Characterizing Web Censorship Worldwide: Another Look at the OpenNet Initiative Data

PHILLIPA GILL, Stony Brook University
MASASHI CRETE-NISHIHATA, The Citizen Lab
JAKUB DALEK, The Citizen Lab
SHARON GOLDBERG, Boston University
ADAM SENFT, The Citizen Lab
GREG WISEMAN, The Citizen Lab

In this study, we take another look at five years of Web censorship data gathered by the OpenNet Initiative in 77 countries using user-based testing with locally-relevant content. Prior to our work, this data had been analyzed with little automation, focusing on what content had been blocked, rather than how blocking was carried out. In this study, we use more rigorous automation to obtain a longitudinal, global view of the technical means used for Web censorship. We also identify blocking that had been missed in prior analyses. Our results point to considerable variability in the technologies used for Web censorship, across countries, time, types of content and even across ISPs in the same country. In addition to characterizing Web censorship in countries that, thus far, have eluded technical analysis, we also discuss the implications of our observations on the design of future network measurement platforms and circumvention technologies.

KEYWORDS. Censorship, Internet filtering, measurement.

Categories and Subject Descriptors: C.2.3 [**Computer-Communication Networks**]: Network Operations

General Terms: Measurement, Security

Additional Key Words and Phrases: Censorship, network measurement

ACM Reference Format:
Phillipa Gill, Masashi Crete-Nishihata, Jakub Dalek, Sharon Goldberg, Adam Senft, and Greg Wiseman. Under Submission. Characterizing Web Censorship Worldwide: Another Look at the OpenNet Initiative Data *ACM Trans. Web* 0, 0, Article 0 (2014), 29 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Politically motivated, country-wide Internet outages in Libya [Cowie 2011b], Egypt [Cowie 2011a], and Syria [Cowie 2012] highlight the role of the Internet as a tool for governments to exert control over their populations. While dramatic outages at this scale are observable by the outside world, there is less visibility into more subtle forms of Internet control that have become the status quo in many networks worldwide. In reaction to this, the OpenNet Initiative (ONI) has been tracking censorship of Web content worldwide since 2003, performing thousands of user-based tests in tens of countries and hundreds of distinct ISPs. The results of the ONI tests have thus

Author's addresses: P. Gill, Department of Computer Science, Stony Brook University; M. Crete-Nishihata, J. Dalek, A. Senft and G. Wiseman, The Citizen Lab, Munk School of Global Affairs, University of Toronto; S. Goldberg, Department of Computer Science, Boston University. P. Gill current address, Rm. 1418 Computer Science, Stony Brook University, Stony Brook, NY, 11794-4400.

DOI: http://dx.doi.org/10.1145/0000000.0000000

far been presented in qualitative journalistic style [Deibert et al. 2008; 2010; OpenNet Initiative 2014], with a focus on understanding *what* content is being censored in each country or ISP, as well as the political and legal context surrounding censorship of the content. The ONI reports on what content is being blocked have recently been complemented by other efforts, including [Zittrain et al. 2014; Google 2014].

Meanwhile, there have been numerous measurement studies that have looked into *how* censorship is executed [Zittrain and Edelman 2003; Clayton et al. 2006; Xu et al. 2011; Anonymous 2012; Murdoch and Anderson 2008; Dornseif 2004; Clayton 2006; ICANN Security and Stability Advisory Committee (SSAC) 2012; Weaver et al. 2009; Wolfgarten 2006; Sfakianakis et al. 2011; Verkamp and Gupta 2012]. While these studies contain the technical details that the ONI reports usually lack, they tend to focus on a single blocking technology [Clayton 2006; Lowe et al. 2007; Clayton et al. 2006; ICANN Security and Stability Advisory Committee (SSAC) 2012; Weaver et al. 2009; Xu et al. 2011], a single country (usually China [Anonymous 2012; Crandall et al. 2007; Park and Crandall 2010; Xu et al. 2011; Fallows 2008; Clayton et al. 2006; Lowe et al. 2007; Wolfgarten 2006; Dornseif 2004], but sometimes others [Clayton 2006; Dornseif 2004]) or a single set of Websites (*e.g.,* Google transparency [Google 2014]). The few technical studies [Sfakianakis et al. 2011; Verkamp and Gupta 2012] that considered multiple countries and technologies rely on limited vantage points and meager information about what content is likely to be censored in a given region.

In this work, we consider data collected by the ONI over the last five years to understand how the technologies used for *Web censorship* vary across time, countries, and ISPs. These data are unique in terms of performing measurements on residential ISPs in many countries that are known to perform Web censorship. Further, through discussions with the individuals performing the measurements, the ONI group has rich contextual information about the political situation and nuances within each country. However, we must also contend with the fact that the ONI measurement client was *not* designed to expose the underlying technical means used to execute censorship. Thus, we design a methodology to infer this information from application log data by correlating results over time and across tests (§3). To enable future studies and ensure the repeatability of our results, we have also released the ONI data used in our study to the community[1].

1.1. Results and Implications

We analyze tests of over 90K distinct URLs gathered in 77 countries and 286 distinct ISPs, and uncover a number of global trends that have an important impact on the design of future network measurement platforms, security, and circumvention technologies. We also present new observations about the technologies used for Internet filtering in many under-studied countries, including Azerbaijan, Burma, Iran, the United Arab Emirates, Kyrgyzstan, South Korea, Vietnam, and Yemen.

- *Censorship technology varies widely.* The strongest trend we encounter is the significant variation in blocking technologies used, both across and within countries, an observation (also made by others) that highlights the need for ongoing user-based measurement that can track censorship over time [Filasto and Appelbaum 2012; Sfakianakis et al. 2011].
- *Academic networks do not provide a representative view of censorship.* We provide the first hard evidence that traditional platforms like PlanetLab are not up to the task of measuring censorship. Specifically, PlanetLab lacks vantage points in countries that

[1]See this URL: http://www.cs.stonybrook.edu/~phillipa/papers/ONIAnaly.html

perform the most extensive Web censorship. Further, we observe academic networks blocking 26% less content, on average, than user-based testing (§5.3).

- *Local context impacts observed censorship.* We also point out the importance of developing appropriate lists of content to trigger censorship mechanisms. Our results show that locally-relevant content can elicit up to 3-5X more blocking in authoritarian countries (§5.4).
- *Countries are selectively transparent about censorship.* We observe and highlight cases of "stealthy" censorship by governments. Specifically, in Yemen the government openly blocks social and Internet-related content (e.g., pornography and proxy services) but also uses TCP reset packets to censor political content that the government is not open about censoring (§4.2). This observation highlights the importance of designing measurement techniques that can identify stealthy forms of censorship that are not obvious to end users.
- *Implications for circumvention.* Our results also indicate that a "one-size-fits all" approach to censorship circumvention is unlikely to succeed. We find that the set of blocking technologies used in China are *not* representative of how content is blocked in the rest of the world. For example, we show that Middle East and North African (MENA) countries most commonly accomplish censorship by delivering explicit block pages (§5.1), instead of the more stealthy approaches (*e.g.,* TCP resets [Crandall et al. 2007; Fallows 2008; Clayton et al. 2006]) used in China.
  We distinguish two types of censorship architectures: centralized and decentralized, which can impact the effectiveness of circumvention technology in a country. Unlike China, where ISPs are known to deploy blocking technologies independently, we give evidence for centralized censorship in Iran (§4.5) and South Korea (§4.8). We also give evidence for decentralized censorship in countries other than China, including Vietnam (§4.7), Kyrgyzstan (§4.8) and UAE (§4.3).
- *Extending ONI results.* We also make a number of new observations that have been overlooked by earlier ONI reports. We find potential surveillance in South Korea (§4.8), investigate blocking of the .il TLD in the United Arab Emirates (§4.3), and provide new analyses of blocking in Kyrgyzstan (§4.8) and Vietnam (4.7).

Organization. We discuss the ONI dataset in §2, and overview blocking technologies and how we identify them in §3. We then use country-specific case studies to illustrate specific observations about how censorship is implemented around the world §4. Finally, in §5 we consider trends across countries, and discuss their implications on future censorship measurement research.

2. THE OPENNET INITATIVE DATASET

We now overview the ONI's data collection procedure and discuss the ONI dataset.

2.1. Data collection

Data was collected by performing synchronized HTTP requests in both a *lab* and *field* location. A field location is a location where Web censorship is suspected. The lab is located at the University of Toronto, a site that does not censor the type of content tested by the measurement software. Field locations were obtained via dialog with regional groups of local researchers, advocates and practitioners that the ONI helped form and support.

Tests were conducted on URL lists that consisted of *globally sensitive URLs*, used in all regions, and *locally sensitive URLs* that were specifically chosen for the region under study. URL lists contained URLs that the ONI classified into the four broad themes shown in Table I. Local groups helped the ONI curate local URL lists, as well

as pinpoint important periods for testing (*e.g.,* protests, elections) and provided context for interpreting test results. Since the local lists were manually curated, their length varied from fewer than 10 URLs to hundreds of URLs. The global list contains approximately 1,500 URLs and was modified over the course of the measurements as relevance of sites changed. While the evolution of the URL lists over time is natural, given that sensitive issues and Web site popularity evolve significantly over a five year period, it also presents challenges for our post-hoc longitudinal data analysis. We control for this variation using a number of techniques, including temporal clustering and a variation on the Jaccard similarity metric; §4.1, §4.3 and §4.5 have examples of how we did this.

Table I. Samples from global URL lists.

Political: opposition to government, human rights, freedom of expression, minority rights, religious movements				
ijm.org	efsha.co.uk	islamicity.com	iico.org	acdi-cida.gc.ca
shia.org	hrw.org	martus.org	imf.org	law-lib.utoronto.ca
Social: sexuality, gambling, and illegal drugs and alcohol, topics perceived as offensive, socially-sensitive topics				
mate1.com	marijuana.nl	aidsonline.com	budweiser.com	ageofempires3.com
gay.com	gayhealth.com	drugsense.org	survive.org.uk	agentprovocateur.com
Internet Tools: e-mail, Internet hosting, search, translation, VoIP, circumvention methods				
dogpile.com	wordpress.com	ultimate-anonymity.com	ask.com	translate.google.com
tinyurl.com	securenym.net	proxytools.sourceforge.net	piolet.com	groups.google.com
Conflict: conflicts, border disputes, separatist movements, militant groups				
instituteforcounterterrorism.org		jdl.org	geocities.com/jklf_uk_europe/fpage.html	
aleph.to	ehj-navarre.org	kurtuluscephesi.com	fisWeb.org	arabrenewal.com

2.2. Ethics and client-based testing

The software client was distributed to researchers and volunteers in countries of interest. The client-based testing involved accessing a large number of potentially sensitive Web sites in quick succession, a prospect that may pose security concerns for testers depending on the country being tested. Before users engaged in testing, an informed consent meeting was held, where the risks posed by the research were explained in plain terms, and consented to by the user. Moreover, the decision of where to test was driven by the ONI's concerns for safety and practicality. Often countries with the potential for interesting data were considered too dangerous for user-based testing *e.g.,* during Syria's conflict, or in countries like Cuba and North Korea. Since the goal of the measurements was to reproduce the experience of the average Internet user in the country, the software did not use censorship-circumvention or anonymity technology.

2.3. Dataset contents & preprocessing

The set of lab and field results collected by a user at a given point in time for a set of URLs is called a *run*. We use the term *test* to refer to an individual test of a single URL within a run. For each run, the dataset logged the client's ISP,[2] for each test in the run the following was logged: URL, timestamp, server IP address, and received HTTP headers and body. HTTP replies were parsed by the Python `urllib2` library, and any HTTP errors detected by the library were also logged. Each test often had metadata annotation added by ONI researchers (*e.g.,* BLOCKED (the site was deliberately blocked in the field, because an explicit blockpage was found), ACCESSIBLE (the site was accessible in the field), INACCESSIBLE (the site was not accessible in

[2]The client's ISP was logged by mapping their IP address to an AS using [Team Cymru IP to ASN Lookup v1.0 2013]; for privacy reasons, the IP address was not logged.

Table II. Countries with more than 50 runs.

Country (Section)	Runs	Country (Section)	Runs	Country (Section)	Runs
Kyrgyzstan (kg) (§4.1)	738	China (cn) (§5.2)	415	Thailand (th)	216
Yemen (ye) (§4.2)	146	Malaysia (my)	138	Belarus (by)	134
Indonesia (id) (§4.6)	129	Azerbaijan (az) (§5.2)	117	Korea (kr) (§4.8)	113
Ukraine (ua)	112	Russia (ru)	102	Italy (it)	100
Nepal (np)	90	Burma (mm) (§4.4)	86	Egypt (eg)	84
Georgia (ge)	79	Vietnam (vn) (§4.7)	74	Israel (il)	73
Iran (ir) (§4.5)	69	Turkey (tr)	66	UAE (ae) (§4.3)	63
		Philippines (ph)	52		

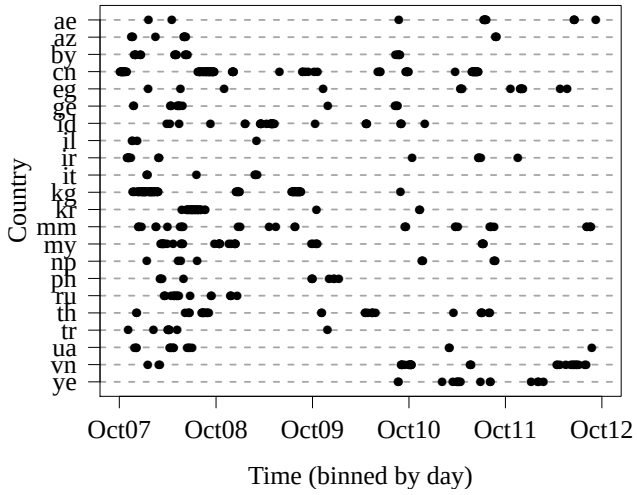


Fig. 1. Overview of runs over time for each country. We only show countries that had more than 50 runs in the ONI dataset.

the field, but no blockpage was observed)) determined using regular expressions developed via manual inspection of block pages. Most previous analysis of the ONI data relied on this manual analysis; the goal of our automated analysis (§3) was to take a closer look at the data in order to identify the technologies used for blocking. Traceroutes and packet captures were collected only for a small fraction of tests. We rely on application-log data, IP addresses, and HTTP captures for our automated analysis (see §3).

The opportunistic manner in which the dataset was collected also creates challenges for longitudinal research. First, the ISPs tested varied over time. Next, measurements were conducted at periods that were deemed interesting by the ONI (*e.g.,* during political, social, or publication events), rather than at regular intervals. There was a tendency to try to capture censorship around sensitive times for countries (*e.g.,* the anniversary of the 1989 Tiananmen square protests in China and the 2009 Elections in Iran). Figure 1 overviews the temporal distribution of testing in countries with at least 50 runs (Table II). Our analysis therefore takes care to control for these variations, by clustering tests according to time period and ISP; see §4.1 for an example of how we did this.

We preprocessed the ONI dataset as follows:

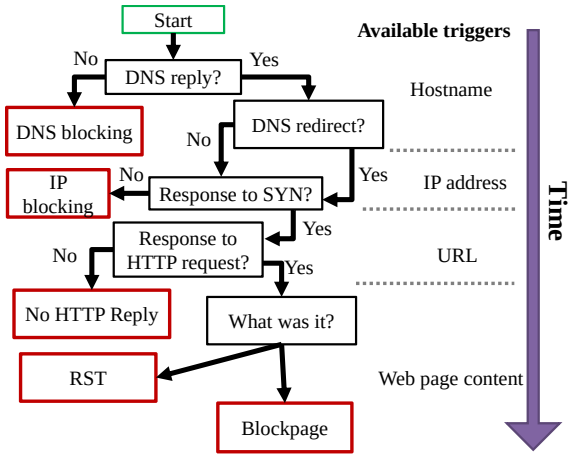


Fig. 2. Flow chart of blocking methods.

(1) Each URL list contained the innocuous train hobbyist Web site abpr2.railfan.net as a control[3]. Given the assumption that the train-hobbyist Web site should not be filtered anywhere, we discarded runs that were unable to access this Web site from either the field or the lab locations. We also discard runs that did not complete an entire URL list, or had less than 100 tests.

(2) Since our main focus here is on state-imposed Internet filtering at ISPs, we separated out runs that were performed at (a) academic networks and (b) enterprise networks and characterized them separately (see §5.3). Moreover, we also remove cases of apparent "reverse blocking" , *i.e.,* where the owner of the content restricts those who may access it. This type of blocking was prevalent on `.mil` domains that were inaccessible from numerous countries, including those with very little censorship (*e.g.,* Italy, France). We omit these tests from our characterization.

Validation. Throughout our analysis we aimed to validate our observations where ever possible. In cases where packet traces were available, we leveraged them to validate our interpretation of the application-layer log data. Further, we use manual analysis to validate unexpected results. Indeed, our methodology allowed us to find blockpages and other events (*e.g.,* §4.8) that were previously missed during the ONI's first pass over the data.

3. INFERRING HOW BLOCKING IS EXECUTED

Our post-hoc analysis of the ONI dataset seeks to shed light on the history of the technologies used for web censorship worldwide. Because the ONI dataset was not specifically collected with this objective in mind, we develop a methodology for extracting this information from the *application-log* data recorded by the ONI. (Refer back to §2.3 for a description of the dataset.)

Our methodology is based on a set of six criteria, summarized in Table III, that we use to classify each test in the ONI dataset; if a test matched one or more criteria in

[3]ONI chose this URL as a control in 2007, and it remained so throughout the project. While future work should identify a more robust set of control URLs, given that our analysis is post-hoc, we unfortunately cannot not go back in time to change this experimental design decision. Instead, we use a data analysis methodology that includes several additional criteria to distinguish between blocking and transient errors; see Sections 3.4 3.2.1.

Table III. Criteria used to identify different blocking methods.

Method	Criteria	Validation
No DNS Result	IP address cell is blank in field vantage point test, but not blank from lab vantage point test. This occurs ≥ 3 in a week for the URL in the ISP from which it was tested.	
DNS Redir.	IP address cell in field vantage point is associated with a reserved private IP per Table VIII **OR** at least 32 lab ASNs according to the methodology detailed in Appendix A.1.	Appendix A.2
No HTTP Response	HTTP response observed in lab vantage point test but not in field vantage point test; in field, URL resolved to routable IP address and no RST error was logged. This occurs ≥ 3 times in a week for the URL in the field ISP from which it was tested.	
LCRST	HTTP status < 400 (success) logged in lab vantage point test. In the field test, the client application logs a RST error, and the Web page is not observed to be successfully delivered.	Appendix A.3
RST	LCRST occurs at ≥ 3 times in a week for the URL in the field ISP from which it was tested.	Appendix A.3
Block page	HTTP content returned in the field matches a regular expression or block page template.	Manual.

Table III, we say that a test is "blocked". It is important to note that blocking techniques can be applied inconsistently and a single accessible test does not necessarily mean that a site is not blocked (*e.g.,* Yemen's censorship ceases when they run out of licenses for their filtering product [ONI Research Profile: Yemen 2009]). Thus, we instead focus on correlating instances where the content was not successfully retrieved to identify cases of blocking.

Once we have identified cases of blocking, we carefully correlate tests classified as "blocked" across time and ISP, in order to shed light on the filtering behaviors in various countries; descriptions of how we do this in various country-wide case studies are in Section 4. In what follows, we explain how we chose each criteria in Table III, and the level of confidence we have in each of them. Whenever possible, we validated the criteria in Table III using packet captures and external information. We discuss validation in Appendix A.2-A.3.

3.1. Overview.

The ONI dataset tracks Web censorship only. We therefore focus on technologies that are known to disrupt an HTTP transaction, summarized in Figure 2. (The ONI dataset lacks sufficient information to allow us to identify IP blocking (as has been measured by [Verkamp and Gupta 2012; Sfakianakis et al. 2011]), but we include them in Figure 2 for completeness.) Figure 2 also shows when different triggers become available for blocking.

3.2. Blocking via DNS

Blocking can occur at the very beginning of a Web transaction — when a client uses DNS to resolve a domain name to an IP address [Zittrain and Edelman 2003; ICANN Security and Stability Advisory Committee (SSAC) 2012; Anonymous 2012; Clayton 2006; Zmijewski 2010]. While the ONI dataset lacks DNS packet captures, it does log the IP address that each tested URL resolves to. We now discuss how we use this information to identify blocking that is likely to be accomplished via DNS, and how we distinguish suspected blocking from transient errors. Our methodology distinguishes two types of tampering with DNS: unresolved domain names, and DNS redirects.

3.2.1. Unresolved domain names

A domain name may fail to resolve to an IP address if the client's DNS query was answered with an error message (*e.g.,* REFUSED, SERVFAIL, NOTIMPL, FORMERR),

or no response at all; these sorts of responses can cause the client's operating system to remove a recursive resolver from its list of name servers [ICANN Security and Stability Advisory Committee (SSAC) 2012], and thus prevent future blocking. As such, the most likely explanation for an unresolved domain name (apart from a transient error) is the injection of NXDOMAIN message by a middlebox or the resolver itself.

To distinguish transient network failures from deliberate blocking in the ONI dataset, we classify unresolved domain names as "blocking" for a given test only if: (1) No IP address was logged from the field vantage point, and (2) the IP address was correctly logged from the lab vantage point, and (3) three observations meeting criteria (1)-(2) occur within a one week period for a given URL in the ISP where the tests were conducted.

3.2.2. DNS redirection

In 2003, Edelman and Zittrain [Zittrain and Edelman 2003] noticed that Chinese filters would sometimes use DNS to redirect the client to an unintended IP address. Since then, many works have discussed DNS redirection (usually in China) [Zittrain and Edelman 2003; ICANN Security and Stability Advisory Committee (SSAC) 2012; Anonymous 2012; Clayton 2006; Zmijewski 2010; Lowe et al. 2007].

To detect DNS redirection in the ONI dataset, we can check for discrepancies between the IP address to which a URL resolves in the field versus its resolution in the lab. However, this approach is complicated by the prevalence of DNS-based load balancing, where different resolvers will (legitimately) resolve the same domain to different IP address in order to reduce latency and spread traffic across servers. We thus use two approaches to identify DNS redirections:

Redirection to a private-reserved IP address. Certain types of DNS redirects prevent the client's HTTP request from leaving its local machine or network: redirects to the loopback block 127.0.0.0/8, local identification addresses in 0.0.0.0/8, link-local addresses in 169.254.0.0/16, or private IPv4 address space[4] in 10.0.0.0/8 or 192.168.0.0/16. Therefore, a test was classified as a redirect to a private-reserved IP when the field IP corresponded to one of the above addresses. We refer to IP addresses meeting these criteria as 'not routable'.

Redirection to a non-private address. Our methodology for detecting redirection to a routable IP address is complicated by the need to differentiate between (a) blocking and (b) DNS-based load balancing or CDNs. To do this, we exploit the idea that a set of URLs resolving to IP addresses hosted at many different ASNs in the lab are very unlikely to be hosted on a single IP address in the field. The methodology is detailed in Appendix A.1, and validated in Appendix A.2.

While DNS redirects to private IPs typically lead to blocking because traffic fails to leave the user's local machine or network, redirects to a routable IP addresses can lead to blocking because: (1) traffic is redirected towards a middlebox that will take filtering action (*e.g.,* drop the HTTP request, deliver a blockpage), (2) traffic is redirected to an IP address that corresponding to an unrelated (but blocked) Web site to trigger IP-based blocking mechanisms (*e.g.,* routers that filter on an address or prefix).

We discuss DNS-based blocking in our case studies (§4), and global trends in DNS redirection in §5.2.

3.3. Blockpages

Blockpages explicitly indicate Internet filtering (Figure 3). Some blockpages indicate why the content was blocked (including how the content was categorized), the com-

[4]Note that there is some ambiguity in whether redirects to private addresses in 10.0.0.0/8 and 192.168.0.0/16 end up at a middlebox or just get dropped; this depends on how the ISP configures its internal network.

Fig. 3. Example of a Qatari blockpage.

mercial product used to implement the filtering, or offer contact information should a user wish to have the filtering of the Web site reassessed. Of the blocking techniques we consider here, blockpages are the only one that explicitly indicates to the *user* that blocking has taken place. In contrast, the other methods are indistinguishable from network failures by the average user; thus we refer to them as less "transparent", or "stealthy".

Detecting blockpages in censorship measurements is surprisingly challenging. A naive approach would be to simply compare HTML content between the field and the lab, but factors such as content localization and dynamic Web content cause this to be highly inaccurate. As a result, ONI automatically flags tests with content variation between lab and field for manual analysis. Once a blockpage is identified with manual analysis, a regular expression is created and applied to the set of historical results in the database. While this technique is highly robust to false positives, it is prone to false negatives as there is bias to investigate ISPs with a higher number of tests. Therefore, the results we report should should be considered a lower bound on the usage of blockpages.

We distinguish two methods for delivering blockpages:

Blockpages with DNS redirects. When a DNS redirection (per §3.2.2) is observed in conjunction with a blockpage for a given test, it suggests that a DNS redirect is used to send the HTTP query to a middlebox that serves blockpages. Traffic interception is accomplished at the DNS phase, where the hostname is used as a trigger to identify content to be blocked. Meanwhile, no special effort is required to intercept the HTTP request, which is sent directly to the IP address of the middlebox serving blockpages.

Blockpages without DNS redirects. On the other hand, when a blockpage is observed in the absence of DNS redirection, we can infer that a client's Web traffic was intercepted directly (*e.g.,* by a Web proxy or other middlebox). While Appendix A.1 describes our methodology for inferring when a DNS redirect *does* occur, we use a different methodology to infer when a DNS redirect *does not* occur: specifically, we require that the tested URL resolves to the *exact same* IP addresses from both the field and lab vantage points. This conservative methodology allows us to identify blockpages delivered via traffic interception (*i.e.,* without DNS redirects) with high confidence.

We discuss global trends in blockpage delivery in §5.1. Some examples of the different ways in which blockpages can be delivered are in §4.3 and §4.2.

3.4. No HTTP reply

Even if the client resolves its URL to a routable IP address, the HTTP request may be disrupted (*e.g.,* by dropping the HTTP GET request or the response).

The lack of packet traces for most tests in the ONI dataset means that we cannot diagnose a root cause when there is no HTTP reply, unless there is evidence of DNS redirection. Specifically, we cannot say if the connection was disrupted during the TCP handshake, or after the HTTP request was made (*e.g.,* as is done by [Verkamp and Gupta 2012]). Classifying a lack of HTTP reply as "blocking" is therefore the most challenging part of our data analysis. To limit false positives, we say a lack of HTTP reply is suspicious only if (1) in the lab, an HTTP reply was observed and (2) in field, the URL did resolve to a *routable* IP address, and (3) in the field, there was no application-log error indicating that a TCP RST had been received (see discussion in §3.5), and (4) three tests meeting criteria (1),(2), and (3) occur within a one week period for the same URL on a given ISP.

There are several ways to disrupt an HTTP request:

DNS redirect to middlebox. Middleboxes may drop HTTP GETs, rather than delivering a blockpage. To infer when this occurs, we look for tests where there is both (a) no HTTP response, and (b) a DNS redirection was inferred according the methodology in §3.2.2.

The ONI dataset does *not* allow us to distinguish between the following two other blocking technologies that could cause a lack of HTTP response:

Interception by on-path middlebox. As discussed in §3.3, an HTTP GET request may be intercepted by a middlebox; however, instead of returning a blockpage, the middlebox can simply drop the request.

IP blocking. A lack of HTTP response may be caused by routing-based blocking that filters traffic to specific IP addresses. Since IP blocking only requires a match on header fields (rather than packet payload), it can be implemented in a variety of ways, including at an in-path middlebox, at routers themselves via access control lists, and even by BGP prefix hijacking (*e.g.,* when Pakistan Telecom hijacked YouTube's prefix in 2008 [Rensys Blog] and there is some evidence that this approach is also being used in China [Anderson 2012]). While IP-based blocking can be distinguished from HTTP-based blocking by checking if no response was received after the TCP SYN packet was sent (Figure 2), the ONI dataset lacks the traces required to do this.

We discuss the techniques we used to distinguish unrequited HTTP responses due to transient errors from those that result from deliberate censorship in §4; a particularly interesting example is our case study of Kyrgyzstan §4.1.

3.5. TCP resets

Much has been written about blocking via TCP resets, especially in China [Clayton et al. 2006; Weaver et al. 2009; Crandall et al. 2007; Xu et al. 2011; Fallows 2008; Wolfgarten 2006]. Blocking with TCP resets is accomplished by an on-path middlebox that observes network traffic, and injects spoofed TCP resets to the client and server. These RSTs race the legitimate traffic, and terminate the connection. However, injected TCP RSTs suffer from a race condition where they cannot match header values (*e.g.,* TCP sequence number) of the connection, and thus are easily detectable in raw packet traces [Weaver et al. 2009].

RSTs are typically associated with application-level triggers (detected by on-path middleboxes), which makes it among the most targeted of the blocking techniques we have discussed. In contrast to the DNS-based approaches discussed in §3.2, they can be used to block specific terms in a URL (*e.g.,* en.wikipedia.org/wiki/falun) or specific words on the webpage itself.

The ONI dataset logs application error conditions that indicate when a TCP RST was received. However, determining whether the TCP RST was indicative of blocking was complicated by the fact that many benign conditions (*e.g.,* Web browser implemen-

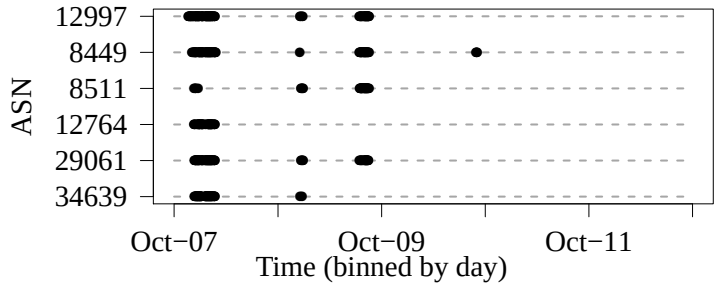


Fig. 4. Distribution of runs in Kyrgyzstan over time and across ISPs.

tations [Arlitt and Williamson 2005] and middleboxes [Weaver et al. 2009]) can cause a TCP connection to terminate via a RST packet rather than a FIN handshake. Given the lack of packet traces for most tests in our dataset, we used the following criteria to distinguish between "benign" and "suspicious" RST packets. An observed RST was as suspicious with "low confidence" (LCRST) if (1) the Web page was accessible in the lab (HTTP status 2xx or 3xx) but (2) in the field we saw (a) a Python error indicating a RST and (b) either no HTTP response or an HTTP status 4xx [Fielding et al. 1999]. Further, to rule out benign cases where the Web server uses RSTs to close TCP connections, we only consider tests where (3) the Web page is not observed being successfully delivered when a RST has occurred in *any test*. We classify a RST as suspicious with "high confidence" (RST) if (4) three tests meeting the low confidence criteria occur within a week for a given URL at a given ISP.

We validated these criteria using a number of tests in the ONI datasets that did record packet traces along with application log errors; see Appendix A.3. While it is well-known that China uses RSTs [Clayton et al. 2006], we also observe their use in other countries (*e.g.,* Yemen (§4.2)).

4. CASE STUDIES

In this section, we present a few representative case studies of censorship in a number of different countries studied by the ONI. We observe considerable variation, across countries, over time, and across ISPs in the *same country*, or even across different types of *content* that were tested in the same ISP. We can sometimes also infer that some countries implement censorship in a decentralized manner (*i.e.,* individual ISPs were tasked with implementing their own censorship), while other case studies give evidence for a centralized infrastructure. The next seven case studies take a deeper look into these two common themes.

4.1. Transient errors vs. deliberate censorship in Kyrgyzstan

We begin with analysis of censorship in Kyrgyzstan, where an enthusiastic network of regional testers contributed tens of thousands of tests to the ONI dataset. Before we get into detailed observations of blocking in Krygyzstan, we first present the methodology we used to distinguish blocking from transient failures, as well as to control for variations in tested URL lists and ISPs over time.

Figure 4 shows the distribution of measurements run in Kyrgyzstan over time and across ISPs. The plot clearly shows that tests are clustered into four time periods; tests run in a given country in a given time "cluster" are part of the same investigation (since ONI testing tends to center around investigations of specific events *e.g.,* elections). We therefore make the first-order assumption that the list of URLs being tested in a given time interval are stable *e.g.,* relating to the political event under consideration. We will take a more fine-grained look at the URL lists later on in this case study.

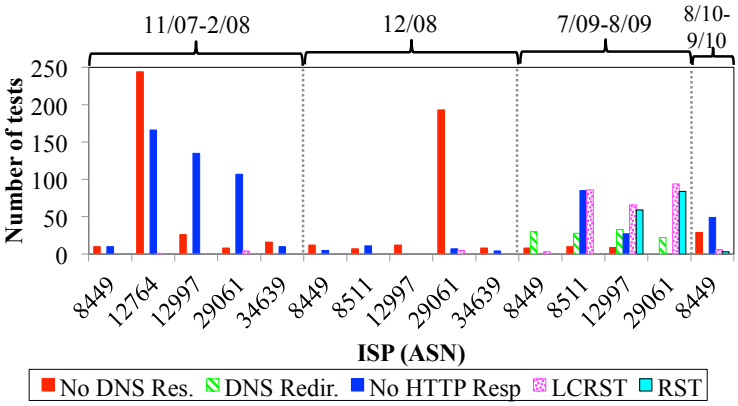
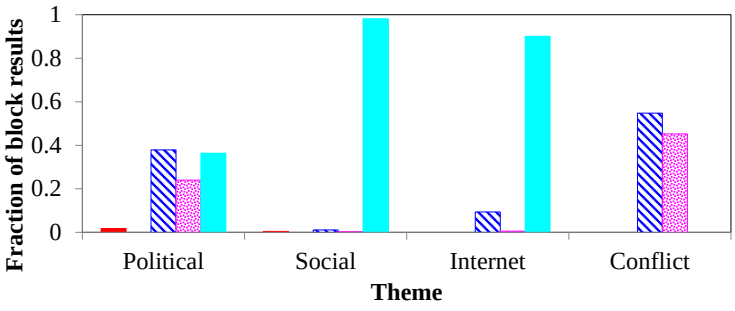


Fig. 5. Distribution of tests in Kyrgyzstan that we classified "blocked" according to the criteria in Table III. (We only show ISPs with $\geq$ 10 blocked URLs per time cluster in this figure.) From the user perspective, most of these "blocked" tests appear to be transient failures; we therefore take a deeper look at the data to distinguish between tests that were blocked due to transient errors, and those that resulted from deliberate censorship; see §4.1.

Working within the first-order assumption that URL lists are stable, Figure 5 breaks down the results from Figure 4 by grouping together all tests in all runs on a single ISP in a single investigation time period, and then classifies each test as "blocked" or not according to the criteria in Table III. For example, the first group of bars in Figure 5 indicates that 11 of the tests classified as blocking in AS 8449 between 11/2007-2/2008 were due to lack of DNS response (first row of Table III) and 9 of the tests classified as blocked were due to a lack of HTTP response (third row of Table III). Figure 5 indicates that all tests classified as "blocked" in Kyrgyzstan looked like transient failures from the perspective of the user; most were cases of "no HTTP response" or "no DNS resolutions". No blockpages explicitly provided evidence of blocking during any of the tests in Kyrgyzstan. Therefore, we need to take a more fine-grained look at the data in order to distinguish between tests that actually represented transient failures, and tests that are indicative of deliberate censorship (that was effectuated in a "stealthy" manner).

To do this, we compare all tests of a single URL during a single time period, and compare how often they were classified as "blocked" at each Kyrgyzstani ISP tested during that time period. This allows us to better distinguish between transient failures and 'stealthy' censorship techniques. For example, in AS 8449, a Russian news site (www.flb.ru), was classified as 'blocked' for only 1 out of the 113 tests during the 11/07-2/08 time period; this 'blocked' test is more likely the result of a transient failure. On the other hand, the same URL was blocked for 46 out of 53 tests at AS 12764, which is more suggestive of deliberate censorship.

Using this technique we can infer multiple interesting cases of deliberately censored URLs. For example, the critical oppositional website eurasia.org.ru was blocked (likely for political reasons) by AS 8511 using dropped HTTP requests, whereas AS 29061 and AS 12997 used RST packets to block the same URL. We also find evidence of deliberate blocking of a Kazakhstani political blog inkar.info at all ISPs except AS 8449, using either unrequited GET requests (AS 8511) or TCP resets (AS 29061, AS 12997). This is particularly interesting because this is a Kazakh opposition site, that is known to be blocked in Kazakhstan. The two mostly likely reasons for this site to be blocked in Kyrgyzstan are: (1) censorship leakage from an upstream Kazakh ISP, or (2) because



	Political	Social	Internet	Conflict
# blocked / # test	509 /18246	1083/10819	1460/15571	42/917

Fig. 6. Tests classified as "blocked" (per Table III) in the YemenNet ISP, broken down by the type of tested content. Stealthy techniques, such as blocking HTTP replies and sending reset packets, are used to block political and conflict-related content.

the site published sensitive opposition content relevant to Kyrgyzstan (since the two countries share a close political relationship).

These observations highlights three key points. First, running repeated tests for censorship is important to distinguish deliberate censorship from transient failures. Second, we can infer that decentralized censorship is occurring by observing that different techniques are used to block the same content; this implies that discussing censorship only at the national level does not give the full picture, since practices may vary even within a single country. Finally, our observation of blocking of a Kazakh blog raises questions of censorship leakage between neighboring countries.

4.2. Stealthy blocking in Yemen

The following case study of Yemen shows how a national ISP evades detection of certain censorship practices by effectuating censorship using stealthy techniques. In Yemen, the national ISP, YemenNet, was transparent about blocking some content (explicitly presenting blockpages to the user). However, YemenNet blocked other sites using 'stealthy' techniques that seem like transient failures to users. This discrepancy stems from national policy in Yemen, where the government is open about blocking content deemed inappropriate according to Sharia law; blocking political content, however, is against the constitution [ONI Research Profile: Yemen 2009].

ONI categorizes web sites as *social*, *Internet*, *political* and *conflict*; some examples are as shown in Table I. Figure 6 shows the fraction of each content type that was blocked using the techniques from Table III. Since the ONI only has visibility into a single Yemeni ISP (YemenNet) and testing was done on a fairly short timespan (2010-2011; *cf.,* Figure 1) we group all tests for the ISP together and break out the results by the type of URL tested and the type of blocking observed. Interestingly, blockpages are presented for 'social' and 'Internet' content, 'stealthy' techniques (*i.e.,* TCP RSTs, no HTTP reply) are used in the politics and conflict categories, which should not be blocked according to the constitution.

Taking an even closer look at the data, we find that sites blocked using 'stealthy' techniques include a number of sites in the .il TLD, including the Ethiopian embassy in Israel http://www.ethioemb.org.il, Israeli technology companies like http://www.gemini.co.il and www.lahat.co.il/ and an Israeli gossip and entertainment web-

Table IV. Fraction of blocking on ISPs in the UAE.

Interval	ISP	Fraction "Blocked"
12/07-2/08	Etisalat	0.14
12/07-2/08	Du	0.05
7/11	Etisalat	0.20
7/11	Du	0.13

site http://www.pnaiplus.co.il. Stealthy techniques were also used to block access to political sites (*e.g.,* www.al-masdar.com, www.alhadath-yemen.net). On the other hand, blockpages were shown for pornographic websites and censorship circumvention technologies (including http://psiphon.civisec.org, http://www.publicproxyservers.com and http://tor.eff.org).

These observations have technical implications for studies of web censorship; specifically, measurement studies need to design techniques that contend with censors that wish to deliberately evade detection.

4.3. Increasing censorship in United Arab Emirates

This case study, into two different United Arab Emirates (UAE) ISPs, Etisalat AS 5384, and Du AS 15802, shows how censorship changes over time. Specifically, a longitudinal look at the data shows how Du moves from providing unfettered Internet access to openly blocking certain web content.

In UAE, the Telecommunication Regulatory Authority openly acknowledges restrictions on certain Internet content, describing the national blocking infrastructure as a "solid wall that does not allow any irrelevant content to be displayed" [Zain 2008]. Blocking is therefore usually accomplished in a transparent manner, via blockpages delivered without DNS redirects. This suggests that UAE is using filtering Web-proxies, which is confirmed by reports that the national ISP, Etisalat, is using the SmartFilter proxy for this task [Noman b].

The Du ISP (AS 15802) was launched in 2007 to provide unfettered Internet access and encourage economic development in Dubai [Deibert et al. 2008]. To get a sense for how these two ISPs block content over time, we group together all tests from all runs at each ISP in a given investigation time period, and determine the fraction of tests classified as 'blocked' per the criteria in Table III. Results are shown in Table IV.

We see that Du begins to aggressively block content in the later time period, which is consistent with reports that Du began filtering traffic in 2008 [Zain 2008; Noman a]. Indeed, a closer look at the data indicates the even those tests categorized as 'blocking' at Du could have been the result of transient errors. Indeed, we have little evidence of deliberate blocking at Du during this time; the ONI dataset has at most 4 tests for each URL tested at Du during this time frame, and for most URLs only a subset of these 4 tests are classified as 'blocking' due to a lack of HTTP reply or DNS errors. Meanwhile, at this time, all URLs that we classified as 'blocked' at Etisalat were the result of explicit blockpages.

On the other hand, in 2011, 181 distinct sites were classified as 'blocked' during *every* test run on Du (most of these URLs were tested 8 times during this time period). Now, sites were explicitly blocked using blockpages. These blocked sites mostly covered pornography, drugs, censorship circumvention (*e.g.,* tor.eff.org, www.anonymizer.com), social networks (orkut.com) but also a number of political and religious websites as well (*e.g.,* http://www.savezackshahin.com, http://www.faithfreedom.org).

Finally, to demonstrate the value of revisiting the ONI data with our new methodology, we attempt to confirm previous reports that indicate that UAE blocks access to "all Web sites" on the Israeli country code TLD .il [Deibert et al. 2008]. We arrive at conflicting results. Specifically, we see Etisalat, AS 5384, primarily (but not exclusively)

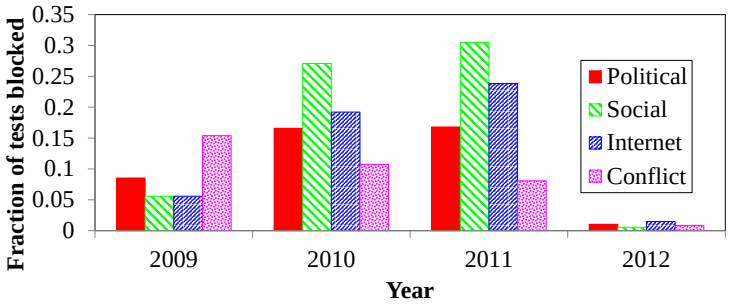


Fig. 7. Tests classified as 'blocked' per Table III in Burma over time. We see that blocking decreases sharply in 2012.

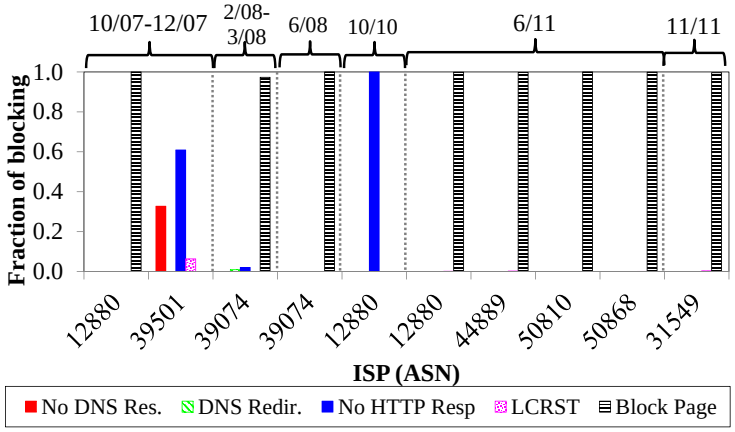


Fig. 8. Summary of blocking in Iran (ISPs with at least 10 blocked URLs per year in at least 2 years).

using block pages to censor .il domains prior to 2010; however, in 2010 this blocking ceased. In contrast, we see no blocking on .il at Du, because tests of .il domains at Du were done in January 2008, before Du began Internet filtering, and after 2010, when blocking of .il domains ceased.

4.4. A decrease in blocking in Burma

We illustrate the opposite trend with a short case study of Burma. Prior to 2012, blocking in Burma was extensive, with more than 5% of tests blocked using blockpages without DNS redirection; this suggests the use of filtering web proxies. Blocking sharply decreased in 2012, when, after years of military rule, the country shifted to a civilian government [ONI Research Profile: Burma 2012]. Figure 7 shows the fraction of tests that were classified as 'blocking' in Burma. (We created this figure using the methodology in §4.8, but this time breaking out tests results by the type of content tested, see Table I.) Figure 7 illustrates the decrease in censorship. Moreover, the only content we classified as "blocked" in 2012 relates to pornography, LGBT issues, alcohol and drugs, and not political issues (as in prior years).

This further highlights the non-monotonic nature of Internet filtering; political shifts can cause countries to cease blocking, regardless of their previous investments in deploying filtering technologies.

Table V. Comparison of the set of URLs classified as "blocked" (per Table III) in the Iranian national ISP (AS 12880, A) and the set of "blocked" URLs in another Iranian ISP (B).

ASN (B):	AS 44889	AS 50810	AS 50868
Union:	579	589	589
Intersection:	381	555	558
Jaccard	0.66	0.94	0.95

4.5. A shift towards centralization in Iran

We present the following longitudinal analysis of Iran to show how a country can shift from decentralized blocking to centralized blocking. In 2001, Iran mandated that all ISPs must connect to the Internet via AS 12880, and reports [ONI Research Profile: Iran 2009] indicate that Iran has since moved to a centralized filtering architecture. We can confirm this using tests conducted on Iran in mid-2011 (Figure 8).

We focus on tests that were conducted in four different Iranian ISPs in June 2011 when testing was done in many ISPs in a single month. To analyze the consistency of blocked content across these ISPs, we introduce a new technique, based on (a variation) of the Jaccard similarity coefficient. The Jaccard coefficient $A \cap B / A \cup B$ measures the difference between two sets, where 1 indicates a perfect matching and 0 indicates two disjoint sets. To control for variation in time and URLs lists, we compare tests of a given URL conducted from the Iranian national ISP AS 12880, to tests of that same URL from a different Iranian ISP *two weeks* before and and after the AS 12880 test. A test was marked as "blocked" in a given time period if at least one test in that network matched a criteria in Table III.

The Jaccard similarity between the set of sites observed blocked in the national ISP AS 12880 (A) and a second ISP (B) is computed for ISPs tested in June 2011 and presented in Table V. The table shows that AS 50810 and AS 50868 are very similar to the national ISP AS 12880. Moreover, almost all blocking observed in this time period (across all ISPs tested) was accompanied by an explicit blockpage containing an iframe to redirect the client to content located at IP 10.10.34.34, which is further evidence of a centralized infrastructure. There were only a small number of exceptions (less than 30 URLs) that received no HTTP response.

Our dataset only has visibility into multiple ISPs in a single other time period: 11/2007, when testing was performed in AS 12880 (the national ISP) and AS 39501 (a residential ISP which was not tested in 2011). However, there is little evidence of blocking at AS 39501 in this time period. Meanwhile the national ISP AS 12880 presented blockpages for 285 different URLs, almost all of which were sites for porn, drugs, and censorship circumvention apart from a few notable exceptions (*e.g.,* http://cyber.law.harvard.edu/ and http://www.citizenlab.org/). Indeed, repeating the Jaccard computation for this pair of ISPs in 11/2007 finds a Jaccard coefficient of zero, illustrating a lack of centralization during this time period.

4.6. Decentralized blocking in Indonesia

Indonesia has been known to use DNS-based blocking to make Internet access "clean and safe" [DNSnawala ; Bu 2010]. Indeed, collateral damage as a result of DNS-based filtering was felt by Internet users in 2009, when censorship of a blog depicting the Prophet Mohammad resulted in the entire blogspot.com domain being unavailable [Sutrisno 2010; ONI Research Profile: Indonesia 2012].

In contrast to our case study in §4.1, where we mostly saw lower-confidence blocking, we observe much more high-confidence blocking in Indonesia. For example, if we group together all tests run in Indonesia, and classify those that meet the criteria in

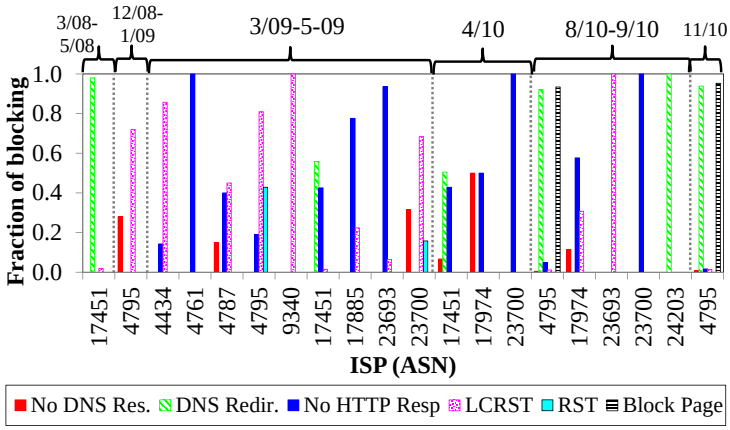


Fig. 9. Summary of blocking in Indonesia (ISPs with at least 10 blocked URLs per year in at least 2 years).

Table III as "blocking", we find that 68% of these tests were classified as "blocking" because of DNS redirection. Block pages are the second most common technique, employed in 45% of the blocked tests (of these 98% also matched the DNS redirection criteria of Table III). Because detection of block pages relies on manual analysis, this observation is robust but should only be considered a lower bound; in fact, manual analysis confirmed the some cases of DNS redirection in Indonesia actually contained blockpages that ONI had previously missed.

Further, Figure 9 illustrates the significant decentralization in how ISPs implement blocking. Figure 9 was generated in the manner described in §4.1. To control temporal and URL-list variation, we clustered tests at a given ISP based on their timing, and for each ISP-time cluster, we classified a test as "blocking" if it matched at least one criteria in Table III. Finally, we broke out each "blocking" technique according to (the one or more) criteria it met in Table III, and plot the fraction of each.

In AS 4795 and 24203 we observe blockpages delivered via DNS redirection to IP addresses hosted by each ISP. The blockpage returned by AS 4795 was identified by the OpenNet Initiative's initial analysis of the data, however, for AS 24203 our observation of DNS redirection led to further analysis and identification of a previously unobserved blockpage (hence, why there is no blockpage bar for this AS). We note that AS 4795's use of blockpages begins in late 2010 showing a shift from RST packets which were observed in tests conducted in 2008 and 2009. Similarly, we observe AS 24203 implementing a combination of DNS redirection and blockpages in 2010, but lack longitudinal data about this ISP. This evolution of blocking may have been spurred by the Indonesian government's development of "Trust+", a platform to facilitate Internet censorship built on top of the Squid Web proxy open source software [ONI Research Profile: Indonesia 2012; trustpositive 2013]. In contrast to AS 4795 and 24203 who display blockpages, AS 17451 implemented DNS redirects that went to non-routable IPs, and therefore look like transient failures from the user's perspective. In 2008, redirects went to IP 0.0.0.1, but in 2009 and onwards we saw this shift to the link-local IP 169.254.1.1 address. Finally, we note that in 16 different runs over three years at First Media (AS 23700) showed no evidence at all of DNS redirection.

4.7. Decentralized DNS-based blocking in Vietnam

This case study of Vietnam illustrates the decentralized nature of Internet filtering in certain countries. Interestingly, while the Vietnamese government is open about

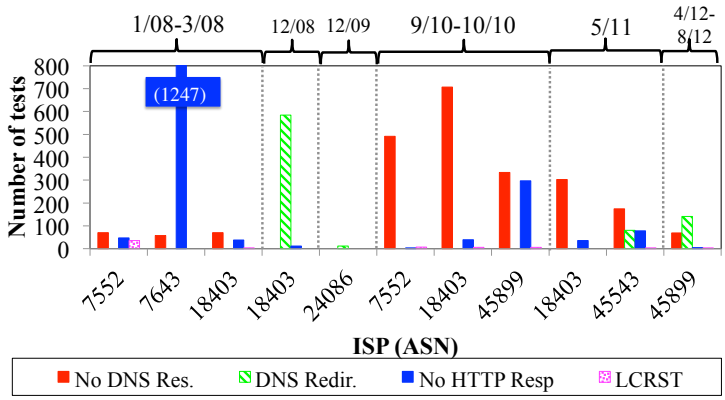


Fig. 10. Summary of tests classified as "blocked" (per Table III) in Vietnam. (We only show results for ISPs with at least 10 "blocked" URLs per time clusters.) DNS manipulations are the dominant form of censorship in this country.

filtering what it deems to be socially- and culturally-offensive content [Deibert et al. 2008], all the blocking we observe looks like transient failures to the user; no evidence of blockpages were found in the ONI dataset. Instead, we mainly observed the following blocking techniques: DNS redirects to "localhost" (*i.e.,* 127.0.0.1), unresolved DNS queries, and sometimes also no HTTP reply (without a DNS redirect).

We created Figure 10 using the methodology discussed in §4.1, to show how blocking technologies evolve over time in Vietnamese ISPs. In 2008, we observed an anomalously high number of tests with no HTTP response. Upon closer inspection these were from five runs within a two hour span (likely a tester repeatedly testing while having networking problems). Aside from this outlier, the vast majority of blocking techniques we observe are no DNS response, DNS redirection and sometimes also no HTTP response. Differences in blocking between ISPs stems from the fact that Vietnam's Internet censorship is largely decentralized, with the government issuing orders to ISPs stating what content to block, but leaving implementation up to the ISPs [Deibert et al. 2008].

The observation that Vietnamese ISPs rely heavily on DNS manipulations for censorship could explain why namehelp [Otto et al. 2012], a tool that optimizes DNS performance, has a disproportionately high rate of adoption in Vietnam[5]. Interestingly, the namehelp tool was designed for performance optimization, not censorship circumvention; we are engaging in ongoing investigation to validate whether this technology has been re-purposed by users a circumvention tool, as a direct response to government censorship.

4.8. Evidence for surveillance in South Korea

Finally, we show how measurements of web censorship can sometime also uncover evidence of surveillance. Our dataset only has good visibility into blocking in South Korea in 2008, which like Vietnam, uses mainly DNS-based approaches. In contrast to the decentralization in Vietnam, DNS redirection in Korea is centralized, with DNS redirects in four ISPs (AS 3786, AS 9701, AS 4766 and AS 9318) sent to a *single* IP address hosted in SK Broadband, AS 9318. The redirection to this address did not return any blockpages; instead, it drops HTTP GET requests. Moreover, redirection to this address was primarily used to block content related to foreign relations (*e.g.,*

[5]http://aqualab.cs.northwestern.edu/projects/151-namehelp

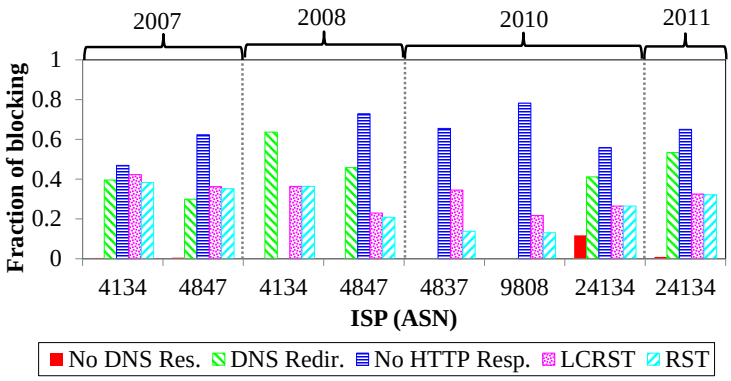


Fig. 11. Summary of how *domain names* containing the "falun" are blocked in China.

North Korea), pornography, and gambling. Interestingly, we see a completely different form of DNS redirect used for content related to Internet (e.g., "warez"); instead of redirection to the address at SK Broadband, these sites were blocked by redirecting to localhost, so that the user's HTTP request never leaves the local machine. This dual use of DNS redirection is highly suspicious, and may indicate surveillance of requests for certain types of content.

The centralized nature of the blocking we see is consistent with the fact that a centralized government agency, the Korean Communications Standards Commission, is responsible for regulating access to Internet content through Internet filtering, deletion of content and closing down Web sites, *etc.* [Deibert et al. 2008]. Our historical observations also contrast with more recent results [Verkamp and Gupta 2012], that find that DNS redirects are are still prevalent in South Korea in 2012; however, the redirections now result in blockpage delivery instead of unrequited HTTP GETs.

4.9. The impact of censorship triggers in China

Censorship in China has received much attention in the research literature (*e.g.,* [Anonymous 2012; Crandall et al. 2007; Park and Crandall 2010; Xu et al. 2011; Fallows 2008; Clayton et al. 2006; Lowe et al. 2007]). While the ONI dataset contains a variety of information about censorship in China, we limit ourselves to a few observations of how URLs can trigger the Chinese censorship system.

Our posthoc analysis on the ONI data allows us to look at how URLs with the same semantic content can elicit different blocking behavior depending on how they trigger the Chinese censorship system. A number of studies [Xu et al. 2011; Crandall et al. 2007] of censorship in China use HTTP requests containing "falun" to elicit blocking by TCP resets, so we investigate more closely how URLs containing the term "falun" are blocked in our data set. A URL containing "falun" in the domain name (*e.g.,* www.falun.com) can trigger IP address, DNS or keyword blocking. In contrast, for a URL where "falun" is in the URL path (*e.g.,* en.wikipedia.org/Falun_gong) is unlikely to elicit DNS-based (or IP-based) blocking, because this would that block the entire domain (*e.g.,* en.wikipedia.org) (or all domains sharing the same IP) as collateral damage, an undesirable outcome from the point of view of the censor. Instead, we expect a URL like en.wikipedia.org/Falun_gong to trigger keyword-based blocking; in the language of §3 this manifests as a TCP reset, an unrequited HTTP GET or a blockpage, but not a DNS redirect or DNS error.

We test this hypothesis by distinguishing URLs where "falun" is in the hostname, and those where it is in the URL path. Unsurprisingly, when falun is in the URL path

Table VI. Blockpages delivered without HTTP redirection in the Middle East and North Africa.

Country	#Tests in country with blockpage	blkpg & no redirect	Total Tests
UAE	9,224	9,224	82,462
Tunisia	4,218	3,807	46,051
Oman	1,358	1,265	19,332
Iran	6,567	5,554	106,288
Qatar	9,56	893	18,835
Yemen	2,563	2,431	48,228
Burma	3,758	3,050	69,166
Kuwait	381	351	8,923
Saudi Arabia	544	512	62,463

(*e.g.,* en.wikipedia.org/Falun_gong), we only observe tests that match the TCP reset criteria in Table III, which confirms the hypothesis and earlier observations [Clayton et al. 2006; Xu et al. 2011]. However, for URLs with falun in the domain name (*e.g.,* www.falun.com), we observe a combination of DNS redirection, dropped HTTP GETs and reset blocking; see Figure 11, which was generated according to the methodology of §4.1 using only URLs tested in China with "falun" in the domain name. This subtle difference highlights the importance of testing for more than just the expected types of blocking; indeed, in China a wide range of techniques are used to block the *same* content. Moreover, we even see variation in blocking rates for the *same* content blocked with different technologies: 99% of tests with "falun" in the HTTP path were classified as "blocking" according to a criteria in Table III, as compared to only 81% of tests with "falun" in the domain name.

5. INTERNATIONAL TRENDS IN CENSORSHIP

We start by discussing global trends in blockpage delivery and DNS redirection, and then move on to some analysis of how measurement vantage points and tested content can have an impact on censorship measurement studies.

5.1. Global trends in blockpage delivery

Blockpages were observed in fifteen countries. We most commonly saw blocking in the absence of DNS redirection, which suggests that transparent Web proxies that perform deep inspection of packet payloads (to the HTTP level) are being used for blocking, rather than the less computationally-costly process required to intercept and tamper with DNS packets.

In Indonesia and Thailand, we saw blockpages delivered both with *and* without DNS redirection. However, the broad trend we observed was that blockpages were delivered without DNS redirection, with the strongest evidence of this in Burma and countries in the Middle East North Africa (MENA) region. To illustrate this, in Table VI we aggregate all test results in each country (last column), and determine the number of tests that elicited blockpages (second column), and show the number of such tests where blockpages were delivered in the absence of DNS redirection (third column); Table VI shows the result for countries where $> 80\%$ of blockpages were delivered in the absence of DNS redirection.

In a concurrent study [Dalek et al. 2013], we used a fingerprinting methodology to identify specific URL filtering products being used for censorship. In that study, we were able to identify McAfee SmartFilter being used in Qatar, Saudi Arabia and UAE; and Netsweepr being used in Qatar, UAE, and Yemen. (Since these products may be used in combination we sometimes observed multiple products on the same network.) This repurposing of a few network management technologies for censorship,

Table VII. DNS redirects.

DNS redirects were observed in the following countries: cn, de, lb, kw, ir, tr, ve, az, kg, jo, ru, kz, ae, th, mm, pe, fr, ma, eg, dz, vn, in, la, gt, ph, my, cu, ng, id, pk, kr, cz

Table VIII. DNS redirects to private IPs.

Private IP	Country (Year observed)
0.0.0.0	Egypt (2008), Nigeria (2008)
0.0.0.1	Indonesia (2008), India (2008), Kuwait (2009)
10.0.0.1	Pakistan (2008)
10.1.1.254	Turkey (2008)
127.0.0.1	UAE (2011), Algeria (2009), Egypt (2008), France (2007-9), Guatemala (2008), India (2008-11), Korea (2008-9), Morocco (2009), Malaysia (2008), Peru (2008), Thailand (2010), Venezuela (2010-11), Vietnam (2008-12)
127.0.0.251	Germany (2008)
169.254.1.1	Indonesia (2009-10), Laos (2008),
192.168.0.1	India (2009)
192.168.1.2	Kazakhstan (2010)
192.168.2.23	Russia (2008)
192.168.56.1	Kyrgyzstan (2009)

highlighted in Table VI, simplifies the task of circumvention, as there are fewer potential censorship actions to consider (*e.g.,* the set of actions these products are known to take when censoring URLs).

5.2. Global trends in DNS redirection

Using the methodology in §3.2, we observed DNS redirection in 32 countries as shown in Table VII. We now discuss a few global trends.

First, consider DNS redirection to a private reserved IP address. It turns out this blocking technique is used in a number of different countries, as shown in Table VIII; redirection to localhost 127.0.0.1 is by far the most prevalent, used in a total of 14 countries. To reduce false positives, where users are redirected to cache servers on the local network, we remove instances where the server header indicates a cache (*e.g.,* 2wireGateway observed in China) or content was delivered successfully from the internal IP.

Next, we consider DNS redirection to a routable IP address. DNS redirects can be used to direct traffic towards a middlebox that takes filtering action (*e.g.,* drops the HTTP request, delivers a blockpage), and we found evidence for this in many countries (*e.g.,* South Korea §4.8, Indonesia, and others).

DNS redirection can also be used to direct traffic to an unrelated (but blocked) Web site or IP address, that trigger IP-based blocking mechanisms (*e.g.,* routers that filter on an address or prefix). However, we only found evidence for the latter in China. To identify this type of DNS redirection, we required that the IP address resolved in the field fit the criteria of Appendix A.1, and also that this IP address was hosted by an AS in a different country than that of the ISP being tested. We used RIPE's BGP data [RIPE Network Coordination Center] to map each IP address to the AS that originated it at the time of the test. Table IX summarizes out-of-country DNS redirection observed in Chinese ISPs in 2011. The set of organizations reflects a variety of large networks, and includes a block registered to the US Department of Defense

Table IX. DNS redirection in China in 2011.

URLs	IP	Reg. Organization
87	8.7.198.45	Level 3 Communications
86	203.98.7.65	TelstraClear Ltd
85	46.82.174.68	Deutsche Telekom AG
85	59.24.3.173	Korea Telecom
81	93.46.8.89	Fastweb SpA
78	78.16.49.15	BT Ireland Backbone
77	159.106.121.75	DoD Network Information Center

that was not globally routable at the time of testing[6]. Tests redirected to these IP addresses received no HTTP response, which suggest that these requests triggered IP-based blocking mechanisms. DNS redirection of this type was also documented in China in 2007 [Lowe et al. 2007]; our dataset has observations from 2007-11 that are validated by lists published by Lowe *et al.* and other online sources [ViewDNS.info 2011; GreatFire.org 2012].

The only other case of DNS redirection to an IP address hosted outside the country was in Azerbaijan; we found that blocked tests in Azerbaijan in 2007 pointed to an IP address that corresponds to the Web site myfamily.com. For these tests, the HTTP status was 200 (successful), but the HTTP body contained no content. This suggests that DNS was used to redirect traffic to a middlebox that intercepted the HTTP request and returned a blank web page.

These observations highlight the need for future censorship measurement platforms to check for DNS-based manipulations; they also have implications for the deployment of DNSSEC, a technology that was specifically designed to prevent DNS manipulations.

5.3. Academic vs. national networks

Placing measurement probes in academic networks is often more convenient than gaining access to users in a country of interest, especially in locations where PlanetLab is present [Sfakianakis et al. 2011; Verkamp and Gupta 2012]. However, PlanetLab has two key limitations when studying censorship. First, PlanetLab is not present in many countries where we found strong evidence of blocking. For example, of fifteen countries where we identified blockpages only Turkey, Tunisia, Thailand, Malaysia and Pakistan have PlanetLab nodes. Similarly, 19 of the 32 countries where we observe DNS redirection lack a PlanetLab node (these countries had tests that were identified as redirects per the criteria in the second row of Table III). Finally, of the countries discussed in our case studies, only China and South Korea have PlanetLab nodes.

Further, the representativeness of PlanetLab nodes, which usually reside in well provisioned academic networks, is debatable. Lower rates of keyword blocking have been observed on paths to academic Web servers [Crandall et al. 2007]. We confirm this observation across the set of ten countries where we have results in both academic and non-academic networks. To do this, we compare the fraction of content blocked in academic and non-academic networks by reprising the technique based on Jaccard similarity described in §4.5. We compare tests of given URL conducted from an academic network, to tests of that same URL from a non-academic network within the same country for *two weeks* before and after the academic vantage point test. A test was marked as "blocked" in an academic network for a given time period (and thus added to set A) if at least one test in that network matched a criteria in Table III. The

[6]We use ARIN rather than historical BGP data from RIPE to determine this.

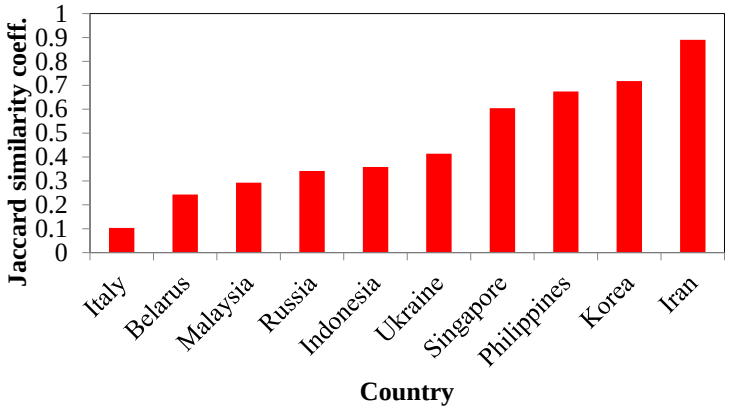


Fig. 12. Jaccard similarity coefficient comparing tests classified as 'blocked' (per Table III) in academic and non-academic networks. See discussion in §5.3.

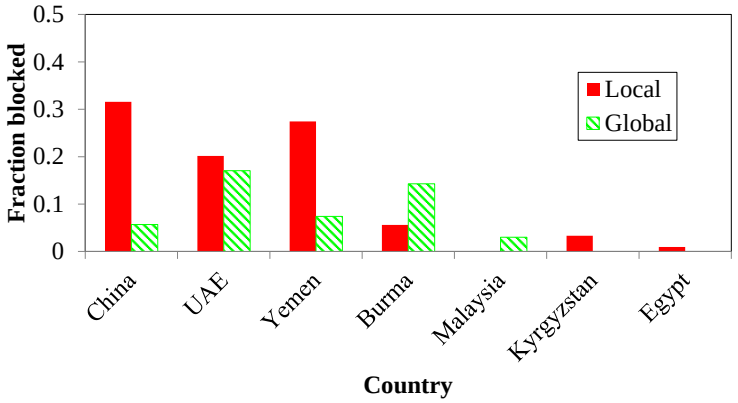


Fig. 13. Rates of tests classified as "blocking" (per Table III) for local and global URLs. (We only show results for countries with at least 10 local URLs.)

same approach was used to add "blocked" tests in non-academic networks to set B, and the Jaccard coefficient $\frac{|A \cap B|}{|A \cup B|}$ was computed and plotted in Figure 12. Averaging over the ten countries where we had academic and non-academic results, we find that the Jaccard coefficient is 0.59, which is indicative of dissimilarity.

Moreover, we identify blocking in non-academic networks more often. The ONI dataset contained 1,947,691 tests where the same URL was tested in academic and non-academic networks within the 4 week period. Of these, 72,454 tests from non-academic vantage points were identified as "blocking" according to at least one criteria in Table III, with only 52,921 "blocking" tests in academic networks, corresponding to 36% more blocking in non-academic networks.

5.4. Locally relevant vs. global content

An often overlooked part of censorship measurement, is that one must know *what* content is likely to trigger censorship. Prior work relied on specific keywords to trigger censorship [Xu et al. 2011; Crandall et al. 2007], crowd-sourced domain names [Verkamp and Gupta 2012] or search engine results for specific types of content [Sfakianakis et al. 2011] to trigger censorship. An advantage of our dataset is that we have lists of locally-sensitive URLs for many of the countries we study. These lists are curated by

local groups and are designed to capture content that (1) has been reported as blocked or (2) will likely be blocked. We note here that the notion of "local" is based on the content of the site, and does not necessarily mean that the site is hosted in the given country. Globally sensitive URLs include content likely to be blocked in many countries (*e.g.,* international human rights organizations).

Local context is crucial for understanding censorship. Figure 13 illustrates the importance of collecting locally-relevant URL lists for measuring Internet filtering, by comparing the blocking rates between URLs on local and global lists. To avoid bias in countries with short local lists, we limit ourselves to countries with at least 10 local URLs. We aggregate together all tests run on local URLs in a given country, and compute the fraction of these tests identified as blocking according to one of the criteria in Table III. We do the same for all global URLs in a given country, and plot the result in Figure 13. As suggested in prior work [Deibert et al. 2008], most countries we tested (with the exception of Burma) show higher blocking rates on local content. The effect is especially pronounced in China and Yemen, where local content is blocked 3-5X more than global content; this is likely due to language issues and the fact that China filters local political content most actively of all four content types tested. In contrast, we observe UAE blocking more global content; because they primarily censor social content with less local bias (*e.g.,* pornography).

To aid future censorship measurement studies to run tests on locally-relevant content, we have released the ONI dataset to the research community. We hope that this data will enable future researchers to get a more accurate picture of censorship practices worldwide.

6. RELATED WORK

Censorship measurement. There is a growing literature that focuses on *what* is censored [Deibert et al. 2008; 2010], and *how* it is censored [Zittrain and Edelman 2003; Clayton et al. 2006; Xu et al. 2011; Anonymous 2012; Dornseif 2004; Clayton 2006; ICANN Security and Stability Advisory Committee (SSAC) 2012; Weaver et al. 2009; Wolfgarten 2006; Fallows 2008; Sfakianakis et al. 2011; Verkamp and Gupta 2012], and how censorship can be circumvented [Leberknight et al. 2012a; 2012b; Elahi and Goldberg].

The literature on *what* is censored is extensive, multidisciplinary, and covers multiple countries (see [Deibert et al. 2008; 2010] for a small sample); indeed, the ONI dataset we analyze was also collected with this objective in mind. However, our second look at the ONI dataset correlated measurements across time, in contrast to ONI reports and blog posts that focus on a set of measurements made at a single point in time. Moreover, our inference methodology §3 has uncovered a number of observations that the ONI missed, including the possibility of surveillance in South Korea §4.8, the fact the UAE stopped blocking the .il domain after 2010 §4.3, methodologies for stealthy blocking in Yemen §4.2, filtering techniques in Azerbaijan §5.2, the global trends in §5.1-§5.2, and others. Moreover, by releasing the ONI the research community, we hope to enable future longitudinal analysis of web censorship.

While analysis of the technical means used for censorship [Wolfgarten 2006; Fallows 2008], including TCP RSTs [Clayton et al. 2006; Crandall et al. 2007; Weaver et al. 2009] and DNS blocking [ICANN Security and Stability Advisory Committee (SSAC) 2012] are available, most work in this space has been focused on censorship in China (see *e.g.,* [Zittrain and Edelman 2003; Clayton et al. 2006; Crandall et al. 2007; Fallows 2008; Xu et al. 2011; Anonymous 2012]), with a few notable early measurement studies about censorship in the UK [Clayton 2006] and Germany [Dornseif 2004]. In recent years, this literature has expanded: ToR now offers statistics on censorship

events in multiple countries [TorProject 2014], large scale Internet disconnections in Egypt and Libya have been analyzed [Dainotti et al. 2011], and studies that look at web censorship in multiple countries are now available [Verkamp and Gupta 2012; Sfakianakis et al. 2011]. However, some web censorship studies mostly rely on PlanetLab nodes [Verkamp and Gupta 2012; Sfakianakis et al. 2011] to obtain vantage points in the countries under study, and our analysis of §5.3 suggests that testing from user vantage points gives us a more complete view of blocking events. Moreover, our analysis gives a complementary, historical view of web censorship.

Measurement platforms. The need to access vantage points for measurements is a key challenge in measuring information controls, however it is not unique to these types of measurements. The need for a global measurement platform led many measurement researchers to use PlanetLab [PlanetLab 2013], which provides access to hosts at institutions around the globe. However, as we have observed (§5.3), PlanetLab is not necessarily representative of end user connectivity. Differences between PlanetLab and end users has led to the development of a myriad of measurement tools and platforms to gain access to representative vantage points in recent years [Sundaresan et al. 2011; Sanchez et al. 2013; RIPE Atlas 2014; The ICSI Netalyzr 2014; Kreibich et al. 2010]. However, despite similarities to network measurement, studies of censorship face many challenges that are not addressed by existing platforms. Specifically, censorship measurements require accessing content that either the government or ISP has deemed inappropriate, which can pose a risk to end users [Burnett and Feamster 2013; Wright et al. 2011].

Our results also suggest that new measurement platforms to detect and analyze censorship on an ongoing basis are sorely needed. CensMon is one measurement platform, but thus far has only been deployed on PlanetLab nodes [Sfakianakis et al. 2011], while OONI [Filasto and Appelbaum 2012] and ICLab [ICLab 2014] are promising platforms under development. Our results suggest that these platforms should take care to perform repeated measurements from representative vantage points, especially in countries where stealthy censorship is the norm (*e.g.,* Kyrgyzstan §4.1, Yemen §4.2). To aid these efforts, we have made public the ONI's testing URL lists.

7. CONCLUSIONS

Our results highlight considerable variability observed in Internet filtering worldwide. Future measurement studies should account for decentralization of censorship infrastructure that can lead to variation between ISPs in the same country, political events that cause censorship behaviors to change or even cease over time, and the type of content that is used to test for blocking, and the types of blocking technologies that it triggers. Indeed, the considerable variability we observed here cautions against making sweeping statements about censorship activities globally, or even nationally. We hope that our work motivates more interdisciplinary measurement work, especially in the design of URL testing lists, the timing and interpretation of measurements, as well as more longitudinal research that addresses the conflicting challenges of scale and representativeness of measurements. These studies can have significant impact on our understanding of the information controls that are being used in countries worldwide. This understanding is critical, not only for circumvention technologies, but also for understanding how new technologies may facilitate or circumvent censorship and how government policy impacts censorship infrastructure.

Acknowledgments

We thank Martin Arlitt for comments on a draft of this paper. Phillipa Gill was supported by a Munk School post-doctoral fellowship. We especially thank the numerous

collaborators and testers involved in the OpenNet Initiative over the past 10 years. The OpenNet Initiative is financially supported by the John D. and Catherine T. MacArthur Foundation.

REFERENCES

Daniel Anderson. 2012. Splinternet Behind the Great Firewall of China. *Queue* 10, 11 (2012), 40.

Anonymous. 2012. The collateral damage of internet censorship by DNS injection. *ACM SIGCOMM Computer Communication Review* 42, 3 (2012).

M. Arlitt and C. Williamson. 2005. An Analysis of TCP reset behavior on the Internet. *ACM CCR* (2005).

Donny Bu. 2010. Indonesia Internet Censorship. ONI Global Summit. http://www.slideshare.net/donnybu/indonesian-internetcensorshipreport2010. (2010).

Sam Burnett and Nick Feamster. 2013. Making sense of internet censorship: a new frontier for internet measurement. *ACM SIGCOMM Computer Communication Review* (2013).

Richard Clayton. 2006. Failures in a hybrid content blocking system. In *Privacy Enhancing Technologies*. DOI:http://dx.doi.org/10.1007/11767831_6

R. Clayton, S. Murdoch, and R. Watson. 2006. Ignoring the great firewall of China. In *Privacy Enhancing Technologies*.

J. Cowie. 2011a. Egypt Leaves the Internet. *Renesys Blog* (2011).

J. Cowie. 2011b. Libyan Disconnect. *Renesys Blog* (2011).

J. Cowie. 2012. Syrian Internet is off the air. *Renesys Blog* (2012).

J.R. Crandall, D. Zinn, M. Byrd, E. Barr, and R. East. 2007. ConceptDoppler: a weather tracker for Internet censorship. In *14th ACM Conference on Computer and Communications Security*. 1–4.

A. Dainotti, C. Squarcella, E. Aben, K.C. Claffy, M. Chiesa, M. Russo, and A. Pescapé. 2011. Analysis of country-wide internet outages caused by censorship. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 1–18.

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

R.J. Deibert, J.G. Palfrey, R. Rohozinski, and J. Zittrain. 2008. *Access denied: The practice and policy of global Internet filtering*. MIT Press.

R.J. Deibert, J.G. Palfrey, R. Rohozinski, and J. Zittrain. 2010. *Access controlled: The shaping of power, rights, and rule in cyberspace*. MIT Press.

DNSnawala. DNS Nawala. http://www.nawala.org/. (????).

Maximillian Dornseif. 2004. Government mandated blocking of foreign Web content. *CoRR* cs.CY/0404005 (2004).

T. Elahi and I. Goldberg. Technical Report.

James Fallows. 2008. The Connection Has Been Reset. *The Atlantic* (March 2008).

R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. 1999. Hypertext Transfer Protocol – HTTP/1.1. RFC 2616 (Draft Standard). (1999). http://www.ietf.org/rfc/rfc2616.txt Updated by RFCs 2817, 5785, 6266, 6585.

A. Filasto and J. Appelbaum. 2012. OONI: Open observatory of network interference. In *USENIX FOCI*.

Google. 2014. Transparency Report. http://www.google.com/transparencyreport/. (2014).

GreatFire.org. 2012. (2012). https://twitter.com/GreatFireChina/status/266898482298773505.

ICANN Security and Stability Advisory Committee (SSAC) 2012. *SSAC Advisory on Impacts of Content Blocking via the Domain Name System*. Technical Report. ICANN Security and Stability Advisory Committee (SSAC).

ICLab 2014. ICLab. (2014). http://iclab.org.

C. Kreibich, N. Weaver, B. Nechaev, and V. Paxson. 2010. Netalyzr: Illuminating the edge network, In The Internet Measurement Conference. *Unknown Journal* (2010).

C.S. Leberknight, M. Chiang, and F.M.F. Wong. 2012a. A Taxonomy of Censors and Anti-Censors: Part I-Impacts of Internet Censorship. *International Journal of E-Politics (IJEP)* 3, 2 (2012), 52–64.

C.S. Leberknight, M. Chiang, and F.M.F. Wong. 2012b. A Taxonomy of Censors and Anti-Censors Part II: Anti-Censorship Technologies. *International Journal of E-Politics (IJEP)* 3, 4 (2012), 20–35.

G. Lowe, P. Winters, and M.L. Marcus. 2007. The Great DNS Wall of China. *MS, New York University* 21 (2007).

S. Murdoch and R. Anderson. 2008. Tools and technology of Internet filtering, In Access Denied: The practice and policy of global Internet filtering. *Unknown Journal* (2008).

H. Noman. Dubai free zone no longer has filter-free Internet access. ONI Blog: http://opennet.net/blog/2008/04/dubai-free-zone-no-longer-has-filter-free-internet-access. (????).

H. Noman. Middle East Censors Use Western Technologies to Block Viruses and Free Speech. (????). http://opennet.net/blog/2009/07/middle-east-censors-use-western-technologies-block-viruses-and-free-speech.

ONI Research Profile: Burma 2012. ONI Research Profile: Burma. http://opennet.net/research/profiles/burma. (2012).

ONI Research Profile: Indonesia 2012. ONI Research Profile: Indonesia. http://opennet.net/research/profiles/indonesia. (2012).

ONI Research Profile: Iran 2009. ONI Research Profile: Iran. http://opennet.net/research/profiles/iran. (2009).

ONI Research Profile: Yemen 2009. ONI Research Profile: Yemen. http://opennet.net/research/profiles/yemen. (2009).

OpenNet Initiative. 2014. OpenNet Initiative. (2014). https://opennet.net/.

J. Otto, M. Sanchez, T. Stein J. Rula, and F. Bustamante. 2012. namehelp: intelligent client-side DNS resolution. In *SIGCOMM 2012 Poster*.

J.C. Park and J.R. Crandall. 2010. Empirical study of a national-scale distributed intrusion detection system: Backbone-level filtering of html responses in China. In *ICDCS*.

PlanetLab 2013. PlanetLab: An open platform for developing, deploying, and accessing planetary-scale services. (2013). http://www.planet-lab.org.

Rensys Blog. Pakistan Hijacks YouTube. (????). http://www.renesys.com/blog/2008/02/pakistan_hijacks_youtube_1.shtml.

RIPE Atlas 2014. RIPE Atlas. (2014). https://atlas.ripe.net/.

RIPE Network Coordination Center. RIPE Routing Information Service. (????). http://www.ripe.net/data-tools/stats/ris/routing-information-service.

M. Sanchez, J. Otto, Z. Bischof, D. Choffnes, F. Bustamante, B. Krishnamurthy, and W. Willinger. 2013. Dasu: Pushing Experiments to the Internet's Edge, In Network Systems Design and Implementation (NSDI). *Unknown Journal* (2013).

A. Sfakianakis, E. Athanasopoulos, and S. Ioannidis. 2011. CensMon: A Web censorship monitor. In *USENIX FOCI*.

S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, and A. Pescape. 2011. Broadband Internet Performance: A View From the Gateway, In ACM SIGCOMM. *Unknown Journal* (2011).

W. Sutrisno. 2010. Indonesia Internet censorship. (2010). http://www.sutrisno.me/2010/05/indonesia-internet-censorship.html.

Team Cymru IP to ASN Lookup v1.0 2013. Team Cymru IP to ASN Lookup v1.0. (2013). http://asn.cymru.com/cgi-bin/whois.cgi.

The ICSI Netalyzr 2014. The ICSI Netalyzr. (2014). http://netalyzr.icsi.berkeley.edu/.

TorProject. 2014. Tor Metrics Portal: Users. https://metrics.torproject.org/users.html?graph=direct-users&country=sy&events=on#direct-users. (2014).

trustpositive 2013. Trust+ Positif. (2013). http://trustpositif.kominfo.go.id/?lang=en.

J.P. Verkamp and M. Gupta. 2012. Inferring Mechanics of Web Censorship Around the World. In *Usenix FOCI*.

ViewDNS.info. 2011. DNS Cache Poisoning in the People's Republic of China. (2011). http://viewdns.info/research/dns-cache-poisoning-in-the-peoples-republic-of-china/.

N. Weaver, R. Sommer, and V. Paxson. 2009. Detecting forged TCP reset packets. *In Proc. of NDSS. Citeseer* (2009).

Sebastian Wolfgarten. 2006. *Investigating large-scale Internet content filtering*. Ph.D. Dissertation. Security & Forensic Computing, Dublin City University, Ireland.

J. Wright, T. de Souza, and I. Brown. 2011. Fine-grained censorship mapping: information sources, legality and ethics, In USENIX Workshop on Free and Open Communication on the Internet. *Unknown Journal* (2011).

X. Xu, Z. Mao, and J. Halderman. 2011. Internet censorship in China: Where does the filtering occur?. In *Passive and Active Measurement*.

Asma Ali Zain. 2008. Internet users stumble across 'adult' content. *Khaleej Times* (April 11 2008).

Jonathan Zittrain, Ryan Budish, and Rob Faris. 2014. Herdict: Help spot web blockages. http://www.herdict.org/. (2014).

J. Zittrain and B. Edelman. 2003. Internet filtering in China. *IEEE Internet Computing* 7, 2 (2003), 70–77.

Earl Zmijewski. 2010. Accidentally Importing Censorship. Renesys blog. http://www.renesys.com/blog/2010/03/fouling-the-global-nest.shtml. (2010).

A. METHODOLOGY & VALIDATION

A.1. Criteria for identifying DNS redirects

As discussed in §3.2.2, we needed a robust technique to distinguish "suspicious" DNS redirects from the DNS-load balancing typically used by CDNs and other services. Recall that for every test URL, our dataset has a mapping to an IP address both in the field and in the lab. To decide when a discrepancy between the IP resolved in the field and in the lab corresponds to a suspicious DNS redirect, we use the idea with benign DNS-based load balancing, a set of URLs that resolves to IPs hosted by *multiple different* ASNs in the controlled lab setting is unlikely to resolve to an IP hosted by a *single* AS in the field tests.

Therefore, we created the following mapping. We collected all tests run from a single field AS vantage point, and then grouped together all tested URLs that resolved to the same IP address *in the field*. For each test in the group, we determine the IP address the tested URL resolved to *in the lab*, and finally use RIPE's BGP data [RIPE Network Coordination Center] to map each IP address to the AS that originated it at the time of the test. We thus mapped from each IP address observed at a given field ISP vantage point to a *set* of ASNs observed in the lab. For example, that set of test URLs run by clients located at field vantage point AS 4847 that all resolved to 72.14.205.104 (Google), those same tests, when conducted in the lab, resolved to IPs hosted by AS 25653, AS 19262, AS 8972, AS 6939 and more than 45 additional ASes; we therefore say that the field IP address 72.14.205.104 corresponds to more than 50 lab ASNs.

Under benign DNS-based load-balancing conditions, we expect this set of ASNs to be small; large sets are indicative of suspicious DNS redirection. Next, we needed a way to determine how many lab ASNs an IP should be associated with before declaring the result suspicious. As such, we used a threshold to distinguish between suspicious and non-suspicious tests. Figure 14 shows the fraction of results classified as no HTTP reply, reset or blockpage (see techniques in Section 3), where the IP address resolved in the field corresponded to more than N ASNs in the lab. We observe that as N decreases, the fraction of results classified as blocking increases until $N = 32$ and $N = 52$, for no HTTP response and RSTs, respectively; the spike around $N = 32$ is caused by a large number of blocking results in AS 9318. Based on Figure 14, we decided on a conservative threshold of $N = 32$ to separate DNS redirection from CDNs and load balancing.

A.2. Validation of DNS redirection criteria

We validate our criteria for identifying DNS redirection to routable IP addresses by checking that tests satisfying this criteria also satisfy another type of blocking criteria. We do this because DNS redirection alone is not enough to block access; it must be paired with a mechanism that disrupts the connection (e.g., a block page). Of all the tests in the ONI dataset that matched our DNS redirection criteria, only 17% result in a successful HTTP transaction (i.e., HTTP status 2xx or 3xx). These 17% of tests are redirected to a set of only 28 distinct IP addresses. Manual analysis of these addresses confirmed that they were indeed used for censorship. One of these address was a redirect to an IP address that corresponds to the Web site myfamily.com, which was returned for all blocked content in Azerbaijan. We also found an IP address that was

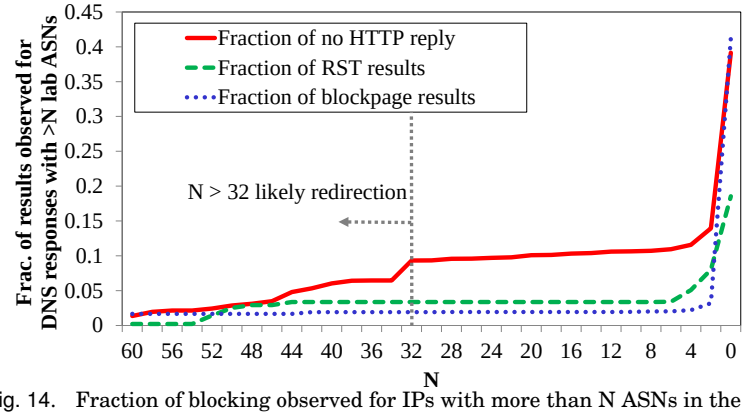


Fig. 14. Fraction of blocking observed for IPs with more than N ASNs in the lab.

Table X. RSTs matching signatures [Weaver et al. 2009]

Signature	Total Results	RST	LCRST
RSTs associated with injecting middleboxes			
IPID = 64	259	206 (77%)	221 (85%)
Seq+=1460	528	481 (91%)	483 (91%)
Benign RSTs			
RST seq = 0	372	0 (0%)	32 (9%)

serving a blockpage in Indonesia (but the block page had not been previously identified by the ONI's manual analysis).

A.3. Validation of TCP reset criteria

Validating our inferences about RSTs was made challenging by the fact that the ONI's measurement software does not collect packet traces by default. We use the few tests where packet captures were collected (via tcpdump) to validate our RST inferences. In total, we have 6,716 tests with packet captures containing RSTs collected in Ethiopia and China[7]

We validate our heuristic for detecting suspicious RST events, by comparing RSTs observed in packet captures to against the distinctive IP header signatures that Weaver *et al.* [Weaver et al. 2009] found to be used by middleboxes that perform "traffic shaping" via TCP RSTs. Our goal is to verify that RSTs which match the signatures found by Weaver *et al.* are correctly inferred to be injected RSTs using our heuristics. We note that we do not expect all of the injected RSTs to match these signatures, nor do we aim to do an exhaustive study of signatures, rather we use the existing signatures for validation purposes only.

Weaver *et al.* [Weaver et al. 2009] observed signatures associated with middleboxes that inject TCP RSTs (*i.e.,* IPID=64 and Seq+=1460) and well as signatures for 'benign' RSTs that are not associated with middleboxes (seq=0). Our validation, in Table X, shows that the our criteria identify RSTs that match Weaver *et al.* 's signatures for RST-injecting middleboxes 77-91% of the time. Moreover, our high-confidence RSTs *never* coincide with benign resets per [Weaver et al. 2009].

[7]The set of countries is a product of the limited by countries that use RST injection for censorship as well as the set of countries where tests were performed with packet captures.