Free and Open
COMMUNICATIONS
https://foci.community on the Internet

Advancing the Art of Censorship Data Analysis

Ram Sundara Raman
University of Michigan

Apurva Virkud
University of Michigan

Sarah Laplante
Google Jigsaw

Vinicius Fortuna
Google Jigsaw

Roya Ensafi
University of Michigan

Free and Open Communications on the Internet 2023(1), 14-23

Advancing the Art of Censorship Data Analysis

Ram Sundara Raman* Apurva Virkud* Sarah Laplante† Vinicius Fortuna† Roya Ensafi*
*University of Michigan †Google Jigsaw

Abstract

A decade of research into collecting censorship measurement data has resulted in the introduction and continued operation of several censorship measurement platforms that collect large-scale, longitudinal censorship data. However, collecting data is only part of the process of understanding Internet censorship phenomena; interpreting this data requires a large amount of effort in data analysis, including removing false positives, adding information from external sources, and exploring aggregated data. The lack of a standardized data analysis process that performs such operations leads to incomplete and inaccurate characterizations of censorship.

In this work, we present a detailed breakdown of the challenges involved in analyzing censorship measurement data, supported by examples from public censorship datasets such as OONI and Censored Planet. The key challenges identified in this paper encompass finding accurate measurement metadata, and accounting for unexpected causes of network interference other than Internet censorship, and we highlight findings from previous work that suffer from these challenges. To address these challenges, we design and implement an open-source data analysis pipeline for a currently active censorship measurement platform, Censored Planet, and motivate and validate each component of the pipeline by demonstrating censorship case studies that can be accurately characterized using the pipeline. We hope that our paper sheds light on the complexity of censorship data analysis and brings systematization to the process.

1 Introduction

Internet traffic is increasingly being disrupted, tampered with, and monitored by governments, ISPs, advertisers, and other actors. Advances in censorship technology and recurring instances of censorship events all over the world [2, 32, 43] have necessitated high-quality, large-scale Internet censorship data that can help researchers, journalists, policymakers, and advocacy groups characterize censorship mechanisms and ensure accountability for censoring authorities. Thus far, most community efforts and previous work has focused on building tools that can collect representative censorship measurement data with good coverage over time and space [4, 11, 36, 43]. However, collecting censorship measurement data is only part of the process of understanding Internet censorship. Parsing, analyzing, and exploring censorship measurement data is complex because *(1)* the Internet's vast size, number of stakeholders, and overall routing complexity make it difficult to characterize what happens to users' traffic as it travels to a destination, even in the absence of an adversary; *(2)* network intermediaries are powerful actors whose capabilities are not fully known; *(3)* some intermediaries hide their actions from existing measurement and monitoring techniques, which cannot detect stealthy behavior; and *(4)* researchers cannot reliably collect ground truth on the counterfactual traffic that would exist without manipulation, making it difficult to calibrate measurements and to attribute findings to specific actors.

Because of this complexity, analyzing censorship measurement data requires a large amount of effort in order to remove false positives, add information from external data sources, and explore aggregated data. For example, researchers using censorship measurement data have to account for CDN localization effects causing measurements to behave unexpectedly [41, 44]. Finding accurate metadata information (such as Autonomous System and geolocation mappings) from external sources is also a challenge, as data sources have been known to contain inaccuracies [24]. So far, such analysis has been performed in an ad-hoc and case-by-case basis, and as we show in this study, this can cause inaccuracies in the reporting of censorship outcomes. The lack of a standardized data analysis workflow that overcomes challenges in data analysis prevents researchers, including domain experts, from accurately characterizing censorship phenomena, and introduces inaccuracies in the reporting of censorship phenomena. In an area where results can have far-reaching implications, it is crucial that the analysis and interpretation of censorship data are performed accurately.

In this paper, based on our experience of working with censorship measurement data over ten years, we present a detailed breakdown of the key challenges involved in analyzing censorship measurement data, using motivating examples from previous work and public data provided by censorship measurement platforms such as OONI [36] and Censored Planet [43]. We highlight several critical steps in the analysis process that are often overlooked by researchers, including finding accurate and representative measurement metadata, accounting for unexpected factors such as Internet shutdowns, server-side blocking, and CDN localization, and accurately interpreting and presenting results.

Based on the identified challenges, we design and implement an open-source iterative data analysis pipeline for data produced by Censored Planet [14]. The pipeline completely separates the analysis process from the measurements themselves, allowing the analysis process to benefit from new and improved methods. The pipeline enables parallel processing of all Censored Planet data in less than 24 hours, accounting for more than 6 terabytes of 65 billion measurement data points collected over 46 months, and produces analyzed data for exploration in near real-time. The data analysis process involves adding metadata from a variety of data sources including CAIDA [10, 12], DB-IP [19], and Censys [20], processing control measurements and page fingerprints to identify unexpected responses, and mapping measurements to human-readable outcomes. We showcase several interesting cases of censorship phenomena that can be easily and accurately characterized using the data analysis pipeline, such as changes in censorship mechanisms and detection of commercial firewalls performing DNS and HTTP blocking.

By open-sourcing our analysis pipeline [14], we aim to improve the state of censorship detection and characterization, and help the censorship measurement community adopt similar best practices and improve the quality of reports on Internet censorship. We conclude the paper with important open challenges that warrant attention from the research community.

2 Background and Related Work

In this paper, we define "network censorship" as the phenomenon through which a network intermediary restricts access to specific content on the Internet for a user. A censor might inhibit communication in different stages of a network connection. A censor may interfere with the DNS resolution process, either preventing a client from obtaining an IP address, or providing a client with the wrong IP address for a domain [5, 27, 38]. A censor may also prevent a client from establishing a transport-layer (e.g. TCP) or application-layer (e.g. HTTP, HTTPS, FTP) connection with a server based on visible content exchanged during the connection by dropping or injecting packets [3, 39, 42, 44–46].

There have been a plethora of reports, news, and measurement studies that show an increasing trend in the censorship of different types of websites, mobile applications, and Internet protocols by many actors around the world [7, 9, 21, 26, 29, 36, 39, 43, 47, 50, 51]. Influenced by these events, there is an increasing interest in collecting and analyzing censorship measurement data. Addressing this need, a number of censorship measurement platforms, complementary to each other, have been developed to collect valuable data on website censorship in countries around the world. The following are some active censorship measurement platforms with longitudinal open-access data on content-based censorship:

- **OONI.** The Open Observatory of Network Interference specializes in direct measurements from volunteer devices [36]. Their open source data collection software, *OONI Probe*, is designed to measure various forms of Internet censorship. OONI obtains informed consent from volunteers, reports measurements at the AS level to avoid risk to volunteers, and the data they collect is automatically processed and published on the OONI website [36].

- **Censored Planet.** Censored Planet specializes in remote measurements to thousands of public infrastructural machines on the Internet (e.g. routers, open DNS resolvers, and webservers) and infers censorship based on responses received from these machines [43]. Censored Planet collects measurements on 6 Internet protocols (DNS, TCP, Echo, Discard, HTTP, and HTTPS) to test reachability to around 2,000 popular and sensitive websites on a bi-weekly basis, and the data collected is published on the Censored Planet website [13].

- **ICLab.** The Information Controls Lab specializes in direct measurements using VPN servers available in different countries [4].

- **GFWatch.** GFWatch measures the DNS filtering performed by the Great Firewall of China longitudinally [27] using direct measurements from inside China, and the data collected is available on the GFWatch website [23].

The goals of these censorship measurement platforms have been to simplify the process of data collection and provide easily accessible data. Arriving at this stage has required a decade of effort, and there is now large-scale censorship data available for researchers to quickly investigate questions related to censorship. In recent years, many research studies investigating specific censorship phenomenon have used data from these measurement platforms [8, 31, 32, 37, 39, 42, 44, 49]. In this paper, we use observations from these previous work and publicly available data from these platforms to highlight key challenges in data analysis.

3 Challenges in Analysis

Accurately characterizing Internet censorship is a multi-step, complex process, starting from a research question (e.g. "Is social media blocked in Belarus?") to arriving at processed data that can provide a clear answer to the research question that supports a particular theory ("Facebook and Twitter are blocked in Belarus").

Overall, there are three general parts to characterizing Internet censorship: (1) The *Data Collection* step involves collecting Internet measurement data using established methods that trigger censorship. (2) The *Data Analysis* step augments the collected data with new features and processes the data to remove noise (3) Finally, the *Data Exploration* step involves aggregating the data and extracting insights. In this paper, we focus on improving and standardizing the *Data Analysis* step. We separate our analysis process from the data collection itself, since the data analysis process can be iteratively improved while the data collected is immutable and cannot be retroactively obtained. However, insights from the data analysis could be used to perform better measurements in the future.

3.1 Data Limitations

In order to create representative insights, the analysis process needs to consider the continuity, coverage, and scale aspects of the collected data. Analysis methods working on large-scale, longitudinal data need to consider whether the data has been collected from multiple ISPs in a country and whether the same websites have been tested frequently in the same networks. Some measurement methods (such as those employed by OONI) perform tests on different protocols sequentially, and this could lead to inaccurate analysis of censorship systems that may block access to websites at different levels of the network stack [9, 15].

In a specific case, previous work by Padmanabhan et al. [37] investigates blocking of popular social media websites in Myanmar between February 2021 and April 2021. The authors report that ISPs in Myanmar use TCP/IP blocking and DNS blocking selectively, with some measurements experiencing DNS blocking and others TCP/IP blocking (See Figure 4 in [37]). However, we find that ISPs in Myanmar *apply both types of blocking concurrently* rather than selectively. Closer inspection of the data suggests that the difference was due to certain volunteers bypassing the DNS tampering by using public DNS resolvers such as Cloudflare Public DNS and Google Public DNS, and thereafter experiencing IP blocking [34, 35]. Considering this effect, ideally, measurements using public DNS resolvers should be analyzed and reported separately, and we adopt this approach with our analysis pipeline.

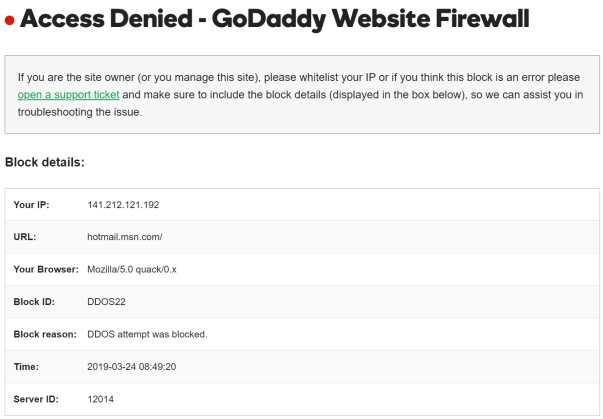

Figure 1: **A GoDaddy CDN hosting server flagging Censored Planet measurements as a DDoS attempt.**

3.2 Metadata Limitations

Extending Internet measurement data with accurate metadata has been a longstanding problem for the Internet measurement community, but the issue becomes even more relevant in censorship data analysis, where incorrect conclusions can have drastic consequences. Since censorship policies are frequently implemented at the ISP or AS level [16, 39], it is crucial that censorship measurement data is annotated with accurate AS information, including traffic volumes which can indicate the impact of censorship. The organization that the IP belongs to is also an important feature to consider apart from AS information, since blocking may be organization-specific. For example, blocking found in a small corporate network does not have the same effect as blocking found across a large residential ISP, and blocking policies may vary among them. However, we find that previous work frequently reports results at the country-level and ignores AS traffic volumes or IP organizations [4, 43].

Moreover, we also observe that other metadata such as categories of websites, blockpage and middlebox fingerprints, and ground truth information are crucial in removing false positives, confirming censorship, and characterizing censorship accurately. An iterative data analysis pipeline such as the one proposed later in this paper can enable constant improvements to metadata added to measurements (refer §4.1).

3.3 Unexpected Network Interference

We observe that censorship measurement data frequently contains instances of website unreachability caused due to factors other than network censorship, and this leads to misinterpretation of results. We highlight three major sources of unexpected network interference, and show why it is crucial that these factors are considered by an analysis pipeline.

3.3.1 Accounting for CDN and hosting configurations

An increasing number of websites are hosted on Content Delivery Networks (CDN), taking advantage of the benefits of localization, load balancing, caching, and protection against DDoS attacks [25, 41]. However, CDN configurations affect censorship measurement datasets and lead to unexpected observations that can be easily misconstrued as censorship without the presence of a standardized analysis process. For example, Cloudflare and Godaddy may block Internet measurements because of DDoS concerns or low IP reputation and inject an "Access Denied" page (see Figure 1) [30, 44].

Measurement methods may also result in unexpected results due to customized CDN configurations. Censored Planet's Hyperquack measurements send HTTP requests for a test domain to a random web server, expecting the web server to respond with an error page (e.g. 404 Not Found errors) [44]. Any deviation from this expected error is often indicative of censorship. This method fails when trying to send measurements to a web server in the Akamai network when the test domain is also hosted by Akamai. Because of Akamai's edge configuration, these measurements end in either a connection timeout or an HTTP status 301 Moved Permanently. Previous work, such as that in [43], have not accounted for cases where test domains and web servers are both hosted on Akamai, leading to an over-estimation of censorship.

To avoid such problems, a few studies have conservatively flagged CDN responses as benign [4, 38, 41]. However, this naive approach may lead to under-reporting censorship. For example, ISPs in China resolve DNS responses of blocked websites to popular CDN IP addresses including those of Facebook and Twitter [6]. There are also cases where blockpages are hosted on CDN IPs [49]. Therefore, considering all CDN responses as benign may lead to false negatives.

Individual websites may also have localization features that cause inconsistencies. Hence, previous work using IP address, ASN and content matching suffer from false positives [38, 43]. For example, `match.com` redirects users automatically based on geolocation to various sub-sites with different content and IPs. For instance, accessing `match.com` from the UK will redirect the user to `uk.match.com`. Additionally, `match.com` resolves to an IP hosted in Match Group's business AS, while `uk.match.com` is hosted on a separate European network. Thus, if DNS measurements for `match.com` from the US and UK are compared, the IP address returned, the ASN of the IP address, and the content of the TLS and HTTP responses, which are hueristics used by previous work [38, 43], would be completely different.

All of the above examples show that it is important to consider the effects of CDNs and hosting configurations in censorship data analysis, especially when the method involves comparing measurements with each other. We account for this in our analysis pipeline by using control measurements and blockpage fingerprints (refer §4.2).

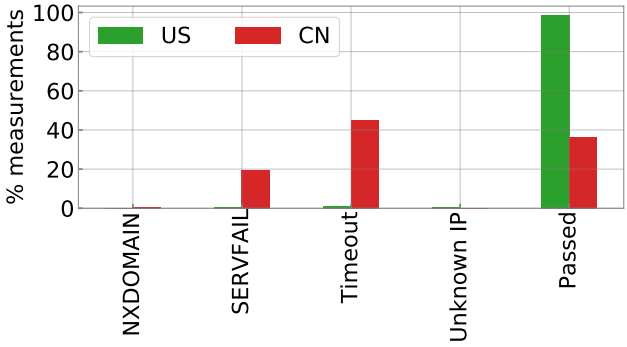

Figure 2: **DNS responses for `.gov` and `.mil` domains in US and CN**—A number of DNS resolutions fail in CN due to SERVFAIL and Timeout errors caused by geoblocking.

3.3.2 Server-side blocking

Server-side blocking is the phenomenon where websites restrict access to users by using features of the source IP address. A common form of server-side blocking is geoblocking, where websites restrict access to users from certain countries [30]. While it is uncertain whether server-side blocking should be considered censorship, the presence of server-side blocking in censorship measurement data may lead to incorrect conclusions regarding Internet freedom in a particular country or region.

For example, Figure 2 shows the outcomes of Censored Planet DNS measurements [38, 41] of 75 domains with *.gov* and *.mil* TLDs on April 11, 2021. From measurements in the United States, 98.35% resolved to the correct IP address. From measurements in China, only 36.06% resolved correctly. Importantly, 19.06% of measurements in China failed with the SERVFAIL DNS code, which has been shown previously to be caused by the US-based nameservers of these websites blocking access from recursive resolvers in China [40]. However, previous studies such as [32, 43, 49] which do not account for geoblocking would consider such cases as DNS failures, leading to an over-estimation of DNS blocking in China. Reports using OONI data [33] showcase the same issue. Thus, the analysis process needs to consider the source of network errors.

3.3.3 Internet shutdowns

There has been an increase in government-directed Internet shutdowns [1, 2, 28], as well as those caused by natural disasters or ISP outages. These events influence data collected by censorship measurement platforms and may lead to false attribution of website censorship in cases where control measurements are not performed or considered for analysis, as we show later in §4.2.

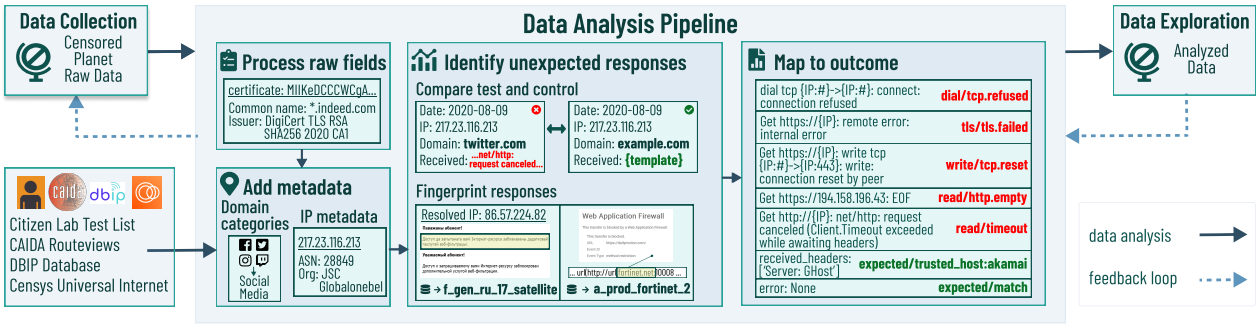

Figure 3: **Data Analysis Pipeline**—The design of our iterative censorship data analysis pipeline, which performs steps such as adding metadata fields, applying fingerprints, and mapping measurements to outcomes.

4 Data Analysis Pipeline

To resolve the challenges laid out in §3, we build an iterative data analysis pipeline for data produced by Censored Planet measurements. The pipeline includes crucial data analysis steps that have been overlooked in prior research. An overview of our data analysis pipeline is shown in Figure 3. The pipeline first parses measurement-specific data (e.g., TLS certificates), and adds metadata fields. Next, the pipeline compares test measurements against control measurements, applies blockpage and non-censorship (e.g., geoblocking) fingerprints to unexpected responses, and maps each measurement to an outcome; these steps reduce the effects of unexpected network interference. Some of the important design features of the pipeline are:

- *Measurements vs Analysis:* It completely separates the analysis process from the measurement itself, providing the ability to introduce new analysis methods that can even improve data collected in the past.

- *Efficiency:* The data analysis pipeline is able to process all of Censored Planet's data sources (over 46 months and 6 Terabytes of 65 billion measurement data points) in less than 24 hours, providing the ability to propagate changes to the data rapidly.

- *Modular:* New metadata and analysis processes are easy to add, and the pipeline can be used incrementally on a subset of the data, enabling the production of analyzed data in near real-time.

Our implementation of the data analysis pipeline is based on Apache Beam and is completely open source [14] enabling the community to process data from Censored Planet. While the pipeline we describe in this study is specific to Censored Planet data, the analysis process and insights from our pipeline are generally applicable to other censorship measurement platforms such as OONI and ICLab. We motivate and describe each step of the pipeline and demonstrate how the steps addresses the challenges discussed in §3 through examples from the Censored Planet data.

Table 1: **Blocking of COVID-19 related websites [49] and APNIC traffic volume [22] in Canada (2020).**

ASN	Name	Block?	APNIC Rank	% of traffic
577	Bell Canada	Yes	1	18.33
812	Roger Communications	Yes	2	14.22
852	Telus Communications	Yes	3	12.08
5769	Videotron Telecom Lte	No	4	10.64
6327	Shaw Communications	Yes	5	3.1
...	...	...	...	...
376	Reseau d'informations scientifiques du Quebec	Yes	70	0.07
62969	Allen business Communications	Yes	177	0.01
17001	University of Manitoba	Yes	N/A	N/A
14472	Roger Communications	Yes	N/A	N/A

4.1 Adding Metadata

In order to contextualize censorship measurements, we need metadata for the domains, IP addresses, and responses, as shown in §3.2. The pipeline augments information from multiple sources immediately after the measurements are published, including the domain category from Citizen Lab [17], and IP metadata from CAIDA, DBIP, and Censys [12, 19, 20] as shown in Figure 3. This IP metadata consists of geolocation, AS information (name, number, class, volume), IP organization, and HTTP body and TLS certificate data.

Case Study: AS Traffic Volumes We highlight a case where an analysis process where the AS information added by our pipeline enables more accurate reporting compared to previous work. Table 1 shows the ASes (and their traffic percentage estimates) in Canada where Vyas et al. recently used Censored Planet data to analyze the blocking of

COVID-related websites categorized as malware [49]. Our pipeline supplements the data with APNIC's AS traffic volume dataset [22], which clearly shows that while the three largest end-user ISPs in the country all observed blocking, many of the networks in which blocking was found are small and belong to universities or corporations. Thus, it is important to provide context about AS traffic volumes by including this data in the analysis.

Case Study: IP Organizations We find that IP organization metadata can be useful to clarify mixed censorship signals within a region. For example, all Hyperquack HTTP measurements for the VPN service `www.hotspotshield.com` in AS 24835 (Vodafone Data) in Egypt indicates blocking on June 16, 2022. However, we observe that some requests experience TCP resets while others observe packet drops. After incorporating the *IP organization*, we find that one organization (Oratech) was responding with TCP resets and the others allowed requests to time out. This difference suggests that the censorship is implemented at an organizational level. We find that such IP metadata is especially important in countries with decentralized censorship policies such as India [53].

Case Study: TLS Certificates We also find that TLS certificate metadata is very useful in accurately detecting censorship, not only in HTTPS measurements, but also as follow-up measurements to DNS queries. We find the presence of DNS filtering products returning poisoned IP addresses that issue certificates which contains the vendor name in the certificate's Common Name field. For instance, we find DNS filtering product *Sky DNS* issuing certificates for blocked domains in Russia, Ukraine, and Kazakhstan, and Safe DNS issuing certificates for blocked domains in the United States, Australia and Netherlands. Our investigation shows that the metadata added by our pipeline can not only accurately detect censorship, but can also help in attributing censorship.

4.2 Identify Unexpected Responses

The pipeline uses Censored Planet's control measurements to compare and identify test measurements that do not behave as expected. The goal is to differentiate censorship from other sources of network interference, including those discussed in §3.3. Any measurements where the control measurement failed are not marked as censorship.

If the control measurement succeeds, and the test measurement fails because of a mismatch between the control measurement response and a test measurement response (i.e. not due to a network error), this indicates an *unexpected response*, either from a network intermediary conducting intentional blocking or from the vantage point IP address itself under measurement. Aside from blocking, unexpected responses could also result from CDN configurations and server-side blocking, as described in §3.3. To add more context and differentiate

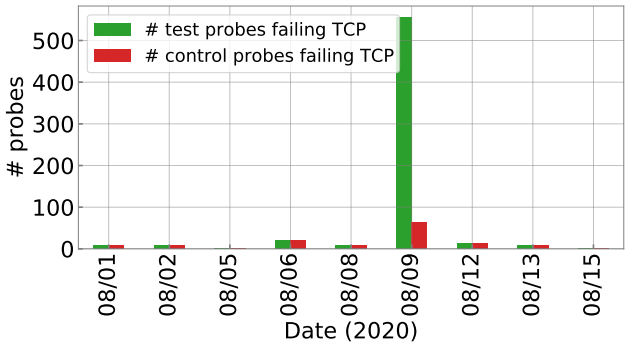

Figure 4: **TCP handshake failures in Censored Planet's Quack Echo measurements in Belarus**—At the start of the Belarus Internet shutdown on August 9, 2020, a large number of Censored Planet probes to Belarus fail to establish a TCP handshake.

these cases, the pipeline checks the responses against a set of fingerprints corresponding to blockpages and non-censorship cases such as geoblocking and bot detection [30]. We use fingerprint datasets from previous work [44] and manual investigation to build and maintain our fingerprint database, which contains HTML patterns that match with known webpages. Although maintaining these fingerprints requires manual effort and only presents a lower bound of confirmation, we find that a large percentage of responses can be confirmed as either a true blockpage or a known non-censorship case using our fingerprints. For instance, more than 60.89% of all data with HTTP responses in Censored Planet's four years of HTTP measurements match with a fingerprint. The fingerprints we develop are completely open-sourced, and we hope to engage the censorship measurement community to crowd-source and better maintain our fingerprint database by updating new signs of blocking.

Case Study: Internet Shutdowns We illustrate the importance of using control measurements to account for Internet shutdowns using Censored Planet Echo measurements during the Belarus Internet shutdown of August 2020 [46, 52]. On the first day of the shutdown (August 9, 2020), there is an increase of two orders of magnitude in the number of test measurements failing during the TCP connection stage (see Figure 4). These failures could be easily misinterpreted as website censorship, however they are caused by measurements failing due to the shutdown. To avoid this, accounting for control measurements that are expected to complete successfully is necessary. Besides the high number of failed TCP connections in test measurements, there was also an order of magnitude increase in failed control measurements on the day of the shutdown, showing that measurements are failing due to reasons other than censorship.

Table 2: **Outcomes per stage for Hyperquack HTTP data from January 2022 to September 2022**—The total number and percentage of measurements matching each outcome is shown.

Stage	Outcome	Num. Measurements	% Measurements
Expected Response (No Blocking)	expected/match	1,772,014,793	94.45%
	expected/akamai	61,943,574	3.30%
Content Mismatch	content/known_not_censorship	16,642,905	0.89%
	content/status_mismatch	13,533,254	0.72%
	content/known_blockpage	743,396	0.04%
	content/body_mismatch	65,577	0.004%
	content/header_mismatch	34,837	0.002%
Read/Write Failure	read/timeout	6,356,637	0.34%
	read/tcp.reset	4,273,880	0.23%
	read/http.empty	180,309	0.01%
	http/http.invalid	176,965	0.01%
	read/http.truncated	71	3.78e-6%
	write/tcp.reset	8	4.26e-7%
Dial Failure	dial/ip.no_route_to_host	28,954	0.001%
	dial/tcp.refused	23,716	0.001%
	dial/tcp.reset	2,104	1.12e-4%
	dial/network_unreachable	436	2.32e-5%
Setup	write/system	1	5.33e-8%

Figure 5: **Blockpage in South Korea**

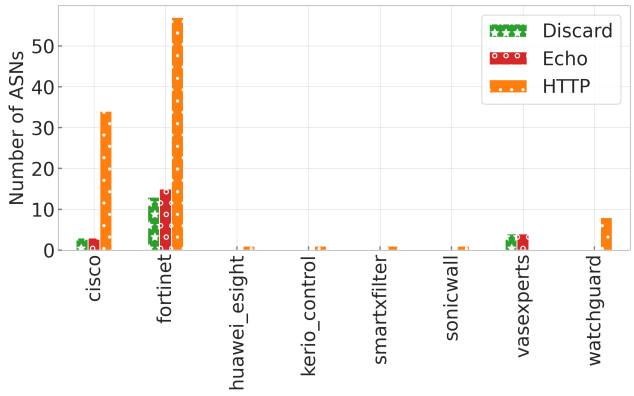

Figure 6: **Commercial Products detected in Censored-Planet HTTP Data in September 2022**—Our specialized fingerprints help in detecting the presence of commercial firewalls that block access to content.

Case Study: Censorship Fingerprints We find that our censorship fingerprints provide explicit confirmation of the entity behind blocking. In South Korea, we observe that 5.6% of Censored Planet's Echo measurements with unexpected responses in May 2022 are matched with a national blockpage fingerprint, shown in Figure 5.

Our censorship fingerprints also help us study the use of commercial firewall software to block access to content in different networks, as done in previous work [18, 44]. Figure 6 shows the commercial products identified by the pipeline while parsing Censored Planet HTTP, Echo, and Discard data in September 2022. We find commercial products manufactured by Fortinet and Cisco being deployed in a large number of ASNs. Arming policymakers with such knowledge quickly can help them raise issues of unfair and unnecessary blocking practices to the right authorities [48, 54].

4.3 Map to Outcomes

Besides unexpected response content, censorship can also result in different types of network errors, such as a TCP reset from an injected packet, or a timeout from dropped packets. However, certain network errors could also be due to factors like network congestion or temporary measurement setup failures. Therefore, the final step of our pipeline is to map each measurement to a human-readable outcome that indicates if the result is expected or the stage and type of error (e.g., read/timeout), which enables efficient and accurate

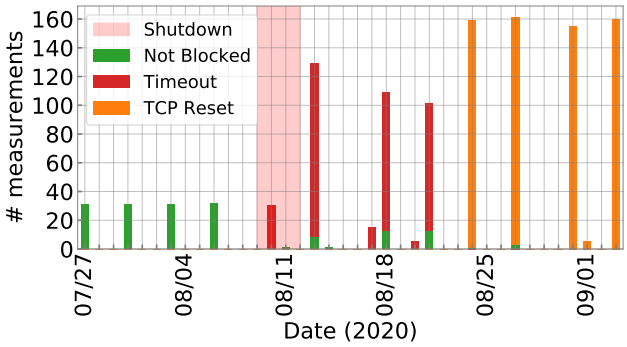

Figure 7: **Accessing psiphon.ca over HTTPS in AS6697 during the Belarus shutdown**—Mapping network errors to outcomes makes changes in censor behavior visible.

aggregation and analysis. We investigate all error strings appearing in the raw Censored Planet data, which correspond to standard network error strings, and observe many errors that did not provide a clear failure reason. For example, we find that the error `readLoopPeekFailLocked: <nil>` actually corresponds to TLS handshake failure. In total, we identify 53 distinct identifiers that cover all appearing errors over different Censored Planet datasets and map them to outcomes with respect to censorship.

An overview of outcomes in HTTP measurements and the percentage of HTTP measurements between January 2022 and September 2022 that match each outcome are shown in Table 2. We define specific outcomes for our fingerprinted responses (e.g., `expected/akamai`, `content/known_blockpage` and `content/known_not_censorship`). We classify over 60 million measurements as expected behavior for the Akamai network due to our fingerprints. Previous work has often misclassified these measurements as censorship, as discussed in §3.3.

Most measurements (94.45%) do not indicate censorship, as censorship is a really rare phenomenon in most parts of the world. A small percentage of measurements fail due to setup errors or errors during the TCP connection (0.002%). Others experience repeated read or write failures during the HTTP request (0.59%), which indicates blocking, or a mismatch between the control and test measurements (1.66%). We hope that our paper encourages censorship measurement platforms to adopt a similar approach to account for all sources of errors.

Case Study: Censorship Mechanisms Our outcome classifications can be used to track changes in censorship mechanisms. For example, Figure 7 displays Censored Planet measurements showing the SNI blocking of psiphon.ca in AS6697 around the August 2020 Belarus shutdown [44, 52]. Separating failed measurements into `connection timeout` and `TCP RST` cases makes it apparent that there are changes in censor behavior over time. Psiphon is first blocked by timeouts during the shutdown. Several weeks after the initial block (and the end of the shutdown), the censorship method changes to injecting TCP RSTs. Our censorship data analysis pipeline enables such accurate and efficient interpretation of censorship data.

5 Discussion & Conclusion

Our work tackles the key challenges currently posing a barrier to the meaningful use of censorship data. We identify several areas where previous work suffer from these challenges, and highlight how the adoption of a standardized analysis process can help characterize censorship practices more accurately. We believe that a good censorship data analysis pipeline must account for the critical challenges we identify, though we do not claim that doing so will eliminate all sources of error. Internet censorship is a constantly evolving phenomenon, and thus the analysis process needs to be modified to account for changes in the future. Many steps in the process (such as adding new page signatures) benefit from the manual context provided by domain knowledge, which is hard to eliminate. Keeping this in mind, we build our data analysis pipeline for Censored Planet data to be iterative and efficient, and open source it so that it can be maintained by the community in a crowd-sourced manner.

Although censorship measurement has garnered much attention over the past years, the availability of large-scale, longitudinal censorship measurement data to analyze is a relatively new advancement. Analyzing censorship measurement data continuously can be prohibitively expensive in terms of computing and storage space. Future work can explore the applicability of machine learning methods that can simplify the analysis process. Another aspect we do not cover explicitly in this work is data exploration, and quickly extracting takeaways from large-scale processed data is a key challenge. We believe further research in censorship data reporting and visualization tools can enable fast analysis by offering the ability to aggregate and investigate at different levels of abstraction.

While the pipeline we propose in this paper is tailored towards censorship data, much of the process is also applicable to other censorship measurements platforms such as OONI, ICLab, and GFWatch, and indeed to other Internet measurement datasets. For example, cases of server-side blocking may appear in datasets containing DNS resolutions, and website localization causes variance in web crawls. We encourage future work to adapt our insights for targeting analysis challenges in other Internet measurement datasets. We hope that our detailed breakdown of challenges motivates researchers to follow best practices and use our data analysis pipeline to provide more accurate and impactful characterization of pervasive Internet censorship.

6 Acknowledgments

The authors thank the anonymous reviewers for their helpful feedback. We are also grateful to Armin Huremagic, Elisa Tsai, and the Google Jigsaw team for their help and support for this work. This work was supported by the Defense Advanced Research Projects Agency under Agreement No. HR00112190127.

References

[1] Access Now. Keep It On. https://www.accessnow.org/keepiton/, 2020.

[2] Access Now. Internet shutdowns report: Shattered dreams and lost opportunities — a year in the fight to #KeepItOn. https://www.accessnow.org/keepiton-report-a-year-in-the-fight/, 03 2021.

[3] S. Afroz and D. Fifield. Timeline of Tor censorship, 2015. http://www1.icsi.berkeley.edu/~sadia/tor_timeline.pdf.

[4] A. Akhavan Niaki, S. Cho, Z. Weinberg, N. P. Hoang, A. Razaghpanah, N. Christin, and P. Gill. ICLab: A Global, Longitudinal Internet Censorship Measurement Platform. In *IEEE Symposium on Security and Privacy (S&P)*, 2020.

[5] Anonymous. Towards a comprehensive picture of the Great Firewall's DNS censorship. In *Free and Open Communications on the Internet (FOCI)*, 2014.

[6] Anonymous, A. A. Niaki, N. P. Hoang, P. Gill, and A. Houmansadr. Triplet censors: Demystifying Great Firewall's DNS censorship behavior. In *Free and Open Communications on the Internet (FOCI)*, 2020.

[7] S. Aryan, H. Aryan, and J. A. Halderman. Internet censorship in Iran: A first look. In *Free and Open Communications on the Internet (FOCI)*, 2013.

[8] K. Bock, A. Alaraj, Y. Fax, K. Hurley, E. Wustrow, and D. Levin. Weaponizing middleboxes for {TCP} reflected amplification. In *USENIX Security Symposium (USENIX Security*, 2021.

[9] K. Bock, Y. Fax, K. Reese, J. Singh, and D. Levin. Detecting and evading censorship-in-depth: A case study of Iran's protocol filter. In *Free and Open Communications on the Internet (FOCI)*, 2020.

[10] CAIDA. AS Rank: A ranking of the largest Autonomous Systems (AS) in the Internet. https://asrank.caida.org/.

[11] CAIDA. Internet Outage Detection and Analysis (IODA). https://ioda.caida.org/ioda/dashboard.

[12] CAIDA. Routeviews Prefix to AS mappings Dataset for IPv4 and IPv6. https://www.caida.org/catalog/datasets/routeviews-prefix2as/.

[13] Censored Planet. Censored planet: An internet-wide, longitudinal censorship observatory, 2022. https://censoredplanet.org/.

[14] Censored Planet Data Analysis Pipeline, 2023. https://github.com/censoredplanet/censoredplanet-analysis.

[15] Z. Chai, A. Ghafari, and A. Houmansadr. On the importance of encrypted-SNI (ESNI) to censorship circumvention. In *Free and Open Communications on the Internet (FOCI)*, 2019.

[16] S. Cho, R. Nithyanand, A. Razaghpanah, and P. Gill. A churn for the better: Localizing censorship using network-level path churn and network tomography. In *International Conference on Emerging Networking EXperiments and Technologies (CoNEXT)*, 2017.

[17] Citizen Lab. Block test list. https://github.com/citizenlab/test-lists.

[18] J. Dalek, B. Haselton, H. Noman, A. Senft, M. Crete-Nishihata, P. Gill, and R. J. Deibert. A method for identifying and confirming the use of URL filtering products for censorship. In *Internet Measurement Conference (IMC)*, 2013.

[19] DB-IP. https://db-ip.com/.

[20] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. A search engine backed by Internet-wide scanning. In *ACM Conference on Computer and Communications Security*, 2015.

[21] R. Ensafi, P. Winter, A. Mueen, and J. R. Crandall. Analyzing the Great Firewall of China over space and time. *Proceedings on Privacy Enhancing Technologies (PETS)*, 2015.

[22] Geoff Huston. How big is that network?, 2014. https://labs.apnic.net/?p=526.

[23] GFWatch. Gfwatch dashboard, 2022. https://gfwatch.org.

[24] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos. A look at infrastructure geolocation in public and commercial databases. In *Internet Measurement Conference (IMC)*. ACM, 2017.

[25] D. Gosain, M. Mohindra, and S. Chakravarty. Too close for comfort: Morasses of (anti-) censorship in the era of cdns. *Proceedings on Privacy Enhancing Technologies (PETS)*, 2021.

[26] N. P. Hoang, S. Doreen, and M. Polychronakis. Measuring I2P censorship at a global scale. In *Free and Open Communications on the Internet (FOCI)*, 2019.

[27] N. P. Hoang, A. A. Niaki, J. Dalek, J. Knockel, P. Lin, B. Marczak, M. Crete-Nishihata, P. Gill, and M. Polychronakis. How great is the great firewall? measuring china's {DNS} censorship. In *USENIX Security Symposium*, 2021.

[28] Internet Society Pulse. Internet Shutdowns. https://pulse.internetsociety.org/shutdowns, 2021.

[29] R. MacKinnon. China's censorship 2.0: How companies censor bloggers. *First Monday*, 2009.

[30] A. McDonald, M. Bernhard, L. Valenta, B. VanderSloot, W. Scott, N. Sullivan, J. A. Halderman, and R. Ensafi. 403 Forbidden: A Global View of CDN Geoblocking. In *Internet Measurement Conference (IMC)*, 2018.

[31] A. McGregor, P. Gill, and N. Weaver. Cache me outside: A new look at dns cache probing. In *Passive and Active Measurement Conference (PAM). Virtual*, pages 427–443, 2021.

[32] OONI. Research reports. https://ooni.org/reports/, 2021.

[33] OONI. New blocks emerge in Russia amid war in Ukraine: An OONI network measurement analysis . https://ooni.org/post/2022-russia-blocks-amid-ru-ua-conflict/, 2022.

[34] OONI Explorer. DNS Tampering in Myanmar, 02 2021. https://explorer.ooni.org/measurement/20210217T163818Z_webconnectivity_MM_58952_n1_jYRBeNMNDaXRkXAo?input=http%3A%2F%2Fwww.facebook.com.

[35] OONI Explorer. TCP/IP blocking in Myanmar, 02 2021. https://explorer.ooni.org/measurement/20210217T170341Z_webconnectivity_MM_58952_n1_owHhJvJ7UD0d6Mhc?input=http%3A%2F%2Fwww.facebook.com.

[36] Open Observatory of Network Interference (OONI). OONI Website. https://ooni.org/, 2021.

[37] R. Padmanabhan, A. Filastò, M. Xynou, R. S. Raman, K. Middleton, M. Zhang, D. Madory, M. Roberts, and A. Dainotti. A multi-perspective view of internet censorship in myanmar. In *Free and Open Communications on the Internet (FOCI)*, 2021.

[38] P. Pearce, B. Jones, F. Li, R. Ensafi, N. Feamster, N. Weaver, and V. Paxson. Global measurement of DNS manipulation. In *USENIX Security Symposium*, 2017.

[39] R. Ramesh, R. Sundara Raman, M. Bernhard, V. Ongkowijaya, L. Evdokimov, A. Edmundson, S. Sprecher, M. Ikram, and R. Ensafi. Decentralized Control: A Case Study of Russia. In *Network and Distributed System Security Symposium (NDSS)*, 2020.

[40] R. Ramesh, R. Sundara Raman, and R. Ensafi. US Government and military websites are geoblocked from Hong Kong and China, 2020. https://censoredplanet.org/hongkong.

[41] W. Scott, T. Anderson, T. Kohno, and A. Krishnamurthy. Satellite: Joint analysis of CDNs and network-level interference. In *USENIX Annual Technical Conference (ATC)*, 2016.

[42] R. Sundara Raman, L. Evdokimov, E. Wustrow, A. Halderman, and R. Ensafi. Investigating Large Scale HTTPS Interception in Kazakhstan. In *Internet Measurement Conference (IMC)*, 2020.

[43] R. Sundara Raman, P. Shenoy, K. Kohls, and R. Ensafi. Censored Planet: An Internet-wide, Longitudinal Censorship Observatory. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2020.

[44] R. Sundara Raman, A. Stoll, J. Dalek, R. Ramesh, W. Scott, and R. Ensafi. Measuring the Deployment of Network Censorship Filters at Global Scale. In *Network and Distributed System Security Symposium (NDSS)*, 2020.

[45] R. Sundara Raman, M. Wang, J. Dalek, J. Mayer, and R. Ensafi. Network measurement methods for locating and examining censorship devices. In *International Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, 2022.

[46] B. VanderSloot, A. McDonald, W. Scott, J. A. Halderman, and R. Ensafi. Quack: Scalable remote measurement of application-layer censorship. In *USENIX Security Symposium*, 2018.

[47] V. Ververis, M. Isaakidis, V. Weber, and B. Fabian. Shedding light on mobile app store censorship. In *User Modeling, Adaptation and Personalization (UMAP)*, 2019.

[48] Vice. Netsweeper removes alternate lifestyle category, 2019. https://motherboard.vice.com/en_us/article/3kgznn/netsweeper-says-its-stopped-alternative-lifestyles-censorship.

[49] A. Vyas, R. Sundara Raman, N. Ceccio, P. M. Lutscher, and R. Ensafi. Lost in Transmission: Investigating Filtering of COVID-19 Websites. In *Financial Cryptography and Data Security (FC)*, 2021.

[50] P. Winter and S. Lindskog. How the Great Firewall of China is blocking Tor. In *Free and Open Communications on the Internet (FOCI)*, 2012.

[51] X. Xu, Z. M. Mao, and J. A. Halderman. Internet Censorship in China: Where Does the Filtering Occur? In *Passive and Active Network Measurement (PAM)*, 2011.

[52] M. Xynou and A. Filastò. Belarus protests: From internet outages to pervasive website censorship. https://ooni.org/post/2020-belarus-internet-outages-website-censorship/, 09 2020.

[53] T. K. Yadav, A. Sinha, D. Gosain, P. K. Sharma, and S. Chakravarty. Where the light gets in: Analyzing web censorship mechanisms in India. In *Internet Measurement Conference (IMC)*. ACM, 2018.

[54] J. York. Websense bars Yemen's government from further software updates, 2009. https://opennet.net/blog/2009/08/websensebars-yemens-government-further-softwareupdates.