show that the instability of network-level paths can act as a substitute for strategically placed internal monitors. Specifically, we show that 25%, 30%, 38%, and 67% of paths between ICLab vantage points and web servers are observed to change over periods of one day, week, month, and year. These changes are found to significantly improve the solvability of our constructed SAT problems.

Identifying censors and censorship leakage. We empirically demonstrate that our approach allows us to reduce the size of the set of potential censoring ASes by 97%, on average. Further, we exactly identify 108 censoring ASes located in 49 different countries. Our study also identifies *leakage* of censorship policies – *i.e.,* cases where censoring ASes blocked access to content even for users outside their network. Specifically, we find that 32 and 18 of the censoring ASes leak censorship to other ASes and countries, respectively.

2 BACKGROUND & RELATED WORK

Censorship measurement. Much of previous work has focused on understanding how censorship is performed by individual network-level entities. Studies have shown that censors may restrict access to content by injecting incorrect DNS replies [1, 25], sending TCP reset packets spuriously [2, 20], using off-the-shelf filtering and blocking tools [10], or throttling connections to censored content [2]. Our study builds off of related work on large-scale longitudinal censorship measurement systems. Specifically, we use data collected by the ICLab platform [28] to detect censorship and generate constraints. This enables us to identify the networks performing such censorship at a global scale. Conceptually, our techniques could be applied to other platforms such as OONI [15] as well.

Fault localization. The problem of identifying the network responsible for implementing censorship can also be recast as a fault localization problem. Other studies have approached the problem of network failure localization with different perspectives. Lifeguard [19] relies on historical control-plane measurements and active probing to automatically identify and route around network failures via crafted BGP messages. Feamster *et al.* [12] measure the effectiveness of reactive routing around node failures. Their approach localizes failures in real-time by analyzing results of active probes, including pings and traceroutes, between vantage points. While other work has focused on identifying the ASes that trigger path changes on the Internet [13, 17, 27, 31, 35], we focus on utilizing such path changes to localize the ASes that perform censorship.

Boolean network tomography. Network tomography [33] typically involves using end-to-end measurements and a set of monitors within the network to uncover hidden node values (which in the case of boolean network tomography, may only take the values True or False). Monitors are used to ensure that appropriate end-to-end measurements may be performed to unveil specific node characteristics. We use the data gathered by the ICLab platform as end-to-end measurements from which we identify nodes (ASes) implementing specific types of censorship. Unlike typical boolean network tomography problems, our study is limited by the absence of strategically located monitors from which end-to-end measurements can be gathered. However, we show that due to the churn of network-level paths, we are still able to use boolean network tomography to identify censoring ASes. Several studies have focused on the problem of error localization through boolean network tomography [8, 9, 11, 23, 32]. Ma *et al.* [23] focus on identifying the conditions, monitor locations, and probing mechanisms that are necessary for fault localization through boolean network tomography. [11] use a boolean network tomography approach with "troubleshooting sensors" (monitors) located within the network to identify misconfigured routers.

2.1 The ICLab Dataset

We rely on data gathered by the ICLab censorship measurement platform [28] as a source for end-to-end measurements. The ICLab platform repetitively performs a variety of measurements between a set of over 1K globally distributed vantage points spread across 539 unique ASes and web servers hosting popular or sensitive web content. The vantage points include end points of popular VPN providers and a small number of Raspberry Pi's running the ICLab client software, while the web servers are a combination of the Alexa Top 500 global websites and regionally sensitive URLs obtained from the Citizen Lab test lists. The platform aims to (1) identify content being censored, (2) understand how censorship is implemented, and (3) record changes in censorship policies over time. Specifically, ICLab identifies the following anomalies as indicative of potential censorship:

TTL and RST anomalies. A packet injector will often have attributes that differ from the legitimate server for a connection. The ICLab platform issues HTTP GET requests and records all responses (while following redirects). The platform then analyzes raw packet captures to identify anomalies. Specifically, we compare the IP TTL header on the SYNACK packet of the connection with subsequent packets. This relies on the assumption that a censor will not be fast enough to act prior to the SYNACK being sent by the server and that the SYNACK TTL is identical for the subsequent packets when we rule out load balanced cases. When the suspicious packets have the RST flag set, we call it a case of RST anomaly. Further, injected packets will often not be able to perfectly mimic the TCP state of the server [34]. We look for cases where there are overlapping sequence numbers between packets or gaps in sequence numbers. These sequence number anomalies, especially when combined with

Period	2016-05 ~ 2017-05
Unique URLs	774
AS Vantage Points	539
Destination ASes	620
All ASes on All Paths	1103
Countries	219
Measurements	4.9M
– w/TTL anomalies	7.2K (0.15%)
– w/RST anomalies	5.0K (0.10%)
– w/Blockpages	1.5K (0.03%)

Table 1: ICLab dataset characteristics.

packets having the RST flag set (to close the connection) are likely indicators of censorship. We increase the reliability of our TTL- and RST-anomaly based censorship detection by only considering cases where no HTML content is returned or a block page is returned.

Block pages. The platform analyzes the responses received to identify blockpages that are returned by a censor. This is done by performing regular expression matching with known examples of blockpages (provided by the OONI project [26]) and by comparing responses with those obtained from censor-free vantage points within the United States. In the latter case, we employ techniques developed by Jones *et al.* to identify block pages [18].

Network paths. In addition to gathering the above data, the platform also records traceroutes from vantage points to the corresponding destinations of each test. In total, we utilize 4.9M measurements (with 14K total identified anomalies) from the platform between vantage points located in 539 different ASes (in 219 countries) and 774 URLs. A summary of the ICLab data that we use in our work is summarized in Table 1.

Ethical considerations and limitations. To mitigate risk, the vast majority of ICLab's vantage points are obtained via commercial VPN providers (many of which are located in ASes classified as content ASes by CAIDA [6]). This allows us to obtain widespread continuous measurements, without putting users in specific regions at risk. A potential limitation of this decision is the inability to observe the same filtering as ASes providing residential connections.

In collaboration with the Citizen Lab, we have worked to deploy a handful of Raspberry Pi nodes running the measurement software. Prior to deploying a node, we discuss with the volunteer about the potential risks and they are further presented with a form that summarizes risks for their given country based on existing metrics (*e.g.,* Freedom House [16]). Since the platform does not collect personally identifiable information, our IRB has determined that this project does not constitute human subjects research. Regardless, we maintain contact with any volunteers and monitor the political situations in different regions. Some regions have been deemed too risky to operate in (*e.g.,* Iran, Syria). In general, we aim to balance risk with potential benefits of the measurements.

3 LOCALIZING CENSORS

At a high-level, our approach works as follows: First, we use the traceroutes gathered by the ICLab platform to construct boolean clauses such that the literals in the clauses represent ASes observed in the traceroute. We then use the censorship measurements associated with the corresponding traceroutes to assign truth values to the clauses. Finally, the clauses are converted to Conjunctive Normal Form (CNF) and used as input to an off-the-shelf SAT solver. The process is repeated for each type of censorship measurement (*i.e.,* HTTP tampering and blockpage detection) and various time slices (*i.e.,* for all measurements performed during the same day, week, and month). Next, we analyze the solutions returned by the SAT solver for each CNF. In cases where there are multiple solutions – *i.e.,* multiple truth assignments for a given CNF formulation – we return all literals (ASes) having *True* assignments as potential censors. In cases where there is a single solution, we return all literals (ASes) having *True* assignments as censors. Finally, we characterize censorship leakage by identifying ASes that observe censorship only when they transit through censoring ASes.

3.1 Constructing a SATisfiability problem

Each record in the ICLab dataset contains: (1) the vantage point AS, (2) the URL being tested, (3) the anomaly being tested (and whether it was detected or not), (4) three traceroutes between the vantage point and the URL at the time of testing, and (5) the time at which the test was performed. We use each of these records to create boolean satisfiability problems as follows:

Clause formulation. First, we use historical IP-to-AS mapping from CAIDA [7] to convert the IP-level traceroutes to AS-level paths. Next, we eliminate cases with inconclusive paths – *i.e.,* cases where one of the following situations occurred: (1) IP-to-AS mapping was not possible for IPs observed in the traceroute and different ASes observed in the previous and subsequent responsive hops, (2) traceroutes were not possible due to errors, (3) AS-inference was not possible due to non-responsive hops and different ASes observed in the previous and subsequent responsive hop, and (4) there was more than one AS-level path obtained after conversion of the three traceroutes. Each of the remaining AS-level paths forms a clause in our SAT formulation, with each observed AS acting as a literal. The truth value attached to the clause is *True* if the measurement detected its corresponding anomaly, and *False* otherwise. For example, if the AS-level path $X \rightarrow Y \rightarrow Z$ observed a blockpage, it is represented by the clause $(X_{block} \vee Y_{block} \vee Z_{block}) = T$.

Time- and URL-based splitting. Our formulation accounts for the fact that censorship policies and techniques may change over time (*e.g.,* Iran is known to increase censorship during political events such as elections [2]). Not doing so introduces the possibility of generating an unsolvable CNF in the event of a policy change – *e.g.,* the measurement $(X_{block} \vee Y_{block} \vee Z_{block}) = T$ is observed on Day 1 and $(X_{block} \vee Y_{block} \vee Z_{block}) = F$ on Day 2. We address this problem by creating CNFs at three time granularities – days, weeks, and months. Additionally, since not all URLs being tested are subject to censorship, we further restrict the CNFs to only include clauses containing measurements to a single URL. Therefore, we generate one CNF per URL per time granularity (day, week, and month). Finally, each CNF is solved using and an Python-based SAT solver, picosat [4]. In total we solve 34,298 CNFs after grouping by time, anomaly type, and destination URL. Each CNF had an average of 43 clauses and required an average of 17.41ms to find a solution.

3.2 Analyzing SAT solutions

Given a CNF, a SAT solver may return no solution, a single solution, or multiple solutions. When no solution is returned, there is no possible truth assignment to ASes that can satisfy the input CNF. These

scenarios may arise due to (1) noise in the ICLab measurements – *i.e.,* incorrect anomaly detection or path inference or (2) changing censorship policies within the specified time granularity. A single solution implies the presence of exactly one satisfying truth assignment to ASes. This ideal scenario allows us to exactly identify ASes that perform the measured censorship related anomalies (*i.e.,* the ASes that are assigned a *True* value in the solution). We label these ASes as *censoring ASes*. Finally, when the CNF does not contain enough clauses to generate a single solution, multiple satisfying assignments may be possible. When this situation arises, we consider every AS as a *potential censor* unless the literal associated with it is assigned a *False* value in all returned solutions.

3.3 Identifying censorship leakage

In order to prevent leakage of censorship (*i.e.,* where regional censorship policies impact users outside the region), censorship policies need to be implemented in ASes that are either stubs or provide transit services only for ASes within the region. To uncover instances of censorship leakage, we use the following approach: First, we only consider all AS-level paths used in CNFs that return exactly one solution. Next, we identify ASes that (1) are assigned a *False* truth value in the returned solution, (2) are located downstream from the identified censors (*i.e.,* closer to the vantage point being used by ICLab) in one of our censored measurements, and (3) are located in a different country from the censoring ASes in the CNF. We label these ASes as victims of censorship leakage due to their inheritance of censorship from censoring ASes in other countries. This inheritance occurs due to their traffic transiting through censoring ASes.

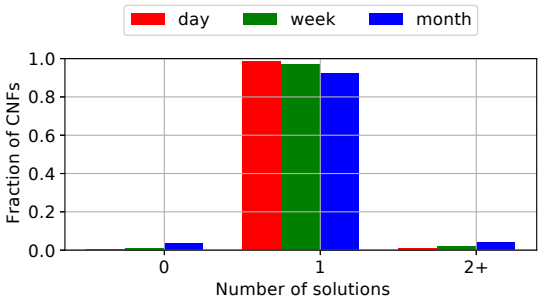
4 EXPERIMENTAL RESULTS

We focus on measuring (1) how often our approach is able to generate solvable SAT instances, (2) the amount of path churn observed and its impact on our SAT instance solvability, and (3) the ASes responsible for implementing and leaking censorship.

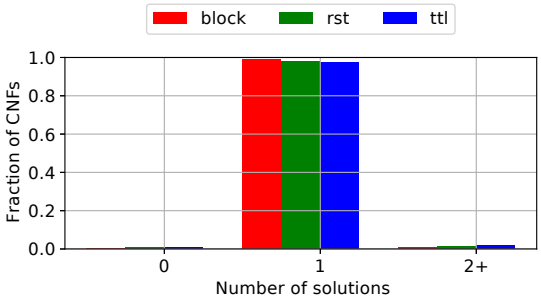
Satisfiability of generated CNFs. As discussed earlier, the CNFs generated by our approach may return (1) no solutions – indicative of changing censorship policies or noise in ICLab measurements, (2) exactly one solution – the ideal scenario, or (3) many solutions – indicative of insufficient number of measurements through diverse paths. In order to understand which scenario occurs most frequently, we analyze the results returned by our SAT solver for CNFs of different time granularities and anomaly types. We find that on average, nearly 97.9% of our CNFs return exactly one solution and less than 0.7% of our CNFs return no solution. This indicates high fidelity in our underlying data and highlights our ability to exactly identify censoring ASes.

Our results, when considering CNFs generated for different time granularities and anomalies, are illustrated in Figure 1. Figure 1a shows that as our CNF granularity becomes coarser, its solvability reduces. This is expected since (1) censorship policies are more likely to change and (2) we are more likely to include a noisy measurement in our CNF form when considering larger time periods.

Figure 1b shows that all anomaly types have high solvability. Due to the difficulty of differentiating between organic and injected RST packets, RST injection measurements from the ICLab platform



(a) By CNF granularity.



(b) By anomaly type.

Figure 1: Number of solutions found for constructed CNFs when split by CNF granularity and anomaly.

have low fidelity; however, we increase its fidelity by checking if there is blockpage or absence of HTML content in addition to the RST. Such conditions are highly indicative of censorship.

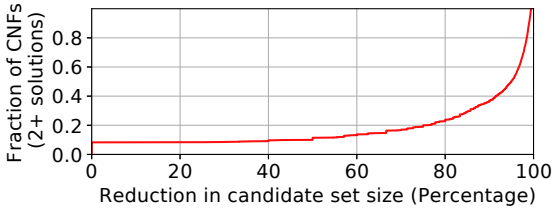


Figure 2: CDF of reduction in number of potential censors in CNFs with 2+ solutions.

Even if the CNFs generated by our approach yield more than one solution, they can be useful to identify ASes that *could not* have been responsible for implementing censorship – *i.e.,* ASes that were assigned a *False* truth value in every solution. We further investigate the 1.4% of scenarios where constructed CNFs yield more than one solution to identify the impact of this reduction. We find that in 8.2% of such cases, the solutions satisfying the CNF do not allow for any elimination of ASes as censors – *i.e.,* we are unable to narrow down the set of possible censors by eliminating definite non-censors. However, on average, 97% of all ASes in a CNF are identified as definite non-censors, leaving only 3% of the observed ASes as potential censoring ASes. Figure 2 shows that 50% of all generated CNFs with multiple solutions have nearly 95% of their ASes eliminated as potential censors.

Impact of path churn. Now, we measure the amount of network-level path churn observed over the course of our measurements

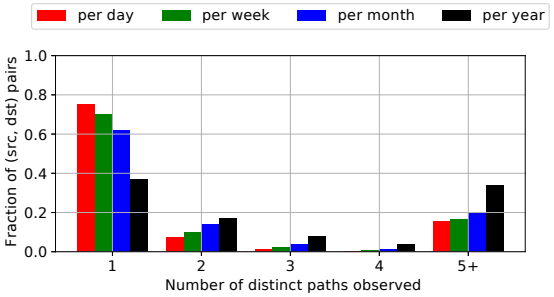


Figure 3: Number of distinct paths observed between a source and destination AS over varying time periods.

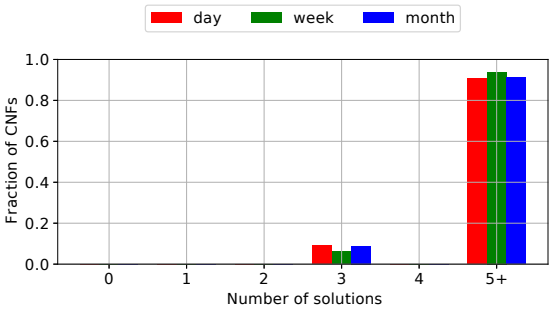


Figure 4: Number of solutions returned by CNFs in the absence of network-level path churn.

and consider the impact it has on the usability of our approach. In order to measure the amount of network-level path churn, we measure the number of distinct network-level paths observed between each source (ICLab vantage point) and destination URL for each day, week, month, and year. Figure 3 shows the fraction of these (source, destination) pairs that observe path changes over varying time periods. We find that nearly 25% of all pairs observe path changes within a single day. This fraction increases to 30%, 38%, and 67% when considering periods of one week, month, and year, respectively. Over the period of one year, 35% of the measured pairs recorded at least five distinct network-level paths. Using CAIDA's AS classification database [6], we found no significant differences in the amount of churn observed when considering specific classes of destination ASes (content, enterprise, or transit AS).

To understand the impact of network-level path churn on the effectiveness of our approach, we analyze the solvability of CNFs constructed in the absence of path churn. We eliminate the impact of path churn by only considering the measurements using the first observed distinct path between a source and destination in each CNF. Figure 4 illustrates the number of solutions returned by such CNFs. We see that nearly 90% of all CNFs return five or more solutions (compared to < 2% in the case of CNFs that include multiple distinct paths). We find that although less than 25% of all paths are impacted by path churn each day, the impact of these path changes on the solvability of the constructed CNFs is significant.

Uncovering censors and censorship leakage. We now analyze the censoring ASes identified by our approach. In total, we identify 108 censoring ASes located in 49 different countries. The countries with the most number of censoring ASes are reported in Table 2. Note that this does not represent the countries that are doing more censorship compared to others. The most commonly observed countries performing censorship are Iran and Cyprus in our results. While there are one and four censoring ASes identified in Iran and Cyprus respectively, we observe their censoring ASes 10 times and 3 times more than the total of all censoring ASes in the United Kingdom respectively.

In our analysis we observe a few regions implementing a wide array of censorship approaches. In particular, we find that censors in Russia, Iran, and India implement all measured censorship approaches. Using the McAfee URL categorization database [24], we find that URLs that are most commonly censored fall in Portal Sites, General News, and Business categories. Further analysis reveals that most ASes perform censorship exclusively on few categories of sites, with the exception of ASes in Cyprus and Iran which censor content across many different categories.

Region	Censoring ASes	Anomalies	Top URL categories
United Kingdom	AS35017, AS5413, AS62217, AS41678, AS9009, AS20860, AS8928, AS61317, AS39451	RST, TTL	Education/Reference, Business, Online Shopping
Germany	AS3320, AS20773, AS3257, AS51167, AS47447, AS33891, AS201011, AS24940	RST, TTL	Portal Sites, General News, Business
China	AS58461, AS4808, AS17621, AS9808, AS4812, AS4134, AS4837, AS37963	RST, TTL	Online Shopping, Auctions/Classifieds Streaming Media
Netherlands	AS1200, AS48684, AS12989, AS5580, AS50673	RST, TTL	Portal Sites, Software/Hardware, Business
Russia	AS35816, AS28840, AS21479, AS199669, AS42610	Block, RST, TTL	Gambling, Pornography, Internet Services

Table 2: Regions with most number of censoring ASes.

To understand the reasons behind our results, we classify censorship into four types based on whether the vantage point and destination are located in countries different from the location of a censor. The most commonly observed censors of each type are reported in Table 3. If the vantage point is outside of the region of the censor, this may represent censorship leakage (Type 1 and Type 2). The difference between these two types comes from the location of destination and the censor. When the censor and destination are located in the same country, we call it server-side filtering (Type 1). In the opposite case, the censor impacts a vantage point which is using the censor as a transit and we call it transit filtering (Type 2). Censorship where governments control policies only within their country are more common (Type 3 and Type 4). The censor might restrict access to specific content either outside of the country (Type 3) or inside it (Type 4).

Table 3 lists the most commonly observed censoring ASes and the ASes responsible for the largest number of censorship leaks (Type 1 and 2). While checking for censorship leakage, to ensure the correctness of our results identifying censors injecting block pages, we performed manual inspection of the recorded packet captures and HTML pages. This manual verification could not be performed for TTL and RST anomalies, however. This is because TTL/RST anomalies occur when the censoring AS emulates the destination

Type	AS	Region	Anomaly	Leaks (AS)	Leaks (Country)	Victims of leakage (Country)	Top URL categories
Type1	AS4134	China	RST, TTL	4	4	South Korea, Germany, Sweden	Online Shopping, Software/Hardware
	AS26615	Brazil	RST, TTL	1	1	United States	Portal Sites
	AS4812	China	RST, TTL	7	5	Singapore, Uganda, United States	Online Shopping, Auctions/Classifieds
Type2	AS4637	Hong Kong	RST, TTL	2	2	United States, Malaysia	General News, Internet Services
	AS1299	Sweden	RST, TTL	16	9	United States, Ukraine, Singapore	Portal Sites, Social Networking
	AS10026	Japan	RST, TTL	1	1	United States	Education/Reference, Online Shopping
Type3	AS48434	Iran	RST, TTL, Block	-	-	-	General News, Internet Services
	AS5384	Emirates	Block	-	-	-	Pornography, Gambling
	AS38001	Singapore	RST, TTL	-	-	-	General News, Internet Services, Web Ads
Type4	AS201011	Germany	RST, TTL	-	-	-	General News
	AS4808	China	RST, TTL	-	-	-	General News
	AS42610	Russia	Block	-	-	-	Potential Illegal Software

Table 3: Classification of most commonly observed censors and the number of countries/ASes impacted by censorship leakage.

using its IP and port number, and thus cannot be confirmed beyond the detection method used to identify them in the first place.

Table 3 also shows a subset of six censoring ASes among a total of 32 ASes that leak their censorship policies to other ASes (Type 1 and Type 2 censorship). To identify where these 32 censoring ASes leak censorship, we record the regions for which they provide downstream transit (§3.3). We also show six ASes among a total of 81 that do not leak censorship (Type 3 and Type 4 censorship).

5 LIMITATIONS

The presented techniques to identify censors and censorship leakage leverage (1) datasets such as censorship measurements from the ICLab platform, the Maxmind geolocation database, and Team Cymru's IP-to-AS mapping, (2) VPN vantage points, and (3) network-level path churn. As a consequence, the applicability of our techniques and the accuracy of the presented results are limited by their fidelity and availability.

Our reliance on VPNs to serve as vantage points for measurements requires that the end points of the VPNs are actually located in the countries that they claim to be in. However, it is possible for VPN providers to use IP addresses from a range of claimed countries while maintaining actual servers in only some of them. Thus, their vantage points might be located in another country, not in the claimed location. This may result in incorrect associations of censorship policies with regions. Currently we address this problem by utilizing VPN end points from multiple commercial VPN providers and working under the assumption that at least one of them has servers in the region that they claim. In concurrent work, we are working to verify the location claims of popular VPN providers.

In addition to the uncertainty of VPN end point locations, our results are also impacted by the fidelity of ICLab's anomaly detection techniques. Poor fidelity of the anomaly detection techniques are likely to result in unsolvable SAT formulations and as a result limit the identification of censors and censorship leakage. In this paper, we exclude low fidelity anomalies that are likely to be attributed to non-censorship events (*e.g.,* sequence number anomalies).

Finally, our technique relies on the occurrence of network-level path churn between two end points. The high path stability from a certain vantage point to a destination server limits the chances of identifying censors along the stable paths.

6 CONCLUSIONS

In this work, we leveraged boolean network tomography and censorship measurements obtained from the ICLab censorship measurement platform to identify the ASes responsible for inducing censorship related anomalies on the Internet. Our results show that even in the absence of strategically selected monitors and vantage points, exact identification of censoring ASes is possible due to network-level path churn. Our approach uncovered 108 censoring ASes located in 49 different countries of which 32 were found to leak censorship into other networks and 18 into other countries. In cases where exact identification of censors was not possible, we were able to reduce the number of potential censoring ASes by 97%.

The results obtained in this work also uncover the need to improve the fidelity of the several anomaly detection techniques used by ICLab and traceroutes gathered by the platform. In addition to improving the robustness of ICLab measurements (*e.g.,* by using tools such as InTrace [29] in conjunction with standard traceroutes), we also plan to use our approach to extend the ICLab censorship measurement platform in several ways. Specifically, we plan to (1) incorporate data obtained from external performance measurement datasets (*e.g.,* data from M-Lab [22]) to identify ASes responsible for throttling the bandwidth made available to specific protocols used for censorship circumvention and (2) identify, at scale, the ASes responsible for blocking access to Tor bridges [14].

ACKNOWLEDGEMENTS

We would like to thank our shepherd Dr. Zied Ben-Houidi for his comments and help in preparing the camera-ready version of this work. We are also thankful to the anonymous reviewers for their feedback and the ICLab development team for their work on the platform that enabled the data for this work.

We acknowledge funding support from the NSF Grants CNS 1350720, CNS 1518845. We also acknowledge partial support by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the "ICT Consilience Creative Program" (IITP-2015-R0346-15-1007) supervised by the IITP (Institute for Information & communications Technology Promotion). The opinions in this paper are those of the authors and do not necessarily reflect the opinions of a sponsor or the United States and Republic of Korea Government.

REFERENCES

[1] Anonymous. 2014. Towards a Comprehensive Picture of the Great Firewall's DNS Censorship. In *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)*. USENIX Association, San Diego, CA.

[2] Simurgh Aryan, Homa Aryan, and J. Alex Halderman. 2013. Internet Censorship in Iran: A First Look. In *Presented as part of the 3rd USENIX Workshop on Free and Open Communications on the Internet*. USENIX, Washington, D.C.

[3] Iljitsch Van Beijnum. 2010. China censorship leaks outside Great Firewall via root server. https://arstechnica.com/tech-policy/2010/03/china-censorship-leaks-outside-great-firewall-via-root-server/. (2010). Online; accessed June 2017.

[4] Armin Biere. 2008. PicoSAT essentials. *Journal on Satisfiability, Boolean Modeling and Computation* 4 (2008), 75–97.

[5] Martin A Brown. 2008. Renesys blog: Pakistan Hijacks YouTube. http://dyn.com/blog/pakistan-hijacks-youtube-1/. (2008). Online; accessed June 2017.

[6] CAIDA. 2017. AS Classification. http://www.caida.org/data/as-classification/. (2017). Online; accessed June 2017.

[7] CAIDA. 2017. IPv4 Routed /24 AS Links Dataset. http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml. (2017). Online; accessed June 2017.

[8] Mark Coates, Alfred O. Hero III, Robert Nowak, and Bin Yu. 2002. Internet tomography. *IEEE Signal Processing Magazine* 19, 3 (May 2002), 47–65.

[9] Mark Coates and Robert Nowak. 2000. Network loss inference using unicast end-to-end measurement. In *ITC Conference on IP Traffic, Modeling and Management*. 28–1.

[10] Jakub Dalek, Bennett Haselton, Helmi Noman, Adam Senft, Masashi Crete-Nishihata, Phillipa Gill, and Ronald J. Deibert. 2013. A Method for Identifying and Confirming the Use of URL Filtering Products for Censorship. In *Proceedings of the 2013 Conference on Internet Measurement Conference (IMC '13)*. ACM, New York, NY, USA, 23–30.

[11] Amogh Dhamdhere, Renata Teixeira, Constantine Dovrolis, and Christophe Diot. 2007. Netdiagnoser: Troubleshooting network unreachabilities using end-to-end probes and routing data. In *Proceedings of the 2007 ACM CoNEXT conference*. ACM, 18.

[12] Nick Feamster, David G Andersen, Hari Balakrishnan, and M Frans Kaashoek. 2003. Measuring the effects of Internet path faults on reactive routing. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 31. ACM, 126–137.

[13] Anja Feldmann, Olaf Maennel, Z Morley Mao, Arthur Berger, and Bruce Maggs. 2004. Locating Internet routing instabilities. *Proceedings of the 2004 conference on SIGCOMM*, 205–218.

[14] David Fifield and Lynn Tsai. 2016. Censors' Delay in Blocking Circumvention Proxies. In *6th USENIX Workshop on Free and Open Communications on the Internet (FOCI 16)*. USENIX Association, Austin, TX.

[15] Arturo Filasto and Jacob Appelbaum. 2012. OONI: Open Observatory of Network Interference.. In *FOCI*. 2nd USENIX Workshop on Free and Open Communications on the Internet.

[16] Freedom House. 2016. Freedom on the Net 2016. https://freedomhouse.org/report/freedom-net/freedom-net-2016. (2016). Online; accessed June 2017.

[17] Umar Javed, Italo Cunha, David Choffnes, Ethan Katz-Bassett, Thomas Anderson, and Arvind Krishnamurthy. 2013. Poiroot: Investigating the root cause of interdomain path changes. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. ACM, 183–194.

[18] Ben Jones, Tzu-Wen Lee, Nick Feamster, and Phillipa Gill. 2014. Automated detection and fingerprinting of censorship block pages. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM, 299–304.

[19] Ethan Katz-Bassett, Colin Scott, David R Choffnes, Ítalo Cunha, Vytautas Valancius, Nick Feamster, Harsha V Madhyastha, Thomas Anderson, and Arvind Krishnamurthy. 2012. LIFEGUARD: Practical repair of persistent route failures. *ACM SIGCOMM Computer Communication Review* 42, 4 (2012), 395–406.

[20] Sheharbano Khattak, Mobin Javed, Philip D. Anderson, and Vern Paxson. 2013. Towards Illuminating a Censorship Monitor's Model to Facilitate Evasion. In *Presented as part of the 3rd USENIX Workshop on Free and Open Communications on the Internet*. USENIX, Washington, D.C.

[21] Philip Levis. 2012. The collateral damage of internet censorship by dns injection. *ACM SIGCOMM Computer Communication Review* 42, 3 (2012), 21–27.

[22] M-Lab. 2017. M-Lab Data. Overview. https://www.measurementlab.net/data/. (2017). Online; accessed June 2017.

[23] Liang Ma, Ting He, Ananthram Swami, Don Towsley, and Kin K Leung. 2017. Network capability in localizing node failures via end-to-end path measurements. *IEEE/ACM Transactions on Networking* 25, 1 (2017), 434–450.

[24] McAfee. 2017. Customer URL Ticketing System. www.trustedsource.org/en/feedback/url?action=checklist. (2017). Online; accessed June 2017.

[25] Zubair Nabi. 2013. The Anatomy of Web Censorship in Pakistan. In *Free and Open Communications on the Internet*. USENIX. http://censorbib.nymity.ch/pdf/Nabi2013a.pdf

[26] OONI. 2017. Open Observatory of Network Interference. https://ooni.torproject.org/. (2017). Online; accessed June 2017.

[27] Dan Pei, Matt Azuma, Dan Massey, and Lixia Zhang. 2005. BGP-RCN: Improving BGP convergence through root cause notification. *Computer Networks* 48, 2 (2005), 175–194.

[28] Abbas Razaghpanah, Anke Li, Arturo Filastò, Rishab Nithyanand, Vasilis Ververis, Will Scott, and Phillipa Gill. 2016. Exploring the Design Space of Longitudinal Censorship Measurement Platforms. *CoRR* abs/1606.01979 (2016). http://arxiv.org/abs/1606.01979

[29] Robert Swiecki. 2017. Enumeration of IP hops using existing TCP connections. https://github.com/robertswiecki/intrace. (2017). Online; accessed June 2017.

[30] Yixin Sun, Anne Edmundson, Laurent Vanbever, Oscar Li, Jennifer Rexford, Mung Chiang, and Prateek Mittal. 2015. RAPTOR: Routing Attacks on Privacy in Tor.. In *USENIX Security Symposium*. 271–286.

[31] Renata Teixeira and Jennifer Rexford. 2004. A measurement framework for pin-pointing routing changes. In *Proceedings of the ACM SIGCOMM workshop on Network troubleshooting: research, theory and operations practice meet malfunctioning reality*. ACM, 313–318.

[32] Yolanda Tsang, Mark Coates, and R.D. Nowak. 2003. Network Delay Tomography. *Trans. Sig. Proc.* 51, 8 (Aug. 2003), 2125–2136.

[33] Yehuda Vardi. 1996. Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data. *J. Amer. Statist. Assoc.* 91, 433 (1996), 365–377.

[34] Nicholas Weaver, Robin Sommer, and Vern Paxson. 2009. Detecting Forged TCP Reset Packets.. In *16th Network and Distributed System Security Symposium (NDSS2009)*. Internet Society.

[35] Jian Wu, Zhuoqing Morley Mao, Jennifer Rexford, and Jia Wang. 2005. Finding a needle in a haystack: Pinpointing significant BGP routing changes in an IP network. In *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2*. USENIX Association, 1–14.