ICLab: A Global, Longitudinal Internet Censorship Measurement Platform

Arian Akhavan Niaki*† Shinyoung Cho*†‡ Zachary Weinberg*§
Nguyen Phong Hoang‡ Abbas Razaghpanah‡ Nicolas Christin§ Phillipa Gill†

†University of Massachusetts, Amherst
{arian, shicho, phillipa}@cs.umass.edu

‡Stony Brook University
{shicho, nghoang, arazaghpanah}@cs.stonybrook.edu

§Carnegie Mellon University
{zackw, nicolasc}@cmu.edu

***Abstract*—Researchers have studied Internet censorship for nearly as long as attempts to censor contents have taken place. Most studies have however been limited to a short period of time and/or a few countries; the few exceptions have traded off detail for breadth of coverage. Collecting enough data for a comprehensive, global, longitudinal perspective remains challenging.**

In this work, we present ICLab, an Internet measurement platform specialized for censorship research. It achieves a new balance between breadth of coverage and detail of measurements, by using commercial VPNs as vantage points distributed around the world. ICLab has been operated continuously since late 2016. It can currently detect DNS manipulation and TCP packet injection, and overt "block pages" however they are delivered. ICLab records and archives raw observations in detail, making retrospective analysis with new techniques possible. At every stage of processing, ICLab seeks to minimize false positives and manual validation.

Within 53,906,532 measurements of individual web pages, collected by ICLab in 2017 and 2018, we observe blocking of 3,602 unique URLs in 60 countries. Using this data, we compare how different blocking techniques are deployed in different regions and/or against different types of content. Our longitudinal monitoring pinpoints changes in censorship in India and Turkey concurrent with political shifts, and our clustering techniques discover 48 previously unknown block pages. ICLab's broad and detailed measurements also expose other forms of network interference, such as surveillance and malware injection.

I. Introduction

For the past 25 years, the Internet has been an important forum for people who wish to communicate, access information, and express their opinions. It has also been the theater of a struggle with those who wish to control who can be communicated with, what information can be accessed, and which opinions can be expressed. National governments in particular are notorious for their attempts to impose restrictions on online communication [29]. These attempts have had unintentional international consequences [9], [15], [56], [78], and have raised questions about export policy for network management products with legitimate uses (*e.g.,* virus detection and protection of confidential information) that can also be used to violate human rights [26].

The literature is rich with studies of various aspects of Internet censorship [6], [7], [8], [9], [10], [15], [17], [23], [25], [26], [34], [36], [37], [43], [48], [56], [61], [64], [65], [66], [72], [76], [78], [79], [88], [89] but a global, longitudinal baseline of censorship covering a variety of censorship methods remains elusive. We highlight three key challenges that must be addressed to make progress in this space:

Challenge 1: Access to Vantage Points. With few exceptions,[1] measuring Internet censorship requires access to "vantage point" hosts within the region of interest.

The simplest way to obtain vantage points is to recruit volunteers [37], [43], [73], [80]. Volunteers can run software that performs arbitrary network measurements from each vantage point, but recruiting more than a few volunteers per country and retaining them for long periods is difficult. Further, volunteers may be exposed to personal risks for participating in censorship research.

More recently, researchers have explored alternatives, such as employing open DNS resolvers [66], [72], echo servers [79], Web browsers visiting instrumented websites [17], and TCP side channels [32], [65]. These alternatives reduce the risk to volunteers, and can achieve broader, longer-term coverage than volunteer labor. However, they cannot perform arbitrary network measurements; for instance, open DNS resolvers can only reveal DNS-based censorship.

Challenge 2: Understanding What to Test. Testing a single blocked URL can reveal that a censorship system exists within a country, but does not reveal the details of the censorship policy, how aggressively it is enforced, or all of the blocking techniques used. Even broad test lists, like those maintained by the Citizen Lab [22], may be insufficient [28]. Web pages are often short-lived, so tests performed in the present may be misleading [84].

Challenge 3: Reliable Detection. Censors can prevent access to content in several different ways. For instance, censors may choose to supply "block pages" for some material, which explicitly notify the user of censorship, and mimic site outages for other material (see §II-C) [27], [43].

Many recent studies focus on a single technique [17], [32], [65], [66], [72], [79]. This is valuable but incomplete, because censors may combine different techniques to filter different types of content.

As the Internet evolves and new modes of access appear (*e.g.,* mobile devices), censorship evolves as well, and monitoring systems must keep up [6], [59]. Ad-hoc detection strategies without rigorous evaluation are prone to false positives [89].

*Authors contributed equally

[1]China filters inbound as well as outbound traffic, making external observation simpler.

DOI 10.1109/SP40000.2020.00014

For example, detecting filtering via DNS manipulation requires care to deal with CDNs [66], [72] and detection of block pages requires taking regional differences in content into account [48].

A. Contributions

We present ICLab, a censorship measurement platform that tackles these challenges. ICLab primarily uses commercial Virtual Private Network servers (VPNs) as vantage points, after validating that they are in their advertised locations. VPNs offer long-lived, reliable vantage points in diverse locations, but still allow detailed data collection from all levels of the network stack. ICLab also deploys volunteer-operated devices (VODs) in a handful of locations.

ICLab is extensible, allowing us to implement new experiments when new censorship technologies emerge, update the URLs that are tested over time, and re-analyze old data as necessary. To date ICLab has only been used to monitor censorship of the web, but it could easily be adapted to monitor other application-layer protocols (*e.g.,* using techniques such as those in Molavi Kakhki et al. [59]). Besides ICLab itself, and its collected data, we offer the following contributions:

Global, longitudinal monitoring. Since its launch in 2016, ICLab has been continuously conducting measurements in 62 countries, covering 234 autonomous systems (ASes) and testing over 45,000 unique URLs over the course of more than two years. The platform has detected over 3,500 unique URLs blocked using a variety of censorship techniques. We discuss our discoveries in more detail in Section V.

Enhanced detection accuracy. ICLab collects data from all levels of the network stack and detects multiple different types of network interference. By comparing results across all the detection techniques, we can discover inaccuracies in each and refine them. We have eliminated all false positives from our block page detector. DNS manipulation detection achieves a false positive rate on the order of 10^{-4} when cross-checked against the block page detector (see Section IV-A). Similar cross-checking shows a negligible false positive rate for TCP packet injection (see Section IV-B).

Semi-automated block page detection. We have developed a new technique for discovering both variations on known block pages and previously unknown block pages. These explicit notifications of censorship are easy for a human to identify, but machine classifiers have trouble distinguishing them from other short HTML documents expressing an error message. Existing systems rely on hand-curated sets of regular expressions, which are brittle and tedious to update.

ICLab includes two novel machine classifiers for short error messages, designed to facilitate manual review of groups of suspicious messages, rather than directly deciding whether each is a block page. Using these classifiers we discovered 48 previously undetected block page signatures from 13 countries. We describe these classifiers and their discoveries in more detail in Section IV-C.

II. Background

Here we briefly review the techniques used to block access to information online, two different options for implementation, and how the censor's goals affect their implementation choices.

A. Network-level blocking techniques

Abstractly, all attempts to interfere with website access are man-in-the-middle (MITM) attacks on communications between a web browser and server. Depending on the location and configuration of their MITM devices, censors may interfere with traffic outside the borders of their own authority [15], [36], [78].

DNS manipulation. When visiting a website, the user's browser must first resolve the web server's IP address using DNS. DNS traffic is unencrypted, and less than 1% of it is authenticated [81]. Using either DNS servers they control, or packet injection from routers, censors can forge responses carrying DNS error codes such as "host not found" (NXDOMAIN), non-routable IP addresses, or the address of a server controlled by the censor [8], [91].

IP-based blocking. Once the browser has an IP address of a web server, it makes a TCP connection to that server. Censors can discard TCP handshake packets destined for IP addresses known to host censored content, reply with a TCP reset packet, or reroute them to a server controlled by the censor [46].

TCP packet injection. Censors can also allow the TCP handshake to complete, and then inject a packet into the TCP stream that either supersedes the first response from the legitimate server, or breaks the connection before the response arrives [82]. For unencrypted websites, this technique allows the censor to observe the first HTTP query sent by the client, and thus block access to individual pages [24].

Transparent proxy. Censors wishing to exercise finer control can use a "transparent proxy" that intercepts all HTTP traffic leaving the country, decodes it, and chooses whether or not to forward it [26]. Transparent proxies act as TCP peers and may modify HTTP traffic passing through, which makes them detectable [83]. They permit fine-grained decisions about *how* to block access to content. However, they are specific to unencrypted HTTP and cannot be used to censor traffic in any other protocol.

B. On-path and in-path censors

Hardware performing DNS manipulation, IP-based blocking, or TCP packet injection can be connected to the network in two different ways. It is not known which option is more commonly used [44], [80].

On-path equipment observes a copy of all traffic passing through a network link. It can react by injecting packets into the link, but cannot modify or discard packets that are already within the flow. While on-path techniques are relatively cheap and easy to deploy, detection is also easy, as injected packets appear alongside legitimate traffic.

In-path equipment operates on the *actual* traffic passing through the network link, and can inject, *modify, or discard*

packets. In-path equipment must operate at the line-rate of a backbone router, so it is more expensive and its features may be limited (*e.g.,* payload inspection may not be an option), but it is harder to detect.

C. Overt and covert censorship

Censorship's visible effects can be either *overt* or *covert*. In overt censorship, the censor sends the user a "block page" instead of the material that was censored. In covert censorship, the censor causes a network error that could have occurred for other reasons, and thus *avoids* informing the user that the material was censored. Censors may choose to be overt for some material and covert for other material. For instance, Yemen has been observed to overtly block pornography, which is illegal there, and to covertly block disfavored, but legal, political content [43].

Overt censorship can be accomplished with a transparent HTTP proxy, an injected TCP packet or DNS response that directs the browser to a server controlled by the censor, or by rerouting TCP traffic to a server controlled by the censor. Covert censorship can be accomplished with a transparent HTTP proxy, an injected TCP reset packet, an injected DNS error or non-routable address, or by discarding packets.

III. System Architecture

ICLab is a platform for measuring censorship of network traffic. As shown in Figure 1, it consists of a central control server and a set of vantage points distributed worldwide. The central server schedules measurements for each vantage point to perform, distributes test lists, and collects measurement results for analysis. The vantage points send and receive network traffic to perform each measurement, and upload their observations to the central server. All analysis is done centrally after the measurements have completed. Raw observations, including complete packet logs, are archived so that new analysis techniques can be applied to old data. There are two types of vantage points: volunteer-operated devices (VODs) configured by us and installed in locations of interest by our volunteers,[2] and VPN-based clients, which forward traffic through commercial VPN proxies located in various countries.

A. Design Goals

We designed ICLab to achieve the following properties:

Global, continuous monitoring. The techniques used for Internet censorship, the topics censored, and the thoroughness with which censorship is enforced are known to vary both among [16], [26], [28], [43], [44], [66] and within [1], [33], [51], [86], [88] countries. Therefore, the system should operate vantage points in multiple locations within each of many countries, to produce a comprehensive global view of censorship. Censorship may ratchet upward over time [29], [39], may change abruptly in response to political events [25] and may even cease after governing parties change [43]. Therefore, the system should perform its measurements continuously over a period of years, to detect these changes as they happen.

[2]Most of these are low-cost Raspberry Pi devices.

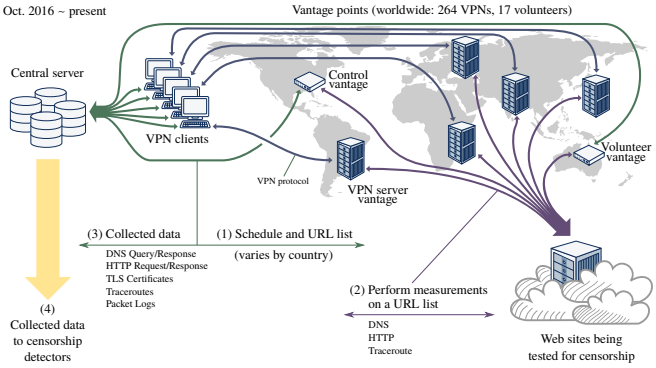

Fig. 1. Architecture of ICLab. (1) The central server sends a measurement schedule along with an associated test list to vantage points. (2) The vantage points perform measurements. (3) Collected data is uploaded to the central server. (4) Censorship detection is done centrally.

Reproducible and extensible. The basic techniques for censoring network traffic (described in §II-A) are well-known [80], [82] but new variations appear regularly [6], [36]. The short lifetime of "long tail" content means that the current content of a website may bear no relationship to what it was when it was originally censored [84]. Therefore, the system needs to be extensible with new types of measurement, and should record as much information as possible with each measurement (*e.g.,* packet traces and detailed contextual information).

Minimal risk to volunteers. Censorship monitoring involves accessing material that is forbidden in a particular country, from that country, and provoking a response from the censor. The response we expect is one of the MITM attacks described in §II-A, but legal or extralegal sanctions aimed at the volunteer operating the vantage point are also possible. The risk may be especially significant for volunteers already engaged in human rights reporting or advocacy. Use of commercial VPNs as vantage points is intended to mitigate these risks. VODs are only deployed in locations where we believe legal or extralegal sanctions are unlikely, and we obtain informed consent from the volunteers who operate them.

B. Vantage Points

Of ICLab's 281 vantage points, 264 are VPN-based, obtaining access to locations of interest via commercial VPN services. 17 vantage points are VODs. The measurement software is the same for both types of vantage; the only difference is that VPN-based vantages route their traffic through a VPN while performing measurements.

VPN-based vantages. ICLab uses VPN-based vantages whenever possible, because of their practical and ethical advantages. We do not need to recruit volunteers from all over the world, or manage physical hardware that has been distributed to them, but we still have unrestricted access to the network, unlike, for instance, phone or web applications [17], [37]. The VPN operator guarantees high availability and reasonable bandwidth, and they often offer multiple locations within a country. For 75% of the countries where we use VPN-based vantages, the

VPNs give us access to at least two ASes within that country (see Appendix A).

On the ethical side, a commercial VPN operator is a company that understands the risks of doing business in each country it operates in. It is unlikely that they would deploy a server in a country where the company or its employees might suffer legal or extralegal sanctions for the actions of its users.

A disadvantage of VPNs is that they only supply a lower bound on the censorship experienced by individuals in each country, because their servers are hosted in commercial data centers. There is some evidence that network censorship is less aggressively performed by data centers' ISPs than by residential ISPs [3], [88]. According to the CAIDA AS classification [18], 41% of the networks hosting our VPN-based vantages are "content" networks, which are the most likely to be subject to reduced levels of censorship. However, we have visibility into at least one other type of AS in 83% of the countries we can observe. In countries where we have both VPNs and VODs, we have observed identical block pages from both, indicating that all types of ASes are subject to similar blocking policies in those countries.

User-hosted VPNs (*e.g.,* Geosurf [42], Hola [47], Luminati [55]) would offer access to residential ISPs, but ICLab does not use them, as they have all the ethical concerns associated with VODs, with less transparency. Also, there are reports of illicit actions by the operators of these VPNs, such as deploying their software as a viral payload, and facilitating distributed denial of service (DDoS) attacks [58], making it even more unethical to use these services.

Since commercial VPNs' advertised server locations cannot be relied on [85], we validate their locations using round-trip time measurements (see Appendix B for details), and we only use the servers whose locations are accurately advertised.

Volunteer-operated device vantages. VODs are more difficult to keep running, and require a local volunteer comfortable with the risks associated with operating the device. Since ICLab does not collect personally identifiable information *about* the volunteers, our IRB has determined that this project is not human subjects research. However, we are guided by the principles of ethical human subjects research, particularly the need to balance potential benefits of the research against risks undertaken by volunteers. Most of our VODs have been deployed opportunistically through collaborations with NGOs and organizations interested in measuring Internet censorship from a policy perspective. For each deployed VOD, we maintain contact with the volunteer, and monitor the political situation in the country of deployment. We have deemed some countries too risky (for now) to recruit volunteers in (*e.g.,* Iran, Syria).

Breadth of coverage. As of this writing, ICLab has VPN-based vantage points in 55 countries, and volunteer-operated clients in 13 countries. 6 countries host both types of clients, so ICLab has vantage points in 62 countries overall. ICLab seeks to achieve both geographic and political diversity in its coverage. Table I summarizes our current geographic diversity by continent, and political diversity by a combination of two scores of political freedom, developed by Freedom House [39] and Reporters Without Borders [68] (see Appendix C).

TABLE I
COUNTRY COVERAGE OF ICLAB. The number of countries and ASes on each continent where we have vantage points with validated locations, since 2017. Oceania includes Australia. VPNs: virtual private network servers. VODs: volunteer-operated devices. NF, PF, F: of the countries with vantage points, how many are politically not free, partially free, or free (see Appendix C).

Continent	VPNs	VODs	Countries	ASes	NF	PF	F
Asia	64	4	14/32	54	5	7	2
Africa	9	10	9/72	19	1	6	2
N. America	87	1	5/17	81	0	1	4
S. America	9	0	5/20	6	1	3	1
Europe	83	2	27/42	64	1	5	21
Oceania	12	0	2/ 6	11	0	0	2
Total	264	17	62/189	234	8	22	32

It is easier to acquire access to vantage points in Europe, North America, and East Asia than in many other parts of the world. We have plans for expanded coverage in Africa and South America in the near future, via additional VPN services. It is also easier to acquire access to vantage points in "free" and "partially free" than "not free" countries, because it is often too risky for either VPN services or volunteer-operated devices to operate in "not free" countries. Expanding our coverage of "not free" countries is a priority for future development of ICLab, provided we can do it safely.

Internet censorship does happen in the "partly free" and "free" countries, and is not nearly as well documented as it is for specific "not free" countries (most notably China). Our broad coverage of these classes of countries gives us the ability to track changes over time, across the full spectrum of censorship policy, worldwide.

C. Test Lists

At present, ICLab's measurements are focused on network-level interference with access to websites. ICLab's vantage points test connectivity to the websites on three lists: the Alexa global top 500 websites (ATL) [5], the websites identified as globally sensitive by the Citizen Lab [22] and the Berkman Klein Center [11][3] (CLBL-G), and, for each country, the websites identified as locally sensitive in that country by Citizen Lab and Berkman Klein (CLBL-C). We only use the global top 500 sites from Alexa's ranking, because its "long tail" is unstable [54], [71]. All test lists are updated weekly.

ICLab has tested a total of 47,000 unique URLs over the course of its operation. Because all of the vantage points test ATL and CLBL-G, there is more aggregated data for these sites: 40% of our data is from sites on ATL, 40% from sites on CLBL-G, and 20% from sites on CLBL-C. Individual vantage points test anywhere from 3,000 to 5,700 URLs per measurement cycle, depending on the size of CLBL-C for the vantage point's country. This is by no means the complete set of sites blocked in any one country [28], and we have plans to broaden our testing, as described further in Section IX.

[3]The lists maintained by Citizen Lab and Berkman Klein are formally independent but have substantial overlap, so we combine them.

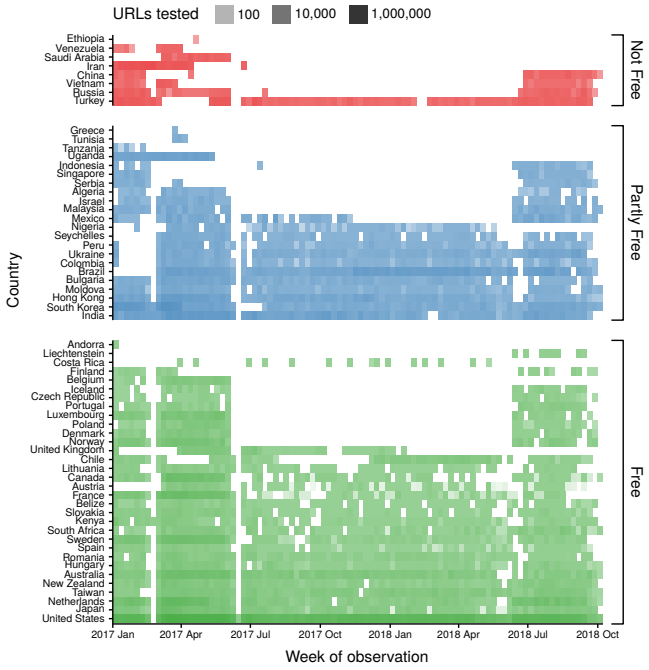

FIG. 2. MEASUREMENTS SINCE 2017 BY COUNTRY. For each of the 62 countries where we have, or had, vantage points since 2017, the total number of measurements per week.

D. Data Collection

A *measurement* of a URL is an attempt to perform an HTTP GET request to that URL, recording information about the results from multiple layers of the network stack: (1) The complete DNS request and response or responses for the server hostname (using both a local resolver and a public DNS resolver). (2) Whether or not a TCP connection succeeded. (3) For HTTPS URLs, the certificate chain transmitted by the server. (4) The full HTTP response (both headers and body). (5) A traceroute to the server. (6) A comprehensive packet trace for the duration of the measurement. This allows us to identify anomalies that would not be apparent from application-layer information alone. For instance, when packets are injected by on-path censors, we can observe both the injected packets and the legitimate responses they conflict with (see Section IV-B).

Each vantage point measures connectivity to all of the sites on its test list at least once every three days, on a schedule controlled by the central server. Depending on the size of the test list, a cycle of measurements typically runs for 1–2 hours.

Figure 2 depicts ICLab's measurements over time in each country. Operating ICLab over a multi-year period has not been easy; several outages are visible in Figure 2. For instance, we lost access to our vantage points in Iran in May 2017 due to a change in the international sanctions imposed on Iran, and we suffered a year-long, multi-country outage due to one VPN provider making configuration changes without notice. The latter incident led us to improve our internal monitoring and our tracking of VPN configuration changes.

Between January 2017 and September 2018, ICLab conducted 53,906,532 measurements of 45,565 URLs in 62 countries and 234 ASes. We publish our data for use by other researchers,[4] with periodic updates as we continue operation.

E. Control Nodes

Many tests of censorship rely on comparison of measurements between the vantage point and a "control" location, where there is not anticipated to be censorship. We repeat all the measurements performed by our vantage points on a *control node* located in an academic network in the USA. This network allows access to all the sites we test for accessibility. The control node has also suffered outages. In this paper, we use public data sets compiled by other researchers to fill in the gaps, as described in Section IV. We have since deployed three more control nodes in Europe, Asia and the USA to improve reliability and geographic diversity.

IV. CENSORSHIP DETECTION

Next, we describe how ICLab detects manipulated DNS responses (§IV-A), packets injected into TCP streams (§IV-B) and HTML-based block pages (§IV-C). All of ICLab's detection algorithms are designed to minimize both false negatives, in which a censored site is not detected, and false positives, in which ordinary site or network outages, or DNS load balancing are misidentified as censorship [32], [43], [48], [82].

A. DNS Manipulation

To access a website, the browser first resolves its IP address with a DNS query. To detect DNS manipulation, ICLab records the DNS responses for each measurement, and compares them with responses to matching DNS queries from our control node, and with DNS responses observed by control nodes OONI [37] operates. ICLab applies the following heuristics, in order, to the observations from the vantage point and the control nodes.

Vantage point receives two responses with different ASes. If a vantage point receives two responses to a DNS query, both with globally routable addresses, but belonging to two different ASes, we label the measurement as DNS manipulation. This heuristic detects on-path censors who inject a packet carrying false addresses [8]. Requiring the ASes to differ avoids false positives caused by a DNS load balancer picking a different address from its pool upon retransmission.

Vantage point receives NXDOMAIN or non-routable address. If a vantage point receives either a "no such host" response to a query (NXDOMAIN, in DNS protocol terms [13]), or an address that is not globally routable (*e.g.,* `10.x.y.z`) [12], but the control nodes consistently receive a globally routable address (not necessarily the same one) for the domain name, over a period of seven days centered on the day of the vantage point's observation, we label the test as DNS manipulation. The requirement for consistency over seven days is to avoid false positives on sites that have been shut down, during the period where a stale address may still exist in DNS caches.

Vantage point receives addresses from the same AS as control nodes. If a vantage point receives a globally routable

[4]Available online at https://iclab.org/.

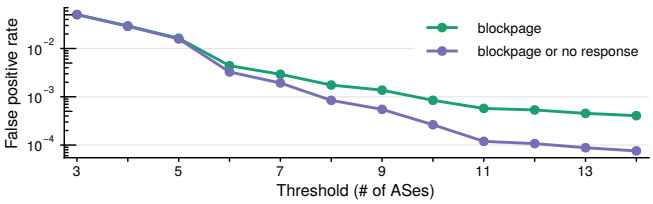

FIG. 3. DNS MANIPULATION FALSE POSITIVES. The false positive rate for the DNS manipulation detector, as a function of the threshold parameter θ.

address, and the control nodes also receive globally routable addresses assigned to the same AS (not necessarily the exact same address), we label the measurement as *not* DNS manipulation. Variation within a single AS is likely to be due to load-balancing over a server pool in a single location.

Vantage point and control nodes receive addresses in different ASes. The most difficult case to classify is when the vantage point and the control nodes receive globally routable addresses assigned to different ASes. This can happen when DNS manipulation is used to redirect traffic to a specific server (*e.g.,* to display a block page). However, it can also happen when a content provider or CDN directs traffic to data centers near the client [66].

We distinguish censors from CDNs using the observation that censors tend to map many blocked websites onto just a few addresses [9], [43]. If a set of websites resolve to a single IP address from the vantage point, but resolve to IPs in more than θ ASes from the control nodes, we count those websites as experiencing DNS manipulation. θ is a tunable parameter which we choose by cross-checking whether these measurements also observed either a block page or no HTTP response at all. Taking this cross-check as ground truth, Figure 3 shows how the false positive rate for DNS manipulation varies with θ. For the results in Section V, we use a conservative $\theta = 11$ which gives a false positive rate on the order of 10^{-4}.

B. TCP Packet Injection

Censors may also allow DNS lookup to complete normally, but then inject packets that disrupt the TCP handshake or subsequent traffic. ICLab detects this form of censorship by recording packet traces of all TCP connections during each test, and analyzing them for (1) evidence of packet injection, and (2) evidence of intent to censor (*e.g.,* block page content or TCP reset flags in injected packets). By requiring both types of evidence, we minimize false positives. Short error messages delivered by the legitimate server will not appear to be injected, and packets that, for innocuous reasons, appear to be injected, will not display an intent to censor.

Evidence of packet injection. If an end host receives two TCP packets with valid checksums and the same sequence number but different payloads, the operating system will generally accept the first packet to arrive, and discard the second [67]. An on-path censor can therefore suppress the server's HTTP response by injecting a packet carrying its own HTTP response (or simply an RST or FIN), timed to arrive first. Because ICLab records packet traces, it records both packets and detects a

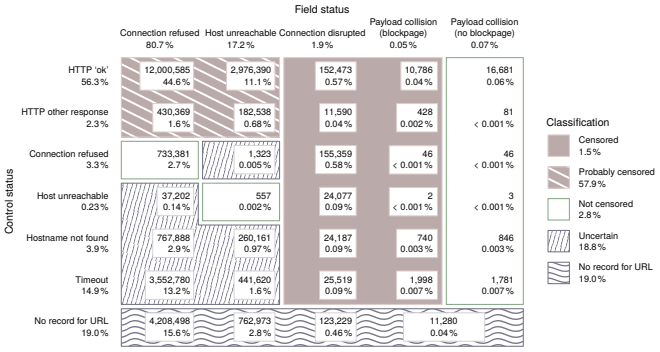

FIG. 4. Classification of packet anomalies by comparison to control observations.

conflict. This is not infallible proof of packet injection; it can also occur for innocuous reasons, such as HTTP load balancers that do not send exactly the same packet when they retransmit.

Intent to censor: RST, FIN, or block page. When we detect a pair of conflicting packets, we inspect them for evidence of intent to censor. An injected packet can disrupt/censor communication by carrying a TCP reset (RST) or close (FIN) flag, causing the client to abort the connection and report a generic error [24], [82]; or it can carry an HTTP response declaring the site to be censored (a "block page," discussed further in Section IV-C), which will be rendered instead of the true contents of the page the client requested [26], [48].

As with DNS manipulation, we compare each observation from a vantage point that shows evidence of packet injection, with matching observations from a control node. We apply the following heuristics, in order, to pairs of observations. The various outcomes of these heuristics are shown in Figure 4.

No matching control observation. When a TCP stream from the vantage point shows evidence of packet injection, but does not seem to correspond to any observation taken by the control node, we abandon any attempt to classify it. This is the "No record for URL" row of Figure 4.

This filtering is necessary because of a limitation in our packet trace analyzer. When a website transfers all of its traffic to another domain name, either via a CNAME record in DNS or using HTTP redirects, the trace analyzer cannot tell that TCP connections to the second domain name are associated with an attempt to test the first domain name. We conservatively do not consider these cases as censorship.

Packet collision after handshake, with RST or FIN. When a TCP stream from the vantage point shows evidence of collisions in TCP sequence numbers after successful completion of the three-way handshake, one side of the collision has its RST or FIN bit set and the other side has neither bit set, we label the measurement as censored by packet injection, regardless of what the control node observed. This is the "connection disrupted" column of Figure 4. We have high confidence that all of these are true positives.

Packet collision after handshake, with payload conflict. When a TCP stream from the vantage point shows evidence of TCP sequence number collisions after successful completion of

the three-way handshake, but neither side of the collision has the RST or FIN bit set, we inspect the contents of the packets for a block page signature, as described in §IV-C. We label the measurement as censored by packet injection only if a known block page signature appears in one of the packets. These cases are the "payload collision (blockpage)" and "payload collision (no blockpage)" columns of Figure 4. Again, we have high confidence that these are true positives and negatives.

Matching RST or ICMP unreachable instead of SYN-ACK. When a TCP SYN from the vantage point receives either a TCP RST or an ICMP unreachable packet, instead of a SYN-ACK, and the control node observes the same network error, we conclude the site is down for everyone, and label the measurement as *not* censored. These cases are the matching "connection refused" and "host unreachable" cells on the left-hand side of Figure 4, and we have high confidence that they are true negatives.

RST or ICMP unreachable instead of SYN-ACK, at vantage only. When a TCP SYN from the vantage point receives either a TCP RST or an ICMP unreachable packet in response, instead of a SYN-ACK, but the control node is able to carry out a successful HTTP dialogue, this *probably* indicates IP-based censorship observed by the vantage point. However, there are other possible explanations, such as a local network outage at the vantage point, or a site blocking access from specific IP addresses on suspicion of malice [57]. Manual spot-checking suggests that many, but not all, of these observations are censorship. These cases are labeled as "probable censorship" in Figure 4, and we discuss them separately in Section V.

Mismatched network errors, or timeout or DNS error at control node. When the vantage point and the control node both received a network error in response to their initial SYN, but not the same network error; when the control node's initial SYN received no response at all; and when the control node was unable to send a SYN in the first place because of a DNS error; we cannot say whether the measurement indicates censorship. These cases are the cells labeled "uncertain" in the lower left-hand corner of Figure 4. We are conservative and do not consider these as censorship in our analysis.

C. Block Page Detection and Discovery

Block page contents vary depending on the country and the technology used for censorship. Known block pages can be detected with regular expressions applied to the TCP payloads of suspicious packets, but these will miss small variations from the expected text, and are no help at all with unknown block pages.

Nonetheless, ICLab uses a set of 308 regular expressions to detect known block pages. We manually verified these match specific, known block pages and nothing else. 40 of them were developed by the Citizen Lab [21], 24 by OONI [37], 144 by Quack [79], and 100 by us.

Anomalous packets that do *not* match any of these regular expressions are examined for block page variations and unknown block pages, as described below; when we discover a block page that was missed by the regular expressions, we write new ones to cover them.

Self-contained HTTP response. To deliver a block page, the protocol structure of HTTP requires a censor to inject a single packet containing a complete, self-contained HTTP response. This packet must conflict with the first data packet of the legitimate response. Therefore, only packets which are both involved in a TCP sequence number conflict, and contain a complete HTTP response, are taken as candidate block pages for the clustering processes described next.

HTML structure clusters. The HTML tag structure of a block page is characteristic of the filtering hardware and software used by the censor. When the same equipment is used in many different locations, the tag structure is often an exact match, even when the text varies. We reduce each candidate block page to a vector of HTML tag frequencies (1 <body>, 2 <p>, 3 , *etc.*) and compare the vectors to all other candidate block pages' vectors, and to vectors for pages that match the known block page regular expressions. When we find an exact match, we manually inspect the matching candidates and decide whether to add a new regular expression to the detection set. Using this technique we discovered 15 new block page signatures in five countries.

Textual similarity clusters. Within one country, the legal jargon used to justify censoring may vary, but is likely to be similar overall. For example, one Indian ISP refers to "a court of competent jurisdiction" in its block pages, and another uses the phrase "Hon'ble Court" instead. Small variations like this are evidently the same page to a human, but a regular expression will miss them. We apply *locality-sensitive hashing* (LSH) [90] to the text of the candidate block pages, after canonicalizing the HTML structure. LSH produces clusters of candidate pages, centered on pages that do match the known block page regular expressions. As with the tag frequency vectors, we manually inspect the clusters and decide whether to add new regular expressions to the curated set. Using this technique, we discovered 33 new block page signatures in eight countries. An example cluster is shown in Appendix D.

URL-to-country ratio. To discover wholly unknown block pages we take each LSH cluster that is *not* centered on a known block page, count the number of URLs that produced a page in that cluster, and divide by the number of countries where a page in that cluster was observed. This is essentially the same logic as counting the number of websites that resolve to a single IP from a test vantage point but not a control vantage point, but we do not use a threshold. Instead, we sort the clusters from largest to smallest URL-to-country ratio and then inspect the entire list manually. The largest ratio associated with a newly discovered block page was 286 and the smallest ratio was 1.0.

V. Findings

Between January 2017 and September 2018, ICLab conducted 53,906,532 measurements of 45,565 URLs in 62 countries. Because we do not have continuous coverage of all

Vantage point DNS \ Public DNS	unmanipulated	NXDOMAIN	SERVFAIL	REFUSED	manipulated
unmanipulated	9,186,154	53,541	4,375	0	174
NXDOMAIN	8,554	1,477	3	0	0
SERVFAIL	5,436	4	75	0	0
REFUSED	2,000	0	0	0	0
manipulated	218	4	0	0	229

FIG. 5. COMPARISON OF DNS RESPONSES FOR THE SAME DOMAIN BETWEEN LOCAL AND PUBLIC NAMESERVERS FROM THE SAME VANTAGE POINT.

these countries (see §III-D), in this paper we present findings only for countries where we successfully collected at least three months' worth of data prior to September 2018. Among those countries, five stand out as conducting the most censorship overall: Iran, South Korea, Saudi Arabia, India, and Kenya. When considering specific subsets of our data, sometimes Turkey or Russia displaces one of these five.

A. Specific Results

We first present details of our observations for each of the three censorship techniques that we can detect.

DNS manipulation. We observe 15,007 DNS manipulations in 56 countries, applied to 489 unique URLs. 98% of these cases received NXDOMAIN or non-routable addresses.

Figure 5 compares DNS responses from a vantage point's local recursive resolver with those received by the same vantage point from a public DNS utility (*e.g.,* `8.8.8.8`). The upper left-hand cell of this chart counts cases where there is no DNS censorship; the other cells in the left-hand column count cases where censorship is being performed by the local DNS recursive resolver. The top rightmost cell counts the number of observations where censorship is being performed only by a public DNS utility, and the bottom rightmost cell counts cases where censorship is being performed by both a local recursive resolver and a public DNS utility. We observe censorship by public DNS utilities only for a few sites from Russia, Bulgaria, and Iran. The middle three columns could be explained as either censorship or as unrelated DNS failures.

Packet injection. We observe 19,493,925 TCP packet injections across 55 countries, applied to 11,482 unique URLs. However, after applying the filtering heuristics described in §IV-B, only 0.7% of these are definitely due to censorship: 143,225 injections, in 54 countries, applied to 1,205 unique URLs. (The numbers in Figure 4 are higher because they do not account for all the filtering heuristics.) Packet injections are usually used to disrupt a connection without delivering a block page; block pages are delivered by only 3.4% of the injections we attribute to censorship.

Another 15,589,882 packet injections—58% of the total—are network errors received instead of a SYN-ACK packet. These are described as "probable censorship" in Figure 4. They could indicate an in-path censor blocking hosts by IP address, but there are many other possible explanations. Our synthetic results (below) might be quite different if we were able to classify these more accurately.

Block pages. We observe 232,183 block pages across 50 countries, applied to 2,782 unique URLs. Iran presents block pages for 24.9% of the URLs it censors, more than any other country. In all of the countries we monitor, block pages are most likely to be used for URLs in the pornography and news categories (see below).

B. Synthetic Analysis

Combining observations of all three types of censorship gives us a clearer picture of what is censored in the countries we monitor, and complements missing events in each.

We use the "FortiGuard" URL classification service, operated by FortiNet [38], to categorize the contents of each test list. This service is sold as part of a "web filter" for corporations, which is the same software as a nation-state censorship system, but on a smaller scale. The URLs on all our lists, together, fall into 79 high-level categories according to this service; the 25 most common of these, for URLs that are censored at least once, are listed in Table IV, along with the abbreviated names used in other tables in this section.

Table II shows the three most censored categories of URLs for the five countries conducting the most censorship, based on the percentage of unique URLs censored over time. It is divided into four columns, showing how the results vary depending on which of our test lists are considered: all of them, only ATL, only CLBL-G, or only CLBL-C. Table VI in Appendix D continues this table with information about the countries ranked 6 through 15.

Iran takes first place in all four columns, and Saudi Arabia is always within the top three. The other countries appearing in Table II are within the top five only for some test lists. The top three categories blocked by each country change somewhat from list to list. For instance, pornography is much less prominent on the country-specific lists than on the global list. Iran's censorship is more uniformly distributed over topics than the other countries, where censorship is concentrated on one or two categories. These results demonstrate how the choice of test lists can change observations about censorship policy.

Table III shows the top five countries conducting the most censorship, for each of the three censorship techniques that ICLab can detect, with the top three categories censored with that technique. This shows how censors use different techniques to censor different types of content, as we mentioned in §I. For example, Turkey uses DNS manipulation for categories ILL and STRM, but uses block pages for PORN and NEWS.

Figure 6 shows how often the various blocking techniques are combined. For instance, in Iran we detect some URLs being redirected to a block page via DNS manipulation (comparing with Table III, we see that these are the URLs in the PORN and BLOG categories), but for many others, we detect only the block page. This could be because Iran uses a technique we cannot detect for those URLs (*e.g.,* route manipulation), or because our analysis of packet injection is too conservative (see Section IV-B).

TABLE II

CENSORSHIP BY TEST LIST AND CATEGORY. For each of the three test lists we use (see §III-C), the five countries censoring the most URLs from that list, the top three FortiGuard categories for their censored URLs (abbreviations defined in Table IV), and the percentage of URLs from that list that are censored.

Overall			Alexa Global (ATL)			Globally Sensitive (CLBL-G)			Per-Country Sensitive (CLBL-C)		
Country	**Category**	**Pct.**	**Country**	**Category**	**Pct.**	**Country**	**Category**	**Pct.**	**Country**	**Category**	**Pct.**
Iran	NEWS	13.1%	Iran	NEWS	14.0%	Iran	PORN	11.6%	Iran	NEWS	21.0%
	PORN	9.2%		PORN	12.7%		NEWS	9.4%		BLOG	17.6%
	BLOG	7.5%		ENT	10.3%		PROX	6.8%		POL	7.2%
South Korea	PORN	15.4%	South Korea	SHOP	14.2%	Saudi Arabia	PORN	31.0%	India	ENT	19.0%
	NEWS	8.4%		PORN	13.7%		GAMB	13.5%		STRM	14.3%
	ORG	7.4%		NEWS	10.8%		PROX	12.2%		NEWS	10.8%
Saudi Arabia	PORN	29.5%	Saudi Arabia	PORN	70.0%	South Korea	PORN	15.6%	Saudi Arabia	NEWS	54.0%
	NEWS	11.3%		ILL	6.6%		ORG	10.4%		POL	7.7%
	GAMB	10.1%		GAMB	6.6%		NEWS	5.7%		RELI	7.7%
India	ENT	13.3%	Turkey	PORN	66.0%	Kenya	PORN	14.5%	Russia	BLOG	16.5%
	STRM	10.8%		ILL	4.0%		GAMB	10.8%		NEWS	14.4%
	NEWS	10.4%		FILE	4.0%		PROX	9.0%		GAMB	12.4%
Kenya	PORN	15.5%	India	ILL	35.5%	Turkey	PORN	47.0%	Turkey	NEWS	29.4%
	GAMB	10.1%		IT	8.8%		GAMB	22.6%		PORN	13.7%
	PROX	8.3%		STRM	6.6%		ILL	3.2%		GAMB	9.8%

TABLE III

CENSORSHIP VARIATION BY TECHNIQUE. For each of the three techniques we can detect, the five countries observed to censor the most URLs using that technique, and the top three FortiGuard categories for those URLs (abbreviations defined in Table IV). Percentages are of all unique URLs tested.

Technique	**Country**	**Categories**	**Pct.**
Block page	Iran	NEWS, PORN, BLOG	24.95%
	Saudi Arabia	PORN, NEWS, GAMB	11.1%
	India	ENT, STRM, NEWS	6.4%
	Kenya	PORN, GAMB, PROX	4.8%
	Turkey	PORN, GAMB, NEWS	4.6%
DNS manipulation	Iran	BLOG, PORN, PROX	5.5%
	Uganda	PORN, ADUL, LING	1.7%
	Turkey	ILL, GAMB, STRM	0.3%
	Bulgaria	ILL, ARM, DOM	0.2%
	Netherlands	ILL, IM, DOM	0.2%
TCP packet injection	South Korea	PORN, ORG, NEWS	9.3%
	India	NEWS, ILL, IT	2.3%
	Netherlands	NEWS, SEAR, GAME	0.9%
	Japan	NEWS, GAME, SEAR	0.9%
	Australia	SEAR, NEWS, ILL	0.8%

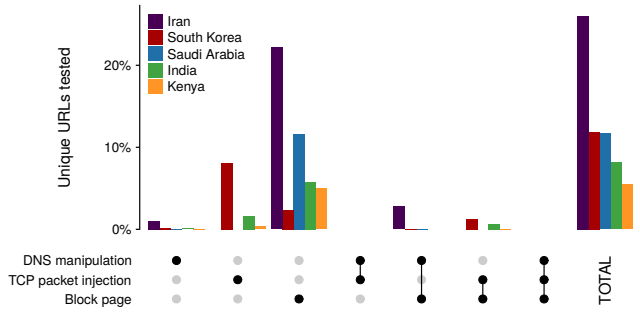

FIG. 6. COMBINATIONS OF CENSORSHIP TECHNIQUES. For the five countries performing the most censorship overall, which combinations of the three phenomena ICLab can detect are observed. Except for "TOTAL," each group of bars is mutually exclusive—URLs counted under "DNS manipulation and packet injection" are not also counted under either "DNS manipulation" or "packet injection."

TABLE IV

FORTIGUARD CATEGORIES AND ABBREVIATIONS. The 25 most common categories for the URLs on our test lists that were censored at least once, with the abbreviated names used in Tables II and III, and the percentage of URLs in each category. CLBL includes both global and per-country test lists.

Abbrev.	**Category**	**ATL %**	**CLBL %**
ADUL	Other Adult Materials	0.91	0.77
ARM	Armed Forces	0.76	0.31
BLOG	Personal websites and blogs	2.00	8.97
DOM	Domain Parking	0.21	0.28
ENT	Entertainment	2.66	2.25
FILE	File Sharing and Storage	1.89	0.55
GAME	Games	2.62	0.83
GAMB	Gambling	1.73	1.18
HEAL	Health and Wellness	2.02	1.04
ILL	Illegal or Unethical	1.85	0.40
IM	Instant Messaging	0.49	0.14
IT	Information Technology	9.31	4.17
ITRA	Internet radio and TV	0.39	0.59
LING	Lingerie and Swimsuit	0.76	0.14
NEWS	News and Media	10.03	18.87
ORG	General Organizations	6.82	4.77
POL	Political Organizations	1.56	5.28
PORN	Pornography	3.87	2.45
PROX	Proxy Avoidance	1.71	0.57
RELI	Global Religion	3.19	2.58
SEAR	Search Engines and Portals	3.93	2.36
SHOP	Shopping	4.86	1.40
SOC	Social Networking	1.19	1.34
SOLI	Society and Lifestyles	0.76	0.97
STRM	Streaming Media and Download	1.83	1.42

C. Longitudinal Analysis

Collecting data for nearly two years gives us the ability to observe changes in censorship over time. Figure 7 shows censorship trends for the six countries ICLab can monitor that block the most URLs from the global test lists (ATL and CLBL-G), plus a global trend line computed from aggregate measurements from all the other monitored countries. We do not have complete coverage for Iran and Saudi Arabia, due to the outages mentioned in §III-D. The large dip in several of the trend lines in February 2017 is an artifact due to month-

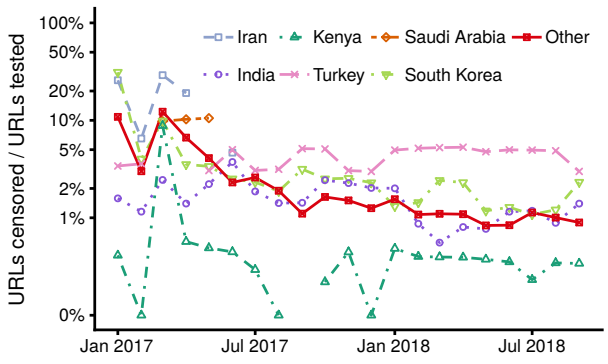

FIG. 7. LOGARITHMIC PLOT OF LONGITUDINAL TRENDS. Changes over time in the level of censorship, within the six countries where we observe the most censorship of URLs from ATL and CLBL-G, plus the aggregate of all other monitored countries.

to-month churn within the Alexa rankings (see Scheitle et al. [71]).

The global trend line shows a steady decreasing trend, which we attribute to the rising use of secure channel protocols (*e.g.,* TLS) and circumvention tools. This trend is also visible for South Korea but not for the other top five countries.

Iran blocks 20–30% of the URLs from ATL, more than any other country. This is due to extensive blocking of URLs in the NEWS and BLOG categories. Saudi Arabia consistently blocks roughly 10% of ATL and CLBL-G URLs, mostly from the PORN and GAMB categories with some NEWS as well. South Korea applies a similar level of blocking for the PORN and GAMB categories; it is a much more democratic nation than Saudia Arabia, but it nonetheless has applied draconian restrictions to "indecent Internet sites" (including both pornography and gambling sites) since before 2008 [63].

Censorship in Kenya is stable at a rate of roughly 0.4% except for March 2017, where the rate spikes to 10%. This is an artifact; for that one month, our VOD in Kenya was connected to a network that applied much more aggressive "filtering" to porn, gambling, and proxy sites than is typical for Kenya, using a commercial product.

At the beginning of 2018, we observe a drop in the level of filtering in India, from 2% to 0.8%, followed by a slow rise back to about 1.5% after about four months. This coincides with political events: India's telecommunications regulator announced support for "net neutrality" at the end of 2017 [70], [89], and most ISPs suspended their filtering in response. However, when a detailed regulation on net neutrality was published in mid-2018 [75], it became clear that the government had not intended to relax its policy regarding content deemed to be illegal, and the filtering was partially reinstated.

Similarly, we see a rise in the level of filtering in Turkey in June 2017, from an earlier level of 3% to 5%. Although it is not visible on this chart, the topics censored also change at this time. Prior to the rise, most of the blocked sites in Turkey carried pornography and other sexual content; after the rise, many more news sites were blocked. This, too, coincides with political events. Following a controversial referendum which increased the power of the Turkish Presidency, the government has attempted to suppress both internal political opposition and news published from other countries. International news organizations took notice of the increased level of Turkish online censorship in May of 2017 [40], [74], while ICLab detected it around the end of April.

D. Heuristic False Positives

We manually reviewed the results of all of our heuristic detectors for errors. Manual review cannot detect false negatives, because we have no way of knowing that we *should* have detected a site as censored, but false positives are usually obvious. Here we discuss the most significant cases we found, and how we adjusted the heuristics to compensate.

DNS Manipulation. We manually verified the detection results identified by each heuristic. The only heuristic producing false positives was the rule for when a vantage point and control nodes receive addresses in different ASes. As we mentioned in Section IV-A, this heuristic gives a false positive rate on the order of 10^{-4} with the value of θ we selected.

Packet injection. As with DNS manipulation, we manually reviewed the results of each heuristic for false positives. We found many false positives for RST or ICMP unreachable instead of SYN-ACK, leading us to reclassify these as only "probable" censorship and not include them in the synthetic analysis above. We also found cases in all of the categories where a packet anomaly was only observed once, for a URL that seemed unlikely to be censored from that vantage (*e.g.,* connection disrupted to an airline website from a VPN vantage in the USA). We therefore discount all cases where a packet anomaly has only been observed once for that URL in that country.

Block pages. Our set of regular expressions did initially produce some false positives, for instance on news reportage of censorship, quoting the text of a block page. We manually reviewed all of the matches and adjusted the regular expressions until no false positives remained. It was always possible to do this without losing any true positives.

VI. OTHER CASES OF NETWORK INTERFERENCE

In this section, we describe three cases of network interference discovered with ICLab, that are different from the form of censorship we set out to detect: Geoblocking by content providers (§VI-A), injection of a script to fingerprint clients (§VI-B), and injected malware (§VI-C).

A. Geoblocking and HTTP 451

HTTP status code 451, "Unavailable for Legal Reasons," was defined in 2016 for web servers to use when they cannot provide content due to a legal obstacle (*e.g.,* the Google restricts access to clients from Iran to enforce US sanctions [57]) or requests from foreign governments [14].

We observe 23 unique websites that return status 451, from vantages in 21 countries. Six of these cases appear to be wordpress.com complying with requests from Turkey and Russia (for blogs related to political and religious advocacy). Along

with the HTTP 451 status, they also serve a block page, explaining that wordpress.com is complying with local laws and court orders. Two more websites (both pornographic) were observed to return status 451 from Russia, with HTTP server headers indicating the error originates from the Cloudflare CDN, but without any explanation. Since the adoption of the GDPR [41] we have observed a few sites returning status 451 when visited from European countries.

Since status 451 is relatively new, the older, more generic status 403 ("Forbidden") is also used to indicate geoblocking for legal reasons. Applying the tag frequency clustering technique described in Section IV-C to the accompanying HTML, we were able to discover six more URLs, in four countries, where status 403 is used with a block page stating that access is prohibited from the client's location. Three of these were gambling sites, with the text of the block page stating that the sites are complying with local regulations.

We also observe a related phenomenon at the DNS level. From a single VPN server located in the USA, we observed netflix.com resolving to an IANA-reserved IP address, `198.18.0.3`. This could be Netflix refusing to provide their service to users behind a VPN.

B. User Tracking Injection

Our detector for block pages (§IV-C) flagged a cluster of TCP payloads observed only in South Korea. Upon manual inspection, these pages contained a script that would fingerprint the client and then load the originally intended page. We observed injections of this script over a five-month period from Oct. 2016 through Feb. 2017, from vantage points within three major Korean ISPs, into 5–30% of all our test page loads, with no correlation with the content of the affected page. By contrast, censorship in South Korea affects less than 1% of our tests and is focused on pornography, illegal file sharing, and North Korean propaganda.

These scripts could be injected by the VPN service, the ISPs, or one or more of their transit providers. The phenomenon resembles techniques used by ad networks for recording profiles of individual web users [2]. This demonstrates the importance of manual checking for false positives in censorship detection. All of the detection heuristics described in Sections IV-B and IV-C triggered on these scripts, but they are not censorship.

C. Cryptocurrency Mining Injection

Our block page detector also flagged a set of suspicious responses observed only in Brazil. The originally intended page would load, but it would contain malware causing the web browser to mine cryptocurrency. (As of mid-2018, this is a popular way to earn money with malware [53].) We were able to identify the malware as originating with a botnet infecting MikroTik routers (exploiting CVE-2018-14847), initially seen only in Brazil [49] but now reported to affect more than 200,000 routers worldwide [31]. Infected routers inject the mining malware into HTTP responses passing through them.

The malware appears in ICLab's records as early as July 21st, 2018—ten days before the earliest public report on the MikroTik botnet that we know of. If ICLab's continuous monitoring were coupled with continuous analysis and alerting (which is planned) it could also have detected this botnet prior to the public report. This highlights the importance of continuously monitoring network interference in general.

VII. Comparison with other Platforms

Other censorship measurement platforms active, at the time of writing, include Encore [17], Satellite-Iris [66], [72], Quack [79], and OONI [37]. Table V shows the high-level features provided by each of these platforms, and a comparison of their country, AS, and URL coverage for the two-month period of August and September 2018. (August 1, 2018 is the earliest date for which data from Quack and Satellite-Iris has been published.) All platforms suffer some variation from day to day in coverage, so we report both a weekly average and the maximum number of covered countries, ASes, and URLs. While many of the platforms described in Table V have chosen to emphasize breadth of country and AS coverage at the expense of detail. ICLab takes the opposite approach, collecting detailed information from a smaller number of vantage points.

A. Quack

Quack relies on public echo servers to measure censorship. It requires at least 15 echo servers within the same country for robust measurements [79]. Currently, these are available to Quack from 75 countries. 95 more countries have at least one echo server, which can still provide some measurements.

Quack aims to detect censorship of websites, but it does not send or receive well-formed HTTP messages. Instead it sends packets that mimic HTTP requests, which the echo server will reflect back to the client. It expects the censor to react to this reflection in the same way that it would to a real HTTP message. The designers of Quack acknowledge the possibility of false negatives when the censor only looks for HTTP traffic on the usual ports (80 and 443). More seriously, manual inspection of the Quack data set reveals that in 32.6% of the tests marked as *blocked*, the client did not successfully transmit a mimic request in the first place. We have reported this apparent bug to the Quack team.

B. Satellite-Iris

Satellite-Iris [19] combines Satellite [72] and Iris [66]. It focuses on DNS manipulation, measuring from open DNS resolvers. It compares the responses received from these vantage points with responses observed from a control node. It also retrieves corresponding TLS certificates from the Censys [30] data set and checks whether they are valid. It applies several heuristics to each response, *all* of which must be satisfied for the response to be judged as censorship. We now highlight two cases where their heuristics lead to false negatives.

If Satellite-Iris can retrieve a TLS certificate from *any* of the IP addresses in the open resolver's response, and that certificate is valid for *any* domain name, it considers the response not to be censored. This means Satellite-Iris will not detect any case where the censor supplies the address of a server for a

TABLE V
HIGH-LEVEL COMPARISON OF ICLAB WITH FIVE OTHER CENSORSHIP MONITORING PLATFORMS.

Platform	Packet Capture	Vantage Point Types VPNs	ORs[a]	VODs	Detection Capabilities DNS	TCP	Blockpage	Coverage (avg/max) Countries	ASes	URLs
Encore [17]			✓					Unknown[b]	Unknown	23
Satellite-Iris [66], [72]			✓		✓			174 / 179	3,261 / 3,617	2,094 / 2,423
Quack [79]			✓					75 / 76	3,528 / 4,135	2,157 / 2,484
OONI [37]	○[c]			✓	✓		✓	113 / 156	670 / 2,015	13,582 / 20,258
ICLab	✓	✓		✓	✓	✓	✓	42 / 50	48 / 62	16.964 / 23,992

[a]Open Relays: Internet hosts that will relay a censorship probe from researchers' computers without any prior arrangement.
[b]Due to privacy concerns, Encore does not record this information.
[c]The OONI client can optionally collect packet traces. However, OONI's servers do not record traces, due to privacy concerns.

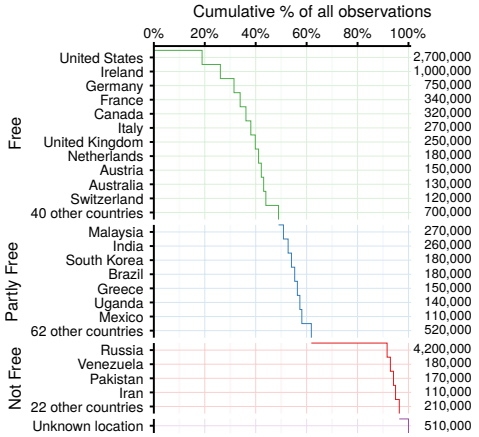

FIG. 8. Distribution of OONI observations by country in August and September of 2018, grouped by Freedom House classification.

different domain in forged DNS responses. In Satellite-Iris's published data set, 0.01% of the measurements are affected by this. Specifically, measurements from China and Turkey show domains belonging to the BBC, Google, Tor and others resolving to an IP address belonging to Facebook's server pool.

Despite the design bias toward false negatives, we also find that 83.8% of the DNS responses considered censored by Satellite-Iris may be false positives. Satellite-Iris depends on the Censys data set to distinguish DNS poisoning from normal IP variation (*e.g.,* due to geotargeting and load balancing by CDNs). When Censys information is unavailable, it falls back on a comparison to a single control resolver, which is inadequate to rule out normal variation, as discussed in §IV-A.

C. OONI

OONI [37] relies on volunteers who run a testing application manually. The application is available on all major desktop and mobile operating systems except Windows. In August and September of 2018, OONI's volunteers conducted 14,000,000 measurements from 156 countries, and reported 29,982 unique URL-country pairs as blocked.

OONI's reliance on volunteers, and on manual operation of the testing application, means that its coverage is not evenly distributed over websites or countries. Their "primary web connectivity" test suite also tests ATL and CLBL-G, but the mobile phone version of the application only tests a short subsample on each run, in order to limit the time and bandwidth consumed by the test. 62% of the measurements for the August and September 2018 period tested fewer than 80 URLs. Figure 8 shows the distribution of countries covered by OONI. 86% of all observations originate from the 23 countries named in this figure, and 48% are from just two countries—Russia and the USA. For another 4% of the observations, no location could be identified. Comparing Figure 8 to Figure 2 shows that OONI does not achieve better coverage of "partly free" and "not free" countries than ICLab does. Volunteers whose Internet access is unreliable or misconfigured may submit inaccurate results. Indeed, more than 100,000 of the observations are tagged inconclusive due to local network errors (*e.g.,* disconnection during the test).

Because OONI's testing application runs without any special privileges, it normally cannot record packet traces. OONI's detection heuristics are rudimentary, leading to a high level of false positives. OONI's DNS consistency test will flag *any disagreement* between the client's local recursive resolver and a public DNS utility as censorship [62]. OONI's block page detector relies on the "30% shorter than uncensored page" heuristic proposed by Jones et al. [48], but innocuous server errors are also short compared to normal page.

Yadav et al. [89] reported very high levels of inaccuracy in OONI's results for India, with an 80% false positive rate and a 11.6% false negative rate. We confirm a high false positive rate for OONI's block page detector: of the 12,506 unique anomalous HTTP responses reported by OONI's volunteers in August and September 2018, our block-page detector only classifies 3,201 of them as censorship, for a 74.4% false positive rate. The most common cause of false positives is a response with an empty HTML body, which can occur for a wide variety of innocuous reasons as well as censorship [4], [57], [77].

VIII. RELATED WORK

China's censorship practices have been studied in detail [23], [35], [64], [86], [88] due to the worldwide reputation of the "Great Firewall" and the relative ease of gaining access to vantage points within the country. Other countries receiving case studies include Iran [6], [10], India [89], Pakistan [51], [60], Syria [20], and Egypt and Libya [25]. Whenever researchers have had access to more than one vantage point within a country,

they have found that the policy is not consistently enforced. There is always region-to-region and ISP-to-ISP variation.

Broader studies divide into two lines of research. One group of studies investigate worldwide variation in censorship: for instance, whether censorship mainly interferes with DNS lookups [66] or subsequent TCP connections, and whether the end-user is informed of censorship [44], [80]. In some cases, it has been possible to identify the specific software in use [26], [48]. Another line of work aims to understand what is censored and why [1], [16], [84], how that changes over time [7], [43], how people react to censorship [52], [87], and how the censor might react to being monitored [16].

We described the difficulties with relying on volunteers in §I and §III. Several groups of researchers have sought alternatives. CensMon [73] used Planet Lab nodes, Anderson et al. [7] used RIPE Atlas nodes, Pearce et al. [66] use open DNS resolvers and VanderSloot et al. [79] use open echo servers. Darer et al. [28] took advantage of the fact that the Chinese Great Firewall will inject forged replies to hosts located outside the country. Burnett and Feamster [17], Ensafi et al. [33], and Pearce et al. [65] all propose variations on the theme of using existing hosts as reflectors for censorship probes, without the knowledge of their operators, at different levels of the protocol stack.

Only a few studies have lasted more than a month. Five prominent exceptions are Encore [17], IRIS [66], OONI [37], Quack[79], and Satellite [72], all of which share goals similar to ICLab. Section VII provides a detailed comparison between ICLab and these projects. Herdict [45] has also been active for years, but simply aggregates user reports of inaccessible websites. It does not test or report why the sites are inaccessible.

IX. Limitations

In this section, we discuss ICLab's limitations and how we have addressed them.

Discrimination against VPN users. Some websites may block access from VPN users [57], [72]. We sometimes observe this discrimination against our VPN clients (see §VI-A for an example), and are careful not to confuse it for censorship.

Malicious VPN Providers. Some VPN providers engage in surveillance and traffic manipulation, for instance to monetize their service by injecting advertisements into users' traffic [50]. We avoid using VPN providers that are known to do this. Our block page detectors are designed not to confuse dynamic content (*e.g.,* advertisements, localization) with censorship, as described in §IV-C. In §VI-B and VI-C we describe surveillance and malware injections that required manual inspection to distinguish from censorship.

VPN providers are also known to falsely advertise the location of their VPN servers [85]. We verify all server locations using the technique described in Appendix B.

Bias in Test Lists. ATL suffers from sampling bias and churn [71]. CLBL-G and CLBL-C may suffer from selection bias, since they are manually curated by activists. We have plans to revise the test lists and add more URLs as needed.

CLBL-G and CLBL-C are updated slowly. It is not unusual for more than half of the sites on a country-specific list to no longer exist [84]. This is not as much of a limitation as it might seem, because censors also update their lists slowly. Several previous studies found that long-gone websites may still be blocked [1], [60], [89].

Coverage of "Not Free" Countries. As discussed in Section §III-B, the risks involved with setting up many vantage points in certain sensitive ("not free") countries prevent us from claiming we can obtain complete coverage at all times. However, the set of countries ICLab covers gives us a good, if imperfect, longitudinal overview of worldwide censorship.

Evading Censorship Detection. Censors are known to try to conceal some of their actions ("covert" censorship). ICLab can detect some covert censorship, as discussed in §IV, but not all of it. The "uncertain" and "probably censored" cases of TCP packet injection (Figure 4 in §IV-B) are priorities for further investigation. Censors could further conceal their actions by disabling filtering for IP addresses that appear to be testing for censorship. Comparing results for vantage points in the same country gives us no reason to believe any country does this today.

X. Conclusions

We presented ICLab, a global censorship measurement platform that is able to measure a wide range of network interference and Internet censorship techniques.

By using VPN-based vantage points, ICLab provides flexibility and control over measurements, while reducing risks in measuring Internet censorship at a global scale. Between January 2017 and September 2018, ICLab has conducted 53,906,532 measurements over 45,565 URLs in 62 countries.

ICLab is able to detect a variety of censorship mechanisms as well as other forms of network interference. Other longitudinal measurement platforms may have more vantage points and accumulated data than ICLab, but also more errors, and/or they only focus on a specific type of censorship. Our platform can more reliably distinguish normal network errors from covert censorship, and our clustering techniques discovered 48 previously unknown block pages.

As we continue to operate ICLab and interact with relevant political science and civil society organizations, ICLab will not only make new technical observations, but also place qualitative work in this area on a firm empirical footing.

Acknowledgements

Nick Feamster, Sathyanarayanan Gunasekaran, Ben Jones, Mose Karanja, Anke Li, the Berkman Klein Center, the Small Media Foundation, and the Citizen Lab, especially Masashi Crete-Nishihata, Jakub Dalek, and Adam Senft, have all provided invaluable assistance with the implementation, testing, and deployment of ICLab.

We would like to thank our shepherd, Leyla Bilge, and all of the anonymous reviewers for their feedback on this paper. We also thank Behtash Banihashemi, Arun Dunna, Pamela Griffith, Steve Matsumoto, Rishab Nithyanad, Pinar Ozisik, Vyas Sekar,

Mahmood Sharif, Rachee Singh, Kyle Soska, Janos Szurdi, Xiao Hui Tai, and Nicholas Weaver for helpful comments and suggestions.

This research was financially supported by the Ministry of Science and ICT, Korea, under award IITP-2019-H8601-15-1011; by the National Science Foundation, United States, under awards CNS-1350720, CNS-1651784, CNS-1700657, CNS-1740895, and CNS-1814817; by a Google Faculty Research Award; and by the Open Technology Fund under an Information Controls Fellowship. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the sponsors, nor of the governments of the Republic of Korea or the United States of America.

References

[1] N. Aase, J. R. Crandall, Á. Díaz, J. Knockel, J. O. Molinero, J. Saia, D. Wallach, and T. Zhu, "Whiskey, Weed, and Wukan on the World Wide Web: On Measuring Censors' Resources and Motivations," in *Free and Open Communications on the Internet*. USENIX, 2012.

[2] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *Computer and Communications Security*. ACM, 2014, pp. 674–689.

[3] G. Aceto, A. Montieri, and A. Pescapè, "Internet Censorship in Italy: A First Look at 3G/4G Networks," in *Cryptology and Network Security*. Springer, 2016, pp. 737–742.

[4] S. Afroz, M. C. Tschantz, S. Sajid, S. A. Qazi, M. Javed, and V. Paxson, "Exploring Server-side Blocking of Regions," 2018.

[5] Alexa Internet, Inc., "How are Alexa's traffic rankings determined?" accessed 2014. [Online]. Available: http://www.alexa.com/faqs/?p=134

[6] C. Anderson, "Dimming the Internet: Detecting Throttling as a Mechanism of Censorship in Iran," 2013.

[7] C. Anderson, P. Winter, and Roya, "Global Network Interference Detection over the RIPE Atlas Network," in *Free and Open Communications on the Internet*. USENIX, 2014.

[8] Anonymous, "Towards a Comprehensive Picture of the Great Firewall's DNS Censorship," in *Free and Open Communications on the Internet*. USENIX, 2014.

[9] ——, "The Collateral Damage of Internet Censorship by DNS Injection," *SIGCOMM Computer Communications Review*, vol. 42, no. 3, pp. 21–27, 2012. [Online]. Available: http://www.sigcomm.org/node/3275

[10] S. Aryan, H. Aryan, and J. A. Halderman, "Internet Censorship in Iran: A First Look," in *Free and Open Communications on the Internet*. USENIX, 2013.

[11] Berkman Klein Center, "Website Inaccessibility Test Lists," 2018. [Online]. Available: https://github.com/berkmancenter/url-lists

[12] R. Bonica, M. Cotton, B. Haberman, and L. Vegoda, "Updates to the Special-Purpose IP Address Registries," RFC 8190, 2017. [Online]. Available: https://tools.ietf.org/html/rfc8190

[13] S. Bortzmeyer and S. Huque, "NXDOMAIN: There Really Is Nothing Underneath," RFC 8020, 2016. [Online]. Available: https://tools.ietf.org/html/rfc8020

[14] T. Bray, "An HTTP Status Code to Report Legal Obstacles," RFC 7725, 2016. [Online]. Available: https://tools.ietf.org/html/rfc7725

[15] M. A. Brown, "Pakistan hijacks YouTube," 2008, accessed 2018. [Online]. Available: https://dyn.com/blog/pakistan-hijacks-youtube-1/

[16] S. Burnett and N. Feamster, "Making Sense of Internet Censorship: A New Frontier for Internet Measurement," *SIGCOMM Computer Communications Review*, vol. 43, no. 3, pp. 84–89, 2013.

[17] ——, "Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests," in *SIGCOMM*. ACM, 2015, pp. 653–667.

[18] CAIDA, "AS Classification," accessed 2017. [Online]. Available: http://www.caida.org/data/as-classification/

[19] Censored Planet, "Satellite," accessed 2018. [Online]. Available: http://www.censoredplanet.com/projects/satellite

[20] A. Chaabane, T. Chen, M. Cunche, E. De Cristofaro, A. Friedman, and M. A. Kaafar, "Censorship in the wild: Analyzing Internet filtering in Syria," in *Internet Measurement Conference*. ACM, 2014, pp. 285–298.

[21] Citizen Lab, "Collection of censorship blockpages," 2015. [Online]. Available: https://github.com/citizenlab/blockpages

[22] ——, "URL testing lists intended for discovering website censorship," 2014. [Online]. Available: https://github.com/citizenlab/test-lists

[23] R. Clayton, S. J. Murdoch, and R. N. M. Watson, "Ignoring the Great Firewall of China," in *Privacy Enhancing Technologies*. Springer, 2006, pp. 20–35.

[24] J. R. Crandall, D. Zinn, M. Byrd, E. Barr, and R. East, "ConceptDoppler: A Weather Tracker for Internet Censorship," in *Computer and Communications Security*. ACM, 2007, pp. 352–365. [Online]. Available: http://www.cs.unm.edu/~crandall/concept_doppler_ccs07.pdf

[25] A. Dainotti, C. Squarcella, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, and A. Pescape, "Analysis of Country-Wide Internet Outages Caused by Censorship," *Transactions on Networking*, vol. 22, pp. 1964–1977, 2013. [Online]. Available: http://www.caida.org/publications/papers/2014/outages_censorship/

[26] J. Dalek, B. Haselton, H. Noman, A. Senft, M. Crete-Nishihata, P. Gill, and R. J. Deibert, "A Method for Identifying and Confirming the Use of URL Filtering Products for Censorship," in *Internet Measurement Conference*. ACM, 2013, pp. 23–30. [Online]. Available: http://www3.cs.stonybrook.edu/~phillipa/papers/imc112s-dalek.pdf

[27] J. Dalek, R. Deibert, S. McKune, P. Gill, A. Senft, and N. Noor, "Information controls during military operations: The case of Yemen during the 2015 political and armed conflict," Citizen Lab, 2015. [Online]. Available: https://citizenlab.ca/2015/10/information-controls-military-operations-yemen/

[28] A. Darer, O. Farnan, and J. Wright, "FilteredWeb: A Framework for the Automated Search-Based Discovery of Blocked URLs," in *Traffic Measurement and Analysis*. IEEE, 2017, pp. 1–9.

[29] R. Deibert, J. Palfrey, R. Rohozinski, and J. Zittrain, Eds., *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*. MIT Press, 2010.

[30] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman, "A search engine backed by Internet-wide scanning," in *Computer and Communications Security*. ACM, 2015, pp. 542–553.

[31] M. Emem, "Monero Cryptomining Attack Affects Over 200,000 ISP-Grade Routers Globally," *CCN Markets*, accessed 2018. [Online]. Available: https://www.ccn.com/monero-cryptomining-attack-affects-over-200000-isp-grade-routers-globally/

[32] R. Ensafi, J. Knockel, G. Alexander, and J. R. Crandall, "Detecting Intentional Packet Drops on the Internet via TCP/IP Side Channels," in *Passive and Active Measurement*. Springer, 2014, pp. 109–118.

[33] R. Ensafi, P. Winter, A. Mueen, and J. R. Crandall, "Large-scale Spatiotemporal Characterization of Inconsistencies in the World's Largest Firewall," 2014.

[34] R. Ensafi, D. Fifield, P. Winter, N. Feamster, N. Weaver, and V. Paxson, "Examining How the Great Firewall Discovers Hidden Circumvention Servers," in *Internet Measurement Conference*. ACM, 2015, pp. 445–458.

[35] R. Ensafi, P. Winter, A. Mueen, and J. R. Crandall, "Analyzing the Great Firewall of China Over Space and Time," in *Privacy Enhancing Technologies*. De Gruyter, 2015, pp. 61–76.

[36] O. Farnan, A. Darer, and J. Wright, "Poisoning the Well: Exploring the Great Firewall's Poisoned DNS Responses," in *Privacy in the Electronic Society*. ACM, 2016, pp. 95–98.

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

[38] "FortiGuard Labs Web Filter," accessed 2018. [Online]. Available: https://fortiguard.com/webfilter

[39] Freedom House, "Freedom on the Net 2017," accessed 2018. [Online]. Available: https://freedomhouse.org/report/freedom-net/freedom-net-2017

[40] ——, "Country Profiles 2017: Turkey," accessed 2018. [Online]. Available: https://freedomhouse.org/report/freedom-net/2017/turkey

[41] "General Data Protection Regulation (2016/679)," *Official Journal of the European Union*, vol. L 119, pp. 1–88, 2016. [Online]. Available: https://gdpr-info.eu/

[42] Geosurf, "Geosurf: Residential and data center proxy network," accessed 2018. [Online]. Available: https://www.geosurf.com

[43] P. Gill, M. Crete-Nishihata, J. Dalek, S. Goldberg, A. Senft, and G. Wiseman, "Characterizing Web Censorship Worldwide: Another Look at the OpenNet Initiative Data," *Transactions on the Web*, vol. 9, no. 1, 2015.

[44] S. Hellmeier, "The Dictator's Digital Toolkit: Explaining Variation in Internet Filtering in Authoritarian Regimes," *Politics & Policy*, vol. 44, no. 6, pp. 1158–1191, 2016.

[45] "Herdict: help spot web blockages," accessed 2018. [Online]. Available: https://www.herdict.org/

[46] N. P. Hoang, P. Kintis, M. Antonakakis, and M. Polychronakis, "An Empirical Study of the I2P Anonymity Network and Its Censorship Resistance," in *Internet Measurement Conference*. ACM, 2018, pp. 379–392.

[47] "Hola!VPN: Access any website," accessed 2018. [Online]. Available: https://hola.org/

[48] B. Jones, T.-W. Lee, N. Feamster, and P. Gill, "Automated Detection and Fingerprinting of Censorship Block Pages," in *Internet Measurement Conference*. ACM, 2014, pp. 299–304.

[49] S. Kenin, "Mass MikroTik Router Infection – First we cryptojack Brazil, then we take the World?" *SpiderLabs Blog*, accessed 2018. [Online]. Available: https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/mass-mikrotik-router-infection-first-we-cryptojack-brazil-then-we-take-the-world/

[50] M. T. Khan, J. DeBlasio, G. M. Voelker, A. C. Snoeren, C. Kanich, and N. Vallina-Rodriguez, "An Empirical Analysis of the Commercial VPN Ecosystem," in *Internet Measurement Conference*. ACM, 2018, pp. 443–456.

[51] S. Khattak, M. Javed, S. A. Khayam, Z. A. Uzmi, and V. Paxson, "A Look at the Consequences of Internet Censorship Through an ISP Lens," in *Internet Measurement Conference*. ACM, 2014, pp. 271–284.

[52] J. Knockel, J. R. Crandall, and J. Saia, "Three Researchers, Five Conjectures: An Empirical Analysis of TOM-Skype Censorship and Surveillance." in *Free and Open Communications on the Internet*. USENIX, 2011.

[53] R. K. Konoth, E. Vineti, V. Moonsamy, M. Lindorfer, C. Kruegel, H. Bos, and G. Vigna, "MineSweeper: An In-depth Look into Drive-by Cryptocurrency Mining and Its Defense," in *Computer and Communications Security*. ACM, 2018, pp. 1714–1730.

[54] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," in *Network and Distributed System Security Symposium*, 2019.

[55] "Luminati: largest business proxy service," accessed 2018. [Online]. Available: http://luminati.io

[56] B. Marczak, N. Weaver, J. Dalek, R. Ensafi, D. Fifield, S. McKune, A. Rey, J. Scott-Railton, R. Deibert, and V. Paxson, "China's Great Cannon," Citizen Lab, 2015. [Online]. Available: https://citizenlab.org/2015/04/chinas-great-cannon/

[57] A. McDonald, M. Bernhard, L. Valenta, B. VanderSloot, W. Scott, N. Sullivan, J. A. Halderman, and R. Ensafi, "403 Forbidden: A Global View of CDN Geoblocking," in *Internet Measurement Conference*. ACM, 2018, pp. 218–230.

[58] X. Mi, Y. Liu, X. Feng, X. Liao, B. Liu, X. Wang, F. Qian, Z. Li, S. Alrwais, and L. Sun, "Resident Evil: Understanding Residential IP Proxy as a Dark Service," in *Symposium on Security and Privacy*. IEEE, 2019, pp. 170–186. [Online]. Available: https://mixianghang.github.io/pubs/rpaas.pdf

[59] A. Molavi Kakhki, A. Razaghpanah, A. Li, H. Koo, R. Golani, D. Choffnes, P. Gill, and A. Mislove, "Identifying Traffic Differentiation in Mobile Networks," in *Internet Measurement Conference*. ACM, 2015, pp. 239–251.

[60] Z. Nabi, "The Anatomy of Web Censorship in Pakistan," in *Free and Open Communications on the Internet*. USENIX, 2013.

[61] A. Nisar, A. Kashaf, I. A. Qazi, and Z. A. Uzmi, "Incentivizing censorship measurements via circumvention," in *SIGCOMM*. ACM, 2018, pp. 533–546.

[62] OONI, "DNS consistency," blog post, accessed 2018. [Online]. Available: https://github.com/ooni/spec/blob/master/nettests/ts-002-dns-consistency.md

[63] OpenNet Initiative, "Country Profiles: South Korea," 2012, accessed 2019. [Online]. Available: https://opennet.net/research/profiles/south-korea

[64] J. C. Park and J. R. Crandall, "Empirical study of a national-scale distributed intrusion detection system: Backbone-level filtering of HTML responses in China," in *Distributed Computing Systems*. IEEE, 2010, pp. 315–326. [Online]. Available: http://iar.cs.unm.edu/~crandall/icdcs2010.pdf

[65] P. Pearce, R. Ensafi, F. Li, N. Feamster, and V. Paxson, "Augur: Internet-Wide Detection of Connectivity Disruptions," in *Symposium on Security and Privacy*. IEEE, 2017, pp. 427–443.

[66] P. Pearce, B. Jones, F. Li, R. Ensafi, N. Feamster, N. Weaver, and V. Paxson, "Global Measurement of DNS Manipulation," in *USENIX Security Symposium*. USENIX, 2017.

[67] J. Postel, "Transmission Control Protocol," RFC 793, 1981. [Online]. Available: https://tools.ietf.org/html/rfc793

[68] Reporters Without Borders, "2018 world press freedom index." [Online]. Available: https://rsf.org/en/ranking/2018

[69] RIPE NCC Staff, "RIPE Atlas: A Global Internet Measurement Network," *The Internet Protocol Journal*, vol. 18, no. 3, pp. 2–26, 2015. [Online]. Available: http://ipj.dreamhosters.com/wp-content/uploads/2015/10/ipj18.3.pdf

[70] P. K. Roy, "India net neutrality rules could be world's strongest," *BBC News*, accessed 2018. [Online]. Available: https://www.bbc.com/news/world-asia-india-42162979

[71] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists," in *Internet Measurement Conference*. ACM, 2018, pp. 478–493.

[72] W. Scott, T. Anderson, T. Kohno, and A. Krishnamurthy, "Satellite: Joint Analysis of CDNs and Network-Level Interference," in *USENIX Annual Technical Conference*. USENIX, 2016.

[73] A. Sfakianakis, E. Athanasopoulos, and S. Ioannidis, "CensMon: A Web Censorship Monitor," in *Free and Open Communications on the Internet*. USENIX, 2011.

[74] E. K. Sozeri, "Inside Turkey's war on Wikipedia," *The Daily Dot*, accessed 2018. [Online]. Available: https://www.dailydot.com/layer8/turkey-bans-wikipedia-censorship/

[75] Telecom Regulatory Authority of India, "Recommendations on Net Neutrality," 2017. [Online]. Available: https://main.trai.gov.in/sites/default/files/Recommendations_NN_2017_11_28.pdf

[76] A. Toonk, "Turkey Hijacking IP addresses for popular Global DNS providers," 2014, accessed 2018. [Online]. Available: https://bgpmon.net/turkey-hijacking-ip-addresses-for-popular-global-dns-providers/

[77] M. C. Tschantz, S. Afroz, S. Sajid, S. A. Qazi, M. Javed, and V. Paxson, "A bestiary of blocking: The motivations and modes behind website unavailability," in *Free and Open Communications on the Internet*. USENIX, 2018.

[78] I. van Beijnum, "China censorship leaks outside Great Firewall via root server," *Ars Technica*, accessed 2018. [Online]. Available: https://arstechnica.com/tech-policy/2010/03/china-censorship-leaks-outside-great-firewall-via-root-server/

[79] B. VanderSloot, A. McDonald, S. Will, J. A. Halderman, and R. Ensafi, "Quack: Scalable Remote Measurement of Application-Layer Censorship," in *USENIX Security Symposium*. USENIX, 2018.

[80] J.-P. Verkamp and M. Gupta, "Inferring Mechanics of Web Censorship Around the World," in *Free and Open Communications on the Internet*, 2012.

[81] M. Wander, "Measurement survey of server-side DNSSEC adoption," in *Network Traffic Measurement and Analysis*. IEEE, 2017.

[82] N. Weaver, R. Sommer, and V. Paxson, "Detecting Forged TCP Reset Packets," in *Network and Distributed System Security*. Internet Society, 2009.

[83] N. Weaver, C. Kreibich, M. Dam, and V. Paxson, "Here Be Web Proxies," in *Passive and Active Measurement*. Springer, 2014, pp. 183–192.

[84] Z. Weinberg, M. Sharif, J. Szurdi, and N. Christin, "Topics of Controversy: An Empirical Analysis of Web Censorship Lists," in *Privacy Enhancing Technologies*. De Gruyter, 2017, pp. 42–61.

[85] Z. Weinberg, S. Cho, N. Christin, V. Sekar, and P. Gill, "How to Catch when Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation," in *Internet Measurement Conference*. ACM, 2018, pp. 203–217.

[86] J. Wright, "Regional Variation in Chinese Internet Filtering," *Information, Communication & Society*, vol. 17, no. 1, pp. 121–141, 2014.

[87] J. Wright, A. Darer, and O. Farnan, "Filterprints: Identifying Localized Usage Anomalies in Censorship Circumvention Tools," 2016.

[88] X. Xu, Z. M. Mao, and J. A. Halderman, "Internet Censorship in China: Where Does the Filtering Occur?" in *Passive and Active Measurement*. Springer, 2011, pp. 133–142.

[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 *Internet Measurement Conference*. ACM, 2018, pp. 252–264.

[90] E. Zhu, F. Nargesian, K. Q. Pu, and R. Miller, "LSH Ensemble: Internet-Scale Domain Search," *VLDB Endowment*, vol. 9, no. 12, pp. 1185–1196, 2016.

[91] J. Zittrain and B. Edelman, "Internet filtering in China," *IEEE Internet Computing*, vol. 7, no. 2, pp. 70–77, 2003.

Appendix A
Access to Different ASes within countries of interest

Censorship policies are known to vary from region to region and network to network within a single country [1], [33], [51], [86], [88]. Therefore, comprehensive monitoring requires vantage points located in diverse locations within a country. Some VPN services offer servers in several physical locations within a single country, making this simple. Even when they don't advertise several physical locations, we have found that they often load-balance connections to a single hostname over IP addresses in several different ASes and sometimes different physical locations as well. When possible, we increase diversity further by subscribing to multiple VPN services. Figure 9 shows a CDF of the number of networks we can access in each country, combining all the above factors; we are able to access two or more networks in 75% of all countries, and three or more networks in 50%.

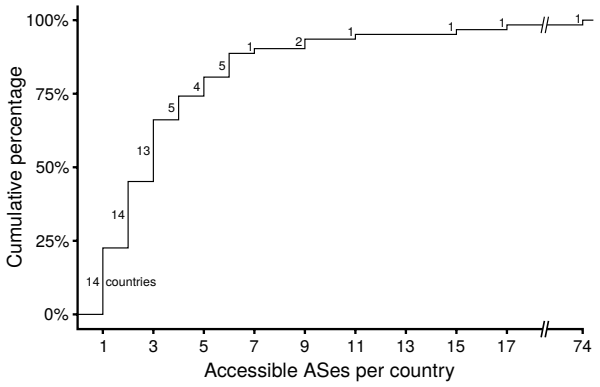

Fig. 9. CDF of number of accessible AS(es) per country

Appendix B
VPN Proxy Location Validation

Commercial VPN services cannot be relied on to locate all of their servers in the countries where they are advertised to be [85]. ICLab therefore checks the location of each VPN server before using it for measurements. We assume that packets are not able to travel faster than 153 km/ms ($0.5104\,c$) over long distances. We measure the round-trip time from each VPN server to a set of landmark hosts in known locations, drawn from the RIPE Atlas measurement constellation [69]. If any packet would have had to travel faster than 153 km/ms to reach the advertised country and return in the measured time, we assume the server is not in its advertised location, and we do not use it as a vantage point.

The VPN services we subscribe to collectively advertise endpoints in 216 countries. Our checker is only able to confirm the advertised location for 55 countries (25.5% of the total). Compared to the results reported by Weinberg et al. [85], who tackled the same problem with more sophisticated techniques, our method rejects significantly more proxies (10% more on average when we experiment across multiple providers). Possibly some of those proxies could be used after all, but we do not want to attribute censorship to the wrong country by accident, so we are being cautious.

Appendix C
Freedom House and Reporters Without Borders Scores

The international organization Freedom House, which promotes civil liberty and democracy worldwide, issues a yearly report on "freedom on the Net," in which they rate 65 countries on the degree to which online privacy and free exchange of information online are upheld in that country [39]. Each country receives both a numerical score and a three-way classification: 16 of the 65 countries are considered "free," 28 are "partly free," and 21 are "not free." Unfortunately, 33 of the countries studied by ICLab are not included in this report.

The international organization Reporters Without Borders (RWB) issues a similar yearly report on freedom of the press. This report covers 189 countries and territories, including all 65 of the countries rated by Freedom House, and all 62 of the countries studied by ICLab [68] Each country receives a numerical score and a color code (best to worst: 16 countries are coded white, 42 yellow, 59 orange, 51 red, and 21 black). Press freedom is not the same as online freedom, and the methodologies behind the two reports are quite different, but the scores from the two reports are reasonably well correlated (Kendall's $\tau = 0.707$, $p \approx 10^{-16}$). We used a simple linear regression to map RWB scores onto the same scale as FH scores, allowing us to label all of the countries studied by ICLab as "free" (72), "partly free" (85), or "not free" (32) in the same sense used by Freedom House.

Appendix D
Detailed censorship results

Table VI continues Table II (Section V), showing countries 6 through 15 in the same ranking, with the top three FortiGuard categories among their censored URLs, and the percentages of all their censored URLs within those categories.

Many of these countries censor only a few of the URLs on the lists we tested, so Table VI may reflect biases of the test lists, such as over-representation of the GAME, IT, NEWS, and SEAR categories on both ATL and CLBL-G.

The presence of countries such as Japan, the Netherlands, Sweden, and the United States in this table could indicate that our detectors still have false positives, or that individual ISPs

TABLE VI
CENSORSHIP BY TEST LIST AND CATEGORY, continued. For each of the three types of test list we use (see Section III-C), the next ten countries performing the most censorship of URLs from that country, the top three FortiGuard categories among their censored URLs (abbreviations defined in Table IV), and the percentage of all censored URLs from that category. We only observe 14 countries to censor anything from CLBL-C.

Overall			Alexa Global (ATL)			Globally Sensitive (CLBL-G)			Per-Country Sensitive (CLBL-C)		
Country	**Category**	**Pct.**	**Country**	**Category**	**Pct.**	**Country**	**Category**	**Pct.**	**Country**	**Category**	**Pct.**
Turkey	PORN	40.2%	Kenya	ILL	28.1%	India	NEWS	10.3%	South Korea	PORN	16.7%
	GAMB	16.6%		PORN	25.0%		ILL	9.2%		NEWS	16.1%
	NEWS	9.2%		GAME	6.2%		IT	8.0%		SHOP	11.8%
Russia	GAMB	23.4%	Russia	PORN	26.3%	United States	NEWS	8.0%	China	NEWS	46.1%
	PORN	10.0%		SHOP	21.0%		IT	6.9%		ORG	46.1%
	NEWS	7.6%		STRM	10.5%		SEAR	6.3%		RELI	7.7%
Uganda	PORN	42.6%	Japan	SEAR	19.0%	Uganda	PORN	42.6%	Hong Kong	ORG	100.0%
	ADUL	11.7%		NEWS	9.5%		ADUL	11.7%			
	LING	10.3%		GAME	9.5%		LING	10.3%			
Netherlands	NEWS	13.4%	Netherlands	ILL	15.3%	Russia	GAMB	39.4%	Poland	GAMB	100.0%
	ILL	8.5%		NEWS	15.3%		PORN	14.9%			
	SEAR	8.5%		SEAR	15.3%		RELI	5.3%			
Japan	NEWS	11.0%	Sweden	SEAR	27.2%	Netherlands	NEWS	13.0%	Singapore	PROX	66.6%
	GAME	9.6%		BLOG	9.1%		ILL	7.2%		GAME	33.3%
	SEAR	9.6%		STRM	9.1%		GAME	7.2%			
Australia	SEAR	15.4%	Hong Kong	STRM	20.0%	Japan	NEWS	11.5%	Ukraine	BLOG	75.0%
	ILL	10.7%		SEAR	20.0%		GAME	9.8%		NEWS	8.3%
	NEWS	9.2%		SHOP	10.0%		ILL	6.5%		IT	8.3%
Sweden	GAME	10.3%	Australia	ILL	30.0%	Australia	SEAR	14.5%	Malaysia	PORN	100.0%
	NEWS	10.3%		SEAR	20.0%		NEWS	10.9%			
	STRM	6.9%		SHOP	10.0%		ILL	7.2%			
New Zealand	GAME	11.5%	New Zealand	SEAR	20.0%	Sweden	GAME	10.6%	Colombia	ITRA	100.0%
	HEAL	9.6%		GAME	20.0%		ILL	6.4%			
	SEAR	9.6%		ILL	10.0%		STRM	6.4%			
China	NEWS	17.0%	United States	SEAR	21.6%	Hong Kong	NEWS	10.9%	Brazil	SOLI	100.0%
	ORG	12.7%		NEWS	10.8%		GAME	10.8%			
	SEAR	6.4%		ILL	8.1%		STRM	8.7%			
Bulgaria	ILL	11.6%	China	SEAR	33.3%	New Zealand	HEAL	9.5%			
	HEAL	9.3%		GAME	16.6%		GAME	9.5%			
	GAME	9.3%		HEAL	16.6%		NEWS	7.1%			

HTML structure

```
ACK+PSH
HTTP/1.1 200 OK
Connection: close
Content-Length: nnnn
Content-Type: text/html; charset="utf-8"
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN">
<html>
<head<title></title></head>
<body><h0><font color="black">
visible message
</font></h0></body></html>
```

Visible message

```
"This URL has been blocked under instructions of a
competent Government Authority or in compliance with
the orders of a Court of competent jurisdiction.

***This URL has been blocked under Instructions of the
Competent Government Authority or Incompliance to
the orders of Hon'ble Court.*** [sic]

*"Error 403: Access Denied/Forbidden"*

404. That's an error.

HTTP Error 404 - File or Directory not found

HTTP Error 404 - File or Directory not found = http://...
```

FIG. 10. EXAMPLE CLUSTER OF BLOCK PAGES. All of the messages in the right-hand column were observed with the HTTP response headers and HTML structure shown on the left.

and/or the VPN services we are using are blocking access to certain sites. The latter is likely for the ILL category, which includes many sites that facilitate copyright infringement.

Figure 10 shows an example group of block pages detected by textual similarity clustering, including two variations on the Indian legal jargon mentioned in Section IV-C, but also messages mimicking generic HTTP server errors. This demonstrates how similarity clustering can detect covert as well as overt censorship.