Modeling and Detecting Internet Censorship Events

Elisa Tsai Ram Sundara Raman Atul Prakash Roya Ensafi
University of Michigan
{eltsai, ramaks, aprakash, ensafi}@umich.edu

***Abstract*—Publicly accessible censorship datasets, such as OONI and Censored Planet, provide valuable resources for understanding global censorship events. However, censorship event detection in these datasets is challenging due to the overwhelming amount of data, the dynamic nature of censorship, and potentially heterogeneous blocking policies across networks in the same country. This paper presents CenDTect, an unsupervised learning system based on decision trees that overcomes the scalability issue of manual analysis and the interpretability issues of previous time-series methods. CenDTect employs iterative parallel DBSCAN to identify domains with similar blocking patterns, using an adapted cross-classification accuracy as the distance metric. The system analyzes more than 70 billion data points from Censored Planet between January 2019 and December 2022, discovering 15,360 HTTP(S) event clusters in 192 countries and 1,166 DNS event clusters in 77 countries. By evaluating CenDTect's findings with a curated list of 38 potential censorship events from news media and reports, we show how all events confirmed by the manual inspection are easy to characterize with CenDTect's output. We report more than 100 ASes in 32 countries with persistent ISP blocking. Additionally, we identify 11 temporary blocking events in clusters discovered in 2022, observed during periods of election, political unrest, protest, and war. Our approach provides informative and interpretable outputs, making censorship data more accessible to data consumers including researchers, journalists, and NGOs.**

I. Introduction

The rise of commodity DPIs and other filtering devices has enabled ISPs to implement censorship policies in a more rapid manner [63], [85], [91]. In 2022 alone, Access Now reported 187 disruptions to Internet access in 35 countries, targeting specific populations during critical times, including humanitarian crises, mass protests, and active conflicts and wars [1]. Consequently, censorship observatories such as the Open Observatory of Network Interference (OONI) [26], Censored Planet [73], and GFWatch [34] have emerged to monitor global Internet censorship, increase transparency, and raise awareness. As of April 2023, OONI and Censored Planet's open-access data includes a staggering 1.38 billion and 78 billion measurements carried out in 241 and 223 countries and regions, respectively.

Typically, censorship event discovery happens when news media report new censorship incidents, prompting researchers to investigate open-access censorship data. The analysis of the events to identify potential causes or methods of blocking remains a largely manual and cumbersome task. Although valuable insights are provided by reports from Access Now [1], OONI [55], Censored Planet [63], [64], [72], [85], and Citizen Lab [15], the scalability of manual analysis remains a challenge. Yet, the unexplored data still holds enormous value for the international community to remain informed and for local activists and NGOs to monitor for transparency and accountability in their networks. Unfortunately, this goes beyond the current capability of manual analysis. Due to the overwhelming large quantity and complexity of data, investigating local censorship data remains challenging for non-experts [24].

Previous use of automatic censorship event detection such as time-series-based anomaly detection and temporal trend analysis focuses on identifying countries with high levels of censorship or increasing censorship trends [73]. However, these attempts fall short of detecting censorship at the local network level or providing easily interpretable results for data consumers. Our goal is to bridge this gap by providing an automatic tool that can offer interpretable results at both the country and local levels to provide more context to users of censorship observatories. In contrast to prior methods based on identifying anomalies, we leverage unsupervised learning techniques based on decision trees, which are known for their interpretability [39]. The application of decision trees to unsupervised censorship event detection poses a significant challenge, as the definition of an event of interest and its unsupervised clustering can be highly subjective.

We introduce *CenDTect*, a system that incorporates novel clustering of decision trees to depict blocking policies across domains. In this study, we rely on Censored Planet data [73], which, alongside OONI [26], are two of the largest currently-running observatories providing open-access global data on domain accessibility. We use OONI data to confirm events discovered by *CenDTect* (§V and §VI). Our approach identifies key censorship events from large amounts of censorship data. Figure 1 shows an example output from our system that consists of decision trees depicting blocking policies (TCP resets in AS20661, in IP organization Ttelecom after 2019 April), along with domains that fall under such policies, providing a more comprehensive understanding of censorship events, including the blocklist, timespan, and geolocation.

To address the challenge of identifying censorship events and clustering them in decision tree-based analysis, we propose a new distance metric called cross-classification accuracy. This metric evaluates whether the decision tree of one domain (e.g., `google.com`) can accurately describe the data of another domain (e.g., `twitter.com`) from measurements within a country. Our empirical results show that this metric effectively identifies domains with similar blocking policies, enabling the discovery of censorship events at both local and country levels. We leverage an iterative parallel DBSCAN (Density-

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.23409
www.ndss-symposium.org

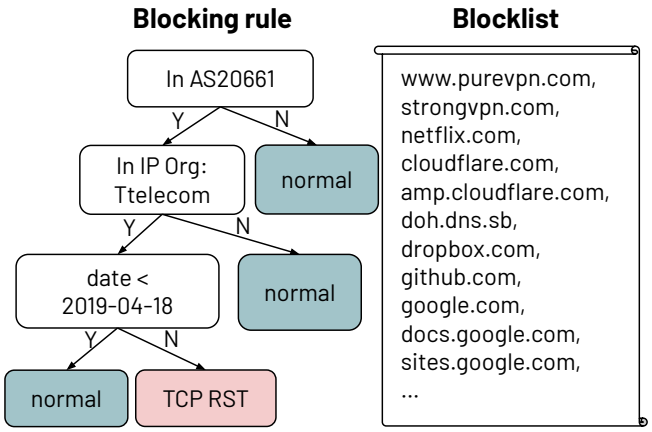


Fig. 1: **Example of *CenDTect*'s output**— A Turkmenistan HTTPS cluster indicating that IP organization Ttelecom in AS20661 blocks domains on the blocklist after 2019 April.

Based Spatial Clustering of Applications with Noise) [22] in conjunction with an adapted prediction scheme (§IV-A) to identify clusters representing censorship events. Our approach overcomes the scalability challenges of manual analysis, the high rate of false negatives (§V), and the lack of interpretability in previous time-series methods. We design *CenDTect* to generate decision trees per domain, preserving interpretability. The event clusters can be used for prominent event discovery (i.e., censorship that impacts a large proportion of a country's population), queried by a search engine, or parsed for new events (§IV-D).

Validating censorship event discovery in real-world datasets is challenging due to the lack of ground truth on a global scale. On the one hand, those who implement censorship policies rarely report them, making it difficult to confirm events with only incomplete evidence [63]. On the other hand, news media typically only cover censorship events with high economic or political impact, such as elections or political unrest [6], [7], [29], [30], [87], making it hard to establish a consistent and reliable benchmark for validating censorship events. Furthermore, the diverse censorship techniques and the dynamic nature of the Internet can lead to both false positives and false negatives [24], [78], further complicating the validation process. To address this issue, we manually compile a Potential Censorship Event List (PCEL) consisting of 38 instances of censorship events from various sources between 2019 and 2020, including OONI reports [55], Google Transparency Report [31], Access Now [53], Internet Society [71] and news media outlets.. We manually check for the presence of blocking signals for the PCEL in Censored Planet's raw data and verify their presence in the event clusters generated by *CenDTect*. Our analysis reveals that all PCEL events confirmed by the manual inspection are present in *CenDTect*'s clusters (§V-A).

We also apply *CenDTect* to analyze Censored Planet HTTP(S) data from January 2019 to December 2022 and DNS data from August 2022 to December 2022, covering a total of 18.07 billion measurements. *CenDTect* discovers 15,360 HTTP(S) clusters in 192 countries and 1,166 DNS clusters in 77 countries. By filtering for prominent events (§IV-D), we report more than 100 ASes in 32 countries with persistent ISP blocking, including 16 previously unreported in other country-specific studies and reports. Additionally, we identify 11 temporary blocking events in clusters discovered in 2022, observed during periods of election, political unrest, protest, and war. Our findings provide insight into the prevalence of organizational blocking and the heterogeneity of blocking practices within countries (§VI-A) [50], [59], [73], [81]. The simplicity and interpretability of *CenDTect*'s output make it persuasive for censorship observatories and data consumers to deploy. We hope that our work enables improved monitoring of censorship events that restrict large-scale access to Internet content.

II. Background

A. Internet Censorship and Measurement Platforms

Internet censorship can happen on multiple protocols. When a user types `example.com` in their browser, a DNS query is sent to resolve the domain name. A censor can either drop this DNS packet (timeout) or manipulate the DNS resolvers to either claim the resolution is failing (nonzero Rcode [65]), inject a private IP, or inject IP addresses hosting a blockpage. Assuming the DNS resolution is not manipulated and the correct IP address is returned, the browser will establish a TCP connection and send an HTTPS request for domain `example.com`. A censor can again intervene by dropping packets, resetting the connection, or injecting responses based on the TLS header or HTTP content.

Numerous studies have delved into Internet censorship, ranging from country-specific reports [5], [9], [21], [33], [34], [47], [48], [51], [63], [68], [80], [84], [86], [89] to multi-country or global-scale measurements [2], [12], [26], [37], [50], [52], [69], [73]–[75]. Global measurements of censorship largely fall into two categories: *in situ* and *remote* measurements. In situ measurements are those conducted by a client device inside the country being studied, e.g., OONI [26] (volunteers) and IClab [50] (VPNs). In contrast, remote measurements typically use public-facing systems to measure disruption, like reflecting queries off of servers, e.g., Censored Planet [73].

B. Censorship Event Detection

Censorship event detection involves identifying anomalous measurements and comprehending the blocklist and scope of networks affected by censorship. Some previous studies view collected measurements as stationary data and identify countries with high levels of censorship by calculating the percentage of anomalous measurements (i.e., $\frac{|domains\_blocks|}{|all\_domains|}$) [59], [78], [81]. Others focus on time series analysis, using either heavily manual [55], [89], or statistical methods such as using a bitmap-based moving windows [73].

For manual analysis, OONI [26] has taken significant steps in activating volunteers on the ground to gather signals of new censorship events for manual analysis. This enables OONI to publish reports of Internet censorship in countries such as Iran, Azerbaijan, Armenia, and Russia [25], [28], [42], [55], [66], [88], [91], pinpointing the blocked domains in measured *ASes*. Despite being accurate, manual analysis suffers from scalability issues. Therefore, it is inevitable that a great number of signals of newly emerging events are overlooked (§VI). This

emphasizes the need for an automatic tool for censorship event discovery.

Sundara Raman et al. [73] conduct an evaluation of four time-series anomaly detection techniques: Median Average Deviation (MAD) [44], likelihood models [77], exponentially weighted moving average models [36], and bitmap-based detection [83]. Among them, bitmap-based detection was found to perform the best on Censored Planet data.

The time-series analyses provide a good starting point for detecting censorship, but they only output a binary judgment on whether a country conducts censorship based on the anomaly threshold. Prior analyses fail to differentiate between ISP and organizational blocking, thus ignoring the heterogeneity of blocking policies across different networks in the same country. As shown in §V, bitmap-based time series anomaly detection is susceptible to false negatives in countries with persistently high levels of blocking rates. In these cases, it fails to produce a noticeable deviation.

Recently, Brown et al. [11] proposed employing supervised learning methods to discern whether a single DNS measurement shows signs of manipulation. While their objective—that of detecting censorship of individual DNS measurements—differs from ours—modeling and detecting censorship events—the features they identify as important are valuable insights for *CenDTect*. Nonetheless, adopting supervised learning for censorship event modeling necessitates comprehensive global ground truth. The challenges of scaling supervised learning to a global scale are elaborated in §III-A and §VI-A.

C. Terminology

***Internet Censorship*:** The censorship we discuss in this paper involves blocking access to *domains* through methods like DNS manipulation and HTTP(S) request filtering. We do not cover Internet blackout datasets such as IODA [27].

***Censorship Event*:** A censorship event is the implementation of new blocking measures by a *controlling authority*, such as a government, ISP, or organization's network administrator. Censorship serves various purposes, including political, moral, ethical, or security-related content regulation. These events can take place at the national, regional, or institutional levels.

***Simultaneous Blocking*:** We term the phenomenon of a group of domains experiencing similar blocking policies within censored regions as "simultaneous blocking". This can occur due to the implementation of new techniques for blocking or the addition of new domains to existing blocklists.

***IP Organizations*:** An IP organization refers to the organization listed in the WHOIS record for an IP address. The IP organization is often the entity that owns or manages the network infrastructure associated with the IP address.

III. Data

In this section, we discuss the challenges of censorship event discovery, the data schema and characteristics, and data preprocessing. We utilize open-access data provided by Censored Planet [73], a global censorship observatory that collects information from various remote infrastructural vantage points worldwide. On a weekly basis, Censored Planet queries over 2,000 domains using global vantage points, detecting censorship on DNS and HTTP(S) protocols.

TABLE II: **Data format of Quack/Hyperquack raw data**

Fields			
Query & Metadata		Response (list)	
vp	vp.ctry_name	tag	anomaly (deprecated)
	vp.ctry_code	response	controls_failed
test_url	start_time	error	stateful_block
protocol	end_time	control_url	matches_template

A. Challenges for Global Censorship Event Discovery

***(Lack of) ground truth*:** In recent years, open-access censorship measurement platforms such as OONI [26] and Censored Planet [73] have shifted away from labeling measurements as censorship based on a fixed set of heuristics [26], [58], [59], [67], [73], [74]. Rather than providing binary judgments based on fixed heuristics, these platforms now report various types of network traffic anomalies that suggest censorship may be occurring [56], [60]. The judgment of whether censorship is actually taking place is left to the data consumers. This shift has been driven by a number of factors. First, the traditional binary approach is limited in its ability to capture the nuances of different censorship techniques. Second, relying on current censorship judgment heuristics has been shown to be error-prone [24], [75], [89], particularly with the rise of CDNs and cloud providers, load-balancing, misconfigurations, and localization, which complicate Internet censorship measurement [78].

Since May 2021, Censored Planet no longer labels censorship in their collected data [60]. Similarly, OONI confirms censorship only when a blockpage is identified in the measurement [56]. A blockpage is a web page that is displayed to users when censorship is in effect and typically cites a legal justification for the blocking (example in Appendix B). OONI does not flag other anomalous measurements as censorship.

Collecting ground truth for every censorship measurement on a global scale is challenging, if not impossible. Blocking policies are rarely disclosed by censors [63]. Instead, activists on the ground often prompt censorship measurement platforms to confirm local censorship events. Researchers then scrutinize recent data in the corresponding country to determine changes in censorship. This limitation makes us rely on unsupervised learning. In §V-A, we manually collect 38 potential censorship events from news media and censorship reports [31], [53], [55], [71] to validate the effectiveness of *CenDTect*. For persistent ISP-level blocking and temporary blocking reported in §VI, we cross-verify it with OONI's open-access data.

***Volatile test lists*:** The challenges of mining censorship datasets are multifaceted. Glitches and churns exist in the dataset, complicating longitudinal event discovery:

- **Domain Test List**: Changes in the test list for censorship measurement platforms occur frequently. While the Citizen Lab Test List [43] changes are typically small, the Tranco top 500 list [23] has been shown to be more volatile.

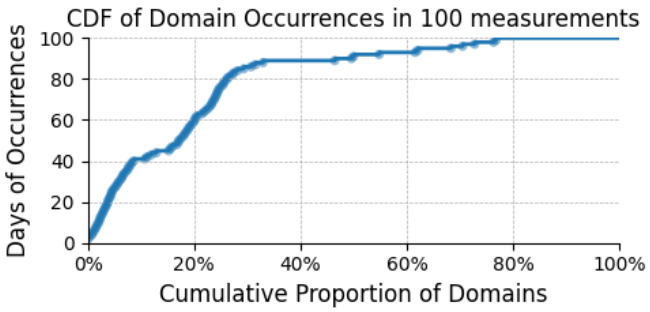


Fig. 2: **CDF of domain occurrences from Jan 01 to April 14, 2022**—The y-axis shows the count of domain occurrences in 100 measurements.

Figure 2 shows that about 20% of domains only occur in 60% of the measurements during a snapshot of 100 HTTPS snapshots of global measurements over the timespan of 4 months in 2022. A heatmap of changes based on category over the same time period can be found in Appendix A.

- **Vantage Point List**: The vantage point list is volatile due to several reasons. Censored Plant uses Zmap [19], Nmap [46], and Censys [18] for initial vantage points selection by scanning the entire IPv4 space for corresponding servers. The IPv4 address space is constantly changing as new devices are added or removed from the internet. For example, the weekly average number of HTTPS servers observed during January 2022 was 20,294, with an average Jaccard similarity coefficient of 71.03%.

***Large volume of data*:** Censored Planet has been operating since August 2018, and has collected more than 13 Terabytes of more than 78 billion data points so far, with about 20 Gigabytes of data increasing on a weekly basis, covering vantage points in 223 countries (see measurement coverage per protocol in Table III). Therefore, we aim to design a model that is computationally efficient for all historical data.

B. Data Collection and Format

***HTTP(S) Data*:** Censored Planet uses Quack [79] and Hyperquack [74] techniques to detect HTTP(S) censorship. Quack employs the Echo [62] and Discard [61] protocols to send HTTP-like requests to remote vantage points. Quack subsequently compares the received response with the "expected" response according to the protocol in order to identify interference. Hyperquack extends the methodology of Quack to the HTTP(S) protocols. Both techniques use Nmap [46], Zmap [19], and Censys [14] to identify vantage points. The HTTP(S) data format is shown in Table II.

***DNS Data*:** Censored Planet detects worldwide DNS manipulation using Satellite/Iris [59], [67] and CERTainty [78] sending queries to resolvers identified through Censys [18], [19]. It collects DNS responses from global resolvers and HTTP(S) pages and certificates hosted on resolved IPs.

C. Data Schema and Preprocessing

The Censored Planet pipeline [75] reads the raw data from various measurement platforms into Big Query and annotates the measurement with IP metadata from CAIDA [13], DBIP [17], and Censys [14].

Protocol	HTTPS	HTTP	ECHO	DISCARD	DNS
VP	56,004	64,200	268,736	147,121	75,122
Country	221	223	188	188	188
ASes	9,354	11,453	7,465	7,467	7,983
IP Orgs	16,869	19,551	17,310	16,341	7,891

TABLE III: **Number of vantage points and covered countries, ASes and IP organizations for each measurement type**—Quack and Hyperquack have yearly averages from 2019 to 2022, while Satellite/Iris has data after August 2022.

As shown below, our work relies on IP metadata (country, ASN, IP organization), timestamps, and parsed classes. "count" reflects the number of vantage points that share identical values across all other entries:

| domain | date | country | AS | IP Org | class | count |

We parse the raw DNS response and HTTP(S) responses and transform the raw response into 4,370 different classes such as connection timeouts, TCP reset, HTTP(S) status mismatches, HTML body mismatches, *etc*.

D. Data Classes Taxonomy

Drawing from Censored Planet [73] and our expertise in censorship data, we created four distinct parsed categories from measurement responses. The "likely censorship" category signifies clear blocking signals, such as a blockpage or a TCP reset, similar to previous work [73], [75].

***HTTP(S)*:** As shown in Table IV, we parse 311 distinct HTTP(S) classes out of the Censored Planet dataset from January 2019 to December 2022. We categorize the observations into different categories.

1) The "benign" category encompasses entries that match with a template or are returned by vantage points on trusted CDNs, meaning that the accessibility of measured domains is not tampered with.
2) The "very unlikely censorship" category represents network reachability anomalies, including instances of dialing errors such as TCP resets, host unavailability, network unreachable errors, and refused connections. These anomalies are deemed highly improbable to be a result of censorship-related activities.
3) The "unlikely censorship" category consists of observations that exhibit HTTP status code mismatches, TLS errors, or mismatches between the HTTP body and TLS. Those are not common signals of network adversaries. Instead, they might occur because of configuration errors or unexpected behavior from hosting providers.
4) Finally, the "likely censorship" category includes responses matching with blockpage fingerprints which indicate potential instances of content blocking, and signals of potential network adversaries such as various read and write errors. We put the matched fingerprint in the parsed classes since the fingerprint name usually indicates either the deployer or origin of censorship or the commercial middleboxes used for censorship, resulting in 225 blockpage classes

	Category	Count	Subcat Count	Type	Example
HTTP(S) - 331 classes	✔ Benign	2	2	Match with template or trusted CDN	match, trusted_host:akamai
	? Very Unlikely	6	6	Network reachability anomalies	dial/tcp.reset, dial/ip.host_no_route, dial/ip.network_unreachable, dial/tcp.refused...
	? Unlikely	73	67 2 4	HTTP status code mismatch TLS error HTTP body and TLS mismatch	content/status_mismatch:210, content/status_mismatch:500,... tls/tls.failed, tls/timeout content/body_mismatch, http/http.invalid, content/tls_mismatch, content/mismatch
	! Likely	230	225 5	Match with blockpage fingerprints Potential network adversaries	content/blockpage:c_isp_ru_sibset, content/blockpage:a_prod_fortinet_1,... read/tcp.reset, read/timeout, read/http.truncated_response, read/http.empty, write/tcp.reset
DNS - 4,059 classes	✔ Benign	3	3	Expected answers	answer:matches_ip, answer:valid_cert, answer:matches_asn
	? Very Unlikely	1	1	Network reachability anomalies	read/ip.host_no_route
	? Unlikely	2,221	2207 11 3	Not validated answer Non zero rcode or no Type A response Read/Write error	answer:not_validated:CHINANET-BACKBONE, answer:not_validated:AS262589, ... answer:no_answer, dns/rcode:ServFail, dns/rcode:FormErr, dns/rcode:Refused, ... read/udp.timeout, read/udp.refused, read/dns.msgsize
	! Likely	1,834	1781 52 1	Invalid Certificates Match with blockpage fingerprints Non zero rcode: NXDOMAIN	answer:cert_not_for_domain:blocked.compnet.ru, answer:invalid_ca_valid_domain: NetAlerts Services, ... page:http_blockpage:c_isp_id_myrepublic_redirect_301_3_satellite, ... dns/rcode:NXDomain

TABLE IV: Taxonomy of parsed classes for Censored Planet HTTP(S) and DNS data in 4 categories—"Benign", "Very Unlikely Censorship", "Unlikely Censorship", and "Likely Censorship". The class template is in the format of <group>/<detailed_info>. The "likely censorship" category signifies clear blocking signals.

(including special status codes such as "423 Locked" and "451 Unavailable For Legal Reasons").

***DNS*:** We have a total number of 4,059 distinct classes in DNS data from August 2022 to December 2022.

1) The "benign" category represents entries that exhibit expected answers, including matching IP addresses, valid certificates for queried domains, and matching ASNs.
2) The "very unlikely censorship" category indicates instances of network reachability anomalies.
3) We identify 3 distinct sub-categories within the "unlikely censorship" category. The first sub-category comprises observations with unverified answers, distinguished by their IP organization name (or AS name, in cases where the IP organization is not available). We retain the AS name and IP organization information to enable future researchers to investigate the underlying reasons behind these anomalies, resulting in 2,217 classes. The second sub-category involves instances of nonzero RCODE (excluding NXDOMAIN) or the absence of a Type A response, indicating potential errors like no answer, server failure, format issues, and refused requests. The third sub-category covers read/write errors, encompassing timeouts, refused connections, and DNS message size discrepancies. Prior research has shown that non-zero RCODE (except NXDOMAIN) and connection errors in DNS packets are typically unrelated to censorship activities [50], [59].
4) The "likely censorship" category within the DNS data consists of three sub-categories. The first sub-category includes instances of invalid certificates, where the certificate either does not correspond to the requested domain or has been signed by an untrusted Certificate Authority (CA). In cases where the certificate is issued for other domains, we denote the class with the certificate domain name (e.g., `cert_not_for_domain:blocked.compnet.ru`). For certificates with untrusted roots, the class is denoted with the CA name (e.g., `invalid_ca_valid_domain:NetAlerts Services`), resulting in a total of 1,781 classes within this sub-category. The second sub-category consists of IPs hosting blockpages, indicating potential instances of DNS-based content blocking. Finally, the third sub-category is identified by non-zero `RCODE=NXDOMAIN`, which signifies that the requested domain does not exist. This particular RCODE value is known to be used by Pakistan for DNS censorship [48].

Given the parsed classes of censorship measurements, our goal is to identify an effective decision boundary between all the classes in the spatiotemporal data. For this paper, we conservatively only report the clusters of the "Likely Censorship" type, which signifies clear blocking signals, and helps our model reflect censorship events accurately. While we focus on detecting highly confident censorship events in this paper, future research can explore other aspects, like large-scale network misconfigurations using *CenDTect*.

E. Ethics

Our analysis in this work relies on historical data collected by Censored Planet, which involves querying vantage points such as open DNS resolvers and HTTP(S) servers, which may potentially trigger a censor and cause potential risk to the operators of these hosts. Such measurements have carefully followed ethical norms and best practices to minimize any risk involved [19], [58], [59], [73], [79]. We acknowledge that the use of this data carries ethical implications and must be handled with caution. Fortunately, previous censorship detection systems, community discussions, and workshops have extensively discussed ethical considerations for censorship measurement [16], [40], [49], [57], [90] resulting in well-defined technical practices to minimize risk which guides our work.

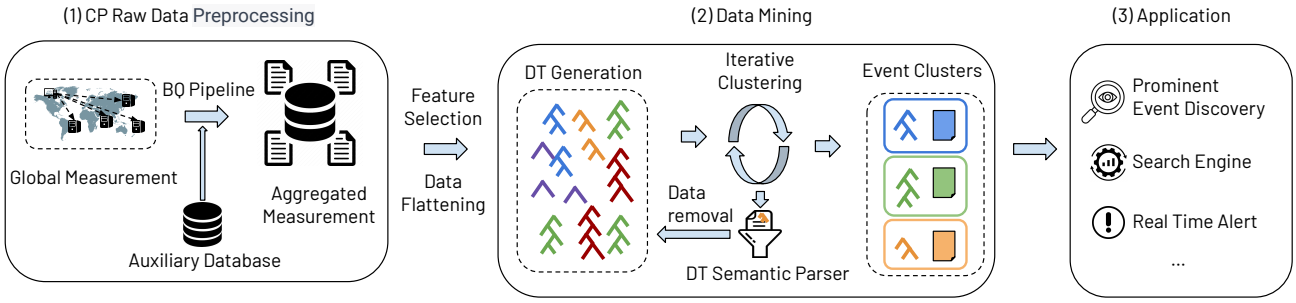


Fig. 3: ***CenDTect* Architecture**–(1) Censored Planet raw data preprocessed; (2) *CenDTect* generates decision trees as blocking policies for each domain. Decision trees are clustered, combined, and transformed into event clusters; (3) applications of event clusters: prominent censorship event discovery, search engine, and real-time alert.

IV. System

CenDTect is composed of three stages as illustrated in Figure 3. In the first stage, global censorship measurement raw data is preprocessed, and augmented with metadata such as IP ownership and domain categories. In the second stage, the system generates decision trees per domain in a given spatiotemporality and clusters them based on the similarity of the blocking rules indicated by the decision trees. Finally, in the third stage, the event clusters can be applied to different use cases such as prominent event discovery, search engines, and alerts.

A. Modeling and Mining Censorship Events

***Assumptions*:** We make two key observation-based assumptions to build our censorship event discovery techniques. Firstly, we assume IP organizations as the currently most acceptable unit of censorship. This is because network administrators (of ISPs, university networks, company networks, *etc*) are able to implement and update blocking policies, which impact the traffic originating from or transmitting their networks. Secondly, we observe that censorship events typically involve the simultaneous blocking of multiple domains, either because new blocking techniques are implemented, or because of blocklist updates. Our assumption is supported by our curated list of censorship events collected from news and reports (§V). In cases where only one major social media platform such as Twitter is blocked, it is highly common that its subdomains such as `mobile.twitter.com`, `t.co` and `analytics.twitter.com` are also blocked [86].

***Goal*:** In the context of identifying a censorship event, it is important to answer four key questions: (1) **where** - the geolocation (country, city, Autonomous System (AS), Internet Service Provider (ISP), or IP organization) of the vantage points that are conducting censorship, (2) **when** - the timespan of the censorship event, (3) **how** - the blocking method (such as DNS poisoning, TCP reset, blockpage injection, *etc*.), and (4) **what** is blocked, i.e., domains on the blocklist. These details should be included to ensure comprehensive reporting of the censorship event. In our work, we define items (1)-(3) as the "blocking rule" of the corresponding censorship event. To put it in formally, censorship event C_θ:

$$C_\theta = (R_\theta, \{d_{\theta_1}, d_{\theta_2}, \ldots, d_{\theta_n}\}) \tag{1}$$

where $d_{\theta_1}, d_{\theta_2}, \ldots, d_{\theta_n}$ are the domains whose signals in the dataset can be described by blocking rule R_θ. Our detection goal is to find such censorship event C_θ in censorship datasets.

***Discover simultaneous blocking through decision trees*:** To extract blocking rules and their corresponding domains, we essentially need to discover a set of vantage points that share the same blocking behavior for a certain set of domains. For domain d_i, its data in a given country c starting from t_1 to t_m can be expressed with:

$$RES_{d_i,c,t_1,t_m} = \begin{array}{c} \begin{matrix} t_1 & t_2 & \ldots & t_m \end{matrix} \\ \begin{pmatrix} res_{i(1,1)} & res_{i(1,2)} & \ldots & res_{i(1,m)} \\ res_{i(2,1)} & res_{i(2,2)} & \ldots & res_{i(2,m)} \\ \vdots & \vdots & \ddots & \vdots \\ res_{i(n,1)} & res_{i(n,2)} & \ldots & res_{i(n,m)} \end{pmatrix} \end{array} \begin{matrix} vp_{c,1} \\ vp_{c,2} \\ \vdots \\ vp_{c,n} \end{matrix} \tag{2}$$

where $(vp_1, \ldots, vp_n)$ are vantage points in country c, and $res_{i(j,k)}$ $(1 \le j \le n,\ 1 \le k \le m)$ is the parsed class label (see §III-D) of vantage point vp_j at time t_k for domain d_i.

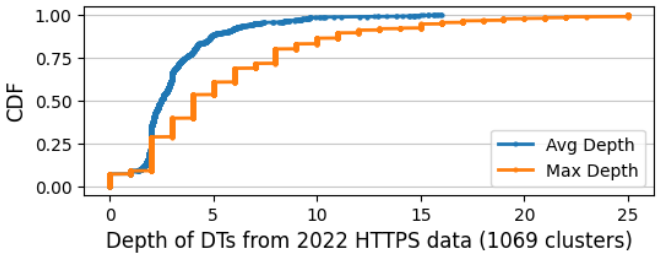
***Aggregation on IP organizations*:** To obtain an aggregated version of Equation 2, we leverage the assumption that IP organization is frequently the atomic unit of Internet censorship.

$$RES_{d_i,c,t_1,t_m} = \begin{array}{c} \begin{matrix} t_1 & \ldots & t_m \end{matrix} \\ \begin{pmatrix} res_{i(1,1)} & \ldots & res_{i(1,m)} \\ res_{i(2,1)} & \ldots & res_{i(2,m)} \\ \vdots & \ddots & \vdots \\ res_{i(n',1)} & \ldots & res_{i(n',m)} \end{pmatrix} \end{array} \begin{matrix} iporg_1 \\ iporg_2 \\ \vdots \\ iporg_{n'} \end{matrix} \tag{3}$$

where $(iporg_1, \ldots, iporg_{n'})$ are the IP organizations that all the vantage points in country c belong to. Suppose there are T classes of parsed responses, then $res_{ia,b}$ is a T-dimensional vector, representing the weighted distribution of all the classes for domain d_i in IP organization $iporg_a$ at time t_b.

Our goal is to identify how to split the spatiotemporal sequences of $RES_{d_1}, RES_{d_2}, \ldots, RES_{d_N}$ for a given country c and time span (t_1, t_m), such that homogeneous blocking patterns emerge. Therefore, our model ought to be able to divide the matrix RES according to how homogeneous the response classes are. Entropy-based models such as decision trees, intuitively,

Fig. 4: **Average and maximum depth of decision trees**—generated using Hyperquack 12-month HTTPS data in 2022.



are the best fit for our needs. Decision trees use a hierarchical structure to recursively partition data into subsets, based on the most informative features, until a stopping criterion is met. This matches our goal of finding regions and timespans with homogeneous blocking policies.

In our approach, we create a decision tree (DT) using Sklearn's Gini impurity decision tree algorithm [10] for each domain to capture its blocking behavior in a specific timespan and geographic location ($d_i \rightarrow DT_i$). We generate a decision tree per domain to preserve the interpretability of decision trees. As illustrated in Figure 4, even when *CenDTect* processes an entire year's worth of measurement data, the depth of the decision trees remains manageable, with an average depth of less than 16 and a maximum depth of less than 25. As a result, the interpretability of decision trees is retained.

In an unsupervised setting, determining if two domains can be described by the same blocking rule requires assessing their proximity based on a distance measure, denoted by $DTDIST(\cdot,\cdot)$. We introduce a novel distance metric for tree structures. To measure the proximity between two domains, we use the decision tree of one domain to classify the other domain in the same spatiotemporality and record the maximum accuracy (minimum dissimilarity) as their distance. The distance measure is shown in Equation 4, which we refer to as cross-classification, and it allows us to determine if two domains exhibit similar behavior:

$$\begin{aligned} &f: d_i \rightarrow DT_i\,, d_j \rightarrow DT_j \\ &DTDIST(d_i,d_j) = 1 - max(DT_i.Pred(RES_{d_j}), \\ &\qquad\qquad DT_j.Pred(RESd_i)) \end{aligned} \tag{4}$$

***Adpoted Prediction and Iterative Clustering*:** In countries with multiple ASes and ISPs, a single domain can belong to different event clusters due to the complexity of network administration, especially in countries where censorship is not deployed at the international Internet Exchange Point but rather at the ISP level [66], [91]. To successfully extract overlapping domain clusters, we propose a special prediction function in conjunction with iterative clustering. It is worth noting that the term "prediction" in Equation 4 does not imply predicting future events. Rather, by classifying the output of one domain under the decision tree generated by data of another domain, we gain insight into the similarity of their blocking rules.

True Label	p_1	p_1	p_1	0	0
Pred Label	p_1	p_2	0	0	p_2
Pred Success	True	False	**True**	True	False

TABLE V: *spclPred* **function**—0 indicates the domain is accessible, and positive integers indicate different kinds of blocking. p_1 and p_2 are positive integers ($p_1 \neq p_2$).

As shown in Table V, we assign the class label "0" to denote normal domain accessibility, while the other positive integers represent different types of network anomalies, such as TCP reset, TCP timeout, or the presence of various blockpages. *CenDTect* begins by filtering out domains whose decision trees show no censorship, which we refer to as "innocent trees". Since censorship is rare on a global scale, with prior research reporting a global blocking rate of only 1%-2% [59], [73], [78], we are consistently able to identify innocent clusters in practice (i.e., domains that are not blocked throughout the year).

Algorithm 1 getDomainTreeClusters

Input: preprocessed data **data**
Output: clusters of decision trees
Let *innoClu* be an empty cluster
Let *clusters*, *events* be an empty sets
innoClu ← **findInnocent**(data)
while True **do**
  data ←**removeInno**(data)
  localClusters ← **genInitialClusters**(*data*, *innoClu*)
  clusters ← *clusters* ∪ (*localClusters* − *innoClu*)
  clusters ← **mergeClusters**(*clusters*)
  localEvents ← **parseEvents**(*localClusters*)
  if *localEvents* ∪ *events* = *events* **then**
    break
  end if
  events ← *events* ∪ *localEvents*
  data ← **removeEvents**(*data*, *localEvents*)
end while
RETURN *clusters*

Algorithm 2 genInitialClusters

Input: df, innoCluster
Output: clusters of decision trees
Let *thres* be the user-defined threshold
Let *label_map* be empty dict: *int* → *Cluster*, where struct *Cluster* has fields: *Cluster.DT*, *Cluster.domains*
Let *clusterList* be an array of *Cluster*
DTs ← **parallelGenDT**(*df*)
DTs ← **bucketUniqueDT**(*DTs*)
// DBSCAN
db ← *DBSCAN*(
    eps = *thres*,
    min_sample = 5,
    metric = *DTDIST*)
for *i* in 1...*length*(*db.labels*) **do**
  put domains and corresponding DT into *clusterList* and *innoCluster*
end for
return *clusterList*, *innoCluster*

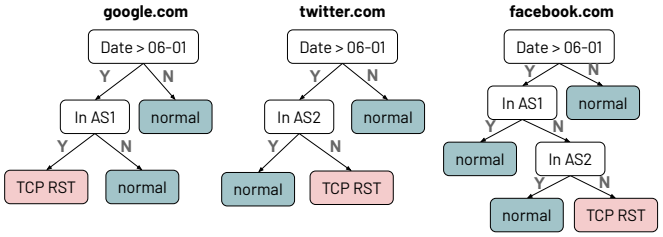


Fig. 5: Toy example of decision trees for domains in a region that only has 3 ASes: AS1, AS2, AS3.

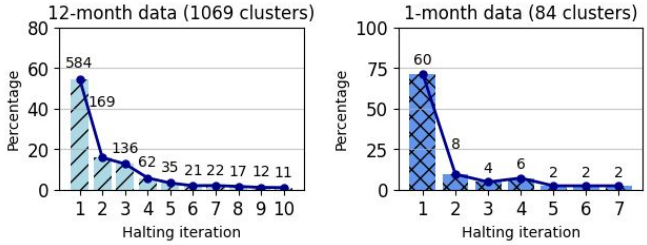


Fig. 6: **The halting iteration of *CenDTect* on different volumes of data** - clusters generated from Hyperquack HTTPS data from 2022 Jan to Dec, and 2023 March, respectively.

Algorithm 1 generates clusters of decision trees. It finds domains that fall under innocent trees (**findInnocent**) and iteratively removes innocent trees from the data (**removeInno**). It then merges local clusters (**mergeClusters**). **parseEvents** extracts the geolocation and timespan of the simultaneous blocking events by traversing the decision tree. If there are no new events from the local cluster, the loop terminates. Otherwise, it removes the lines of data that correspond to the events already found in the data (**removeEvent**). Finally, it returns the resulting clusters.

Algorithm 2 performs clustering. It uses parallel processing to generate decision trees for each domain (**parallelGenDT**) and buckets domains with unique trees (**bucketUniqueDT**). It performs DBSCAN clustering on the decision trees and assigns domains and corresponding decision trees to clusters. The resulting clusters and the innocent cluster are returned.

As shown in Table V, *spclPred* is an adapted prediction function that operates similarly to a regular prediction function. However, when the decision tree classifier predicts a blocking case as accessible (assuming that this tree is not an "innocent tree"), we consider the classification successful. The prediction scheme allows us to deal with domains that are blocked in country c by different IP organizations. In each iteration, we remove the data responsible for the discovery of previously detected events and begin another round of clustering until no new events are found. This approach ensures the detection of all censorship events associated with a particular domain, even in the presence of overlapping events. We refer to this process as iterative clustering, where each clustering iteration operates on a refined data set based on the previous clustering results until no further clusters can be identified. Figure 6 illustrates the stopping iteration when using 12-month and 1-month Hyperquack HTTPS data as input for all the countries covered in our study.

B. Clustering Example

Figure 5 presents simplified examples of decision trees generated for three domains in a specific country. In this hypothetical scenario, there are only 3 ASes (AS1, AS2, and AS3) each with one IP (no further branching). The decision trees depict the blocking rules for each website:

- T_{google}: `google.com` is blocked in AS1 via TCP reset after June 1. The blocking in AS1 is homogeneous.
- $T_{twitter}$: `twitter.com` is blocked via TCP reset after June 1 in AS1 and AS3.
- $T_{facebook}$: `facebook.com` is blocked via TCP reset after June 1 in AS3.

The rule "blocked in AS1 via TCP reset" is a subset of the rule "blocked in all ASes via TCP reset except AS2". To separate sequences RES_{google} and $RES_{twitter}$ efficiently, we split them based on AS1. *CenDTect* automates this process by identifying cluster $C_1 = (R_1 = DT_{google}, \{google.com, twitter.com\})$ using the adopted cross-classification metrics. It then removes data in decision tree DT_{google} to discover C_1. *CenDTect* continues clustering and discovers cluster $C_2 = (R_1 = DT_{facebook}, \{facebook.com, twitter.com\})$. The iterative clustering process halts when no new events are found.

We leverage parallel DBSCAN [22], a density-based clustering algorithm, to implement iterative clustering of decision trees. The decision tree serves as the input to DBSCAN to cluster domains with similar censorship policies together using cross-classification rate as the distance measures, where the distance threshold is a tunable hyperparameter. This approach ensures an iterative refinement of the clustering process and enables us to identify all censorship events associated with a domain, even in cases where events overlap.

C. Implementation

We implement *CenDTect* in Python 3.10 on a Linux server with 32 GB of RAM and a 12-core processor. There are several hyperparameters in our model, which include the date range and the epsilon parameter of the DBSCAN algorithm. For the latter, we set the threshold at 99.5% to account for the rarity of censorship events (prior work reports 1%-2% in global measurements) [26], [50], [59], [73], [78]. Since the dataset is collected twice per week and may miss events lasting less than 3 days, the baseline error margin during a year of data collection is approximately $\frac{3}{365} = 0.8219\%$. All event clusters in this paper are generated from input data covering a whole year, except for alerting systems (§IV-D). The 99.5% threshold provides some tolerance for occasional temporal glitches, without significantly increasing uncertainty in the data mining stage. The Epsilon is tunable for future users of *CenDTect* based on the range of dates they are investigating and the nature of the censorship datasets they are working with. *CenDTect* can generate event clusters using all the data in a given country, or alternatively, for countries with a large number of ASes, data can be split on each AS during the preprocessing stage.

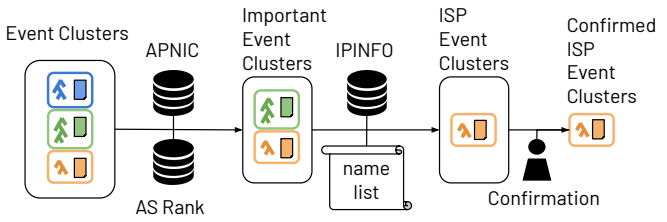


Fig. 7: **Pipeline of *CenDTect*'s prominent event discovery**.

D. Application

The data mining stage of *CenDTect* produces decision trees and a corresponding list of domains. In order to gain insight into the blocking rules and the blocked domains, *CenDTect* performs a semantic parsing of each cluster's representative tree from the bottom up, generating rules that entail geolocation and timespan. The parsed rules and blocked domains can be used as input for various applications, including prominent event discovery, search engines, alerts, blocklist analysis, and measurement optimization. In this work, we implement the following three instances of applications.

***Prominent Data Discovery*:** Not all instances of censorship bear an equal impact on the affected networks and populations. Therefore, we use the APNIC eyeball population estimates [4], which provides the "percent of population" that an AS covers in a given country, and the CAIDA AS-RANK [13], which calculates the AS popularity based on AS customer cone [45]. In order to focus on the most influential ASes, we aggregate the two lists and filter the clusters by the top 20 ASes in all the measured countries, as illustrated in Figure 7.

It is also necessary to differentiate between censorship imposed by ISPs versus that implemented by organizational entities, such as schools and corporations (see §VI-A). To achieve this differentiation, we utilize ipinfo's ASN API [38] to tag ASes as ISP, Education, Business, Hosting, or Inactive. Given that there is no database that categorizes *IP organizations* in this manner to our knowledge, we extract common words from 40,831 unique IP organization names (47.5% in English). We then manually create a list of words that indicate organizational blocking, such as "bank," "hospital," "insurance," "medical," "university," "research," *etc*. To ensure that our list is language-independent, we translate these words to the top occurring languages in IP organization name identified by Python's langdetect. If the IP organization name contains an element in the organizational blocking name list (or their translated version), we downgrade the event's blocking type from ISP-level to organizational. Otherwise, each IP organization inherits the category of its AS's category according to ipinfo. To validate all ISP-level events reported in §VI, we manually verified all ASes and IP organizations to ensure that they are ISPs (see filtering statistics for events in 2022 in Table VIII).

***Search Engine*:** We build a command-line search script that enables users to specify a country, timespan, and additional parameters such as ASes, netblocks, IP organizations, and domains to investigate if *CenDTect* detects clusters corresponding to blocking based on the provided criteria. This tool can be further extended to a web-based search engine. In our evaluation of censorship events in PCEL, we rely on this search script.

***Alerting System*:** *CenDTect* is capable of discovering ongoing censorship events. When deployed for this purpose, *CenDTect* fetches the N most recent scans, where N can be configured, and checks for the emergence of new events. To differentiate between the emergence of new clusters due to the implementation of new censorship measures or the addition of new domains to Censored Planet's test list, the detector sends an alert when a domain's category does not fall within the top K persistently blocked categories in the country. The threshold for this classification is adjustable and we refer to these events as "unique events". In this paper, we set the threshold at 5.

We employ *CenDTect* to cluster the Hyperquack HTTPS data in March 2023, encompassing 77.98 million lines of measurement. As in Figure 6, the halting iterations are fewer when doing alerts. Our implementation (§IV-C) achieves an average clustering time of 38.31 seconds per country, with a maximum of 320 seconds and a standard deviation of 53.14 seconds. This demonstrates the efficiency of *CenDTect* in handling large-scale censorship data.

V. Evaluation

Evaluating global censorship event discovery is challenging because of the lack of ground truth [24], [26], [50], [59], [73], [81]. We tackle this by evaluating *CenDTect*'s output against a list of manually-collected censorship events from 2019 to 2020. Furthermore, we evaluate unique censorship events from 12-month data in 2022, showing *CenDTect*'s ability to uncover previous unreported events.

A. Potential Censorship Events List

We aggregate a list of Potential Censorship Events (PCEL) that occurred in 2019 and 2020 via various sources, including OONI Reports [55], Google Transparency Report [31], Access Now [53], Internet Society [71] and news media outlets such as BBC, Washington Post, New York Times, CNN, and Eurasianet. PCEL identifies a total of 114 events. After filtering out events that are caused by Internet blackouts [27], [32], [41], we narrow down our analysis to 38 events. We manually go through the 38 events to see if the signals of blocking occur in the Censored Planet raw data. Some of the events are not covered in the raw data, as we discuss in §V-C. We identify 12 events that have signals in raw data through manual analysis.

B. Positive Cases

We use PCEL to evaluate *CenDTect*, and a version of the bitmap anomaly detection technique proposed in prior work [73]. The bitmap models require a threshold to determine if an anomaly score is high enough to be considered an anomaly. For each event in the PCEL with timespan (t_s, t_e), we fetch the data for the target countries in $(t_s - 15, t_e + 15)$ to see if the event can be reported by bitmap anomaly detection. Sundara Raman et al. [73] use an alphabet size of 4 and a lead and lag window size of 2% of the time series length.

While *CenDTect* is able to provide informative characterization of all 12 events in PCEL, time series anomaly detections are more limited in certain cases, detecting anomalies in only 7 cases with an anomaly score threshold of 3 as suggested in

Country	Start (y/m/d)	End (y/m/d)	Reference	Manual	Cen	Dates	Cen ASes	Cen Count	Cen Categories
Venezuela	19/01/21	19/01/29	IS/OONI	✓	✓	01/17-02/05	AS28089, AS20312	8	SOCIAL, NEWS
Tajikistan	19/04/23	19/05/08	IS	✓	✓	04/19-05/12	AS43197	15	MEDIA, NEWS, SOCIAL
Kazakhstan	19/05/09	19/05/10	GTR/Netblocks/IS	✓	✓	05/14-05/18	AS35168	5	NEWS, ANON, COMM
Venezuela	19/05/15	19/05/16	IS/NetBlocks	✓	✓	05/07-05/10	AS19192, AS28089	17	HUMANR, NEWS, CLTR
Tajikistan	19/05/21	19/05/21	GTR/IS/Euraisanet	✓	✓	05/18-05/29	AS43197	5	SOCIAL, NEWS, MEDIA
Ecuador	19/10/07	19/10/08	IS	✓	✓	10/08-10/10	AS28006	6	HUMANR, CRIT, FILE
Guinea	20/05/20	20/05/20	GTR/Access Now	✓	✓	05/20-05/22	AS38266	18	BLOG, SOCIAL, SEARCH
Burundi	20/05/20	20/05/21	CPJ/OONI	✓	✓	05/17-05/22	AS25429	19	SOCIAL, COMM, MEDIA
Belarus	20/08/09	20/08/11	GTR/Reuters	✓	✓	08/08-08/11	AS62197, AS202090, AS6697	70	COMM, ANON, BLOG, SEARCH
Azerbaijan	20/09/27	20/11/12	GTR/Netblocks	✓	✓	09/26-11/11	AS39397, AS29049, AS34170	33	COMM, SOCIAL, MEDIA
Guinea	20/10/23	20/10/27	GTR/Netblocks	✓	✓	10/23-10/26	AS37430	10	SOCIAL, COMM, NEWS, SEARCH
Tanzania	20/10/26	20/11/5	GTR/PRI/OONI	✓	✓	10/31-11/08	AS37027, AS33765	18	ANON, COMM, SOCIAL, NEWS

TABLE VI: **Potential Censorship Event List**—The list of reported censorship events that have signals in Censored Planet raw data (through manual analysis). All the duration *CenDTect* reports are in the same year of the event. Appendix A contains the domain categories and corresponding abbreviations. GTR is Google Transparency Report [31]; IS is Internet Society [71]. We report the detected duration, ASes, domains in the event clusters, and corresponding top domain categories.

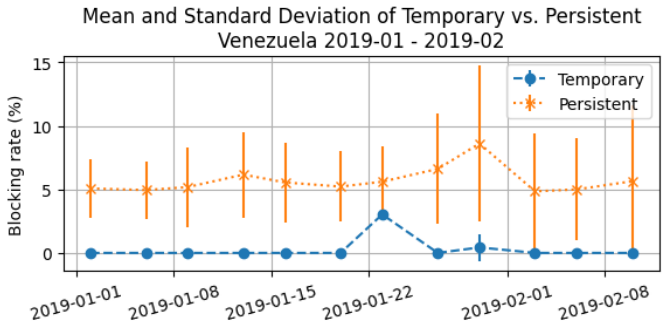


Fig. 8: **Blocking rate and in Venezuela around January 2019**—Temporary events in countries with high persistent blocking are hard to identify. In countries with persistent blocking, time series anomaly detection methods would require a low anomaly threshold to detect temporary events, which would however result in more cases of anomalies to manually explore.

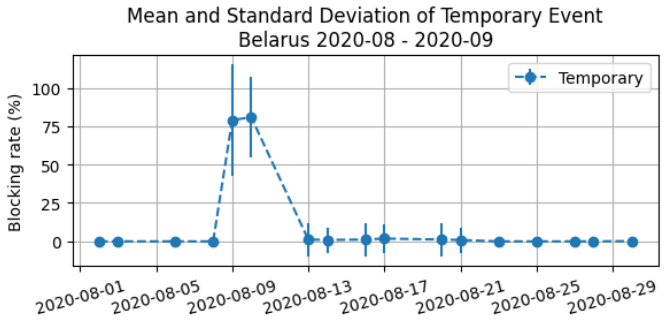


Fig. 9: **Blocking rate in Belarus around August 2020**—Time series anomaly detection methods work well in countries with no persistent blocking to obscure anomalous temporary events.

previous work [73]. The effectiveness of the bitmap anomaly detection method depends on several key factors, including the anomaly score threshold, the chosen time window for analysis, and the hyperparameters like lead and lag window sizes. As shown in Figure 8 and Figure 9, for countries with persistent blocking (such as Venezuela), reducing the threshold to identify new events involving a small blocklist will inevitably yield a higher number of anomaly occurrences requiring manual investigation. Therefore, it's worth noting that utilizing bitmap detection for temporary blocking events poses scalability challenges. Apart from the potential of suffering from false negatives, time series anomaly detection methods lack the human interpretability carried by decision trees. Knowing that there is an anomaly in a time series does not provide information on the exact geolocation, blocking method, or blocked domains.

Country	Start (y/m/d)	End (y/m/d)	Duration	Reference	Negative Reason
Cuba	19/02/24	19/02/24	1d	OONI	No Blocking
Algeria	19/02/21	19/02/21	1d	IS	No New Events
Nauru	19/04/03	19/04/08	6d	GTR/Access Now	No Blocking
Sudan	19/04/07	19/04/08	2d	IS/HRW	No New Events
India	19/04/22	19/04/22	3d	IS	No New Events
Sri Lanka	19/04/21	19/04/29	9d	GTR/BBC/CNN	No New Events
Benin	19/04/28	19/04/28	12h	GTR/Netblocks/ IS/OONI	No Measurements
Venezuela	19/04/30	19/04/30	1d	IS/CNN/Netblocks	No Blocking
Venezuela	19/05/15	19/05/15	1h	IS	No Measurements
Sri Lanka	19/04/21	19/05/06	16d	IS	No New Events
Liberia	19/06/07	19/06/08	2d	IS/CPJ	No Measurements
Indonesia	19/05/22	19/05/25	4d	IS/TechCrunch	No New Events
Venezuela	19/06/17	19/06/17	2h	GTR/Netblocks	No Measurements
Venezuela	19/06/19	19/06/19	1h	GTR/Netblocks	No Measurements
Ethiopia	19/06/22	19/06/27	6d	GTR/Fortune/ OONI	No Measurements
India	19/07/05	19/07/07	3d	IS	No New Events
Egypt	19/09/22	19/09/23	2d	Insights/OONI	No Measurements
Turkey	19/10/11	19/10/11	1d	IS	No New Events
Iraq	19/10/25	19/10/26	2d	IS/Al Jazeera	No Measurements
India	19/11/9	19/11/11	3d	IS	No New Events
Venezuela	19/11/16	19/11/16	2h	GTR/Netblocks	No Measurements
India	19/12/22	19/12/22	12h	IS	No Measurements
Burundi	19/12/13	19/12/16	4d	GTR/RegionWeek	No Measurements
Turkey	20/02/27	20/02/27	17h	GTR/Euronews	No New Events
Myanmar	20/03/23	20/03/24	2d	OONI	No New Events
Chad	21/03/01	21/03/06	6d	GTR	No Measurements

TABLE VII: **PCEL events that Censored Planet raw data does not cover**—GTR is short for Google Transparency Report [31]; IS is short for Internet Society [71].

C. Negative Cases

In the data mining process, *CenDTect* only outputs clusters whose representative tree is not "innocent" (indicating no blocking). Consequently, any spatiotemporal points lacking a covering cluster are considered by *CenDTect* as instances where the Internet remains untampered. However, validating negative cases presents challenges due to the substantial volume of measurements that exhibit no anomalies. Censorship is a rare phenomenon at a global scale, with prior research reporting that only 1%-2% of global measurements are anomalous [50], [73], [78]. As a result, we turn to the PCEL list to investigate the 26 negative cases. Our analysis reveals that these cases indeed qualify as true negatives, implying that despite the media coverage, the Censored Planet raw data does not exhibit any signals of blocking at these specific points in time and location.

As shown in Table VII, the negative cases in PCEL fall into 3 categories: 1) no measurements, 2) no blocking, and 3) no new events.

Protocol	Total	Important ISP		Conf. ISP	Temp. ISP	Uniq Temp.
HTTPS	1,069	478	205	146	47	11
HTTP	962	393	216	168	54	8
ECHO	283	98	82	51	13	1
DISCARD	276	88	75	42	9	0
DNS	1,166	292	171	92	28	0

TABLE VIII: **Number of events**—In 2022, HTTPS and DNS clusters were filtered using the prominent event discovery process described in Figure 7.

***No Measurements*:** The "no measurements" category arises when data is unavailable for a particular country and timespan. This could be because Censored Planet lacks vantage points in certain countries during that period (e.g., Burundi and Chad). Additionally, some events are too short-lived to be captured, such as the 12-hour social media blocking in India on December 22, 2019, or the 2-hour blocking in Venezuela on November 16, 2019. 46.15% (12/26) of the events fall into this category.

***No Blocking*:** All measurements for a country during a specific timespan show accessible measured domains. 11.5% (3/26) of the events fall into this category.

***No New Events*:** Lastly, the "no new events" category indicates the presence of persistent and ongoing blocking in the country during the specified timespan. However, upon examination of the raw data, no new events corresponding to the reported incidents were found. As an illustration, we consider the case of Venezuela on April 30, 2019, where through manual analysis, signals of blocking were detected in AS23007 (Universidad de Los Andes), AS19192 (Universidad Central de Venezuela), and AS27957 (Banco Mercantil C.A.). AS19192 had vantage points in the Censored Planet HTTP(S) raw data in 2019 from January 3, 2019, to August 12, 2019. Throughout this period, we observe persistent organizational blocking in the form of one tree in *CenDTect*'s output covering 16 domains. However, these instances of blocking were identified as persistent organizational blocking in the *CenDTect* output, rather than representing new and temporary events as mentioned in news coverage. This makes up 42.31% (11/26) of the events.

In summary, the negative cases identified in PCEL are attributed to inherent data limitations. Due to Censored Planet's measurement schedule, vantage point and test domain selection, signals of blocking are not included in the raw data, thus preventing *CenDTect* from detecting these specific events.

D. Newly Discovered Events Outside PCEL

To demonstrate *CenDTect*'s ability to detect new events outside PCEL, we report *important unique temporary events* identified in Censored Planet's 12-month HTTP(S) measurements in 2022. Specifically, we focus on unique temporary events that diverge from persistent ISP blocking in the same region (later reported in Table X), if such blocking is present.

Our analysis of temporary blocking events in 2022 consists of 2,590 HTTP(S) clusters and 1,166 DNS clusters (Table VIII). After filtering, we identify 11 *unique ISP temporary* HTTP(S) events with a blocklist that differs from the top five categories of domains in persistent ISP blocking (see Table IX). Our analysis

Country	ASes	Count	Main Categories	Date	Evidence
Nepal	AS4007	15	SOCIAL, REL, GOV	Jan	
Venezuela	AS8048	8	ANON, NEWS, MEDIA	Feb	
Russia	AS28891 AS3216★ *AS60459**	77	NEWS, SOCIAL	Mar	War [6]
Sri Lanka	AS18001	7	SOCIAL, NEWS	Apr-May	Protect [30]
Estonia	AS3327 AS3249	5	NEWS, HUMANR, FILE	May	Russia War [29]
Zimbabwe	AS37204★	17	NEWS, GOV, HUMANR	May-June	
Uganda	AS20294	5	NEWS, HUMANR, FILE	July	
Burkina Faso	AS37721	7	NEWS, SOCIAL, ANON	Aug	
Zambia	AS36959★ AS37146	20	COMM,SOCIAL,ANON	Aug	Election [87]
Armenia	AS43733★ AS12297★	38	NEWS, COMM, ANON	Sep	Conflict [6]
Iran	AS42337★ *AS25184**	22	SOCIAL, ANON, BLOG	Sep-Oct	Protest [7]

TABLE IX: **Temporary Blocking Events**—Superscript ★ indicates the blocking of domains in the same categories are present in OONI's data. *Italic text** indicates organizational vantage points with same blocking behaviors with ISP blocking. The blocking types include TCP reset, timeout, and private IPs.

indicates in contrast to persistent ISP blocking (see §VI-B), temporary blocking often targets news media, social networks, and anonymization tools. For example, during the protests in Iran in September 2022, multiple anonymization tools were blocked. Moreover, several DNS over HTTPS canary domains, including `doh.opendns.com`, `doh-fi.blahdns.com`, and `mozilla.cloudflare-dns.com` were blocked. Our event confirmation suggests these events happened during periods of election, political unrest, protest, and war [7], [82], [88]. Four of the events in Table IX are not confirmed by OONI data or news coverage, showing *CenDTect*'s ability to detect events outside media coverage and highlighting the importance of rapid focus and collaboration with in-country experts, activists, journalists, and researchers from various fields, to fully leverage the potential of the data.

***False positives*:** We attempt to validate the 11 newly discovered events using news media, country-specific reports, and OONI's data (Table IX). Similar to the PCEL, our analysis is conducted as a best-effort and ad hoc process, aiming to collect as much information as possible from public sources. 63.64% (7/11) of temporary blocking instances were confirmed by these sources. The absence of news media or OONI data doesn't necessarily indicate false positives; it could indicate events not yet known to the community. *CenDTect*'s data-driven perspectives can assist researchers and activists in confirming censorship events as well as discovering new events.

VI. Findings

In total, we discover 15,360 HTTP(S) clusters in 192 countries from January 2019 to December 2022, and 1,166 DNS clusters from August 2022 to December 2022 in 77 countries.

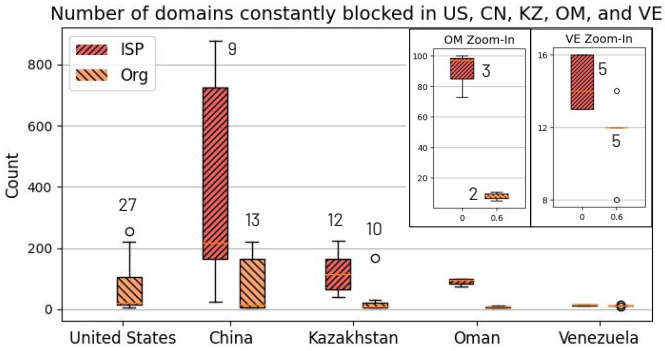


Fig. 10: **Average constant HTTP(S) blocking in 5 countries**—The boxplot shows the range of the number of domains blocked. The numbers on the boxes indicate the number of clusters in each country.

Through data mining, *CenDTect* refines 6 Terabytes of data into 238 Megabytes. Our finding highlights the prevalence of organizational blocking, the heterogeneity of blocking within countries, as well as 32 countries with persistent ISP-level blocking.

The Censored Planet dataset covers vantage points in diverse organizations, such as companies, schools, universities, hospitals, governmental entities, and public Wi-Fi providers, each employing their unique internet filters. While these filters are valuable for gaining insights into global filtering deployment, this research primarily centers on ISP-level blocking rather than organizational filtering. It is worth noting that this study does not furnish evidence regarding the specific actors responsible for the blocking, but instead, it offers insights into the types of ASes experiencing blocking.

A. Censorship Characteristics

***Heterogeneous blocking within countries*:** Our results show that on a global scale, censorship behavior is heterogeneous within countries. Many prior works report censorship based on the blocking ratio of the entire country [50], [59], [81], or calculate the blocking ratio of each Autonomous System (AS) in a given country and weigh each AS by assuming that the overall blocking percentage of a country at a specific scan date should be representative [73], we show that these practices are error-prone. As shown in Figure 10, we observe countries such as China, Kazakhstan, Oman, and Venezuela, exhibit ISP-level, persistent blocking behavior, which can be distinguished from organizational blocking based solely on the magnitude of blocking. We observe that for countries like Egypt, Russia, and India, the blocking behavior is heterogeneous even in ISP-level censorship. Therefore, researchers studying censorship should be aware of the heterogeneity of blocking behavior within countries and ASes, and avoid making generalizations based on country-level blocking ratios alone.

***Organizational Blocking in "free" countries*:** Some previous research documented instances of censorship in countries categorized as "Free" by Freedom House [59], [73], [81]. Based on *CenDTect*'s output, we believe that this type of censorship is mainly organizational blocking. For example,

FH	Country	ASes	Domain Count	Top Categories	Type	Confirmation
Not Free	Afghanistan	AS131284(ℍ★), AS55330(ℍ)	> 100	LGBT, PORN, GMB, AD	RST, 423	
	Azerbaijan	AS29049(ℍ★)	> 100	AD, PORN, XED	RST	OR [88]
	Bahrain	AS5416(ℍ)	> 100	FILE, MEDIA, PORN	RST	
	China	AS4837(ℍ★), AS56048(ℍ★), AS9929(ℍ,𝔻), AS58466(ℍ,𝔻), AS56040(𝔻★), AS56046(ℍ,𝔻), AS24355(𝔻), AS17621(𝔻★)...	> 800	ANON, BLOG, NEWS, SEARCH	RST, CERT	Studies [34], [51]...
	Egypt	AS24835(ℍ★)	> 10	COMM, ANON, MEDIA, BLOG	TO	OR [25]
	Iran	AS42337(ℍ★), AS205647(ℍ★), AS25184(ℍ★), AS39501(ℍ★), AS58224(ℍ★), AS39308(𝔻★),..	> 800	ANON, SOCIAL, HUMANR, NEWS	TO	Studies [5], [9], [20]
	Jordan	AS48832(ℍ★), AS8697(ℍ), AS8697(ℍ)	> 60	NEWS, MEDIA, ANON, GMB	RST, TO	
	Libya	AS37284(ℍ)	> 90	PORN, LGBT, NEWS, HUMANR	RST	
	Kazakhstan	AS41798(ℍ,𝔻★), AS21299(ℍ★), AS9198(𝔻★)	> 100	ANON, GMB, PORN, HUMANR	RST, TO,BP	Studies [72], [76], [92]
	Myanmar	AS136255(ℍ★), AS136255(ℍ★), AS136255(ℍ★)	> 50	NEWS, ANON, SEAECH, HUMANR	RST	Studies [8], [72], [92]
	Oman	AS28885(ℍ★), AS50010(ℍ)	> 100	ANON, COMM, SOCIAL	RST, BP	
	Pakistan	AS38193(ℍ,𝔻), AS17557(ℍ★), AS9260(𝔻), AS45773(𝔻)	> 40	LGBT, ANON, GMB, DATE	RST, LOCAL	Studies [48]
	Qatar	AS8781(ℍ,𝔻★)	> 20	ANON, SOCIAL, AD	RST, CERT	
	Russia	AS8402(ℍ★), AS20485(ℍ★), AS31261(ℍ), AS196695(ℍ) AS31261(ℍ), AS12389(𝔻★), AS34757(𝔻★), ...	> 80	ANON, PORN, BLOG, GMB	RST, TO, CERT, BP, 451	Studies [63], [85]
	Saudi Arabia	AS35753(ℍ★), AS25233(ℍ★), AS14754(ℍ), AS29684(ℍ), AS35819(ℍ)	> 100	MEDIA, PORN, GMB, ANON	RST, TO, BP	
	Tanzania	AS37027(ℍ★), AS33765(ℍ★)	> 200	HACK, ANON, PORN, GMB	RST, TO	
	Thailand	AS45458(ℍ★), AS7470(ℍ★), AS7693(ℍ), AS4750(𝔻★), AS9931(𝔻★),...	> 80	GMB, HUMANR, ANON	RST, BP, CERT	OR [66]
	Turkey	AS34984(ℍ★), AS15924(ℍ),...	> 100	LGBT, PORN, GMB, ANON	RST, TO, BP	
	Turkmenistan	AS20661(ℍ,𝔻★)	> 47	SEARCH, COMM, ANON, BLOG	RST, CERT	Studies [52]
	UAE	AS5384(ℍ★), AS15802(ℍ★)	> 600	ANON, NEWS, LGBT, COMM	RST, BP	
	Vietnam	AS7552(ℍ★), AS135905(ℍ★), AS45903(ℍ)	> 10	NEWS, GAMB, COMM, HUMANR	TO	OR [3]
Partly Free	Armenia	AS12297(ℍ★)	> 80	ANON, COMM, LGBT, NEWS	RST	
	Bangladesh	AS23688(ℍ★), AS24323(ℍ), AS23956(ℍ★), AS55492(ℍ★), AS63969(ℍ★), AS55492(ℍ), AS59362(ℍ), AS17806(ℍ)...	> 80	GMB, PORN, NEWS	RST, TO, BP	
	Bolivia	AS27882(ℍ★)	> 10	ANON, GMB	RST	
	India	AS55824(ℍ,𝔻★),AS9498(ℍ★),AS24186(ℍ★),AS4755(𝔻★)	> 100	ANON, GMB, AD, NEWS	RST, BP, CERT	Studies [70], [89]
	Indonesia	AS7713(ℍ★), AS17451(ℍ,𝔻★), AS9341(𝔻), AS9905(𝔻), AS38758(𝔻),...	> 100	LGBT, ANON, HUMANR, GAM	CERT, BP, LOCAL	OR [91]
	Kuwait	AS21050(𝔻★), AS9155(ℍ)	> 20	ANON, SOCIAL, HACK, GAM	BP	
	Malaysia	AS9930(ℍ,𝔻★), AS9534(ℍ★), AS38182(ℍ), AS4788(ℍ)	> 20	CRIT, COMM, PORN	RST, TO	OR [42]
	Nigeria	AS37282(ℍ★), AS29091(ℍ)	> 200	PORN, ANON, HUMANR	BP, TO	
	Philippines	AS23944(ℍ★), AS55821(ℍ), AS135423(𝔻), AS9658(ℍ)	> 90	ANON, REL, GAM, PORN	RST, LOCAL	OR [28]
Free	Poland	AS12741(ℍ★), AS5617(ℍ★)	> 100	HACK, NEWS	RST	
	Romania	AS8708(ℍ★), AS12302(ℍ★), AS2614(ℍ), AS8953(ℍ), AS2614(ℍ), AS31313(ℍ)	> 100	ANON, GMB, NEWS, HUMANR	RST, BP	

TABLE X: **Countries with persistent ISP blocking**—We report country by Freedom House categories, and corresponding ASes, number of domains in the blocklist, top blocked domain categories, type of blocking, and confirmation (if any). ℍ, 𝔻 denotes HTTPS and DNS blocking, respectively. ★ indicates the blocking of domains in the same categories are present in OONI's data. Blocking types include TCP reset (RST), connection timeout (TO), blockages (BP), 423 (Locked) and 451 (Unavailable For Legal Reasons) status code, invalid certificates for the queried domain (CERT, indicating DNS manipulation), private IP (LOCAL). OR is short for OONI reports [55]. The domain categories and corresponding abbreviations can be found in Table XI.

from *CenDTect*'s output, we see that Edith Cowan University in AS7575 blocks anonymization and circumvention tools such as HMA, Megaproxy, and StrongVPN. While organizational blocking falls under our definition of censorship (§II-C), it is crucial to distinguish between ISP and organizational blocking practices, as they impact different proportions of the population and carry different significance to global Internet censorship. In the same vein, Brown et al. [11] train a supervised model for DNS manipulation detection, assuming that blocking recorded in the United States by Censored Planet is noise since there is no national-level censorship. Our results show that organizational blocking is prevalent on a global scale. In total, we discover 14 free countries with organizational blocking, such as the United States, Australia, Canada, and Japan. In the United States alone, we identify 59 clusters for organizational blocking.

B. Persistent Blocking within ISPs

We present a comprehensive analysis of persistent ISP blocking in 32 countries from January 2019 to December 2022 for HTTP(S) and August 2022 to December 2022 for DNS using *CenDTect*. Table X lists 21 countries designated as "Not Free," 9 countries designated as "Partly Free," and 2 countries designated as "Free" by the Freedom on the Net Report [35]. To our knowledge, this is the first report on persistent ISP blocking on a global scale. To ensure accuracy, we manually verified that all the IP organizations associated with events in Table X correspond to ISPs and we report the AS numbers instead of the names of specific ISPs. We also report the approximate number of unique domains blocked since the blocklists usually differ among ISPs in the same country despite they might share similar blocked categories. Additionally, Censored Planet's domain test list is dynamic (Figure 2), so reporting an approximate count of blocked domains is more meaningful than an exact count for persistent blocking.

We confirm our findings from multiple angles, including previous country-specific studies on Internet censorship [8], [21], [34], [51], [52], [63], [72], [72], [76], [85], [92], [92], OONI reports [55], and OONI's globally-collected data. OONI provides volunteers with the flexibility to customize their domain test list and provide a default test list (the Citizen Lab Test list [43]) if the volunteers opt in. The two open-access censorship measurement platforms differ in their test lists, vantage points, measurement timing, and measurement frequency. Therefore, we consider an AS as "confirmed by OONI" if there exist anomalous measurements for domains *in the same categories* as those domains that are discovered by *CenDTect*, and denote such an AS with a star (★).

While some countries, such as China [21], [21], [34], [47], [51], [84], Russia [54], [63], [64], [85], [86], Iran [5], [9], and Kazakhstan [8], [72], have well-documented Internet censorship mechanisms that exhibit persistent blocking practices, other countries, like Nepal and Romania, lack such studies. Thus, our research underscores the importance of having an automated tool for detecting censorship events.

VII. Limitations

In this work, we focused on data collected by Censored Planet. *CenDTect* misses certain types of events due to some inherent limitations of Censored Planet's data collection. Out of the 114 events, 76 are Internet blackouts, which Censored Planet is not equipped to measure, and 26 are short-lived events beyond its scanning and domain test list coverage. Using complementary datasets that improve the frequency of data collection, protocol coverage and test list coverage may enable *CenDTect* to uncover more events.

The PCEL list is generated through a best-effort manual collection, which may introduce biases, such as relying solely on English media and potentially missing some reports. It captures instances of temporary blocking caused by political unrest or election-related events with media coverage, but may not represent all Internet censorship cases. Additionally, news reports rarely offer a comprehensive list of blocked domains during specific events, further complicating the evaluation. Despite these limitations, PCEL serves as a basis for validation, verifying the effectiveness of the event discovery model and Censored Planet's data quality.

VIII. Discussion and Conclusion

Our work represents one of the first attempts to apply machine learning techniques to censorship event discovery. The simplicity and interpretability of *CenDTect* make it a convincing option for censorship observatories and data consumers. We are currently assessing integrating *CenDTect* into Censored Planet, enabling more accessible and efficient censorship event analysis, leading to better-informed decisions and actions. The output of *CenDTect* can support the development of next-generation censorship measurements. While our event discovery is semi-automated, potential enhancements, such as NLP-based website summaries, may improve ISP identification. We hope to pave the way for integrating machine learning into future censorship research.

We introduce *CenDTect*, an automated system for detecting internet censorship events using decision trees and iterative clustering to identify domains with shared blocking policies. Our results showcase persistent ISP blocking on a global scale, as well as temporary blocking events during periods of political unrest, protest, and war. By providing interpretable results, *CenDTect* makes censorship data more accessible to censorship data consumers. Our findings demonstrate the potential of automated censorship detection, while also highlighting the importance of collaboration with in-country experts and researchers from various fields to fully utilize the data.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments on the paper. We also thank Chad Sharp for his valuable feedback on DBSCAN optimization, and Arham Jain for his feedback on selecting popular ASes. This work was supported by the Defense Advanced Research Projects Agency under Agreement No. HR00112190127 and National Science Foundation grant CNS-2141512 and CNS-2237552.

References

[1] AccessNow, "Keepiton: Fighting internet shutdowns around the world," 2023, https://www.accessnow.org/campaign/keepiton/.

[2] G. Aceto, A. Botta, A. Pescapè, N. Feamster, M. Faheem Awan, T. Ahmad, and S. Qaisar, "Monitoring internet censorship with ubica," in *Traffic Monitoring and Analysis: 7th International Workshop, TMA 2015, Barcelona, Spain, April 21-24, 2015. Proceedings 7*. Springer, 2015, pp. 143–157.

[3] Anonymous, K. Koh, and S. Nurliza Samsudin, "imap state of internet censorship report 2022 - vietnam," 2022, https://ooni.org/post/2022-state-of-internet-censorship-vietnam/.

[4] APNIC, "Customers per as measurements," 2023, https://stats.labs.apnic.net/aspop.

[5] S. Aryan, H. Aryan, and J. A. Halderman, "Internet censorship in iran: A first look." in *FOCI*, 2013.

[6] BBC, "Armenia-azerbaijan: Almost 100 killed in overnight clashes," 2022, https://www.bbc.com/news/world-europe-62888891.

[7] ——, "Fury in iran as young woman dies following morality police arrest," 2022, https://www.bbc.com/news/world-middle-east-62930425.

[8] D. Beisembayeva, E. Papoutsaki, E. Kolesova, and S. Kulikova, "Social media, online activism and government control in kazakhstan," 2013.

[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 *USENIX Workshop on Free and Open Communications on the Internet*, 2020.

[10] L. Breiman, "Random forests," *Machine learning*, 2001.

[11] J. A. M. Brown, X. Jiang, V. Tran, A. N. Bhagoji, N. P. Hoang, N. Feamster, P. Mittal, and V. Yegneswaran, "Augmenting rule-based dns censorship detection at scale with machine learning," *arXiv preprint arXiv:2302.02031*, 2023.

[12] S. Burnett and N. Feamster, "Encore: Lightweight measurement of web censorship with cross-origin requests," in *Proceedings of the 2015 ACM conference on special interest group on data communication*, 2015, pp. 653–667.

[13] CAIDA, "As rank: A ranking of the largest autonomous systems (as) in the internet," 2023, https://asrank.caida.org/.

[14] Censys, "Attack surface management and data solutions," 2023, https://censys.io/.

[15] CitizenLab, "Citizen lab research," 2023, https://citizenlab.ca/category/research/.

[16] J. R. Crandall, M. Crete-Nishihata, and J. Knockel, "Forgive us our syns: Technical and ethical considerations for measuring internet filtering." in *NS Ethics@ SIGCOMM*, 2015, p. 3.

[17] DBIP, "Ip geolocation api & free address database," 2023, https://db-ip.com/.

[18] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman, "A search engine backed by internet-wide scanning," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 542–553.

[19] Z. Durumeric, E. Wustrow, and J. A. Halderman, "{ZMap}: Fast internet-wide scanning and its security applications," in *22nd USENIX Security Symposium (USENIX Security 13)*, 2013, pp. 605–620.

[20] K. Elmenhorst, B. Schütz, N. Aschenbruck, and S. Basso, "Web censorship measurements of http/3 over quic," in *Proceedings of the 21st ACM Internet Measurement Conference*, 2021, pp. 276–282.

[21] R. Ensafi, P. Winter, A. Mueen, and J. R. Crandall, "Analyzing the great firewall of china over space and time." *Proc. Priv. Enhancing Technol.*, 2015.

[22] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, 1996.

[23] V. L. P. et al., "Tranco: A research-oriented top sites ranking hardened against manipulation," 2023, https://tranco-list.eu/.

[24] L. Evdokimov and V. Ververis, "Identifying cases of DNS misconfiguration: Not quite censorship," 2017, https://ooni.org/post/not-quite-network-censorship/.

[25] L. Evdokimov, M. Xynou, M. El-Taher, H. Al-Azhary, and S. Mohsen, "The state of internet censorship in egypt," 2018, https://ooni.org/post/egypt-internet-censorship/.

[26] A. Filasto and J. Appelbaum, "OONI: Open Observatory of Network Interference," in *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2012.

[27] C. for Applied Internet Data Analysis (CAIDA), "Ioda | monitor macroscopic internet outages in near real-time," 2023, https://ioda.inetintel.cc.gatech.edu/.

[28] K. Francisco, K. Zhafri, R. Tan, S. Pacia, V. Lucero, S. N. Samsudin, and K. Koh, "imap state of internet censorship report 2022 - philippines," 2022, https://ooni.org/post/2022-state-of-internet-censorship-philippines/.

[29] FreedomHouse, "Estonia: Freedom on the net 2022 country report," 2023, https://freedomhouse.org/country/estonia/freedom-net/2022.

[30] GlobalVoices, "Sri lanka in crisis," 2022, https://globalvoices.org/special/sri-lanka-in-crisis/.

[31] Google, "Traffic and disruptions to google," 2022, https://transparencyreport.google.com/traffic/overview.

[32] R. Gupta and K. Kumar, "What missing the internet means in digital era: A case study of longest ever internet blackout in jammu & kashmir," 2020.

[33] N. P. Hoang, S. Doreen, and M. Polychronakis, "Measuring i2p censorship at a global scale," *arXiv preprint arXiv:1907.07120*, 2019.

[34] 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 *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 3381–3398.

[35] F. House, "Freedom on the net 2022," 2022, https://freedomhouse.org/sites/default/files/2022-10/FOTN2022Digital.pdf.

[36] J. S. Hunter, "The exponentially weighted moving average," *Journal of quality technology*, 1986.

[37] O. Initiative *et al.*, "Opennet initiative," 2010.

[38] ipinfo, "The trusted source for ip address data, leading ip data provider," 2023, https://ipinfo.io/.

[39] Y. Izza, A. Ignatiev, and J. Marques-Silva, "On explaining decision trees," *arXiv preprint arXiv:2010.11034*, 2020.

[40] B. Jones, R. Ensafi, N. Feamster, V. Paxson, and N. Weaver, "Ethical concerns for censorship measurement," in *Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research*, 2015, pp. 17–19.

[41] R. Kathuria, M. Kedia, G. Varma, K. Bagchi, and R. Sekhani, "The anatomy of an internet blackout: measuring the economic impact of internet shutdowns in india," 2018.

[42] K. Koh and S. N. Samsudin, "imap state of internet censorship report 2022 - malaysia," 2022, https://ooni.org/post/2022-state-of-internet-censorship-malaysia/.

[43] C. Lab and Others, "Url testing lists intended for discovering website censorship," 2014, https://github.com/citizenlab/test-lists.

[44] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of experimental social psychology*, 2013.

[45] M. Luckie, B. Huffaker, A. Dhamdhere, V. Giotsas, and K. Claffy, "As relationships, customer cones, and validation," in *Proceedings of the 2013 conference on Internet measurement conference*, 2013, pp. 243–256.

[46] G. Lyon, "Nmap security scanner," *línea] URL: http://nmap.org/[Consulta: 8 de junio de 2012]*, 2014.

[47] B. Marczak, N. Weaver, J. Dalek, R. Ensafi, D. Fifield, S. McKune, A. Rey, J. Scott-Railton, R. Deibert, and V. Paxson, "An analysis of {China's}{"Great}{Cannon"}," in *5th USENIX Workshop on Free and Open Communications on the Internet (FOCI 15)*, 2015.

[48] Z. Nabi, "The anatomy of web censorship in pakistan," in *3rd USENIX Workshop on Free and Open Communications on the Internet (FOCI 13)*, 2013.

[49] A. Narayanan and B. Zevenbergen, "No encore for encore? ethical questions for web-based censorship measurement," *Ethical Questions for Web-Based Censorship Measurement (September 24, 2015)*, 2015.

[50] A. A. Niaki, S. Cho, Z. Weinberg, N. P. Hoang, A. Razaghpanah, N. Christin, and P. Gill, "Iclab: A global, longitudinal internet censorship measurement platform," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 135–151.

[51] A. A. Niaki, N. P. Hoang, P. Gill, A. Houmansadr *et al.*, "Triplet censors: Demystifying great firewall's dns censorship behavior," in *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20)*, 2020.

[52] S. Nourin, V. Tran, X. Jiang, K. Bock, N. Feamster, N. P. Hoang, and D. Levin, "Measuring and evading turkmenistan's internet censorship: A case study in large-scale measurements of a low-penetration country," *arXiv preprint arXiv:2304.04835*, 2023.

[53] A. Now, "Access now," 2023, https://www.accessnow.org/blog/.

[54] K. Ognyanova, "In putin's russia, information has you: Media control and internet censorship in the russian federation," in *Censorship, surveillance, and privacy: Concepts, methodologies, tools, and applications*. IGI Global.

[55] OONI, "Research reports," 2022, https://ooni.org/reports/.

[56] ——, "Frequently asked questions," 2023, https://ooni.org/support/faq/.

[57] C. Partridge and M. Allman, "Ethical considerations in network measurement papers," *Communications of the ACM*, 2016.

[58] P. Pearce, R. Ensafi, F. Li, N. Feamster, and V. Paxson, "Augur: Internet-wide detection of connectivity disruptions," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 427–443.

[59] P. Pearce, B. Jones, F. Li, R. Ensafi, N. Feamster, N. Weaver, and V. Paxson, "Global measurement of {DNS} manipulation," in *26th USENIX Security Symposium (USENIX Security 17)*, 2017, pp. 307–323.

[60] C. Planet, "Dns data - satellite," 2021, https://docs.censoredplanet.org/dns.html.

[61] J. Postel, "Rfc0863: Discard protocol," 1983.

[62] ——, "Internet control message protocol," Tech. Rep., 1981.

[63] 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 Systems Security (NDSS) Symposium 2020*, 2020.

[64] R. Ramesh, R. Sundara Raman, A. Virkud, A. DirksenΔ, A. Huremagic, D. F. D. Rodenburg, R. Hynes, D. Madory, and R. Ensafi, "Network responses to russia's invasion of ukraine in 2022: A cautionary tale for internet freedom," in *USENIX Security Symposium*, 2023.

[65] RFC, "Rfc6895 domain name system (dns) iana considerations 2.3. rcode assignment," 2023, https://www.rfc-editor.org/rfc/rfc6895.

[66] S. N. Samsudin and K. Koh, "imap state of internet censorship report 2022 - thailand," 2012, https://ooni.org/post/2022-state-of-internet-censorship-thailand/.

[67] W. Scott, T. Anderson, T. Kohno, and A. Krishnamurthy, "Satellite: Joint analysis of {CDNs} and {Network-Level} interference," in *2016 USENIX Annual Technical Conference (USENIX ATC 16)*, 2016, pp. 195–208.

[68] N. Selaiha, "The fire and the frying pan: Censorship and performance in egypt," *TDR*, pp. 20–47, 2013.

[69] A. Sfakianakis, E. Athanasopoulos, and S. Ioannidis, "Censmon: A web censorship monitor," in *USENIX Workshop on Free and Open Communication on the Internet (FOCI)*, 2011.

[70] K. Singh, G. Grover, and V. Bansal, "How india censors the web," in *12th ACM Conference on Web Science*, 2020, pp. 21–28.

[71] I. Society, "Build, promote, and defenf the internet," 2023, https://www.internetsociety.org/.

[72] R. Sundara Raman, L. Evdokimov, E. Wurstrow, J. A. Halderman, and R. Ensafi, "Investigating large scale HTTPS interception in Kazakhstan," in *ACM Internet Measurement Conference (IMC)*, 2020, pp. 125–132.

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

[74] 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 *NDSS*, 2020.

[75] R. Sundara Raman, A. Virkud, S. Laplante, V. Fortuna, and R. Ensafi, "Advancing the art of censorship data analysis," *Free and Open Communications on the Internet (FOCI)*, 2023.

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

[77] G. Tauchen, "Diagnostic testing and evaluation of maximum likelihood models," *Journal of Econometrics*, 1985.

[78] E. Tsai, D. Kumar, R. Sundara Raman, G. Li, Y. Eiger, and R. Ensafi, "Certainty: Detecting dns manipulation at scale using tls certificates," *Proceedings on Privacy Enhancing Technologies*, 2023.

[79] B. VanderSloot, A. McDonald, W. Scott, J. A. Halderman, and R. Ensafi, "Quack: Scalable remote measurement of application-layer censorship," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 187–202.

[80] V. Ververis, M. Isaakidis, V. Weber, and B. Fabian, "Shedding light on mobile app store censorship," in *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, 2019, pp. 193–198.

[81] M. Wander, C. Boelmann, L. Schwittmann, and T. Weis, "Measurement of globally visible dns injection," *IEEE Access*, 2014.

[82] H. R. Watch, "Belarus: Unprecedented crackdown," 2021, https://www.hrw.org/news/2021/01/13/belarus-unprecedented-crackdown.

[83] L. Wei, N. Kumar, V. N. Lolla, E. J. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Assumption-free anomaly detection in time series." in *SSDBM*, 2005.

[84] X. Xu, Z. M. Mao, and J. A. Halderman, "Internet censorship in china: Where does the filtering occur?" in *Passive and Active Measurement: 12th International Conference, PAM 2011, Atlanta, GA, USA, March 20-22, 2011. Proceedings 12*. Springer, 2011, pp. 133–142.

[85] D. Xue, B. Mixon-Baca, A. Ablove, B. Kujath, J. R. Crandall, and R. Ensafi, "Tspu: Russia's decentralized censorship system," in *Proceedings of the 22nd ACM Internet Measurement Conference*, 2022, pp. 179–194.

[86] D. Xue, R. Ramesh, L. Evdokimov, A. Viktorov, A. Jain, E. Wustrow, S. Basso, and R. Ensafi, "Throttling twitter: an emerging censorship technique in russia," in *Proceedings of the 21st ACM Internet Measurement Conference*, 2021, pp. 435–443.

[87] M. Xynou and A. Filastò, "Zambia: Social media blocked amid 2021 general elections," 2022, https://ooni.org/post/2021-zambia-social-media-blocks-amid-elections/.

[88] M. Xynou and A. Geybulla, "Ooni measurements show ongoing internet censorship in azerbaijan," 2023, https://ooni.org/post/2023-azerbaijan-internet-censorship/.

[89] 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 *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 252–264.

[90] B. Zevenbergen, B. Mittelstadt, C. Véliz, C. Detweiler, C. Cath, J. Savulescu, and M. Whittaker, "Philosophy meets internet engineering: Ethics in networked systems research.(gtc workshop outcomes paper)," in *GTC Workshop Outcomes Paper)(September 29, 2015)*, 2015.

[91] K. Zhafri, P. P. Rasidi, D. Kristin, S. Nurliza, Samsudin, and K. Koh, "imap state of internet censorship report 2022 - indonesia," 2022, https://ooni.org/post/2022-state-of-internet-censorship-indonesia/.

[92] J. L. Zittrain, R. Faris, H. Noman, J. Clark, C. Tilton, and R. Morrison-Westphal, "The shifting landscape of global internet censorship," *Berkman Klein Center Research Publication*, no. 2017-4, pp. 17–38, 2017.

APPENDIX

A. Test Domains

Censored Planet's techniques, including Quack [79], Hyperquack [74], and Satellite/Iris [59], [78], rely on the same test list. This list is a combination of the Citizen Lab test list, which contains potentially censored domains collected globally, and the Tranco top 500 [23]. Censored Planet measured a total of 6,640 categorized domains from January 2019 to December 2022, as shown in Table XI. Figure 11 presents a heatmap of the occurrence of different categories in 100 measurements.

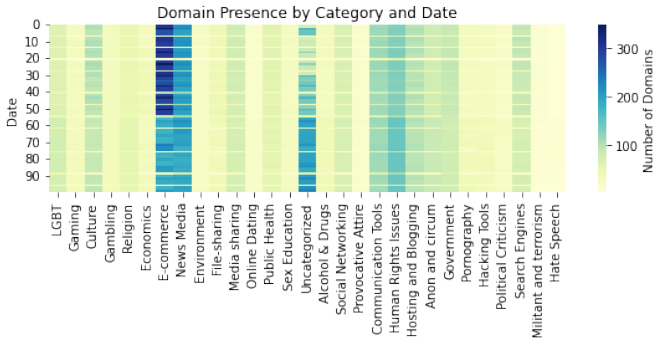


Fig. 11: **Heatmap of domain category occurrences from Jan 01 to April 14, 2022**—The y-axis shows the number of date occurrences in of 100 measurements.

Fig. 12: **Example blockpage**—"Situs Terlarang" ("Forbidden Site"). Served by Hypernet, an ISP in AS38758, Indonesia.

TABLE XI: Categorical breakdown of the domain test list used by Censored Planet from Jan 2019 to Dec 2022.

Category	Abbreviation	Count
Alcohol & Drugs	AD	48
Anonymization and circumvention	ANON	109
Communication Tools	COMM	176
Culture	CLTR	474
Environment	ENV	44
E-commerce	SHOP	1438
Economics	ECON	128
File-sharing	FILE	122
Gambling	GMB	54
Gaming	GAME	124
Government	GOV	193
Hacking Tools	HACK	75
Hate Speech	HS	17
History arts and literature	HIST	9
Hosting and Blogging Platforms	BLOG	224
Human Rights Issues	HUMANR	361
Illegal	ILEG	294
Intergovernmental Organizations	IGO	22
LGBT		105
Media sharing	MEDIA	354
Miscellaneous content	MISC	187
News Media	NEWS	825
Online Dating	DATE	27
Political Criticism	CRIT	74
Pornography	PORN	129
Provocative Attire	PA	26
Public Health	HLTH	109
Religion	REL	74
Search Engines	SEARCH	294
Sex Education	XED	38
Social Networking	SOCIAL	464
Terrorism and Militants	TERR	21

B. Blockpages

As shown in Figure 12, a blockpage is a webpage that clearly informs the user that the intended website has been intentionally blocked and may cite relevant legal justifications for the blocking [50], [56], [74]. This type of censorship, known as *overt censorship* [50], is transparent, and the ISP's intention is clear. Some HTTP status codes such as 451 (Unavailable For Legal Reasons) have the same effect as a blockpage. This form of censorship can be identified with high confidence. However, our detection goal extends beyond blockpage-serving censorship because prior research shows that more than 80% of manipulations are deployed without returning blockpages [78].

Prior work [74], [78] provides HTML fingerprints of censorship blockpages, categorized by the deployer (country, ISPs, and organizations) and middlebox vendor (Fortinet, Cisco, SkyDNS *etc*). As discussed later in §III-D, we use those blockpage fingerprints as a good source for identifying overt censorship.