Augur: Internet-Wide Detection of Connectivity Disruptions

Paul Pearce[†*], Roya Ensafi[§*], Frank Li[†], Nick Feamster[§], Vern Paxson[†]
[†]University of California, Berkeley [§]Princeton University
{pearce, frankli, vern}@berkeley.edu {rensafi, feamster}@cs.princeton.edu

***Abstract*—Anecdotes, news reports, and policy briefings collectively suggest that Internet censorship practices are pervasive. The scale and diversity of Internet censorship practices makes it difficult to precisely monitor where, when, and how censorship occurs, as well as what is censored. The potential risks in performing the measurements make this problem even more challenging. As a result, many accounts of censorship begin—and end—with anecdotes or short-term studies from only a handful of vantage points.**

We seek to instead continuously monitor information about Internet reachability, to capture the onset or termination of censorship across regions and ISPs. To achieve this goal, we introduce *Augur*, a method and accompanying system that utilizes TCP/IP side channels to measure reachability between two Internet locations without directly controlling a measurement vantage point at either location. Using these side channels, coupled with techniques to ensure safety by not implicating individual users, we develop scalable, statistically robust methods to infer network-layer filtering, and implement a corresponding system capable of performing continuous monitoring of global censorship. We validate our measurements of Internet-wide disruption in nearly 180 countries over 17 days against sites known to be frequently blocked; we also identify the countries where connectivity disruption is most prevalent.

I. Introduction

Anecdotes, news reports, and policy briefings collectively suggest that Internet censorship practices are pervasive. Many countries employ a variety of techniques to prevent their citizenry from accessing a wide spectrum of information and services, spanning the range from content sensitive for political or religious reasons, to microblogging, gambling, pornography, and suicide, to the use of censorship circumvention systems themselves. Unfortunately, despite the fact that censorship affects billions of people, our understanding of its practices and techniques remains for the most part pointwise. Studies and accounts heavily focus on the state of censorship in a single country, often as seen at a single point in time. We lack *global* views that comprehensively span the worldwide Internet, and we lack *continual* views that flag the onset of new censorship and relaxation of existing censorship.

To date, efforts to obtain global visibility into censorship practices have required some sort of network presence in each country to monitor. This might mean the use of network proxies, such as ICLab's use of VPN exits [28], or the deployment of dedicated systems, such as by OONI [48]. These approaches remain difficult to deploy in practice: for example, some countries might not have globally available VPN exits within them, or may have censors that block the network access required for the measurements (such as OONI's use of Tor). Another approach is to opportunistically leverage a network presence in a given country using browser-based remote measurement of potential censorship [45]. This method can have difficulties in obtaining fully global views, though, because it is driven by end-user browsing choices. Due to its potential for implicating end users in attempting to access prohibited Internet sites, it can only be used broadly to measure reachability to sites that would pose minimal additional risk to users, which limits its utility for measuring reachability to a broad range of sites.

Fortunately, advances in TCP/IP side-channel measurement techniques offer a new paradigm for obtaining global-scale visibility into Internet connectivity. Ensafi et al. recently developed *Hybrid-Idle Scan*, a method whereby a third vantage point can determine the state of network-layer reachability between two other endpoints [22]. In other words, an off-path measurement system can infer whether two remote systems can communicate with one another, regardless of where these two remote systems are located. To perform these measurements, the off-path system must be able to spoof packets (i.e., it must reside in a network that does not perform egress filtering), and one of the two endpoints must use a single shared counter for generating the IP identifier value for packets that it generates. This technique provides the possibility of measuring network-layer reachability around the world by locating endpoints within each country that use a shared IP ID counter. By measuring the progression of this counter over time, as well as whether our attempts to perturb it from other locations on the Internet, we can determine the reachability status between pairs of Internet endpoints. This technique makes it possible to conduct measurements continuously, across a large number of vantage points.

Despite the conceptual appeal of this approach, realizing the method poses many challenges. One challenge concerns *ethics*: Using this method can make it appear as though a user in some country is attempting to communicate with a potentially censored destination, which could imperil users. To abide by the ethical guidelines set out by the Menlo [19] and Belmont [9] reports, we exercise great care to ensure that we perform our measurements from Internet infrastructure (e.g., routers, middleboxes), as opposed to user machines. A second challenge concerns *statistical robustness* in the face

*Joint first authors.

of unrelated network activity that could interfere with the measurements, as well as other systematic errors concerning the behavior of TCP/IP side channels that sometimes only become apparent at scale. To address these challenges we introduce *Augur*. To perform detection in the face of uncertainty, we model the IP ID increment over a time interval as a random variable that we can condition on two different priors: with and without responses to our attempts to perturb the counter from another remote Internet endpoint. Given these two distributions, we can then apply statistical hypothesis testing based on maximum likelihood ratios.

We validate our Augur measurements of Internet-wide disruption in nearly 180 countries over 17 days against both block lists from other organizations as well as known IP addresses for Tor bridges. We find that our results are consistent with the expected filtering behavior from these sites. We also identify the top countries that experience connectivity disruption; our results highlight many of the world's most infamous censors.

We begin in Section II with a discussion of related work. In Section III, we provide an overview of our method. We present Augur in Section IV, introducing the principles behind using IP ID side channels for third-party measurement of censorship; discussing how to identify remote systems that enable us to conduct our measurements in an ethically responsible manner; and delving into the extensive considerations required for robust inference. In Section V, we present a concrete implementation of Augur. In Section VI, we validate Augur's accuracy and provide an accompanying analysis of global censorship practices observed during our measurement run. We offer thoughts related to further developing our approach in Section VII and conclude in Section VIII.

II. Related Work

Previous work spans several related areas. We begin with a discussion of closely related work on connectivity measurements using side channels. We then discuss previous research which has performed pointwise studies of censorship in various countries, as well as tools that researchers have developed to facilitate these direct measurements. Finally, we discuss previous studies that have highlighted the variability and volatility of censorship measurements over time and across regions, which motivates our work.

Measuring connectivity disruptions with side channels. Previous work has employed side channels to infer network properties such as topology, traffic usage, or firewall rules between two remote hosts. Some of these techniques rely on the fact that the IP identifier (IP ID) field can reveal network interfaces that belong to the same Internet router, the number of packets that a device generates [13], or the blocking direction of mail server ports for anti-spam purposes [43].

The SYN backlog also provides another signal that helps with the discovery of machines behind firewalls [23], [55]. Ensafi et al. [22] observed that combining information from the TCP SYN backlog (which initiates retransmissions of SYN ACK packets) with IP ID changes can reveal packet loss between two remote hosts, including the direction along the path where packet drops occurred; the authors demonstrated the utility of their technique by measuring the reachability of Tor relays from China [24]. Our work builds on this technique by developing robust statistical detection methods to disambiguate connectivity disruptions from other effects that induce signals in these side channels.

Direct measurements from in-country vantage points. Researchers have performed many pointwise measurement studies that directly measure connectivity disruptions in countries including China [5], [16], [56], Iran [7], Pakistan [33], [38], and Syria [12]. These studies have typically relied on obtaining vantage points in target countries, often by renting virtual private servers (VPSs) and performing measurements from that vantage point. These direct measurements have served to reveal censorship mechanisms, including country-wide Internet outages [17], the injection of fake DNS replies [6], [34], the blocking of TCP/IP connections [53], HTTP-level blocking [18], [30], [42], and traffic throttling [3]. In general, studies involving direct measurements can shed more light on specific mechanisms that a censor might employ. By contrast, the techniques we develop rely on indirect side channels, which limits the types of measurements that we can perform. On the other hand, our approach permits a much larger scale than any of these previous studies, as well as the ability to conduct measurements continuously. Although these studies provide valuable insights, their scale often involves a single vantage point for a limited amount of time (typically no more than a few weeks). Our aim is to shed light on a much broader array of Internet vantage points, continuously over time.

Tools to facilitate direct measurements. OONI performs an ongoing set of censorship measurement tests from the vantage points of volunteer participants. It runs on both personal machines and embedded devices such as Raspberry Pis [26]. Although OONI performs a more comprehensive set of tests than we can with our indirect measurement, the tool has deployment at a limited number of vantage points. CensMon [46] only runs on PlanetLab nodes, limiting its visibility to academic networks that can experience different filtering practices than residential or commercial networks within a country. UBICA [1] aimed to increase vantage points by running censorship measurement software on home gateway devices and user desktops. These systems require points of contact within a country to establish and maintain the infrastructure. The OpenNet Initiative [41] leverages social connections to people around the world to perform one-off censorship measurements from home networks. As these measurements are collected opportunistically with no systematic baseline, it can be difficult to draw consistent, repeatable conclusions.

Studies that highlight the temporal and spatial variability of connectivity disruptions. If patterns of censorship and connectivity disruptions hold relatively static, then existing one-off measurement studies would suffice to over time build

up a global picture of conditions. Previous work, however, has demonstrated that censorship practices vary across time; across different applications; and across regions and Internet service providers, even within a single country. For example, previous research found that governments target a variety of services such as video portals (e.g., YouTube) [51], news sites (e.g., `bbc.com`) [8], and anonymity tools (e.g., Tor) [53]. For example, Ensafi [21] showed that China's Great Firewall (GFW) actively probes—and blocks upon confirmation—servers suspected to abet circumvention. Many studies show that different countries employ different censorship mechanisms beyond IP address blocking to censor similar content or applications, such as Tor [50]. Occasionally, countries also deploy new censorship technology shortly before significant political events. For example Aryan [7] studied censorship in Iran before and after the June 2013 presidential election. The observations of variable and volatile filtering practices underscore the need for our work, since none of the existing techniques capture such variations.

III. Method Overview

In this section, we provide an overview of the measurement method that we developed to detect filtering. We frame the design goals that we aim to achieve and the core technique underlying our approach. Then in Section IV we provide a detailed explanation of the system's operations.

A. Design Goals

We first present a high-level overview of the strategy underlying our method, which we base on inducing and observing potential increments in an Internet host's IP ID field. The technique relies on causing one host on the Internet to send traffic to another (potentially blocked) Internet destination; thus, we also consider the ethics of the approach. Finally, we discuss the details of the method, including how we select the specific Internet endpoints used to conduct the measurements.

Ultimately, the measurement system that we design should achieve the following properties:

- *Scalable.* Because filtering can vary across regions or ISPs within a single country, the system must be able to assess the state of filtering from a large number of vantage points. Filtering will also vary across different destinations, so the system must also be able to measure filtering to many potential endpoints.
- *Efficient.* Because filtering practices change over time, establishing regular baseline measurements is important, to expose transient, short-term changes in filtering practices, such as those that might occur around political events.
- *Sound.* The technique should avoid false positives and ensure that repeated measurements of the same phenomenon produce the same outcome.
- *Ethical.* The system design must satisfy the ethical principles from the Belmont [9] and Menlo [19] Reports: respect for people, beneficence, justice, and respect for law and public interest.

We present a brief overview of the scanning method before explaining how the approach satisfies the design goals above.

B. Approach

The strategy behind our method is to leverage the fact that when an Internet host generates and sends IP packets, each generated packet contains a 16-bit IP identifier ("IP ID") value that is intended to assist endpoints in re-assembling fragmented IPv4 packets. Although path MTU discovery now largely obviates the need for IP fragmentation, senders still generate packets with IP ID values. There are only 2^{16} unique IP ID values, but the intent is that subsequent packets from the same host should have different IP ID values.

When an Internet host generates a packet, it must determine an IP ID to use for that packet. Although different hosts on the Internet use a variety of mechanisms to determine the IP ID for each packet (e.g., random, counter-based increment per-connection or per-interface), many hosts use a single global counter to increment the IP ID value for all packets that originate from that host, regardless of whether the packets it generates bear a relationship to one another. In these cases where the host uses a single IP ID counter, the value of the counter at any time reflects how many packets the host has generated. Thus, the ability to observe this counter over time gives an indication of whether a host is generating IP packets, and how many.

The basic method involves two mechanisms:

- *Probing:* A mechanism to observe the IP ID value of a host at any time.
- *Perturbation:* A mechanism to send traffic to that same host from different Internet destinations, which has the property of inducing the initial host to respond, thus incrementing its IP ID counter.

We now describe the basic design for probing and perturbation, in the absence of various complicating factors such as cross-traffic or packet loss. Figure 1 illustrates the process.

To *probe* the IP ID value of some host over time, a measurement machine sends unsolicited TCP SYN-ACK packets to the host and monitors the responses—TCP RST packets—to track the evolution of the host's IP ID. We monitor the IP ID values at the host on one end of the path. We call this host the *reflector*, to denote that the host reflects RST packets from both our measurement machine and the endpoint that a censor may be trying to filter. This reflector is a machine in a network that may experience IP filtering. We call the other endpoint of this connection the *site*, as for our purposes we will commonly use for it a website operating on port 80.

To *perturb* the IP ID values on either end of the path, a measurement machine sends a TCP SYN packet to one host, the site; the TCP SYN packet carries the (spoofed) source IP address of a second machine, the reflector. We term this *injection*. If no filtering is taking place, the SYN packet from the measurement machine to the site will elicit a SYN-ACK from the site to the reflector, which will in turn elicit a RST from the reflector to the site (since the reflector had not previously sent a TCP SYN packet for this connection). When the reflector sends a RST packet to the site, it uses a new IP ID. If the reflector generates IP ID values for packets based on a

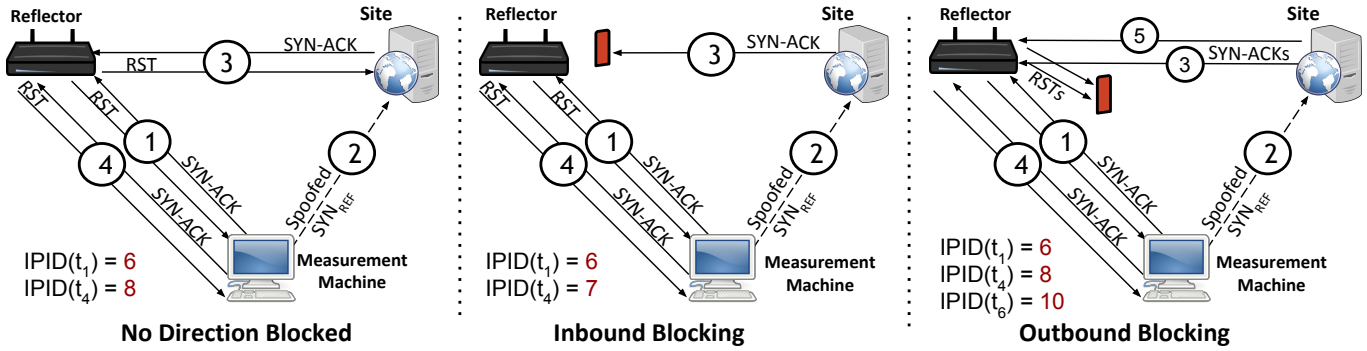

Fig. 1: Overview of the basic method of probing and perturbing the IP ID side channel to identify filtering. Reflectors are hosts on the Internet with a global IP ID. Sites are potentially filtered hosts that respond to SYN packets on port 80. (In the right hand figure, we omit subsequent measuring of the reflector's IP ID by the measurement machine at time t_6). Spoofed SYN packets have a source field set to the reflector.

single counter, the measurement machine can observe whether the reflector generated a RST packet with subsequent probes, because the IP ID counter will have incremented by two (one for the RST to the site, one for the RST to our measurement machine). Figure 1 shows this process in the "no direction blocked" scenario.

Suppose that filtering takes place on the path between the site and the reflector (i.e., one of the other two cases shown in Figure 1). We term blocking that manifests on the path from the site to the reflector as *inbound* blocking. In the case of inbound blocking, the site's SYN-ACK packet will not reach the origin, thus preventing the expected IP ID increment at the reflector. In the absence of other traffic, the IP ID counter will increment by one. We show this in the second section of Figure 1.

Conversely, we call blocking on the path from the reflector to the site *outbound* blocking; in the case of outbound blocking, SYN-ACK packets from the site reach the reflector, but the RST packets from the reflector to the site never reach the site. At this point, the site should continue to retransmit SYN-ACK packets [49], inducing further increments in the IP ID value at the reflector at various intervals, though whether and how it actually does so depends on the configuration and specifics of the site's operating system. The final section of Figure 1 shows the retransmission of SYN-ACK packets and the increment of the global IP ID at two different times. If our measurements reveal a site as inbound-blocked, filtering may actually be bidirectional. We cannot differentiating between the two using this technique because there is no way to remotely induce the reflector to send packets to the site.

C. Ethics

The measurement method we develop generates spoofed traffic between the reflector and the site which might cause an inexperienced observer of these measurements to (wrongly) conclude that the person who operates or owns the reflector was willfully accessing the site. The risks of this type of activity are unknown, but are likely to vary by country. Although the spoofed nature of the traffic is similar to common large-scale denial-of-service backscatter [37] and results in no data packets being exchanged between reflector and site, we nonetheless use extreme caution when selecting each reflector. In this type of measurement, we must first consider *respect for humans*, by limiting the potential harm to any person as a result of this experiment. One mechanism for demonstrating respect for humans is to obtain informed consent; unfortunately, obtaining informed consent is difficult, due to the scope, scale, and expanse of the infrastructure that we employ.

Salganik explains that the inability to obtain informed consent does not by itself reflect a disregard of respect for humans [44]. Rather, we must take other appropriate measures to ensure that we are abiding by the ethical principles from the Belmont [9] and Menlo [19] reports. To do so, we develop a method that reduces the likelihood that we are directly involving any humans in our experiments in the first place, by focusing our measurements on *infrastructure*. Specifically, our method works to limit the endpoints that we use as reflectors to likely Internet infrastructure (e.g., routers in the access or transit networks, middleboxes), as opposed to hosts that belong to individual citizens (e.g., laptops, desktops, home routers, consumer devices). To do so, we use the CAIDA Ark dataset [11], which contains traceroute measurements to all routed /24 networks. We include a reflector in our experiments only if it appears in an Ark traceroute at least two hops away from the traceroute endpoint. The Ark dataset is not comprehensive, as the traceroute measurements are conducted to a randomly selected IP address in each /24 prefix. Restricting the set of infrastructure devices to those that appear in Ark restricts the IP addresses we might be able to discover with a more comprehensive scan.

Although this approach increases the likelihood that the reflector IP addresses are routers or middleboxes as opposed to endpoints, the method is not fool-proof. For example, devices that are attributable to individuals might still be two hops from the network edge, or a network operator might

be held accountable for the *perceived* actions performed by the machines. Our techniques do not eliminate risk. Rather, in accordance with the ethical guideline of *beneficence*, they reduce it to the point where the benefits of collecting these measurements may outweigh the risks of collecting them. In keeping with Salganik's recommendations [44], we aim to conduct measurements that pose a *minimal additional risk*, given both the nature of the spoofed packets and the potential benefits of the research.

The Internet-wide scans we conduct using ZMap [20] to detect possible reflectors introduce concerns related to *respect for law and public interest*. Part of the respect of law and public interest is to reduce the network load we induce on reflectors and sites, to the extent possible, as unnecessary network load could drive costs higher for the operators of reflectors and sites; if excessive, the probing traffic could also impede network performance. To mitigate these possible effects, we follow the approach for ethical scanning behavior as outlined by Durumeric et al. [20]: we signal the benign intent of our scans in the WHOIS entries and DNS records for our scanning IPs, and provide project details on a website hosted on each scanning machine. We extensively tested our scanning methods prior to their deployment; we also respect opt-out requests.

The measurement probes and perturbations raise similar concerns pertaining to respect for law and public interest. We defer the details of the measurement approach to Section IV but note that reflectors and sites receive an average of one packet per second, with a maximum rate of ten SYN packets in a one-second interval. This load should be reasonable, given that reflectors represent Internet infrastructure that should be able to sustain modest traffic rates directed to them, and sites are major websites that see much higher traffic rates than those we are sending. To ensure that our TCP connection attempts do not use excessive resources on sites or reflectors, we promptly reset any half-open TCP connections that we establish.

The ethical principle of *justice* states that the parties bearing the risk should be the same as those reaping the benefits; the parties who would bear the risk (users in the countries where censorship is taking place) may ultimately reap some benefit from the knowledge about filtering that our tools provide through improved circumvention tools and better information about what is blocked.

IV. Augur: Putting the Method to Practice

In this section, we present our approach for identifying reflectors and sites, and then develop in detail how we perform the measurements described in Section III.

A. Reflector Requirements

Suitable reflectors must satisfy four requirements:

1) **Infrastructure machine.** To satisfy the ethical guidelines that we outlined in Section III-C, the reflector should be Internet infrastructure, as opposed to a user machine.
2) **RST packet generation.** Reflectors must generate TCP RST packets when receiving SYN-ACKs for unestablished connections. The RST packets increment the reflector's IP ID counter while ensuring that the site terminates the connection.
3) **Shared, monotonically incrementing IP ID.** If a reflector uses a shared, monotonic strictly increasing per-machine counter to generate IP ID values for packets that it sends, the evolution of the IP ID value—which the measurement machine can observe—will reflect any communication between the reflector and any other Internet endpoints.
4) **Measurable IP ID perturbations.** Because the IP ID field is only 16 bits, the reflector must not generate so much traffic so as to cause the counter value to frequently wrap around between successive measurement machine probes. The natural variations of the IP ID counter must also be small compared to the magnitude of the perturbations that we induce.

Section V describes how we identify reflectors that meet these requirements.

B. Site Requirements

Our method also requires that sites exhibit certain network properties, allowing for robust measurements at reflectors across the Internet. Unlike reflectors, site requirements are not absolute. In some circumstances, failure to meet a requirement requires discarding of a result, or limits possible outcomes, but we can still use the site for some measurements.

1) **SYN-ACK retransmission (SAR).** SYN-ACK retries by sites can signal outbound blocking due to a reflector's RST packets not reaching the site. If a site does not retransmit SYN-ACKs, we can still detect inbound blocking, but we cannot distinguish instances of outbound blocking from cases where there is no blocking.
2) **No anycast.** If a site's IP address is anycast, the measurement machine and reflector may be communicating with different sites; in this case, RSTs from the reflector will not reach the site that our measurement machine communicates with, which would result in successive SYN-ACK retransmissions from the site and thus falsely indicate outbound blocking.
3) **No ingress filtering.** If a site's network performs ingress filtering, spoofed SYN packets from the measurement machine may be filtered if they arrive from an unexpected ingress, falsely indicating inbound blocking.
4) **No stateful firewalls or network-specific blocking.** If a site host or its network deploys a distributed stateful firewall, the measurement machine's SYN packet may establish state at a different firewall than the one encountered by a reflector's RSTs, thus causing the firewall to drop the RSTs. This effect would falsely indicate outbound blocking. Additionally, if a site or its firewall drops traffic from some IP address ranges but not

others (e.g., from non-local reflectors), the measurement machine may falsely detect blocking.

Section V-E describes how we identify sites that satisfy these requirements.

C. Detecting Disruptions

As discussed in Section III, we detect connectivity disruptions by perturbing the IP ID counter at the reflector and observing how this value evolves with and without our perturbation.

Approach: Statistical detection. We measure the natural evolution of a reflector's counter periodically in the absence of perturbation as a control that we can compare against the evolution of the IP ID under perturbation. We then perturb the IP ID counter by injecting SYN packets and subsequently measure the evolution of this counter. We take care not to involve any site or reflector in multiple simultaneous measurements, since doing so could conflate two distinct results.

Ultimately, we are interested in detecting whether the IP ID evolution for a reflector changes as a result of the perturbations we introduce. We can represent this question as a classical problem in statistical detection, which attempts to detect the presence or absence of a prior (i.e., perturbation or no perturbation), based on the separation of the distributions under different values of the prior. In designing this detection method, we must determine the random variable whose distribution we wish to measure, as well as the specific detection approach that allows us to distinguish the two values of the prior with confidence. We choose IP ID *acceleration* (i.e., the second derivative of IP ID between successive measurements) as ideally this value has a zero mean, regardless of reflector. With a zero mean, the distribution of the random variable should be stationary and the distribution should be similar across reflectors. Conceptually, this can be thought of as a reflector, at a random time, being as likely to experience traffic "picking up" as not. However, subtle Internet complexities such as TCP slow start bias this measure slightly. We discuss empirical measures of these priors and their impact on our method in Section V-D.

In contrast, the first derivative (IP ID velocity) is not stationary. Additionally, each reflector would exhibit a different mean velocity value, requiring extensive per-reflector baseline measurements to capture velocity behavior.

Detection framework: Sequential hypothesis testing (SHT). We use sequential hypothesis testing (SHT) [31] for the detection algorithm. SHT is a statistical framework that uses repeated trials and known outcome probabilities (*priors*) to distinguish between multiple hypotheses. The technique takes probabilities for each prior and tolerable false positive and negative rates as input and performs repeated online trials until it can determine the value of the prior with the specified false positive and negative rates. SHT's ability to perform online detection subject to tunable false positive/negative rates, and its tolerance to noise, make it well-suited to our detection task. Additionally, it is possible to compute an expectation for the number of trials required to produce a detection, thus enabling efficient measurement.

We begin with the SHT formulation developed by Jung et al. [31], modifying it to accommodate our application. For this application to hold, the IP ID acceleration must be stationary (discussed more in Section V-D), and the trials must be independent and identically distributed (i.i.d.). To achieve i.i.d., we randomize our trial order and mapping between sites and reflectors and run experiments over the course of weeks.

For a given site S_i and reflector R_j, we perform a series of N trials, where we inject spoofed SYN packets to S_i and observe IP ID perturbations at R_j. We let $Y_n(S_i, R_j)$ be a random variable for the nth trial, such that:

$$Y_n(S_i, R_j) = \begin{cases} 0 & \text{if no IP ID acceleration occurs} \\ 1 & \text{if IP ID acceleration occurs} \end{cases}$$

during the measurement window following injection. We identify two hypotheses: H_0 is the hypothesis that no inbound blocking is occurring (the second derivative of IP ID values between successive measurements should be observed to be positive, which we define as IP ID acceleration), and H_1 is the hypothesis that blocking is occurring (no IP ID acceleration). Following constructions from previous work, we must identify the prior conditional probabilities of each hypothesis, specifically:

$$\Pr[Y_n = 0|H_0] = \theta_0, \quad \Pr[Y_n = 1|H_0] = 1 - \theta_0$$
$$\Pr[Y_n = 0|H_1] = \theta_1, \quad \Pr[Y_n = 1|H_1] = 1 - \theta_1$$

The prior θ_1 is the probability of no observed IP ID acceleration in the case of inbound blocking. We can experimentally measure this prior as the probability of IP ID acceleration during our reflector control measurements, since the IP ID acceleration likelihood during control measurements is the same as during inbound blocking (as no additional packets reach the reflector in both cases). Intuitively, we can think of this value as 0.5 given the prior discussion of second-order value being thought of as zero mean (i.e., in aggregate traffic, with no induced behavior, acceleration is as likely to occur as deceleration).

The prior $1 - \theta_0$ is the probability of observed IP ID acceleration during injection. It can be measured as the probability of IP ID acceleration during an injection period across all reflector injection measurements. Assuming no blockage and perfect reflectors with no other traffic, this value can be thought of as approaching 1. The prior can be estimated from all reflector measurements under the assumption that blocking is uncommon for a reflector. However, even if the assumption does not hold and blocking is common, the prior estimation is still *conservative* in that it drives the prior closer to the θ_1, making detection more difficult, increasing *false negatives*.

From the construction above, we define a likelihood ratio $\Lambda(Y)$, such that:

$$\Lambda(Y) \equiv \frac{\Pr[Y|H_1]}{\Pr[Y|H_0]} = \Pi_{n=1}^{N} \frac{\Pr[Y_n|H_1]}{\Pr[Y_n|H_0]}$$

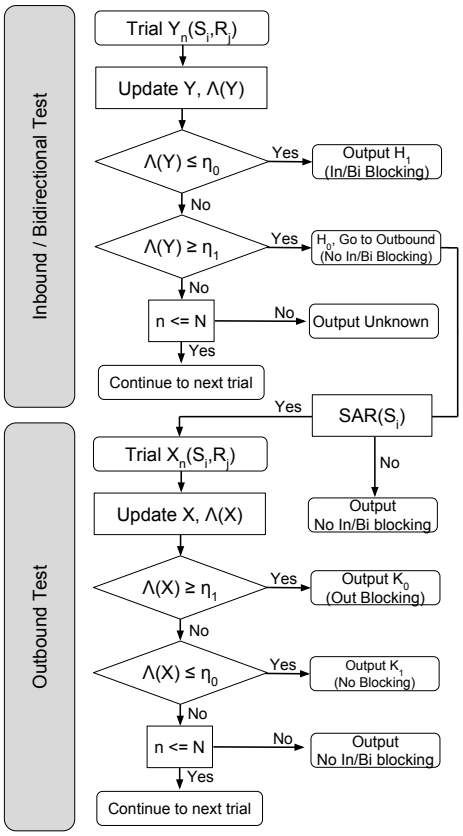

Fig. 2: Flow chart of our algorithm to identify both inbound and outbound blocking using a series of sequential hypothesis tests. Detailed descriptions of the notation and terminology are given in Section IV-C.

where Y is the sequence of trials observed at any point. We derive an upper bound threshold η_1 such that:

$$\frac{\Pr[Y_1, \ldots Y_N | H_1]}{\Pr[Y_1, \ldots Y_N | H_0]} \geq \eta_1$$

and a similar lower bound threshold η_0. Both η_0 and η_1 are bounded by functions of the tolerable probability of false positives and negatives. We elaborate on these bounds and the impact of false positives and negatives later in this section.

Figure 2 illustrates our detection algorithm, which performs a series of sequential hypothesis tests; the rest of this section describes this construction in detail. The Inbound Blocking portion of Figure 2 shows how SHT uses this construction to make decisions. This is extended to include outbound blocking subsequently.

As each trial is observed, we update the likelihood ratio function $\Lambda(Y)$ based on the prior probabilities. Once updated, we compare the value of $\Lambda(Y)$ against the thresholds η_0 and η_1. If $\Lambda(Y) \leq \eta_0$, we accept H_1 and output Input or Bidirectional Blocking.

If $\Lambda(Y) \geq \eta_1$, we accept H_0, which is that IP ID acceleration occurred as a result of no inbound blocking. This does not give us a result, as we still must decide between outbound blocking and no blocking. To make this decision, we proceed to the second SHT phase,"Outbound Test," which is discussed subsequently.

A third output of the system is that $\Lambda(Y)$ did not meet either threshold. If there are more trials we restart the algorithm. If we have exhausted our trials, we output the result blockage that of S_i at R_j is undetermined.

Outbound blocking detection with SHT. Given IP ID acceleration at the reflector, we must distinguish outbound-only blocking from a lack of blocking whatsoever. To do so, we develop a key new insight that relies on a secondary IP ID acceleration that should occur due to subsequent SYN-ACK retries by the site.

To determine a site's eligibility for outbound blocking detection, we must identify whether it retries SYN-ACKs, and that the retries have reliable timing. Section V discusses these criteria further. We abstract this behavior as a function SAR(S_i) (for SYN-ACK Retry) that indicates whether a site is suitable for outbound blocking detection. We define $X_n(S_i, R_j)$ such that:

$$X_n(S_i, R_j) = \begin{cases} 0 & \text{if no IP ID accel. during SAR} \\ 1 & \text{if IP ID accel. during SAR} \end{cases}$$

We now formulate two new hypotheses, K_0 such that outbound blocking is occurring (IP ID acceleration occurs during the SAR time window), and K_1 such that there is no connection blocking (IP ID acceleration does not occur during the SAR window). From this:

$$\Pr[X_n = 0 | K_0] = \theta_0, \quad \Pr[X_n = 1 | K_0] = 1 - \theta_0$$
$$\Pr[X_n = 0 | K_1] = \theta_1, \quad \Pr[X_n = 1 | K_1] = 1 - \theta_1$$

In this construction $1 - \theta_0$ is the measurable probability of observing IP ID acceleration during injection, and θ_1 is the measurable prior probability of seeing no IP ID acceleration during the SAR window across all of the reflector's measurements. Similar arguments hold as above to why these provide conservative estimations of the prior values. (We also discuss the measurable IP ID acceleration during the SAR window in Section V-D.) Figure 2 shows how this construction is used to label S_i, R_j as either outbound-blocked or not blocked. If the thresholds are not met and there are no more trials, we output that we know S_i is not inbound-blocked, but we do not know the outbound-block status.

Expected number of trials. The SHT construction from Jung et al. also provides a framework for calculating the expected number of trials needed to arrive at a decision for H_0 and H_1. The expected values are defined as:

$$E[N|H_0] = \frac{\alpha \ln \frac{\beta}{\alpha} + (1-\alpha) \ln \frac{1-\beta}{1-\alpha}}{\theta_0 \ln \frac{\theta_1}{\theta_0} + (1-\theta_0) \ln \frac{1-\theta_1}{1-\theta_0}},$$
$$E[N|H_1] = \frac{\beta \ln \frac{\beta}{\alpha} + (1-\beta) \ln \frac{1-\beta}{1-\alpha}}{\theta_1 \ln \frac{\theta_1}{\theta_0} + (1-\theta_1) \ln \frac{1-\theta_1}{1-\theta_0}}. \quad (1)$$

where α and β are also bounded by functions of the tolerable false positive and negative rates, discussed subsequently.

Similar constructions hold for K_0 and K_1. We investigate the expected number of trials for both inbound and outbound blocking further in Section V-D.

False positives and negatives. Following the construction from Jung et al., α and β are both tunable parameters which are bounded by our tolerance to both false positives and false negatives. P_F is defined as the false positive probability, and P_D as the detection probability. The complement of P_D, $1 - P_D$ is the probability of false negatives. These values express the probability of a false result for a single SHT experiment. However, for our method, we perform numerous SHT experiments across sites and reflectors. To account for these repeated trials we set both P_F and $1-P_D = 10^{-5}$. Given that as P_F and $1-P_D$ decrease, the expected number of trials to reach a decision increases, our selection of a small value negatively impacts our ability to make decisions. This effect is somewhat mitigated by the distance between experimentally observed priors, and is explored in more detail in Section V-D and Figure 4.

V. Augur Implementation and Experiment Data

In this section, we discuss the deployment of our approach to measure connectivity disruptions across the Internet, as well as the setup that we use to validate the detection method from Section IV.

A. Selecting Reflectors and Sites

Reflector selection. To find reflectors that satisfy the criteria from Section IV, we created a new ZMap [20] probe module that sends SYN-ACK packets and looks for well-formed RST responses. Our module is now part of the open-source ZMap distribution. Using this module, we scan the entire IPv4 address space on port 80 to identify possible reflectors.

We then perform a second set of probes against this list of candidate reflectors to identify a subset that conforms to the desired IP ID behavior. Our tool runs from the measurement machine and sends ten SYN-ACK packets to port 80 of each host precisely one second apart, recording the IP ID of each RST response. We identify reflectors whose IP ID behaviors satisfy the previously outlined requirements: no IP ID wrapping, variable accelerations observed (indicating our packets do induce perturbations in the IP ID dynamics), and a response to all probes. Because the measurement machine induces packet generation at the reflector at a constant rate, any additional IP ID acceleration must be due to traffic from other connections. We further ensure that the measurement machine receives a response for each probe packet that it sends, ensuring that the reflector is stable and reliable enough to support continuous measurements.

This selection method identifies viable reflectors, those that are responsive and exhibit the desired IP ID behavior. We finally filter the viable reflectors that do not correspond to infrastructure, as described in Section III-C, which significantly reduces the number of available reflectors, as described in Section V-E.

Reflector Datasets	Total Reflectors	Num. Countries	Median / Country
All Viable	22,680,577	234	1,667
Ethically Usable	53,130	179	15
Experiment Sample	2,050	179	15

TABLE I: Summary of our reflector datasets. All viable reflectors are identified across the IPv4 address space. Those ethically usable are routers at least two hops away from traceroute endpoints in the Ark data, and we select a random subset as our experiment set.

Reflector Dataset	AF	AS	EU	NA	SA	OC	ME
All Viable	55	50	52	39	23	14	20
Ethically Usable	36	47	46	30	14	6	18
Experiment Sample	36	47	46	30	14	6	18

TABLE II: The distribution of countries containing reflectors across continents. Note the continent coverage of our experiment sample is identical to that of the ethically usable dataset, as we sampled at least one ethically usable reflector per country in that dataset. The continent labels are as follows: AF=Africa, AS=Asia, EU=Europe, NA=North America, SA=South America, OC=Oceana/Australia. We also label ME=Middle East, as a region with frequent censorship.

Site selection. We begin with a list of sites, some of which are expected to be disrupted by network filtering or censorship from a variety of vantage points. We seed our candidate sites with the Citizen Lab list of potentially censored URLs [15], which we call the *CLBL*. This list contains potentially blocked URLs, broken down by category. To further identify sensitive URLs, we use Khattak et al.'s dataset [32] that probed these URLs using the OONI [40] measurement platform looking for active censorship. After filtering the list, we distill the URLs down to domain names and resolve all domains to the corresponding IP addresses using the local recursive DNS resolver on a network in the United States. If a domain name resolves to more than one IP, we randomly select one A record from the answers. To augment this list of sites, we randomly select domains from the Alexa top 10,000 [2]. As with the CLBL, if a host resolves to multiple IPs, we select one at random. Section V-B provides a breakdown of the site population. Section V-E explains how we dynamically enforce site requirements.

B. Measurement Dataset

In this section, we describe the characteristics of the dataset that we use for our experiments.

Reflector dataset. The geographic distribution of reflectors illuminates the degree to which we can investigate censorship or connectivity disruption within each country. Table I summarizes the geographic diversity of our reflector datasets.

The Internet-wide ZMap scan found 140 million reachable hosts. Approximately 22.7 million of these demonstrated use of a shared, monotonically increasing IP ID. These reflectors were geographically distributed across 234 countries around the world, with a median of 1,667 reflectors per country. This initial dataset provides a massive worldwide set of reflectors to potentially measure, yet many may be home routers, servers, or user machines that we cannot use for experimentation due to ethical considerations.

Merging with the Ark to ensure that the reflectors only contain network infrastructure reduces the 22.7 million potential reflectors to only about 53,000. Despite this significant reduction, the resulting dataset contains reflectors in 179 countries, with a median of 15 reflectors per country. Table II gives a breakdown of reflector coverage by continent.

We select a subset of these reflectors as our final experiment dataset, randomly choosing up to 16 reflectors in all 179 countries, yielding 1,947 reflectors (not all countries had 16 infrastructure reflectors). In addition to these reflectors, we added 103 high-reliability (stable, good priors) reflectors primarily from China and the US to ensure good coverage with a stable set of reflectors, resulting in 2,050 reflectors in the final dataset. These reflectors also exhibit widespread AS diversity, with the resulting set of reflectors representing 31,188 ASes. Using the Ark dataset to eliminate reflectors that are not infrastructure endpoints reduces this set to 4,214 ASes, with our final experiment sample comprising 817 ASes.

Site dataset. Merging the CLBL with Khattak et al.'s dataset [32] yields 1,210 distinct IP addresses. We added to this set an additional 1,000 randomly selected sites from the Alexa top 10,000. To this set of sites we also added several known Tor bridges, as discussed in Section VI-C. While this set consists of 2,213 sites, some sites appeared in both the CLBL and Alexa lists. Thus, our site list contains a total of 2,134 unique sites, with a CLBL composition of 56.7%.

C. Experiment Setup

The selection process above left us able to measure connectivity between 2,134 sites and 2,050 reflectors. We collected connectivity disruption network measurements over 17 days, using the method described in Section IV. We call one measurement of a reflector-site pair a *run*, involving IP ID monitoring and one instance of blocking detection. Related, we define an experiment *trial* as the complete measurement of one run for all reflector-site pairs. Over our 17-day window, we collected a total of 207.6 million runs across 47 total trials, meaning we tested each reflector-site pair 47 times.

Each run comprises of a collection of one-second time intervals. For each time interval, we measure the IP ID state of the reflector independent of all other tasks. We begin each run by sending a non-spoofed SYN to the site from the measurement machine. Doing so performs several functions. First, it allows us to ensure that the site is up and responding to SYNs at the time of the measurement. Second, it allows us to precisely measure if the site sends SYN-ACK retries,

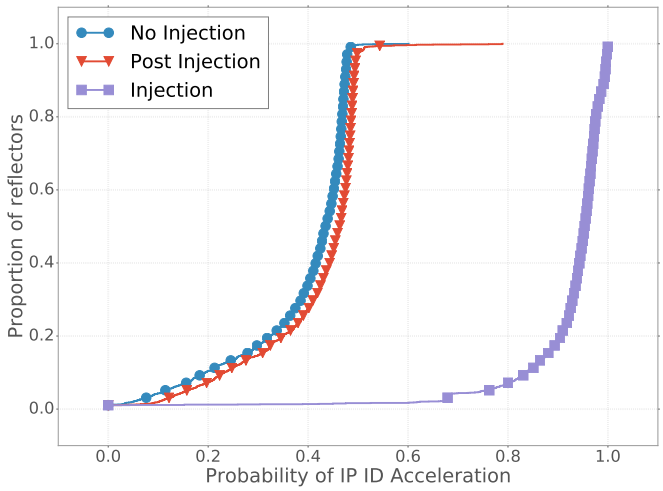

Fig. 3: CDF of probability of IP ID acceleration per reflector across the experiment.

and to characterize the timing of the retries. We record this behavior for each run and incorporate this initial data point into the subsequent SHT analysis. We then wait four seconds before injecting spoofed SYN packets towards the site. The reflector measurements during that window serve as *control* measurements. During the injection window, we inject 10 spoofed SYN packets towards the site.

For each run, we denote the SYN-ACK retry behavior and at what subsequent window we expect SYN-ACK retries to arrive at the reflector, and use this information to identify which window to look for follow-on IP ID acceleration. At the end of the run, we send corresponding RST packets for all SYNs we generated, to induce tear-down of all host state. We then cool down for 1 second before starting a new run. We randomize the order of the sites and reflectors for testing per trial. We test all reflector-site pairs before moving on to a new trial. For reasons discussed earlier, we never involve the same reflector and site in two independent simultaneous measurements between endpoints.

After each run, we ensure that (1) the reflector's IP ID appeared to remain monotonically increasing; (2) no packet loss occurred between the measurement machine and the reflector, and (3) the site is up and responding to SYN packets. Additionally, we ensure that the IP ID does not wrap during either the injection window or the SAR window. We discard the measurements if any of these conditions fails to hold. After these validity checks, our dataset contains 182.5 million runs across 1,960 reflectors and 2,089 sites. The reduction in number of sites and reflectors corresponds to unstable or down hosts. We then apply SHT (Section IV) to analyze the reachability between these site-reflector pairs.

D. Measured Priors and Expectations

A critical piece in the construction of our SHT framework is formulating the prior probabilities for each of our hypotheses.

Figure 3 shows CDFs of the measured prior probabilities of IP ID acceleration for three different scenarios.

The IP ID acceleration of reflectors matches our intuition, where the acceleration decreasing as frequently as it increases across the dataset. We show this with the "No Injection" CDF, with nearly all reflectors having a probability of IP ID acceleration without injection of less than 0.5. Many reflectors have a probability of acceleration far lower, corresponding to reflectors with low or stable traffic patterns. We then use this per-reflector prior for θ_1 in our SHT construction for detecting inbound blocking. While we could instead estimate the value as 0.5, the expected number of trials depends on the separation between the injection and non-injection priors, so if we are able to use a smaller θ_1 (per reflector), this greatly speeds up detection time.

Figure 3 also shows the probability of IP ID acceleration under injection. This value approaches 1 for many reflectors and is above 0.8 more than 90% of reflectors. Noticeably, it is, however, quite low, and even 0 for a handful of reflectors. These correspond to degenerate or broken reflectors that we can easily identify due to their low priors, removing them from our experiment (discussed more in Section V-E). We use this experimentally measured prior as $1 - \theta_0$ in both of our sequential hypothesis tests. This distribution provides a lower bound for the actual probability of IP ID acceleration, as the experimentally measured value includes inbound blocking (i.e., if some sites experience blocking, those values would *lower* the measured value). Inbound-blocked runs lower the overall probability of acceleration. This still reflects a *conservative* measurement, as a prior closer to control increases the likelihood of false negatives, not false positives.

Lastly, we also measure the probability of IP ID acceleration at the SYN-ACK retry point of each run. We dynamically determine where this falls in each run using the properties the site manifests during that run.[1] As expected, the distribution closely matches the control distribution. The differences in the curve are explained by the dataset containing outbound blocking. Such blocking raises the probability of acceleration at that point, pulling the distribution slightly closer to the injection case. We use this prior as θ_1 during the outbound SHT test.

Once we have computed the priors, we can compute the expected number of trials to reach each of our output states (on a per-reflector basis) using Equation 1. Figure 4 presents CDFs of these results. More than 90% of reflectors have 40 or fewer expected trials needed to reach one of the states. The remaining reflectors have a large tail and correspond to unstable or degenerate reflectors. We do not need to explicitly remove these reflectors from the dataset, but must refrain from making decisions based on them in some cases.

[1]If a SYN-ACK retry occurs in the window adjacent to injection, we discard that and look for the next retry. If we did not discard that measurement, the retry would correspond to non-acceleration rather than acceleration.

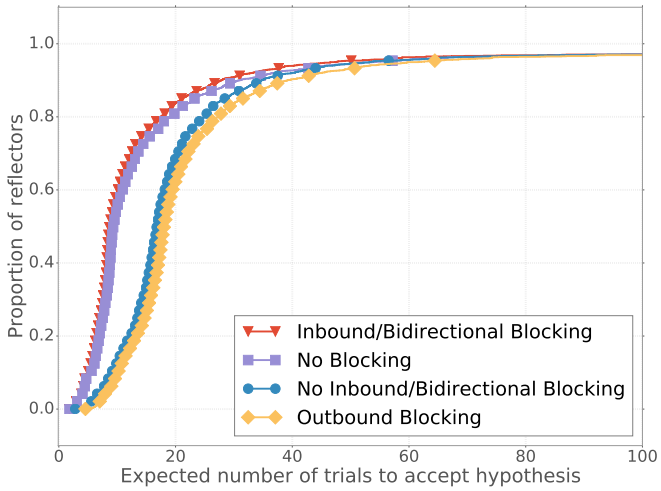

Fig. 4: CDF of expected number of trials at false positive and negative probability of 10^{-5} to accept one of the four SHT hypothesis outcomes, per reflector. "No Inbound/Bidirectional Blocking" means we passed our first SHT and did not detect inbound blocking, but have not yet attempted to differentiate between no blocking and outbound blocking.

E. Identifying and Removing Systematic Effects

Our initial selection of sites did not address some of our site requirements from IV, such as network filtering or anycast IP addresses. Failure to identify these sites generates *systematic effects* within our results dataset. Recall that we only wish to filter these sites when necessary. For example, in the case of anycast sites, we can still classify them as inbound-blocked or not blocked, but we cannot detect scenarios where the site is outbound-blocked.

Problematic sites. We identify sites that fail to meet these requirements by conducting a set of experiments with nine geographically diverse vantage points. These hosts reside in cloud service providers and universities, all of which have limited to no network blocking as vantage points. We perform these measurements concurrently with our primary blockage measurements. For each site, we perform two measurements for each vantage point. The diversity of the vantage points enables us to identify these network effects rather than identify censorship or blockage. These tests do not need to be globally complete as the network effects manifest readily.

The first measurement ensures that a vantage point can have bidirectional communication with a site. From a vantage point, we send five SYN packets to a site, evenly distributed over the experiment run (approximately an hour). We monitor for SYN-ACK replies, which demonstrate two-way communication. If a vantage point cannot reliably establish bidirectional communication with a site, we exclude it from our further vantage-point measurements.

In the second measurement, the measurement machine sends a *spoofed* SYN packet to the site with the IP address of

a vantage point. Since we previously confirmed the vantage point can communicate with the site, any missing SYN-ACKs or retransmissions are the result of sites not conforming to our requirements, rather than blockage. If the vantage point does not receive a SYN-ACK response from the site, ingress filtering or network origin discrimination may be occurring. If the vantage point does receive a SYN-ACK, it responds with a RST packet. If the vantage point continues to receive multiple SYN-ACKs, the site is not correctly receiving the vantage point's RST packets, suggesting the site host (or its network provider) may be anycast, employing a distributed stateful firewall, or discriminating by traffic origin. We repeat this experiment three times to counter measurement errors introduced by random packet loss. If a vantage point never receives a SYN-ACK, or only ever receives multiple SYN-ACK retries, we conservatively conclude the site exhibits one of the unacceptable network properties from that vantage point. Thus, we disregard its blockage results, *except* if the observed measurement results cannot be a false signal due to the site's properties. For example, if vantages observe only multiple SYN-ACK retries for a site (indicating our measurements with that site may falsely identify outbound blocking), but our measurements detect no blocking or only inbound blocking, we can still consider these results.

We find that this relatively small number of vantage points suffices to characterize sites, as experiment results typically remained consistent across all vantage points. All online sites that we tested were reachable from at least three vantage points, with 98.4% reachable at five or more. This reachability affords us with multiple geographic vantage points to assess each site. For 98.6% of sites, all reachable vantage points consistently assessed the site requirement status, indicating that we can detect site network properties widely from a few geographically distinct locations. This approach is ultimately best effort, as we may fail to detect sites whose behavior is more restricted (e.g., filtering only a few networks).

Through our site assessment measurements, we identified 229 sites as invalid for inbound blocking detection due to ingress filtering or network traffic discrimination. These sites were widely distributed amongst 135 ASes, each of which may employ such filtering individually or may experience filtering occurring at an upstream ISP.

We also flagged 431 sites as invalid for outbound blocking detection as they either lacked a necessary site property (discussed in in Section IV-B) or did not respect RST packets (perhaps filtering them). To distinguish between the two behaviors, we probed these sites with non-spoofed SYN and RST packets using vantage points, similar to the experiments described earlier in this section. For each site, we sent a SYN packet from a well-connected vantage, and responded with a RST for any received SYN-ACK. If we continued receiving multiple SYN-ACK retries, the site did not respect our RST packets. Otherwise, the site does properly respond to RST packets in the non-spoofing setup, and might be exhibiting an undesirable site property (as listed in Section IV-B) in our spoof-based connectivity disruption experiments. We iterate this measurement three times for robustness against sporadic packet loss, concluding that a site ignores RST packets if any vantage point observes multiple SYN-ACK retransmissions in all trials.

Using this approach, we identified that 64 sites (14.8% of sites invalid for outbound blocking detection) exhibited a non-standard SYN-ACK retransmission behavior, and conclude that the remaining 367 sites (85.2%) are either anycast, deploying stateful firewalls, or discriminating by network origin. These sites were distributed amongst 62 ASes. The majority are known anycast sites, with 75% hosted by CloudFlare and 7% by Fastly, both known anycast networks.

We additionally checked all sites against the Anycast dataset produced by Cicalese et al. [14]. Our technique identified all but 3 IP addresses. We excluded those 3 sites from our results.

Problematic reflectors. A reflector could be subject to filtering practices that differ based on the sender of the traffic, or the port on which the traffic arrives. This systematic effect can manifest as a reflector with significant inbound or outbound blocking. From manual investigation, we identify several reflectors that demonstrate this property independent of spoofed or non-spoofed traffic. In all cases, such reflectors were outliers within their country. To remove these systematic effects, we ignore reflectors in the 99th percentile of blockage for their country. Sites blocked by these reflectors do not show a bias to the CLBL list (discussed more in Section VI). This process removed 91 reflectors from our dataset.

VI. Validation and Analysis

The value of the method we develop ultimately rests on the ability to accurately measure connectivity disruption from a large number of measurement vantage points. Validating its findings presents challenges, as we lack widespread ground truth, presenting a chicken-and-egg scenario. One approach, presented in Sections VI-A and VI-B, is to analyze the aggregate results produced and confirm they accord with reasonable assumptions about the employment of connectivity disruption. While doing so does not guarantee correctness, it increases confidence in the observations. The other approach is to corroborate our findings against existing ground truth about censored Internet traffic. In Section VI-C, we perform one such analysis, providing a limited degree of more concrete validation.

A. Disruption Bias

Conceptually, one would expect the set of sites disrupted by a network censor to be biased towards sites that are known to be commonly censored. From this notion, we can examine the set of sites blocked by each reflector and ask how that population compares to the input population.

Figure 5 shows, in aggregate, the bias of connectivity disruption towards commonly censored websites. 56.7% of websites in the input site dataset are from the CLBL, demarcated in the plot with a vertical dotted line (which we call the

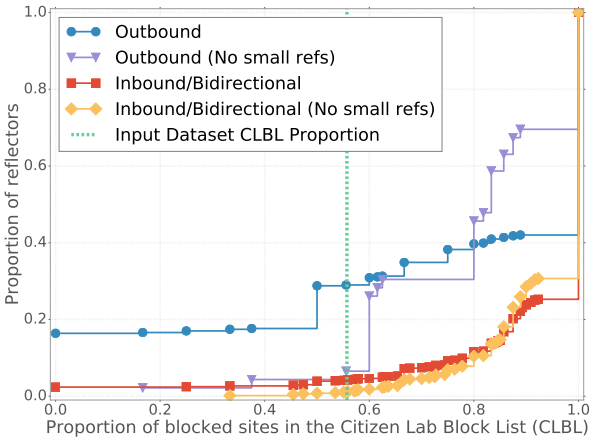

Fig. 5: Bias of blocked sites towards CLBL sites. CLBL sites consist of 56.7% of our sites, demarcated at the dotted vertical line. To reduce small value effects, we remove reflectors with fewer than 5 blocked websites in curves labeled with "No small refs".

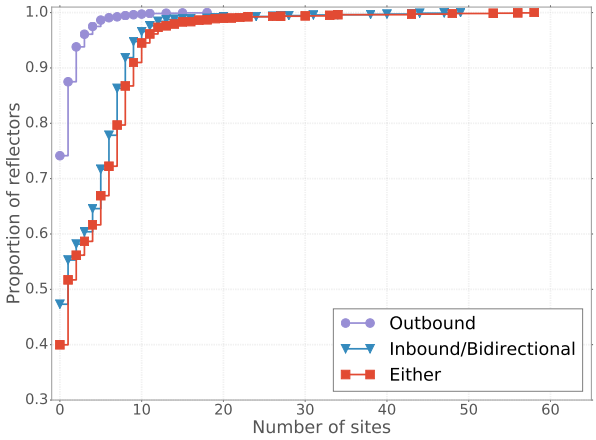

Fig. 6: CDF of site filtering per reflector, separated by inbound/bidirectional and outbound filtering.

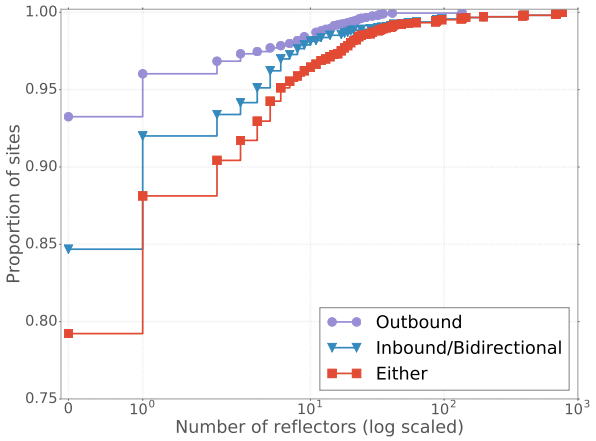

Fig. 7: CDF of site filtering across reflectors, separated by inbound/bidirectional and outbound filtering. Note the log-scaled x-axis.

CLBL bias line). If the detection we observed was unrelated to censorship, we would expect to find roughly 56.7% of that reflector's blocked sites listed in the CLBL. The results, however, show a considerable bias towards CLBL sites for both inbound and outbound filtering. We see this with the bulk of the graph volume lying to the right of the vertical dotted CLBL bias line. Excluding reflectors with fewer than 5 blocked sites to avoid small number effects, we observe that for 99% of reflectors, more than 56.7% of inbound filtering is towards CLBL sites. Similarly, we find 95% of outbound filtering biased towards the CLBL. This observed bias agrees with our prior expectations that we should find CLBL sites more widely censored.

B. Aggregate Results

Site and reflector results. We first explore the extent of connectivity disruption from both the site and reflector perspective. We might naturally assume that filtering will not manifest ubiquitously. We do not expect to find a site blocked across the majority of reflectors; similarly, we should find most sites not blocked for any given reflector. This should particularly hold since approximately half of our investigated sites come from the Alexa top 10K most visited websites. Although some popular Alexa websites contain potentially sensitive content (e.g., adult or social media sites), many provide rather benign content and are unexpected targets of disruption.

We observe the degree of filtering from the reflector perspective in Figure 6. Approximately 99% of reflectors encounter connectivity impediments in either direction for 20 or fewer sites, with no reflector blocked for more than 60 sites. This finding concurs with the assumption that site filtering at reflectors is not ubiquitous. On the other hand, connection disruption appears widespread, as 60% of reflectors experience some degree of interference, corroborating anecdotal observations of pervasive censorship.

We find inbound/bidirectional disruption occurs more commonly compared to outbound-only filtering. In total, fewer than 30% of reflectors experience any outbound-only filtering, while over 50% of reflectors have blocked inbound packets from at least one site. This contrast is unsurprising, because bidirectional filtering of a blacklisted IP address is a simple and natural censorship policy; as a result, most results will appear as either inbound or bidirectional filtering.

Figure 7 depicts a similar outlook on connectivity disruption from the site viewpoint. We again witness that inbound or bidirectional filtering affects more sites than outbound filtering. Over 15% of sites are inbound-blocked along the path to at least one reflector, while only 7% of sites are ever outbound-blocked. In total, connections to 79% of websites never appear disrupted, and over 99% of sites exhibit inaccessibility by

No.	Site	Class	% Refs	% Cnt.
1.	hrcr.org	HUMR	41.7	83.0
2.	alstrangers.[LJ].com	MILX	37.9	78.8
3.	varlamov.ru	ALEXA	37.7	78.0
	nordrus-norna.[LJ].com	HATE		
4.	www.stratcom.mil	FREL	37.5	78.6
5.	www.demonoid.me	P2P	21.7	58.5
6.	amateurpages.com	PORN	21.2	57.9
	voice.yahoo.jajah.com	VOIP		
	amtrak.com	ALEXA		
7.	desishock.net	P2P	10.8	32.7
8.	wzo.org.il	REL	7.9	17.6
9.	hateit.ru	HATE	7.3	14.5
10.	anonymouse.org	ANON	5.3	16.4

TABLE III: Summary of the top 10 sites by the percent of reflectors experiencing *inbound* blocking. Rows sharing rank reflect domains that share an IP address. [LJ] denotes `livejournal`. We list a categorization of each website using the definitions provided in Appendix A. We additionally report the percent of countries for which we find a site inbound-blocked by at least one reflector.

No.	Site	Class	% Refs	% Cnt.
1.	nsa.gov	USMIL	7.4	23.3
2.	scientology.org	MINF	2.2	6.9
3.	goarch.org	MINF	1.9	4.4
4.	yandex.ru	FEXP	1.8	3.8
5.	hushmail.com	EMAIL	1.8	4.4
6.	carnegieendowment.org	POLR	1.6	4.4
7.	economist.com	FEXP	1.6	2.5
8.	purevpn.com	ANON	1.4	1.9
9.	freedominfo.org	FEXP	1.3	3.1
10.	wix.com	HOST	1.3	0.6

TABLE IV: Summary of the top 10 sites by the percent of reflectors experiencing *outbound* blocking. We provide a categorization of each website using definitions provided in Appendix A. We additionally report the percent of countries for which we find a site inbound-blocked by at least one reflector.

100 reflectors (5%) or less. As before, these results agree with our expectation that sites are typically not blocked across the bulk of reflectors.

Several sites show extensive filtering, as listed in Tables III and IV. Here, we have determined reflector country-level geolocation using MaxMind [35]. We found six sites inbound-blocked for over 20% of reflectors across at least half the countries, with the human rights website *hrcr.org* inaccessible by 41.7% of reflectors across 83% of countries. The top 10 inbound-blocked sites correspond closely with anticipated censorship, with 9 found in the Citizen Lab Block List (CLBL). A surprisingly widely blocked Alexa-listed site is *varlamov.ru*, ranked third; in fact, it actually redirects to LiveJournal, a frequent target of censorship [39], [52]. On a related note, the IP address for *amtrak.com* is the sixth most inbound-blocked site—but it is co-located with two CLBL websites, underscoring the potential for collateral damage that IP-based blacklisting can induce.

The top outbound-blocked sites tell a similar tale, although with less pervasive filtering. The most outbound disrupted site is *nsa.gov*, unreachable by 7.4% of reflectors across 23.3% of countries. Given the nature of this site, perhaps the site performs the filtering itself, rather than through reflector-side disruption. All top 10 sites are known frequently blocked websites, listed in the CLBL.

This aggregate analysis of connectivity disruption from both site and reflector perspective accords with our prior understanding that while disruption is not ubiquitous, it may be pervasive. It affects a large proportion of reflectors, and can widely suppress access to particular sites. The sites for which our method detects interference closely correspond with known censored websites. This concordance bolsters confidence in the accuracy of our method's results.

Country-level connectivity disruption. Analysis of aggregate connectivity disruption across countries provides another perspective for validation. Using reflector country geolocation provided by MaxMind [35], Table V ranks the top 10 countries by percentage of blocked sites across any reflectors in the country. Figure 8 portrays this at a global scale, illustrating that some degree of connectivity disruption is experienced by hosts in countries around the world.

We see that many of the most disruptive countries correspond closely with countries known to heavily censor, such as China, Iran, Sudan, Russia, and Turkey [41]. Of the top 10 countries, the OpenNet Initiative [41] has reported Internet censorship of political or social material in every country except Latvia and the United Kingdom.[2] More recently, reports have documented Latvia as heavily censoring gambling websites and political content [4], [47]. Our results appear plausible for the United Kingdom as well, which has a history of filtering streaming and torrent sites [10] and adult content [36].

While we are aggregating at a country granularity, these disruptions may actually be implemented in different ways within a single country. These differences result in non-uniform filtering policies, as has been observed with the Great Firewall of China [24], [54] and UK adult content filtering [36]. In Figure 9, we plot the variation in the number of sites blocked for reflectors within each country. We remove countries without any site filtering. We observe that for most countries, there exists some variation in the disruption experienced by reflectors within a country, suggesting that interference indeed often differs across networks even within a country. The extent of this behavior is widespread and highlights the importance of connectivity measurements from many vantage points, since findings may differ across nearby networks and geolocations.

[2]We list Hong Kong separately from China, although traffic from Hong Kong may traverse Chinese networks and experience disruption.

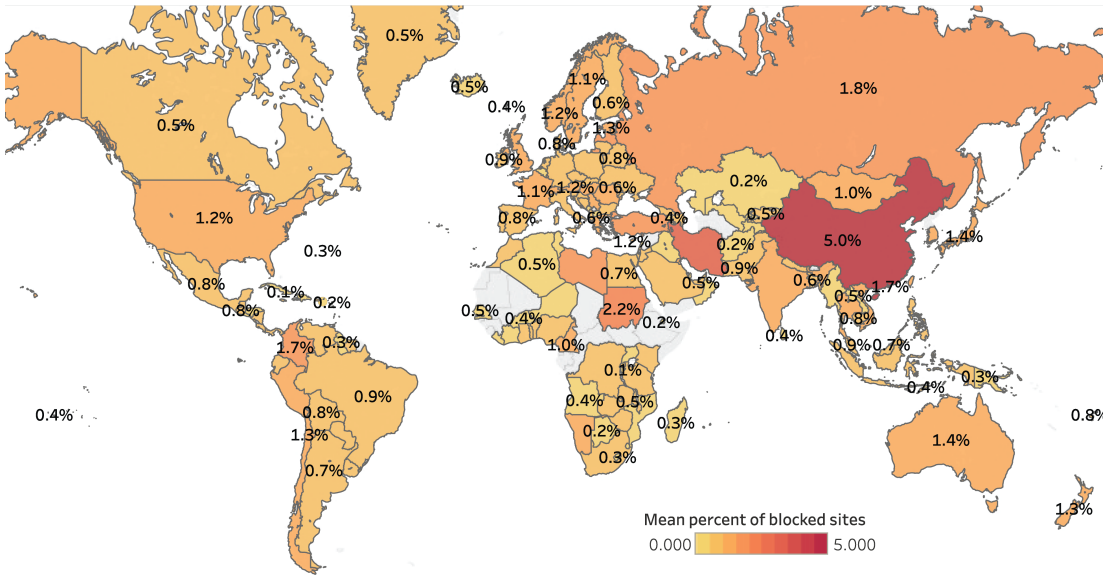

Fig. 8: Global heat map showing the percentage of sites filtered for any reflector in countries around the world. China experiences the highest average amount of filtering, at 5% of measurable sites filtered by a resolver within the country.

No.	Country	Num. Reflectors	Block %	CLBL %	Mean Blocked In/Out	Med. Blocked In/Out	Total Num. Blocked In/Out
1.	China	36	5.0	70.9	11.2 / 1.8	1.5 / 0.0	70 / 33
2.	Iran	14	3.4	55.7	10.8 / 1.4	0.0 / 0.0	53 / 17
3.	Sudan	12	2.2	54.3	6.5 / 0.0	1.0 / 0.0	46 / 0
4.	Russia	17	1.8	78.9	4.8 / 1.4	0.0 / 0.0	18 / 20
5.	Latvia	14	1.8	81.6	3.3 / 1.6	2.0 / 0.0	22 / 19
6.	Turkey	15	1.8	83.8	2.1 / 1.5	0.0 / 0.0	23 / 14
7.	Hong Kong	16	1.7	88.9	2.8 / 1.4	0.0 / 0.0	14 / 22
8.	Columbia	16	1.7	85.7	4.2 / 1.2	6.0 / 0.0	17 / 18
9.	Libya	10	1.5	77.4	8.4 / 3.2	9.5 / 3.0	16 / 15
10.	United Kingdom	16	1.4	90.0	3.1 / 0.8	2.0 / 0.0	19 / 11

TABLE V: Summary of the top 10 countries ranked by the percentage of sites blocked at any reflectors within each country (shown in the "Block %" column). Additionally, we list for each country the number of reflectors within that country, the blockage bias towards CLBL sites, and statistics on inbound versus outbound blockage.

C. Tor Bridge Case Study

In the previous section, we analyzed our method's results in aggregate, finding them in line with reasonable assumptions and existing reports of Internet censorship. Here, we use several known Tor bridges as a case study providing an additional (though limited) check of correctness. This validation increases confidence in our method, as we are able to replicate previous findings with regards to which sites experience blocking, the country of censorship, and the directional nature of disruption.

Our set of sites contains three Tor Obfuscation4 (obfs4) Bridges open on port 80, for which we have some ground truth on their censorship. A prior study [25] tested all three bridges from vantage points in the U.S., China, and Iran, over a five-month period. The first two bridges (TB1 and TB2) were included in the Tor Browser releases. Fifield and Tsai detected that only China frequently *inbound*-blocked these, albeit inconsistently, likely due to the federated nature of the Great Firewall of China. The third bridge (TB3) had been only privately distributed, and remained unblocked throughout the study.

Our findings are consistent with this ground truth. Both TB1 and TB2 experienced inbound filtering in China only, while connectivity to TB3 was never disrupted. Of the 36 reflectors in China, we detected inbound filtering of TB1 for 8 reflectors, no filtering for 8 reflectors, and inconclusive evidence for the remaining 20 (due to lack of a statistically significant signal during our hypothesis testing). For TB2, 9 reflectors were inbound-blocked, 11 were unblocked, and 16 were undecided. TB3, expected to be unblocked, was accessible by 22 reflectors, with the remaining 14 undetermined. These findings accord with prior results regarding the distributed and disparate nature of Chinese Tor filtering.

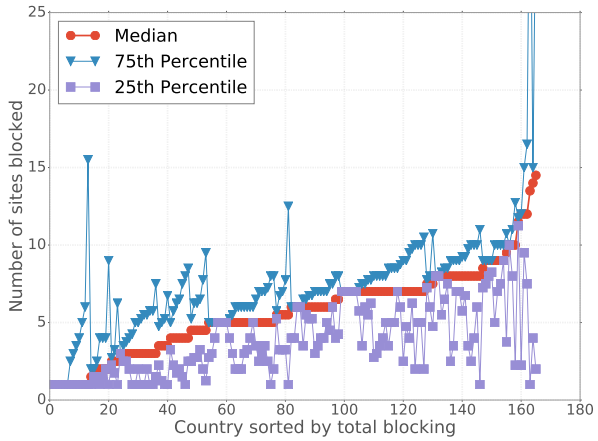

Fig. 9: Plot of the variations in site filtering experienced by reflectors within countries. We elide countries without any disruption.

VII. Discussion

In this section, we discuss various aspects concerning the coverage, granularity, and accuracy of the current measurements.

Coverage limitations. Ethical considerations when performing our measurements restricted the reflectors from which we measure to a set of hosts that we can confidently conclude represent Internet infrastructure in the interior of the network. Recall that we do so by measuring the Internet topology and only using reflectors at least two traceroute hops into the network. This approach drastically reduces the number of hosts that we can use as reflectors. In the future, more exhaustive techniques to identify Internet infrastructure could increase the set of IP addresses that we might use as reflectors.

Evasion Augur relies on the injection of spoofed SYN-ACK packets. A natural evasion mechanism could use a stateful firewall to drop SYN-ACKs that do not correspond to a previously sent SYN. Implementing such firewalls at scale poses significant challenges. Large networks frequently have multiple transit links resulting in asymmetric routing; SYN packets may traverse a different path than the SYN-ACKs. The censor would need to coordinate state across these links. Any errors in state management would lead to blocking benign connections, resulting in collateral damage.

Alternatively, censors could switch to allowing through TCP control packets and only disrupting *data* packets. Such an approach might complicate the censor's own monitoring of their blocking efforts as it runs counter to assumptions commonly made by diagnostic tools. Similarly, it may introduce management burdens because it does not accord with common forms of packet filtering.

Ambiguity in location and granularity of filtering. The current measurements only indicate whether packets became filtered somewhere along the end-to-end path between a reflector and a site; they do not indicate the *location* where that filtering might take place. As a result, our techniques cannot disambiguate the scenario where a remote site blocks access from all reflectors in an entire region from the scenario where an in-country censor filters traffic along that path. For example, financial and commerce sites may block access from entire countries if they have no customers in those regions.

Additionally, the current measurements only employ TCP packets using port 80. Thus, they do not disambiguate filtering of IP addresses versus filtering of only port 80 traffic associated with that IP address. An extension of our system might perform follow-up measurements on different ports to determine whether filtering applies across all ports. On a related note, our techniques only measure TCP/IP-based filtering; future work may involve correlating the measurements that we observe with tools that measure global filtering at other layers or applications (e.g., HTTP, DNS).

Other sources of inaccuracy. Existing IP geolocation tools have known inaccuracies [27], particularly for Internet infrastructure (i.e., IP addresses that do not represent end hosts). As a result, some of our results may not reflect precise characterizations of country-level filtering. As IP geolocation techniques improve, particularly for IP addresses that correspond to Internet infrastructure, we can develop more confidence in the country-level characterizations from Section VI. Additionally, various network mechanisms, including anycast, rerouting, traffic shaping, and transient network failures, may make it difficult to disambiguate overt filtering actions from more benign network management practices. Some of these effects may even operate dynamically: for example, network firewalls may observe our probes over time, come to view them as an attack, and begin to block our probes; in this case, our own measurements may give rise to filtering, rendering it difficult to disambiguate reactive filtering of our measurements from on-path filtering between a site and reflector, particularly since the latter may also change over time.

VIII. Conclusion

Despite the pervasive practice of Internet censorship, obtaining widespread, continuous measurements from a diversity of vantage points has proved elusive; most studies of censorship to-date have been limited both in scale (i.e., concerning only a limited number of vantage points) and in time (i.e., covering only a short time span, with no baseline measurements). The lack of comprehensive measurements about Internet censorship stems from the difficulty of recruiting vantage points across a wide range of countries, regions, and ISPs, as most previous techniques for measuring Internet censorship have required some type of network presence in the network being monitored.

In this paper, we tackle this problem with a fundamentally different type of approach: instead of relying on in-country monitoring points for which we have no direct access, we exploit recent advances in TCP/IP side-channel measurement techniques to collect measurements between pairs of endpoints

that we do not control. This ability to conduct measurements from "third-party" vantage points that we control allows us to continuously monitor many more paths than was previously possible. Previous work introduced the high-level concept of these third-party side-channel measurements; in this work, we transition the concept to practice through a working system that abides by ethical norms and produces sound measurements in the presence of the measurement artifacts and noise that inevitably manifest in real-world deployments.

The continuous, widespread measurements that we can collect with these techniques can ultimately complement anecdotes, news reports, and policy briefings to ensure that we can back future assessments of Internet filtering with sound, comprehensive data. Part of this transition to practice involves further developing the system that we have developed to facilitate ongoing operation, including automating the validation of the measurements that we collect. We aim to ultimately correlate this data with other datasets that pertain to application-layer [45] and DNS-based [6], [29] filtering.

IX. Acknowledgements

The authors are grateful for the assistance and support of Randy Bush, Jed Crandall, David Fifield, Sarthak Grover, and Brad Karp. This work was supported in part by National Science Foundation Awards CNS-1237265, CNS-1518878, CNS-1518918 CNS-1540066 and CNS-1602399.

References

[1] G. Aceto, A. Botta, A. Pescapè, N. Feamster, M. F. Awan, T. Ahmad, and S. Qaisar. Monitoring Internet censorship with UBICA. In *International Workshop on Traffic Monitoring and Analysis*. Springer, 2015.

[2] Alexa Top Sites. http://www.alexa.com/topsites.

[3] C. Anderson. Dimming the Internet: Detecting throttling as a mechanism of censorship in Iran. *arXiv preprint arXiv:1306.4361*, 2013.

[4] G. Angioni. PokerStars, Full Tilt, and 888Poker Blacklisted in Latvia, August 2014. https://www.pokernews.com/news/2014/08/pokerstars-full-tilt-and-888poker-blacklisted-in-latvia-18971.htm.

[5] Anonymous. GreatFire.org. https://en.greatfire.org/.

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

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

[8] BBC's website is being blocked across China. http://www.bbc.com/news/world-asia-china-29628356, October 2014.

[9] The Belmont Report - Ethical Principles and Guidelines for the protection of human subjects of research. http://ohsr.od.nih.gov/guidelines/belmont.html.

[10] D. Bolton. Putlocker Blocked in the UK by Internet Service Providers after High Court Order, May 2016. https://goo.gl/s8Hb43.

[11] CAIDA. Archipelago (Ark) Measurement Infrastructure. http://www.caida.org/projects/ark/.

[12] A. Chaabane, T. Chen, M. Cunche, E. D. Cristofaro, A. Friedman, and M. A. Kaafar. Censorship in the wild: Analyzing Internet filtering in Syria. In *Internet Measurement Conference (IMC)*. ACM, 2014.

[13] W. Chen, Y. Huang, B. F. Ribeiro, K. Suh, H. Zhang, E. de Souza e Silva, J. Kurose, and D. Towsley. Exploiting the IPID Field to Infer Network Path and End-System Characteristics. In *Passive and Active Network Measurement (PAM)*. Springer, 2005.

[14] D. Cicalese, D. Z. Joumblatt, D. Rossi, M. O. Buob, J. Aug, and T. Friedman. Latency-based anycast geolocation: Algorithms, software, and data sets. *IEEE Journal on Selected Areas in Communications*, 34(6):1889–1903, June 2016.

[15] Citizen Lab. Block Test List. https://github.com/citizenlab/test-lists.

[16] R. Clayton, S. J. Murdoch, and R. N. M. Watson. Ignoring the Great Firewall of China. In *Privacy Enhancing Technologies (PETS)*, Cambridge, England, 2006. Springer.

[17] A. Dainotti, C. Squarcella, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, and A. Pescapé. Analysis of country-wide Internet outages caused by censorship. In *Internet Measurement Conference (IMC)*. ACM, 2011.

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

[19] D. Dittrich and E. Kenneally. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. Technical report, U.S. Department of Homeland Security, Aug 2012.

[20] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-Widfe Scanning and its Security Applications. In *USENIX Security Symposium*, 2013.

[21] 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 (IMC)*. ACM, 2015.

[22] 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 Measurements Conference (PAM)*. Springer, 2014.

[23] R. Ensafi, J. C. Park, D. Kapur, and J. R. Crandall. Idle port scanning and non-interference analysis of network protocol stacks using model checking. In *USENIX Security Symposium*, 2010.

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

[25] D. Fifield and L. Tsai. Censors' Delay in Blocking Circumvention Proxies. In *Free and Open Communications on the Internet (FOCI)*. USENIX, 2016.

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

[27] B. Huffaker, M. Fomenkov, and k. claffy. Geocompare: a comparison of public and commercial geolocation databases - Technical Report . Technical report, Cooperative Association for Internet Data Analysis (CAIDA), May 2011.

[28] ICLab. Iclab: a censorship measurement platform. https://iclab.org/, 2014.

[29] B. Jones, N. Feamster, V. Paxson, N. Weaver, and M. Allman. Detecting DNS root manipulation. In *Passive and Active Measurement (PAM)*. Springer, 2016.

[30] B. Jones, T.-W. Lee, N. Feamster, and P. Gill. Automated detection and fingerprinting of censorship block pages. In *Internet Measurement Conference (IMC)*. ACM, 2014.

[31] J. Jung, V. Paxson, A. W. Berger, and H. Balakrishnan. Fast Portscan Detection Using Sequential Hypothesis Testing. In *IEEE Symposium on Security and Privacy (S&P)*, 2004.

[32] S. Khattak, D. Fifield, S. Afroz, M. Javed, S. Sundaresan, V. Paxson, S. J. Murdoch, and D. McCoy. Do you see what I see? differential treatment of anonymous users. In *Network and Distributed System Security Symposium (NDSS)*, 2016.

[33] S. Khattak, M. Javed, P. D. Anderson, and V. Paxson. Towards illuminating a censorship monitor's model to facilitate evasion. In *Free and Open Communications on the Internet (FOCI)*. USENIX, 2013.

[34] G. Lowe, P. Winters, and M. L. Marcus. The great DNS wall of China. Technical report, New York University, 2007.

[35] MaxMind. https://www.maxmind.com/.

[36] S. Mitchell and B. Collins. Porn blocking: What the Big Four ISPs Actually Did, August 2015. http://www.alphr.com/networking/20643/porn-blocking-what-the-big-four-isps-actually-did.

[37] D. Moore, G. Voelker, and S. Savage. Inferring Internet Denial-of-Service Activity. In *USENIX Security Symposium*, 2001.

[38] Z. Nabi. The anatomy of web censorship in Pakistan. In *Free and Open Communications on the Internet*. USENIX, 2013.

[39] Q. Northon. China Blocks LiveJournal, March 2007. https://www.wired.com/2007/03/china-blocks-livejournal/.

[40] ooniprobe: a network interference detection tool. https://github.com/thetorproject/ooni-probe.

[41] OpenNet Initiative. https://opennet.net/.

[42] 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*, pages 315–326. IEEE, 2010.

[43] Z. Qian, Z. M. Mao, Y. Xie, and F. Yu. Investigation of triangular spamming: A stealthy and efficient spamming technique. In *IEEE Symposium on Security and Privacy (S&P)*, 2010.
[44] M. Salganik. Bit by Bit: Social Research for the Digital Age, 2016. http://www.bitbybitbook.com/.
[45] Sam Burnett and Nick Feamster. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *ACM SIGCOMM*, 2015.
[46] A. Sfakianakis, E. Athanasopoulos, and S. Ioannidis. CensMon: A web censorship monitor. In *Free and Open Communications on the Internet*. USENIX, 2011.
[47] A. Spence. Russia Accuses Latvia of "blatant censorship" after Sputnik News Site is Shut Down, Mar. 2016. https://goo.gl/wIUUCr.
[48] The Tor Project. OONI: Open observatory of network interference. https://ooni.torproject.org/, 2014.
[49] Transmission Control Protocol. RFC 793, Sept. 1981.
[50] M. C. Tschantz, S. Afroz, Anonymous, and V. Paxson. SoK: Towards grounding censorship circumvention in empiricism. In *IEEE Symposium on Security and Privacy (S&P)*, 2016.
[51] G. Tuysuz and I. Watson. Turkey blocks youtube days after twitter crackdown. http://www.cnn.com/2014/03/27/world/europe/turkey-youtube-blocked/, Mar. 2014.
[52] S. Wilson. The Logic of Russian Internet Censorship, Mar. 2014. https://www.washingtonpost.com/news/monkey-cag⟩e/wp/2014/03/16/the-logic-of-russian-internet-censorship/.
[53] P. Winter and S. Lindskog. How the Great Firewall of China is blocking Tor. In *Free and Open Communications on the Internet (FOCI)*. USENIX, 2012.
[54] X. Xu, Z. M. Mao, and J. A. Halderman. Internet censorship in China: Where does the filtering occur? In *Passive and Active Measurement Conference (PAM)*. Springer, 2011.
[55] X. Zhang, J. Knockel, and J. R. Crandall. Original syn: Finding machines hidden behind firewalls. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 720–728. IEEE, 2015.
[56] J. Zittrain and B. G. Edelman. Internet filtering in China. *IEEE Internet Computing*, 7(2):70–77, Mar. 2003.

APPENDIX

Below are the definitions for website classes as specified by the CLBL [15]:

Class	**Definition**
ANON	Anonymizers and censorship circumvention
EMAIL	Free email
FEXP	Freedom of expression and media freedom
FREL	Foreign relations and military
HATE	Hate speech
HOST	Web hosting services
HUMR	Human rights
MILX	Militants extremists and separatists
MINF	Minority faiths
P2P	Peer-to-peer file sharing
POLR	Political reform
PORN	Pornography
REL	Religious conversion, commentary and criticism
USMIL	US government-run military website
VOIP	Voice over Internet Protocol (VoIP)