# Toward Automated Worldwide Monitoring of Network-level Censorship

Submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Zachary Weinberg

B.A. Chemistry, Columbia University
M.S. Cognitive Science, University of California, San Diego

Carnegie Mellon University
Pittsburgh, PA

December, 2018

# Abstract

Although Internet censorship is a well-studied topic, to date most published studies have focused on a single aspect of the phenomenon, using methods and sources specific to each researcher. Results are difficult to compare, and global, historical perspectives are rare. Because each group maintains their own software, erroneous methods may continue to be used long after the error has been discovered. Because censors continually update their equipment and blacklists, it may be impossible to reproduce historical results even with the same vantage points and testing software. Because "probe lists" of potentially censored material are labor-intensive to compile, requiring an understanding of the politics and culture of each country studied, researchers discover only the most obvious and long-lasting cases of censorship.

In this dissertation I will show that it is possible to make progress toward addressing all of these problems at once. I will present a proof-of concept monitoring system designed to operate continuously, in as many different countries as possible, using the best known techniques for detection and analysis. I will also demonstrate improved techniques for verifying the geographic location of a monitoring vantage point; for distinguishing innocuous network problems from censorship and other malicious network interference; and for discovering new web pages that are closely related to known-censored pages. These techniques improve the accuracy of a continuous monitoring system and reduce the manual labor required to operate it.

This research has, in addition, already led to new discoveries. For example, I have confirmed reports that a commonly-used heuristic is too sensitive and will mischaracterize a wide variety of unrelated problems as censorship. I have been able to identify a few cases of political censorship within a much longer list of cases of moralizing censorship. I have expanded small seed groups of politically sensitive documents into larger groups of documents to test for censorship. Finally, I can also detect other forms of network interference with a totalitarian motive, such as injection of surveillance scripts.

In summary, this work demonstrates that mostly-automated measurements of Internet censorship on a worldwide scale are feasible, and that the elusive global and historical perspective is within reach.

# Acknowledgments

This dissertation has benefited from the assistance of dozens of people over the years it has been in preparation. First among these are of course my advisor, Nicolas Christin, and the other members of my committee, Lujo Bauer, Vyas Sekar, and Phillipa Gill. The other members of Dr. Christin's research group at CMU, particularly Luís Brandão, Aya Kunimoto, Srujana Peddada, Mahmood Sharif, Kyle Soska, and Janos Szurdi, and Dr. Gill's research group at UMass-Amherst, particularly Shinyoung Cho, Nguyen Phong Hoang, Arian Niaki, Abbas Razaghpanah, and Rachee Singh, have also been regular sounding boards and sources of suggestions and criticism throughout the process.

Chapter 2 is based on a paper coauthored by Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. Chapter 3 includes material from an unpublished paper coauthored by Arian Niaki, Shinyoung Cho, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, and Phillipa Gill. Chapter 4 is based on a paper coauthored by Mahmood Sharif, Janos Szurdi, and Nicolas Christin. They have all benefited from feedback from anonymous reviewers at the various conferences to which they have been submitted.

The experiments described in Chapter 2 were assisted by 40 anonymous volunteers and 150 anonymous paid workers. I also thank Sumana Harihareswara for assistance in recruiting volunteers, and Michael Gargiulo and Mohammad Taha Khan for providing me with data on the commercial VPN ecosystem. The 15 "potentially censored" lists discussed in Chapter 4 were compiled by various anonymous whistleblowers and watchdog groups. The "negative control" list was compiled by Pamela Griffith. Ariya Hidayat provided technical assistance with the retrieval of uncensored pages, and Karen Lindenfelser provided technical assistance with the classification process. The "*de novo* manual classification" described in Chapter 5 was prepared by Tianyi Ma, Yanhao Li, and Yuhong Zha.

Although none of my previous work on steganographic communication appears in this dissertation, I would not have begun the research projects that do appear here if I had not done that work first, and so I also wish to acknowledge my previous co-authors and assistants, variously from CAIDA, CMU, the Internet Systems Consortium, SRI International, Stanford University, and the Tor Project: Arjun Athreya, Dan Boneh, Linda Briesemeister, Brian Caswell, Steven Cheung, kc claffy, Drew Dean, Bruce DeBruhl, Roger Dingledine, Paul Hick, Daira Hopwood, George Kadianakis, Eric Kline, Andrew Lewman, Patrick Lincoln, Ian Mason, Jeroen Massar, Nick Mathewson, Phillip Porras, Steve Schwab, William Simpson, Paul Vixie, Mike Walker, Frank Wang, Jeffrey Wang, and Vinod Yegneswaran.

All of my friends and relations, but particularly Michael Ellsworth, Pamela Griffith, Sumana

Harihareswara, Riana Pfefferkorn, Nathaniel Smith, and Dara Weinberg, have provided editorial and technical assistance, moral support, snacks, answers to peculiar questions, distractions, tea, and good cheer throughout the long process of research and writing.

Financial support for my and my coauthors' research has come, at various times, from the National Science Foundation (awards DGE-1147470, CCF-0424422, and CNS-1223762); the Department of Homeland Security Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD); the Government of Australia and SPAWAR Systems Center Pacific (through BAA-11.02, contract number N66001-13-C-0131); the Defense Advanced Research Project Agency (DARPA) and SPAWAR Systems Center Pacific (contract number N66001-11-C-4022); the MSIT (Ministry of Science and ICT), Korea, under the ICT Consilience Creative Program (IITP-2017-R0346-16-1007) supervised by the IITP (Institute for Information & Communications Technology Promotion); and the Open Technology Fund (Information Controls Fellowship Program). This work does not represent an official position of any of the aforementioned funders.

This dissertation is dedicated to the memory of Mia Alexander.

# Contents

# List of Tables

# List of Figures

# 1. Introduction

For 25 years, the Internet has been the site of a struggle between people who wish to access information, express opinions, and communicate with each other, and people who wish to control what information can be accessed, what opinions can be expressed, and who can communicate with whom. National governments are especially fond of imposing restrictions on online communication [50]. Some of these restrictions have had international consequences [12, 19, 28, 123]. Activists regularly raise concerns about exporting "network management" products from countries with strong human rights protections, to countries that will use them to violate human rights [47].

The literature is rich with studies of individual aspects of Internet censorship (see Section 1.3 for a detailed review), but most are limited to a single country and a relatively short period of time. Censorship has become commonplace worldwide, but China and its infamous Great Firewall still receive a disproportionate amount of attention, and cross-border comparisons are few and far between. Each group has developed its own set of measurement tools, and their code is not often shared, leading to duplication of effort, persistence of errors, and incomparable results. Finally, far more attention has been paid to mechanism (how is the censorship carried out?) than to policy (what are the censor's goals and how does the execution reflect that?)

I suggest that there are four basic challenges to be overcome before we can achieve a comprehensive, global and historical understanding of Internet censorship:

**Challenge 1: Access to Vantage Points.** With few exceptions, measuring Internet censorship requires access to "vantage point" hosts within the region of interest. Countries where access is (relatively) easy to come by have received more research attention than those where it is difficult. Relying on volunteers limits breadth and duration, may make it outright impossible to access some countries, and raises ethical problems [30, 70, 132]. Most of the proposed alternatives cannot capture nearly as much detail [62, 140].

**Challenge 2: Reliable Worldwide Detection.** There are several different ways for a censor to prevent access to content, and they may use different techniques for different purposes [76], so studies that focus on a single technique are inherently incomplete. Detection strategies developed without careful, broad evaluation are prone to error, both because there are innocuous causes for every network-level anomaly one can think of [141, 156, 196] and because variation from place to place or year to year will invalidate tests that are too specific [96].

**Challenge 3: Understanding What to Test.** Testing a single blocked URL can reveal that a censorship system exists within a country, but does not reveal the details of the censorship policy or how aggressively it is enforced. Many studies rely on a "probe list" of keywords, URLs, and/or

domain names that are suspected to be censored, but the contents of that list are often left unspecified, rendering the research unreproducible. Even the broadest and best-maintained probe lists, such as the set maintained by the Citizen Lab [39] only capture a small fraction of the websites censored in each country [48, 49, 92]. Knowing what is censored does not always reveal *why* it is censored [43].

**Challenge 4: Maintenance and automation.** To collect data on Internet censorship over a long period of time, someone must continuously operate and maintain a monitoring system. It must be adapted to changes over time, and it should record enough detail with each measurement that it is possible to re-analyze old data using new criteria [147]. As much of the monitoring and analysis process as possible should be automated, simply because the volume of data produced by the system will be unmanageable otherwise.

In this dissertation I will show how to address all of these challenges at once:

> *It is possible to build a system that can detect network-level censorship of many varieties, and monitor globally for changes over time, with minimal human intervention.*

While I do not claim to have completely solved any of these challenges, I have made advances upon all of them. I will present four specific advances: Chapter 2 describes a way to verify the physical location of vantage points. This will allow a monitor to rely more on commercial VPN services than had previously been possible, mitigating the ethical and practical problems with using volunteer labor. Chapter 3 describes improved methods for detecting censorship. These reduce error rates by looking at data from all levels of the network stack. It also describes a semi-automated method for identifying when newly discovered network anomalies are due to censorship. Chapter 4 uses standard techniques of automated natural language processing to assess the contents of an existing probe list, and compare it to leaked actual blacklists and to lists of websites compiled via other means. This reveals where effort needs to be concentrated when updating the probe list.

Finally, in Chapter 5, I conclude by describing future plans for refining probe lists, using the same natural-language processing techniques to direct keyword searches and web crawls.

## 1.1 Background

In this section, I briefly review the techniques used to block access to information online, describe two different ways they can be implemented in networking equipment, and finally discuss how the censor's intentions can affect their choice of method.

This thesis only discusses censorship carried out by man-in-the-middle interference at the level of IP, TCP, and DNS, and it concentrates on applications of this kind of censorship to the Web. However, the techniques I describe are applicable to other forms of censorship. The network-layer

probing and monitoring techniques described in Chapter 3 could be applied to detect network interference with any application protocol, and this is a near-term goal for future work. Application-layer censorship, as seen in e.g. Chinese chat clients [72] could be studied with the topic analysis methods described in Chapters 4 and 5, as long as there is a source of censored content and a way to detect censorship events.

### 1.1.1 Network-level blocking techniques

A network attacker—in the context of censorship, usually this is a backbone router equipped with "deep packet inspection" capabilities [42]—can interfere with access to websites in one of the following ways:

**DNS manipulation.** When visiting a website, the user's browser first sends a DNS request to resolve the website's domain name to an IP address. DNS traffic is cleartext, and less than 1 % of it is authenticated [177]. Censors can inject DNS responses intended to reach the client before the reply from the legitimate DNS resolver, carrying DNS error messages (e.g. NXDOMAIN), non-routable IP addresses, or the address of a server controlled by the censor, which the browser will attempt to access instead of the legitimate server [13, 201].

**IP-based blocking.** Once the browser has an IP address of a web server, it attempts to make a TCP connection to that server. Censors can discard TCP SYN packets destined for IP addresses known to host censored content, or reroute them to a server controlled by the censor [89]. Discarding packets is only suitable for covert censorship (see Section 1.1.3 below), but rerouting them risks interfering with traffic outside the borders of the censor's authority [28].

**TCP packet injection.** Censors can also allow TCP connections to proceed normally, but then inject a packet into the TCP stream that either supersedes the first response from the legitimate server, or breaks the connection before the response arrives [184]. For unencrypted websites, this technique allows the censor to observe the first HTTP query sent by the client, and thus block access to individual pages [43].

**Transparent proxy.** Censors wishing to exercise finer control than is possible with the above three techniques can use a "transparent proxy" that intercepts all HTTP traffic attempting to leave the country, decodes it, and chooses whether or not to forward it [47]. These devices act as TCP peers and may modify HTTP traffic they allow to pass through, which makes them detectable [183]. They facilitate fine-grained decisions about *how* to block access to content, e.g. delivering a generic "page not found" message for some content and a "access forbidden due to local laws" message for others [46]. However, they are specific to HTTP; they cannot be used for, e.g., VoIP censorship.

### 1.1.2 On-path and in-path censors

The network equipment responsible for implementing censorship can operate in two different ways, with different visible symptoms. *On-path* equipment observes a copy of all traffic passing through a bottleneck network link. It can react to the traffic by injecting more packets into the link, but it cannot modify or discard packets that are already within the flow. *In-path* equipment operates on the actual traffic passing through the network link, not a copy. It can therefore modify or discard packets as well as injecting new packets.

On-path equipment is cheaper and easier to deploy, but because it cannot modify or discard packets, its actions are easier to detect: injected packets will appear alongside packets from the legitimate server. Which of the two techniques is more common is not well understood [83, 174].

### 1.1.3 Overt and covert censorship

Depending on the censor's intentions, blocking can be conducted in either an *overt* or a *covert* fashion. In overt blocking, the censor informs the user that something was censored, by arranging for them to see a "block page" instead of the material that was censored. For instance, intercepting an HTTP connection and redirecting the browser to a server controlled by the censor, which displays a message stating that access to this material is forbidden for legal reasons, is an act of overt censorship. In covert blocking, the censor *avoids* informing the user that something was censored, usually by arranging a network error that could have occurred for other reasons. For instance, intercepting a DNS lookup and replying with a "no such host" (NXDOMAIN) error is an act of covert censorship.

Overt blocking of web content can be accomplished with a transparent HTTP proxy, an injected TCP packet or DNS response that directs the browser to a server controlled by the censor, or by rerouting TCP traffic from a legitimate server's IP address to a server controlled by the censor. Covert blocking can be accomplished with a transparent HTTP proxy, an injected TCP RST packet, an injected DNS NXDOMAIN or non-routable address, or by discarding packets intended for a legitimate server's IP address.

Censors can engage in both overt and covert blocking. For instance, Yemen has been observed to use overt blocking for pornography (which is illegal there) and covert blocking for disfavored political content (which is legal) [76].

## 1.2 Architecture of a continuous censorship monitor

I will present my advancements upon the four challenges I outlined in Section 1 in the context of a proof-of-concept system for continuous censorship monitoring and discovery. This system builds

upon an existing system for continuous censorship monitoring, ICLab [147], with whose principals I have been collaborating. ICLab has been continuously collecting data since 2016, in 63 countries, covering 246 ASes and testing over 46 798 unique URLs over the course of more than two years. ICLab currently focuses on web censorship by network interference, but its principals have plans to expand it to other applications (e.g. VoIP and instant messaging).

My technique for verifying the physical location of vantage points (Chapter 2), and my methods for detecting censorship (Chapter 3) improve simpler techniques previously in use by ICLab. My assessment of the contents of existing probe lists (Chapter 4), and my future plans for refining probe lists (Chapter 5), rely on data collected by ICLab and, in the future, will feed the refined lists back into ICLab's monitoring.

### 1.2.1 Censorship vantage points

Censorship monitoring involves accessing material that is forbidden in a particular country, from that country, and provoking a response from the censor. The expected response is relatively harmless—a "block page" or a forged TCP RST—but there may be real world risks associated with running these tests, especially for volunteers already engaged in human rights reporting or advocacy. (See Section 1.3.4 for how other researchers have addressed these issues.)

**VPN-based clients.** ICLab uses VPN-based clients whenever possible, because of their practical and ethical advantages. It is not necessary to recruit volunteers from all over the world, or manage physical hardware that has been distributed to them, but the measurement software still has unrestricted access to the network, unlike, for instance, phone or web applications [30, 70]. The VPN operator guarantees high availability and reasonable bandwidth. There is no need to worry about consuming volunteers' transfer quota, which would both inconvenience them and disrupt the experiment.

On the other hand, VPN servers tend to be hosted in commercial data centers (specifically, in "content" rather than "transit/access" ASes according to the CAIDA classification [33]). These ASes may not be able to observe all of the censorship applied to, or by, last-hop ISPs serving individuals [5, 195]. Some VPNs are reported to engage in surveillance and traffic manipulation that could interfere with ICLab's measurements [104]. Websites may refuse service to client IP addresses known to be VPNs [126, 156]; we must take care not to confuse this for censorship.

User-hosted VPNs (e.g. Geosurf [21], Hola [90], Luminati [122]) would offer access to residential ISPs, but ICLab currently does not use them, as they have all the ethical concerns associated with volunteer-operated clients, with less transparency. Also, there are reports of illicit actions by the operators of these VPNs, such as deploying their software as a viral payload, and facilitating distributed denial of service (DDoS) attacks [127].

On the ethical side, a commercial VPN operator is a company that understands the risks of doing business in each country it operates in. It is unlikely that they would deploy a server in a country where the company or its employees might suffer legal or extralegal sanctions for the actions of its users. They might have accepted a legal obligation to censor or surveil their users on behalf of the government; if this matches the legal obligations that a residential ISP in that country would have to accept, it would actually be helpful to us, since it would mean we *were* observing the same sort of censorship applied to residential ISPs, perhaps with some variation. On the other hand, the obligations might be intentionally different—for instance, a government that suspects commercial VPNs of being used by foreigners to monitor their censorship policies, might order the VPN company to impose *no* censorship on certain accounts. We would then have to stop using that VPN.

Some VPN operators are suspected of claiming to have servers in countries where they do not have a physical presence; for instance, a server advertised as located in North Korea, and which does appear to be in North Korea according to popular IP-to-location databases, but is actually in the Czech Republic. Falsely located servers are useless for censorship monitoring, and must be excluded, using the techniques described in Chapter 2.

**Volunteer-operated clients.** Volunteer-operated clients (VOCs) enable access to countries and ASes where ICLab cannot use a VPN. However, they are more difficult to keep running, and require a local volunteer comfortable with the risks associated with operating the device. ICLab obtains informed consent from each volunteer before shipping them a device, maintains contact with them for as long as they operate it, and monitors the political situation in each country where a volunteer-operated client has been deployed. Some countries have been deemed too risky to operate volunteer-operated clients in at present, and/or impossible to deploy equipment to because of an embargo (e.g. Iran, Syria)

### 1.2.2 Breadth of worldwide monitoring

As of September 2018, ICLab maintains vantage points in 63 countries. Naturally, some countries are easier to acquire vantage points in than others. Unfortunately, the countries that do the most censorship, and therefore are of the most interest to censorship researchers, are also the most difficult to access. Therefore, for a complete understanding of how broad ICLab's coverage is, we must consider both which countries it can access, and how aggressively those countries apply censorship.

The international organization Freedom House, which promotes civil liberty and democracy worldwide, issues a yearly report on "freedom on the Net," in which they rate 65 countries on the degree to which online privacy and free exchange of information online are upheld in that country [101]. Each country receives both a numerical score and a three-way classification: 16 of

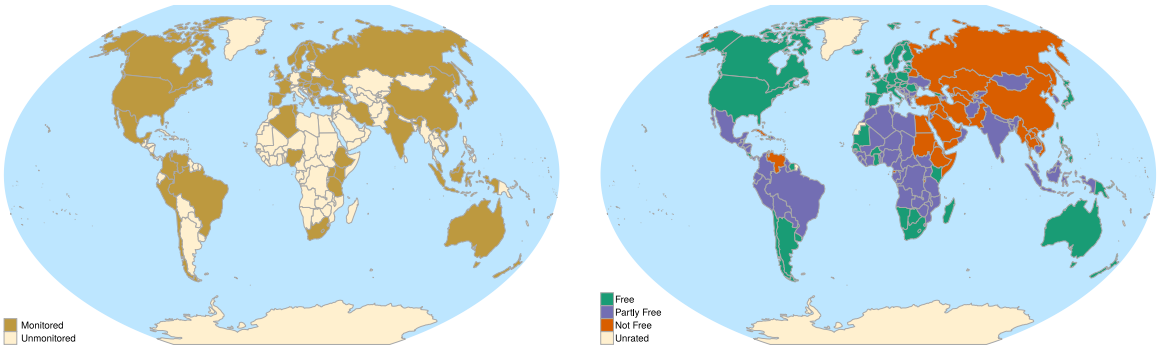



Figure 1.1: Countries monitored by ICLab (left) and the Freedom House/Reporters Without Borders classification of online freedom per country (right)

| | |
|---|---|
| **Measurement duration** | January 2017–September 2018 |
| **Number of measurements** | 53 906 532 |
| **Countries** | 63 |
| **URLs** | 46 798 |
| **ASes** | 246 |
| **Vantage Points** | 314 |

Table 1.1: Summary of data collected by ICLab to date

the 65 countries are considered "free," 28 are "partly free," and 21 are "not free." Unfortunately, 33 of the countries monitored by ICLab are not included in this report.

The international organization Reporters Without Borders (RWB) issues a similar yearly report on freedom of the press. This report covers 189 countries and territories, including all 65 of the countries rated by Freedom House, and all 63 of the countries monitored by ICLab [149]. Each country receives a numerical score and a color code (best to worst: 16 countries are coded white, 42 yellow, 59 orange, 51 red, and 21 black). Press freedom is not the same as online freedom, and the methodology behind the two reports is quite different, but the scores from the two reports are reasonably well correlated (Kendall's $\tau = 0.707$, $p = 1.17 \times 10^{-16}$) Using a simple linear regression to map RWB scores onto the same scale as FH scores, we can assign "free", "partly free", and "not free" labels to the countries not scored by Freedom House.

Figure 1.1 shows how the 63 countries monitored by ICLab are distributed worldwide, and the freedom classification of all 189 countries rated by Freedom House and/or Reporters Without Borders.

### 1.2.3 Censorship measurements

At present, ICLab's measurements are focused on network-level interference with access to websites. Each vantage point performs a cycle of measurements at least once every three days, on a schedule

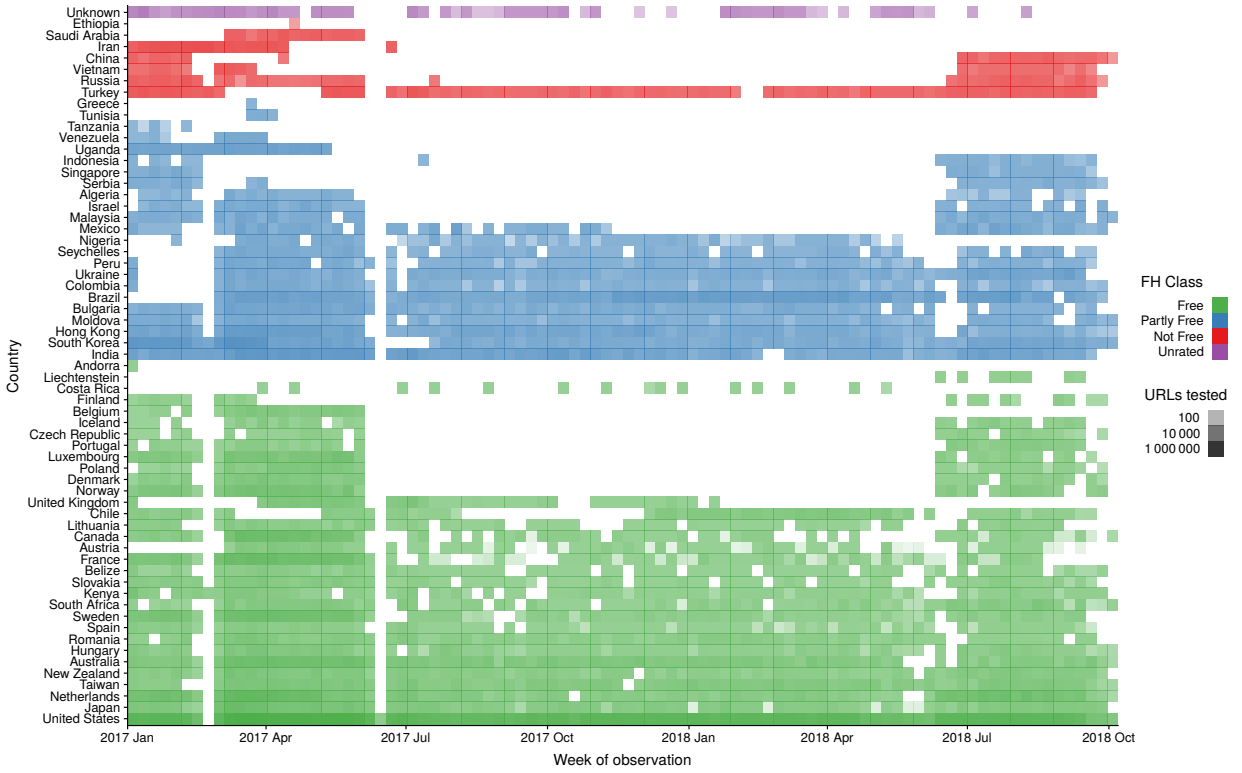



Figure 1.2: Distribution of ICLab's measurements in time and space

controlled by a central server. The set of URLs tested on each cycle is also distributed from the central server. Raw observations are transmitted back; all analysis is done centrally. This design allows new analysis techniques to be applied to old data as they are developed.

A *measurement* of a specific website or web page is an attempt to perform an HTTP GET request to a particular URL, recording information about the results from multiple layers of the network stack: (1) The complete DNS request and response or responses for the server hostname (using both a local resolver and a public DNS resolver), (2) Whether or not a TCP connection succeeded, (3) For HTTPS URLs, the certificate chain transmitted by the server, (4) The full HTTP response (both headers and body), (5) Traceroute information to the URLs being tested, and (6) A comprehensive packet trace for the duration of the measurement.

Many tests of censorship rely on comparison with matching observations recorded from a location where no censorship is anticipated. Therefore, all measurements are repeated from a *control node*, located in an academic network in the USA, which is known not to block access to any of the sites under test.

Table 1.1 summarizes the aggregate data collected by ICLab as of this writing, and Figure 1.2 shows how that data is distributed over the 63 countries that ICLab has access to. The color-coding in Figure 1.2 indicates the freedom classification of each country.

### 1.2.4 Censorship detection modules

In ICLab's architecture, a censorship detection module is a self-contained program that examines the aggregate measurements collected by the vantage points, identifies cases where censorship appears to have occurred, and records them in a database. Detection modules run on a compute cluster attached to the central server. They do not have direct access to vantage points, and their results are expected to be deterministic, for reproducibility's sake.

ICLab presently has three detection modules, each detecting a different signal of censorship: suspicious DNS responses, suspicious injected TCP packets, and text characteristic of block pages. In Chapter 3 I will describe improvements I made to the latter two modules. I was only peripherally involved with the development of the DNS module, so it is not described in this thesis.

### 1.2.5 Uncensored page collection and topic assignment

My additions to ICLab aim at discovery of unsuspected cases of censorship—web pages that are censored, but are not included in the existing sets of URLs to test. I do this by developing topic classifications for pages that are known to be censored, and using the classification to direct keyword searches and web crawls. There is no reason why this process could not make use of censorship observations from other probe platforms, e.g. Iris [141], OONI [70], or Quack [173] as well as from ICLab.

To analyze the topics of censored web pages, we must have access to their uncensored contents. We use an automated web browser, hosted in a commercial data center in the USA, to collect uncensored copies of web pages; we also check the Internet Archive [94] for historical copies of each page, to determine whether its topic has changed. This process is described in detail in Chapter 4.

Using an automated web browser, instead of a simple HTTP client, improves our chances of extracting useful text from web pages that rely heavily on JavaScript. It also handles markup errors and ambiguities up front, simplifying downstream processing.

Once the pages are downloaded, we preprocess them to remove uninteresting material, e.g. navigation boilerplate, copyright statements, and advertisements. The remaining text is mechanically translated into English and then clustered using Latent Dirichlet Allocation (LDA) [24]. The clustering algorithm produces a set of keywords associated with each cluster. We manually assign meaningful labels to the clusters, based on the keywords. This process is also described in Chapter 4.

### 1.2.6 Page discovery

Given a starting sample of known-censored material in a particular country, it is not difficult to discover much more material that is also censored [48, 49, 109, 157]. However, it is not practical to monitor all of that material continuously, so probe list curators must make decisions about which pages, keywords, etc. to include. Developing rigorous criteria to guide those decisions, and tools that will automate as much of the process as possible, is future work. However, Chapter 5 closes this dissertation with some suggestive preliminary results on how manual topic classification compares to LDA classification, and to what extent the topic of a web page predicts the topics of the pages it links to.

## 1.3 Previous work by others

Early adopters believed that the Internet's decentralized basic design and lack of "gatekeepers" (in contrast to traditional publishing) would render it difficult or even impossible to censor, but this has proven to be wishful thinking. While it usually remains possible for a determined and technically skilled individual to evade access restrictions, over the past twenty years, Internet censorship has steadily become more effective, sophisticated, and widespread. The mechanism of censorship is content-neutral—it only takes a configuration change to convert a "filter" aimed at blocking distribution of broadly condemned material (e.g., spam, child pornography, neo-Nazi propaganda) into one that will suppress political opposition. For historical retrospectives, see Bambauer [17], Gallagher [73], and Subramanian [162] and the *Access* book series published by the OpenNet Initiative (ONI) [50, 51, 52].

### 1.3.1 Single-country case studies

The most common type of academic study of online censorship is a case study of a single country at a single point in time. The motivations of these studies vary widely; the following review is not meant to be exhaustive, but rather to give an impression of the breadth of the field.

China, with its aggressive and widely publicized "Great Firewall," has received the most attention. Some studies are entirely about the mechanism of censorship. For instance, in 2006 Clayton, Murdoch, and Watson [42] observed that it was possible for a client computer to evade the Great Firewall by ignoring its forged TCP RSTs. Ten years later, Wang *et al.* [182] repeated this study in more detail; implementation details had changed enough to render older evasion techniques unworkable, but they found new techniques and ways to keep updating them. Winter and Lindskog [189] and Ensafi *et al.* [61] investigated Chinese active probes for circumvention servers, such as

Tor bridges, and Farnan, Darer, and Wright [68] analyzed the effects of the Great Firewall's forged responses on DNS caches within China.

Another group of studies investigate policy variation within China. Xu, Mao, and Halderman [195] and Wright [191] observed significant policy variations from ISP to ISP, and speculated that detailed policy decisions are left to local authorities and/or ISPs. Ensafi *et al.* [63], however, found that the filtering routers are centralized in IXPs (Internet eXchange Points). This is not a contradiction: the purpose of an IXP is to provide a centralized physical location for border routers controlled by several different organizations. Park and Crandall [138] reported that in late 2008, China ceased to block web pages based on keywords in HTML responses; they speculated that this was too expensive and unreliable to be worth doing.

A related line of research explores "collateral damage" due to Chinese networks acting as transit providers: Anderson [10] and an anonymous report [12] both reported that traffic from clients in China's geographic neighbors, to servers worldwide, may be disrupted by the Great Firewall even though the Chinese government has no authority over those clients. China is not the only country in a position to cause collateral damage; Gosain *et al.* [79] observe that many countries in Southeast Asia purchase transit from Indian networks and could be affected by future increases in the level of Indian censorship.

Chinese censorship has also been studied through a more sociological lens. Link [121] compared the techniques of thought control used by the Soviet Union with those used by the People's Republic of China. In their account, both nations' rulers were primarily motivated by a desire to maintain their own power, but Chinese leadership chose to do this via psychological uncertainty, intended to promote self-censorship: "questions of risk—how far to go, how explicit to be, with whom to ally, and so on—are to be judged by each writer and editor.... If you calculate incorrectly you can lose your job, be imprisoned, or, in the worst case, [be executed].... But most censorship does not directly involve such happenings. It involves *fear* of such happenings."

This theory is consistent with Xu, Mao, and Halderman [195]'s observation of significant policy variation from ISP to ISP, and with Knockel, Crete-Nishihata, and Ruan [111]'s observation that chat client authors are legally responsible for deciding what to censor. It is directly echoed by Crandall *et al.* [43]'s observation that keyword filtering by the Great Firewall does not block "every illicit word," but perhaps only "enough to promote self-censorship." Relatedly, Aase *et al.* [1] point out that when detailed policy is left to individuals and organizations, they will not produce an exhaustive blacklist, because that would be too expensive. King, Pan, and Roberts [107] present what they describe as a counterargument, observing that "posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored. Instead,... the censorship program is aimed at curtailing collective action by silencing comments that represent, reinforce, or spur social mobilization, regardless of content." This is not really a contradiction: Chinese leadership could

well have decided that only collective action is a serious threat to their authority.

Wang and Mark [179] surveyed Chinese Internet users, investigating both how aware they are of their government's censorship policies, and how much they approve of them. Unsurprisingly, awareness is directly correlated with technical sophistication, but also with socioeconomic status. Approval, among users who are aware, is predicted by "authoritarian personality," as defined in Altemeyer [7].

Other autocratic nations also practice aggressive Internet censorship, but not with the same technical sophistication as China. Only a few of these countries have received any targeted case studies. Dainotti *et al.* [45] reported on the techniques used by Egypt and Libya to shut down regional access to the Internet during a period of civil instability. Anderson [8] and Aryan, Aryan, and Halderman [15] described the more subtle approach taken by Iran, using bandwidth throttling to discourage use of websites hosted outside the country, combined with blacklisting of sites with objectionable content. Nabi [131] presented a basic overview of the censorship mechanism in Pakistan as of 2013; Khattak *et al.* [106] described how the Pakistani censorship policy changed over the course of a two-year period, and how Internet users reacted to the changes. Chaabane *et al.* [34] analyzed a set of "leaked" logs from the firewall routers in Syria, covering a nine-day period in 2011; of interest is that both the `.il` top-level DNS domain, and all IP address ranges allocated to Israel, are blocked regardless of their content.

In "Western" countries, censorship is considered incompatible with the strong native tradition of free expression, but this does not mean it never happens. Hintz and Milan [88] and Webb [185] both identify a steady trend toward authoritarianism in these countries, and a concurrent trend of "outsourcing" surveillance and censorship decisions to industry, which is often less transparent and less responsible to the general public. As a concrete example, the USA has a long line of legal cases denying the government power to censor *publishers*, such as *New York Times Co. v. United States* [133], but the rights of *readers* are not as well established and there is often local political pressure for schools and libraries to prevent their computers being used to view pornography. Richardson *et al.* [150] investigated whether the commercial software for this purpose might, accidentally or otherwise, also block access to sexual health information that teenagers may need.

France and Germany both have long-standing policies of suppressing neo-Nazi movements by any means necessary, and the public policy literature contains some discussion of whether Internet censorship is an effective tool for this [27, 57]. In Italy, the courts can order ISPs to block access to sites, usually because of copyright infringement, unlicensed gambling, or sale of counterfeit goods. Aceto, Montieri, and Pescapè [5] reported that these orders are inconsistently enforced by five major Italian ISPs. India has similar laws, used for similar purposes but also in an attempt to combat politically-motivated rioting. Two recent case studies [79, 196] have investigated regional and per-ISP variation, finding (as with China) that many decisions are being made at a local level.

As mentioned in the introduction, and discussed further in Section 1.3.5, most of these studies take the set of sites, keywords, and so on to test for censorship as a given.

### 1.3.2 Cross-national comparisons

Another common research angle is to compare the mechanism, policy, and implementation of censorship across several different countries. Much of the work in this vein is aimed at policymakers rather than academics. The ONI maintains a set of policy-focused per-country profiles based on continuous monitoring, comparing all of them against an ideal of free expression [167]. Hellmeier [83] summarized the ONI's data into a "toolkit" of strategies known to be used by autocrats to suppress online political opposition. Singh *et al.* [159] used the structure of the global routing graph to predict the Freedom House "freedom on the net" index [100], based on the observation that censorship is easiest when all traffic passes through a bottleneck node. Bailey and Labovitz [16] observed, from incident analyses, a trend toward greater "untrustworthiness" of backbone providers. Clark, Faris, and Jones [41] study both worldwide and per-country trends in the accessibility of Wikipedia, using server access logs provided by the Wikimedia Foundation. Tschantz *et al.* [171] investigated the related phenomenon of Web servers themselves denying access to clients based on their apparent geographic location. This can indicate that a foreign organization is not prepared to comply with local laws, and in some cases the relevant laws impose censorship.

More academically focused cross-national comparisons divide into two lines of research. One group of studies investigate worldwide variation in the mechanism of censorship: for instance, whether censorship mainly interferes with DNS lookups [141] or subsequent TCP connections, and whether the end-user is informed of censorship [174]. In some cases, it has been possible to identify the specific "filter" software in use [47, 96]. Another line aims to understand what is censored and why [1, 29], how that changes over time [9, 76], how people react to censorship [110, 192], and how the censor might react to being monitored [29].

These studies, especially the "what is censored and why" group, have generally devoted somewhat more attention to the contents of their probe lists, but they are still vague on details.

### 1.3.3 Non-Web censorship

The rise of Internet censorship coincides with the rise to dominance of the Web over other protocols. Some academic attention has been paid to the wholesale blockade of VPN protocols, peer-to-peer file-sharing protocols, and anonymizing protocols such as Tor [54], which are commonly used to circumvent censorship. For instance, Winter [188] describe a privacy-preserving way for Tor users to report when Tor is inaccessible in their country, and Wright, Darer, and Farnan [192, 193] take country-specific anomalies in the level of Tor usage as a sign that people in that country are trying

to circumvent something. Elahi and Goldberg [59] and Tschantz *et al.* [170] are meta-analyses of circumvention and counter-circumvention techniques used in various countries. Several Chinese case studies target the censorship policies enforced at the application layer by chat clients popular in China [109, 110, 111] and similarly social media sites [72, 135, 200].

### 1.3.4 Measurement methodology

Despite the breadth of goals and hypotheses in the studies listed above, most of them used the same basic methodology: from a "vantage point" within the network subject to censorship, attempt to access censored material and observe what happens, to varying levels of detail. The primary exception is when researchers gain access to "leaks" of inside information, allowing them to see a (possibly incomplete) list of censored material; recently this has occurred for backbone filters in Syria [34] and Pakistan [106], the TOM-Skype chat client's internal keyword filter [110], and a variety of Chinese-language mobile applications whose source was available online [109].

Acquiring access to appropriate vantage points is a central methodological problem. The most common technique is to rely on volunteers [5, 55, 70], but of course this requires one to find volunteers in each country of interest. Several groups of researchers have sought alternatives. CensMon [157] used Planet Lab nodes, Anderson, Winter, and Roya [9] used RIPE Atlas nodes, Pearce *et al.* [141] use open DNS resolvers and VanderSloot *et al.* [173] use open echo servers. Commercial VPN providers are another obvious alternative, but they may advertise nodes in more countries than they actually have (see Chapter 2). All of these alternatives risk missing censorship that is applied only to residential networks [5, 195]. Darer, Farnan, and Wright [48] took advantage of a quirk in the Chinese Great Firewall: a DNS query originating from anywhere, directed to any IP address allocated to China, will receive a forged reply if the domain name is censored.

Many more vantage points would be available if researchers could use third parties' computers without their owners' cooperation. The proposal that goes the furthest in this direction is Encore [30], which would embed censorship monitoring software within unrelated web pages, to be executed, invisibly, by each browser visiting that page. If deployed on popular websites, this could enlist vantage points anywhere in the world. It would also access potentially-censored sites exactly as normal browsing would, which avoids false positives due to sites blocking automated access. However, it would be difficult to be certain of, and impossible to control, where the vantage points were, and it has serious ethical problems [132]. The censor will observe people in their jurisdiction visiting censored websites that they would not otherwise have, and both the censored websites and the monitoring software could be subverted to distribute malware that the clients would not otherwise have been exposed to.

Ensafi *et al.* [63] also proposed to use third parties' computers, but much more passively; their

"hybrid idle scan" technique causes these computers to perform reflected SYN probes. Taking advantage of an information leak in many implementations of TCP/IP ("predictable IP-ID numbers"), they can tell, from off-path, whether the reflected SYNs received a forged RST from the censor. This is very lightweight, so it can be applied to thousands of third-party IP addresses to study within-country variation in detail. It avoids some of Encore's ethical quandaries; it does not risk exposing the third parties to malware, but it still does make the censor observe people in their jurisdiction attempting to access censored sites that they would not otherwise have. However, it can only detect forged RSTs, not any other method of censorship. Pearce *et al.* [140] addressed the ethical concern by selecting reflector IP addresses known to be assigned to routers rather than clients, but this could cause them to miss censorship happening very close to the edges of the network (perhaps on the very router they are using as a reflector).

Only a few studies have lasted more than a month. However, since 2012 several groups have operated continuous monitoring systems: OONI [70], Encore [30], Satellite [156], IRIS [141], Quack [173], and ICLab, which is described in Chapter 3. Herdict [20] has also been active for many years, but it only aggregates reports of inaccessible websites from around the world, without determining why they are inaccessible.

### 1.3.5 Probe list development

Despite the central position of keyword and URL probe lists in all of these studies, relatively little attention has been paid to the contents of the lists, or how they are developed. Ongoing monitoring projects such as OONI [70], Herdict [20], and the OpenNet Initiative [167] rely primarily on volunteer submissions [166]. Some countries have activist groups who maintain "watchdog" sites, publishing a list of web pages that are censored in that country, and other countries have experienced one-time "leaks" of the list. In all the above cases, researchers receive a raw list of unverified allegations. It may contain a tremendous amount of uninteresting detail (e.g. every image file used by a gambling website). It may also contain many sites censored for uninteresting reasons (e.g. the censor bought an off-the-shelf filter originally intended to discourage people from watching porn on a public-access terminal, then added a handful of locally politically sensitive sites to its blacklist). Websites may have ceased to exist since the list was published, or their names may have been reused for something unrelated. Cleaning all such detritus out of these lists presently requires a great deal of manual effort, and must be done with great caution to avoid malware, not to mention things one would rather not have seen.

To date, there have been only a few efforts to automatically find pages that might be censored. CensMon [157] attempted to auto-generate lists of censored keywords, using a latent semantic model, but ran into problems due to the large number of homographs in Chinese. Hounsel, Mittal,

and Feamster [92] improve on CensMon by using bigrams and trigrams as well as single words, which disambiguates many of the homographs. FilteredWeb [48] proposes to extract identifying "tags" from known-censored pages and use them to search for more pages which may be censored, and Darer, Farnan, and Wright [49] checks whether the outbound hyperlinks from blocked pages are likely to lead to more blocked pages. All of these projects were able to discover hundreds or even thousands of censored pages that did not appear in existing probe lists, but they offer no criteria for deciding which of the new discoveries to add to existing probe lists; in Chapter 5 we will discuss some possibilities for developing such criteria.

# 2. Validation of VPN Proxy Locations

As we mentioned in Section 1.2.1, ICLab uses VPN-based clients whenever possible. Commercial VPN services compete to offer the highest speed, the strongest privacy assurances, and the broadest possible set of server locations. There is no difficulty in finding a VPN service that *advertises* servers in any location you like—including implausible locations such as North Korea, Vatican City, and Pitcairn Island. They offer no proof of their claims. On several occasions in 2015 and 2016, we were unable to reproduce censorship observations reported by volunteers, using a VPN service claiming to host a server in the same country.

There are innocuous explanations for this; for instance, data center networks might experience less censorship than residential networks [195]. However, there is also a sinister possibility. VPN services that consolidate their servers in a smaller number of locations than they advertise can choose those locations for better performance, reliability, and reduced operational expenses. This gives them a competitive advantage over services that strive for true location diversity. If they can manipulate IP-to-location databases, they can still provide the *appearance* of location diversity, which will be enough for many customers, e.g. those seeking to defeat geographic restrictions on online media streaming [2]. IP-to-location databases have been shown to be full of errors [75, 143, 158]. Worse, they rely on information that VPN providers may be able to manipulate, such as location codes in the names of routers [36].

In this chapter, we apply *active geolocation* to check the advertised locations of VPN servers. Active geolocation estimates the location of an Internet host by measuring packet round-trip times between it and other hosts in known locations. It has been demonstrated to work at the scale of a large country or small continent (e.g. China, Europe, and the USA), with varying levels of accuracy, depending on how efficient the regional network is [38, 53, 66, 119]. However, it has not been thoroughly tested at the scale of the entire world.

Using active geolocation, we can usually locate a VPN server to within 1 000 $km^2$, anywhere in the world. Our results are more precise in more densely connected regions and/or when landmarks are nearby, but even when we are uncertain about where a server actually is, we can still disprove blatant inaccuracies in marketing claims. For instance, if we know that a server is in Belgium, Netherlands, or Germany, but not which, that still proves it is not in North Korea. We tested 2 269 servers operated by seven VPN services, including five of the top 20 by number of claimed countries. *At least a third of all the servers we tested are not in their advertised country.*

The material in this chapter was previously published as "How to Catch when Proxies Lie" at IMC 2018 [186].

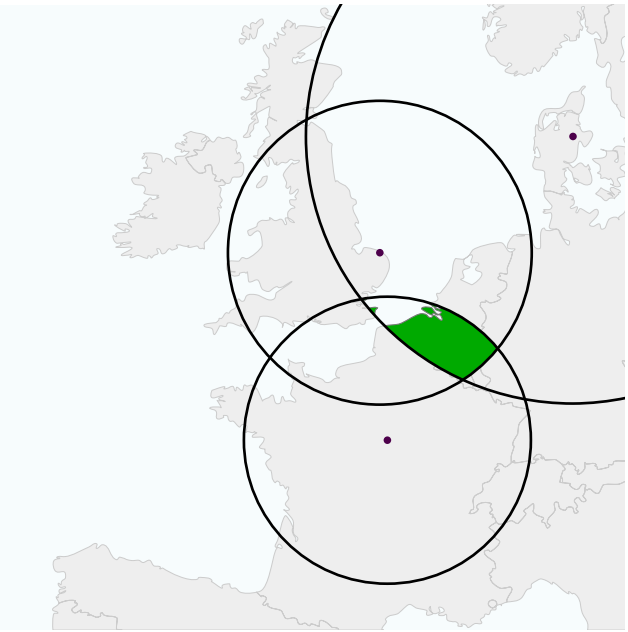


Figure 2.1: The principle of multilateration. If something is within 500 km of Bourges, 500 km of Cromer, and 800 km of Randers, then it is in Belgium (roughly).

## 2.1 Background

Existing methods for finding the physical location of Internet hosts can be divided into two general classes. Passive methods collect location information from regional Internet registries, location information encoded in router hostnames, and private consultation with individual ISPs [36], and produce a database mapping IP addresses to locations. These databases are widely used, but notorious for their errors [75, 143, 158], some of which are significant enough that they make the news [87].

Active methods, on the other hand, rely on measurements of packet round-trip time between a *target* host, which is to be located, and a number of *landmark* hosts, which are in known locations. The simplest active method is to guess that the target is in the same place as the landmark with the shortest round-trip time [38, 137, 202]. This breaks down when the target is not near any of the landmarks. The next step up in complexity is to estimate, for each landmark, the maximum distance that a packet could have traveled in the time measured, and draw disks on a map, bounded by these distances. The target must be in the region where the disks all intersect. This process is called *multilateration*. Figure 2.1 shows an example: measurements taken from Bourges in France, Cromer in the UK, and Randers in Denmark produce an intersection region roughly covering Belgium.

The central problem for network multilateration is that network packets do not travel in straight lines. Cables are laid on practical paths, not great circles. Network routes are optimized for

bandwidth rather than latency, leading to "circuitous" detours that can add thousands of kilometers to the distance traveled [115, 118, 124]. Intermediate routers can add unbounded delays [119]. Distance and delay do still correlate, but not in a straightforward way. Much research on active methods focuses on increasingly sophisticated models of the delay-distance relationship [14, 56, 64, 81, 117, 118, 124, 137, 190]. One common refinement is to assume a minimum travel distance for any given delay, as well as a maximum.

**Challenges of global geolocation** When both landmarks and targets are in the same subcontinental region, sophisticated models improve accuracy—if that region is Europe or the USA. On the other hand, for China, several papers report that *simple* models are more accurate [38, 53, 119]. They propose that simple models are more robust in the face of severe congestion. Li *et al.* [119] specifically points out that a minimum travel distance assumption is invalid in the face of large queuing delays at intermediate routers. A second possibility is that sophisticated models are more reliable when there are more possible paths between landmarks and targets, as is the case in Europe and North America, but not China [66]. A third is that models tested on PlanetLab nodes [142] gain an unfair advantage due to the generally better connectivity enjoyed by academic networks. In Section 2.4, we test four algorithms, covering a range of model complexity, on hosts crowdsourced from all over the world. We also find that simple models are more effective, overall, and our data is more consistent with the congestion explanation.

Increasing the number of landmarks improves accuracy but also slows down the measurement process, since all of the landmarks must send packets to the target and wait for replies (or vice versa). If they all do this simultaneously, they may create enough extra network congestion to invalidate the measurement [91]. Several researchers have observed that landmarks far away from the target are less useful, and proposed a two-stage process, in which a small number of widely dispersed landmarks identify the subcontinental region where the target lies, and then a larger group of landmarks within that region pin down the target's position more accurately [53, 93, 105, 202].

**Challenges of geolocating proxies** Less than ten percent of the proxies we are interested in testing will respond to pings, and we do not have the ability to run measurement programs on the proxies themselves. We can only send packets *through* the proxies, which means the apparent round-trip time to each landmark is the sum of the round-trip time from the proxy to the landmark, and the round-trip time from our measurement client to the proxy. This is similar to the problem faced by Castelluccia *et al.* [31] when attempting to geolocate botnet command-and-control servers, and we adopt the same solution, as discussed further in Section 2.4.3.

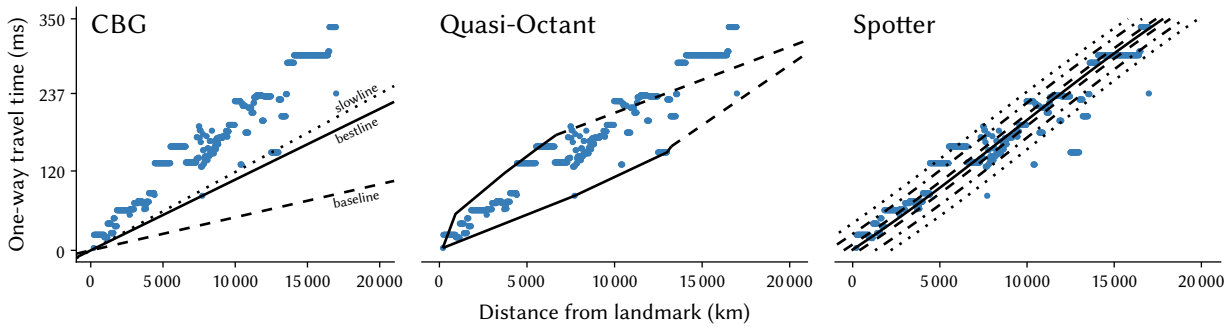



Figure 2.2: Example calibration scatter plots for CBG, (Quasi-)Octant, and Spotter.

## 2.2 Algorithm selection

Since the proxies we are investigating could be spread all over the world, we must find an active geolocation algorithm that will work at the scale of the whole world. We reimplemented four active geolocation algorithms from earlier papers: CBG [81], Octant [190], Spotter [117], and an Octant/Spotter hybrid of our own invention. We did not have access to the original implementations, and we had to fill in gaps in all their published descriptions. All the software we developed for this project is open-source and available online.[1]

Eriksson *et al.* [65] recommend considering external facts about where a server could plausibly be, such as "on land, and not in Antarctica." We take this advice and exclude all terrain north of 85° N and south of 60° S from the final prediction region for each target and algorithm. Using the 2012 Natural Earth [139] map of the world, we also exclude oceans and lakes. We do not, however, exclude any islands, no matter how small or remote, because some of the proxy providers do claim to have servers on remote islands (e.g. Pitcairn).

### 2.2.1 Constraint-Based Geolocation

Constraint-Based Geolocation (CBG) is one of the oldest and simplest multilateration algorithms. It uses a linear model for the delay-distance relationship, limited by a "baseline" speed of 200 km/ms, or $\frac{2}{3}c$, which is approximately how fast signals propagate in fiber-optic cable. For each landmark, CBG computes a "bestline" from the calibration data, which is as close as possible to all of the data points on a scatter plot of delay as a function of distance, while remaining below all of them, and above the baseline. This will be a speed *slower* than 200 km/ms, and will therefore give a *smaller* estimate of how far a packet could have gone in a given time. Each landmark's bestline gives the maximum distance for a round-trip measurement to that landmark.

[1] https://github.com/zackw/active-geolocator

The left panel of Figure 2.2 shows an example calibration for CBG. The blue dots are round-trip time measurements taken by one RIPE anchor. The bestline (solid) is above the baseline (dotted); it passes through exactly two data points, with all the others above. It corresponds to a speed of 93.5 km/ms—less than half the theoretical maximum. The "slowline" will be explained in Section 2.4.1.

### 2.2.2 Quasi-Octant

Octant elaborates on CBG in two ways. First, it estimates both the maximum and the minimum distance to each landmark, and draws rings on the map, not disks. Second, Octant uses piecewise-linear curves for both distance models. These are defined by the convex hull of the scatter plot of delay as a function of distance, up to 50 % and 75 % of all round-trip times, respectively. Observations beyond those cutoffs are considered unreliable, so Octant uses fixed empirical speed estimates for longer round-trip times. The middle panel of Figure 2.2 shows an example Octant calibration, with the same data as the CBG calibration to its left. The convex hull is drawn with solid lines and the fixed empirical speeds with dashed lines.

Octant includes features that depend on route traces, such as a "height" factor to eliminate the effect of a slow first hop from any given landmark. Since we cannot collect route traces (see Section 2.3.2), these have been omitted from our re-implementation, and we call it "Quasi-Octant" to denote that change.

### 2.2.3 Spotter

Spotter [117] uses an even more elaborate delay-distance model. It computes the mean and standard deviation of landmark-landmark distance as a function of delay, and fits "a polynomial" to both. Unlike CBG and Octant, a single fit is used for all landmarks. The paper does not specify the degree of the polynomial, or the curve-fitting procedure; we use cubic polynomials, fit by least squares, and constrain each curve to be increasing everywhere (anything more flexible led to severe overfitting in pilot tests).

Spotter also uses a probabilistic multilateration method. It estimates the distance from each landmark to the target as a Gaussian distribution, with mean $\mu$ and standard deviation $\sigma$ given by the fitted curves. This produces a ring-shaped probability distribution over the surface of the Earth; the rings for each landmark are combined using Bayes' Rule to form the final prediction region.

The right panel of Figure 2.2 shows an example Spotter calibration. The solid line is the best cubic fit for the mean $\mu$ of the distance-delay relationship; dashed, dash-dot, and dotted lines are drawn at $\mu \pm \sigma$, $\mu \pm 3\sigma$, and $\mu \pm 5\sigma$ respectively.

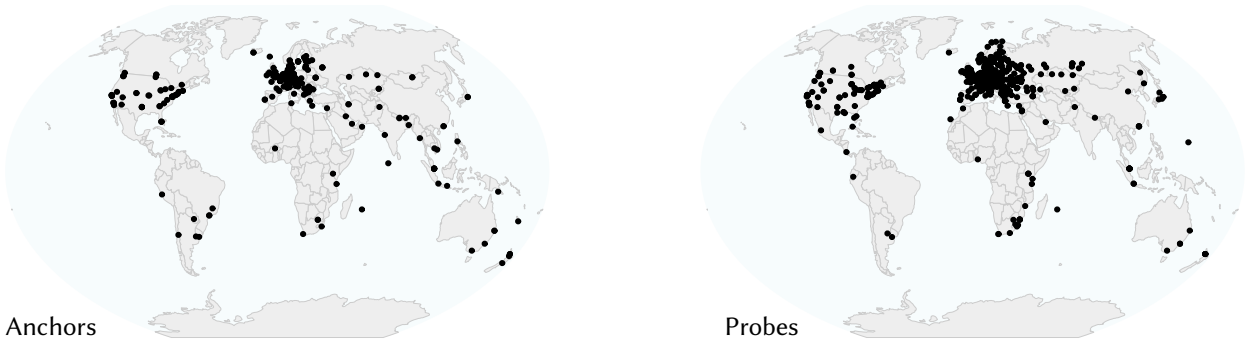



Figure 2.3: Locations of the RIPE Atlas anchors (left) and probes with stable IPv4 addresses, as of April 2018.

### 2.2.4 Quasi-Octant/Spotter hybrid

To separate the effect of Spotter's probabilistic multilateration from the effect of its cubic-polynomial delay model, we also implemented a hybrid that uses Spotter's delay model, but Quasi-Octant's ring-based multilateration. The minimum and maximum radii of the ring are set to $\mu - 5\sigma$ and $\mu + 5\sigma$, respectively.

## 2.3 Measurement method

The experiments described in this chapter ran from July 2016 through April 2018.

For all our experiments, we used the "anchor" hosts of RIPE Atlas [151] as landmarks. RIPE Atlas is a worldwide constellation of hosts dedicated to Internet measurement, composed of "probes" and "anchors;" there are fewer anchors, but they are more convenient for use as landmarks. They are reliably available 24/7, their documented locations are accurate, and they all continuously ping each other and upload the round-trip times (RTT) to a publicly accessible database. At the beginning of the experiment, there were 207 usable anchors; by the end of it, 12 had been decommissioned and another 61 added. Figure 2.3 (left side) shows all the anchors' locations. The majority are in Europe; North America is also well-represented. While there are fewer anchors in Asia and South America, and only a few in Africa, their geographic distribution is adequate—the most difficult case for active geolocation is when all of the landmarks are far away from the target, in the same direction [66, 124].

### 2.3.1 Two-phase measurement

It takes several minutes to ping all 250 of the anchors. Landmarks far from the target do not contribute much useful information, as we will discuss further in Section 2.4.2. We speed up

the process with a two-phase measurement, as proposed by Khan, Naveed, and Cottrell [105] and others [53, 93, 202]. We first measure RTTs to three anchors per continent, and use these measurements to deduce which continent the target is on. We then randomly select and measure RTTs to 25 more landmarks on that continent, from a list including all of the anchors, plus all the probes that have been online for the past 30 days with a stable IPv4 address. These probes are shown in Figure 2.3 (right side).

Random selection of landmarks in the second phase spreads out the load of our measurements, reducing their impact on concurrent experiments [91]. Using stable probes as well as anchors spreads the load even in parts of the world where there are few anchors.

We maintain a server that retrieves the list of anchors and probes from RIPE's database every day, selects the probes to be used as landmarks, and updates a delay-distance model for each landmark, based on the most recent two weeks of ping measurements available from RIPE's database. Our measurement tools retrieve the set of landmarks to use for each phase from this server, and report their measurements back to it. Some of the landmarks have both IPv4 and IPv6 addresses, but the commercial proxy servers we are studying offer only IPv4 connectivity, so the server resolves the landmarks' hostnames itself and sends only IPv4 addresses to the tools.

### 2.3.2 Measurement tools

Commercial proxy providers aggressively filter traffic through their proxies. Of the VPN servers we tested, roughly 90 % ignore ICMP ping requests. Similarly, 90 % of the default gateways for VPN tunnels (i.e. the first-hop routers for the VPN servers) ignore ping requests and do not send time-exceeded packets, which means we cannot see them in a `traceroute` either. Roughly a third of the servers discard *all* time-exceeded packets, so it is not possible to `traceroute` through them at all. Some servers even drop UDP and TCP packets with unusual destination port numbers.

In short, the only type of network message we can reliably use to measure round-trip time is a TCP connection on a commonly used port, e.g. 80 (HTTP). We implemented two measurement tools that use this method to measure round-trip times to each landmark.

**Command-line** For measurements of VPN proxies' locations (Section 2.5), we used a standalone program, written in Python and C. It can take measurements either directly or through a proxy, and it can process a list of proxies in one batch.

This tool uses the POSIX sockets API to make TCP connections. It measures the time for the `connect` primitive to succeed or report "connection refused," and then closes the connection without sending or receiving any data. We verified that `connect` consistently returns as soon as the second packet of the TCP three-way handshake arrives (i.e. after a single round-trip to a landmark) on both Linux and NetBSD. (Linux was used as the client OS for all the measurements of VPN

proxies; some pilot tests involved clients running NetBSD.) If a connection fails with an error code other than "connection refused," the measurement is discarded. "Network unreachable" errors, for instance, originate from intermediate routers, so they do not reflect the full round-trip time.

**Web-based** For algorithm validation (Section 2.4) we crowdsourced hosts in known locations from around the world. We could not expect volunteers from the general public, or Mechanical Turk workers, to download, compile, and run a command-line tool, so we implemented a second measurement tool as a Web application. Anyone can access the website hosting this application,[2] and it requires no "plug-ins" or software installation. It presents a live demonstration of active geolocation, displaying the measurements as circles drawn on a map, much as in Figure 2.1. After this demonstration is complete, it offers an explanation of the process, and invites the user to upload the measurements to our database, if they are willing to report their physical location.

The price of user-friendliness is technical limitations. Web applications are only allowed to send well-formed HTTP(S) messages; we cannot close connections immediately upon establishment, without sending or receiving any data, as the command-line tool does.

In principle, web applications are not allowed to communicate with arbitrary hosts, only with the server hosting their website [95]. However, this rule has a loophole. When a web application attempts to communicate with a server that is not hosting its website, the browser will send an HTTP request, but will not return a successful response unless the server allows it, using special HTTP response headers. Errors are still reported. Since we only care about the time to connect, we make a request that we know will fail, and measure the time to failure. Ideally, we would connect to a TCP port that was not blacklisted by any VPN provider, and was closed (not blackholed) on all the RIPE Atlas nodes we use, but there is no such port.

Instead, the web application makes encrypted (HTTPS) connections to the usual TCP port for *unencrypted* HTTP (80). This will fail after one round-trip if the landmark is not listening on port 80. However, if it is listening on port 80, the browser will reply to the SYN-ACK with a TLS ClientHello packet. This will trigger a protocol error, and the browser will report failure, but only after a second round-trip. Thus, depending on whether the landmark is listening on port 80 (which depends on the version of the RIPE Atlas node software it is running; we cannot tell in advance) the web application will measure the time for either one or two round-trips, and we cannot tell which.

### 2.3.3 Tool validation

Figure 2.4 shows the abstract difference in the network traffic generated by the two tools. Figure 2.5 compares the round-trip times measured by the command-line tool and the web application running under two different browsers, all three on a computer in a known location running Linux, to a

[2] https://research.owlfolio.org/active-geo

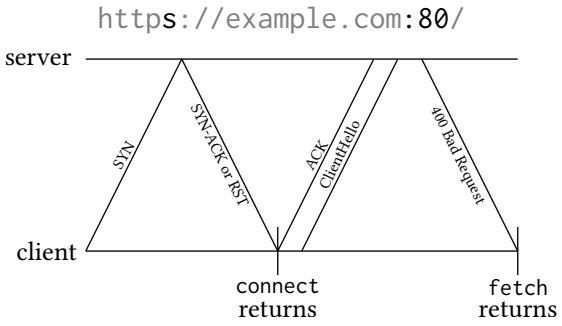



Figure 2.4: Both measurement tools make TCP connections to port 80 on each landmark. The CLI tool can use the low-level `connect` API; the web application must use the higher-level `fetch` API. We instruct `fetch` to send encrypted (HTTPS) traffic to the usual port for *unencrypted* HTTP, forcing a protocol error. The CLI tool reliably measures one round-trip time; the web application measures one or two round-trips, depending on whether the landmark is listening on port 80.

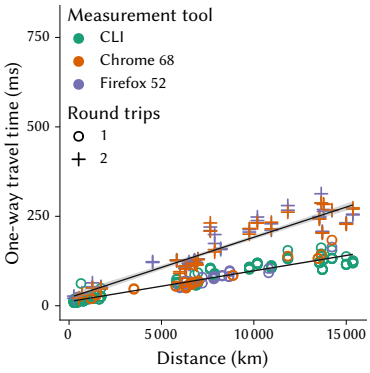



Figure 2.5: Comparison of the CLI geolocation tool with the web application in two browsers, all running on Linux.

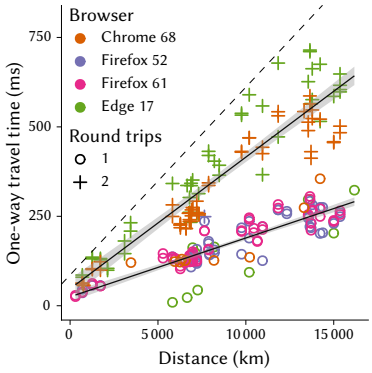



Figure 2.6: Comparison of four browsers running on Windows 10.

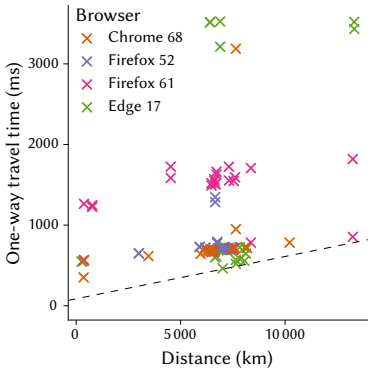



Figure 2.7: High outliers removed from Figure 2.6. The dashed line has the same slope in both figures.

collection of landmarks as described in Section 2.3.1. We manually partitioned the measurements taken by the web application into groups suspected to be one round trip and two round trips, and estimated the distance-delay relationship for each by linear regression, shown with black lines and gray 95 % confidence intervals. The slope of the two-round-trip line is 1.96 times the slope of the one-round-trip line; adjusted $R^2$ (considering both lines as a single model) is 0.9942. After accounting for the effects of distance and whether we measured one or two round trips, ANOVA finds no significant difference among the three tools (two additional degrees of freedom, $F = 0.8262$, $p = 0.44$) which is a testament to the efficiency of modern JavaScript interpreters.

Figure 2.6 compares the round-trip times measured by the web application running under four additional browsers, on the same computer that was used for Figure 2.5, but running Windows 10 instead. (The command-line tool has not been ported to Windows.) Measurements on Windows are much noisier than on Linux. We can still distinguish a group of one-round-trip data points and a group of two-round-trip data points, but there is a third group, "high outliers," separately shown in Figure 2.7 so that Figures 2.5 and 2.6 can have the same vertical scale. The diagonal dashed line on Figures 2.6 and 2.7 has the same absolute slope. The high outlier measurements are much slower than can be attributed to even two round-trips, and their values are primarily dependent on the browser they were measured with, rather than the distance.

Excluding the high outliers, the remaining data points for Windows can also be modeled by a division into groups for one or two round-trips, but not as cleanly as on Linux. The ratio of slopes is 2.29, adjusted $R^2 = 0.8983$, and ANOVA finds the model is significantly improved by considering the browser as well (three more degrees of freedom, $F = 13.11$, $p = 6.1 \times 10^{-8}$). Equally concerning, if we combine the two models, we find that the operating system has a significant effect on the slopes of the lines (four additional degrees of freedom, $F = 693.56$, $p < 2.2 \times 10^{-16}$) and the regression line for two round-trips measured on Linux ($t = 0.03375d + 45.52$, distance in km, time in ms) is about the same as the line for *one* round-trip measured on Windows ($t = 0.03288d + 49.92$).

In section 2.4, we will speak further of how these limitations affect our assessment of which algorithm is most suitable for estimating the location of a proxy that could be anywhere in the world.

## 2.4 Algorithm testing

In order to test our geolocation algorithms on hosts they had not been calibrated with, we crowd-sourced a second set of hosts in known locations.[3] 40 volunteers, recruited from a variety of mailing lists and online forums, and another 150 paid contributors, recruited via Mechanical Turk for 25¢ each, provided us with the approximate physical location of their computers (rounded to

[3]This study was approved by our university's IRB.

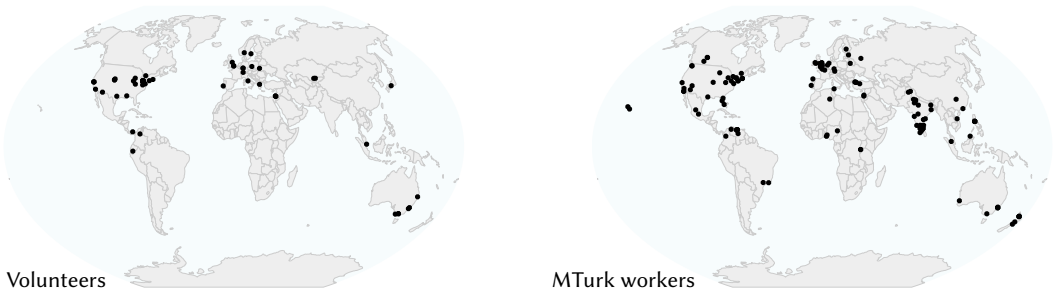



Figure 2.8: Locations of the crowdsourced hosts used for algorithm validation, with volunteers on the left and Mechanical Turk workers on the right.

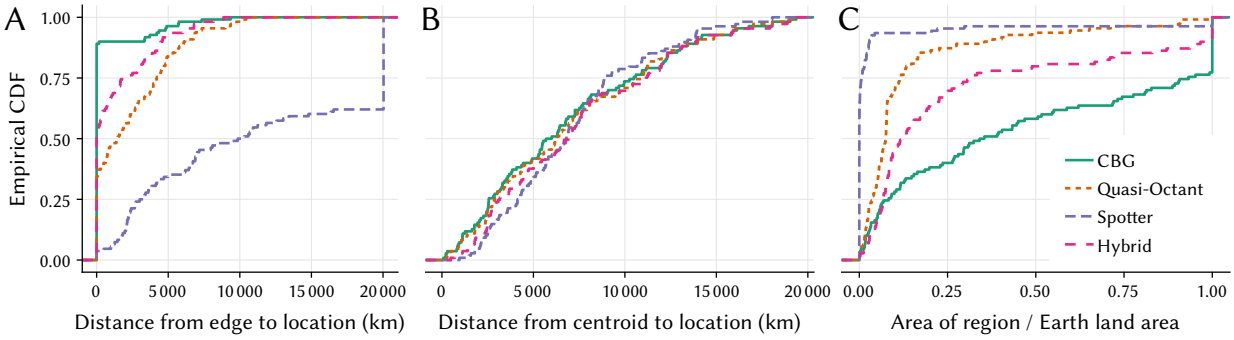



Figure 2.9: Precision of predicted regions for crowdsourced test hosts.

two decimal places in latitude and longitude, or roughly 10 km of position uncertainty) and a set of round-trip times to RIPE Atlas anchors and probes, using the Web-based measurement tool described in Section 2.3.2. Their self-reported locations are shown in Figure 2.8. Like the RIPE anchors, the majority are in Europe and North America, but we have enough contributors elsewhere for statistics.

Our priority was to find an algorithm that would always include the true location of each host in its predicted region, even if this meant the region was fairly imprecise. To put it another way, when investigating the locations of commercial proxies, we want to be certain that the proxy is where we say it is, even if that means we cannot assure that it is not where the provider says it is.

In figure 2.9, panel A, we plot an empirical CDF of how far outside the predicted region each true location is, for each of the four algorithms. This is a direct measure of each algorithm's failure to live up to the above requirement. None of the algorithms are perfect, but CBG does better than the other three, producing predictions that do include the true location for 90 % of the test hosts, and are off by less than 5 000 km for 97 % of them. Hybrid and Quasi-Octant's predictions miss the mark for roughly 50 % of the test hosts, but they are off by less than 5 000 km for roughly 90 %. Fully half of Spotter's predictions are off by more than 10 000 km.

In panels B and C of Figure 2.9, we look into *why* the predictions miss the true region. Panel B

shows that the distances from the *centroid* of each algorithm's predictions, to the true locations, are about the same for all four algorithms, and panel C shows that CBG produces predictions that are much larger than the other three. We conclude that none of the algorithms can reliably center their predicted region on the true location, but CBG's predictions are usually big enough to cover the true location anyway, whereas the other three algorithms' predictions are not big enough.

Why should CBG be so much more effective? Looking again at the calibration data in Figure 2.2, we observe that most of the data points are well above CBG's bestline. Quasi-Octant and Spotter draw more information from these points than CBG does. If most of those points are dominated by queuing and circuitousness delays, rather than the great-circle distance between pairs of landmarks, that would lead Quasi-Octant and especially Spotter to underestimate the speed packets can travel, therefore predicting regions that are too small. Large queuing delays also invalidate the assumption, shared by both Quasi-Octant and Spotter, that there is a *minimum* speed packets can travel [119].

Most of our crowdsourced contributors used the web application under Windows. As we described in Section 2.3.3, this introduces extra noise and "high outliers" into the measurements. CBG has an inherent advantage in dealing with measurements biased upward, since it always discards all but the quickest observation for each landmark, its bestlines are the fastest travel time consistent with the data, and it does not assume any minimum travel speed when multilaterating. Crowdsourced measurements using only the command-line tool might have allowed Quasi-Octant and Spotter to do better. However, measurements taken through proxies are liable to suffer extra noise and queuing delays as well. We could thus argue that the web application's limitations make the crowdsourced test a better simulation of the challenges faced by active geolocation of proxies.

### 2.4.1 Eliminating underestimation: CBG++

Regardless of the reasons, CBG clearly is the most effective algorithm in our testing, but it still does not always cover the true location with its predictions. We made two modifications in order to eliminate this flaw, producing a new algorithm we call CBG++.

CBG's disks can only fail to cover the true location of the target if some of them are too small. A disk being too small means the corresponding bestline underestimates the distance that packets could travel. This can easily happen, for instance, when the network near a landmark was congested during calibration [113]. Not only can an underestimate make the prediction miss the target, it can make the intersection of all the disks be empty, meaning that the algorithm fails to predict *any* location for the target.

To reduce the incidence of underestimation, we first introduced another physical plausibility constraint. CBG's bestlines are constrained to make travel-speed estimates no faster than 200 km/ms as packets can travel no faster than this in undersea cables. We also constrain them to make travel-

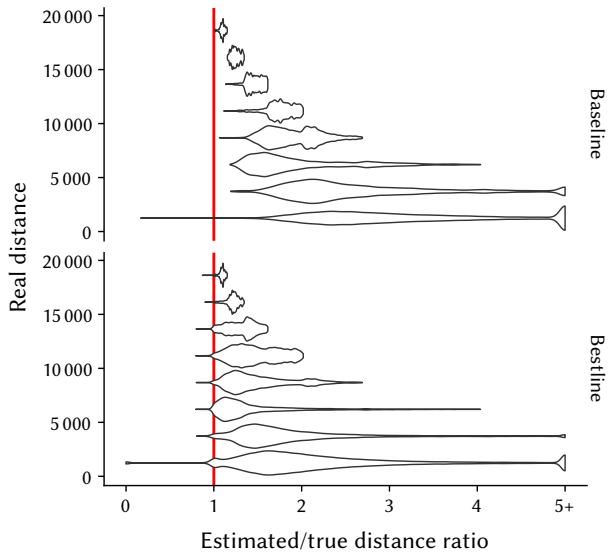



Figure 2.10: CBG bestline and baseline estimates compared to the true distance.

speed estimates no *slower* than 84.5 km/ms; this is the "slowline" in the CBG panel of Figure 2.2. The logic behind this number is: No landmark can be farther than half the equatorial circumference of the Earth, 20 037.508 km, from the target. One-way travel times greater than 237 ms could have involved a geostationary communications satellite, and one such hop can bridge any two points on the same hemisphere, so they provide no useful information. 20 037.508 km/237 ms = 84.5 km/ms.

The slowline constraint is not enough by itself. Figure 2.10 shows the distribution of ratios of bestline and baseline distance estimates to the true distances, for all pairs of landmarks, with the slowline constraint applied. We use the landmarks themselves for this analysis, rather than the crowdsourced test hosts, because we know their positions more precisely and the ping-time measurements they make themselves are also more accurate. A small fraction of all bestline estimates are still too short, and for very short distances this can happen for baseline estimates as well.

We weed out the remaining underestimates with a more sophisticated multilateration process. For each landmark, we compute both the bestline disk, and a larger disk using the baseline. We find the largest subset of all the baseline disks whose intersection is nonempty; this is called the "baseline region." Any bestline disk that does not overlap this region is discarded. Finally we find the largest subset of the remaining bestline disks whose intersection is nonempty; this is the "bestline region." These subsets can be found efficiently by depth-first search on the powerset of the disks, organized into a suffix tree. Retesting on the crowdsourced test hosts, we found that this algorithm eliminated all of the remaining cases where the predicted region did not cover the true location.

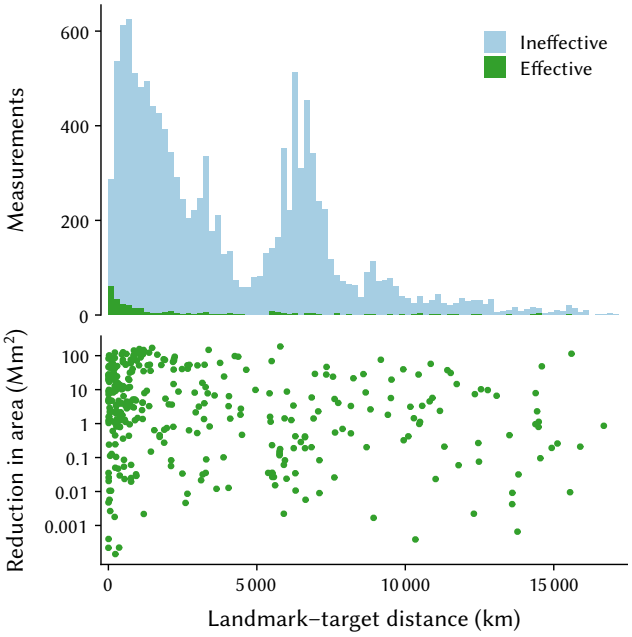



Figure 2.11: Proportion of geolocation measurements that had an effect on the final prediction region, as a function of distance between landmark and target; for effective measurements, the amount by which they reduced the size of the final region. The total land area of Earth is roughly 150 square megameters ($Mm^2$), and the land area of Egypt is roughly 1 $Mm^2$.

### 2.4.2 Effectiveness of landmarks

To check the observations of Khan, Naveed, and Cottrell [105] and others, that landmarks closer to the target are more useful, we measured the round-trip time between all 250 RIPE Atlas anchors and the target for all of the crowdsourced test hosts. A large majority of all measurements lead to disks that radically overestimate the possible distance between landmark and target. Multilateration produces the same final prediction region even if these overestimates are discarded. We call these measurements *ineffective*. As shown in Figure 2.11, effective measurements are more likely to come from landmarks close to the target, but among the effective measurements, there is no correlation between distance and the amount by which the measurement reduced the size of the final prediction. This is because a distant landmark may still have only a small overlap with the final prediction region, if it is distant in just the right direction.

### 2.4.3 Adaptations for proxies

When taking measurements through a network proxy, each measured round-trip time is the sum of the RTT from the client to the proxy, and the RTT from the proxy to the landmark. To locate the proxy, we need to measure and subtract the RTT from the client to the proxy. We cannot measure this directly, because the proxy services usually configure their hosts not to respond to ICMP ping

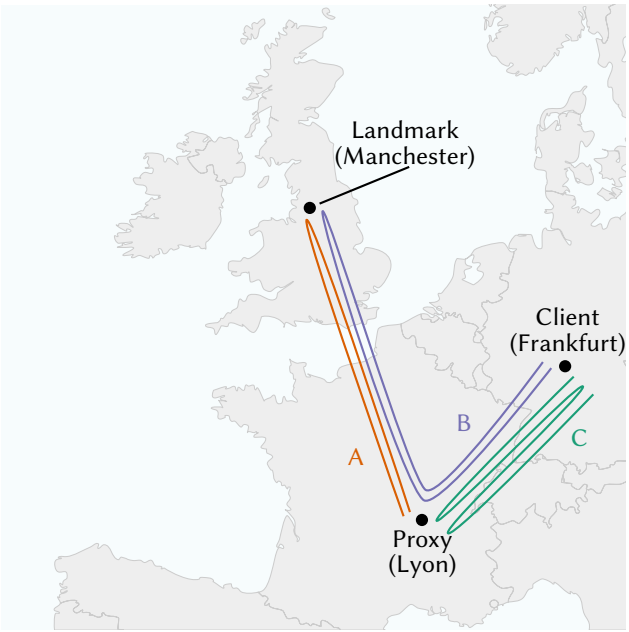



Figure 2.12: The RTT from a proxy to a landmark, A, must be derived from the RTT through a proxy to a landmark, B, and the RTT through a proxy back to the client, C: $A = B - \eta C$.

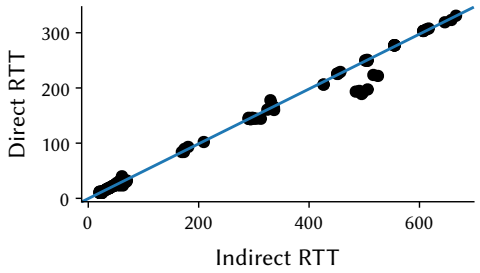



Figure 2.13: The relationship between direct and indirect round-trip times, $\eta$, is almost exactly 1/2.

packets, and aggressively rate-limit incoming TCP connections.

Instead, we take inspiration from Castelluccia *et al.* [31] and have the client ping *itself*, through the VPN, as illustrated in Figure 2.12. This should take slightly more than twice as long as a direct ping. Figure 2.13 shows the relationship between direct and indirect pings for all of the proxies in the study that can be pinged both ways. The blue line is a robust linear regression, whose slope $\eta$ is the inverse of the `RTT_factor` described by Castelluccia *et al.* In our case, the slope is 0.49 with $R^2 > 0.99$.

## 2.5 Locating VPN proxies

We used the two-phase, proxy-adapted CBG++ to test the locations of proxies from seven VPN providers. This paper's purpose is not to call out any specific provider for false advertising, so we are not naming the seven providers that we tested; however, figure 2.14 shows their rankings by number of countries and dependencies claimed, with 150 of their competitors for comparison. Providers A through E are among the 20 that make the broadest claims, while F and G make more modest and typical claims. Notice that providers who claim only a few locations, tend to claim more or less the same locations; this is what one would expect if it were much easier to lease space in a data center in some countries than others.

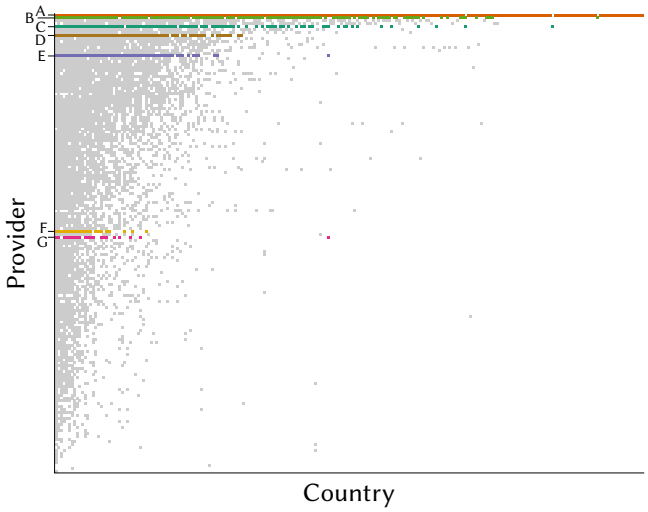



Figure 2.14: The countries where 157 VPN providers claim to have proxies. Providers included in this study are colored and labeled. Data provided by VPN.com [74].

All of the VPN providers we tested use round-robin DNS for load balancing; to avoid the possibility of unstable measurements, we looked up all of the server hostnames in advance, from the same host that would run the command-line measurement tool, and tested each IP address separately. We used a single client host for all of the measurements, located in Frankfurt, Germany. Because of this, we cannot say whether the VPN providers might be using DNS geotargeting or anycast routing to direct clients in different parts of the world to different servers. In total, we tested 2 269 unique server IP addresses, allegedly distributed over 222 countries and territories.

None of the providers advertise exact locations for their proxies. At best they name a city, but often they only name a country. City claims sometimes contradict themselves; for instance, we observed a configuration file named "`usa.new-york-city.cfg`" directing the VPN client to contact a server named "`chicago.vpn-provider.example`." Therefore, we only evaluate country-level claims.

CBG++ tells us only that a proxy is within some region. If that region is big enough to cover more than one country, we cannot be certain where the server really is. However, we might still be certain that it is *not* where the proxy provider said it was; for instance, a predicted region that covers Canada and the USA still rules out the entire rest of the world. We say that the provider's claim for a proxy is *false* if the predicted region does not cover any part of the claimed country. We say that it is *credible* if the predicted region is entirely within the claimed country, and we say that it is *uncertain* if the predicted region covers both the claimed country and others. For false and uncertain claims, we also checked whether any of the countries covered by the prediction region were on the same continent as the claimed country.

Some uncertain predictions can be resolved by referring to a list of known locations of data

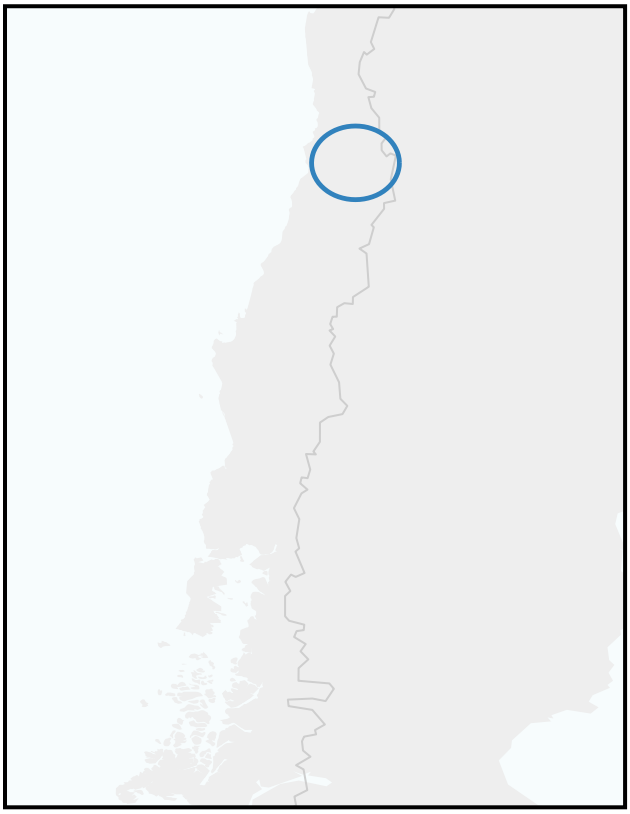


Figure 2.15: Disambiguation by data center locations: the only data centers in this region are in Chile, not Argentina.

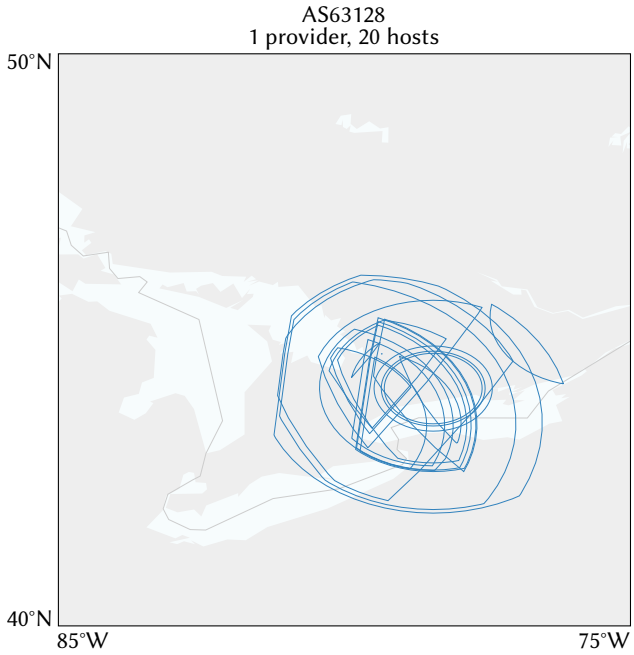



Figure 2.16: Disambiguation by metadata: all these hosts belong to the same provider, the same AS, and the same /24, so they are likely to be in the same physical location.

centers, such as the one maintained by the University of Wisconsin [172]. For example, the prediction shown in Figure 2.15 is uncertain because it covers Argentina as well as Chile. However, the only data centers within the region are in Chile, so we can conclude that this server is in Chile. When data center locations are not enough, cross-checking with network metadata may help. For example, in Figure 2.16, the largest of the 20 predicted regions cover data centers on both sides of the USA-Canada border, but all of the hosts share a provider, an autonomous system (AS), and a 24-bit network address, which means they are practically certain to be in the *same* data center. Since all of the regions cover part of Canada, but only some of them cross into the USA, we ascribe all of these hosts to Canada. Overall, these techniques allow us to reclassify 353 uncertain predictions as credible or not-credible.

Putting it all together, we find that the claimed location is credible for 989 of the 2 269 IP addresses, uncertain for 642, and false for 638. For 401 of the false addresses, the true location is not even on the same continent as the claimed location; however, for 462 of the uncertain addresses, the true location *is* somewhere on the same continent as the claimed location. (See Section 2.6 for how we defined continental boundaries, and some discussion of which countries and continents are most likely to be confused.)

Figure 2.17 shows which countries, overall, are more likely to host credibly-advertised proxies, and where the servers for the false claims actually are. The ten countries with the largest number of claimed proxies account for 84 % of the credible cases, and only 11 % of the false cases. (Uncertain cases are nearly evenly split between the top ten and the remainder.) False claims are spread over the "long tail" of countries, with only a few advertised servers each. Figure 2.17 also shows the overall effect of using data center and AS information to disambiguate predictions. It is particularly effective when the prediction region crosses continents; 55 % of those cases were completely resolved for our purposes. Only 23 % of the regions covering multiple countries within the same continent could be disambiguated.

Figure 2.18 shows another perspective on the same observation, by relating credibility to the country ranking in Figure 2.14. The credible claims are concentrated in the countries where many other VPN providers also claim to host proxies. This is evidence for our original intuition that proxies are likely to be hosted in countries where server hosting is easy to acquire.

We might also like to know if some providers are more reliable than others. Figure 2.19 shows, for each provider, a map of the world with each country color-coded according to the overall honesty of the provider's claims for that country. If a country is drawn in white, the provider did not claim to have any proxies in that country to begin with. Bright green means all of the claimed proxies' CBG++ predictions overlap the country at least somewhat—that is, the "yes" or "uncertain" categories in Figure 2.17, after taking data center locations into account. Dark purple means none of the predictions overlap the country at all. Colors in between mean CBG++ backs up the claim

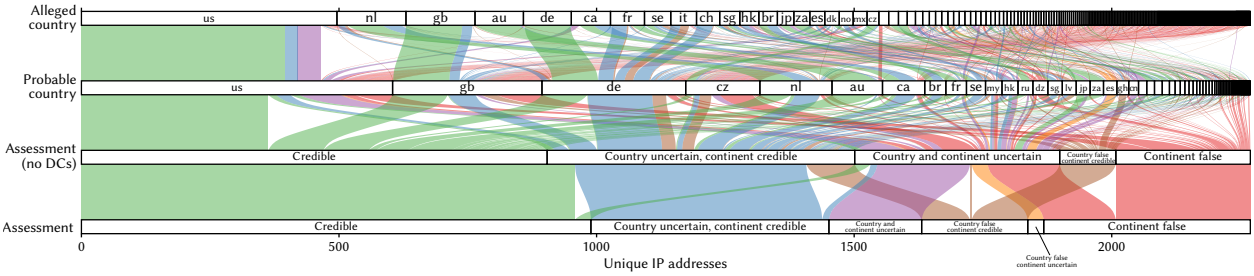



Figure 2.17: Overall credibility of VPN providers' claims to have proxies in specific countries.

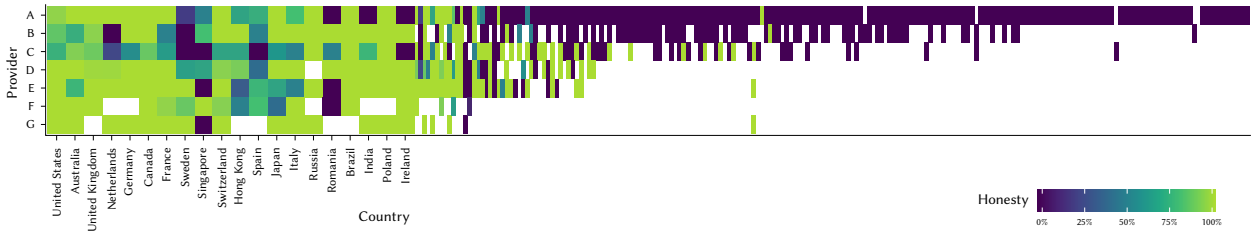



Figure 2.18: Credible claims are concentrated in the most commonly claimed countries.

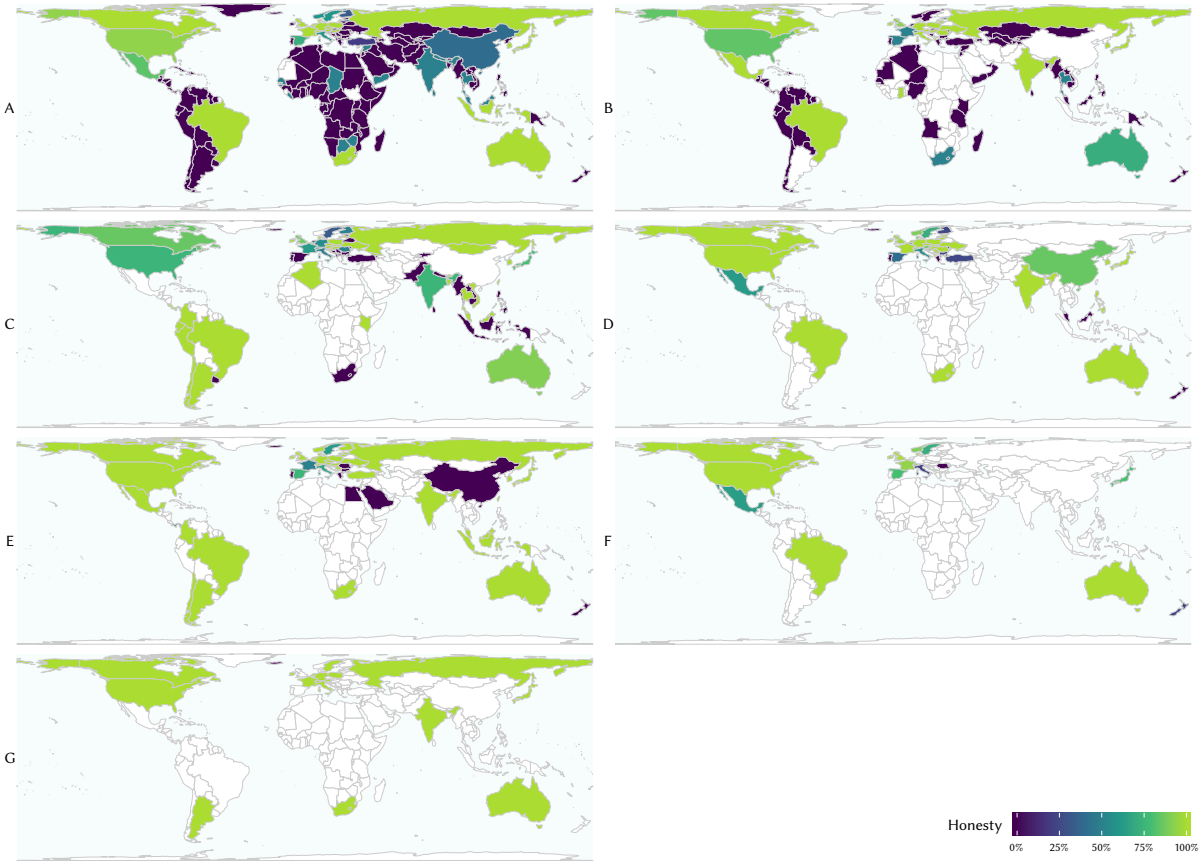



Figure 2.19: The credibility of each provider's claims for specific countries.

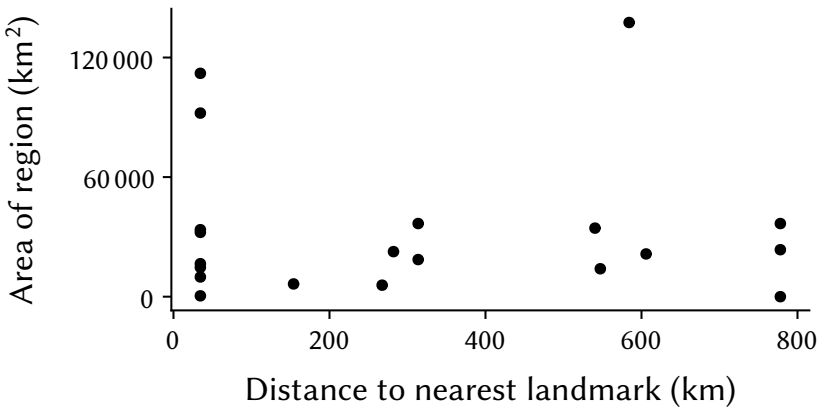



Figure 2.20: For AS63128, the size of the prediction region is not correlated with the distance to the nearest landmark.

for some but not all of the proxies claimed to be in that country.

There is some variation among the providers; for instance, C and E are actually hosting servers in more than one country of South America, whereas providers A and B just say they are. However, claimed locations in countries where server hosting is difficult are almost always false. Even in regions like Western Europe, where hosting *is* available in any country one would like, providers seem to prefer to concentrate their hosts in a few locations.

### 2.5.1 Data centers and prediction error

Groups of proxies that we believe are all in the same data center can be used for another check on the accuracy of our geolocation methods. We are not yet certain enough of our data center groups to run this analysis on all of the grouped proxies, but we can discuss the results for a clear-cut case like AS63128. If geolocation worked perfectly, all of the regions shown in Figure 2.16 ought to be the same, but clearly they are not, and there is no single sub-region that they all cover. Since our two-phase procedure uses a different randomly-selected group of landmarks for each measurement, variation is to be expected. Figure 2.20 shows that there is no correlation between the size of each prediction region, and the distance to the nearest landmark for that region from the centroid of all the predictions taken together. This means the variation is not simply due to geographic distance; it may instead be due to some landmarks experiencing more congestion or routing detours than others.

### 2.5.2 Comparison with ICLab and IP-to-location databases

ICLab had previously been using a simple "sanity check" for proxy locations. It only attempts to prove that each proxy is *not* in the claimed country. It assumes that it is impossible for a packet to have traveled faster than a configurable speed limit; their actual tests used 153 km/ms (0.510 4 $c$)

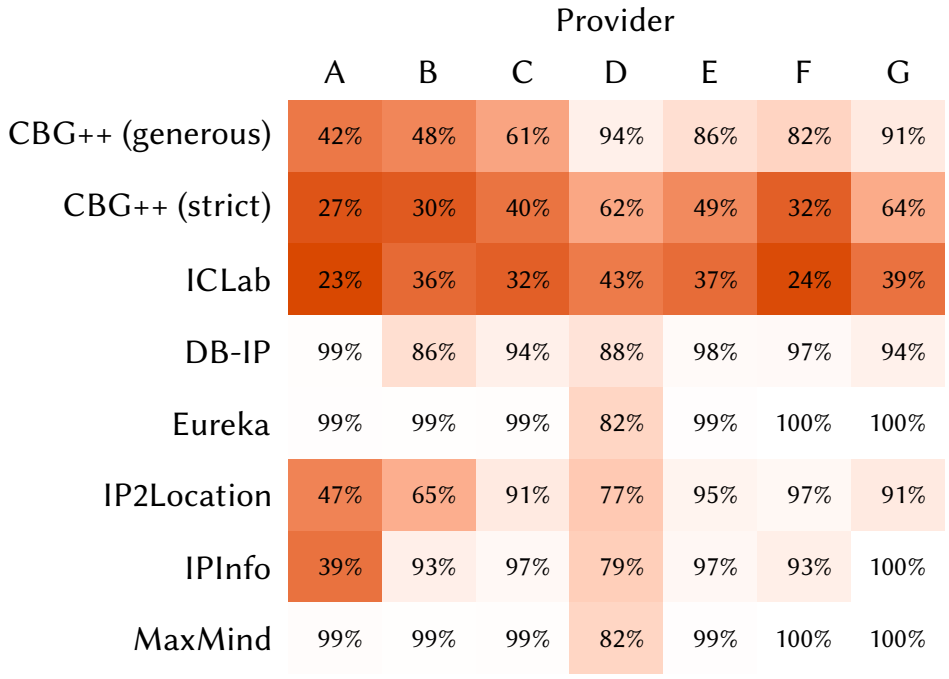



Figure 2.21: The percentage of each provider's proxies for which our validation (two different ways), ICLab's validation, and five popular geolocation databases agree with the advertised location.

for this limit (slightly faster than the "speed of internet" described in Katz-Bassett *et al.* [99]). Given a country where a host is claimed to be, and a set of round-trip measurements, the "sanity check" calculates the minimum distance between each landmark and the claimed country, then checks how fast a packet would have had to travel to cover that distance in the observed time. The claimed location is only accepted if none of the packets had to travel faster than the limit.

Figure 2.21 shows the percentage of overall claims by each proxy provider that our algorithm, ICLab's "sanity check," and five popular IP-to-location databases agree with. The numbers for CBG++ are calculated two ways: "generous" means we assume that all of the "uncertain" cases are actually credible, and "strict" means we assume they are all false. ICLab's algorithm is even stricter than ours, but most of that is explained by our more subtle handling of uncertain cases. Our "strict" numbers are usually within 10 % of the sanity check. Looking more deeply into the disagreements reveals that CBG++ almost always predicts a location close to a national border—just the situation where either algorithm could be tripped up by an underestimate.

All five of the IP-to-location databases are more likely to agree with the providers' claims than either active-geolocation approach is. As discussed earlier, we are inclined to suspect that this is because the proxy providers have influenced the information in these databases. We have no hard evidence backing this suspicion, but we observe that there is no pattern to the countries for which the IP-to-location databases disagree with provider claims. This is what we would expect to see if the databases were being influenced, but with some lag-time. As the proxy providers add servers,

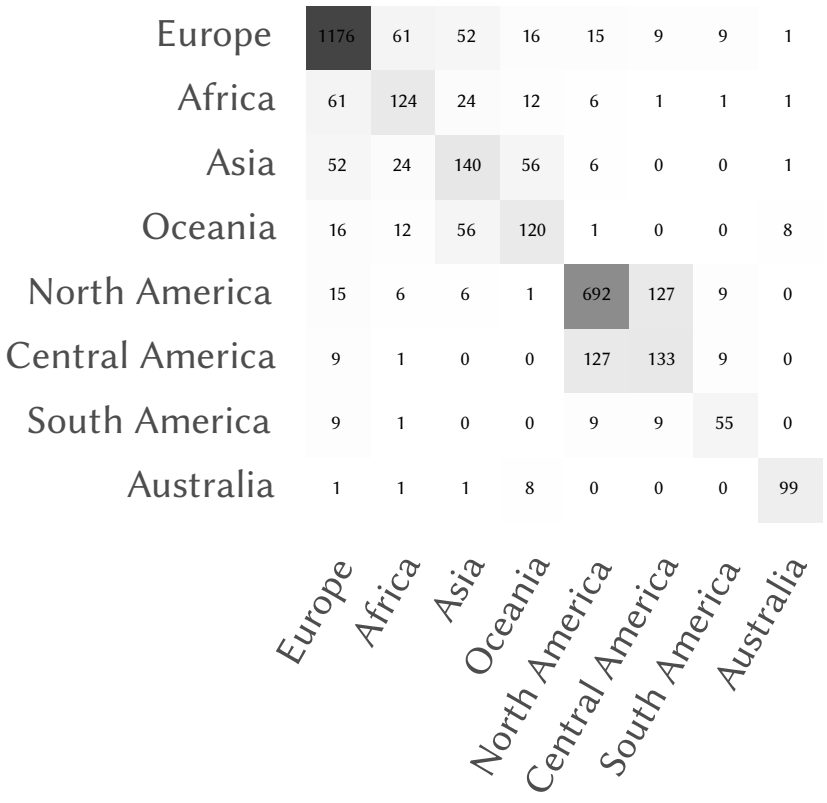



Figure 2.22: Confusion matrix among continents

the databases default their locations to a guess based on IP address registry information, which, for commercial data centers, may be reasonably close to the truth. When the database services attempt to make a more precise assessment, this draws on the source that the providers can influence.

## 2.6 Uncertainty and continents

Uncertain prediction regions include more than one country, or even more than one continent. Since a prediction region is always contiguous, we expect uncertainty among groups of neighboring countries, but which groups? We briefly examine this question with a pair of confusion matrices, one for continents (Figure 2.22) and the other for countries (Figure 2.23). All data is for the proxies, not the crowdsourced test hosts.

The lines separating continents are somewhat arbitrary. For this analysis, we chose to include Mexico with Central America, Turkey and Russia with Europe, all of the Middle East with Africa, and all of Malaysia and New Zealand with Oceania.

Intercontinental uncertainty is as one would expect: Europe/Africa/Asia, Asia/Oceania/Australia, North/Central and to a lesser extent South America. The country matrix, however, reveals that

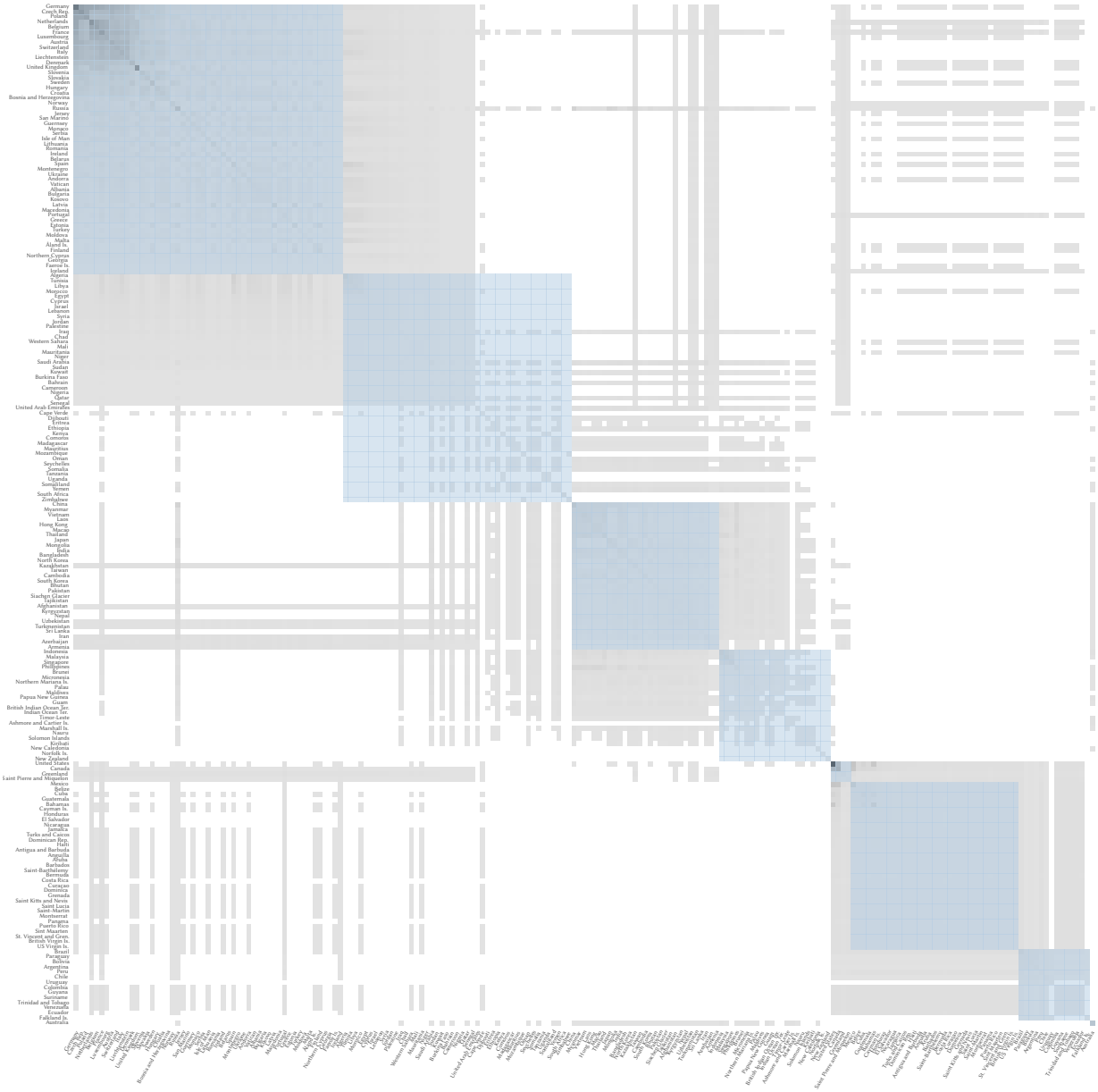


Figure 2.23: Confusion matrix among countries. Blue tinting marks groups of countries on the same continent.

just about every country within a continent *can* share a prediction region; the exceptions are more interesting. Many southern African countries seem more likely to be confused with Asia than their neighbors, and not just the Indian subcontinent, but all the way to Japan. Similar effects appear for Oceania. This may reflect neighboring countries or islands in these areas not being connected directly, only through a more developed hub.

## 2.7 Related work

Existing "measurement networks" such as PlanetLab [142], RIPE Atlas [151], or CAIDA Ark [32] have poor coverage outside Europe and North America, and at their best they only offer enough landmarks for city-scale geolocation. Wang *et al.* [181] propose to increase landmark density to the point where "street-level" geolocation is feasible, by enlisting small businesses' Web servers as additional landmarks, on the assumption that each server is physically located at the street address of the business. They apply heuristics to exclude sites on shared hosting and centralized corporate networks. Chen *et al.* [38] improves on this by using constrained mean-square error optimization to validate and fine-tune the location of each street-level server.

As we mentioned in Section 2.1, researchers in this field have devoted considerable effort to modeling the minimum and maximum feasible distance for each round-trip time measurement. Another line of research involves incorporating other sources of information as well as end-to-end round-trip time measurements. The original Octant (not the reimplementation in this paper) assumes that the target's LAN is probably small and any time spent within it is pure overhead, so it uses traceroute information to subtract off elapsed time up to the point where the routes begin to diverge. Komosny, Simek, and Kathiravelu [112] elaborated on this by using the Vivaldi [44] "network coordinate system" to model all of the observed distances between intermediate routers, but did not find an improvement over CBG. On the other hand, TBG [99] reports substantial improvements over CBG by using constrained optimization to do the same thing.

Eriksson *et al.* [65] recommends taking into account prior knowledge (in the Bayesian sense) about where a target host can plausibly be, such as geography ("must be on land"), population density ("more likely to be in a large city"), or known locations of data centers. Alidade [36] builds on this concept, drawing on both active measurements and passive data sources to compile a database that can be queried as easily as a traditional IP-to-location database, but with improved accuracy. OpenIPMap [85] has similar goals and also makes use of crowdsourced location reports.

## 2.8 Discussion

We have demonstrated the viability of a simple algorithm for active geolocation, CBG++, at global scale, especially when it is possible to use a crude location estimate to select landmarks within the same continent as the target. We have also confirmed that it is possible to geolocate proxy servers, even when they cannot be directly pinged. Our implementation of the four geolocation algorithms, as well as our measurement code, is publicly available at https://github.com/zackw/active-geolocator.

We have also put to the test the location claims of seven major commercial proxy operators. Our findings are dire: advertised server locations cannot be relied upon, especially when the operators claim to have servers in locations where server hosting is difficult. At most 70 % of the servers are where their operators say they are, and that is giving them the full benefit of the doubt; we can only confidently confirm the providers' claims for about 50 % of the servers, and all of those are in countries where hosting is easy. Provider A is especially misleading, but all seven of the providers we evaluated had at least a couple of questionable hosts. We shared our results with the providers and asked for an explanation, but all of them declined to respond.

Our results call into question the validity of any network measurement that used VPNs to gain location diversity, especially to diversify beyond Europe and North America. Also, despite a steady stream of reports that IP-to-location databases are unreliable (e.g. [75, 143, 158]) they are still relied upon in numerous contexts; we add our voices to those earlier notes of caution.

As we mentioned in the introduction, many of a VPN provider's customers might be content to appear to be in a specific country. We are not aware of anyone having investigated what VPN customers think they are buying, when they subscribe to a provider that advertises servers in many countries. It would be interesting to find out. Relatedly, while it is well-known that commercial IP-to-location databases contain *errors*, we are not aware of anyone having investigated the possibility of their containing deliberately false information (perhaps because the database compilers themselves were deceived).

One might also wonder whether the VPN operators could actively mislead investigators about the true location of their servers, by interfering with round-trip time measurements. They have no particular reason to do this now, but if active location validation becomes common, they might be motivated to try it. Previous work has found that hostile geolocation targets can indeed foul a position estimate. Gill *et al.* [77] and others [2, 130] report that *selective* added delay can displace the predicted region, so that its centroid is nowhere near the target's true location; more sophisticated delay-distance models are more susceptible to this, especially if they derive minimum as well as maximum feasible distances from delay measurements. Abdou, Matrawy, and Van Oorschot [3] go further, describing two methods for modifying ICMP echo replies so that some landmarks

compute *smaller* round-trip times than they should; with this ability, an adversarial target can shift the predicted region to be anywhere in the world, irrespective of its true location.

Our measurements use TCP handshakes, which include anti-forgery measures, rather than ICMP echo exchanges; also, we can trust both the landmarks and the host running the measurement tool. It is the VPN proxy, in the middle, that is the target of geolocation and not trusted. Unfortunately, being in the middle means it is *easier* for a proxy target to manipulate RTTs both up and down, than it was for an end-host target as considered by Abdou, Matrawy, and Van Oorschot It can selectively delay packets, and it can also selectively forge early SYN-ACKs without needing to guess sequence numbers, since it sees the SYNs. Conceivably, we could prevent this by using landmarks that report their own idea of the time, unforgeably, e.g. authenticated NTP servers [58]—if we could be sure that our measurement client and all of the landmarks already had synchronized clocks, which is a substantial engineering challenge in itself.

Finally, our Web-based measurement technique could be used to geolocate any visitor to a malicious website without their knowledge or consent. This would be foiled by the use of a proxy, VPN, Tor, or similar, in much the same way that IP-based geolocation is foiled by masking one's IP address with these tools. However, it is still an argument against allowing Web applications to record high-precision information about page load timings, and we plan to discuss this with the major browser vendors.

### 2.8.1 Future work

We were only able to include seven VPN providers in this study; there are at least 150 others, some of which make claims nearly as extravagant as provider A. We intend to expand the study to cover as many additional providers as possible, in cooperation with researchers and consumer watchdog organizations looking into other ways commercial VPN providers may fail to live up to their users' expectations. This will also allow us to repeat the measurements over time, and report on whether providers become more or less honest as the wider ecosystem changes.

In order to understand the errors added to our position estimates by the indirect measurement procedure described in Section 2.4.3, we are planning to set up test-bench VPN servers of our own, in known locations worldwide, and attempt to measure their locations both directly and indirectly.

While our two-phase measurement process is fast and efficient, it also produces noisy groups of measurements like those shown in Figure 2.16. We think this can be addressed with an iterative refinement process, in which additional probes and anchors are included in the measurement as necessary to reduce the size of the predicted region.

We are experimenting with an additional technique for detecting proxies in the same data center, in which we measure round-trip times to each proxy from each other proxy. Pilot tests indicate

that some groups of proxies (including proxies claimed to be in separate countries) show less than 5 ms round-trip times among themselves, which practically guarantees they are on the same local network.

It would be valuable to have a measurement tool that is as user-friendly as the existing Web-based tool, but as accurate as the command-line tool. The Web-based tool could reliably record the time for a single round-trip, and perhaps also avoid some of the Windows-specific overhead and noise, if it could use the W3C Navigation Timing API [178]. This API gives Web applications access to detailed information about the time taken for each stage of an HTTP query-response pair. Unfortunately, it can only be used if each server involved allows it, and currently none of the RIPE Atlas anchors and probes do. We plan to discuss the possibility with the RIPE team. Of course, as we mentioned above, the fact that active geolocation from a Web application is possible at all arguably constitutes a privacy leak in Web browsers, and we also plan to discuss that with the browser vendors.

RIPE Atlas anchors tend to be on sub-networks with more stable, less congested connectivity to the global backbone than is typical for their locale. That could mean each anchor's CBG++ bestline, calibrated from measurements of round-trip times to the other anchors, overestimates the distance packets can typically travel from that anchor. Overestimation leads only to greater uncertainty in predicted locations, whereas underestimation leads to failure (as discussed in Section 2.4.1). Still, this is a source of error that should be quantified. We are considering adding other measurement constellations, such as the CAIDA Archipelago [32] and PlanetLab [142], to our landmark set. This would allow us to compare the delay-distance relationships observed across constellations to those observed within a single constellation, and thus investigate the degree of overestimation. Additional constellations would also improve our landmark coverage outside Europe and North America. All of the above are also concentrated in the "developed world," but in sparse networks, each new landmark helps a great deal [66].

# 3. Censorship Detection

As I mentioned in Section 1.2, my proof-of-concept monitoring system builds on an existing, somewhat less ambitious monitoring system, ICLab, with whose principals I have been collaborating. ICLab seeks to monitor for censorship continuously and worldwide, to generate accurate, reproducible results and data sets useful to other researchers, and to minimize risk to volunteers. It does not, presently, seek to improve researchers' shared understanding of what to test.

In this chapter, I will describe the improved detection algorithms for suspicious TCP packet injection and unsuspected block pages that I developed to take advantage of ICLab's data. ICLab can also detect censorship by DNS manipulation, but I did not contribute to that algorithm, so it is not included as part of this thesis.

## 3.1 Censorship detection

The existing detection algorithms for network-level censorship are known to suffer from high levels of both false negatives and false positives [62, 76, 96, 184]. False negatives occur because of variation in how censorship is implemented, and false positives occur because every phenomenon that could be due to censorship also happens for innocuous reasons.

It is easiest to illustrate the problem with examples from block page detection. Jones *et al.* [96] observe that two different countries may block the same website, but are practically certain to use different block pages in doing so. Within a country, the same can happen from region to region or ISP to ISP. This means one cannot reliably detect block pages by matching on their expected contents. However, the alternative heuristics proposed by Jones *et al.* have proved to be too sensitive. For example, they measured block pages to be consistently at least 30 % shorter than the corresponding uncensored page observed from a control vantage, but innocuous server errors are also short compared to normal pages [196], so this is not a sufficient test for censorship by itself.

In collaboration with ICLab, I have developed improved detection algorithms which minimize false positives and negatives. I will describe here the algorithms I developed for detecting suspicious TCP packet injection and discovering unsuspected block pages; these are now in regular use as part of ICLab's continuous monitoring effort.

### 3.1.1 TCP packet injection

My algorithm for detecting TCP packets injected by a censor uses a double check to minimize false positives. The packet traces collected by ICLab are analyzed for both evidence of packet injection,

and evidence of intent to censor. Short error messages delivered by the legitimate server will not appear to be injected, and packets that, for innocuous reasons, appear to be injected, will not display an intent to censor.

**Payload conflict.** If a TCP peer receives two packets with valid checksums and the same sequence number but different payloads, whichever packet arrived first is accepted and the other is discarded [145]. An on-path censor can therefore suppress the server's HTTP response by injecting a packet carrying its own HTTP response (or simply an RST or FIN), timed to arrive first. Because ICLab records packet traces before TCP processing, it records both packets and detects a conflict. This is not infallible proof of packet injection; it can occur for innocuous reasons as well, such as HTTP load balancers that do not send exactly the same packet when they retransmit.

**TTL fluctuation.** The time-to-live (TTL) field of the IP header is a precaution against routing loops. It is initialized to a small power of two (usually 64, 128 or 256) and decremented by each router on the network path. When it reaches zero, the packet is discarded [146]. Normally all of the packets received by the client from one server should have the same TTL, as long as the routing is stable. Injected packets, however, frequently have a different TTL value, since the censor is closer to the client than the server is. Again, this is not infallible proof of packet injection; routing sometimes does change mid-session, and load balancers may handle the TCP handshake themselves but forward subsequent packets to a server that's a few hops further away.

**Intent to censor: RST, FIN, or block page.** When one of the above anomalies is detected, the anomalous packets are inspected in detail for evidence of intent to censor. An injected packet only accomplishes the censor's goals if it disrupts communication between the client and server, and this can only be done in a few ways. It can carry a TCP reset (RST) or close (FIN) flag, causing the client to abort the connection and report a generic error [43, 184]. Or it can carry an HTTP response declaring the site to be censored (a "block page," discussed further in Section 3.1.2), which will be rendered instead of the true contents of the page the client requested [47, 96]. If ICLab detects one of these three things within an anomalous packet, it counts the website as censored by TCP packet injection.

### 3.1.2 Block page discovery

As we mentioned in Section 1.1.1, block pages are used for overt censorship, where the censor wishes it to be known that a site is censored. Block pages' contents vary depending on the country and the technology used for censorship. Block pages that are already known can be detected with regular expressions applied to the TCP payloads of suspicious packets, but this can easily miss even small variations from the expected text, and is no help at all with unknown block pages.

Nonetheless, ICLab's primary technique for detecting block pages is a set of hand-curated

regular expressions, because these can be written to eliminate all false positives. The set began as a combination of the regular expressions collected by the Citizen Lab [40], OONI [70], and Quack [173], augmented with the block-page discovery techniques described below. We obtained 24, 40, and 126 regular expressions from OONI, Citizen lab and Quack respectively.

Anomalous packets that do *not* match any of the hand-curated regular expressions are examined for block page variations and unknown block pages, using a combination of four heuristics.

**Self-contained HTTP response.** The protocol structure of HTTP, as it is used in practice,[1] requires a censor who wishes to display a block page to inject a single packet containing a complete, self-contained HTTP response message. This packet must arrive before, and have the same sequence number as, the first data packet of the legitimate response. Anything else may garble the legitimate response, but will not suppress it altogether. Therefore, given a pair of packets involved in a payload conflict, if one of them is not a complete, self-contained HTTP response, it is unlikely to have been transmitted by the censor.

This heuristic is used as a preliminary filter. Collisions that pass this heuristic are considered candidate block pages to be processed by the next three heuristics. Note that if *both* sides of the collision are self-contained HTTP responses, both sides are considered candidates.

**HTML tag frequency vector clusters.** The text of a block page is written by the censor, but the HTML tag structure of the page is often more characteristic of the filtering equipment they are using. When the same filtering equipment is used in many different locations—either different countries, or different networks within one country—the tag structure is often an exact match, even when the text varies. We reduce each candidate block page to a vector of HTML tag frequencies (one BODY, two P, three EM, etc.) and compare the vectors to all other candidate block pages' vectors, and to vectors for pages that match the known block page expressions. When we find an exact match, we manually inspect the matching candidates and decide whether to add a new regular expression to the curated set that will detect them. Using this heuristic, we discovered 15 new block page signatures in five countries.

**Textual similarity clusters.** Within one country, the legal jargon used to justify censoring a page may vary from locale to locale, but is likely to be similar overall. For example, one Indian ISP refers to "a court of competent jurisdiction" in its block pages, and another uses the phrase "Hon'ble Court" instead. Small variations like this are evidently the same page to a human, but a regular expression will miss them. We apply *locality-sensitive hashing* (LSH) [199] to the text of the candidate block pages, after stripping out the HTML structure but preserving the URLs associated with any hyperlinks. LSH produces clusters of candidate pages, centered on pages that do match the known block page expressions. As with the tag frequency vectors, we manually inspect the

[1] The HTTP 1.1 feature of "pipelining" [69] would have complicated matters, but it proved too unreliable for any major browser to deploy [155].

| HTML structure | Visible message |
|---|---|
| ACK+PSH<br>HTTP/1.1 200 OK<br>Connection: close<br>Content-Length: *nnnn*<br>Content-Type: text/html;<br>charset="utf-8"<br><!DOCTYPE html PUBLIC<br>"-//W3C//DTD HTML 4.01//EN"><br><html><br><head<title></title></head><br><body><br><h0><font color="black"><br>*visible message*<br></font></h0><br></body><br></html> | "This URL has been blocked under instructions of a competent Government Authority or in compliance with the orders of a Court of competent jurisdiction.<br>***This URL has been blocked under Instructions of the Competent Government Authority or Incompliance to the orders of Hon'ble Court.*** *[sic]*<br>*"Error 403: Access Denied/Forbidden"*<br>404. That's an error.<br>HTTP Error 404 - File or Directory not found<br>HTTP Error 404 - File or Directory not found = *http://...* |

Figure 3.1: Example cluster of block pages. All of the messages in the right-hand column were observed with the HTTP response headers and HTML structure shown on the left.

clusters and decide whether to add new regular expressions to the curated set. Using this technique, we discovered 33 new block page signatures in eight countries.

Figure 3.1 shows an example cluster, including both variations on the Indian legal jargon mentioned above, but also four other messages, all of which are attempting to mimic generic HTTP server errors—thus, this technique can detect covert as well as overt censorship.

**URL-to-country ratio.** To discover wholly unknown block pages, rather than just variations on previously known pages, we take each LSH cluster that is *not* centered on a known block page, count the number of URLs that produced a page in that cluster, and divide by the number of countries where a page in that cluster was observed. We sort the clusters from largest to smallest URL-to-country ratio and then read through the entire list manually. The largest ratio associated with a newly discovered block page was 286 and the smallest ratio was 1.0.

Although manual inspection is still necessary, this heuristic reliably sorts clusters which are not block pages to the bottom of the list. The most common two cases of non-block page clusters are due to collisions where both sides are self-contained single-packet HTTP responses. When a group of censored websites have all been purchased by the same conglomerate and merged into one site, we will see the censor's redirects on one side of the packet collision and the conglomerate's redirects on the other. Similarly, when a group of censored websites all use the same content delivery network, "domain parking" service, or HTTP server software, we may observe the censor's redirects on one side of the packet collision and nearly-identical generic error pages on the other.

| Country | Alexa Global Category | Percentage | Country | Citizen Lab Global Category | Percentage | Country | Country Sensitive Category | Percentage |
|---|---|---|---|---|---|---|---|---|
| Iran | News and Media | 13.6 | Iran | Pornography | 11.5 | Iran | News and Media | 20.0 |
| | Pornography | 12.4 | | News and Media | 9.4 | | Personal Websites | 17.2 |
| | Entertainment | 10.0 | | Proxy Avoidance | 6.7 | | Political Organizations | 7.3 |
| Turkey | Pornography | 67.3 | Turkey | Pornography | 46.6 | India | Entertainment | 16.5 |
| | Illegal or Unethical | 4.0 | | Gambling | 21.9 | | Streaming Media | 14.1 |
| | Gambling | 4.0 | | File Sharing | 4.5 | | Social Networking | 13.7 |
| South Korea | Pornography | 79.4 | South Korea | Pornography | 46.6 | Russia | Personal Websites | 16.5 |
| | Illegal or Unethical | 5.8 | | Gambling | 7.4 | | News and Media | 14.4 |
| | Other Adult Materials | 3.0 | | News and Media | 4.2 | | Gambling | 12.3 |
| India | Illegal or Unethical | 55.2 | Uganda | Pornography | 35.4 | Turkey | News and Media | 29.1 |
| | Streaming Media | 6.9 | | Lingerie and Swimsuit | 8.5 | | Pornography | 13.6 |
| | File Sharing | 6.9 | | Other Adult Material | 8.5 | | Gambling | 9.7 |
| Russia | Pornography | 40.0 | India | Illegal or Unethical | 15.0 | South Korea | News and Media | 16.6 |
| | Instant Messaging | 10.0 | | File Sharing | 14.0 | | Pornography | 15.3 |
| | Gambling | 10.0 | | Streaming Media | 6.5 | | Shopping | 10.9 |

Table 3.1: Variance in censorship observations by test list

## 3.2 Detection results

ICLab has historically used three probe lists: the Alexa top 500 (updated periodically) and the global and country-sensitive lists compiled by the Citizen Lab [39]. Table 3.1 shows how the five countries with the most observed censorship vary if we only include one of these three lists when ranking the amount of censorship performed. (The classification used here was developed by FortiGuard Labs [71]. In Section 5.1.1 we will compare their classification with the mechanical classification I have developed.)

While there is a great deal of overlap, India and Russia both score higher on their country-specific list than on the global lists, and Uganda displaces Russia from the top five when considering only the Citizen Lab global list. The top three categories blocked by each country also varies somewhat from list to list; in particular, pornography and illegal content are much less prominent on the country-specific lists than on the global lists, and while other countries heavily censor specific classes of content, Iran has a more uniform distribution across block lists. This demonstrates how one's choice of test lists can lead to different conclusions about censorship policy.

### 3.2.1 Trends over time

Figure 3.2 shows the filtering trend of the five countries with the most censorship observed overall. The y-axis gives the percentage of blocked URLs relative to total tested URLs in each country.

ICLab lost access to its Iranian vantage points in mid-2017 due to sanctions imposed by the USA. Prior to that, the observed blocking rate in this country is always above 20 %, except for February. This fluctuation is because of churn in the Alexa top 500 (as described by Scheitle *et al.* [154]) causing only 590 URLs to be tested in February 2017, as opposed to more than 700 URLs in other months. This affects Iran more than the other countries in Figure 3.2 because only Iran blocks

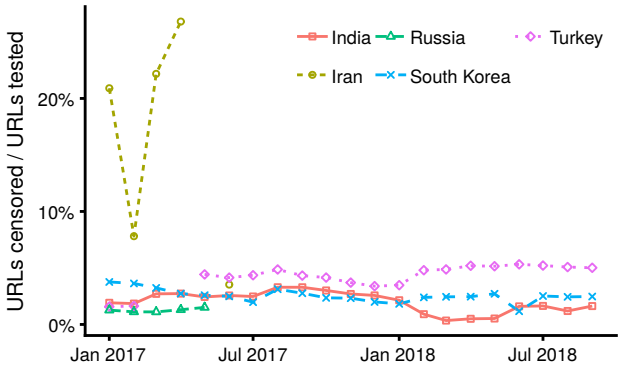



Figure 3.2: Filtering trends observed by ICLab for the five countries with the most censorship observed overall.

so many of the sites included in the Alexa top 500.

In India, ICLab observes a consistent 2.5–3.5 % blocking rate for all of 2017, dropping to less than 0.5 % at the beginning of 2018. This correlates with an official policy change: India's telecommunications regulators announced a new set of regulations, requiring "net neutrality," in November 2017 [153].

At the beginning of 2017, ICLab observed an overall blocking rate of about 1.5 % in Turkey, mostly comprised of sites carrying pornographic or other sexual content. Following a series of controversial political events in April 2017, the blocking rate rose to 5 % and has remained there ever since. The additions are almost all from the "News and Media" category, and many of them are sites whose editorial policies conflict with the government's interests, consistent with reporting elsewhere [102, 108].

### 3.2.2 Combinations of censorship techniques

Table 3.2 can be used to further understand which categories are blocked in an overt or covert fashion. Block pages are considered a means for overt censorship, we can observe that Iran serves block pages for news websites more often than personal blogs. Further, Figure 3.3 allows us to compare which censorship techniques are mostly used in countries and how countries combine these techniques to block access to web content. We can easily observe that block pages are more common than DNS manipulation and TCP packet injections in all countries. Further, we see instances where two techniques are used alongside each other. As an example, we see Iran is redirecting users to a block page using DNS manipulation.

| Censorship Technique | Country | Categories | Percentage |
|---|---|---|---|
| Block page | Iran | NEWS, PORN, BLOG | 24.2 |
| | India | ENT, STRM, SOC | 6.6 |
| | Turkey | PORN, GAMB, NEWS | 4.5 |
| | South Korea | PORN, NEWS, SHOP | 3.5 |
| | Russia | GAMB, PORN, BLOG | 2.4 |
| DNS manipulation | Iran | BLOG, NEWS, PORN | 7.0 |
| | Uganda | PORN, ADUL, LING | 1.8 |
| | Turkey | ILL, NEWG, GAMB | 0.4 |
| | South Korea | ORG, STRM, FILE | 0.4 |
| | India | FILE, PORN, ILL | 0.4 |
| TCP packet injection | Iran | NEWS, PORN, RELI | 5.0 |
| | India | ENT, NEWS, STRM | 3.8 |
| | South Korea | PORN, NEWS, SHOP | 2.2 |
| | New Zealand | STRM, IT, GOV | 0.4 |
| | Russia | GAMB, PORN, MAR | 0.4 |

Table 3.2: How the "five countries with the most censorship observed" varies among the three classes of network interference that ICLab can detect. Some countries appear in more than one class, but with different categories of material censored. Abbreviations are defined in Table 3.3.

| Abbreviations | Category |
|---|---|
| ADUL | Other Adult Materials |
| BLOG | Personal websites and Blogs |
| ENT | Entertainment |
| FILE | File Sharing and Storage |
| GAMB | Gambling |
| GOV | Government and Legal Organizations |
| ILL | Illegal or Unethical |
| IT | Information Technology |
| LING | Lingerie and Swimsuit |
| MAR | Marijuana |
| NEWG | Newsgroups and Message Boards |
| NEWS | News and Media |
| ORG | General Organizations |
| PORN | Pornography |
| RELI | Global Religion |
| SHOP | Shopping |
| SOC | Social Networking |
| STRM | Streaming Media and Download |

Table 3.3: Abbreviations and full category names from FortiGuard, used in Table 3.2.

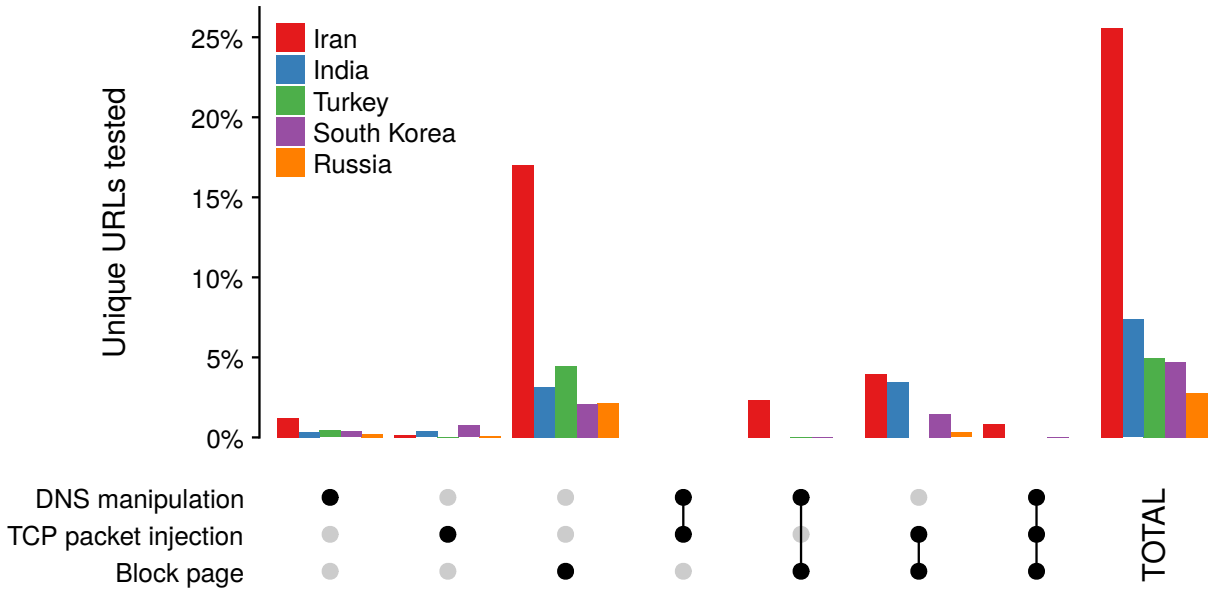



Figure 3.3: Prevalence of each censorship technique that ICLab can detect, for the five countries with the most censorship observed overall.

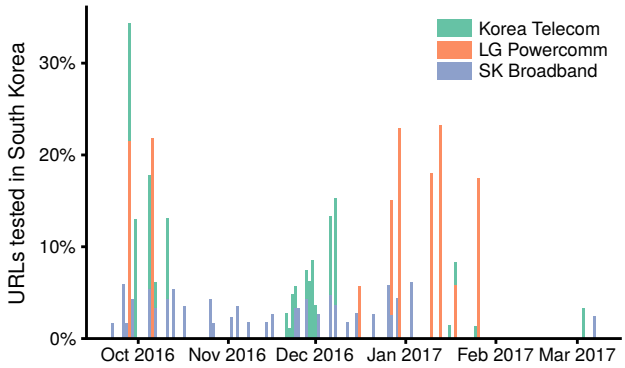



Figure 3.4: For several months, a browser-fingerprinting and tracking script was injected into 5–30% of all test page loads from vantages in three Korean ASes.

### 3.2.3 User tracking injection

The detector for unknown block pages (Section 3.1.2) flagged a cluster of injected TCP payloads observed only in South Korea. Upon manual inspection, these were not block pages. Instead, they contained a script that would fingerprint the client and then allow the originally intended page to load. As shown in Figure 3.4, ICLab observed these scripts being injected over a five-month period from Oct. 2016 through Feb. 2017, from vantage points within three Korean ASes, into 5–30 % of all our test page loads, with no correlation with the content of the affected page. By contrast, actual censorship by the South Korean government affects less than 1 % of tests and is focused on pornography, illegal file sharing, and North Korean propaganda.

We do not know who was injecting these scripts; it could be the VPN service, the ISPs owning the ASes listed in Figure 3.4, or one or more of their transit providers. We do not know their motives either, but the phenomenon resembles techniques used by ad networks for recording profiles of individual web users and then serving them targeted ads [4]. This demonstrates the importance of manual checking for false positives in censorship detection. All of the detection heuristics described in Sections 3.1.1 and 3.1.2 triggered on these scripts, but they are not censorship. On the other hand, they are invasive of user privacy, and could facilitate surveillance.

### 3.2.4 Cryptocurrency mining injection

The block page detector also flagged a set of suspicious responses observed only in Brazil, that were not block pages; instead, they were malware that would have caused the client's browser to mine cryptocurrency (an increasingly popular way to earn money with malware [114]). We were able to identify this malware as originating with a botnet infecting MikroTik routers (exploiting CVE-2018-14847), initially seen only in Brazil [103] but now reported to affect more than 200 000 routers worldwide [60]. Infected routers inject the mining malware into HTTP responses passing through them.

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

# 4. Assessment of Existing Probe Lists

ICLab presently relies on the probe list maintained by Citizen Lab [39]. The process for updating this list is informal, and entirely reliant on volunteer contributions [166]. One might wonder whether it is missing anything important. In this chapter, we assess the quality of the Citizen Lab probe list and compare it with 21 other lists of URLs that may or may not be censored in any particular country, using mechanical topic analysis. The material in this chapter was previously published as "Topics of Controversy" at PETS 2017 [187].

We use the following five criteria for a high-quality probe list:

**Breadth** A good list includes many different types of potentially-censored material. Hand-compiled probe lists reflect the developers' interests, so they may over-investigate some types of potentially censored material and under-investigate others. Deviations from a country's official policy will only be discovered by a probe list that is not limited to the official policy.

**Depth** A good list includes many sites for each type of material, so that it will reveal how thoroughly that material is censored in each target country, and the boundaries of the category. This is especially important when one list is to be used to probe the policies of many different countries, because even when two countries declare the same official policy, the actual set of sites blocked in each country may be different.

**Freshness** A good list includes sites that are currently active, and avoids those that are abandoned. Sophisticated censors devote more effort to recently published content. China's "Great Firewall," for instance, not only adds sites to its blacklist within hours of their becoming newsworthy, but drops them again just as quickly when they stop being a focus of public attention [1, 43, 200]. Thus, an outdated probe list would underestimate the effectiveness of censorship in China.

Conversely, less sophisticated censors may be content to use off-the-shelf, rarely-updated blacklists of porn, gambling, etc. sites, perhaps with local additions. A recent crackdown on pornography in Pakistan led to a 50% reduction in consumption, but the remaining 50% simply shifted to sites that had escaped the initial sweep—and the censors did not update their blacklist to match [131]. Thus, an outdated probe list would overestimate the effectiveness of censorship in Pakistan.

**Efficiency** A good list can be probed in a short time, even over a slow, unreliable network connection. This is most important when attempting to conduct fine-grained measurements, but a list that is too large or bandwidth-intensive might not be usable at all.

Efficiency, unfortunately, is in direct tension with breadth and depth: the easiest way to make a probe list more efficient is to remove things from it. As we discuss in Section 4.2, efficiency also

suffers if one seeks to collect more detailed information from each probed site.

**Ease of maintenance** A good list requires little or no manual adjustment on an ongoing basis. This is obviously in tension with freshness, and provides a second reason to keep lists short.

## 4.1 Data sources

We studied 758 191 unique URLs drawn from 22 lists (shown in Table 4.1). Only one was created to be used as a probe list [39], but another 15 are (allegedly) actual blacklists used in specific countries, and two more have algorithmic selection criteria that should be positively correlated with censorship. The remaining four are control groups. One should be *negatively* correlated with censorship, and the others are neutral.

Due to the sheer size and diversity of the global Web, and the large number of pages that are not discoverable by traversing the link graph [22], any sample will inevitably miss something. We cannot hope to avoid this problem, but drawing our sample from a wide variety of sources with diverse selection criteria should mitigate it.

### 4.1.1 Potentially censored

Pages from these lists should be more likely than average to be censored somewhere.

**Blacklists and pinklists** These documents purport to be (part of) actual lists of censored URLs in some countries. Most are one-time snapshots; some are continuously updated. They must be interpreted cautiously. For instance, the leaked "BlueCoat" logs for Syria [34] list only URLs that someone tried to load; there is no way of knowing whether other URLs are also blocked, and one must guess whether entire sites are blocked or just specific pages.

This study includes 15 lists from 12 countries, for a total of 331 362 URLs. Eight of them include overwhelmingly more pornography than anything else; we will refer to these as *pinklists* below. (All eight do include some non-pornographic sites, even though six of them are from countries where the ostensible official policy is *only* to block pornography.) The other seven do not have this emphasis, and we will refer to them as *blacklists* below.

**Citizen Lab probe list** At the time of the study described in this chapter, the Citizen Lab probe list (also known as the "Open Net Initiative" probe list, and referred to as such in all of the figures in this chapter) consisted of 12 107 hand-curated URLs discussing sensitive topics [39]. The principle is that these are more likely to be censored than average, not that they necessarily *are* censored somewhere. We take this list as representative of the probe lists used by researchers in this field. 1 227 of the URLs are labeled as globally relevant, the rest as relevant to one or more specific countries.

Hand-curated lists will inevitably reflect the concerns of their compilers. This probe list, for instance, has more "freedom of expression and media freedom" sites on the list than anything else.

**Herdict** Herdict [20] is a service which aggregates worldwide reports that a website is inaccessible. A list of all the URLs ever reported can be downloaded from a central server; this comes to 76 935 URLs in total. The browser extension for making reports is marketed as a censorship-reporting system, but they do not filter out other kinds of site outage. This list includes a great deal of junk, such as hundreds of URLs referring to specific IP addresses that serve Google's front page.

**Controversial Wikipedia articles and their references** Yasseri *et al.* [198] observe that controversy on Wikipedia can be mechanically detected by analyzing the revision history of each article. Specifically, if an article's history includes many "mutual reverts," where pairs of editors each roll back the other's work, then the article is probably controversial. (This is a conservative measure; as they point out, Wikipedia's edit wars can be much more subtle.) They published lists of controversial Wikipedia articles in 13 languages. We augmented their lists with the external links from each article. This came to a total of 105 181 URLs.

### 4.1.2 Controls

These lists were selected to reflect the Web at large.

**Pinboard** We expect pages on this list to be *less* likely to be censored than average. It is a personal bookmark list with 3 276 URLs, consisting mostly of articles on graphic design, Web design, and general computer programming, with the occasional online shopfront.

**Alexa 25K** Alexa Inc. claims that these are the 25 019 most popular websites worldwide; their methodology is opaque, and we suspect it over-weights the WEIRD (Western, Educated, Industrialized, Rich, and Democratic [84]) population. Sensitive content is often only of interest to a narrow audience, and the popularity of major global brands gives them some protection from censorship, so sites on this list may also be less likely to be censored than average.

**Twitter** Another angle on popularity, we use a small (less than 0.1 %) sample of all the URLs shared on Twitter from March 17 through 24, 2014, comprising 30 487 URLs shared by 27 731 user accounts. Twitter was chosen over other social networks because, at the time of the sample, political advocacy and organization via Twitter was fashionable.

**Common Crawl** Finally, this is the closest available approximation to an unbiased sample of the entire Web. The Common Crawl Foundation continuously operates a large-scale Web crawl and publishes the results [165]. Each crawl contains at least a billion pages. We sampled 177 109 pages from the September 2015 crawl uniformly at random.

Table 4.1: Jaccard coefficients for list similarity, by URL

| | | (size) | aus | dnk | fin | deu | ita | nor | th1 | tur | in1 | in2 | rus | syr | th2 | th3 | gbr | oni | hdk | wki | pin | alx | twi | ccr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pinklist | Australia 2009 | 5 130 | 1 | < | .01 | .02 | .05 | .03 | .01 | < | | | < | < | | | < | < | < | | | < | ≪ | |
| | Denmark 2008 | 7 402 | < | 1 | .08 | < | < | .12 | < | < | | | ≪ | | | | ≪ | | ≪ | | | | | |
| | Finland 2009 | 1 336 | .01 | .08 | 1 | < | .03 | .04 | < | < | | | < | | | | ≪ | | < | | | | | |
| | Germany 2014 | 13 174 | .02 | < | < | 1 | < | < | .02 | .01 | | | < | | | | < | < | < | ≪ | | < | ≪ | |
| | Italy 2009 | 1 078 | .05 | < | .03 | < | 1 | .03 | < | < | | | < | | | | < | < | < | | | < | | |
| | Norway 2009 | 14 022 | .03 | .12 | .04 | < | .03 | 1 | < | < | | | < | | | | < | | < | | | | | |
| | Thailand 2007 | 26 789 | .01 | < | < | .02 | < | < | 1 | .01 | | | < | < | | | < | < | < | ≪ | | < | ≪ | |
| | Turkey 2015 | 172 971 | < | < | < | .01 | < | < | .01 | 1 | ≪ | | < | ≪ | | | .02 | < | < | ≪ | ≪ | < | < | ≪ |
| blacklist | India 2012 (Anonymous) | 214 | | | | | | | | ≪ | 1 | | | | | | < | < | < | | | | | |
| | India 2012 (Assam riots) | 103 | | | | | | | | | | 1 | | | | | | < | | | | | | |
| | Russia 2014 | 4 482 | < | ≪ | < | < | < | < | < | < | | | 1 | | | | < | < | < | | | < | | |
| | Syria 2015 | 12 428 | < | | | | | | < | ≪ | | | | 1 | | | < | < | < | < | ≪ | < | ≪ | |
| | Thailand 2008 | 1 298 | | | | | | | | | | | | | 1 | < | | | < | | | | | |
| | Thailand 2009 | 408 | | | | | | | | | | | | | < | 1 | | | | | | | | |
| | Great Britain 2015 | 87 032 | < | ≪ | ≪ | < | < | < | < | .02 | < | | < | < | | | 1 | < | .03 | < | < | < | < | ≪ |
| probe list | OpenNet Initiative 2014 | 12 107 | < | | | < | < | | < | < | < | < | < | < | | | < | 1 | .02 | < | < | .01 | < | ≪ |
| crowdsourced | Herdict 2014 | 76 935 | < | ≪ | < | < | < | < | < | < | < | | < | < | < | | .03 | .02 | 1 | < | < | .04 | < | ≪ |
| sampled | Wikipedia controv. 2015 | 105 181 | | | | ≪ | | | ≪ | ≪ | | | | < | | | < | < | < | 1 | ≪ | < | < | ≪ |
| neg. control | Pinboard 2014 | 3 876 | | | | | | | | ≪ | | | | ≪ | | | < | < | < | ≪ | 1 | < | ≪ | ≪ |
| popular | Alexa 2014 | 25 019 | < | | | < | < | | < | < | | | < | < | | | < | .01 | .04 | < | < | 1 | < | ≪ |
| | Tweets 2014 | 40 198 | ≪ | | | ≪ | | | ≪ | < | | | | ≪ | | | < | < | < | < | ≪ | < | 1 | ≪ |
| generic | Common Crawl 2015 | 177 109 | | | | | | | | ≪ | | | | | | | ≪ | ≪ | ≪ | ≪ | ≪ | ≪ | ≪ | 1 |

<: smaller than 0.01. ≪: smaller than 0.0001. Blank: zero.

### 4.1.3 Overlap between lists

We begin our investigation by comparing the probe lists to each other, using the Jaccard index of similarity: $J = \frac{|A \cap B|}{|A \cup B|}$ for any two sets $A$ and $B$. It ranges from 0 (no overlap at all) to 1 (complete overlap).

Table 4.1 shows the Jaccard indices for each pair of lists, comparing full URLs. It is evident that, although there is some overlap (especially among the pinklists, in the upper left-hand corner), very few full URLs appear in more than one list. There is more commonality if we look only at the hostnames, as shown in Table 4.2. The pinklists continue clearly to be more similar to each other than to anything else. The blacklists, interestingly, continue not to have much in common with each other. And, equally interestingly, all the other lists—regardless of sampling criteria—have more in common with each other than they do with most of the blacklists and pinklists. This already suggests that manually curated lists such as ONI's may not be digging deeply enough into the "long tail" of special-interest websites.

While we can see that there are patterns of similarities, Tables 4.1 and 4.2 do not reveal *what* some lists have in common with each other. To discover that, we must study the content of each page.

We collected both *contemporary* and *historical* snapshots of each page we investigated. As the

Table 4.2: Jaccard coefficients for list similarity, by hostname

| | | (size) | aus | dnk | fin | deu | ita | nor | th1 | tur | in1 | in2 | rus | syr | th2 | th3 | gbr | oni | hdk | wki | pin | alx | twi | ccr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **pinklist** | **Australia 2009** | 1 752 | 1 | .01 | .04 | .03 | .08 | .04 | .02 | < | | < | < | < | < | < | < | < | < | < | < | < | < | < |
| | **Denmark 2008** | 7 402 | .01 | 1 | .08 | < | < | .19 | < | < | | | < | | | | ≪ | | < | | | | | |
| | **Finland 2009** | 1 336 | .04 | .08 | 1 | < | .04 | .07 | < | < | | | < | < | | | ≪ | | < | | | | | ≪ |
| | **Germany 2014** | 6 199 | .03 | < | < | 1 | < | .01 | .02 | .01 | | < | < | < | < | < | < | < | < | < | < | < | < | < |
| | **Italy 2009** | 539 | .08 | < | .04 | < | 1 | .03 | < | < | | | < | | | | < | < | < | | | < | | ≪ |
| | **Norway 2009** | 7 011 | .04 | .19 | .07 | .01 | .03 | 1 | < | < | | | < | | | | < | | < | ≪ | | | | ≪ |
| | **Thailand 2007** | 11 880 | .02 | < | < | .02 | < | < | 1 | .01 | | ≪ | < | < | < | < | < | < | < | < | < | < | < | < |
| | **Turkey 2015** | 172 971 | < | < | < | .01 | < | < | .01 | 1 | ≪ | ≪ | < | < | ≪ | ≪ | .02 | < | < | < | ≪ | < | < | < |
| **blacklist** | **India 2012 (Anonymous)** | 203 | | | | | | | | ≪ | 1 | | | < | | | < | < | < | ≪ | < | | < | < |
| | **India 2012 (Assam riots)** | 21 | < | | | < | | | ≪ | ≪ | | 1 | < | < | < | < | ≪ | < | < | < | < | < | < | < |
| | **Russia 2014** | 1 994 | < | < | < | < | < | < | < | < | | < | 1 | < | < | < | < | < | < | < | < | < | < | < |
| | **Syria 2015** | 6 526 | < | | < | < | | | < | < | < | < | < | 1 | < | < | < | .02 | < | < | < | < | < | < |
| | **Thailand 2008** | 104 | < | | | < | | | < | ≪ | | < | < | < | 1 | .21 | ≪ | < | < | < | < | < | < | < |
| | **Thailand 2009** | 94 | < | | | < | | | < | ≪ | | < | < | < | .21 | 1 | ≪ | < | < | < | < | ≪ | < | ≪ |
| | **Great Britain 2015** | 79 510 | < | ≪ | ≪ | < | < | < | < | .02 | < | ≪ | < | < | ≪ | ≪ | 1 | < | .03 | < | < | < | < | < |
| **probe list** | **OpenNet Initiative 2014** | 10 016 | < | | | < | < | | < | < | < | < | < | .02 | < | < | < | 1 | .02 | .02 | < | .02 | < | < |
| **crowdsourced** | **Herdict 2014** | 70 528 | < | < | < | < | < | < | < | < | < | < | < | < | < | < | .03 | .02 | 1 | .02 | < | .04 | .01 | .02 |
| **sampled** | **Wikipedia controv. 2015** | 27 410 | < | | | < | | ≪ | < | < | ≪ | < | < | < | < | < | < | .02 | .02 | 1 | < | .04 | .01 | .03 |
| **neg. control** | **Pinboard 2014** | 2 495 | < | | | < | | | < | ≪ | < | < | < | < | < | < | < | < | < | < | 1 | < | < | < |
| **popular** | **Alexa 2014** | 24 977 | < | | | < | < | | < | < | | < | < | < | < | ≪ | < | .02 | .04 | .04 | < | 1 | .02 | .05 |
| | **Tweets 2014** | 12 504 | < | | | < | | | < | < | < | < | < | < | < | < | < | < | .01 | .01 | < | .02 | 1 | < |
| **generic** | **Common Crawl 2015** | 47 042 | < | | ≪ | < | ≪ | ≪ | < | < | < | < | < | < | < | ≪ | < | < | .02 | .03 | < | .05 | < | 1 |

<: smaller than 0.01. ≪: smaller than 0.0001. Blank: zero.

names imply, contemporary snapshots show the page as it currently is, whereas historical snapshots look back in time. Contemporary data is sufficient to evaluate web page availability and topic. Historical data reveals pages whose topic has changed since they were entered onto a blacklist; it helps us discover the topic of pages that no longer exist; and it allows us to compute page availability over time, as well as topic changes over time.

All of the snapshots were collected from commercial data centers in the USA. There have been occasional moves toward blocking access to "obscene" material in the USA [169], but we are not aware of any filtering imposed on the data centers we used.

## 4.2 Contemporary data collection

We used an automated Web browser, PhantomJS [86] to collect contemporary snapshots of each page. We record the complete set of HTTP requests and responses in an HTTP Archive (HAR) [136]. Images and multimedia content are not downloaded or recorded, but their URLs are logged. We also record HTML corresponding to the main document after any JavaScript programs have had a chance to modify it. This HTML is preprocessed as described in Section 4.4 and then analyzed for its topic as described in Section 4.5.

PhantomJS is based on WebKit, and supports roughly the same set of features as Safari 6.0.

Relative to "bleeding edge" browsers, the most significant missing features involve multimedia content (video, audio, etc.), which we would not collect anyway, for legal reasons (see below). A controller program started a new instance of PhantomJS for each page load, with all caches, cookie jars, etc. erased.

An automated browser offers major advantages over traditional "crawling" using an HTTP client that does not parse HTML or execute JavaScript. An increasing number of pages rely on JavaScript to the point where a client that does not run scripts will see none of the intended content. Also, markup ambiguities and errors are handled exactly as a human-driven browser (Safari) would. Downstream processing receives only well-structured, canonicalized HTML documents. Finally, it is harder for the server to detect that it is being accessed by an automated process, which might cause it to send back different material than a human would receive [180].

This approach also has disadvantages. The most significant is its cost in time and computer power. The data-collection host could sustain an average page-load rate of approximately 4 pages per second, with the limiting factor being PhantomJS's substantial RAM requirements. A well-tuned traditional crawler, by contrast, can sustain an average page-load rate of 2 000 pages per second with roughly equivalent hardware resources [6]. We also suffer from a much larger set of client bugs. 6 980 attempted page loads (0.92 %) caused PhantomJS to crash. Finally, it is still possible for sites to distinguish PhantomJS from a "real" browser. Some sites block this kind of close mimic, while allowing obvious web crawlers access. For instance, LinkedIn blocked us from accessing user profile pages and job listings.

Our collector ignores `robots.txt`, because a human-driven browser would do the same. Instead, we avoid disruptive effects on websites by randomizing the order of page loads, so that no website sees a large number of accesses in a short time. Also, the collector by itself does not traverse outbound hyperlinks from any page, which reduces the odds of *modifying* sites by accessing them. For legal reasons, our collector does not load images and videos, nor does it record how the pages would be rendered. While HTML sources are safe, there exist images that are illegal to possess, even unintentionally, in the USA.

Ideally, contemporary data collection should occur at a single point in time, but this is impractical, given the volume of data we are acquiring. Most of our contemporary data was collected over a two-month period ranging from September 21 through December 3, 2015. For efficiency, we imported HTML directly from Common Crawl's data release rather than re-crawling each page ourselves. These pages were collected between July 28 and August 5, 2015. More importantly, Common Crawl is a traditional crawler, so these pages' contents may be less accurately recorded.

Table 4.3 shows statistics on how many of the pages were successfully retrieved, and for those that were not, why not. The most common problem is that the entire domain either no longer exists, or has been decommissioned and "parked" (see Section 4.4.1). It is also quite common for a single

Table 4.3: The proportion of URLs from both the initial and the snowball samples that were successfully and unsuccessfully retrieved, with a high-level breakdown of the reasons for unsuccessful retrieval.

| | Total | | Initial | | Snowball | |
|---|---|---|---|---|---|---|
| **All URLs** | 911 058 | | 803 665 | | 137 494 | |
| **Successfully retrieved** | 628 823 | 69 % | 576 598 | 72 % | 74 779 | 54 % |
| **Page lost or inaccessible** | 191 929 | 21 % | 176 440 | 22 % | 22 691 | 17 % |
| Host not found | 77 764 | 8.5 | 74 666 | 9.3 | 6 275 | 4.6 |
| Parked domain | 33 137 | 3.6 | 33 733 | 4.2 | 1 268 | 0.92 |
| Page not found (404/410) | 32 592 | 3.6 | 23 097 | 2.9 | 9 669 | 7 |
| Timeout | 23 274 | 2.6 | 22 164 | 2.8 | 1 925 | 1.4 |
| Server error | 21 768 | 2.4 | 20 528 | 2.6 | 2 373 | 1.7 |
| Access denied (401/403) | 4 687 | 0.51 | 3 866 | 0.48 | 956 | 0.7 |
| Bad certificate | 321 | 0.035 | 95 | 0.012 | 243 | 0.18 |
| Unavailable for legal reasons (451) | 8 | <0.001 | 7 | <0.001 | 1 | <0.001 |
| **Could not retrieve** | 95 933 | 11 % | 56 825 | 7.1 % | 40 381 | 29 % |
| Browser crashed | 46 750 | 5.1 | 13 736 | 1.7 | 33 122 | 24 |
| Empty after pruning boilerplate | 44 817 | 4.9 | 39 657 | 4.9 | 6 197 | 4.5 |
| Network fault | 4 581 | 0.5 | 3 649 | 0.45 | 1 085 | 0.79 |
| Invalid URL | 122 | 0.013 | 111 | 0.014 | 11 | 0.008 |

page to have been taken down, or the server to be malfunctioning, or for the page to be empty after pruning boilerplate (see Section 4.4.2). However, we were able to retrieve something meaningful for 72 % of the pages. (The "snowball" columns of this table will be discussed in Section 5.2.)

## 4.3 Historical data collection

Our historical snapshots were all collected by the Internet Archive's "Wayback Machine" [94]. The Archive began recording Web pages in 1996. They offer HTTP-based APIs for retrieving all the dates where they have a snapshot of a particular page, and then for retrieving the page as they saw it on a particular date. We used these APIs to retrieve snapshots at one-month intervals (whenever possible), running backward in time from the date of our contemporary snapshot to at least one year before the earliest date that the page appears in any of our lists. Like Common Crawl, the Wayback Machine uses a traditional crawler, so there is some loss of fidelity in historical snapshots.

The Wayback Machine has snapshots for 423 265 of the 758 192 pages in this study (55.8 %), but these are not evenly distributed across all of the lists. Figure 4.1 shows how many of the pages on each list have historical snapshots, and how those are distributed over the 10 years before our contemporary data was collected. The Wayback Machine is more likely to collect popular and long-lived websites, and, unfortunately, this means it has less data for the sites on the pinklists and blacklists. As we will discuss in Section 4.7, we have enough data to predict the lifetime of a page

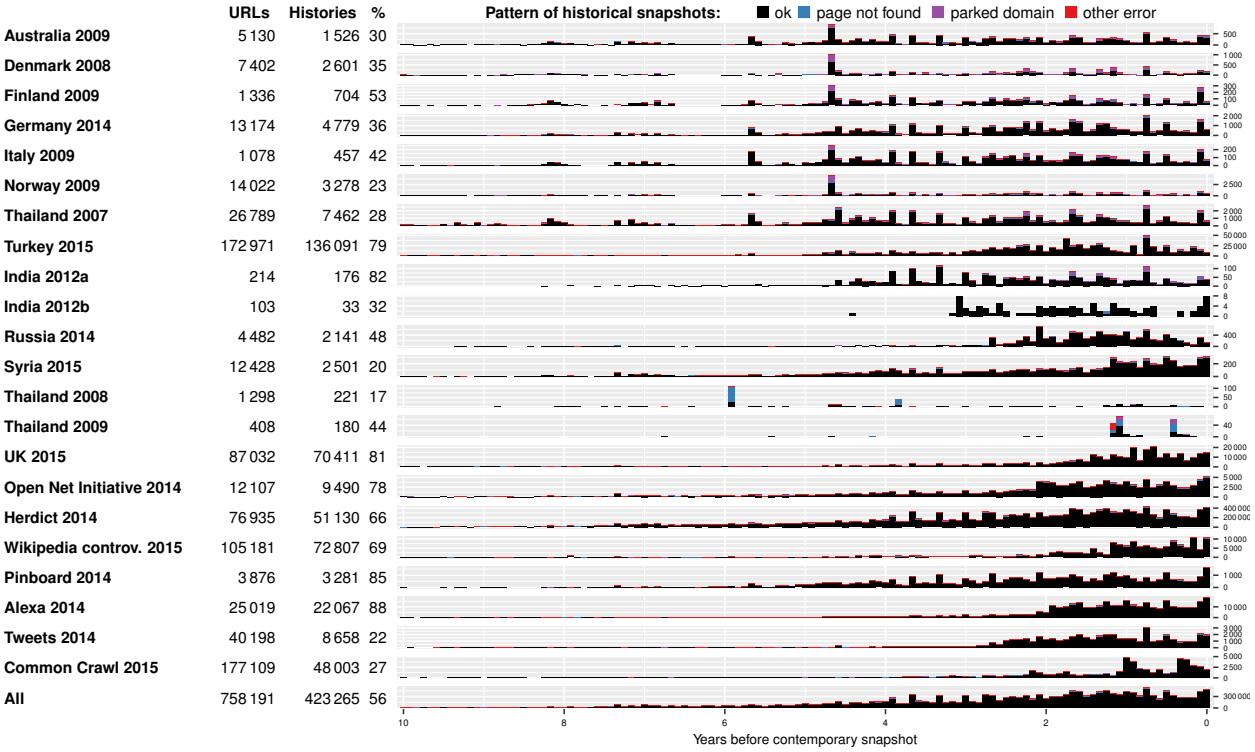



Figure 4.1: Availability of historical snapshots for the pages on each source list

as a function of its source category, but not individual sources, topics or languages.

"India 2012a" appears to be well-collected, but this is an artifact. That list consists mostly of YouTube videos; YouTube is extremely popular, so the Wayback Machine has good coverage of it. Most of the videos have been removed from YouTube (we suspect this is a case of "DMCA takedown abuse," in which a legal process intended to combat copyright infringement is applied to suppress controversial material [80]), but YouTube's "This video is no longer available" error message is served as a *successful* HTTP transaction (200 OK). Thus, the pages have not truly survived, but the Wayback Machine's API reports that they have. Fortunately, there are only a few hundred pages affected by this artifact.

For 81 988 of the pages (10.8 %), the Wayback Machine records at least one snapshot within 30 days of the earliest date when the page was entered onto one of our lists. It is at this time that the page's topic is most likely to be relevant to its chances of being censored.

## 4.4 Document preprocessing

Having collected pages, we wish to reason about their contents. For example, we wish to assign a topic to each page, and detect when this topic changes. With hundreds of thousands of pages collected, this process must be automated as much as possible.

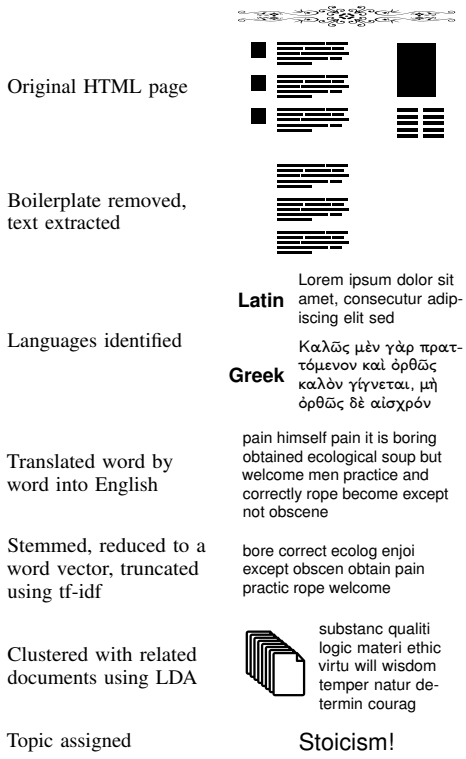



Figure 4.2: Document processing pipeline

The principal technique we use is Latent Dirichlet Allocation (LDA) clustering [24]. Our analysis pipeline (illustrated in Figure 4.2) includes several heuristic filtering steps before LDA, which remove irrelevant "boilerplate" and reduce the cost of model training. These are described in this section. LDA, and the uses we put it to, are described in Chapter 4.

### 4.4.1 Parked domain detection

A *parked domain* is a placeholder website, operated by a *domain parker*, who hopes to sell the domain name eventually. It contains no meaningful content, only advertising banners and links [175]. Domain parkers often claim the names of abandoned websites, hoping to gain from visits by people looking for the former occupant. The placeholder site will bear little or no relationship to the content of the website that used to be there. However, it may parrot keywords from search queries that lead to the site. This confuses LDA, which cannot tell that the words are being repeated meaninglessly.

Therefore, we identify parked domains using a dedicated, heuristic classifier and exclude them from topic analysis. We tested two such classifiers from the literature [164, 175] and selected the one that performed best on our data. Vissers, Joosen, and Nikiforakis [175] developed a random-forest classifier [26] based on features extracted from the HTML and the HTTP transactions at page load time. It relies on structural differences between a parked domain and a normal domain, such as

Table 4.4: Performance of the two parked-domain detectors on three data sets.

| Algorithm | Random forest | | | Regexps | | |
|---|---|---|---|---|---|---|
| Dataset | PS | LT | Cen | PS | LT | Cen |
| Accuracy | 97.9 | 93.1 | 89.0 | 95.0 | 89.6 | 99.0 |
| Precision | 99.2 | 89.9 | 42.9 | 99.9 | 96.9 | 100.0 |
| Recall | 96.9 | 92.1 | 30.0 | 90.4 | 79.1 | 90.0 |

the ratio of text to markup, the ratio of internal to external links, and the number of nested pages ("frames"). Szurdi *et al.* [164], developed a set of regular expressions based on the templates used by specific domain parkers, while investigating the related practice of *typosquatting*. Typosquatters place often-malicious sites at misspellings of the names of popular websites, e.g. `googel.com` for Google.

We evaluated our classifiers on three data sets, "PS," "LT," and "Cen." PS is the set used by Vissers, Joosen, and Nikiforakis [175] to assess their classifier. It includes 3 047 non-parked domains taken from Alexa (see Section 4.1), and 3 227 parked domains operated by 15 parkers. LT was used by Szurdi *et al.* [164] to evaluate their classifier. It consists of 2 674 pages collected from typosquatted domains, and manually labeled; 996 are parked and 1 678 not parked. Finally, Cen consists of 100 pages randomly selected from our contemporary data.

To train the random-forest classifier, we combined PS and LT, and then split the combination 80/20 for training and testing. Neither LT nor Cen includes HTTP transaction information, so the features depending on this data were disabled. Despite this, we reproduce Vissers, Joosen, and Nikiforakis's results on PS, which indicates that those features are not essential. The regexp classifier does not require training, but we augmented the original battery of regular expressions with new rules derived from PS.

Table 4.4 shows the performance of both classifiers on all three datasets. The random-forest classifier performs reasonably well on PS and LT, with precision and recall both 90 % or above, but poorly on Cen: precision drops to 42.9 % and recall to 30.0 %. (Accuracy remains high because Cen is skewed toward non-parked pages.) The (improved) regular-expression classifier, on the other hand, performs well on all three; its worst score is 79.1 % recall for LT.

To better understand why the random-forest classifier performs poorly on Cen, we constructed a larger version of it containing 7 422 pages. Both classifiers agreed on 6 869 of these: 81 parked, 6 788 not parked. 447 pages were classified as parked only by the regular-expression classifier, and 106 pages only by the random-forest classifier. We manually verified a subsample of 25 pages in each category. In all cases, the regular-expression classifier was correct; where they disagreed, the random-forest classifier was invariably wrong. The most common cause of errors was pages using frames to load most of their content. The random-forest classifier treats this as a strong signal that

the page is parked, but this is inaccurate for Cen.

### 4.4.2 Boilerplate removal

Nearly all HTML documents contain "boilerplate" text which has little or nothing to do with the topic of the document itself, such as site navigation menus, copyright notices, and advertising. It may not even be in the same language as the document's main content [161]. Boilerplate varies only a little from site to site, so it can confuse semantic analysis algorithms into grouping documents that are unrelated. This problem has been recognized since 2002 [18] and the solution is to strip the boilerplate from each document prior to semantic analysis. Unfortunately, the most widely used algorithms for stripping boilerplate, such as Readability [148] and the similar "reader view" features in Chrome, Firefox, and Safari, depend on the standard semantics of HTML elements. In a large sample of not necessarily well-structured documents, this is not a safe approach. Some algorithms also make strong assumptions about the document language [67] or require several pages from the same site [161].

We developed a hybrid of the boilerplate removal algorithms described in Lin, Chen, and Niu [120] and Sun, Song, and Liao [163]. These are completely language-neutral, use HTML element semantics only as hints, and in combination, require no manual tuning. Their basic logic is that heavily marked-up text is more likely to be boilerplate.

The hybrid algorithm merges subtrees of the parsed HTML into a tree of "blocks," each of which represents a contiguous run of text. Blocks are bigger than paragraphs but usually smaller than sections. Each block is assigned a *text density* score, which is the total number of text characters in the block, divided by the logarithm of the total number of markup characters in the block. Stylistic markup (bold, italic, etc.) does not count, and invisible HTML elements (scripts, etc.) are completely discarded. After the entire page has been scored, the algorithm identifies the *least* dense block that contains the *most* dense block. This block's density score is the "threshold." Every block that is less dense than the threshold (that is, it contains more markup and less text) is removed from the page. Finally, all remaining markup is stripped, leaving only text.

### 4.4.3 Language identification and translation

LDA topic models detect semantic relationships between words based on their co-occurrence probabilities within documents. Therefore, it is necessary for all documents to be in the same language. Multi-lingual versions of LDA exist, but they are either limited to two languages [25], or they require all documents to be available in all languages, with accurate labeling [128]. Our data meets neither condition, so instead we mechanically translated as much text as possible into English.

After boilerplate removal, we used CLD2 [160] to determine the languages of each document and divide multilingual documents into runs of text in a single language. We then used Google Translate's API [78] to translate as much text as possible into English. At the time of writing, CLD2 can detect 83 languages with accuracy higher than 97 %, and Google Translate can translate 103 languages into English; neither set is a superset of the other. 11.5 % of all words were unrecognized or untranslatable; the bulk of these were nonwords (e.g. long strings of digits) and errors on CLD2's part. In a bilingual document, for instance, CLD2 frequently gets each split point wrong by a couple words, or tags small runs of one language as "unknown." Only 29 234 documents (0.8 %) were completely untranslatable.

Google charges US$20 to translate a million characters. After boilerplate removal, the 4 355 234 unique pages in our database (including both contemporary and historical snapshots) add up to 13.3 *trillion* characters; translating each document in full would have cost $260 000, which was beyond our budget. Instead, we reduced each document to a "bag of words," and then translated each word in isolation, which cost only $3 700. This required us to "segment" text into words, which is nontrivial in languages written without spaces between the words. For Chinese we used the Stanford segmenter [37]; Japanese, Mugabe [116]; Vietnamese, dongdu [11]; Thai, libthai [98]; Arabic and related languages, SNLP [129]; all others, NLTK [23].

Because our data set is so large, we needed to truncate the translated word vectors to complete training in a reasonable amount of time. After translation, we reduced all words to morphological stems using the Porter stemmer [144]. We then used *term frequency-inverse document frequency* (tf-idf, [152]) to select terms with a high degree of salience for each document, preserving terms whose combined *tf-idf* constituted (at least) 90 % of the total. After pruning, the median size of a word vector was 37 words.

### 4.4.4 Language biases of sources

Figure 4.3 shows, for each source list, what proportion of its non-boilerplate text is in each of the most commonly used 21 languages (plus "other," "unrecognized," and "untranslatable"). English, unsurprisingly, dominates nearly all of the lists—the interesting case is when it does not dominate, as in the Russian blacklist. We suspect this might also occur for Chinese, if we had a Chinese blacklist. Where a single language dominates non-English text, it is also unsurprising: German for Germany, Russian for Russia, Arabic for Syria, Thai for Thailand. ONI, Herdict, Wikipedia, Alexa and Twitter show no dominant second language, again as expected.

Four lists have hardly anything *but* English. India 2012b and Thailand 2009 are dominated by videos posted to YouTube and similar sites, for which the common language for leaving comments is English; the videos themselves may well have featured other languages. Most of these videos

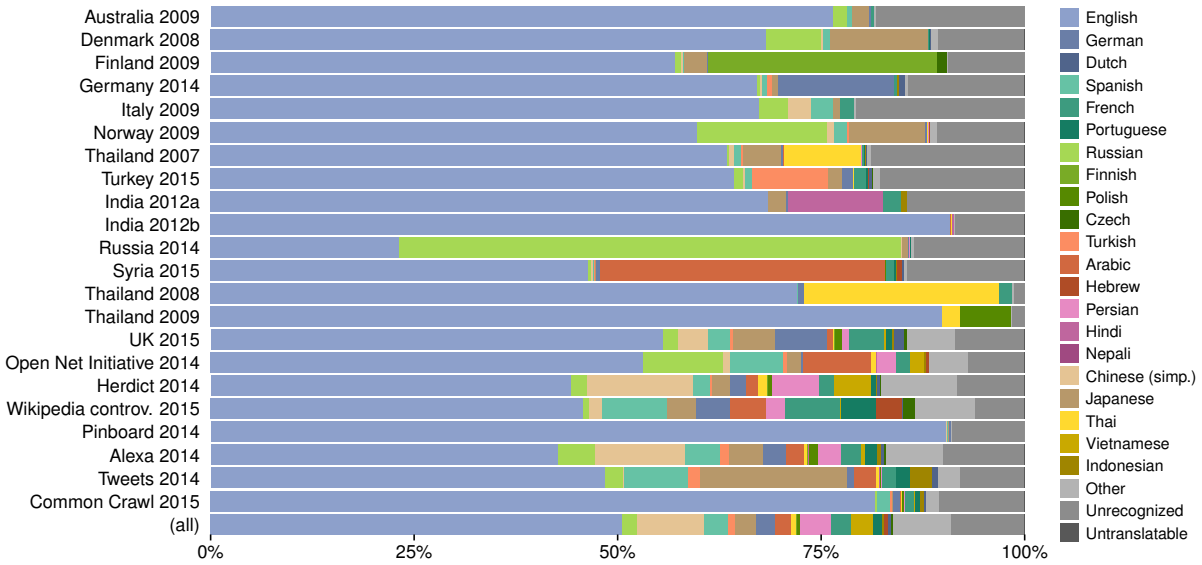



Figure 4.3: The proportion of each source list devoted to text in each of the most common 21 languages.

have since been taken down, so we could not spot-check them. Pinboard, the negative control, was compiled by someone who is only fluent in English. This may mean that our topic model is largely ignorant of *innocuous* material in languages like Russian and Arabic, but this is harmless for now. It will become a problem in the future, when we attempt to make automated judgments about whether pages are worth including in a continuously updated probe list.

Finally, the dominance of English in the Common Crawl data is inexplicable and disturbing. It may indicate a methodological error, either on the part of Common Crawl itself, or in our selection of a subsample. One possibility is that there are too few *sites* in our subsample, and those sites are largely Anglophone. Another is that their crawler may not have started from the right "seed" locations within the hyperlink graph to find much material in other languages. Regardless of the cause, though, this casts doubt on our assumption that this subsample is a good baseline for cross-list comparison. Anything that is in other languages may seem more unusual than it is, by comparison.

## 4.5 Topic assignment

We used the MALLET implementation of LDA [125, 134, 197] to cluster documents into topics.

We used the contemporary data for training and selecting the topic models. Half of the collected documents were used for training, and the remainder were used for model-selection. We trained models with $N \in \{100, 150, \ldots, 250\}$ topics and $\alpha \in \{0.1, 0.5, 1, 5, 10, 100\}$ ($\alpha$ controls the sparsity of the topic assignment), and selected the max-likelihood model, following the procedure described by Wallach *et al.* [176]. We found the parameters $N = 100$ and $\alpha = 5$ to be optimal.

After model training, two researchers reviewed the top words associated with each topic and labeled the topics. A colleague not otherwise involved with the research scored inter-coder agreement between the labels, which came to 87 %. Disagreements were resolved by discussion between the researchers.

To capture the complexities of modern web pages (e.g., dynamically updated contents, mashups, etc.), rather than assigning a single topic to each given page, we assigned it a vector of probabilities over all $N$ topics. For instance, a news website front-page containing article snippets about sports and politics would have those topics ("sports," "news," "politics") assigned relatively high probabilities, perhaps 0.2, 0.4 and 0.35. Other topics would receive probabilities very close to zero.

LDA found several topics with identical labels. This is a known limitation of LDA when the training data set is skewed toward certain topics. The algorithm will split those topics arbitrarily in order to make all of the clusters roughly the same size [194]. We solved this problem by manually merging topics that have similar labels and summing their probabilities. For instance, suppose that topics 26 and 61 were both labeled "news," and that a page has probability 0.24 for topic 26 and 0.56 for topic 61. These topics would be combined into a single "news" topic, and the page's weight for the combined topic would be 0.8.

Using this procedure, our initial set of 100 topics was reduced to 64 merged topics, and then further grouped into nine categories. Two artificial topics were added to account for documents that could not be processed by LDA at all. The final set of topics and categories is shown in Table 4.5 along with measures of the bias of lists and languages toward each topic, which will be discussed further in the next section.

## 4.6 Topic-source correlations

To examine the correlation of topics with source lists and languages, we apply $\chi^2$ tests of independence to the contemporary data set. Overall tests strongly confirm the hypotheses that the distribution of documents over topics is correlated with source list and with language. Coincidentally, both tests have 1 365 degrees of freedom. For topic × source, $\chi^2 = 6.45 \times 10^5$ ($p < 0.001$), for topic × language, $\chi^2 = 1.33 \times 10^6$ ($p < 0.001$). We then perform post-hoc $\chi^2$ tests on each combination of list and topic, or language and topic, using a $2 \times 2$ contingency table of the form

$$\begin{array}{c|c} w_{gt} = \sum_{u \in g} \mathbf{T}_{ut} & w_{g\neg t} = |g| - \sum_{u \in g} \mathbf{T}_{ut} \\ \hline w_{rt} = \sum_{u \in r} \mathbf{T}_{ut} & w_{r\neg t} = |r| - \sum_{u \in r} \mathbf{T}_{ut} \end{array} \tag{4.1}$$

where $g$ is the selected list or language, $r$ is the *reference* list or language (see below), and $\mathbf{T}_{ut}$ is the probability that page $u$ belongs to topic $t$. (Recall from Section 4.5 that each page is assigned a

probability vector over all topics.) There are a total of 2 904 such combinations. After Bonferroni correction, 585 of the topic-list correlations and 580 of the topic-language correlations are significant at the usual $\alpha = 0.05$ level.

However, significant correlations might still be too small to be interesting. Rather than show significance by itself, therefore, we compute the *odds ratio* for each significant cell. This statistic can be computed directly from the $2 \times 2$ contingency table above:

$$r_{t;g,r} = \frac{w_{gt}/w_{g\neg t}}{w_{rt}/w_{r\neg t}} \tag{4.2}$$

It is one when there is no difference between the source and the reference, greater when the group has more pages on a topic than the reference does, and smaller when it has fewer. In Table 4.5, we show the odds ratio for each significant comparison. Again, this considers contemporary page contents only. Blank cells are non-significant. Shades of red indicate that the topic is positively correlated with the list or language, and shades of blue indicate that it is anti-correlated.

In the left half of Table 4.5, we correlate the topics with the source lists, taking Common Crawl as the reference (we believe this to be the most topic-uniform of our lists; but see Section 4.4.4 for evidence to the contrary). When two lists have red cells for the same topic, that indicates a commonality between the lists. However, when two lists have blue cells for the same topic, that means only that neither is correlated with that topic, which does not qualify as something they have in common.

We can immediately see the same three clusters of source lists that appeared in Tables 4.1 and 4.2. The blacklists have more in common with the potentially-censored lists and the controls, but when it comes to the most politically controversial categories (news, politics, religion, etc.) they tend to be concentrated on one or two specific topics, whereas the potentially-censored lists are spread over many such topics. In some cases it is apparent that a country is censoring news related to *itself*, but not other news. The Syria 2015 list includes a surprisingly large number of software-related sites; spot checking indicates that this is due to indiscriminate blocking of websites with Israeli domain names.

The blacklists also devote more attention to specific entertainment topics than the potentially-censored lists do. Social media in particular stands out, consistent with external evidence that this is increasingly seen as a threat by censors [52]. Blocking access to video-sharing and other entertainment sites may also be meant to suppress copyright infringement and/or support local businesses over global incumbents [106].

Pornographic topics are concentrated in the pinklists and underrepresented elsewhere. All of the pinklists have some non-pornographic pages. Some of these can be explained by poor classification of image-heavy pages, and by debatable classification of, for instance, “mail-order bride” sites.

Table 4.5: Correlation of topics with languages and source lists.

However, we do see a genuine case of political censorship under cover of porn filtering: many of the pages on the Thailand 2007 list that were filed under Japan, Vietnam, or social media are discussing Southeast Asian regional politics. This was known from previous case studies of Thailand [168] and is exactly the phenomenon we designed our system to detect.

The negative control (Pinboard) is almost perfectly anti-correlated with the blacklists and pinklists. There is some overlap on software topics. This is largely due to the negative control being strongly biased toward those topics, so any software topics at all in any of the blacklists will show as overlap. Also, software-industry-focused news sites tend to be hostile to attempts to censor the Internet.

The other controls have more in common with Common Crawl than with the blacklists or pinklists. They also have more in common with the "probably censored" lists than the blacklists or pinklists; here we see that popular pages are more likely to get cited on Wikipedia or have someone bother to report an outage to Herdict. The over-weighting of some regional news topics in this group of lists may also indicate biases in Common Crawl (see Section 4.4.4).

In the right half of Table 4.5, we correlate topics with the 21 most commonly used languages in our data set (and with "other languages"), taking English (which is far and away the most common) as the reference. The same caution about paired red versus blue cells applies.

News topics for specific countries are very strongly correlated with the languages spoken in those countries, and Islam correlates with languages spoken in countries where it is the most or second-most common religion. Many of the more "commercial" topics are dominated by English; this may be an artifact of data collection from the USA, since commercial sites often change their language depending on the apparent location of the client.

The "junk" topics at the bottom of the table collect various documents that we could not interpret meaningfully. Despite our efforts to weed them out early, some error messages (perhaps served with the wrong HTTP response code, so the crawler does not detect an unsuccessful page load) and web page boilerplate creep through. Mistranslated, unintelligible, and empty documents are self-explanatory. Finally, documents that we were unable to translate are in languages that Google Translate does not support, or (in the case of Japanese) suffered from a character encoding problem that made them *appear* untranslatable to the automation. The high concentration of error messages on the pinklists probably reflects the short lifetime of pages on these lists (see Section 4.7 for more on this); the untranslatable documents on the pinklists may be an artifact of porn sites carrying far more imagery than text; the mistranslations on the blacklists probably indicate weak support for colloquial Russian and Arabic in particular.

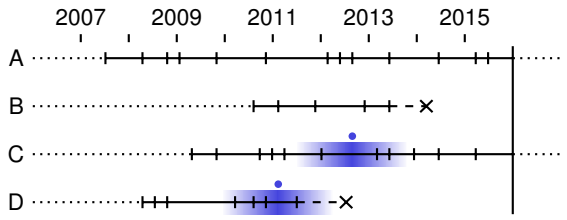



Figure 4.4: The life cycles of four hypothetical websites. Tick marks are Wayback Machine snapshots; the vertical bar is our contemporary data capture; blue shaded areas indicate possible censorship events. In survival analysis jargon, we must account for "delayed entry" because we do not know how long the sites existed before their first snapshot (dotted lines at left) and "right-censoring" because some of the sites still exist at the end of the study (dotted lines at right).

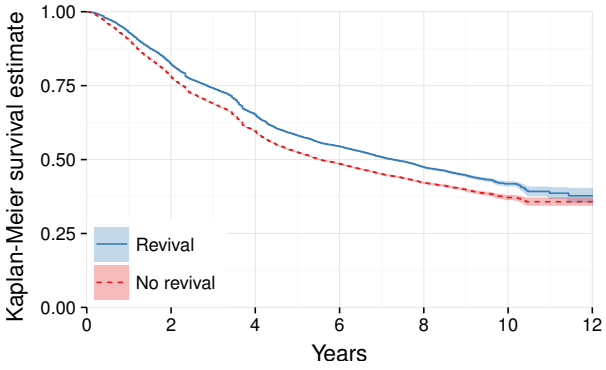



Figure 4.5: Comparison of the two approaches to page revival. Shading shows confidence intervals.

## 4.7 Survival analysis

Our data on the life cycle of websites is unavoidably incomplete. Figure 4.4 shows four hypothetical cases which illustrate the problem. In no case do we know when a page was created, only when it first came to the attention of the Wayback Machine. If a page survives to the present (A, C), we do not know how much longer it will continue to exist. If it was abandoned (B, D), we only know that this happened within an interval between two observations. If a page appears on a censorship blacklist (C, D), we know when this happened (blue dot) but we can only guess at how long the page was censored (blue shaded area).

*Survival analysis* [97] is a set of statistical techniques originally developed for predicting the expected lifespan of patients with terminal illnesses. Because medical studies often suffer from exactly the same kinds of gaps in their data—one does not usually know how long a tumor was present before it was diagnosed, for instance—survival analysis is prepared to deal with them. However, to use these techniques we must define what it means for a page to "die." Clearly, if the site is shut down or turns into a parked domain, that should qualify. Less obviously, we also count topic change as "death." This is because, after a censored page has changed topic, it no longer

provides the same kind of sensitive material that it used to, so it can no longer be considered "fresh." For example, the blog `http://amazighroots.blogspot.com` was taken over by spammers in 2014, and its apparent topic changed from "news and politics" to "food." Probing such pages longer reveals whether the censor cares about that kind of sensitive material. It may instead reveal how diligent the censor is about updating their blacklist, but this was not the original goal.

We modeled page survival curves using Kaplan-Meier estimators [97], allowing for delayed entry (survival analysis jargon for not knowing when the pages were actually created) and right-censoring (jargon for some of the pages surviving into the future—an unfortunate case of homonymy). When death events were only known to have occurred within some interval, we substituted the midpoint of the interval. We compared survival curves using log-normal tests and Cox proportional hazard models.

Unlike medical patients, web pages may be "dead" only temporarily, due to server crashes, vandalism, the owners forgetting to renew the domain registration, and so on. Stock survival analysis does not allow for this possibility. To handle it, we calculated every survival curve two ways: first, assuming that revivals never happen (once a site "dies" it is treated as staying that way, even if we have evidence to the contrary) and second, allowing sites to revive even after an arbitrarily long hiatus. The first approach gives an underestimate of survival times, the second, an overestimate. Figure 4.5 shows, over all pages we monitored, that on average, the differences between both approaches are relatively small; and that our error ranges are small.

### 4.7.1 Detection of topic changes

As discussed in Section 4.7, pages can cease to be relevant to a censorship probe list by being taken down entirely, or by changing their topic to the point where they no longer contain sensitive material. Therefore, to evaluate the freshness of a probe list we need to detect topic changes.

Existing algorithms for detecting *any* change in a web page (e.g. [35]) are too sensitive for our purposes. Even looking for changes in the most probable topic chosen by LDA is too sensitive. The most probable topic assigned to the front page of a news website changes several times every day, as new articles appear, but it is still the front page of a news website.

Instead, we compare the entire sets of probable topics that were assigned to a pair of observations. Specifically, if $T_1$ and $T_2$ are the topic probability vectors assigned to a pair of observations, let $S_1 = \{i : T_{1i} \geq p\}$ and $S_2 = \{i : T_{2i} \geq p\}$, that is, the respective sets of topic indices for which the assigned probabilities are greater than $p$. Then, the page's topic is judged to have changed if $|S_1 \cap S_2| < m$, that is, the intersection of the topic sets is smaller than $m$.

This algorithm has two parameters, $p$ and $m$. Its performance is also affected by the LDA sparsity parameter $\alpha$ and the number of topics $N$. To tune these parameters, we used a manually

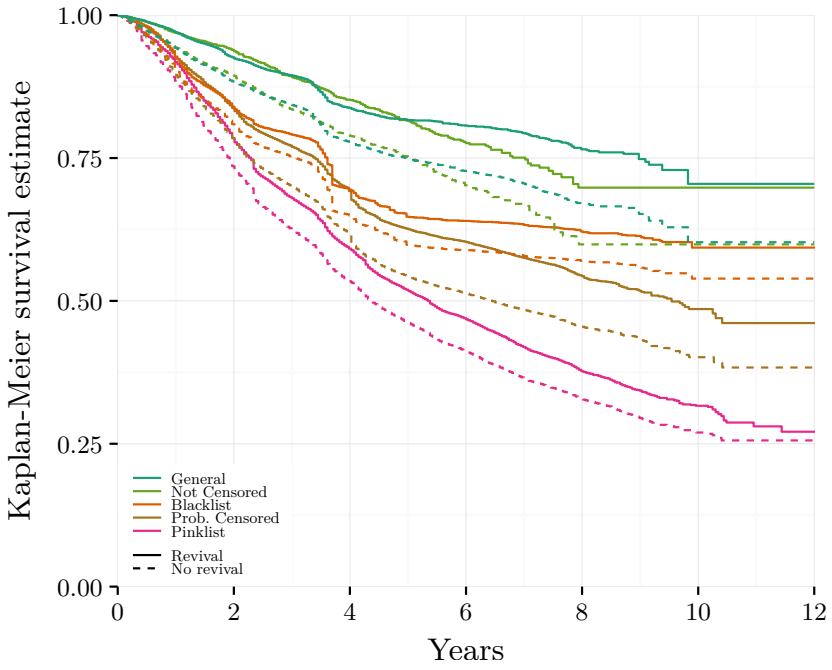



Figure 4.6: Kaplan-Meier curves for different lists.

chosen development set of 30 pairs of observations whose topics were the same, and another set of similar size of observation pairs whose topics were different. We found that for $p = 0.05$, $m = 1$, $\alpha = 10$, and $N = 100$ the algorithm achieved perfect accuracy on the development set. On a second randomly-chosen evaluation set, with 50 pairs of observations whose topics were the same and 50 whose topics were different, the algorithm achieved 97.8 % recall and 86 % precision.

### 4.7.2 Results

As mentioned in Section 4.3, the limited data we have from the Wayback Machine is insufficient to compute Kaplan-Meier curves for each topic, or each source list. Only 55.8 % of the URLs have any historical data at all, and the median number of historical snapshots per URL is only 3, with large gaps between observations. We do have enough data to compute K-M curves for each group of source lists (Figure 4.6) and each category of topics (Figure 4.7). These larger-scale clusters correspond to the horizontal and vertical divisions of the left half of Table 4.5, plus an extra topic category just for HTTP errors. It should be said, however, that the large gaps mean that all these curves probably overestimate survival times.

From these curves, we can see that pages hosting sensitive material (pinklist, blacklist, and probably censored; porn, software, entertainment, video, news) are significantly shorter-lived than less sensitive web pages, with the pinklist pages faring worst. (The especially short lifetime of the "error" category reflects that once pages start turning into error messages, the entire site is likely to

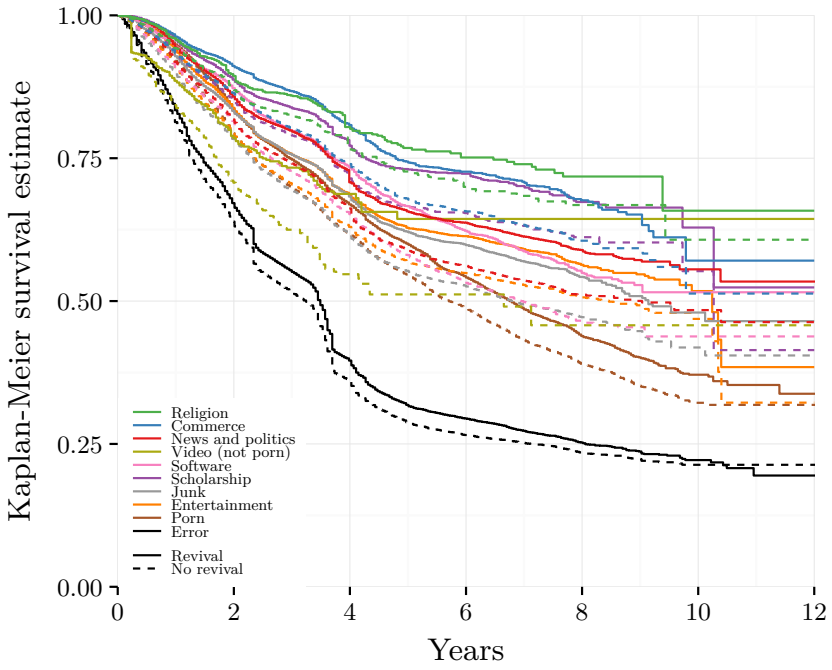



Figure 4.7: Kaplan-Meier curves for different categories of topic.

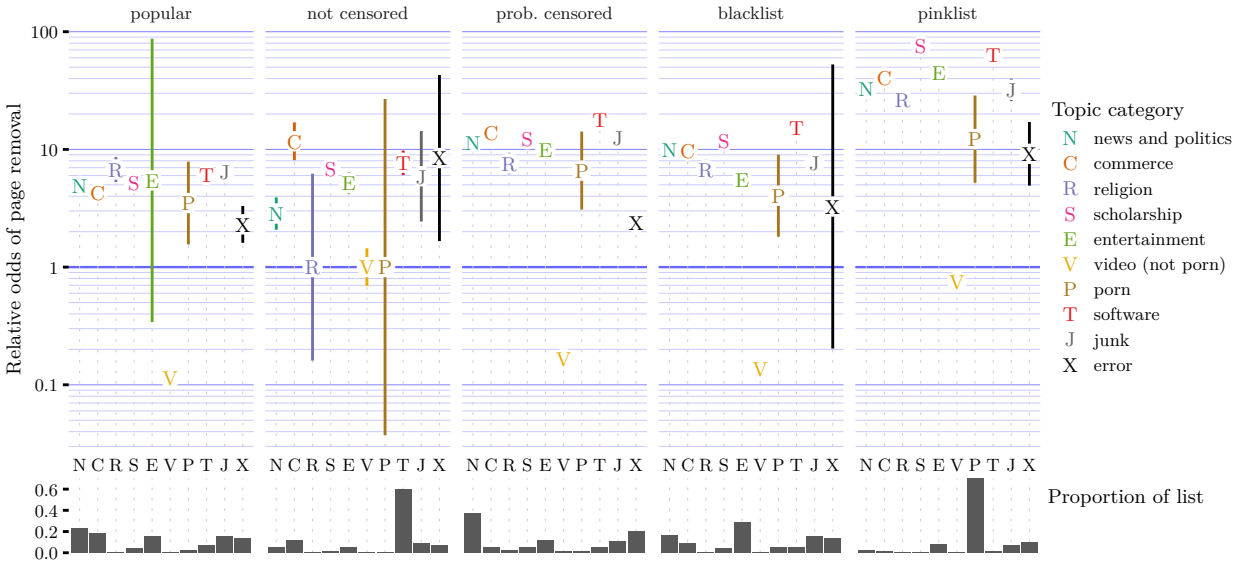



Figure 4.8: How likely pages are to be taken down compared to Common Crawl pages. Error bars show 95 % confidence intervals.

go away.) This in itself demonstrates the need for frequent updates to probe lists.

To reveal how lists and topics interact in determining page lifetime, we use a two-variable Cox proportional hazard model:

$$h_i(t) = h_0(t)e^{\beta_1 L_i + \beta_2 T_i} \tag{4.3}$$

where $L_i$ is the type of page (blacklisted, etc), $T_i$ is the topic category, $h_0$, $\beta_1$, and $\beta_2$ regression coefficients, and $h_i$ the hazard rate at time $t$. Using this model, Figure 4.8 compares the odds of death of each type and category of page with those in the Common Crawl list. Each panel of this figure compares a group of source lists to Common Crawl; within each panel, there is a letter indicating the odds ratio, with 95 % confidence interval, for each group of topics. Larger numbers on the y-axis indicate greater chances of a page being removed from the net. When the confidence interval is large, this means either that we have very little data (for instance, the "not censored" list group has fewer than 10 sites in the "religion" and "porn" categories) or that the lifetimes of pages in some category vary widely (for instance, "popular/entertainment").

This analysis confirms that, regardless of their topic, pages listed on the pinklists are more likely to be removed than pages on any other set of lists. Popular pages are somewhat less likely to be removed, but not as much as one would expect; this is probably because these lists include a fair number of sites that were only popular for the proverbial fifteen minutes. Curiously, non-pornographic video is *less* likely to be taken down than anything else; this may reflect the durability of video-sharing sites, which have to make a substantial infrastructure investment just to get started.

Another curious result is that, while pinklisted porn has a very short lifetime ($\beta_1 = 0.715$, $p < 0.01$), porn in general has a *longer* lifetime than the average ($\beta_2 = -0.123$, $p < 0.01$) (page revival allowed in both cases). This may reflect a fundamental dichotomy within the category. Legal pornography (in the USA) is a large and well-funded industry, capable of maintaining stable websites for years on end. However, there is also a substantial gray market, consisting of many small-time operations looking to make money fast and not so much concerned with law or ethics. These sites turn over very quickly indeed, and are perhaps also more likely to come to the negative attention of a censorship bureau.

## 4.8 Future work

This study demonstrates the power of natural language processing to reason about the contents of collected web pages, even using crude approximations such as bag-of-words document representations and dictionary-lookup translation. However, context-aware translation would almost certainly improve our classification, considering that people often use metaphors and ellipses to get around keyword blacklists [43]. We can also refine our topic model by using information which is

already collected but not analyzed, such as words found in the URL of the site and in its outbound links. And the "web page boilerplate" and "error message" topics demonstrate that our various preprocessing heuristics could still be improved.

The large number of languages present in our data set poses unique challenges. 11.5 % of all the words were either unrecognized by CLD2, untranslatable by Google, or both. This is a larger fraction of the data set than any single language other than English. Some of it is nonwords (e.g. strings of symbols and numbers) but, obviously, more comprehensive language resources would be better. In addition, segmentation tools are not available for all of the languages that are written without spaces between the words. In our data set, the most prominent lacuna is Tibetan.

If the legal obstacles can be resolved, augmenting the topic model with information from images might be an interesting experiment. The state of the art in machine classification of images is well behind that for text, but is advancing rapidly. We suspect that many of the presently unclassifiable pages, especially those where no text survives boilerplate removal, are image-centric.

Finally, our statistical analysis of topic correlation with sources relies on the assumption that Common Crawl is topic-neutral. Unfortunately, Common Crawl is more strongly biased toward English than most of our other sources (see Section 4.4.4) so this assumption is suspect. Alternatives exist, but they have their own deficiencies. For example, one can uniformly sample hostnames from the top-level DNS zones, but this only discovers website front pages. Developing a better "uniform" sample of the Web may well be a project in itself.

# 5. Toward Discovery of New Cases

The lists of URLs described in Section 4.1 are not easy to compare, but patterns emerge when the pages are downloaded and analyzed for their topics. Cross-country patterns of censorship are readily detectable by comparing topics; pornography features prominently, but so do social media, music (copyright infringement?), and regional news. Survival analysis of web pages within each topic and each source provides convincing evidence that potentially controversial pages tend to have shorter lifetimes than less sensitive pages. The topic of a page is a significant predictor of its lifespan, and appearing on certain types of lists is also an effective predictor, even when controlling for topic.

Most of the blacklists are heavily weighted toward topics relevant to the countries they came from. This is a point in favor of ONI's split list design, with one set of URLs to test everywhere and then additional sets to test in each country. However, the rapid decay of controversial pages demonstrates that it is imperative to update probe lists frequently.

To achieve both depth and breadth of coverage, one should start with a topic balance across web pages of interest that is consistent with the web at large. If a censor is known to object to specific topics, these topics may deserve to be weighted more heavily; however, the odds of noticing censorship are proportional to the size of the *intersection* of the probe list with the blacklist. Thus, when a censor attempts to block a given topic comprehensively, probe lists need not weight that topic heavily.

A few researchers have experimented with discovery of previously-unsuspected censored pages, mostly using web search for sensitive keywords, possibly extracted from known-censored pages [43, 48, 92]. Darer, Farnan, and Wright [49] takes a different approach, traversing hyperlinks from blocked page and checking whether the destination pages are also censored. All of these projects were able to discover many more censored pages than appear in existing probe lists. However, it is not feasible for censorship monitors to test *every* censored page on a regular basis, so we must decide which of the discoveries should be added to probe lists.

In this chapter, we investigate the possibility of leveraging topic classification to aid in this decision.

## 5.1 Reclassification

As a preliminary step, we revisited the page classification developed in Section 4.5. First, we retrained the LDA model using the same procedure and LDA parameters (100 topics, sparsity

parameter $\alpha = 5$), but this time the training document set was limited to pages that were successfully loaded (HTTP status code 200), not parked, and not empty after boilerplate removal. (Including other pages in the training set was an error in the original experiment.)

The new model's categories and topics are listed in Table 5.1. Reducing the amount of junk in the training set allowed LDA to create more meaningful clusters: the same nine categories appear, but there are now 73 distinguishable topics instead of 64. Comparing with Table 4.5, the "news and politics" category is much the same, but the "commerce," "entertainment," and "scholarship" categories have all grown. Some of the new categories are quite narrow: cars, guns, camping gear, climate change, and Japanese idol singers. The "junk" category still exists, but is now populated with specific types of boilerplate that the generic boilerplate-removal algorithm (Section 4.4.2) did not recognize, such as names of languages, library-catalog jargon, and HTML markup (spot-checking indicates that there are quite a few pages with markup typos, causing tags to be interpreted as text).

Table 5.1: Manually labeled LDA categories for the snowball sample

| Category | Topic | Most salient five words |
|---|---|---|
| news and politics | general | sports health travel weather business |
| | world | australia africa republic kingdom argentina[1] |
| | extremism | war nation democracy communist nazi |
| | finance | invest market finance bank company |
| | government | economy nation government state organization |
| | europe | refugee policy minister europe scandal |
| | balkans | hungary budapest croatia romania viktor |
| | central europe | czech prague republic brno slovak |
| | france | franc french paris pierre jacques |
| | germany | german berlin munich hamburg austria |
| | russia | russia moscow ukraine putin crimea |
| | east asia | japan china beijing taiwan yuan |
| | southeast asia | china thailand lama buddhist king |
| | indonesia, malaysia | indonesia jakarta malaysia java bali |
| | korea | korea seoul lee kim jin |
| | japan | japan tokyo prefecture country minister |
| | vietnam | vietnam nguyen newspaper minh country |
| | india | india delhi mumbai tamil hindi |
| | middle east | arab allah egypt saudi iraq |
| | iran | iran tehran islam republic revolution |
| | israel | israel jerusalem jewish palestine knesset |
| | spain, latin america | spain carlos mexico argentina madrid |
| | usa | florida york obama congress investigation |
| commerce | jobs | offer experience profession work quality |
| | marketing | market business technology solution company |
| | online shopping | ship order price product purchase |
| | social media | email twitter facebook site page |
| | software | server software file windows linux |
| | telecomms | location router content orange vodafone |
| | web design | image html code css design |
| | cars | car vehicle wheel engine auto |
| | cosmetics | skin hair beauty makeup cosmetic |
| | electronics | iphone device tablet phone camera |
| | fashion | shoe shirt accessory bag jacket |
| | guns | gun rifle shoot pistol hunt |
| | home decoration | decor gift furniture craft kitchen |
| | home improvement | light steel water heat storage |
| | outdoors gear | merrel wigwam camper durango teva |

[1]The full keyword list included many more country names.

Table 5.1: Manually labeled LDA categories for the snowball sample

| Category | Topic | Most salient five words |
|---|---|---|
| scholarship | agriculture | plant species seed garden farm |
| | biology | gene genome dna protein virus |
| | climate change | climate earth temperature ocean carbon |
| | education | student school college teacher library |
| | engineering | process system method structure effect |
| | genealogy | archive cemetery william richard john |
| | health | health medicine disease treatment patient |
| | history (european) | century ancient roman greek period |
| | law | law court case regulation office |
| | philosophy | fact argument evidence reason belief |
| | psychology | people life speak understand feeling |
| religion | christian | church god christian faith jesus |
| | islam | allah islam prophet muhammad mosque |
| entertainment | celebrities | celebrity star photo fashion actress |
| | celebrities (japan) | tokyo idol princess aya yuki |
| | food | food recipe cook drink fruit |
| | movies | film movie series actor drama |
| | music | music song album artist band |
| | sports | player league sport champion football |
| | television | diary vampire throne dead dome[2] |
| | travel | hotel safari explore travel flight |
| | video games | game xbox adventure playstation wii |
| | video (not porn) | view duration youtube video autoplay |
| | wine | wine winery tasting vineyard bottle |
| porn | general | porn sex cam girl horny |
| | fetish | bdsm corset exotic speculum lactation |
| | stage names | alexi ava brooke lexi nikki |
| | age gates | adult sexual explicit enter minor |
| junk | date indices | september august wednesday sunday post |
| | forum boilerplate | forum post thread register message |
| | html tags | http src img url iframe |
| | language names | deutsch italiano english nederland chinese |
| | library catalog jargon | library worldcat oclc book publish |
| | unintelligible | thing hand good man back |
| | classification failure | |

[2]The full keyword list included many more titles of current TV shows.

### 5.1.1 Comparison with manual classification

The topics listed in Table 5.1 are plausible enough, but they may not be what a human would produce, looking at the same document. We next compare the (new) LDA classification scheme with three different manual classifications of subsamples. One classification was freshly developed by a group of three CMU master's students, as part of a class project. They were given 1 186 URLs, all of which had been detected as censored at some point by ICLab, and instructed to develop a classification scheme which would, in their opinion, best capture the probable motive for the page having been censored. (948 of these URLs were in the probe lists used for the study in Chapter 4, and an additional 143 URLs were discovered by the snowball sample which will be described below.) The set of 20 topics they developed is summarized in Table 5.2.

The Citizen Lab's probe list also provides a manual classification of its URLs; this is, naturally, focused on topics of interest to a human rights watchdog. Their topics are listed in the y-axis labels of Figures 5.1 and 5.2 (note that the sort order is different in each figure).

Finally, we queried the "FortiGuard" URL classification service operated by FortiNet [71] for all of the URLs included in either the manual sample or the Citizen Lab probe lists. This service is sold as part of a "web filter" for corporations, which is the same software as a nation-state censorship system, but on a smaller scale and (presumably) more concerned with content that is "not safe for work" than with political control. Its coverage of sites in non-Western languages can be erratic: for instance, `https://www.bintang.com/`, which is a lifestyle-and-glamour magazine with Alexa rank #161 in Indonesia, is classified as "Meaningless Content" by Fortiguard, probably because no one at FortiNet speaks Indonesian. (By way of comparison, `https://www.cosmopolitan.com/`, a similar publication in English, is classified as "News and Media.")

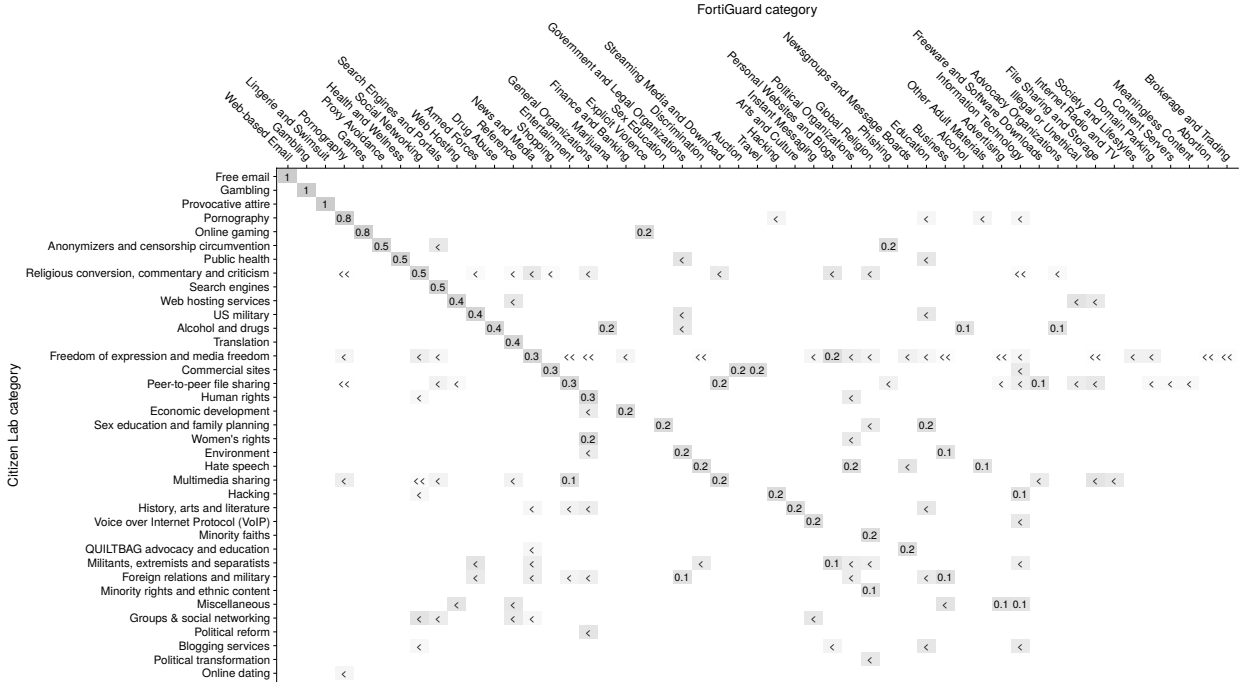
Figures 5.1, 5.2, and 5.3 compare the three manual classifications to each other, using Jaccard indices of similarity for each pair of topics (see Section 4.1.3 for the definition of the Jaccard index). Each matrix has been sorted left-to-right and top-to-bottom by decreasing value of the index, so perfect agreement between two classifications would be visible as nonzero values only on the main diagonal. We do see perfect or near-perfect agreement for some topics, notably gambling, pornography, online gaming, illicit drugs, and anonymizers and circumvention tools. The CMU students did not give "free email service" its own category, but if they had, they would probably have been in complete agreement with Citizen Lab and FortiGuard about that one too.

The cases where the three manual classifications are in poor agreement can be divided into two subcases. Sometimes one side of the comparison uses a single broad topic where the other has many narrow ones: this is most noticeable with Citizen Lab's "religious conversion, commentary and criticism" and "freedom of expression and media freedom" topics, which cover dozens of religion- and news-related topics in both the FortiGuard and CMU manual classifications. Another important

Table 5.2: Manually developed classification scheme for known-censored URLs

1. News and Politics: (Affiliation, Language, Topic)
   e.g. News and Politics: (France, French, Sports).
2. Website selling a product (not including financial services and entertainment. Seven non-exclusive subcategories:
   (a) Does it get pirated? (e.g. music)
   (b) Is it possible to make fakes? (e.g. fashionable clothing)
   (c) Are there lots of individual sellers? (e.g. Ebay, Etsy)
   (d) Is it purely digital? (not a physical good; e.g. software)
   (e) Is it purely a cloud service? (nothing to download)
   (f) Is it an addictive drug?
   (g) Is it medical? (e.g. prescription drugs, mobility aids)
3. Financial services
4. Entertainment: (movie, music, radio, gaming, dating)
5. Copyright Infringement: (torrent, music, movie, software)
6. Gambling
7. Pornography: (straight, gay)
8. Religion: (name of religion, group/personal)
9. Human rights: (organization or country name)
10. Militants, Extremists, Separatists
11. Education: (sex, public health, general)
12. Malware
13. Anonymizers and censorship circumvention tools
14. Government: (country name)
15. Non-religious organizations
16. History, arts, literature
17. Shock site
18. No valuable content
19. No Access: (e.g. 404, 403, host not found, domain for sale)
20. Incomprehensible

Figure 5.1: Jaccard similarity between Citizen Lab and FortiGuard classification



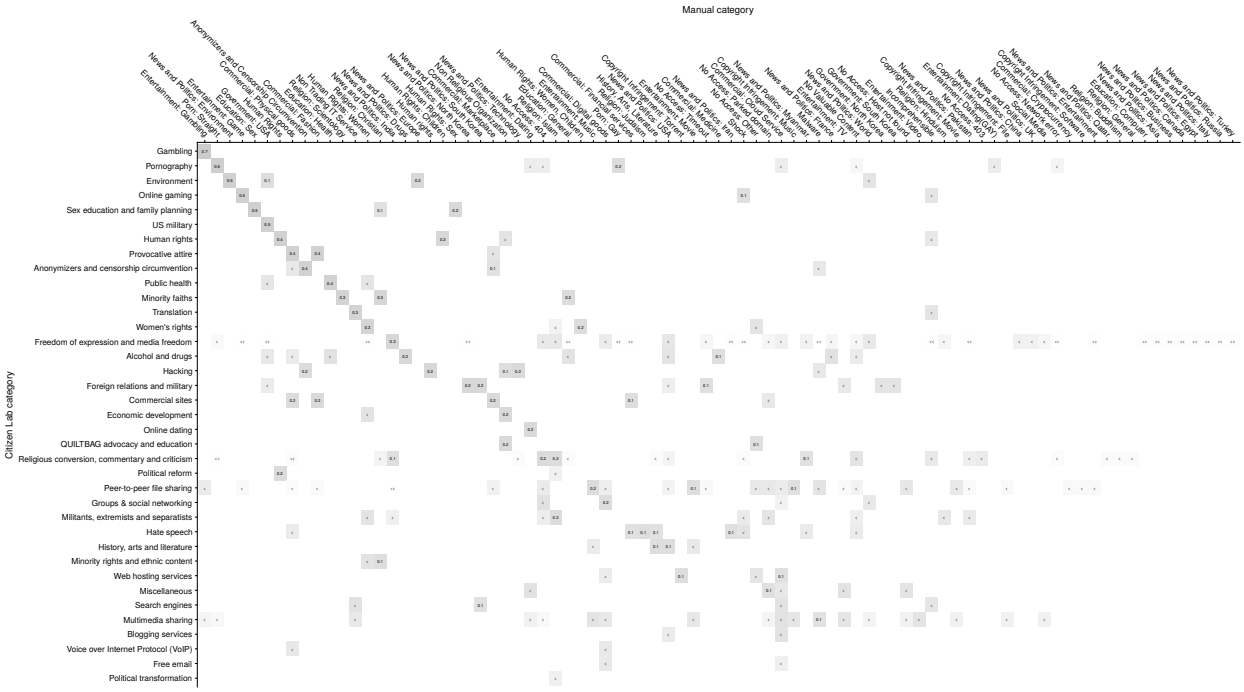




Figure 5.2: Jaccard similarity between Citizen Lab and manual classification

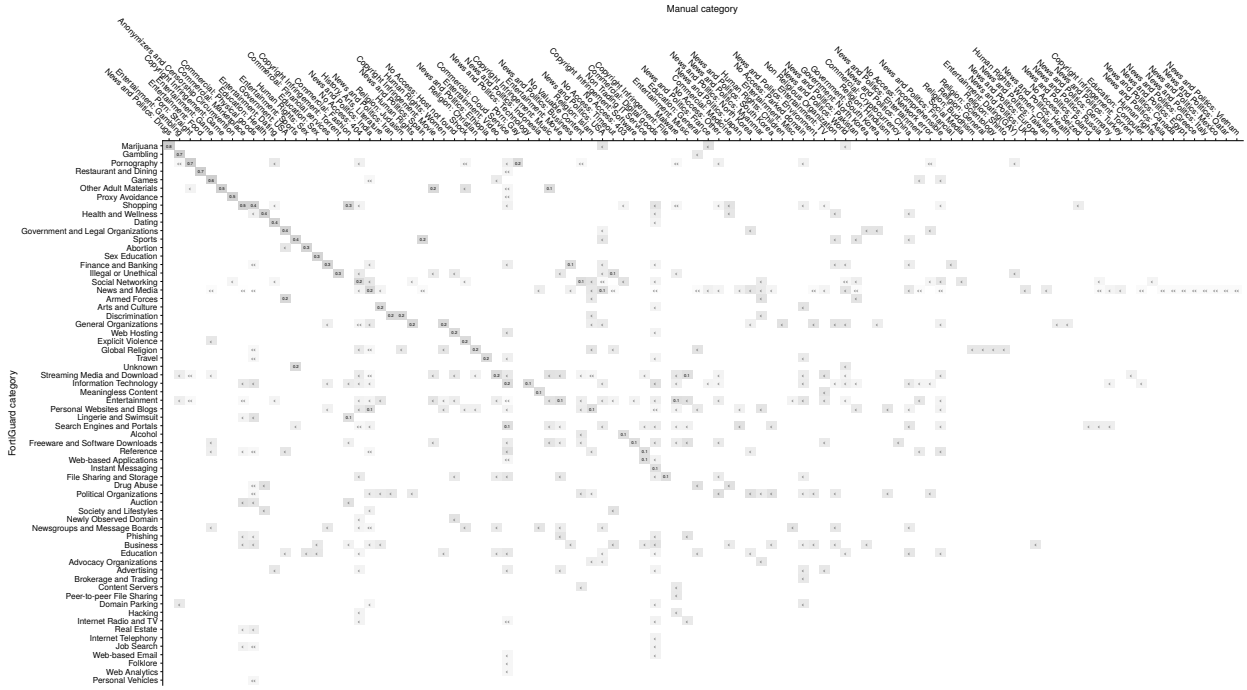



Figure 5.3: Jaccard similarity between FortiGuard and manual classification

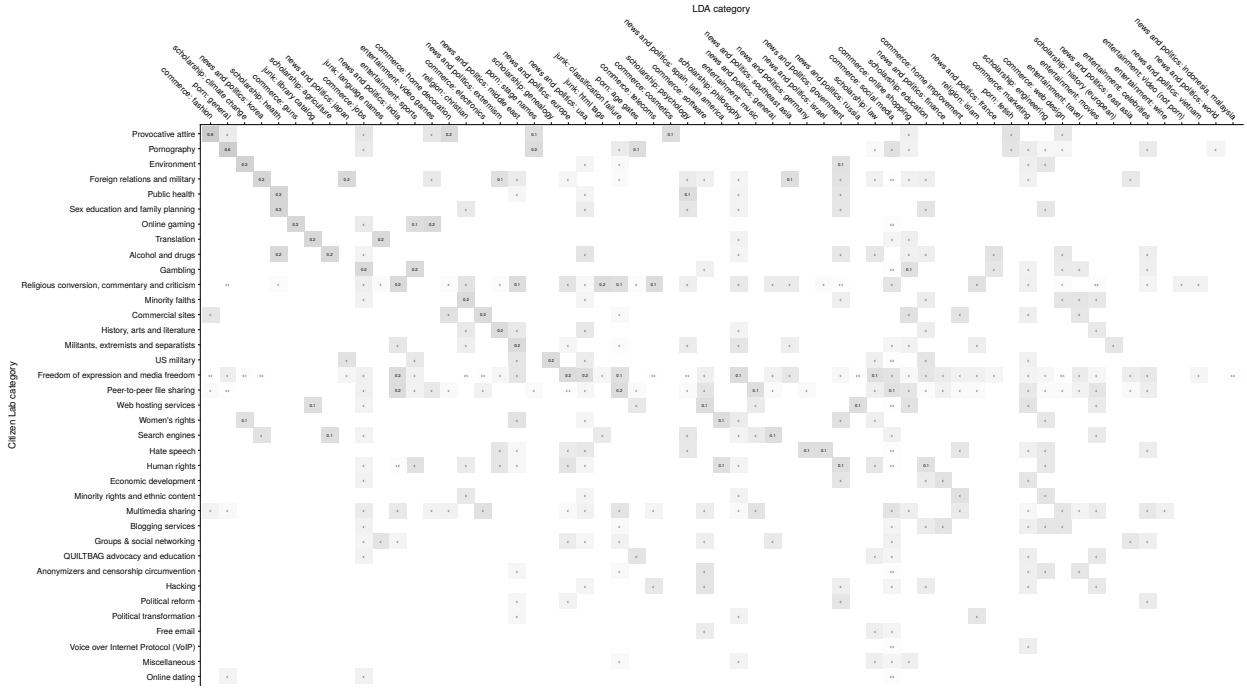



Figure 5.4: Jaccard similarity between Citizen Lab and LDA classification

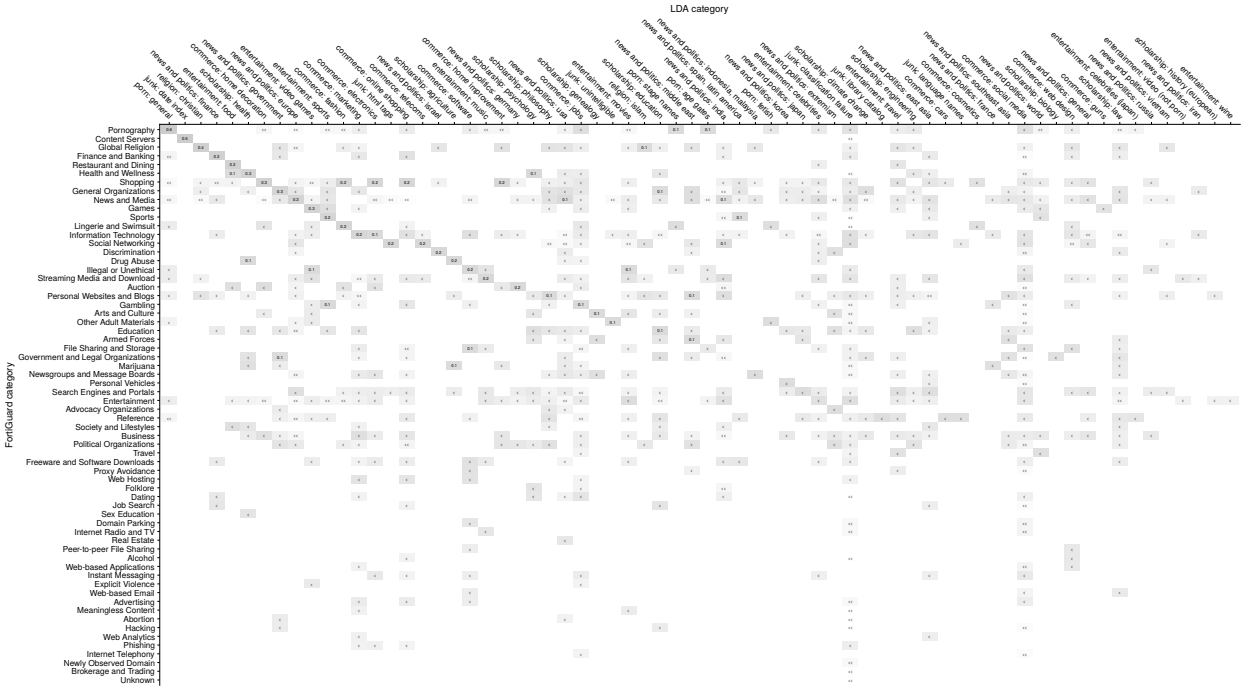



Figure 5.5: Jaccard similarity between FortiGuard and LDA classification

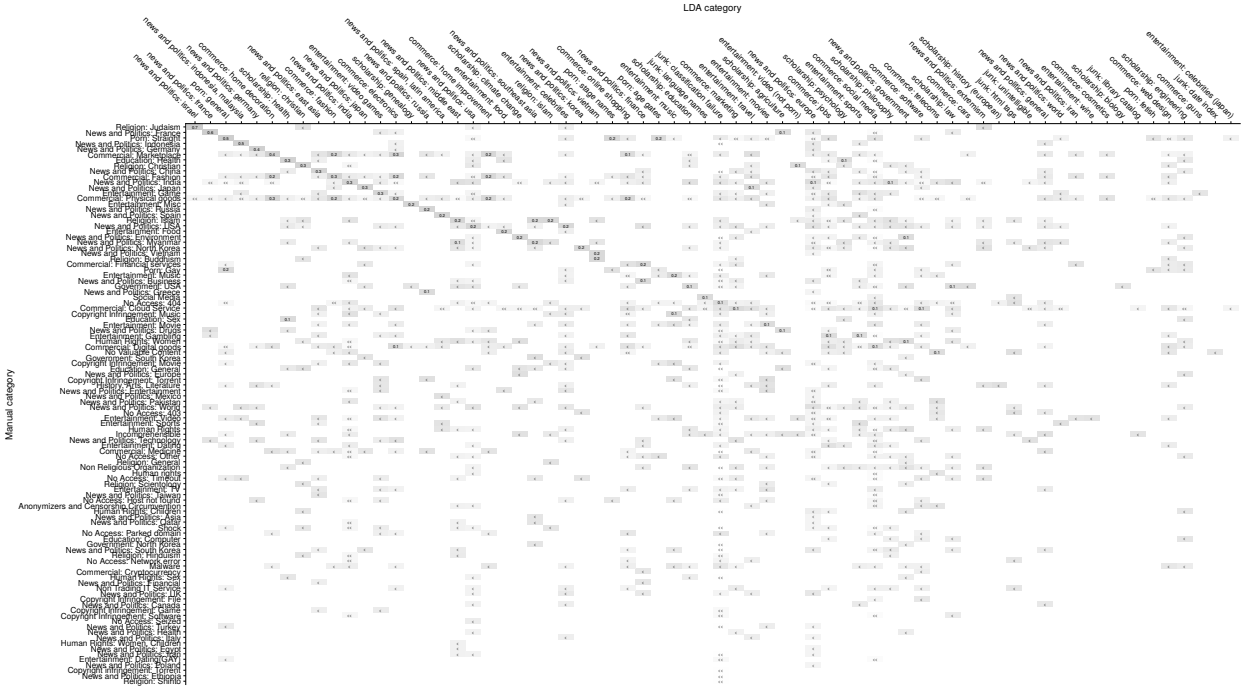



Figure 5.6: Jaccard similarity between manual and LDA classification

case is when sites are short-lived: the Citizen Lab classification used in this study dates to 2012, whereas the FortiGuard classification is continuously updated and the CMU manual classification was done in fall 2018. Many of the sites classified as "multimedia sharing" by Citizen Lab *were* file-sharing sites in 2012, but are now parked domains, spam, or something else entirely. (This effect should be considered whenever a site on a censorship probe list, blacklist, or pinklist is now something that does not seem likely to draw a censor's attention, as discussed at more length in Section 4.7.2.)

Figures 5.4, 5.5, and 5.6 compare the three manual classifications to the LDA-based machine classification. The LDA topics are often narrower than any of the manual topics; we see Citizen Lab's "religious conversion, commentary and criticism" and "freedom of expression and media freedom" topics, and FortiGuard's "Shopping," "General Organizations," and "Entertainment" topics smeared over many LDA topics. This happens substantially less with the CMU manual classification, which was, after all, trained on a very similar data set. It's also noteworthy here that LDA can assign many topics to one document. The vertical stripe seen under the "commerce: social media" LDA category in all three figures, for instance, reflects the appearance of social media keywords, buttons, and transclusions on many sites whose primary topic is something else. Similarly, vertical stripes under "commerce: marketing" and "scholarship: law" reflect the appearance of marketing buzzwords and legal boilerplate on practically everything.

We would not expect perfect agreement between a mechanical classifier and any human classifier, but the appearance of a main diagonal in all three of these figures is a good sign, as is the general consistency between pairs of labels on the main diagonal. Porn maps to porn, finance news to finance news, food to food, Christianity to religion, and so forth. This confirms the meaningfulness and interpretability of the keyword vectors produced by LDA. Sometimes the match is not exact, but there is an obvious connection: CMU manual topic "Religion: Judaism" associates strongly with LDA topic "news and politics: israel," for instance.

## 5.2 Snowball sampling

We extracted all of the outbound hyperlinks from the pages collected by our uncensored-page collector. On average, each page contributed 1.41 unique links to the collection, for a total of 754 128 outbound links. We then prepared a balanced subsample of 137 500 of these links, with, as nearly as possible, an equal number of links from pages belonging to each of the LDA topics described in the previous section. Junk was included, because the junk topics in the revised classification are pages which we could not interpret, not pages with no meaning at all. Links to a different site were preferred over links within a single site, but links within a site were included when necessary to make the sample balanced.

We collected uncensored copies of the sampled pages, using almost but not quite the same procedure used for the original uncensored page collection (Section 4.2), over a period of five months in 2018 (from June 5 through November 2). The most important change from the old procedure was that, in the three years since the original uncensored page collection, PhantomJS ceased to be maintained, meaning that it was no longer adequately up-to-date with the ever-changing "web platform" and had become *less* effective at accurately capturing the contents of webpages than a simple crawler would have been. We therefore replaced it with a newer automated browser, Headless Chromium [82]; this is an officially maintained alternative mode of operation of the Chrome browser.

Refer back to Table 4.3 to see how many of the pages in the snowball sample were successfully retrieved, compared to the initial sample. The only major difference is that the browser crashed on only 1.7 % of the original sample, but 24 % of the snowball sample. (In other words, Headless Chromium is significantly more prone to crashing than PhantomJS was.) This may be random, or it may be that these pages have something in common which is causing crashes. Identifying the problem and making the headless browser more reliable is a high priority for future work.

## 5.3 Page topic and linked topic

Recall from Chapter 4 that a well-designed probe list has both breadth and depth. When considering outbound links from censored pages as new pages to add to the list, therefore, one should balance additions between same-topic and off-topic pages. Adding same-topic pages to a list improves depth, and provides replacements for pages that no longer exist. Adding off-topic pages to a list, on the other hand, improves breadth. The exact proportions to choose are a matter for future research, but we would point out that unsuspected cases of censorship can only be found by expanding the breadth of a probe list.

Figure 5.7 shows, for each LDA topic, the average number of outbound links (in the subsample) per page in that topic, divided into three classes: the link goes to a page with the same topic (green), or to a page with a different topic but still the same larger category (blue), or a different category altogether (orange). For this figure, each page was assigned to the single LDA topic that received the highest score, except that if the highest score was a junk topic and the second-highest score was a non-junk topic, the non-junk topic was used instead.

There is a strong tendency for commerce pages to link to other commerce pages, and porn pages even more so. News is more mixed. Scholarship, religion, and entertainment pages, on the other hand, link out to other categories more than they link to their own topic or even their own category. Video pages seem almost never to link to other video pages—this may be a consequence of PhantomJS not supporting video; YouTube and other high-end video hosting sites may have

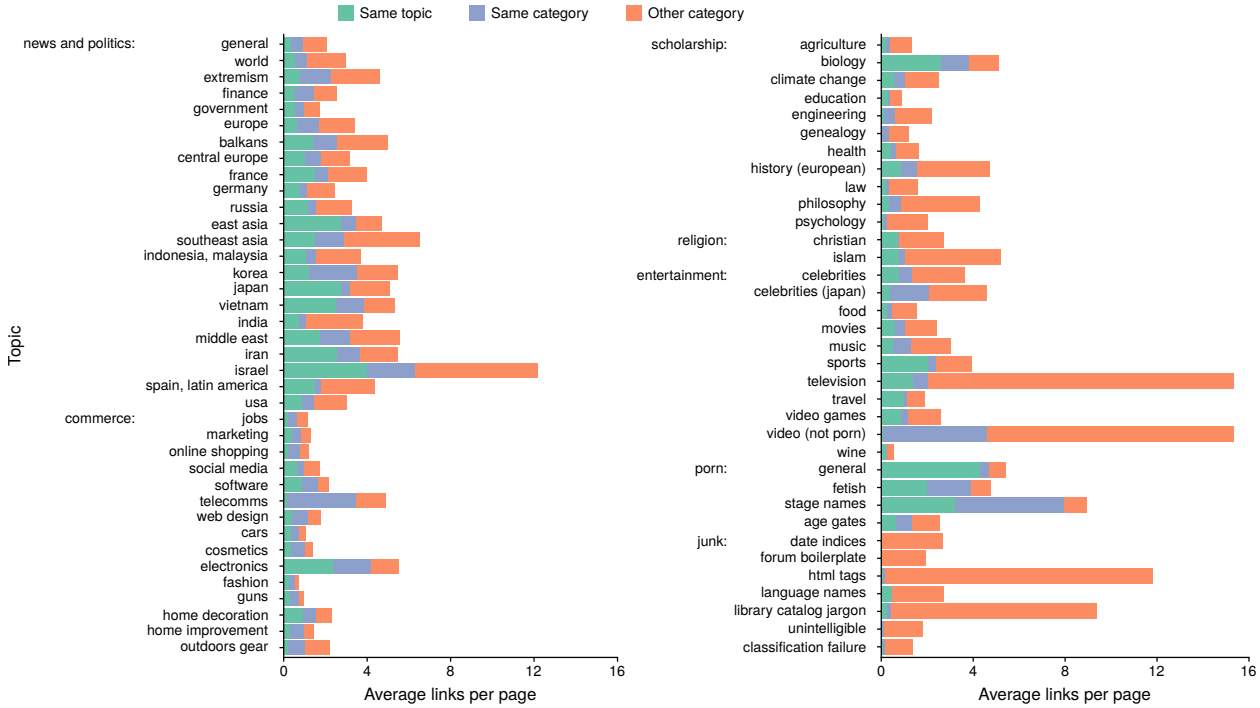



Figure 5.7: Proportion of outbound links to pages in the same topic

showed it a “browser not supported” message rather than video recommendations.

Figure 5.8 uses Jaccard similarity, again, to compare the full topic vectors of the pages in each topic to the topic vectors of their outbound links. The off-diagonal entries appear fairly uniform, which means the topic of a page usually does not predict the topics of its off-topic links.

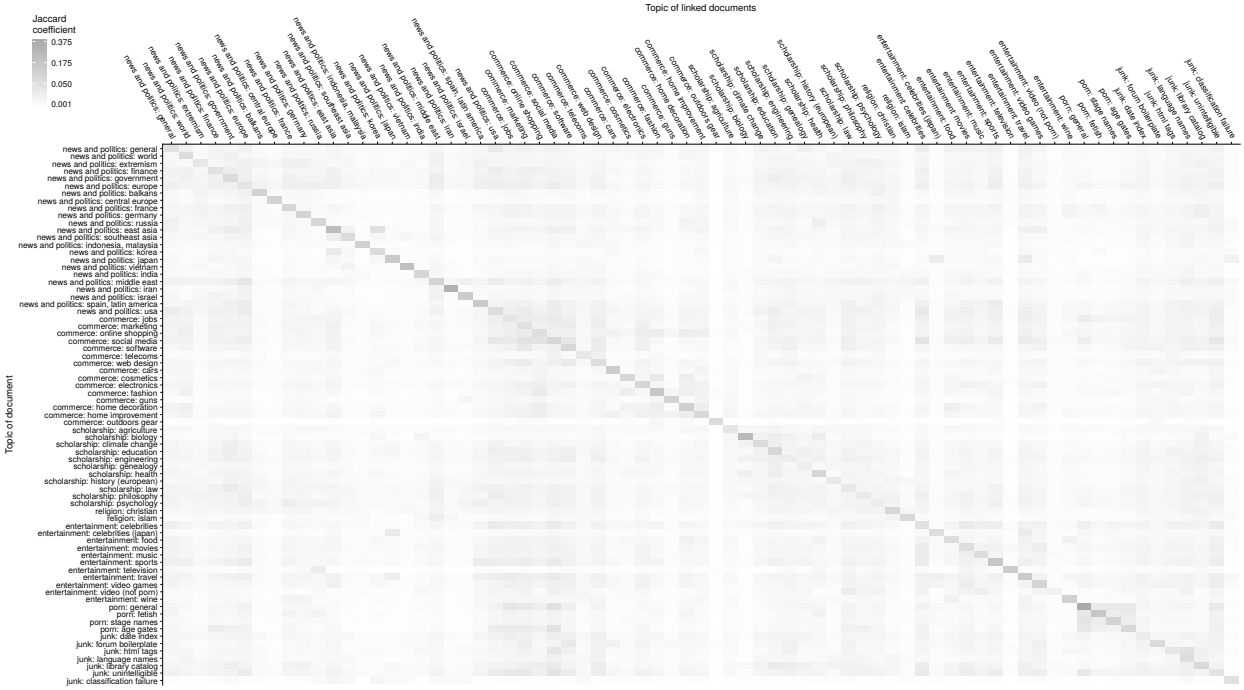



Figure 5.8: Jaccard similarity of documents' topics to their outbound links' topics

# 6. Conclusion

In this dissertation I have presented four improvements upon the state of the art in worldwide, continuous censorship monitoring.

I have shown that the physical locations of censorship vantage points can be confirmed with active geolocation, at least when those vantage points are in data centers, which enables monitors to make use of VPN services with confidence. In the process, I have identified a systemic tendency for commercial VPN services to inflate the number of physical locations where they host servers.

I have demonstrated that by combining information from multiple levels of the network stack, such as both the TCP- and HTTP-level semantics of a packet that may have been injected, one can devise heuristics that detect censorship with fewer false positives.

I have developed a method for identifying unsuspected fingerprints of censorship, using unsupervised clustering, scored by the URL-to-country ratio. This method allowed me to discover 33 previously unknown block page signatures.

Finally, I have assessed the quality of an existing censorship probe list by retrieving the uncensored contents of its pages, mechanically identifying their topics, and using records from the Internet Archive to estimate the lifetime of pages that no longer exist.

My work sheds light on major gaps in our understanding of Internet censorship as a worldwide phenomenon, but also points toward filling those gaps in the future.

# Bibliography


[1] N. Aase, J. R. Crandall, Á. Díaz, J. Knockel, J. O. Molinero, J. Saia, D. Wallach, and T. Zhu, "Whiskey, Weed, and Wukan on the World Wide Web: On Measuring Censors' Resources and Motivations," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2012. URL: https://www.usenix.org/system/files/conference/foci12/foci12-final17.pdf.

[2] A. M. Abdou, A. Matrawy, and P. C. Van Oorschot, "CPV: Delay-based Location Verification for the Internet," *Transactions on Dependable and Secure Computing*, vol. 14, no. 2, pp. 130–144, 2015. DOI: 10.1109/TDSC.2015.2451614. URL: http://www.scs.carleton.ca/~paulv/papers/CPV-TDSC-authorcopy.pdf.

[3] ——, "Accurate Manipulation of Delay-based Internet Geolocation," in *Asia Conference on Computer and Communications Security*, New York: ACM, 2017, pp. 887–898. DOI: 10.1145/3052973.3052993. URL: http://www.scs.carleton.ca/~paulv/papers/asiaccs-2017.pdf.

[4] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *Computer and Communications Security*, Scottsdale, Arizona: ACM, 2014, pp. 674–689. DOI: 10.1145/2660267.2660347. URL: https://pdfs.semanticscholar.org/01db/c5466cce6abd567cc5b34a481f5c438fb15a.pdf.

[5] G. Aceto, A. Montieri, and A. Pescapè, "Internet Censorship in Italy: A First Look at 3G/4G Networks," in *Cryptology and Network Security*, S. Foresti and G. Persiano, Eds., ser. Lecture Notes in Computer Science, vol. 10052, Berlin, Heidelberg: Springer, 2016, pp. 737–742. DOI: 10.1007/978-3-319-48965-0_53.

[6] S. T. Ahmed, C. Sparkman, H.-T. Lee, and D. Loguinov, "Around the Web in Six Weeks: Documenting a Large-Scale Crawl," in *INFOCOM*, Piscataway, NJ: IEEE, 2015, pp. 1598–1606. URL: http://irl.cse.tamu.edu/people/tanzir/papers/infocom2015a.pdf.

[7] B. Altemeyer, *Right-Wing Authoritarianism*. Winnipeg: University of Manitoba Press, 1981, ISBN: 0-88755-124-6.

[8] C. Anderson, *Dimming the Internet: Detecting Throttling as a Mechanism of Censorship in Iran*, 2013. arXiv: 1306.4361 [cs.NI].

[9] C. Anderson, P. Winter, and Roya, "Global Network Interference Detection over the RIPE Atlas Network," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2014. URL: https://www.usenix.org/system/files/conference/foci14/foci14-anderson.pdf.

[10] D. Anderson, "Splinternet Behind the Great Firewall of China," *ACM Queue*, vol. 10, no. 11, 2012. URL: http://queue.acm.org/detail.cfm?id=2405036.

[11] L. T. Anh and K. Yamamoto, *DongDu — Vietnamese Word Segmenter*, Software library, 2012. URL: http://eng.jnlp.org/dongdu.

[12] Anonymous, "The Collateral Damage of Internet Censorship by DNS Injection," *SIGCOMM Computer Communications Review*, vol. 42, no. 3, pp. 21–27, 2012. DOI: 10.1145/2317307.2317311. URL: http://www.sigcomm.org/node/3275.

[13] ——, "Towards a Comprehensive Picture of the Great Firewall's DNS Censorship," in *Free and Open Communications on the Internet*, San Diego, CA: USENIX, 2014. URL: https://www.usenix.org/conference/foci14/workshop-program/presentation/anonymous.

[14] M. J. Arif, S. Karunasekera, and S. Kulkarni, "GeoWeight: Internet Host Geolocation Based on a Probability Model for Latency Measurements," in *Australasian Computer Science Conference*, vol. 102, Sydney: ACS, 2010, pp. 89–98. URL: http://crpit.com/confpapers/crpitv102arif.pdf.

[15] S. Aryan, H. Aryan, and J. A. Halderman, "Internet Censorship in Iran: A First Look," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2013. URL: https://www.usenix.org/system/files/conference/foci13/foci13-aryan.pdf.

[16] M. Bailey and C. Labovitz, "Censorship and Co-option of the Internet Infrastructure," University of Michigan, Tech. Rep. CSE-TR-572-11, 2011. URL: http://tangle.eecs.umich.edu/publications/CSE-TR-572-11.pdf.

[17] D. E. Bambauer, "Censorship v3.1," *IEEE Internet Computing*, vol. 17, no. 3, pp. 26–33, 2013. DOI: 10.1109/MIC.2013.23. URL: https://ssrn.com/abstract=2144004.

[18] Z. Bar-Yossef and S. Rajagopalan, "Template Detection via Data Mining and Its Applications," in *World Wide Web*, New York: ACM, 2002, pp. 580–591. DOI: 10.1145/511446.511522.

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

[20] Berkman Center for Internet and Society, *Herdict: help spot web blockages*, Web site, 2009–. URL: https://www.herdict.org/.

[21] BI Science Ltd., *Geosurf: Residential and data center proxy network*, Web site, 2009–. URL: https://www.geosurf.com.

[22] C. Biemann, F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski, and T. Zesch, "Scalable Construction of High-Quality Web Corpora," *Journal for Language Technology and Computational Linguistics*, vol. 28, no. 2, pp. 23–59, 2013. URL: http://www.jlcl.org/2013_Heft2/2Biemann.pdf.

[23] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media Inc., 2009. URL: http://www.nltk.org/.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1002, 2003. URL: http://cs.berkeley.edu/~jordan/papers/blei03a.ps.

[25] J. L. Boyd-Graber and D. M. Blei, "Multilingual Topic Models for Unaligned Text," in *Uncertainty in Artificial Intelligence*, J. A. Bilmes and A. Y. Ng, Eds., Arlington, VA: AUAI Press, 2009, pp. 75–82. URL: http://uai.sis.pitt.edu/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1644&proceeding_id=25.

[26] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[27] Y. Breindl and J. Wright, "Internet Filtering Trends in Western Liberal Democracies," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2012. URL: https://www.usenix.org/system/files/conference/foci12/breindl2012foci.pdf.

[28] M. A. Brown, "Pakistan hijacks YouTube," *Dyn.com Vantage Point*, Feb. 2008. URL: https://dyn.com/blog/pakistan-hijacks-youtube-1/.

[29] S. Burnett and N. Feamster, "Making Sense of Internet Censorship: A New Frontier for Internet Measurement," *SIGCOMM Computer Communications Review*, vol. 43, no. 3, pp. 84–89, 2013. DOI: 10.1145/2500098.2500111. URL: http://www.sigcomm.org/sites/default/files/ccr/papers/2013/July/2500098-2500111.pdf.

[30] ——, "Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests," in *SIGCOMM*, New York: ACM, 2015, pp. 653–667. DOI: 10.1145/2785956.2787485. arXiv: 1410.1211 [cs.CY]. URL: http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p653.pdf.

[31] C. Castelluccia, M. A. Kaafar, P. Manils, and D. Perito, "Geolocalization of Proxied Services and its Application to Fast-Flux Hidden Servers," in *Internet Measurement Conference*, New York: ACM, 2009, pp. 184–189. DOI: 10.1145/1644893.1644915. URL: https://planete.inrialpes.fr/~ccastel/PAPERS/imc09.pdf.

[32] Center for Applied Internet Data Analysis, *Archipelago (Ark) Measurement Infrastructure*, 2006. URL: http://www.caida.org/projects/ark/.

[33] ——, *The CAIDA UCSD AS Classification Dataset*, Data set, 2017. URL: http://www.caida.org/data/as-classification/.

[34] A. Chaabane, T. Chen, M. Cunche, E. De Cristofaro, A. Friedman, and M. A. Kaafar, "Censorship in the Wild: Analyzing Internet Filtering in Syria," in *Internet Measurement Conference*, Vancouver, BC: ACM, 2014, pp. 285–298. DOI: 10.1145/2663716.2663720. URL: http://conferences2.sigcomm.org/imc/2014/papers/p285.pdf.

[35] S. Chakravarthy and S. C. H. Hara, "Automating change detection and notification of Web pages," in *Database and Expert Systems Applications*, Piscataway, NJ: IEEE, 2006, pp. 465–469. DOI: 10.1109/DEXA.2006.34.

[36] B. Chandrasekaran, M. Bai, M. Schoenfield, A. Berger, N. Caruso, G. Economou, S. Gilliss, B. Maggs, K. Moses, D. Duff, K.-C. Ng, E. G. Sirer, R. Weber, and B. Wong, "Alidade: IP Geolocation without Active Probing," Department of Computer Science, Duke University, Tech. Rep. CS-TR-2015.001, 2015. URL: https://pdfs.semanticscholar.org/258c/729741648b5762aa3a9b8e8b111460d15f73.pdf.

[37] P.-C. Chang, M. Galley, and C. D. Manning, "Optimizing Chinese word segmentation for machine translation performance," in *Statistical Machine Translation*, Stroudsburg, PA: ACL, 2008, pp. 224–232. URL: http://statmt.org/wmt08/pdf/WMT36.pdf.

[38] J. Chen, F. Liu, X. Luo, F. Zhao, and G. Zhu, "A landmark calibration-based IP geolocation approach," *EURASIP Journal on Information Security*, vol. 2016, 2016. DOI: 10.1186/s13635-015-0029-5.

[39] Citizen Lab *et al.*, *URL testing lists intended for discovering website censorship*, Git repository, 2014–. URL: https://github.com/citizenlab/test-lists.

[40] ——, *Blockpages as collected by various sources*, Git repository, 2015–. URL: https://github.com/citizenlab/blockpages.

[41] J. Clark, R. Faris, and R. H. Jones, "Analyzing Accessibility of Wikipedia Projects Around the World," Berkman Klein Center for Internet & Society, Tech. Rep., 2017. URL: https://cyber.law.harvard.edu/publications/2017/04/WikipediaCensorship.

[42] R. Clayton, S. J. Murdoch, and R. N. M. Watson, "Ignoring the Great Firewall of China," in *Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science, vol. 4258, Berlin, Heidelberg: Springer, 2006, pp. 20–35. DOI: 10.1007/11957454_2.

[43] J. R. Crandall, D. Zinn, M. Byrd, E. Barr, and R. East, "ConceptDoppler: A Weather Tracker for Internet Censorship," in *Computer and Communications Security*, New York: ACM, 2007, pp. 352–365. DOI: 10.1145/1315245.1315290. URL: http://www.cs.unm.edu/~crandall/concept_doppler_ccs07.pdf.

[44] F. Dabek, R. Cox, F. Kaashoek, and R. Morris, "Vivaldi: A Decentralized Network Coordinate System," in *SIGCOMM*, New York: ACM, 2004, pp. 15–26. DOI: 10.1145/1015467.1015471. URL: https://pdos.csail.mit.edu/papers/vivaldi:sigcomm/paper.pdf.

[45] A. Dainotti, C. Squarcella, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, and A. Pescape, "Analysis of Country-Wide Internet Outages Caused by Censorship," *IEEE/ACM Transactions on Networking*, vol. 22, pp. 1964–1977, 6 2013, ISSN: 1063-6692. DOI: 10.1109/TNET.2013.2291244. URL: http://www.caida.org/publications/papers/2014/outages_censorship/.

[46] J. Dalek, R. Deibert, S. McKune, P. Gill, A. Senft, and N. Noor, "Information Controls during Military Operations," CitizenLab, Tech. Rep., 2015. URL: https://citizenlab.ca/2015/10/information-controls-military-operations-yemen/.

[47] J. Dalek, B. Haselton, H. Noman, A. Senft, M. Crete-Nishihata, P. Gill, and R. J. Deibert, "A Method for Identifying and Confirming the Use of URL Filtering Products for Censorship," in *Internet Measurement Conference*, New York: ACM, 2013, pp. 23–30. DOI: 10.1145/2504730.2504763. URL: http://www3.cs.stonybrook.edu/~phillipa/papers/imc112s-dalek.pdf.

[48] A. Darer, O. Farnan, and J. Wright, "FilteredWeb: A Framework for the Automated Search-Based Discovery of Blocked URLs," in *Network Traffic Measurement and Analysis*, Dublin: IEEE, 2017. arXiv: 1704.07185 [cs.CY].

[49] ——, "Automated Discovery of Internet Censorship by Web Crawling," in *Web Science*, Amsterdam, Netherlands: ACM, 2018, pp. 195–204. DOI: 10.1145/3201064.3201091. arXiv: 1804.03056 [cs.CY].

[50] R. Deibert, J. Palfrey, R. Rohozinski, and J. Zittrain, Eds., *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*, ser. ONI Access 2. Boston: MIT Press, 2010, ISBN: 0-262-51435-4. URL: http://access.opennet.net/controlled/.

[51] R. Deibert, J. Palfrey, R. Rohozinski, and J. Zittrain, Eds., *Access Denied: The Practice and Policy of Global Internet Filtering*, ser. ONI Access 1. Boston: MIT Press, 2008, ISBN: 0-262-54196-3. URL: http://access.opennet.net/denied/.

[52] R. Deibert, J. Palfrey, R. Rohozinski, and J. Zittrain, Eds., *Access Contested: Security, Identity, and Resistance in Asian Cyberspace*, ser. ONI Access 3. Boston: MIT Press, 2011, ISBN: 0-262-01678-8. URL: http://access.opennet.net/controlled/.

[53] S. Ding, X. Luo, M. Yin, Y. Liu, and F. Liu, "An IP Geolocation Method Based on Rich-Connected Subnetworks," in *International Conference on Advanced Communication Technology*, Piscataway, NJ: IEEE, 2015, pp. 176–181. DOI: 10.1109/ICACT.2015.7224779.

[54] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The Second-Generation Onion Router," in *USENIX Security Symposium*, Berkeley, CA: USENIX, 2004, pp. 303–320. URL: https://www.usenix.org/legacy/events/sec04/tech/full_papers/dingledine/dingledine.pdf.

[55] M. Dischinger, M. Marcon, S. Guha, K. P. Gummadi, R. Mahajan, and S. Saroiu, "Glasnost: Enabling End Users to Detect Traffic Differentiation," in *Networked Systems Design and Implementation*, Berkeley, CA: USENIX, 2010. URL: https://www.usenix.org/legacy/events/nsdi10/tech/full_papers/dischinger.pdf.

[56] Z. Dong, R. D. W. Perera, R. Chandramouli, and K. P. Subbalakshmi, "Network measurement based modeling and optimization for IP geolocation," *Computer Networks*, vol. 56, no. 1, pp. 85–98, 2012. DOI: 10.1016/j.comnet.2011.08.011. URL: http://www.academia.edu/download/31092865/ip-geo-1.pdf.

[57] M. Dornseif, "Government mandated blocking of foreign Web content," in *DFN-Arbeitstagung über Kommunikationsnetze*, Berlin: Deutsche Forschungsnetz, 2003, pp. 617–648. arXiv: cs/0404005 [cs.CY].

[58] B. Dowling, D. Stebila, and G. Zaverucha, "Authenticated Network Time Synchronization," in *USENIX Security Symposium*, Berkeley, CA: USENIX, 2016, pp. 823–840. URL: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/dowling.

[59] T. Elahi and I. Goldberg, "CORDON: A Taxonomy of Internet Censorship Resistance Strategies," Centre for Applied Cryptographic Research, University of Waterloo, Tech. Rep. 2012-33, 2012. URL: http://cacr.uwaterloo.ca/techreports/2012/cacr2012-33.pdf.

[60] M. Emem, "Monero Cryptomining Attack Affects Over 200,000 ISP-Grade Routers Globally," *CCN.com*, Aug. 2018. URL: https://web.archive.org/web/20180819120151/https://www.ccn.com/monero-cryptomining-attack-affects-over-200000-isp-grade-routers-globally/.

[61] R. Ensafi, D. Fifield, P. Winter, N. Feamster, N. Weaver, and V. Paxson, "Examining How the Great Firewall Discovers Hidden Circumvention Servers," in *Internet Measurement Conference*, New York: ACM, 2015, pp. 445–458. DOI: 10.1145/2815675.2815690. URL: https://censorbib.nymity.ch/pdf/Ensafi2015b.pdf.

[62] R. Ensafi, J. Knockel, G. Alexander, and J. R. Crandall, "Detecting intentional packet drops on the Internet via TCP/IP side channels," in *Passive and Active Measurement*, ser. Lecture Notes in Computer Science, vol. 8362, Berlin, Heidelberg: Springer, 2014, pp. 109–118. DOI: 10.1007/978-3-319-04918-2_11.

[63] R. Ensafi, P. Winter, A. Mueen, and J. R. Crandall, *Large-scale Spatiotemporal Characterization of Inconsistencies in the World's Largest Firewall*, 2014. arXiv: 1410.0735 [cs.NI].

[64] B. Eriksson, P. Barford, B. Maggs, and R. Nowak, "Posit: A Lightweight Approach for IP Geolocation," *SIGMETRICS Performance Evaluation Review*, vol. 40, no. 2, pp. 2–11, 2012, ISSN: 0163-5999. DOI: 10.1145/2381056.2381058. URL: http://www.brianeriksson.com/static/pdf/2012/erikssonPosit12.pdf.

[65] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A Learning-Based Approach for IP Geolocation," in *Passive and Active Measurement*, A. Krishnamurthy and B. Plattner, Eds., Berlin, Heidelberg: Springer, 2010, pp. 171–180. DOI: 10.1007/978-3-642-12334-4_18. URL: http://pam2010.ethz.ch/papers/full-length/18.pdf.

[66] B. Eriksson and M. Crovella, "Understanding Geolocation Accuracy using Network Geometry," in *INFOCOM*, Piscataway, NJ: IEEE, 2013, pp. 75–79. DOI: 10.1109/INFCOM.2013.6566738. URL: http://www.brianeriksson.com/static/pdf/2013/erikssonGeoBounds13.pdf.

[67] S. Evert, "A lightweight and efficient tool for cleaning Web pages," in *International Conference on Language Resources and Evaluation*, Paris: European Language Resources Association, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/summaries/885.html.

[68] O. Farnan, A. Darer, and J. Wright, "Poisoning the Well: Exploring the Great Firewall's Poisoned DNS Responses," in *Workshop on Privacy in the Electronic Society*, New York: ACM, 2016, pp. 95–98. DOI: 10.1145/2994620.2994636. URL: https://censorbib.nymity.ch/pdf/Farnan2016a.pdf.

[69] R. Fielding and J. Reschke, Eds., *Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing*, RFC 7230, Jun. 2014. URL: https://tools.ietf.org/html/rfc7230.

[70] A. Filastò and J. Appelbaum, "OONI: Open Observatory of Network Interference," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2012. URL: https://www.usenix.org/system/files/conference/foci12/foci12-final12.pdf.

[71] FortiNet, *FortiGuard Labs Web Filter*, Online service, 2017–. URL: https://fortiguard.com/webfilter/categories.

[72] K. W. Fu, C. H. Chan, and M. Chau, "Assessing Censorship on Microblogs in China: Discriminatory Keyword Analysis and the Real-Name Registration Policy," *IEEE Internet Computing*, vol. 17, no. 3, pp. 42–50, 2013. DOI: 10.1109/MIC.2013.28.

[73] S. Gallagher, "Big Brother on a budget: How Internet surveillance got so cheap," *Ars Technica*, 2012. URL: http://arstechnica.com/information-technology/2012/09/big-brother-meets-big-data-the-next-wave-in-net-surveillance-tech/.

[74] M. Gargiulo, "List of VPN Locations by Provider," in *VPN Reviews & Free Comparison Charts*, Cupertino, CA: VPN.com, Feb. 2018. URL: https://www.vpn.com/.

[75] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos, "A Look at Router Geolocation in Public and Commercial Databases," in *Internet Measurement Conference*, New York: ACM, 2017, pp. 463–469. DOI: 10.1145/3131365.3131380.

[76] P. Gill, M. Crete-Nishihata, J. Dalek, S. Goldberg, A. Senft, and G. Wiseman, "Characterizing Web Censorship Worldwide: Another Look at the OpenNet Initiative Data," *ACM Transactions on the Web*, vol. 9, no. 1, 2015. DOI: 10.1145/2700339. URL: http://spin2013.cs.sunysb.edu/~phillipa/papers/TWeb.pdf.

[77] P. Gill, Y. Ganjali, B. Wong, and D. Lie, "Dude, where's that IP?: Circumventing measurement-based IP geolocation," in *USENIX Security*, Berkeley, CA: USENIX, 2010. URL: https://www.usenix.org/legacy/events/sec10/tech/full_papers/Gill.pdf.

[78] Google Inc., *Translate API*, Online service, 2014–. URL: https://cloud.google.com/translate/.

[79] D. Gosain, A. Agarwal, S. Shekhawat, H. B. Acharya, and S. Chakravarty, "Mending Wall: On the Implementation of Censorship in India," in *Security and Privacy in Communication Networks*, Cham: Springer, 2018, pp. 418–437. DOI: 10.1007/978-3-319-78813-5_21. arXiv: 1806.06518 [cs.CR].

[80] J. C. Gratz, M. Ammori, and L. Ammori, "Brief of *amici curiae* Automattic Inc.; Google Inc.; Twitter Inc.; and Tumblr, Inc," in *Lenz v. Universal Music*, ser. 801 F.3d 1126, 2015. URL: https://www.eff.org/files/2013/12/17/048_automattic_google_twitter_tumblr_amicus_brief_12.13.13.pdf.

[81] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based Geolocation of Internet Hosts," in *Internet Measurement Conference*, New York: ACM, 2004, pp. 288–293. DOI: 10.1145/1028788.1028828.

[82] *Headless Chromium*, Open source software, 2016–. URL: https://chromium.googlesource.com/chromium/src/+/lkgr/headless/README.md.

[83] S. Hellmeier, "The Dictator's Digital Toolkit: Explaining Variation in Internet Filtering in Authoritarian Regimes," *Politics & Policy*, vol. 44, no. 6, pp. 1158–1191, 2016. DOI: 10.1111/polp.12189.

[84] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?" *Behavioral and Brain Sciences*, vol. 33, no. 2-3, pp. 61–83, 2010, ISSN: 1469-1825. DOI: 10.1017/S0140525X0999152X. URL: http://www2.psych.ubc.ca/~anlab/Manuscripts/Weird_People_BBS_Henrichetal_FullPackage.pdf.

[85] J. den Hertog and M. Candela, *OpenIPMap: A Collaborative Approach to Mapping Internet Infrastructure*, 2018. URL: https://labs.ripe.net/Members/jasper_den_hertog/openipmap-a-collaborative-approach-to-mapping-internet-infrastructure.

[86] A. Hidayat *et al.*, *PhantomJS*, Software application, 2010–. URL: http://phantomjs.org/.

[87] K. Hill, "How an internet mapping glitch turned a random Kansas farm into a digital hell," *FUSION*, 2016. URL: http://fusion.net/story/287592/internet-mapping-glitch-kansas-farm/.

[88] A. Hintz and S. Milan, "Through a Glass, Darkly: Everyday Acts of Authoritarianism in the Liberal West," *International Journal of Communication*, vol. 12, pp. 3939–3959, 2018. URL: https://ijoc.org/index.php/ijoc/article/view/8537.

[89] N. P. Hoang, P. Kintis, M. Antonakakis, and M. Polychronakis, "An Empirical Study of the I2P Anonymity Network and Its Censorship Resistance," in *Internet Measurement Conference*, Boston: ACM, 2018, pp. 379–392. DOI: 10.1145/3278532.3278565. arXiv: 1809.09086 [cs.NI].

[90] Hola!VPN: Access any website, Web site, 2013–. URL: https://hola.org/.

[91] T. Holterbach, C. Pelsser, R. Bush, and L. Vanbever, "Quantifying Interference Between Measurements on the RIPE Atlas Platform," in *Internet Measurement Conference*, New York: ACM, 2015, pp. 437–443. DOI: 10.1145/2815675.2815710.

[92] A. Hounsel, P. Mittal, and N. Feamster, "Automatically Generating a Large, Culture-Specific Blocklist for China," in *Free and Open Communications on the Internet*, Baltimore, MD: USENIX, 2018. arXiv: 1806.03255 [cs.CY].

[93] Z. Hu, J. Heidemann, and Y. Pradkin, "Towards Geolocation of Millions of IP Addresses," in *Internet Measurement Conference*, New York: ACM, 2012, pp. 123–130. DOI: 10.1145/2398776.2398790. URL: http://www.isi.edu/~johnh/PAPERS/Hu12a.pdf.

[94] Internet Archive, *Wayback Machine*, Online service, 1996–. URL: https://archive.org/web/web.php.

[95] C. Jackson, A. Bortz, D. Boneh, and J. C. Mitchell, "Protecting Browser State from Web Privacy Attacks," in *World Wide Web*, New York: ACM, 2006, pp. 737–744. URL: http://www.stanford.edu/people/jcm/papers/sameorigin.pdf.

[96] B. Jones, T.-W. Lee, N. Feamster, and P. Gill, "Automated Detection and Fingerprinting of Censorship Block Pages," in *Internet Measurement Conference*, New York: ACM, 2014, pp. 299–304. DOI: 10.1145/2663716.2663722. URL: http://conferences2.sigcomm.org/imc/2014/papers/p299.pdf.

[97] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, pp. 457–481, 1958. DOI: 10.2307/2281868.

[98] T. Karoonboonyanan, P. Kiatisevi, V. Ampornaramveth, P. Veerathanabutr, and C. Silpa-Anan, *LibThai*, Software library, 2001–2013. URL: http://linux.thai.net/projects/libthai.

[99] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP Geolocation Using Delay and Topology Measurements," in *Internet Measurement Conference*, New York: ACM, 2006, pp. 71–84. DOI: 10.1145/1177080.1177090. URL: http://conferences.sigcomm.org/imc/2006/papers/p7-bassett.pdf.

[100] S. Kelly, M. Earp, L. Reed, A. Shahbaz, and M. Truong, *Freedom on the Net 2014*, 2014. URL: https://freedomhouse.org/report/freedom-net/freedom-net-2014.

[101] S. Kelly, M. Truong, A. Shahbaz, M. Earp, and J. White, *Freedom on the Net 2017*, 2017. URL: https://freedomhouse.org/report/freedom-net/freedom-net-2017.

[102] ——, *Freedom on the Net 2017: Turkey*, 2017. URL: https://freedomhouse.org/report/freedom-net/freedom-net/2017/turkey.

[103] S. Kenin, "Mass MikroTik Router Infection — First we cryptojack Brazil, then we take the World?," *SpiderLabs Blog*, Aug. 2018. URL: https://web.archive.org/web/20181101052031/https://www.trustwave.com/Resources/SpiderLabs-Blog/Mass-MikroTik-Router-Infection-%5C%E2%5C%80%5C%93-First-we-cryptojack-Brazil,-then-we-take-the-World-/.

[104] M. T. Khan, J. DeBlasio, G. M. Voelker, A. C. Snoeren, C. Kanich, and N. Vallina-Rodriguez, "An Empirical Analysis of the Commercial VPN Ecosystem," in *Internet Measurement Conference*, Boston: ACM, 2018, pp. 443–456. DOI: 10.1145/3278532.3278570.

[105] R. A. A. Khan, A. Naveed, and R. L. Cottrell, "Adaptive Geolocation of Internet Hosts," SLAC National Accelerator Laboratory, Tech. Rep. SLAC-PUB-16463, 2016. URL: https://www.slac.stanford.edu/pubs/slacpubs/16250/slac-pub-16463.pdf.

[106] S. Khattak, M. Javed, S. A. Khayam, Z. A. Uzmi, and V. Paxson, "A Look at the Consequences of Internet Censorship Through an ISP Lens," in *Internet Measurement Conference*, New York: ACM, 2014, pp. 271–284. DOI: 10.1145/2663716.2663750. URL: http://conferences2.sigcomm.org/imc/2014/papers/p271.pdf.

[107] G. King, J. Pan, and M. E. Roberts, "How censorship in China allows government criticism but silences collective expression," *American Political Science Review*, vol. 107, no. 2, pp. 326–343, 2013. DOI: 10.1017/s0003055413000014. URL: http://nrs.harvard.edu/urn-3:HUL.InstRepos:11878767.

[108] P. Kingsley, "Turkey Purges 4,000 More Officials, and Blocks Wikipedia," *The New York Times*, Apr. 2017. URL: https://www.nytimes.com/2017/04/30/world/europe/turkey-purge-wikipedia-tv-dating-shows.html.

[109] J. Knockel, "Measuring Decentralization of Chinese Censorship in Three Industry Segments," PhD thesis, The University of New Mexico, 2018. URL: http://digitalrepository.unm.edu/cs_etds/90.

[110] J. Knockel, J. R. Crandall, and J. Saia, "Three Researchers, Five Conjectures: An Empirical Analysis of TOM-Skype Censorship and Surveillance.," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2011. URL: https://www.usenix.org/legacy/events/foci11/tech/final_files/Knockel.pdf.

[111] J. Knockel, M. Crete-Nishihata, and L. Ruan, "The effect of information controls on developers in China: An analysis of censorship in Chinese open source projects," in *Natural Language Processing for Internet Freedom*, Santa Fe, NM: ACL, 2018, pp. 1–11. URL: http://www.aclweb.org/anthology/W18-4201.

[112] D. Komosny, M. Simek, and G. Kathiravelu, "Can Vivaldi Help in IP Geolocation?" *Przegląd Elektrotechniczny*, vol. 2013, no. 5, pp. 100–106, 2013. URL: http://www.pe.org.pl/articles/2013/5/20.pdf.

[113] D. Komosny, M. Voznak, G. Kathiravelu, and H. Sathu, "Estimation of Internet Node Location by Latency Measurements—The Underestimation Problem," *Information Technology and Control*, vol. 44, no. 3, pp. 279–286, 2015. URL: http://hdl.handle.net/10084/110524.

[114] R. K. Konoth, E. Vineti, V. Moonsamy, M. Lindorfer, C. Kruegel, H. Bos, and G. Vigna, "MineSweeper: An In-depth Look into Drive-by Cryptocurrency Mining and Its Defense," in *Computer and Communications Security*, Toronto: ACM, 2018, pp. 1714–1730. DOI: 10.1145/3243734.3243858. URL: https://pdfs.semanticscholar.org/a746/f6bc51be73aa14981a9e9b80b9b9833ef715.pdf.

[115] R. Krishnan, H. V. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao, "Moving Beyond End-to-End Path Information to Optimize CDN Performance," in *Internet Measurement Conference*, New York: ACM, 2009, pp. 190–201. DOI: 10.1145/1644893.1644917. URL: http://www.academia.edu/download/40030850/imc191.pdf.

[116] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," in *Empirical Methods in Natural Language Processing*, Stroudsburg, PA: ACL, 2004, pp. 230–237. URL: http://chasen.org/~taku/publications/emnlp2004-2.pdf.

[117] S. Laki, P. Mátray, P. Hága, T. Sebők, I. Csabai, and G. Vattay, "Spotter: A Model Based Active Geolocation Service," in *INFOCOM*, Piscataway, NJ: IEEE, 2011, pp. 3173–3181. DOI: 10.1109/INFCOM.2011.5935165. URL: http://complex.elte.hu/~matray/publications/spotter_infocom2011.pdf.

[118] R. Landa, R. G. Clegg, J. T. Araújo, E. Mykoniati, D. Griffin, and M. Rio, "Measuring the Relationships between Internet Geography and RTT," in *International Conference on Computer Communications and Networks*, Piscataway, NJ: IEEE, 2013, pp. 1–7. DOI: 10.1109/ICCCN.2013.6614151. URL: http://richardclegg.org/sites/default/files/papers/raul_icccn_2013_0.pdf.

[119] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z. Zhang, and Y. Zhang, "IP-Geolocation Mapping for Moderately Connected Internet Regions," *Transactions on Parallel and Distributed Systems*, vol. 24, no. 2, pp. 381–391, 2013. DOI: 10.1109/TPDS.2012.136. URL: http://nasp.cs.tsinghua.edu.cn/GeoGet-TPDS.pdf.

[120] S. Lin, J. Chen, and Z. Niu, "Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction," *Tsinghua Science and Technology*, vol. 17, no. 3, pp. 256–264, 2012. DOI: 10.1109/TST.2012.6216755.

[121] P. Link, "China: The Anaconda in the Chandelier," *New York Review of Books*, 2002. URL: http://www.chinafile.com/library/nyrb-china-archive/china-anaconda-chandelier.

[122] *Luminati: largest business proxy service*, Web site, 2014–. URL: http://luminati.io.

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

[124] P. Mátray, P. Hága, S. Laki, G. Vattay, and I. Csabai, "On the spatial properties of internet routes," *Computer Networks*, vol. 56, no. 9, pp. 2237–2248, 2012, ISSN: 1389-1286. DOI: 10.1016/j.comnet.2012.03.005.

[125] A. McCallum, *{MALLET: A MACHINE LEARNING FOR LANGUAGE TOOLKIT}*, Software library, 2002. URL: http://mallet.cs.umass.edu/.

[126] A. McDonald, M. Bernhard, L. Valenta, B. VanderSloot, W. Scott, N. Sullivan, J. A. Halderman, and R. Ensafi, "403 Forbidden: A Global View of CDN Geoblocking," in *Internet Measurement Conference*, Boston: ACM, 2018, pp. 218–230. DOI: 10.1145/3278532.3278552.

[127] X. Mi, Y. Liu, X. Feng, X. Liao, B. Liu, X. Wang, F. Qian, Z. Li, S. Alrwais, and L. Sun, "Resident Evil: Understanding Residential IP Proxy as a Dark Service," in *Symposium on Security and Privacy*, Piscataway, NJ: IEEE, 2019, pp. 170–186. DOI: 10.1109/SP.2019.00011. URL: https://mixianghang.github.io/pubs/rpaas.pdf.

[128] D. M. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual Topic Models," in *Empirical Methods in Natural Language Processing*, P. Koehn and R. Mihalcea, Eds., Stroudsburg, PA: ACL, 2009, pp. 880–889, ISBN: 978-1-932432-59-6. URL: http://www.aclweb.org/anthology/D09-1092.

[129] W. Monroe, S. Green, and C. D. Manning, "Word segmentation of informal Arabic with domain adaptation," in *Annual Meeting of the Association for Computational Linguistics: Short Papers*, Stroudsburg, PA: ACL, 2014, pp. 206–211. URL: http://acl2014.org/acl2014/P14-2/pdf/P14-2034.pdf.

[130] J. A. Muir and P. C. Van Oorschot, "Internet Geolocation: Evasion and Counterevasion," *ACM Computing Surveys*, vol. 42, no. 1, 4:1–4:23, 2009. DOI: 10.1145/1592451.1592455.

[131] Z. Nabi, "The Anatomy of Web Censorship in Pakistan," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2013. URL: https://www.usenix.org/conference/foci13/workshop-program/presentation/nabi.

[132] A. Narayanan and B. Zevenbergen, "No Encore for Encore? Ethical Questions for Web-Based Censorship Measurement," Data & Society, Tech. Rep., 2015. DOI: 10.2139/ssrn.2665148.

[133] *New York Times Co. v. United States*, 403 U.S. 713, 1971. URL: https://supreme.justia.com/cases/federal/us/403/713/case.html.

[134] D. Newman, A. U. Asuncion, P. Smyth, and M. Welling, "Distributed Algorithms for Topic Models," *Journal of Machine Learning Research*, vol. 10, pp. 1801–1828, 2009. DOI: 10.1145/1577069.1755845. URL: http://doi.acm.org/10.1145/1577069.1755845.

[135] K. Y. Ng, A. Feldman, and C. Leberknight, "Detecting Censorable Content on Sina Weibo: A Pilot Study," in *Hellenic Conference on Artificial Intelligence*, Patras, Greece: ACM, 2018. DOI: 10.1145/3200947.3201037.

[136] J. Odvarko, *HTTP Archive 1.2 Specification*, Web page, 2007. URL: http://www.softwareishard.com/blog/har-12-spec/.

[137] V. N. Padmanabhan and L. Subramanian, "An Investigation of Geographic Mapping Techniques for Internet Hosts," in *SIGCOMM*, New York: ACM, 2001, pp. 173–185. DOI: 10.1145/964723.383073. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2001/08/sigcomm2001.pdf.

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

[139] T. Patterson, N. V. Kelso, *et al.*, *Natural Earth*, Free vector and raster map data, 2012. URL: http://www.naturalearthdata.com/.

[140] P. Pearce, R. Ensafi, F. Li, N. Feamster, and V. Paxson, "Augur: Internet-Wide Detection of Connectivity Disruptions," in *Symposium on Security and Privacy*, Piscataway, NJ: IEEE, 2017, pp. 427–443. DOI: 10.1109/SP.2017.55. URL: http://www.icir.org/vern/papers/oakland_2017_augur.pdf.

[141] P. Pearce, B. Jones, F. Li, R. Ensafi, N. Feamster, N. Weaver, and V. Paxson, "Global Measurement of DNS Manipulation," in *USENIX Security Symposium*, Vancouver, BC: USENIX, 2017, pp. 307–323. URL: http://www.icir.org/vern/papers/iris-dns-usesec17.pdf.

[142] *PlanetLab: an open platform for developing, deploying, and accessing planetary-scale services*, 2007. URL: http://www.planet-lab.org/.

[143] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "IP Geolocation Databases: Unreliable?" *SIGCOMM Computer Communications Review*, vol. 41, no. 2, pp. 53–56, 2011. DOI: 10.1145/1971162.1971171.

[144] M. F. Porter, "An Algorithm for Suffix Stripping," in *Readings in Information Retrieval*, K. Sparck Jones and P. Willett, Eds., San Francisco: Morgan Kaufmann, 1997, pp. 313–316, ISBN: 1-55860-454-5. URL: http://dl.acm.org/citation.cfm?id=275537.275705.

[145] J. Postel, Ed., *Transmission Control Protocol*, RFC 793, Sep. 1981. URL: https://tools.ietf.org/html/rfc793.

[146] J. Postel, Ed., *Internet Protocol*, RFC 791, Sep. 1981. URL: https://tools.ietf.org/html/rfc791.

[147] A. Razaghpanah, A. Li, A. Filastò, R. Nithyanand, V. Ververis, W. Scott, and P. Gill, *Exploring the Design Space of Longitudinal Censorship Measurement Platforms*, 2016. arXiv: 1606.01979 [cs.NI].

[148] *Readability: Read Comfortably—Anytime, Anywhere*, Online service, 2009–2016. URL: https://readability.com/about.

[149] Reporters Without Borders, *World Press Freedom Index, 2018*, 2018. URL: https://rsf.org/en/ranking/2018.

[150] C. R. Richardson, P. J. Resnick, D. L. Hansen, H. A. Derry, and V. J. Rideout, "Does Pornography-Blocking Software Block Access to Health Information on the Internet?" *The Journal of the American Medical Association*, vol. 288, no. 22, pp. 2887–2894, 2002. DOI: 10.1001/jama.288.22.2887.

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

[152] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004. DOI: 10.1108/00220410410560582.

[153] P. K. Roy, "India net neutrality rules could be world's strongest," *BBC News*, Nov. 2017. URL: https://www.bbc.com/news/world-asia-india-42162979.

[154] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists," in *Internet Measurement Conference*, Boston: ACM, 2018, pp. 478–493. DOI: 10.1145/3278532.3278574. arXiv: 1805.11506 [cs.NI].

[155] F. Schneider, B. Ager, G. Maier, A. Feldmann, and S. Uhlig, "Pitfalls in HTTP Traffic Measurements and Analysis," in *Passive and Active Measurement*, ser. Lecture Notes in Computer Science, vol. 7192, Berlin, Heidelberg: Springer, 2012, pp. 242–251. DOI: 10.1007/978-3-642-28537-0_24.

[156] W. Scott, T. Anderson, T. Kohno, and A. Krishnamurthy, "Satellite: Joint Analysis of CDNs and Network-Level Interference," in *USENIX Annual Technical Conference*, Denver, CO: USENIX, 2016, pp. 195–208. URL: https://www.usenix.org/system/files/conference/atc16/atc16_paper-scott.pdf.

[157] A. Sfakianakis, E. Athanasopoulos, and S. Ioannidis, "CensMon: A Web Censorship Monitor," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2011. URL: https://www.usenix.org/legacy/event/foci11/tech/final_files/Sfakianakis.pdf.

[158] Y. Shavitt and N. Zilberman, "A Geolocation Databases Study," *Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011. DOI: 10.1109/JSAC.2011.111214. URL: https://www.cl.cam.ac.uk/~nz247/publications/JSAC2011-Geolocation.pdf.

[159] R. Singh, H. Koo, N. Miramirkhani, F. Mirhaj, P. Gill, and L. Akoglu, "The Politics of Routing: Investigating the Relationship Between AS Connectivity and Internet Freedom," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2016. arXiv: 1606.02230 [cs.NI].

[160] D. Sites, *Compact Language Detection 2*, Software library, 2013–. URL: https://code.google.com/p/cld2/.

[161] K. Soska and N. Christin, "Automatically Detecting Vulnerable Websites Before They Turn Malicious," in *USENIX Security Symposium*, Berkeley, CA: USENIX, 2014, pp. 625–640. URL: https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-soska.pdf.

[162] R. Subramanian, "The Growth of Global Internet Censorship and Circumvention: A Survey," *Communications of the IIMA*, vol. 11, no. 2, 2011. URL: http://scholarworks.lib.csusb.edu/ciima/vol11/iss2/6/.

[163] F. Sun, D. Song, and L. Liao, "DOM Based Content Extraction via Text Density," in *Research and Development in Information Retrieval*, New York: ACM, 2011, pp. 245–254. DOI: 10.1145/2009916.2009952. URL: http://www.ofey.me/papers/cetd-sigir11.pdf.

[164] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich, "The Long "Taile" of Typosquatting Domain Names," in *USENIX Security Symposium*, Berkeley, CA: USENIX, 2014, pp. 191–206. URL: https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-szurdi.pdf.

[165] The Common Crawl Foundation, *Common Crawl*, Web site, 2008–. URL: http://commoncrawl.org.

[166] The Open Observatory of Network Interference, *Contributing to test lists*, 2016. URL: https://ooni.torproject.org/get-involved/contribute-test-lists/.

[167] The OpenNet Initiative, *ONI Country Profiles*, Web site, 2007–. URL: https://opennet.net/research/profiles.

[168] ——, *ONI Country Profiles: Thailand*, Web report, 2007–. URL: https://opennet.net/research/profiles/thailand.

[169] ——, *United States and Canada Overview [of Censorship]*, Web report, 2007–. URL: https://opennet.net/research/regions/namerica.

[170] M. C. Tschantz, S. Afroz, Anonymous, and V. Paxson, "SoK: Towards Grounding Censorship Circumvention in Empiricism," in *Symposium on Security and Privacy*, Piscataway, NJ: IEEE, 2016, pp. 914–933. DOI: 10.1109/SP.2016.59.

[171] M. C. Tschantz, S. Afroz, S. Sajid, S. A. Qazi, M. Javed, and V. Paxson, "A Bestiary of Blocking: The Motivations and Modes behind Website Unavailability," in *Free and Open Communications on the Internet*, Baltimore, MD: USENIX, 2018. arXiv: 1806.00459 [cs.NI]. URL: https://www.usenix.org/conference/foci18/presentation/tschantz.

[172] University of Wisconsin, *Internet Atlas (DS-468)*, Continuously updated data set, 2011. DOI: 10.23721/110/1353976.

[173] B. VanderSloot, A. McDonald, W. Scott, J. A. Halderman, and R. Ensafi, "Quack: Scalable Remote Measurement of Application-Layer Censorship," in *USENIX Security Symposium*, Baltimore, MD: USENIX, 2018, pp. 187–202. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/vandersloot.

[174] J.-P. Verkamp and M. Gupta, "Inferring Mechanics of Web Censorship Around the World," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2012. URL: https://www.usenix.org/conference/foci12/workshop-program/presentation/verkamp.

[175] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking Sensors: Analyzing and Detecting Parked Domains," in *Network and Distributed Security Symposium*, Reston, VA: Internet Society, 2015. URL: http://www.internetsociety.org/sites/default/files/01_2_2.pdf.

[176] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. M. Mimno, "Evaluation methods for topic models," in *International Conference on Machine Learning*, A. P. Danyluk, L. Bottou, and M. L. Littman, Eds., vol. 382, New York: ACM, 2009, pp. 1105–1112, ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553515.

[177] M. Wander, "Measurement survey of server-side DNSSEC adoption," in *Network Traffic Measurement and Analysis*, Piscataway, NJ: IEEE, 2017, pp. 1–9. DOI: 10.23919/TMA.2017.8002913. URL: http://tma.ifip.org/wp-content/uploads/sites/7/2017/06/tma2017_paper58.pdf.

[178] Z. Wang, Ed., *Navigation Timing*, W3C Recommendation, 2012. URL: http://www.w3.org/TR/2012/REC-navigation-timing-20121217/.

[179] D. Wang and G. Mark, "Internet Censorship in China: Examining User Awareness and Attitudes," *Transactions on Computer-Human Interaction*, vol. 22, no. 6, 31:1–31:22, 2015. DOI: 10.1145/2818997. URL: https://escholarship.org/uc/item/48x7k7j2.

[180] D. Y. Wang, S. Savage, and G. M. Voelker, "Cloak and Dagger: Dynamics of Web Search Cloaking," in *Computer and Communications Security*, New York: ACM, 2011, pp. 477–490. URL: http://cseweb.ucsd.edu/~voelker/pubs/cloaking-ccs11.pdf.

[181] Y. Wang, D. Burgener, M. Flires, A. Kuzmanovic, and C. Huang, "Towards Street-Level Client-Independent IP Geolocation," in *Networked Systems Design and Implementation*, Berkeley, CA: USENIX, 2011, pp. 365–379. URL: https://www.usenix.org/legacy/events/nsdi11/tech/full_papers/Wang_Yong.pdf.

[182] Z. Wang, Y. Cao, Z. Qian, C. Song, and S. V. Krishnamurthy, "Your State is Not Mine: A Closer Look at Evading Stateful Internet Censorship," in *Internet Measurement Conference*, London, UK: ACM, 2017, pp. 114–127. DOI: 10.1145/3131365.3131374. URL: https://conferences.sigcomm.org/imc/2017/papers/imc17-final59.pdf.

[183] N. Weaver, C. Kreibich, M. Dam, and V. Paxson, "Here Be Web Proxies," in *Passive and Active Measurement*, Los Angeles, CA, USA: Springer, 2014, pp. 183–192. DOI: 10.1007/978-3-319-04918-2_18.

[184] N. Weaver, R. Sommer, and V. Paxson, "Detecting Forged TCP Reset Packets," in *Network and Distributed System Security*, Reston, VA: Internet Society, 2009. URL: http://www.icsi.berkeley.edu/pubs/networking/ndss09-resets.pdf.

[185] M. Webb, *Illusions of Security: Global Surveillance and Democracy in the Post-9/11 World*. San Francisco: City Lights, 2007, ISBN: 0-87286-476-6.

[186] Z. Weinberg, S. Cho, N. Christin, V. Sekar, and P. Gill, "How to Catch when Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation," in *Internet Measurement Conference*, Boston: ACM, 2018, pp. 203–217. DOI: 10.1145/3278532.3278551.

[187] Z. Weinberg, M. Sharif, J. Szurdi, and N. Christin, "Topics of Controversy: An Empirical Analysis of Web Censorship Lists," in *Privacy Enhancing Technologies*, Berlin: De Gruyter, 2017, pp. 42–61. DOI: 10.1515/popets-2017-0004.

[188] P. Winter, "Towards a Censorship Analyser for Tor," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2013. URL: https://www.usenix.org/conference/foci13/workshop-program/presentation/Winter.

[189] P. Winter and S. Lindskog, "How the Great Firewall of China is Blocking Tor," in *Free and Open Communications on the Internet*, Berkeley, CA: USENIX, 2012. URL: https://www.usenix.org/conference/foci12/workshop-program/presentation/Winter.

[190] B. Wong, I. Stoyanov, and E. G. Sirer, "Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts," in *Networked Systems Design and Implementation*, Berkeley, CA: USENIX, 2007. URL: https://www.usenix.org/legacy/events/nsdi07/tech/full_papers/wong/wong.pdf.

[191] J. Wright, "Regional Variation in Chinese Internet Filtering," *Information, Communication & Society*, vol. 17, no. 1, pp. 121–141, 2014. DOI: 10.1080/1369118X.2013.853818. URL: http://ssrn.com/abstract=2265775.

[192] J. Wright, A. Darer, and O. Farnan, *Filterprints: Identifying Localized Usage Anomalies in Censorship Circumvention Tools*, 2016. arXiv: 1507.05819 [cs.CY].

[193] ——, "On Identifying Anomalies in Tor Usage with Applications in Detecting Internet Censorship," in *Web Science*, Amsterdam, Netherlands: ACM, 2018, pp. 87–96. DOI: 10.1145/3201064.3201093. URL: https://ora.ox.ac.uk/objects/uuid:31bc9ebf-457f-48d4-b527-e1b2325318bf/.

[194] P. Xie, Y. Deng, and E. Xing, "Diversifying restricted boltzmann machine for document modeling," in *Knowledge Discovery and Data Mining*, New York: ACM, 2015, pp. 1315–1324. DOI: 10.1145/2783258.2783264.

[195] X. Xu, Z. M. Mao, and J. A. Halderman, "Internet Censorship in China: Where Does the Filtering Occur?" In *Passive and Active Measurement*, ser. Lecture Notes in Computer Science, vol. 6579, Berlin, Heidelberg: Springer, 2011, pp. 133–142. DOI: 10.1007/978-3-642-19260-9. URL: https://censorbib.nymity.ch/pdf/Xu2011a.pdf.

[196] T. K. Yadav, A. Sinha, D. Gosain, P. K. Sharma, and S. Chakravarty, "Where The Light Gets In: Analyzing Web Censorship Mechanisms in India," in *Internet Measurement Conference*, Boston: ACM, 2018, pp. 252–264. DOI: 10.1145/3278532.3278555. arXiv: 1808.01708 [cs.CY].

[197] L. Yao, D. M. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Knowledge Discovery and Data Mining*, J. F. Elder IV, F. Fogelman-Soulié, P. A. Flach, and M. J. Zaki, Eds., New York: ACM, 2009, pp. 937–946, ISBN: 978-1-60558-495-9. DOI: 10.1145/1557019.1557121. URL: http://doi.acm.org/10.1145/1557019.1557121.

[198] T. Yasseri, A. Spoerri, M. Graham, and J. Kertész, "Global Wikipedia: International and cross-cultural issues in online collaboration," in *The Most Controversial Topics in Wikipedia: A multilingual and geographical analysis*, P. Fichman and N. Hara, Eds., Lanham, MD: Rowman & Littlefield, 2014, ISBN: 0-8108-9101-8. arXiv: 1305.5566 [cs.CL].

[199] E. Zhu, F. Nargesian, K. Q. Pu, and R. Miller, "LSH Ensemble: Internet-Scale Domain Search," *Proceedings of the VLDB Endowment*, vol. 9, no. 12, pp. 1185–1196, 2016. arXiv: 1603.07410 [cs.DB].

[200] T. Zhu, D. Phipps, A. Pridgen, J. R. Crandall, and D. S. Wallach, "The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions," in *USENIX Security Symposium*, Berkeley, CA: USENIX, 2013, pp. 227–240. URL: https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/zhu.

[201] J. Zittrain and B. Edelman, "Internet filtering in China," *IEEE Internet Computing*, vol. 7, no. 2, pp. 70–77, 2003. DOI: 10.1109/MIC.2003.1189191. URL: http://www.academia.edu/download/4984692/10.1.1.98.8690.pdf.

[202] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. B. Duarte, "Improving the accuracy of measurement-based geographic location of Internet hosts," *Computer Networks*, vol. 47, no. 4, pp. 503–523, 2005. DOI: 10.1016/j.comnet.2004.08.013.