Global Measurement of DNS Manipulation

Paul Pearce, *UC Berkeley;* Ben Jones, *Princeton;* Frank Li, *UC Berkeley;*
Roya Ensafi and Nick Feamster, *Princeton;* Nick Weaver, *ICSI;* Vern Paxson, *UC Berkeley*

https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/pearce

This paper is included in the Proceedings of the
26th USENIX Security Symposium

August 16–18, 2017 • Vancouver, BC, Canada

ISBN 978-1-931971-40-9

Global Measurement of DNS Manipulation

Paul Pearce[◇] *Ben Jones*[†] *Frank Li*[◇] *Roya Ensafi*[†]
Nick Feamster[†] *Nick Weaver*[‡] *Vern Paxson*[◇]

[◇]*University of California, Berkeley* [†]*Princeton University*
[‡] *International Computer Science Institute*

{*pearce, frankli, vern*}*@cs.berkeley.edu* {*bj6, rensafi, feamster*}*@cs.princeton.edu*
nweaver@icsi.berkeley.edu

Abstract

Despite the pervasive nature of Internet censorship and the continuous evolution of how and where censorship is applied, measurements of censorship remain comparatively sparse. Understanding the scope, scale, and evolution of Internet censorship requires global measurements, performed at regular intervals. Unfortunately, the state of the art relies on techniques that, by and large, require users to directly participate in gathering these measurements, drastically limiting their coverage and inhibiting regular data collection. To facilitate large-scale measurements that can fill this gap in understanding, we develop Iris, a scalable, accurate, and ethical method to measure global manipulation of DNS resolutions. Iris reveals widespread DNS manipulation of many domain names; our findings both confirm anecdotal or limited results from previous work and reveal new patterns in DNS manipulation.

1 Introduction

Anecdotes and reports indicate that Internet censorship is widespread, affecting at least 60 countries [29, 39]. Despite its pervasive nature, empirical Internet measurements revealing the scope and evolution of Internet censorship remain relatively sparse. A more complete understanding of Internet censorship around the world requires *diverse* measurements from a wide range of geographic regions and ISPs, not only across countries but also within regions of a single country. Diversity is important even within countries, because political dynamics can vary internally, and because different ISPs may implement filtering policies differently.

Unfortunately, most mechanisms for measuring Internet censorship currently rely on volunteers who run measurement software deployed on their own Internet-connected devices (e.g., laptops, phones, tablets) [43, 49]. Because these tools rely on people to install software and perform measurements, it is unlikely that they can ever achieve the scale required to gather continuous and diverse measurements about Internet censorship. Performing measurements of the scale and frequency necessary to understand the scope and evolution of Internet censorship calls for fundamentally new techniques that do not require human involvement or intervention.

We aim to develop techniques that can perform widespread, longitudinal measurements of global Internet manipulation without requiring the participation of individual users in the countries of interest. Organizations may implement censorship at many layers of the Internet protocol stack; they might, for example, block traffic based on IP address, or they might block individual web requests based on keywords. Recent work has developed techniques to continuously measure widespread manipulation at the transport [23,42] and HTTP [45] layers, yet a significant gap remains in our understanding of global information control concerning the manipulation of the Internet's Domain Name System (DNS). Towards this goal, we develop and deploy a method and system to detect, measure, and characterize the manipulation of DNS responses in countries across the entire world.

Developing a technique to *accurately* detect DNS manipulation poses major challenges. Although previous work has studied inconsistent or otherwise anomalous DNS responses [32, 34], these methods have focused mainly on identifying DNS responses that could reflect a variety of underlying causes, including misconfigurations. In contrast, our work aims to develop methods for accurately identifying DNS manipulation indicative of an intent to restrict user access to content. To achieve high detection accuracy, we rely on a collection of metrics that we base on the underlying properties of DNS domains, resolutions, and infrastructure.

One set of detection metrics focuses on *consistency*—intuitively, when we query a domain from different locations, the IP addresses contained in DNS responses should reflect hosting from either a common server (i.e., the same IP address) or the same autonomous system.

Another set of detection metrics focuses on *independent verifiability*, by comparison to independent information such as the identity in the TLS certificate for the website corresponding to the domain. Each of these metrics naturally lends itself to exceptions: for example, queries from different locations utilizing a content distribution network (CDN) will often receive different IP addresses (and sometimes even different CDNs). However, we can use violations of *all* of the metrics as a strong indicator of DNS manipulation.

In addition to achieving accurate results, another significant design challenge concerns *ethics*. In contrast to systems that explicitly involve volunteers in collecting measurements, methods that send DNS queries through open DNS resolvers deployed across the Internet raise the issue of potentially implicating third parties who did not in fact agree to participate in the measurement. Using "open resolvers" is potentially problematic, as most of these are not actual resolvers but instead DNS forwarders in home routers and other devices [46]. A censor may misattribute requests from these resources as individual citizens attempting to access censored resources.

Reasoning about the risks of implicating individual citizens requires detailed knowledge of how censors in different countries monitor access to censored material and how they penalize such actions. These policies and behaviors may be complex, varying across time, region, individuals involved, and the nature of the censored content; such risks are likely intractable to accurately deduce. To this end, our design takes steps to ensure that, to the extent possible, we only query open DNS resolvers hosted in Internet *infrastructure* (e.g., within Internet service providers or cloud hosting providers), in an attempt to eliminate any use of resolvers or forwarders in the home networks of individual users. This step reduces the set of DNS resolvers that we can use for our measurements from tens of millions to only a few thousand. However, we find that the resulting coverage still suffices to achieve a global view of DNS manipulation, and—importantly—in a safer way than previous studies that exploit open DNS resolvers.

Our work makes the following contributions. First, we design, implement, and deploy Iris, a scalable, ethical system for measuring DNS manipulation. Second, we develop analysis metrics for disambiguating natural variation in DNS responses for a domain from nefarious manipulation. Third, we perform a global measurement study that highlights the heterogeneity of DNS manipulation, across countries, resolvers, and domains. We find that manipulation varies across DNS resolvers even within a single country.

2 Related Work

Country-specific censorship studies. In recent years many researchers have investigated the whats, hows, and whys of censorship in particular countries. These studies often span a short period of time and reflect a single vantage point within a target country, such as by renting virtual private servers. For example, studies have specifically focused on censorship practices in China [55], Iran [7], Pakistan [38], Syria [12], and Egypt [8]. Studies have also explored the employment of various censorship methods, e.g., injection of fake DNS replies [5, 36], blocking of TCP/IP connections [54], and application-level blocking [19, 33, 41]. A number of studies suggest that countries sometimes change their blocking policies and methods in times surrounding political events. For example, Freedom House reports 15 instances of Internet shutdowns—where the government cut off access to Internet entirely—in 2016 alone [29]. Most of these were apparently intended to prevent citizens from reaching social media to spread unwanted information.

Other studies have demonstrated that government censorship covers a broad variety of services and topics, including video portals (e.g.,`youtube.com`) [51], blogs (e.g., `livejournal.com`) [3], and news sites (e.g., `bbc.com`) [9]. Censors also target circumvention and anonymity tools; most famously, the Great Firewall of China has engaged in a decade-long cat-and-mouse game with Tor [24, 53]. Although these studies provide important data points, each reflects a snapshot at a single point in time and thus cannot capture ongoing trends and variations in censorship practices.

Global censorship measurement tools. Several research efforts developed platforms to measure censorship by running experiments from diverse vantage points. For instance, CensMon [48] used PlanetLab nodes in different countries, and UBICA [1] aimed to increase vantage points by running censorship measurement software on home gateway devices and user desktops. In practice, as far as we know, neither of these frameworks are still deployed and collecting data. The OpenNet Initiative [39] has used its public profile to recruit volunteers around the world who have performed one-off measurements from home networks each year for the past ten years. OONI [49] and ICLab [30], two ongoing data collection projects, use volunteers to run both custom software and custom embedded devices (such as Raspberry Pis [26]).

Although each of these frameworks can perform a extensive set of tests, they rely on volunteers who run measurement software on their Internet-connected devices. These human involvements make it more challenging—if not impossible—to gather continuous and diverse measurements.

Pearce et al. recently developed Augur, a method to perform longitudinal global measurement using TCP/IP side channels [42]. Although Augur examines a similar set of domains and countries as Iris, it focuses on identifying IP-based disruption rather than DNS-based manipulation.

Measuring DNS manipulation. The DNS protocol's lack of authentication and integrity checking makes it a prime target for attacks. Jones et al. presented techniques for detecting unauthorized DNS root servers, though found little such manipulation in practice [32]. Jiang et al. identified a vulnerability in DNS cache update policies that allows malicious domains to stay in the cache even if removed from the zone file [31].

Several projects have explored DNS manipulation using a limited number of vantage points. Weaver et al. explored DNS manipulation with respect to DNS redirection for advertisement purposes [52]. The authors also observed incidents in which DNS resolvers redirected end hosts to malware download pages. There are many country-specific studies that show how different countries use a variety of DNS manipulation techniques to exercise Internet censorship. For example, in Iran the government expects ISPs to configure their DNS resolvers to redirect contentious domains to a censorship page [7]. In Pakistan, ISPs return NXDOMAIN responses [38]. In China, the Great Firewall injects forged DNS packets with seemingly arbitrary IP addresses [5]. These studies however all drew upon a small or geographically limited set of vantage points, and for short periods of time.

Using open resolvers. A number of studies have explored DNS manipulation at a larger scale by probing the IPv4 address space to find open resolvers. In 2008, Dagon et al. found corrupt DNS resolvers by running measurements using 200,000 open resolvers [18]; they do not analyze the results for potential censorship. A similar scan by anonymous authors [4] in 2012 showed evidence of Chinese DNS censorship affecting non-Chinese systems.

Follow-on work in 2015 by Kührer et al. tackled a much larger scope: billions of lookups for 155 domain names by millions of open resolvers [34]. The study examined a broad range of potentially tampered results, which in addition to censorship included malware, phishing, domain parking, ad injection, captive portals, search redirection, and email delivery. They detected DNS manipulation by comparing DNS responses from open resolvers with ground truth resolutions gathered by querying control resolvers. They then identified legitimate unmanipulated answers using a number of heuristic filtering stages, such as treating a differing response as legitimate if its returned IP address lies within the same AS the ground truth IP address.

We tried to use their method for conducting global measurements specifically for detecting censorship. However, censorship detection was not a focus of their work, and the paper does not explicitly describe the details of its detection process. In particular, other than examining HTTP pages for "blocked by the order of ..." phrasing, the paper does not present a decision process for determining whether a given instance of apparent manipulation reflects censorship or some other phenomenon. In addition, their measurements leverage open resolvers *en masse*, which raises ethical concerns for end users who may be wrongly implicated for attempting to access banned content. In contrast, we frame an explicit, reproducible method for globally measuring DNS-based manipulation in an ethically responsible manner.

In 2016, Scott et al. introduced Satellite [47], a system which leverages open resolvers to identify CDN deployments and network interference using collected resolutions. Given a bipartite graph linking domains queried with IP address answers collected from the open resolvers, Satellite identifies strongly connected components, which represent domains hosted by the same servers. Using metrics for domain similarity based on the overlap in IP addresses observed for two domains, Satellite distinguishes CDNs from network interference as components with highly similar domains (additionally, other heuristics help refine this classification).

3 Method

In this section we describe Iris, a scalable, lightweight system to detect DNS manipulation. We begin by scoping the problem space, identifying the capabilities and limitations of various measurement building blocks, and stating our assumptions about the threat model. We explain the process by which we select (1) which domain names to measure, and (2) the vantage points to measure them from, taking into consideration questions of ethics and scalability. We then describe, given a set of measurement vantage points and DNS domain names, how we characterize the results of our measurements and use them to draw conclusions about whether DNS manipulation is taking place, based on either the *consistency* or the *independent verifiability* of the responses that we receive. Next, we consider our technical approach in light of existing ethical norms and guidelines, and explain how various design decisions help us adhere to those principles as much as possible. Finally, we discuss the implicit and technical limitations of Iris.

3.1 Overview

We aim to identify DNS manipulation, which we define as the instance of a DNS response both (1) having attributes (e.g., IP addresses, autonomous systems, web content) that are not consistent with respect to a well-defined control set; and (2) returning information that is demonstrably incorrect when compared against independent information sources (e.g., TLS certificates).

Approach. Detecting DNS manipulation is conceptually simple: At a high-level, the idea entails performing DNS queries through geographically distributed DNS resolvers and analyzing the responses for activity that suggests that the responses for a DNS domain might be manipulated. Despite its apparent simplicity, however, realizing a system to scalably collect DNS data and analyze it for manipulation poses both ethical and technical challenges. The ethical challenges concern selecting DNS resolvers that do not implicate innocent citizens, as well as ensuring that Iris does not induce undue load on the DNS resolution infrastructure; §3.2 explains the ethical guidelines we use to reason about design choices. §3.3 describes how Iris selects a "safe" set of open DNS resolvers; The technical challenges center around developing sound methods for detecting manipulation, which we describe in §3.4 and §3.5.

Identifying DNS names to query. Iris queries a list of sensitive URLs compiled by Citizen Lab [14]. We call this list the Citizen Lab Block List (CLBL). This list of URLs is compiled by experts based on known censorship around the world, divided by category. We distill the URLs down to domain names and use this list as the basis of our dataset. We then supplement this list by adding additional domain names selected at random from the Alexa Top 10,000 [2]. These additional domain names help address geographic or content biases in the the CLBL while not drastically increasing the total number of queries.

Assumptions and focus. First, Iris aims to identify widespread manipulation at the scale of Internet service providers and countries. We cannot identify manipulation that is targeted at specific individuals or populations or manipulation activities that exploit high-value resources such as valid but stolen certificates. Second, we focus on manipulation tactics that do not rely on stealth; we assume that adversaries will use DNS resolvers to manipulate the responses to DNS queries. We assume that adversaries do not return IP addresses that are incorrect but within the same IP prefix as a correct answer [5, 7, 38]. Finally, when attributing DNS manipulation to a particular country or dependent territory, we rely on the country information available from Censys [21] supplemented with MaxMind's [37] dataset to map a resolver to a specific country (or dependent territory).

3.2 Ethics

The design of Iris incorporates many considerations regarding ethics. Our primary ethical concern is the risks associated with the measurements that Iris conducts, as issuing DNS queries for potentially censored or manipulated DNS domains through resolvers that we do not own could potentially implicate otherwise innocent users. A second concern is whether the DNS queries that we generate introduce undue query load on authoritative DNS nameservers for domains that we do not own. With these concerns in mind, we consider the ethics of performing measurements with Iris, using the ethical guidelines of the Belmont Report [10] and Menlo Report [20] to frame our discussion.

One important ethical principle is *respect for persons*; essentially, this principle states that an experiment should respect the rights of humans as autonomous decision-makers. Sometimes this principle is misconstrued as a requirement for informed consent for all experiments. In many cases, however, informed consent is neither practical nor necessary; accordingly, Salganik [44] characterizes this principle instead as "some consent for most things". In the case of Iris, obtaining the consent of all open DNS resolver operators is impractical.

In lieu of attempting to obtain informed consent, we turn to the principle of *beneficence*, which weighs the benefits of conducting an experiment against the risks associated with the experiment. Note that the goal of beneficence is not to *eliminate* risk, but merely to *reduce* it to the extent possible. Iris's design relies heavily on this principle: Specifically, we note that the benefit of issuing DNS queries through tens of millions of resolvers has rapidly diminishing returns, and that using only open resolvers that we can determine are unlikely to correspond to individual users greatly reduces the risk to any individual without dramatically reducing the benefits of our experiment. We note that our consideration of ethics in this regard is a significant departure from previous work that has issued queries through open DNS resolver infrastructure but has not considered ethics.

The principle of *justice* states that the beneficiaries of an experiment should be the same population that bears the risk of that experiment. On this front, we envision that the beneficiaries of the kinds of measurements that we collect using Iris will be wide-ranging: designers of circumvention tools, as well as policymakers, researchers, and activists who are improving communications and connectivity for citizens in oppressive regimes all need better data about the extent and scope of Internet censorship. In short, even in the event that some entity in a country that hosts an open DNS resolver might bear some risk as a result of the measurements we conduct, we envision that those same entities may ultimately benefit from the research, policy-making, and tool development that Iris facilitates.

A final guideline concerns *respect for law and public interest*, which essentially extends the principle of beneficence to all relevant stakeholders, not only the experiment participants. This principle is useful for reasoning about the externalities that our DNS queries create by increasing DNS query load on the nameservers for various DNS domains. To abide by this principle, we rate-limit our DNS queries for each DNS domain to ensure that the owners of these domains do not face large expenses as a result of the queries that we issue. This rate limit is necessary because some DNS service providers charge based on the peak or near peak query rate.

3.3 Open DNS Resolvers

To obtain a wide range of measurement vantage points, we use *open DNS resolvers* deployed around the world; such resolvers will resolve queries for any client.

Measurement using open DNS resolvers is an ethically complex issue. Previous work has identified tens of millions of these resolvers around the world [34]. Given their prevalence and global diversity, open resolvers are a compelling resource, providing researchers with considerable volume and reach. Unfortunately, open resolvers also pose a risk not only to the Internet but to individual users.

Open resolvers can be the result of configuration errors, frequently on end-user devices such as home routers [34]. Using these devices for measurement can incur monetary cost, and if the measurement involves sensitive content or hosts, can expose the owner to harm. Furthermore, open resolvers are also a common tool in various online attacks such as Distributed Denial-of-Service (DDoS) amplification attacks [35]. Despite efforts to reduce both the prevalence of open resolvers and their potential impact [40], they remain commonplace.

Due to these and the ethics considerations that we discussed in §3.2, we restrict the set of open resolvers that we use to the few thousand resolvers that we are reasonably certain are part of the Internet infrastructure (e.g., belonging to Internet service providers, online cloud hosting providers), as opposed to attributable to any single individual. Figure 1 illustrates the process by which Iris finds safe open DNS resolvers. We now explain this process in more detail. Conceptually, the process comprises two steps: (1) scanning the Internet for open DNS resolvers; or (2) pruning the list of open DNS resolvers that we identify to limit the resolvers to a set that we can reasonably attribute to Internet infrastructure.

By using DNS resolvers we do not control, we cannot differentiate between country-wide or state-mandated censorship and localized manipulation (e.g., captive portals, malware [34]) at individual resolvers. Therefore we must aggregate and analyze results at ISP or country scale.

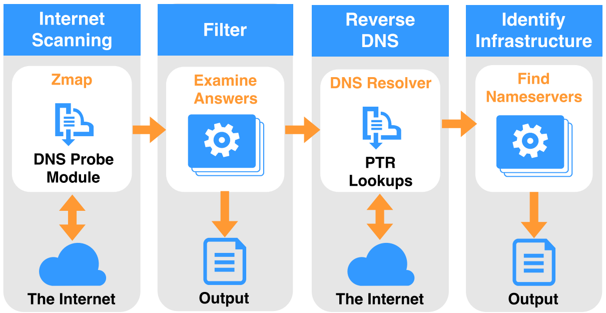

Figure 1: *Overview of Iris's DNS resolver identification and selection pipeline. Iris begins with a global scan of the entire IPv4 address space, followed by reverse DNS PTR lookups for all open resolvers, and finally filtering resolvers to only include DNS infrastructure.*

Step 1: Scanning the Internet's IPv4 space for open DNS resolvers. Scanning the IPv4 address space provides us with a global perspective on all open resolvers. To do so, we developed an extension to the ZMap [22] network scanner to enable Internet-wide DNS resolutions[1]. This module queries port 53 of all IPv4 addresses with a recursive DNS A record query. We use a purpose-registered domain name we control for these queries to ensure there is a known correct answer. We conduct measurements and scans from IP addresses having a `PTR` record identifying the machine as a "research scanner." These IP addresses also host a webpage identifying our academic institution and offering the ability to opt-out of scans. From these scans, we select all IP addresses that return the correct answer to this query and classify them as open resolvers. In §4.1, we explore the population of open DNS resolvers that we use for our study.

Step 2: Identifying Infrastructure DNS Resolvers. Given a list of all open DNS resolvers on the Internet, we prune this list to include only DNS resolvers that can likely be attributed to Internet infrastructure. To do so, we aim to identify open DNS resolvers that appear to be authoritative nameservers for a given DNS domain. Iris performs reverse DNS `PTR` lookups for all open resolvers and retains only the resolvers that have a valid `PTR` record beginning with the subdomain `ns[0-9]+` or `nameserver[0-9]*`. This filtering step reduces the number of usable open resolvers—from millions to thousands—yet even the remaining set of open DNS resolvers provides broad country- and network-level coverage (characterized further in §4.1).

Using `PTR` records to identify infrastructure can have

[1] Our extension has been accepted into the open source project and the results of our scans are available as part of the Censys [21] system.

both *false negatives* and *false positives*. Not all infrastructure resolvers will have a valid PTR record, nor will they all be authoritative nameservers. These false negatives limit the scope and scale of our measurement, but are necessary to reduce risk. Similarly, if a user operated their own authoritative nameserver on their home IP or if a PTR record matched our naming criteria but was not authoritative, our method would identify that IP as infrastructure (false positives).

3.4 Performing the Measurements

Given a list of DNS domain names to query and a global set of open DNS resolvers from which we can issue queries, we need a mechanism that issues queries for these domains to the set of resolvers that we have at our disposal. Figure 2 shows an overview of the measurement process. At a high level, Iris resolves each DNS domain using the global vantage points afforded by the open DNS resolvers, annotates the response IP addresses with information from both outside datasets as well as additional active probing, and uses *consistency* and *independent verifiability* metrics to identify manipulated responses. The rest of this section outlines this measurement process in detail, while §3.5 describes how we use the results of these measurements to ultimately identify manipulation.

Step 1: Performing global DNS queries. Iris takes as input a list of suitable open DNS resolvers, as well as the combined CLBL and Alexa domain names. In addition to the DNS domains that we are interested in testing, we include 3 DNS domains that are under our control to help us compute our consistency metrics when identifying manipulation.

Querying tens of thousands of domains across tens of thousands of resolvers required the development of a new DNS query tool, because no existing DNS measurement tool supports this scale. We implemented this tool in Go [27]. The tool takes as input a set of domains and resolvers, and coordinates random querying of each domain across each resolver. The tool supports a variety of query types, multiple of which can be specified per run, including A, AAAA, MX, and ANY. For each (domain, resolver) pair, the tool crafts a recursive DNS request and sends it to the resolver. The recursive query requests that the resolver resolve the domain and return the ultimate answer, logging all responses, including timeouts. The tool follows the set of responses to resolve each domain to an IP address. For example, if a resolver returns a CNAME, the tool then queries the resolver for resolution of that CNAME.

To ensure resolvers are not overloaded, the tool includes a configurable rate-limit. For our experiments, we limited queries to resolvers to an upper bound of 5 per second. In practice, this rate tends to be much lower due to network latency in both reaching the resolver, as well as the time it takes the resolver to perform the recursive response. To cope with specific resolvers that are unstable or timeout frequently, the tool provides a configurable failure threshold that halts a specific resolver's set of measurements should too many queries fail.

To ensure the domains we query are not overloaded, the tool randomizes the order of domains and limits the number of resolvers queried in parallel such that in the worst case no domain experiences more than 1 query per second, in expectation.

Step 2: Annotating DNS responses with auxiliary information. Our analysis ultimately relies on characterizing both the *consistency* and *independent verifiability* of the DNS responses that we receive. To enable this classification we first must gather additional details about the IP addresses that are returned in each of the DNS responses. Iris annotates each IP address returned in the set of DNS responses with additional information about each IP address's geolocation, autonomous system (AS), port 80 HTTP responses, and port 443 HTTPS X.509 certificates. We rely on the Censys [21] dataset for this auxiliary information; Censys provides daily snapshots of this information. This dataset does not contain every IP address; for example, the dataset does not include IP addresses that have no open ports, or adversaries may intentionally return IP addresses that return error pages or are otherwise unresponsive. In these cases, we annotate all IP addresses in our dataset with AS and geolocation information from the Maxmind service [37].

Additional PTR and TLS scanning. For each IP address, we perform a DNS PTR lookup to assist with some of our subsequent consistency characterization (a process we detail in §3.5). Another complication in the annotation exercise relates to the fact that in practice a single IP address might host many websites via HTTP or HTTPS (i.e., virtual hosting). As a result, when Censys retrieves certificates via port 443 (HTTPS) across the entire IPv4 address space, the certificate that Censys retrieves might differ from the certificate that the server would return in response to a query via TLS's Server Name Indication (SNI) extension. Such a discrepancy might lead Iris to mischaracterize virtual hosting as DNS inconsistency. To mitigate this effect, for each resulting IP address we perform an additional active HTTPS connection using SNI, specifying the name originally queried. We annotate all responses with this information, which we use for answer classification (examined further in §5.1).

3.5 Identifying DNS Manipulation

To determine whether a DNS response is manipulated, Iris relies on two types of metrics: *consistency* metrics

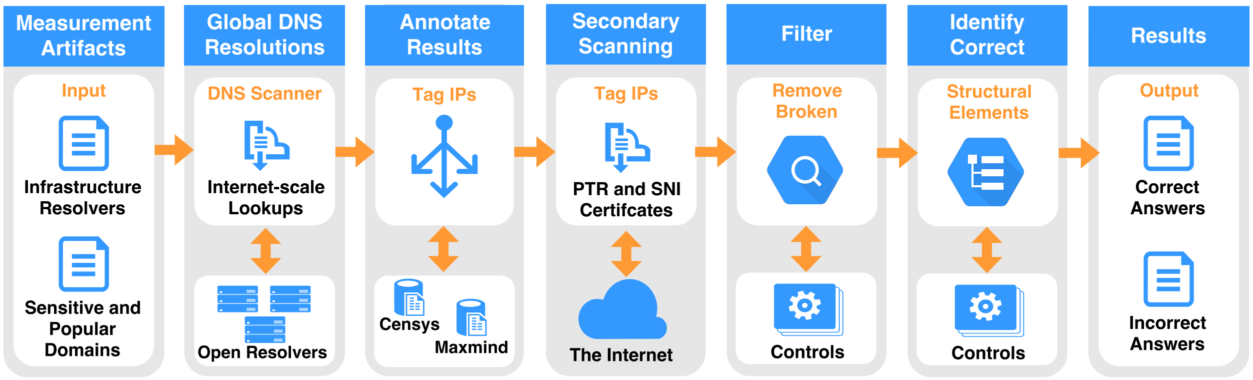

Figure 2: *Overview of DNS resolution, annotation, filtering, and classification. Iris inputs a set of domains and DNS resolvers and outputs results indicating manipulated DNS responses.*

and *independent verifiability* metrics. We say that a response is correct if it satisfies *any* consistency or independent verifiable metric; otherwise, we classify the response as *manipulated*. In this section, we outline each class of metrics as well as the specific features we develop to classify answers. The rest of this section defines these metrics; §5.1 explores the efficacy of each of them.

3.5.1 Consistency

Access to a domain should have some form of *consistency*, even when accessed from various global vantage points. This consistency may take the form of network properties, infrastructure attributes, or even content. We leverage these attributes, both in relation to control data as well as across the dataset itself, to classify DNS responses.

Consistency Baseline: Control Domains and Resolvers. Central to our notion of consistency is having a set of geographically diverse resolvers we control that are (presumably) not subject to manipulation. These controls give us a set of high-confidence correct answers we can use to identify consistency across a range of IP address properties. Geographic diversity helps ensure that area-specific deployments do not cause false-positives. For example, several domains in our dataset use different content distribution network (CDN) hosting infrastructure outside North America. As part of our measurements we insert domain names we control, with known correct answers. We use these domains to ensure a resolver reliably returns unmanipulated results for non-sensitive content (e.g., not a captive portal).

For each domain name, we create a set of consistency metrics by taking the union of each metric across all of our control resolvers. For example, if Control A returns the answer `192.168.0.10` and `192.168.0.11` and Control B returns `192.168.0.12`, we create a set of consistent IP set of (`192.168.0.10, 192.168.0.11, 192.168.0.12`). We say the answer is "correct" (i.e., not manipulated) if, for each metric, the answer is a non-empty subset of the controls. Returning to our IP example, if a global resolver returns the answer (`192.168.0.10, 192.168.0.12`), it is identified as correct. When a request returns multiple records, we check all records and consider the reply good if any response passes the appropriate tests.

Additionally, unmanipulated passive DNS [6] data collected simultaneously with our experiments across a geographically diverse set of countries could enhance (or replace) our consistency metrics. Unfortunately we are not aware of such a dataset being available publicly.

IP Address. The simplest consistency metric is the IP address or IP addresses that a DNS response contains.

Autonomous System / Organization. In the case of geographically distributed sites and services, such as those hosted on CDNs, a single domain name may return different IP addresses as part of normal operation. To attempt to account for these discrepancies, we also check whether different IP addresses for a domain map to the same AS we see when issuing queries for the domain name through our control resolvers. Because a single AS may have multiple AS numbers (ASNs), we consider two IP addresses with either the same ASN or AS *organization name* as being from the same AS. Although many responses will exhibit AS consistency even if individual IP addresses differ, even domains whose queries are not manipulated will sometimes return inconsistent AS-level and organizational information as well. This inconsistency is especially common for large service providers whose infrastructure spans multiple regions and continents and is often the result of acquisitions. To account for these inconsistencies, we need additional consistency metrics at higher layers of the protocol stack (specifically HTTP and HTTPS), described next.

HTTP Content. If an IP address is running a webserver on port 80, we include a hash of the content returned as an additional consistency metric. These content hashes come from a port 80 IP address Censys crawl. This metric effectively identifies sites with limited dynamic content. As discussed in §5.1, this metric is also useful in identifying sites with dynamic content but shared infrastructure. For example, as these hashes are based on `HTTP GET` fetches using an IP address as the `Host` in the header, this fetch uniquely fingerprints and categorizes CDN failures or default host pages. In another example, much of Google's web hosting infrastructure will return the byte-wise identical redirection page to `http://www.google.com/` for `HTTP GET`s without a valid Google host header. These identical pages allow us to identify Google resolutions as correct even for IP addresses acting as a Point-of-Presence.

HTTPS Certificate. We label a response as correct if the hash of the `HTTPS` certificate presented upon connection matches that of an IP returned via our controls. Note this is not an independent verifiability metric, as the certificates may or may not be trusted, and may not even be correct for the domain.

PTRs for CDNs. From our control data, we classify domains as hosted on particular CDNs based on `PTR`, AS, and certificate information. We consider a non-control response as consistent if the `PTR` record for that response points to the same CDN.

3.5.2 Independent Verifiability

In addition to consistency metrics, we also define a set of metrics that we can independently verify using external data sources, such as the HTTPS certificate infrastructure. We describe these methods below.

HTTPS Certificate. We consider a DNS response to be correct, independent of controls, if the IP address presents a valid, browser-trusted certificate for the correct domain name when queried without SNI. We further extend this metric to allow for common configuration errors, such as returning certificates for `*.example.com` when requesting `example.com`.

HTTPS Certificate with SNI. We add an additional metric that checks whether the certificate returned from our follow-up SNI-enabled scans returns a valid, browser-trusted certificate for the correct IP address.

3.6 Limitations

To facilitate global coverage in our measurements, our method has limitations that impact our scope and limit our results.

Localized Manipulation. Since Iris relies entirely on open infrastructure resolvers that we do not control, in regions with few resolvers, we cannot differentiate between localized manipulation by the resolver's operator and ISP or country-wide manipulation. Analysis of incorrect results focusing on consistency across ISP or country, or examination of webpage content, could aid in identifying localized manipulation.

Domain Bias. From our set of infrastructure resolvers, we measure manipulation of the CLBL and a subset of Alexa top sites. Although the CLBL is a community-based effort to identify sensitive content globally, by its very nature it is not *complete*. URLs and domains are missing, and sensitive content may change faster than the list is updated. Similarly, the list may exhibit geographic *bias* based on the language of the project and who contributes to it. This bias could affect the relative volume and scope of manipulation that Iris can detect.

Evasion. Although we focus on manipulation at ISP or country scale, an active adversary can still attempt to evade our measurements. Upstream resolvers could use EDNS Client Subnet [16] to only manipulate results for certain target IP ranges, or ISP resolvers could choose to manipulate only their own customers. Country-wide firewalls that perform injection could identify our scanning IP addresses and either not inject results or block our communication entirely. An adversary could also exploit our consistency metrics and inject incorrect IP addresses within the same AS as the targets.

Geolocation Error. We rely on Censys [21] and Maxmind [37] for geolocation and AS labeling of infrastructure resolvers to perform country or ISP-level aggregation. Incorrect labeling would identify country-wide manipulation as incomplete (false negatives), or identify manipulation in countries where it is not present (false positives).

4 Dataset

In this section, we characterize the data collected and how we processed it to obtain the results used in our analysis.

4.1 Open Resolver Selection

We initially identified a large pool of open DNS resolvers through an Internet-wide ZMap scan using our DNS extension to ZMap in January 2017. In total, 4.2 million open resolvers responded with a correct answer to our scan queries. This number excludes resolvers that replied with valid DNS responses but had either a missing or incorrect IP resolution for our scan's query domain.

Resolver Datasets	Total Resolvers	Number Countries	Median / Country
All Usable	4,197,543	232	659.5
Ethically Usable	6,564	157	6.0
Experiment Set	6,020	151	6.0

Table 1: *DNS resolver datasets. We identify all correctly functioning open resolvers are across the IPv4 address space. The experiment set consists of resolvers that passed additional functional tests beyond our basic scan. Note that the number of countries includes dependent territories.*

Resolver Dataset	AF	AS	EU	NA	OC	SA
All Usable	55	49	52	41	21	14
Ethically Usable	29	42	42	25	8	11
Experiment Set	26	41	41	24	8	11

Table 2: *Number of countries (and dependent territories) containing usable resolvers by continent. AF=Africa, AS=Asia, EU=Europe, NA=North America, OC=Oceana/Australia, SA=South America.*

The degree to which we can investigate DNS manipulation across various countries depends on the geographic distribution of the selected DNS resolvers. By geolocating this initial set of resolvers using Censys [21] and MaxMind [37], we observed that these resolvers reside in 232 countries and dependent territories[2], with a median of 659 resolvers per country. Due to the ethical considerations we outlined in §3.2, we restrict this set of resolvers to 6,564 infrastructure resolvers, in 157 countries, again with a median of 6 resolvers per country. Finally, we remove unstable or otherwise anomalous resolvers; §4.3 describes this process in more detail. This filtering reduces the set of usable resolvers to 6,020 in 151 countries, with a median of 6 resolvers in each. Table 1 summarizes the resulting population of resolvers; Table 2 shows the breakdown across continents. We also use 4 geographically diverse resolvers for controlled experiments; the 2 Google Public DNS servers [28], a German open resolver hosted on Amazon AWS, and a resolver that we manage at the University of California, Berkeley.

4.2 Domain Selection

We investigate DNS manipulation for both domains known to be censored and domains for popular websites. We began with the Citizen Lab Block List (CLBL) [14], consisting of 1,376 sensitive domains. We augment this list with 1,000 domains randomly selected from the Alexa Top 10,000, as well as 3 control domains we manage that should not be manipulated. Due to overlap between the two domain sets, our combined dataset consists of 2,330 domains. We excluded 27 problematic domains that we identified through our data collection process, resulting in our final population of 2,303 domains.

Response Datasets	Total Responses	Number Resolvers	Number Domains
All Responses	14,539,198	6,564	2,330
After Filtering	13,594,683	6,020	2,303

Table 3: *DNS response dataset before and after filtering problematic resolvers, domains, and failed queries.*

4.3 Response Filtering

We issued 14.5 million DNS `A` record queries for our 2,330 pre-filtered domains, across 6,564 infrastructure and control open resolvers during a 2 day period in January 2017. We observed various erroneous behavior that required further filtering. Excluding these degenerate cases reduced our dataset collection to 13.5 million responses across 2,303 domains and 6,020 resolvers, as summarized in Table 3. The rest of this section details this filtering process.

Resolvers. We detected that 341 resolvers stopped responding to our queries during our experiment. An additional 202 resolvers incorrectly resolved our control domain names, despite previously answering correctly during our Internet-wide scans. The common cause of this behavior was rate limiting, as our Internet-wide scans queried resolvers only once, whereas our experiments necessitated repeated queries. We identified another problematic resolver that exhibited a query failure rate above 70% due to aggressive rate limiting. We eliminated these resolvers and their associated query responses from our dataset, reducing the number of valid responses by 510K.

Domains. Our control DNS resolvers could not resolve 15 domain names. We excluded these and their associated 90K query responses from our dataset. We removed another 12 domains and their 72K corresponding query responses as their DNS resolutions failed an automated sanity check; resolvers across numerous countries provided the same incorrect DNS resolution for each of these domains, and the IP address returned was unique per domain (i.e., not a block page or filtering appliance). We did not expect censors to exhibit this behavior; a single censor is not likely to operate across multiple countries or geographic regions, and manipulations such as block pages that use a single IP address across countries should also be spread across multiple domains. These domains do not support HTTPS, and exhibit geograph-

[2] Countries and dependent territories are defined by the ISO 3166-1 alpha-2 codes, the granularity of Maxmind's country geolocation.

ically specific deployments. With increased geographic diversity of control resolvers or deployment of HTTPS by these sites, our consistency or verifiability metrics would account for these domains.

Queries. We filtered another 256K queries that returned failure error codes; 93.7% of all errors were timeouts and server failures. Timeouts denote connections where the resolver did not respond to our query within 15 seconds. Server failures indicate when a resolver could not recursively resolve a domain within its own pre-configured time allotment (10 seconds by default in BIND). Table 4 provides a detailed breakdown of error responses.

Failure Type	Count	% of Responses
Timeout	140,551	0.97%
Server Fail	107,826	0.74%
Conn Refused	7,823	0.05%
Conn Error	3,686	0.03%
Truncated	3,451	0.02%
NXDOMAIN	1,713	0.01%

Table 4: *Breakdown of the 265,050 DNS responses that returned a non-success error code.*

Returning an NXDOMAIN response code [38], which informs a client that a domain does not exist, is an obvious potential DNS censorship mechanism. Unfortunately, some CDNs return this error in normal operations, presumably due to rate limiting or client configuration settings. We found that the most prevalent NX behavior occurred in the countries of Tonga and Pakistan; both countries exhibited censorship of multiple content types, including adult and LGBT. Previous studies have observed NXDOMAIN blocking in Pakistan [38]. These instances comprise a small percentage of overall NXDOMAIN responses. Given the many non-censorship NXDOMAIN responses and the relative infrequency of their use for censorship, we exclude these from our analysis. Another 72K responses had a `SUCCESS` response code, but contained no IP address in the response. This failure mode frequently coincide with `CNAME` responses that could not be resolved further. We excluded these queries. Table 5 provides a geographic breakdown of NXDOMAIN responses.

After removing problematic resolvers, domains, and failed queries, the dataset comprises of 13,594,683 DNS responses. By applying our *consistency* and *independent verifiability* metrics, we identify 41,778 responses (0.31%) as manipulated, spread across 58 countries (and dependent territories) and 1,408 domains.

Country	% NXDOMAIN
Tonga	2.93%
Pakistan	0.37%
Bosnia/Herzegovina	0.12%
Isle of Man	0.04%
Cape Verde	0.04%

Table 5: *The top 5 countries / dependent territories by the percent of queries that responded with NXDOMAIN.*

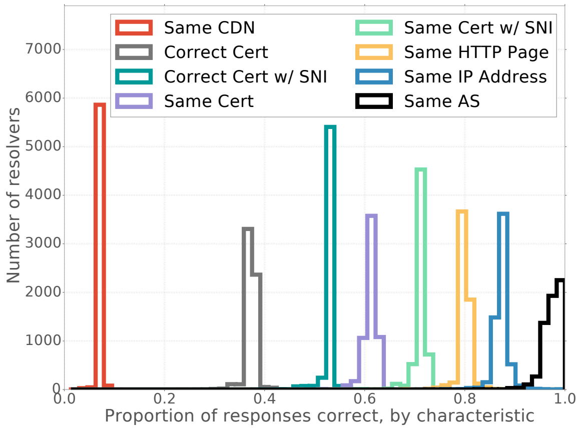

Figure 3: *The ability of each correctness metric to classify responses as correct. Table is ordered (top to bottom, left to right) by the lines on the graph (left to right).*

5 Results

We now evaluate the effectiveness of our DNS manipulation metrics and explore manipulated DNS responses in the context of Internet censorship.

5.1 Evaluating Manipulation Metrics

To assess the effectiveness of the *consistency* and *independent verifiability* metrics, we quantify the ability of each metric to identify unmanipulated responses (to exclude from further investigation). Figure 3 shows each metric's efficacy. The horizontal axis represents the fraction of responses from a particular resolver that are classified as correct by a given metric. The vertical axis indicates the number of resolvers that exhibit that same fraction of correct responses (again under the given metric). For example, almost 6,000 resolvers had roughly 8% of their responses identified as correct under the "Same CDN" metric. A narrow band indicates that many resolvers exhibit similar fractions of correct responses under that metric (i.e., it is more *stable*). The closer the center mass of a histogram lies to 1.0, the more *effective* its corresponding metric, since a larger fraction of responses are classified as correct (i.e., not manipulation) using that metric.

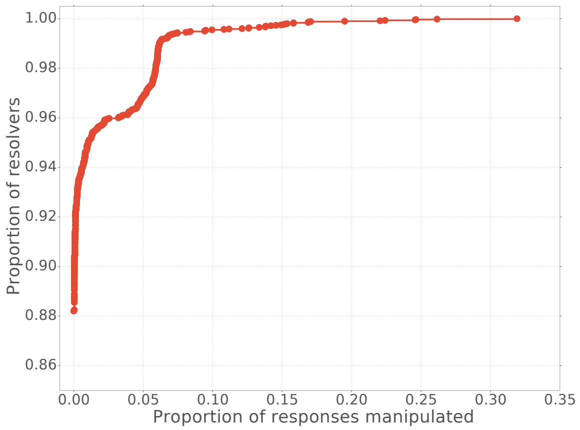

Figure 4: *The fraction of responses manipulated, per resolver. For 89% of resolvers, we observed no manipulation.*

The AS consistency metric ("Same AS") is the most effective: it classified 90% of the DNS responses as consistent. Similarly, identifying matching IP addresses between responses from our control resolvers and our experiment resolvers flagged about 80% of responses as correct across most resolvers. "Same HTTP Page" is also relatively effective, as many geographically distributed deployments of the same site (such as with Points-of-Presence) have either identical content or infrastructure error characteristics (see §3.5.1). This figure also illustrates the importance of SNI, increasing the effectiveness of correct and valid HTTPS certificates from 38% to 55%. The same HTTPS certificate ("Same Cert") metric turns out to be more effective than simply having a correct certificate ("Correct Cert"), because so many sites incorrectly deploy HTTPS.

5.2 Manipulated DNS Responses

We detect nearly 42,000 manipulated DNS responses; we now investigate the distribution of these responses across resolvers, domains, and countries.

Manipulated responses by resolver. Figure 4 shows the cumulative fraction of results that return at least a certain fraction of manipulated responses: 88% of resolvers exhibited no manipulation; for 96% of resolvers, we observe manipulation for fewer than 5% of responses. The modes in the CDF highlight differences between resolver subpopulations, which upon further investigation we discovered reflected differing manipulation practices across countries. Additionally, 62% of domains are manipulated by at least one resolver, which is expected given that more than half of our selected domains are sensitive sites on the CLBL. We explore these variations in more detail later in this section.

Country (# Res.)	Median	Mean	Max	Min
Iran (122)	6.02%	5.99%	22.41%	0.00%
China (62)	5.22%	4.59%	8.40%	0.00%
Indonesia (80)	0.63%	2.81%	9.95%	0.00%
Greece (26)	0.28%	0.40%	0.83%	0.00%
Mongolia (6)	0.17%	0.18%	0.36%	0.00%
Iraq (7)	0.09%	1.67%	5.79%	0.00%
Bermuda (2)	0.04%	0.04%	0.09%	0.00%
Kazakhstan (14)	0.04%	0.30%	3.90%	0.00%
Belarus (18)	0.04%	0.07%	0.30%	0.00%

Table 6: *Top 10 countries by median percent of manipulated responses per resolver. We additionally provide the mean, maximum, and minimum percent for resolvers in each country. The number of resolvers per country is listed with the country name.*

Manipulated responses by country. Previous work has observed that some countries deploy nation-wide DNS censorship technology [5]; therefore, we expected to see groups of resolvers in the same country, each manipulating a similar set of domains. Table 6 lists the percent of manipulated responses per resolver, aggregated across resolvers in each country. Resolvers in Iran exhibited the highest degree of manipulation, with a median of 6.02% manipulated responses per Iranian resolver; China follows with a median value of 5.22%. These rankings depend on the domains in our domain list, and may merely reflect that the CLBL contained more domains that are censored in these countries.

The top 10 countries shown in Table 6 all have at least one resolver that does not manipulate *any* domains; IP address geolocation inaccuracy may partially explain this surprising finding. For example, uncensored resolvers in Hong Kong may be incorrectly labeled as Chinese. Additionally, for countries that do not directly implement the technical manipulation mechanisms but rather rely on individual ISPs to do so, the actual manifestation of manipulation may vary across ISPs within a single country. Localized manipulation by resolver operators in countries with few resolvers could also influence these results. §5.3 investigates these factors further.

Figure 5 shows the representation of responses in our dataset by country. For example, the leftmost pair of bars shows that, while less than 5% of all responses in our dataset came from Iranian resolvers, the responses that we received accounted for nearly 40% of manipulated responses in the dataset. Similarly, Chinese resolvers represented 1% of responses in the data but contributed to 15% of the manipulated responses. In contrast, 30% of our DNS responses came from resolvers in the United States, but accounted for only 5% of censored responses.

Table 7 shows the breakdown of the top manipulated

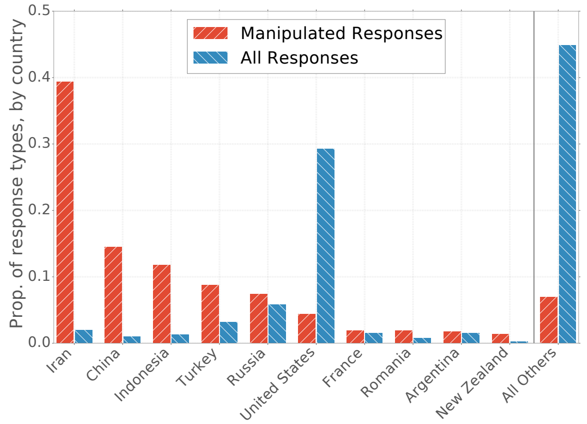

Figure 5: *The fraction of all responses in our dataset from each country (blue), and the fraction of all manipulated responses in our dataset from the corresponding country (red).*

Answer	Results	Names	Category
10.10.34.36	12,144	140	Private
10.10.34.34	4,566	776	Private
146.112.61.106	3,495	801	OpenDNS Adult
195.175.254.2	3,137	129	HTTP Error Page
93.46.8.89	1,571	88	China*
118.97.116.27	1,212	155	Safe / Filtering
243.185.187.39	1,167	88	China*
127.0.0.1	876	267	Private
95.53.248.254	566	566	Resolver's Own IP
95.53.248.254	565	565	Resolver's Own IP
8.7.198.45	411	75	China*
202.169.44.80	379	113	Safe / Filtering
212.47.252.200	371	371	Resolver's Own IP
212.47.254.200	370	370	Resolver's Own IP
213.177.28.90	352	22	Gambling Blockpg
208.91.112.55	349	320	Blockpg
180.131.146.7	312	145	Safe / Filtering
203.98.7.65	303	78	China*
202.182.48.245	302	100	Adult Blockpg
93.158.134.250	258	86	Safe / Filtering

Table 7: *Most common manipulated responses by volume, with manual classification for public, non-resolver IP addresses. The category "China*" are IP addresses previously observed by Farnan et al. in 2016 [25].*

responses, by the IP address that appears in the manipulated answer. The top two special-purpose (i.e., private) IP addresses appear in the majority of responses within Iran. The third most common response is an OpenDNS (a DNS filtering and security product [13]) blockpage indicating adult content. The fourth most frequent response is an IP address hosting an HTTP error page known to be used in Turkey DNS manipulation [11].

Private and special-purpose IPv4 addresses in manipulated DNS responses. Of the roughly 42,000 manipulated DNS responses, 17,806 correspond to special-purpose IPv4 addresses as defined by RFC 6890 [17]; the remaining 23,972 responses corresponded to addresses in the public IP address space. Table 8 shows the extent to which countries return private IP addresses in responses, for the top 10 countries ranked by the relative amount of DNS manipulation compared to the total number of results from that country. For example, we observed more manipulated responses from Turkey than Iraq, but Iris used more open DNS resolvers in Turkey, so observed frequencies require normalization. Here, we notice that countries that manipulate DNS tend to either return only special-purpose IP addresses in manipulated responses (as in the case of Iran, Iraq, and Kuwait) or only public IP addresses (China).

Figure 6 presents the distribution of observed public IP addresses across manipulated responses in our dataset. The most frequently returned public IP address, an OpenDNS blockpage, constituted almost 15% of all manipulated responses containing public IP addresses. The top ten public IP addresses accounted for nearly 60% of responses.

Many IP answers have been observed in previous studies on Chinese DNS censorship [5, 25]. These addresses are seemingly arbitrary; they host no services, not even a fundamental webpage. The 10 most frequent Chinese responses constituted almost 75% of Chinese responses. The remaining 25% are spread over a long tail of nearly 1,000 seemingly arbitrary non-Chinese IP addresses.

5.3 Manipulation Within Countries

Figure 7 shows the DNS manipulation of each domain by the fraction of resolvers *within* a country, for the 10 countries with the most normalized amount of manipulation. Each point represents a domain; the vertical axis represents the fraction of resolvers in that country that manipulate it. Shading shows the density of points for that part of the distribution. The plot reveals several interesting phenomena. One group of domains is manipulated by about 80% of resolvers in Iran, and another group is manipulated by fewer than 10% of resolvers. This second group of domains is manipulated by a smaller fraction of resolvers, also returning non-public IP addresses. These effects are consistent with previously noted blackholing employed by DNS manipulation infrastructure [7]; this phenomenon deserves further investigation.

Similarly, one set of domains in China experiences manipulation by approximately 80% of resolvers, and another set experiences manipulation only half the time. In contrast, manipulation in Greece and Kuwait is more homogeneous across resolvers.

Country (# Res.)	% Incor.	% Pub.
Iran (122)	6.02%	0.01%
China (62)	4.52%	99.46%
Indonesia (80)	2.74%	95.08%
Iraq (7)	1.68%	1.49%
New Zealand (16)	1.59%	100.00%
Turkey (192)	0.84%	99.81%
Romania (45)	0.77%	100.00%
Kuwait (10)	0.61%	0.00%
Greece (26)	0.41%	100.00%
Cyprus (5)	0.40%	100.00%

Table 8: *Percent of public IP addresses in manipulated responses, by country. Countries are sorted by overall frequency of manipulation.*

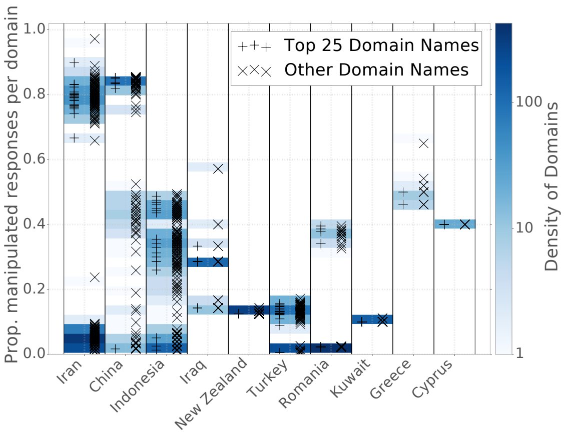

Figure 7: *The fraction of resolvers within a country that manipulate each domain.*

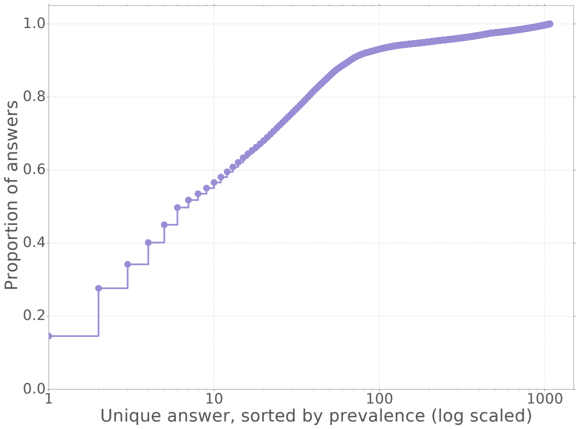

Figure 6: *Manipulated but public IP addresses in our dataset. The horizontal axis is sorted by the most common IP.*

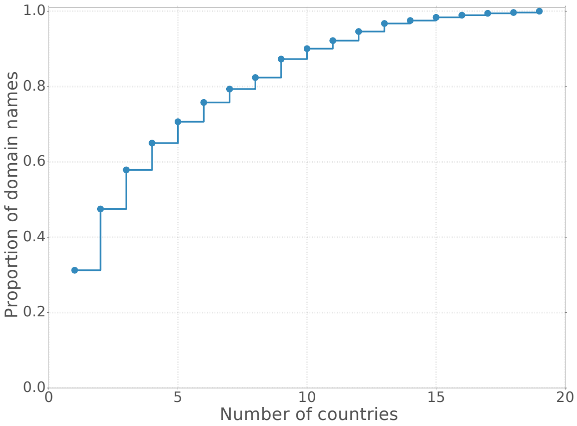

Figure 8: *The number of countries (or dependent territories) that block each domain with observed manipulated responses, sorted by manipulation prevalence.*

Heterogeneity across a country may suggest a situation where different ISPs implement filtering with different block lists; it might also indicate variability across geographic region within a country. The fact that manipulation rates vary even among resolvers in a certain group within a country may indicate either probabilistic manipulation, or the injection of manipulated responses (a phenomenon that has been documented before [5]). Other more benign explanations exist, such as corporate firewalls (which are common in the United States), or localized manipulation by resolver operators.

Ceilings on the percent of resolvers within a country performing manipulation, such as no domain in China experiencing manipulation across more than approximately 85% of resolvers, suggest IP geolocation errors are common.

5.4 Commonly Manipulated Domains

Commonly manipulated domains across countries. Many domains experienced manipulation across a range of countries. Figure 8 shows a CDF of the number of countries (or dependent territories) for which at least one resolver manipulated each domain. 30% of domains were manipulated in only a single country, while 70% were manipulated in 5 or fewer countries. No domain was manipulated in more than 19 countries.

Table 9 highlights domains that experience manipulation in many countries (or dependent territories). The 2 most manipulated domains are both gambling websites, each experiencing censorship across 19 different countries. DNS resolutions for pornographic websites are similarly manipulated, accounting for the next 3 most commonly affected domains. Peer-to-peer file sharing

Rank	Domain Name	Category	# Cn	# Res
1	*pokerstars.com	Gambling	19	251
2	betway.com	Gambling	19	234
3	pornhub.com	Pornography	19	222
4	youporn.com	Pornography	19	192
5	xvideos.com	Pornography	19	174
6	thepiratebay.org	P2P sharing	18	236
7	thepiratebay.se	P2P sharing	18	217
8	xhamster.com	Pornography	18	200
9	*partypoker.com	Gambling	17	226
10	beeg.com	Pornography	17	183
80	torproject.org	Anon. & cen.	12	159
181	twitter.com	Twitter	9	160
250	*youtube.com	Google	8	165
495	*citizenlab.org	Freedom expr.	4	148
606	www.google.com	Google	3	56
1086	google.com	Google	1	5

Table 9: *Domain names manipulated in the most countries (or dependent territories), ordered by number of countries with manipulated responses. Domains beginning with * begin with "www.".*

Rank	Domain Category	# Cn.	# Resolv.
1	Alexa Top 10k	36	442
2	Freedom of expr.	35	384
3	P2P file sharing	34	394
4	Human rights	31	288
5	Gambling	29	377
6	Pornography	29	342
7	Alcohol and drugs	28	274
8	Anon. & censor.	24	303
9	Hate speech	22	158
10	Multimedia sharing	21	293
20	Google	16	234
34	Facebook	10	175
38	Twitter	9	160

Table 10: *Top 10 domain categories, ordered by number of countries (or dependent territories) with manipulated answers.*

sites are also commonly targeted, particularly The Pirate Bay. The Tor Project [50] DNS domain is the most widely interfered with domain amongst anonymity and censorship tools, manipulated across 12 countries. Citizen Lab [15] also experienced manipulation across 4 countries. We note that `www.google.com` is impacted across more countries than `google.com`, unsurprising since all HTTP and HTTPS queries to `google.com` immediately redirect to `www.google.com`; for example, China manipulates `www.google.com` queries but disregards those for `google.com`. This result underscores the need for domain datasets that contain complete domains and subdomains, rather than simply second-level domains.

We also note that commonly measured sites such as The Tor Project, Google, and Twitter, experience manipulation across significantly fewer countries than some sites. Such disparity points to the need for a diverse domain dataset.

China focuses its DNS manipulation not just on adult content but also major English news outlets, such as `nytimes.com`, `online.wsj.com`, and `www.reuters.com`. China is the only country observed to manipulate the DNS responses for these domains; it also censored the Chinese language Wikipedia domain.

Commonly manipulated categories. Table 10 shows the prevalence of manipulation by CLBL categories. We consider a category as manipulated within a country if any resolver within that country manipulates a domain of that category. Domains in the Alexa Top 10K experienced the most manipulation; these domains did not appear in the CLBL, which highlights the importance of measuring both curated lists from domain experts as well as broad samples of popular websites. Although no single domain experiences manipulation in more than 19 countries, several categories experience manipulation in more than 30 countries, indicating that while broad categories appear to be commonly targeted, the specific domains may vary country to country.

To study how manipulated categories vary across countries, we analyzed the fraction of resolvers within each country that manipulate a particular category. The top categories vary extensively across countries. Table 11 shows the most frequently manipulated categories for the top 10 countries by normalized amounts of manipulation. The top category of manipulated content in Iran, "provocative attire," is not a category across any of the other top 10 countries. Manipulation of domains randomly selected from Alexa but not in the CLBL ("Alexa Top 10k") is prevalent across numerous countries, again reinforcing the need for diverse domain datasets. Anonymity and censorship tools are manipulated extensively across 85% of resolvers in China, but not across the rest of the top 10. Pornography and gambling sites are manipulated throughout.

6 Summary

Internet censorship is widespread, dynamic, and continually evolving; understanding the nature of censorship thus requires techniques to perform continuous, large-scale measurement. Unfortunately, the state-of-the-art techniques for measuring manipulation—a common censorship technique—rely on human volunteers, limiting the scale and frequency of measurements. This work introduces a method for measuring DNS manipulation on

Country	Domain Category	% of Resolv.
IR	Provocative attire	90.98%
	Alexa Top 10k	90.16%
	Freedom of expr.	90.16%
CN	Alexa Top 10k	85.48%
	Freedom of expr.	85.48%
	Anon. & censor.	85.48%
ID	Pornography	57.50%
	Alexa Top 10k	56.25%
	P2P file sharing	52.50%
IQ	Political Blog	57.14%
	Alexa Top 10k	28.57%
	Freedom of expr.	28.57%
NZ	Alexa Top 10k	12.50%
	Freedom of expr.	12.50%
	P2P file sharing	12.50%
TR	Alexa Top 10k	18.23%
	Freedom of expr.	17.71%
	Pornography	16.67%
RO	Alexa Top 10k	37.78%
	Gambling	37.78%
	Freedom of expr.	2.22%
KW	Alexa Top 10k	10.00%
	Freedom of expr.	10.00%
	P2P file sharing	10.00%
GR	Gambling	50.00%
	Alexa Top 10k	46.15%
CY	Alexa Top 10k	40.00%
	Gambling	40.00%

Table 11: *Breakdown of the top 3 domain categories experiencing manipulation, per country. Countries are ordered by the relative amount of manipulated responses for that country. Both Greece (GR) and Cyprus (CY) only experience manipulated responses across 2 categories.*

a global scale by using as vantage points open DNS resolvers that form part of the Internet's infrastructure.

The major contributions of our work are: (1) Iris: a scalable, ethical system for measuring DNS manipulation; (2) an analysis technique for disambiguating natural variation in DNS responses (e.g., due to CDNs) from more nefarious types of manipulation; and (3) a large-scale measurement study that highlights the heterogeneity of DNS manipulation, across countries, resolvers, and domains. Notably, we find that manipulation is heterogeneous across DNS resolvers even within a single country. Iris supports regular, continuous measurement, which will ultimately facilitate tracking DNS manipulation trends as they evolve over time; our next step is to operationalize such measurements to facilitate longitudinal analysis.

Acknowledgements

The authors are grateful for the assistance and support of Manos Antonakakis, Randy Bush, Jed Crandall, Zakir Durumeric, and David Fifield. This work was supported in part by National Science Foundation Awards CNS-1237265, CNS-1406041, CNS-1518878, CNS-1518918 CNS-1540066 and CNS-1602399.

References

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

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

[3] C. Anderson, P. Winter, and Roya. Global Network Interference Detection Over the RIPE Atlas Network. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2014.

[4] Anonymous. The Collateral Damage of Internet Censorship by DNS Injection. *SIGCOMM Computer Communication Review*, 42(3):21–27, June 2012.

[5] Anonymous. Towards a Comprehensive Picture of the Great Firewall's DNS Censorship. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2014.

[6] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster. Building a Dynamic Reputation System for DNS. In *USENIX Security Symposium*, 2010.

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

[8] M. Bailey and C. Labovitz. Censorship and Cooption of the Internet Infrastructure. Technical Report CSE-TR-572-11, University of Michigan, Ann Arbor, MI, USA, July 2011.

[9] BBC. BBC's Website is being Blocked across China. `http://www.bbc.com/news/world-asia-china-29628356`, October 2014.

[10] The Belmont Report - Ethical Principles and Guidelines for the Protection of Human Subjects of Research. `http://ohsr.od.nih.gov/guidelines/belmont.html`.

[11] S. Bortzmeyer. Hijacking through routing in turkey. https://ripe68.ripe.net/presentations/158-bortzmeyer-google-dns-turkey.pdf.

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

[13] Cisco OpenDNS. https://www.opendns.com/.

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

[15] Citizen Lab. https://citizenlab.org.

[16] C. Contavalli, W. van der Gaast, D. C. Lawrence, and W. Kumari. Client Subnet in DNS Queries. RFC 7871.

[17] M. Cotton, L. Vegoda, R. Bonica, and B. Haberman. Special-Purpose IP Address Registries. RFC 6890.

[18] D. Dagon, N. Provos, C. P. Lee, and W. Lee. Corrupted DNS Resolution Paths: The Rise of a Malicious Resolution Authority. In *Network & Distributed System Security Symposium (NDSS)*, 2008.

[19] 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 *ACM Internet Measurement Conference (IMC)*, 2013.

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

[21] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. A Search Engine Backed by Internet-Wide Scanning. In *ACM Conference on Computer and Communications Security (CCS)*, 2015.

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

[23] R. Ensafi, J. Knockel, G. Alexander, and J. R. Crandall. Detecting Intentional Packet Drops on the Internet via TCP/IP Side Channels. In *Passive and Active Measurements Conference (PAM)*, 2014.

[24] R. Ensafi, P. Winter, A. Mueen, and J. R. Crandall. Analyzing the Great Firewall of China Over Space and Time. *Privacy Enhancing Technologies Symposium (PETS)*, 1(1), 2015.

[25] O. Farnan, A. Darer, and J. Wright. Poisoning the Well – Exploring the Great Firewall's Poisoned DNS Responses. In *ACM Workshop on Privacy in the Electronic Society (WPES)*, 2016.

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

[27] The Go Programming Language. https://golang.org/.

[28] Google Public DNS. https://developers.google.com/speed/public-dns/.

[29] F. House. Freedom on the Net. 2016.

[30] ICLab. ICLab: a Censorship Measurement Platform. https://iclab.org/.

[31] J. Jiang, J. Liang, K. Li, J. Li, H. Duan, and J. Wu. Ghost Domain Name: Revoked yet Still Resolvable. In *Network & Distributed System Security Symposium (NDSS)*, 2012.

[32] B. Jones, N. Feamster, V. Paxson, N. Weaver, and M. Allman. Detecting DNS Root Manipulation. In *Passive and Active Measurement (PAM)*, 2016.

[33] B. Jones, T.-W. Lee, N. Feamster, and P. Gill. Automated Detection and Fingerprinting of Censorship Block Pages. In *ACM Internet Measurement Conference (IMC)*, 2014.

[34] M. Kührer, T. Hupperich, J. Bushart, C. Rossow, and T. Holz. Going Wild: Large-Scale Classification of Open DNS Resolvers. In *ACM Internet Measurement Conference (IMC)*, 2015.

[35] M. Kührer, T. Hupperich, C. Rossow, and T. Holz. Exit from Hell? Reducing the Impact of Amplification DDoS Attacks. In *USENIX Security Symposium*, 2014.

[36] G. Lowe, P. Winters, and M. L. Marcus. The Great DNS Wall of China. Technical report, New York University, 2007.

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

[38] Z. Nabi. The Anatomy of Web Censorship in Pakistan. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2013.

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

[40] Open Resolver Project. http://openresolverproject.org/.

[41] 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 *IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2010.

[42] P. Pearce, R. Ensafi, F. Li, N. Feamster, and V. Paxson. Augur: Internet-Wide Detection of Connectivity Disruptions. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

[43] A. Razaghpanah, A. Li, A. Filastò, R. Nithyanand, V. Ververis, W. Scott, and P. Gill. Exploring the Design Space of Longitudinal Censorship Measurement Platforms. Technical Report 1606.01979, ArXiv CoRR, 2016.

[44] M. Salganik. Bit by Bit: Social Research for the Digital Age, 2016. http://www.bitbybitbook.com/.

[45] Sam Burnett and Nick Feamster. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *ACM SIGCOMM*, 2015.

[46] K. Schomp, T. Callahan, M. Rabinovich, and M. Allman. On Measuring the Client-Side DNS Infrastructure. In *ACM Internet Measurement Conference (IMC)*, 2013.

[47] W. Scott, T. Anderson, T. Kohno, and A. Krishnamurthy. Satellite: Joint Analysis of CDNs and Network-Level Interference. In *USENIX Annual Technical Conference (ATC)*, 2016.

[48] A. Sfakianakis, E. Athanasopoulos, and S. Ioannidis. CensMon: A Web Censorship Monitor. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2011.

[49] The Tor Project. OONI: Open observatory of network interference. https://ooni.torproject.org/.

[50] The Tor Project. https://www.torproject.org/.

[51] G. Tuysuz and I. Watson. Turkey Blocks YouTube Days after Twitter Crackdown. http://www.cnn.com/2014/03/27/world/europe/turkey-youtube-blocked/, Mar. 2014.

[52] N. Weaver, C. Kreibich, and V. Paxson. Redirecting DNS for Ads and Profit. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2011.

[53] P. Winter. The Philippines are blocking Tor? Tor Trac ticket, June 2012. https://bugs.torproject.org/6258.

[54] P. Winter and S. Lindskog. How the Great Firewall of China is Blocking Tor. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2012.

[55] X. Xu, Z. M. Mao, and J. A. Halderman. Internet Censorship in China: Where Does the Filtering Occur? In *Passive and Active Measurement Conference (PAM)*, 2011.