Many Roads Lead To Rome: How Packet Headers Influence DNS Censorship Measurement

Abhishek Bhaskar and Paul Pearce, *Georgia Institute of Technology*

https://www.usenix.org/conference/usenixsecurity22/presentation/bhaskar

This paper is included in the Proceedings of the 31st USENIX Security Symposium.

August 10–12, 2022 • Boston, MA, USA

978-1-939133-31-1

Open access to the Proceedings of the 31st USENIX Security Symposium is sponsored by USENIX.

Many Roads Lead To Rome: How Packet Headers Influence DNS Censorship Measurement

Abhishek Bhaskar Paul Pearce

Georgia Institute of Technology
{abhaskar, pearce}@gatech.edu

Abstract

Internet censorship is widespread, impacting citizens of hundreds of countries around the world. Recent work has developed techniques that can perform widespread, longitudinal measurements of global Internet manipulation remotely and have focused largely on the scale of censorship measurements with minimal focus on reproducibility and consistency.

In this work we explore the role packet headers (e.g., source IP address and source port) have on DNS censorship. By performing a large-scale measurement study building on the techniques deployed by previous and current censorship measurement platforms, we find that choice of ephemeral source port and local source IP address (e.g., x.x.x.7 vs x.x.x.8) influence routing, which in turn influences DNS censorship. We show that 37% of IPs across 56% ASes measured show some change in censorship behavior depending on source port and local source IP. This behavior is frequently *all-or-nothing*, where choice of header can result in no observable censorship. Such behavior mimics and could be misattributed to geolocation error, packet loss, or network outages. The scale of censorship differences can more than *double* depending on the lowest 3 bits of the source IP address, consistent with known router load balancing techniques. We also observe smaller-scale censorship variation where only a few domains experience censorship differences based on packet parameters. We lastly find that these variations are persistent; packet retries do not control for observed variation. Our results point to the need for methodological changes in future DNS censorship measurement, which we discuss.

1 Introduction

Internet censorship affects hundreds of countries around the world [24, 30, 31, 37]. Despite this prevalence, empirical Internet measurements revealing the scope and behavior of censorship remain comparatively nascent. Recent work has developed techniques that can perform widespread, longitudinal measurements of global Internet manipulation remotely, without requiring the participation of individual users in the countries of interest [30, 31, 37].

Remote measurement methods rely on the construction of measurement packets which are sent through network devices that perform manipulation and censorship. While prior work [24, 30, 31, 37] has focused largely on the scale of censorship measurements—how do we measure as many *things* (i.e., censored domains) as possible—relatively little work exists exploring the reproducibility and consistency of these measurements. More troubling, prior work has identified non-determinism [2, 31, 33, 39] in censorship results, which are typically attributed to issues of load in censorship devices or geographic variations, but without significant understanding of these phenomena.

Unrelated to censorship, Internet routing between endpoints is known to change both over time [1, 5] and based on the construction of the packet [4]. In order to evenly distribute load, routers will examine portions of the IP and TCP/UDP header and send packets on different paths based on those values [4]. Bits of source IP and source port are some of the fields known to cause routing changes [4, 12]. Despite this phenomenon being well known in the *routing* literature, we are unaware of any prior censorship measurement work that has accounted for such routing variations.

Our work explores the intersection of router load balancing based on packet headers (e.g., source IP address and port) and DNS censorship. Our key insight is that the choice of ephemeral source port and local source IP address *within a subnet* (e.g., x.x.x.7 vs x.x.x.8) results in Internet route differences which when combined with country-level Internet censorship apparatuses, change DNS censorship behavior. These routing variations are the result of load balancing done throughout the Internet orthogonal to censorship activities.

Using Chinese DNS censorship [19, 23, 25, 42] as a lens to understand this phenomena, we perform a large-scale measurement study building on the techniques deployed by previous and current censorship measurement platforms [31, 36, 37], augmented with our own measurement methodology to identify and control for packet headers and routing changes.

We find that changes to ephemeral source port and IP address do indeed influence routing, which in turn influences DNS censorship measurement results. We find that 37% of IPs across 56% ASes measured show some change in censorship behavior depending on source port and source IP. We also find that the most common form of variation is *all-or-nothing behavior*, where a source IP or source port either experiences *no* censorship or "expected" censorship activity. Such behavior mimics and could be misattributed to non-determinism,

load shedding, or geolocation errors. The extent of these censorship differences can more than *double* depending on the lowest 3 bits of the source IP address. We document and describe how this bit-level operation is consistent with known router load balancing techniques [11–13]. Further, we observe similar censorship changes based on the lowest 3 bits of the destination IP address. We lastly show smaller-scale censorship variation where only a few domains experience censorship differences. We find that all these variations are persistent; packet retries do not control for these variations.

Our results point to the need to carefully control for IP address and source port selection to ensure the correctness of censorship measurement studies, as well as the need to understand these phenomena when performing localized or longitudinal analysis of the results.

Our contributions include:

- Describing and exploring the impact router load balancing has on DNS censorship measurement.
- Developing BreadCrumb, a measurement methodology and associated tool to quantitatively understand the prevalence of censorship changes due to router load balancing across source port, source IP, and destination IP.
- Demonstrating the methods previous studies have used for DNS censorship measurement are subject to routing-induced censorship differences.
- Finding that router load balancing based on IP address patterns and source port results in observed censorship differences across 37% of measured IPs and 56% measured ASes. We further show these differences can more than double based on the lower 3-bits of source IP.
- Quantifying small-scale censorship measurement differences over a subset of newly censored domains and a subset of destination IPs.
- Providing guidance for future measurement studies to account for this phenomena.

2 Related Work

In the last decade there have been many studies focused on understanding censorship at different scales. Some focus on specific countries, such as China [42], Iran [3,8], Pakistan [26], Russia [34], India [22], Syria [10], and Egypt [6]. Others focus on large scale global measurement [20, 24, 31, 33, 37, 38]. These censorship studies answer various questions like: (1) what is censored? (2) how does the censorship work? (3) how does censorship change over a long time scale? None of these studies have explored the impact of router load balancing resulting from packet source parameters on censorship. *BreadCrumb* explores the consistency of these results and how they change in the face of router load balancing.

Remote Censorship Measurement. Since most censorship systems are deployed in the form of middle-boxes that intercept traffic and perform an action based on it, many studies utilize some form of remote (i.e., outside-in) measurement technique [3, 9, 18, 19, 23, 27, 28, 31, 33, 34, 37, 38, 40, 42] to understand censorship. These outside-in, remote, or external measurements typically involve sending a probe packet containing some content that will trigger censorship behavior towards a vantage within a censored country, and analyzing the results of the responses to understand censorship.

Remote censorship measurement usually takes place at one of the network stack levels: (1) DNS, where mainly manipulation at the DNS level is studied with some auxiliary information [19, 23, 27, 31, 40], (2) TCP, where some form of TCP connection is attempted at a vantage point with some sensitive contents [38]), or (3) some form of hybrid censorship [34, 37]. In any of these cases, the measurement relies on the stability of the path traversed by the packet to obtain consistent results. *BreadCrumb*'s exploration of how remote censorship results change with packet construction is useful in understanding and validating not only all of these studies, but also future studies that make use of these techniques.

Chinese Internet Censorship. There have been many studies that have focused on understanding censorship behavior in China directly [2, 19, 23, 25, 29, 42] or as part of a larger global-scale censorship study [24, 31, 37]. *BreadCrumb* is directly relevant to and builds on the understanding (and in some cases methodology [31]) of these studies, as much of the censorship explored does not account for potential changes based on packet construction.

Network Load Balancing. Router load balancing is prevalent across the Internet. Augustin et al. [5] explored the prevalence of load balancing in the Internet and found that close to 72% of paths traverse through a load balancing router (in 2011). Their earlier work Paris Traceroute [4], explores the contents of a packet that can influence routing. In addition, router documentation [11, 13] indicates the use of various packet parameters like source port, source IP and destination IP to make load balancing routing decisions. Therefore, any censorship measurement techniques that rely on the stability of a network path to obtain accurate results are potentially affected by the choice of input parameters used in the experiment. This is the core problem *BreadCrumb* explores.

Non-Determinism in Censorship Results. Previous work on censorship has observed unstable results or non-determinism that, while referenced, have not been explored in great detail [2,31,33,39,40]. Raman et al. [33] experienced infrequent absences in their results and they speculate this could be caused by routing changes in the path over time (not due to packet construction). Weinberg et al. [40] found varying levels of censorship depending on the origin network and destination of their probes and also speculated the variations could depend on route, but did not analyze further. Wang et al. [39] noticed that among a certain set of client and servers in their measurement, the `TCP RST` employed by the Great Firewall (GFW) was not successful in tearing down the connection

which they presume is caused by load balancing. Pearce et al. [31] also identified variation in censorship within a country. Anonymous [2] observed that changing the destination and port of their injecting packet resulted in different paths resulting in different injecting interfaces, but did not explore the phenomena. *BreadCrumb* presents an opportunity to connect a common thread across the anecdotes and speculated causes of censorship non-determinism, providing specific measurements, understanding, and cause of the phenomena, as well as recommendations for future studies.

3 Method

In this section we describe *BreadCrumb*, our methodology and supporting tool to understand differences in DNS censorship measurement resulting from changes in packet header information (e.g., source IP and source port). We begin by reviewing the problem space and then describe the design considerations of *BreadCrumb* required to explore both routing changes and censorship results. *BreadCrumb* builds on the DNS censorship measurement methodology used by prior studies and systems [31, 36, 37].

We begin by discussing the ethical guidelines and principles we adhered to when developing our methodology and conducting our study. We then explore the design considerations for (1) choosing the destination IPs we perform the experiments on, and (2) choosing the input parameters of the packets we vary. Finally, we describe how given a set of destination IPs and source parameters, we identify routing and censorship changes, and use methodology to develop understanding of how changes in the source parameters influences paths differences and observed censorship behavior.

3.1 Overview

BreadCrumb aims to understand the impact of changing various packet parameters on: (1) the route taken by a DNS probe packet, and (2) observed DNS censorship behavior. It can be leveraged to understand the behavior in a specific country, e.g., China, or in a broader global DNS injection measurement study. Figure 1 presents an overview of *BreadCrumb*.

Previous censorship measurement methodologies focus on obtaining *response* packets and then understanding if they are *correct* [31, 36, 37]. Our goal is different. Namely we need to explore the route taken by packets that result in such censorship. This involves developing our own traceroute measurement methodology that utilizes censorship measurement packets for route discovery. *This distinction is critical as differences in packet header information between censorship measurement packets and traceroute packets could themselves lead to route changes [4]. BreadCrumb* understands this problem and generates traceroute information utilizing censorship measurement packets that ensure the route between traceroute and censorship measurement is unchanged.

Approach. *BreadCrumb* utilizes the core remote DNS censorship measurement methodologies of prior studies [31, 36, 37]. Namely, we send DNS packets from an external measurement vantage point towards IP addresses in a censored region and observe the result. Where *BreadCrumb* differs from prior works is that the goal is not to identify censorship itself but rather observe how changes in the packet header change the censorship measurement results.

A key insight is since *BreadCrumb* is interested in *changes* in censorship based on packet construction, rather than the actual censorship result itself, the challenging [31] problem of detecting if the IP address in a response is correct, is simplified. We further solve this problem by focusing our measurements on *symmetric* DNS censorship, meaning censorship apparatuses that inject responses on both inbound and outbound (from the perspective of the censor) DNS packets. As discussed subsequently, focusing on symmetric censorship aids the scale of our measurements, supports the ethical principles we adhere to (Section 3.2), and simplifies understanding response packets.

We utilize Chinese DNS censorship as a lens for understanding the phenomena of routing-induced censorship changes. We selected China for two reasons: (1) DNS censorship in China is well studied [19, 23, 25, 31, 36, 37, 42], and (2) Chinese DNS censorship is symmetric [23, 31]. This symmetry allows us to develop measurement methodologies that adhere to our ethical guidelines while still replicating the methodologies of previous studies [31, 36, 37].

By leveraging this symmetric behavior and the need to only capture changes in censorship, we can measure DNS censorship by probing *any* IP address in a given network range—we do not need to limit our measurements to actual DNS resolvers, as the censors will inject responses for packets destined for *any* address. To that end we focus our measurements on IP addresses that do not have any open ports observable via Censys [16]. Selecting such IPs allows *any* response observed to be a censorship response (since the destination IP does not respond), and ensures that no subsequent queries are generated by the measured IP addresses. This allows us extensive flexibility in our measurements while also helping to address ethical concerns (Section 3.2).

Building on prior censorship measurement [31, 36, 37] and fixed-path traceroute [4] methodologies, we generate fixed-construction DNS packets for censored domain names. We then vary the source IP address across a `/24` subnet and across source ports while also varying the TTL. This set of packets is then tested against a set of measurement destination IPs. The varying TTLs allows us to reconstruct the routes the packets take, while the varying source IP and source port allow us to discover differences in routes and censorship. Once the measurement packets and responses are collected, we perform analysis to identify routing and censorship differences. We describe this process in more depth subsequently.

Assumptions. *BreadCrumb*'s goal is to identify changes in censorship measurement behavior that results from changes to packet headers. Our goal is not to identify censorship be-

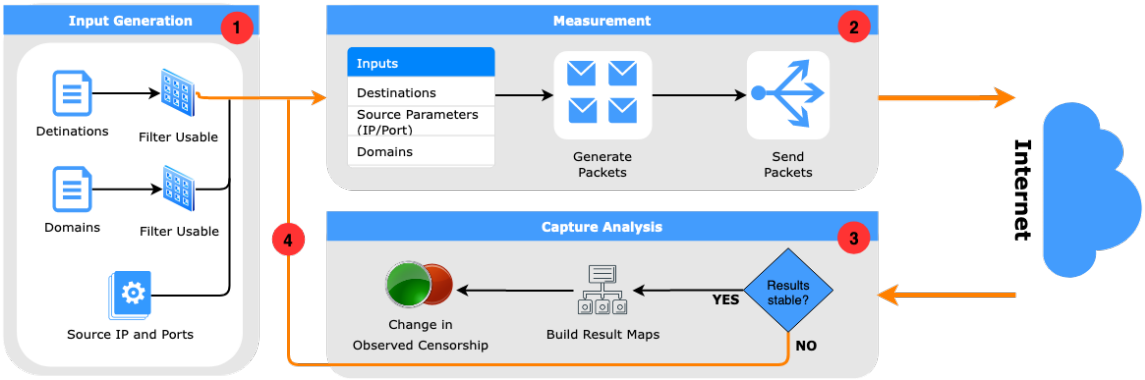


Figure 1: *Overview of* **BreadCrumb**. Our system takes a range of source ports, source IPs, destination IPs, and potentially censored domains, and ultimately identifies changes in censorship behavior due to changes in source IP and source port.

havior but instead to capture differing censorship behavior. We assume that measurement systems vary the source port they send measurement packets from, and the measurement packets may come from a variety of source IP addresses. We also assume that the censorship systems we are measuring inject packets symmetrically to inbound or outbound traffic, regardless of the existence of services on either endpoint.

3.2 Ethics

Our measurement methodology is designed around considering issues of ethics. Measuring Internet censorship carries potential risks. We consider these risks across two axes: (1) if issuing DNS queries for censored domain names via resolvers we do not control could potentially implicate unrelated parties, and (2) if issuing our queries could create load on DNS servers we do not control. We begin to address these issues by limiting our censorship measurement to IP addresses that have no observable open common ports. More specifically we send DNS packets to a closed port that will not respond to us. We also limit our measurements to the minimal volume of IP addresses and domain names we need to ensure our results are representative; our goal is not exhaustive exploration but rather demonstrating the existence of a phenomena in order to improve future studies. Despite these steps to mitigate risk it may still exist. We therefore consider the ethics of performing our experiments based on the structure established by prior studies [30,31], namely considering the ethical guidelines and principles from the Belmont [7] and Menlo [15] reports.

We begin by considering the principle of *justice*. Justice encompasses the notion that those who bear the risk of an experiment should also be those who would benefit from it. The direct beneficiaries of *BreadCrumb* are broad and include policy makers, censorship measurement researchers, and circumvention tool designers. Improvements across each of these fronts will in turn directly benefit those who potentially bear the risk of these experiments.

We next consider *respect for persons*, a principle which aims to protect humans as autonomous decision makers. Respect for persons can be misinterpreted as informed consent [31]. Rather Salganik describes this principle as "some consent for most things [35]." While we aim to respect this principle by limiting which IPs we measure, for our study it is impractical to obtain the consent of the owners of each IP address which may or may not be in use, turning our attention to our next principle.

Given that we cannot obtain informed consent, we look to *beneficence*. Beneficence weighs the benefit of performing research versus its inherent risk. Beneficence does not attempt to eliminate risk, rather it seeks to reduce it. We rely heavily on beneficence both by the selection of IP addresses that do not have commonly open ports as well as limiting the extent of what we measure to representative IPs and domains. We note, as have other studies [31], that there are diminishing returns on exhaustive measurements, and these returns likely do not justify the given risk. We also note our methodology of using exclusively IPs that are *not* resolvers is a significant departure (and reduction in risk) from prior methodologies.

Lastly, we consider *respect for law and public interest*. This can be considered the natural extension of beneficence to all stakeholders [31], not just the subject of an experiment. Considering the potential increase in DNS query load caused by censorship measurements falls under this principle. We address this principle both by limiting *what* we measure (IPs and domains) but also by significantly limiting our query rate. The only increased load in our measurements would be similar to that of Internet scans of a closed port.

3.3 Building and Deploying *BreadCrumb*

Conceptually the problem of identifying censorship changes due to routing can be broken down into several distinct tasks: input generation, packet generation, and analysis. Practically, the problems of analysis and packet generation are iterative to address packet loss and short-term routing fluctuations (discussed further in Section 3.3.3). Figure 1 shows an overview of *BreadCrumb* broken down by these tasks. We now describe each task in further depth.

3.3.1 Task 1: Input Generation

Input generation can be broken down around the sub-tasks of selecting destination IPs in a given region to explore, identifying domain names to look for censorship changes across, and selecting source IPs and ports to run the experiments over. This process is shown in component 1 in Figure 1.

Selecting Destination IPs. To obtain a comprehensive view of routing and censorship changes, we need to identify an extensive set of geographically distributed vantage points inside the region or country of interest. This problem is a direct analog to the problem of vantage selection when conducting remote measurements in prior studies [30,31,36,37]. Our goal is to understand how changes in packet construction can influence censorship measurement results, therefore we replicate the destination IP selection methods of prior studies [31,36] to the extent possible.

BreadCrumb begins by utilizing the Censys [16] system to identify all open DNS resolvers within a given region. For the purposes of this study, that is China. [1] By further leveraging Censys we then select another IP address on the same /24 subnet as the open resolvers that has *no known open ports*, including port 53 (DNS) not being open. We perform additional filtration and verification of the experimental IPs each time before they are used to address stale data and churn, while also ensuring Censys correctness. Before and during each experiment we send out DNS requests to each destination for known uncensored domains (i.e., `example.com`, `afekv.com` [2]) and if we obtain any responses to these requests, we discontinue use of that IP for any experiment (explained further in Section 3.3.3).

The benefit of selecting a "non responsive" IP in the same subnet as a vantage point is twofold. First, this method keeps with our ethical principles (Section 3.2) by limiting risk as much as possible—our queries elicit no responses from the destination IP, and the destination IPs do not in turn issue subsequent queries. Second, any DNS responses we receive should be the result of censorship, thus simplifying analysis. From this list of viable candidate destination addresses we randomly sample weighting by the autonomous system in an effort to obtain network diversity. The number of IPs selected varies and is kept as low as possible based on the experiment. This is discussed further in Section 4.

Identifying Domains to Query. Similar to the selection of destination IPs, we attempt to replicate the domain selection methodology of prior studies [30,31,37]. We focus on measuring changes in censorship to domain names from the Citizen Lab Block List (CLBL) [14]. Since our goal is discovering routing induced censorship changes and not studying comprehensive censor behavior, we select a minimum sample of domains from each category for our measurements that focus on domain scale. For other experiments that focus on routing instead of changes in censorship we use 2 well known uncensored domains (i.e., `example.com`, `afekv.com`). Section 4 provides per-experiment domain selection details.

Selecting Source IPs and Ports. The hypothesis *BreadCrumb* seeks to explore is if the selection of source IP address or source port ultimately impacts the form of remote vantage point censorship measurement utilized by previous studies [30,31,36,37]. To explore this hypothesis *BreadCrumb* selects source IPs at random from a research /24 IPv4 address range. *BreadCrumb* also selects source ports at random from the ephemeral port range. The number of IPs/ports selected depends on the experiment as discussed in Section 4.

3.3.2 Task 2: Packet Generation

BreadCrumb seeks to understand two interconnected phenomena both related to the construction of packets. First *BreadCrumb* aims to understand how packet construction influences routing in scenarios similar to previous studies. Second, *BreadCrumb* needs to understand how those changes in route map to changes in censorship.

Performing Censorship Traceroute. Our first step is to reconstruct the approximate router hop path traveled by our censorship measurements. A key challenge here is typical `traceroute` tools do not account for changes in packet construction that can influence routing [4]. Further, we cannot directly use tools such as Paris Traceroute [4] as it does not natively support several features necessary for our work. For example it has: (1) no integrated functionality to generate application-level packets, (2) no integrated ability to vary source IP within an experiment, and (3) limited ability to control packet volume. Moreover, while Paris Traceroute produces routes, we are (mainly) interested in censorship.

BreadCrumb addresses these challenges by combining fixed-route traceroute tooling with censorship measurement methodologies. Given a particular input combination (i.e., destination IP, source IP, source port, and domain), *BreadCrumb* sends censorship measurement DNS queries to the destination IP at incrementing TTLs (similar to `traceroute`), while carefully controlling for fields in the packet header that are known to be used by routers to perform load balancing [4]. We record all DNS and ICMP `TTL Expired` responses we receive during the experiment, using the ICMP responses to reconstruct a path, and the DNS responses to identify changes in censorship.

A second key challenge is disambiguating ICMP responses since many of the packet header fields typically used to disambiguate experiments (e.g., source port) must be fixed as TTL increases in order to ensure consistent routing. To solve this problem we disambiguate response ICMP packets via the UDP checksum field inside the UDP header embedded in the ICMP response. The UDP checksum is an ideal field as both routers are not known to utilize that field for load balancing [4], and the UDP checksum field is within the first

[1] Censys provides geolocation data using the Maxmind GeoIP2 dataset.
[2] `afekv.com` is a domain created specifically for Censys DNS scans.

8-bytes of IP packet data required to be embedded within the ICMP packet [32]. This UDP checksum technique is also utilized by the Paris Traceroute fixed-route tool [4].

BreadCrumb performs DNS Traceroute only for experiments aimed at understanding paths. For experiments aimed at understanding changes in censorship, *BreadCrumb* forgoes all traceroute and TTL measurements to limit the overall number of packets sent, in keeping with our ethical guidelines.

Sending DNS Queries. *BreadCrumb*'s core measurement functionality takes in source IPs, source ports, destination IPs, domain names, and experiment parameters (e.g., censorship changes or traceroute mode), and generates a *packet schedule*. The packet schedule randomizes all parameters and distributes queries evenly across destination IPs to avoid overloading while simultaneously ensuring rate-limiting (locally and globally). *BreadCrumb*'s packet generation and capture tooling is implemented in the Go programming language [21].

3.3.3 Task 3: Analysis

After we have generated our first packet schedule and conducted our first round of measurements, we begin an analysis phase which not only provides the ultimate understanding of changes in censorship, but also directs an iterative process of subsequent measurements to ensure correct results.

Establishing Result Consistency. *BreadCrumb* aims to understand subtle changes in censorship behavior based on packet construction. To achieve this we must establish confidence in the consistency and robustness of our results. Given our methodology we expect any response packet to be a censorship event, and any non-response to be lack of censorship. If we send a packet and do not get a response, that could be a lack of censorship, or it could be a dropped packet. Moreover if we then send a second packet and get a response, was that a change in route, or recovery from a packet drop? Packet drops commonly occur on the Internet for a variety of reasons, and routes could change regardless of packet construction. Both of these scenarios could appear as a *change in censorship* and must be taken into account by our methodology.

We address these challenges by taking an iterative experimental approach and using a conservative decision metric. Steps 3 and 4 of Figure 1 show this process. After each experiment iteration *BreadCrumb* accumulates all queries and responses and identifies if some responses were missing. A missing response could denote a lack of censorship, a change in censorship (if other queries generated a response), or a dropped packet. For each missing packet, *BreadCrumb* repeats the specific experiment (source IP & port, destination IP, domain) in the next iteration. This process repeats several times until a response occurs or we reach maximum retries. The convergence of this method is discussed in Section 4.5.

BreadCrumb takes a conservative approach to identifying changes in censorship by assuming that if a response is *ever* received for a given experiment, that response is correct (and excluded from future measurements). For example, if we perform the same experiment three times and the first two times no response is seen, and the third time we receive a response, there are two possible scenarios: (1) the route changed between the second and third packets, or (2) packet loss occurred. Our decision metric would assume the latter and identify this experiment as "unchanged." This metric is conservative as we potentially undercount routing changes, making our measures lower bounds on the total routing changes observed.

Censorship Profiles. Our goal of studying changes in censorship behavior for a particular vantage point necessitates constructing a representation of what is censored from vantage to vantage. To this end we define a *censorship profile* to be the set of all domains censored at a specific destination IP. Since we only measure "inactive" IP addresses (IPs with no known open ports, confirmed by controls) this set is easy to generate as any response we obtain for a particular domain directly indicates that particular domain is being censored.

Identifying Changes in Censorship Profile. Not all of the domains we test are uniformly censored across all destinations [31, 41]. We need a way of identifying changes in censorship on a per-destination basis. Ideally a particular destination is expected to have just one censorship profile (if measured from a single source), but in reality changing the probe packet structure results in a number of different censorship profiles stemming from routing changes. We pick out the profile that is most common among all measurements and record all deviations from the majority profile.

Controls. As part of our input generation we include control domain names (known to be widely censored or never censored) and destination IP addresses of public resolvers outside the censorship region. We utilize these controls to ensure "inactive" IPs remain inactive and that all responses are censorship responses. We also use our control IPs to ensure our measurements are functioning correctly and not subject to packet loss or transient network outage.

4 Dataset and Experiments

Our exploration of packet construction driven censorship changes spans three distinct and iterative experiments: (1) router path exploration, (2) assessing the scope of changes across IPs, and (3) assessing the scope of changes across domain names. In this section we explain the data collected for each experiment and how we processed the data to obtain our results. In keeping with our ethical principles, we limit the scale of each experiment to what we believe are the minimal viable set of domains and/or IP addresses to successfully evaluate our research questions. Table 1 provides an overview of input parameters used for all experiments.

Router Path Exploration: aims to understand *path changes* resulting from changing various packet parameters. *BreadCrumb* performed DNS traceroutes on 363 destination IPs (one for each AS, as discussed in Section 3.3.1) using

Experiment	Destinations	# of Source IPs/Ports	Domains
Router Paths	363	1/1, 196/1, 1/196, 14/14	2
Explore. Changes across IPs	363	100/50	25
Changes across IPs	10000	200/100	4
Changes across Domains	492	5882 Total IP:Port Combos	75

Table 1: *Overview of our experimental dataset.* For Changes across Domains, we used a collection of source IP/port combinations from Changes across IPs that exhibited the most variation in order to explore (only) domain differences.

two well-known uncensored domains (`example.com`, `afekv.com`) while varying packet construction per our methodology.

Exploratory Changes Across IPs: aims to establish a minimal viable set of domains to use for the following, larger, Changes Across IPs experiment.

Changes Across IPs: aims to understand *censorship changes* remote measurements would observe from varying the source IP and source ports of measurement packets. We measure a geographically diverse set of IP addresses across China, selected using the previously outlined methodology.

Changes Across Domains: looks to understand if routing changes also result in more subtle changes across what is censored across a larger set of domains. We use the results from previous experiments to identify optimal destination IP and source parameters most likely to yield differences in censorship behavior. This set is biased towards change and is only used to identify domains (RQ4, Section 5.4), not the prevalence of routing changes.

4.1 Destination IP Selection

We began by identifying a set of vantage points inside China using the Censys dataset as described in Section 3.3.1. Using an August 2021 snapshot we began with 13.6 million IPs that were geolocated within China. Of those we identified 175K that were open on port 53 (i.e., open resolvers). From these we used our "inactive" methodology to select 175K IPs in the same subnet which had no known open ports. We selected IPs near these open resolvers in an effort to provide results consistent with open resolver measurement studies [31, 36], while still adhering to our ethical guidelines. These 175K IPs were spread across 363 autonomous systems (ASes). We verified before and during our experimentation that each IP address remained unresponsive on port 53. All experiments were performed across the month of September 2021.

Weighing the need for a comprehensive picture of routing and censorship changes against ethical concerns, we randomly selected a subset of IPs appropriate for each experiment, controlling for balance across ASes. For our Router Paths experiment we randomly selected 363 IPs (one per AS), and to explore Changes Across IPs we selected 10,000 IPs (randomly, but weighted by AS). For Changes Across Domains, we selected 492 destinations that exhibited the most variation in previous experiments to explore (only) domain differences.

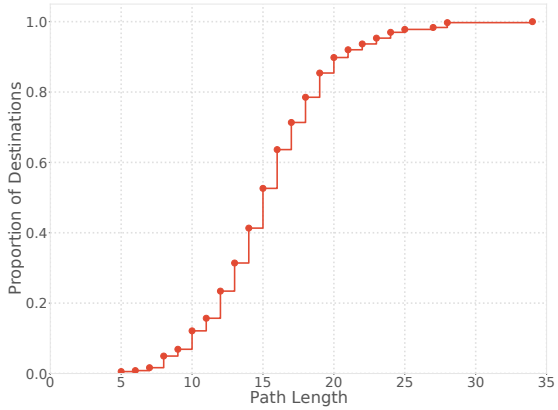


Figure 2: *Cumulative Distribution Function (CDF) of observed path lengths during trial experiments.* This distribution was used to decide the TTL range to perform experiments on in order to limit the number of packets sent. We observed a mean path length of 15 and roughly 98% of the paths have a length less than 25.

4.2 Selecting TTL Range

Before conducting our Router Paths experiment we need to understand the distribution of route lengths across our dataset in order to limit the number of measurements needed to quantify route changes. Such understanding is needed as we must ultimately generate packets for each possible TTL value, while also varying multiple other source packet parameters. Limiting the max TTL ensures we limit the overall number of packets sent to a minimum. We used *BreadCrumb* to perform non-censorship related DNS measurements to a representative subset of the initial seed open resolvers in the target region, querying an uncensored domain. We use uncensored domains to ensure we explore the full path between our measurement machine and various destinations, not just to some subset of censorship infrastructure. Figure 2 shows the distribution of those path lengths. We find that 98% of destinations have a path of less than 25 hops. We use this information to test TTLs in the range of 2 to 25.

4.3 Selecting Domains

We began with the Citizen Lab Block List (CLBL) [14] consisting of 567 sensitive domains across 27 different categories for China. Our goals were to understand the extent and different forms of censorship variation across a large set of destinations. Given our ethical goal to minimize risk combined

with the likelihood of diminishing returns [31], we selected a subset of domains from this list spread across categories.

We selected one domain from each category (excluding Pornography and Terrorism) and performed a preliminary "Exploratory Changes across IPs" experiment across the same 363 destination IPs used in the "Router Paths" experiment. From this we selected three domains that showed changes in behavior due to routing which we utilized for subsequent experimentation. These domains were a US-based social media platform, a US-based news website, and a US-based think tank (See Appendix A). In addition to these three we included a control domain we own that has no history of observed censorship. For our final experiment on Changes across Domains we selected 75 total domains, three from each category.

4.4 Selecting Source IPs and Ports

We utilized a contiguous research /24 IPv4 address range for our experiments. For these experiments all IPs were routed to the same device and all had the same upstream first-hop router. Keeping with community norms [17] the address range had reverse DNS PTR names and `WHOIS` records that indicated it was used for research Internet scans, and each IP hosted webpages with contact and opt-out information. We selected IP addresses at random from this subnet, avoiding .0 and .255. For source port we selected ports at random from the ephemeral range.

4.5 Result Convergence and Filtering

As discussed in Section 3.3.3, we performed iterations of experiments to establish confidence in the consistency and robustness of our results, removing consistent experiments as they were generated. Figure 3 shows the destinations measured and coverage at each iteration. We observe that with each iteration, we significantly reduce the number of destinations measured and at the same time increasing the number of potential responses received.

Across the 10,000 destination IPs that we tested, 9430 generated censored responses when expected and remained "inactive." We began with this set and iterated, removing experiments from the set when they produced a censored response to the well-known censored domains. This method accounts for potential packet loss or short-term transient outages, while also being conservative by immediately terminating if any response is received. By iteration 12, results had stabilized both in terms of response and remaining experiments, and *BreadCrumb* terminated. The remaining experiments that lacked censorship responses across a (destination IP, source IP, and source port) tuple (but for which other experiments at that destination IP showed censorship) we expect, and verify, represent routing-induced censorship changes. We note that it is possible that a routing change during iterative measurement would change a particular experiment from "not responding" to "responding." Given that our default state is assuming consistency for a destination IP across all experiments, such a shift would move an experiment *into* the consistency state, thus under counting routing changes, and yielding our conservative estimation.

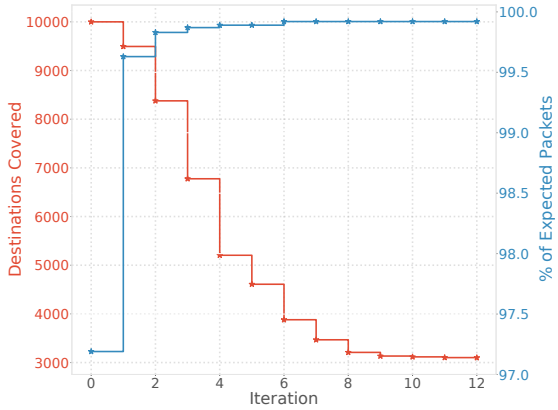


Figure 3: *Result coverage per iteration.* Percentage of responses and number of destinations tested at each iteration to obtain results. We see that we get a higher coverage with each iteration but at the same time send packets to a fewer number of destinations. Results stabilize by iteration 12.

5 Results

In this section we show the impact of changing packet source parameters on: (1) changes in path taken by a DNS probe, and (2) changes in observed censorship behavior and the extent of those changes. This section is structured as posing and answering a series of research questions on the effects of input parameters in a DNS censorship measurement scenario. The section concludes with a case study visually documenting how changes in route result in measurement packets bypassing censorship devices.

5.1 RQ1: Does Varying Source Port and IP Change the Path of Censorship Measurement Packets?

Previous work has shown that routers balance load based on packet construction [4, 5]. The goal for this research question is to understand the presence of these changes in the context of DNS censorship measurement. Our goal is not to thoroughly explore the magnitude of multipath load-balancing, but instead to demonstrate how changes in source parameters cause path differences in a censorship measurement packet. To this end, we employ two metrics - *number of nodes* and *number of paths*. A *node* is an IP address that indicates a router hop and *number of nodes* is the set of all nodes observed by *BreadCrumb* for a (destination, domain) pair across all changes in the source parameters. A *path* is a set of all nodes observed in a DNS traceroute with *one* set of source parameters and *number of paths* is a set of all paths for a (destination, domain) pair across all changes in the source parameters.

Experiment	Mean Number of Nodes	Mean Number of Paths
Fix IP & Port	15	2
Fix IP, Vary Port	55	110
Vary IP, Fix Port	89	134
Vary IP & Port	75	129

Table 2: *Path Metrics.* Mean *number of paths* and *number of nodes* for the different experiments. We utilize the same total number of measurements to ensure results are normalized (e.g., test fewer IPs when varying both IP and port). We observe that simply changing the source port significantly increases the number of paths, and additionally changing both the source IP causing even more variation.

We apply these metrics to the dataset collected in the Router Paths experiment (described in Section 4). Table 2 presents a summary of results from the different experiments performed. Figures 4 and 5 also show the results of these experiments. We now outline the parameters of the experiments and discuss the results in more depth.

Constant (Fixed) Parameters. We performed repeated experiments with a set of fixed source parameters. Performing repeated measurements *without* changing source parameters provides us with a baseline to understand changes in path due to changing source parameters in the later experiments. Across all destinations, we observe a mean of 15 *number of nodes* and 2 *number of paths* (with 196 repeated measurements). This indicates that a fixed set of source parameters *does not* yield significant differences paths, over a fixed period of time, with repeated experiments.

Varying Source Port. Next, we performed the same experiment but with 196 randomly selected source ports in the ephemeral port range and fixing all other source parameters (thus keeping the total number of probes consistent). We observe a mean of 55 *number of nodes* and 110 *number of paths*, across all destinations. Given that we performed a total of 196 experiments, the maximum *number of paths* we could have observed is 196.

We can see that varying only one source parameter, the source port, results in a considerable number of differing paths for DNS censorship measurement packets.

Varying Source IP. We next performed an experiment across 196 randomly selected source IPs from our /24 IPv4 address range and fixed all other source parameters. We observe a mean 89 *number of nodes* and 134 *number of paths*, across all experiments. Varying source IPs appears to have a higher impact in causing differing paths with the DNS probe packet with almost 60% more nodes discovered compared to just varying source port.

Varying Both Source IP and Port. Next we performed an experiment across 14 source IPs and ports (totaling 196 combinations, keeping total volume consistent). We observe 75 *number of nodes* and 129 *number of paths* on average. As expected, these numbers appear to lie in between varying ports and varying IPs, as we sample from both distributions.

In answering RQ1 we established that varying different source parameters of a DNS censorship measurement packet has a strong impact on the path traversed by the packet and to differing levels, based on different source parameters. Figure 4 shows a distribution of these metrics across destinations. Note that in varying the source IP, a group of the *number of paths* are in the rightmost extreme indicating that for many of the destinations, varying 196 source IPs produced 196 unique paths. Figure 5 shows a CDF of the *number of nodes* and *number of paths* metrics across all the destinations. Performing repeated experiments with fixed source parameters does not produce significantly differing paths but changing the source parameters independently or together causes varying modes of differing paths across the dataset.

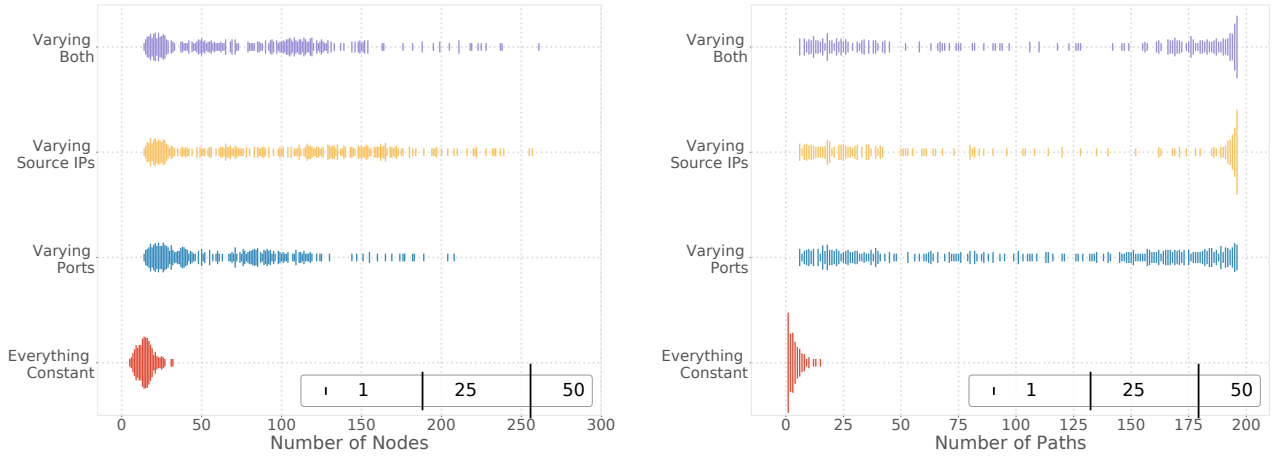
5.2 RQ2: Does Varying Source IP and Port Change Measured Censorship?

We have established that changing packet source IP (within a subnet) and source port causes significant routing changes. We now explore the impact of route changes on censorship behavior. As described in Section 4, we use results from the large-scale *Changes across IPs* and *Changes across Domains* experiments to understand changes in observed censorship.

We explore the change in censorship due to routing changes via the Changes across IPs experiment. In this experiment we select 10,000 destination IPs at random within China using the aforementioned methodology, and measure three known-censored domains. We construct censorship profiles for each destination IP and identify changes in censorship on a per-destination basis (see Section 3.3.3). Figure 6 shows these results. In this plot we focus on an observation that changes in censorship were overwhelmingly *all-or-nothing*—either a particular (source IP, source port) experienced "expected" censorship, or *no censorship at all*. We find 37% of destinations across 56% of the ASes exhibited some form of change in observed censorship based on input parameters in this randomized experiment.

Of note is that a significant portion of destination IPs for which there are only a handful of experimental parameters that yield changes in censorship behavior. At first glance this could be attributed to measurement error or dropped packets, however, that is not the case. For these parameters the following properties are true: (1) All domains tested exhibited the same behavior for these experimental parameters, (2) These results remained constant over numerous experiment iterations, across half a day, and (3) manual inspection of a random sample of the results days later yielded the same outcome. If such experimental parameters are in fact some form of routing error or dropped packets, given the extent to which they exist and remain consistent, we argue they *are* routing induced censorship measurement changes.

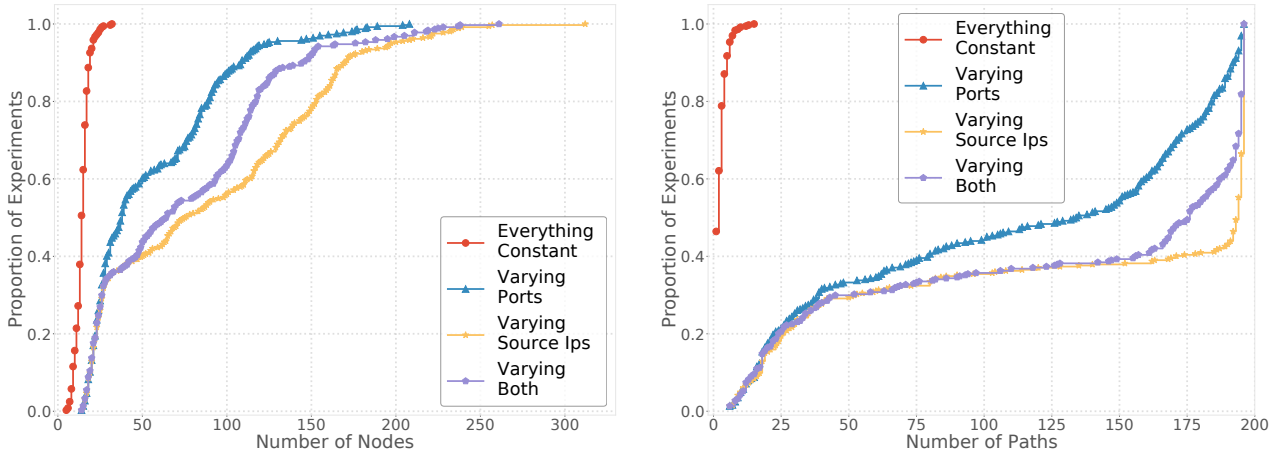
Exploring the remainder of the results, we find that 3%



(a) Number of Nodes

(b) Number of Paths

Figure 4: *Distribution of number of nodes and paths across all the destinations for each experiment.* Marker size does not scale linearly. The *number of paths* for the case where source IP is varied is more distributed towards the extremes whereas with varying ports, they are more evenly distributed. Of note is that for a large portion of the destinations, each source IP resulted in a completely different path.



(a) Number of Nodes

(b) Number of Paths

Figure 5: *CDF of number of nodes and paths observed for all destinations for the different experiments.* For the Constant (Fixed) experiment, the number of nodes seen is at most 30 and number of paths within 13. When we vary the source port and source IP, the number of nodes and paths vary to different extents at several different modes. This figure shows route variation does exist, and it could potentially influence censorship (explored further in RQ2).

of the destination IPs and 17% of ASes observed changes in censorship over 10% of our experimental parameters. We will subsequently show (Section 5.3) that these effects are based on the bit patterns of the source IP *and* destination IP, as well being randomly distributed over source port. Given these distributions, *changes in censorship tend to accumulate* over large measurements. This has two critical effects: (1) as censorship measurements scale-up (such as in prior studies [31, 37]), packet parameter-induced censorship changes are more likely to appear, and (2) comparisons within an AS are more likely to observe censorship changes than individual IP measurements (as seen in Figure 6).

Given that variation can be highly-localized to specific IP and port combinations, as well as domains, in randomized experiments the observed effects may be small and could be mistaken for packet loss or similar transient phenomena. For example, in the *Changes across IPs* experiment, across all destinations that exhibit censorship, we received censorship

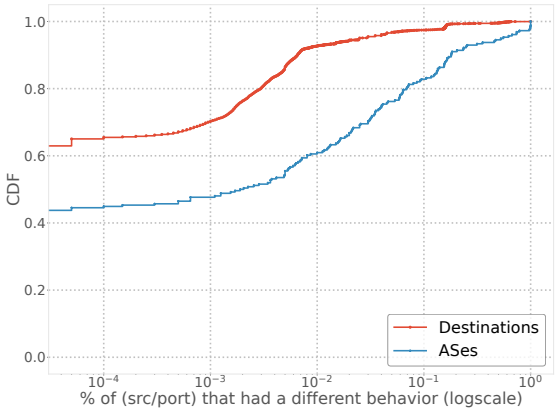


Figure 6: *CDF of (source IP, source port) pairs which had censorship changes, per destination.* X-Axis is log scaled. We observe that 37% of the destinations and 56% of ASes had persistent censorship differences across some set of source IP and source port combinations. As the effect is based on source IP, source port, and *destination IP*, change is cumulative, leading to increased prevalence across ASes.

responses for 66% of the queries we sent to known sensitive domains. However, if we instead limit our experimentation to just two known widely censored domains that were less likely to experience variation, we received censorship responses to 97% of DNS queries sent when randomizing packet parameters. This underscores the need for careful construction of packets in conjunction with the selection of domains.

5.2.1 All-or-Nothing Censorship

We find that 95% of all censorship changes were *all-or-nothing*, meaning that depending on the source parameters (IP and port), we either observed no censorship or "expected" censorship activity. The extent of change varied for each destination and also with source parameters. As our experiment was conducting over 200 source IPs and 100 source ports, we now explore the influence varying each source IP and port had on changes in censorship.

Source Parameters. We begin exploring this behavior in Figure 7, a CDF of the percentage of source ports and IPs that exhibited the *all-or-nothing* behavior. For source IPs, we observe that: (1) roughly 5% of the destinations had a change across all source IPs, (2) 20% of the destinations had one source IP which resulted in no censorship, and (3) for the average destination IP, 3 source IPs exhibit no censorship. For source ports, we see that (1) for roughly 18% of the destinations, all ports experience no censorship (across the 200 source IPs), and (2) for the average destination, around 50 ports experienced no censorship across all IPs. From this we see that the number of paths vary with both changing source port and IPs but change in censorship behavior is influenced more by change in source IPs rather than ports. These results could also however point to multiple independent routing decisions, some based on IP, some on port, influencing change.

5.2.2 Other Forms of Variation

Even though *all-or-nothing* was the predominant form of variation observed, there were changes in observed censorship among the other domains we tested as a consequence of change in packet parameters. We observed a change in censorship for the US-based think tank at 288 destinations and the US-based news website at 161 destinations. These results were manually confirmed to be accurate. The latter was particularly interesting since the US-based news website was only censored at 161 destinations and *all* of them had a change in observed censorship as a consequence of change in packet parameters. We speculate this is a result of the distributed nature of the great firewall, whereby some routing change lead to a path that exercised a completely different piece of censorship infrastructure. We also note that the scope of this experiment was only a handful of domains. In Section 5.4 we explore a larger set of domains and find further variation.

5.3 RQ3: Do Particular Source IPs or Ports Cause More Censorship Changes?

We have established that measured censorship changes occur as a consequence of selecting different packet parameters. We now focus on understanding whether particular source IPs or ports cause more change than others, and why.

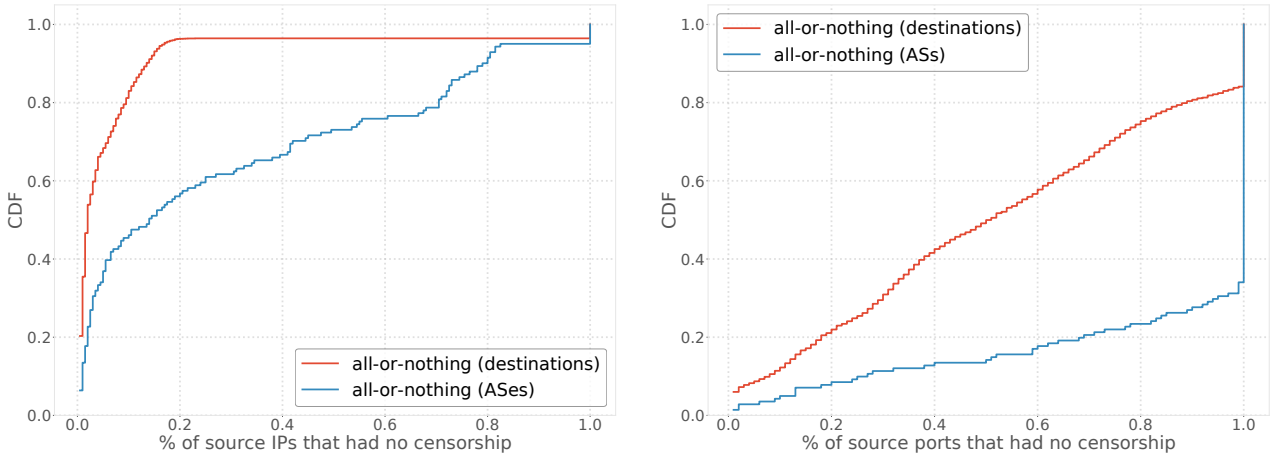
Figure 8(a) shows the distribution of destinations for which a particular source IP resulted in no censorship, with the colors representing the lowest 3 bits of the source IP. Surprisingly, we observe that: (1) Some source IPs seem to elicit more non censorship behavior than others and are not as uniformly distributed as source ports, (2) Depending on the last 3 bits of the source IP, the number of destinations for which the particular source IP experiences no censorship almost doubles, and (3) the last three bits of the source IP has a direct impact on the number of destinations where the particular source IP causes change.

Figure 8(b) is the same dataset as (a) but instead colored by the last 3 bits of the destination IP. We observe that: (1) not only do certain source IPs cause more changes, the destination IP itself also influences the amount of change in observed censorship behavior, and (2) we see that the amount of observed censorship behavior is a function of the combination of source and destination IPs' last three bits.

Upon investigation we discovered that router load balancing algorithms [11–13] have been known to *XOR* the last few bits of the source and destination IPs to perform routing decisions. Observing the patterns in Figure 8, we can deduce that a form of known load-balancing based on packet headers influences changes in censorship results.

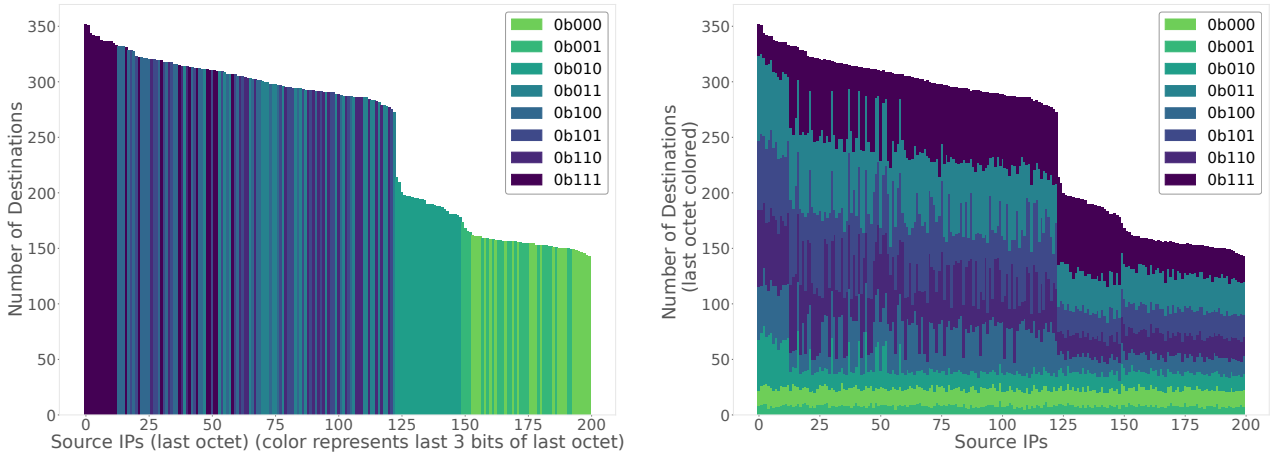
5.4 RQ4: How Prevalent are Censorship Changes Across Domains?

We have shown that changing packet parameters results in different paths and also in differing observed censorship be-



(a) Source IPs **(b)** Source Ports

Figure 7: *CDF of proportion of source IPs/Ports that had different censorship behavior (for all destinations that had some change).* Plot (a) shows the median destination saw changes across 5% of source IPs. 10% of destination ASes saw changes across 80% or more of source IPs. 5% of destination IPs saw changes across all source IPs. Plot (b) shows change is distributed roughly evenly over the ports, with roughly 5% of destinations showing changes across only 1% of ports. Of note is 18% of destinations saw changes for all source ports (across source IPs).



(a) Source IP bits colored **(b)** Destination IP bits colored

Figure 8: *Distribution of experiments with censorship change across source IPs.* (a) results are colored by the lowest 3 bits of source IP, (b) results are colored by the lowest 3 bits of destination IP. X-axis is sorted (descending order) based on changes caused by a source IP. We see that the extent of censorship changes nearly *double* depending on the lowest 3 bits of the source IP. In (b) we see that the destination IP lower order bits also influence censorship changes significantly. We note that routers are known to load balance by these bit patterns [11–13].

havior for some destinations. We now look at whether subtle changes across what is censored arise due to routing changes. For our final Changes across Domains experiment with 492 destinations and 75 sensitive domains we observed that: (1) the *all-or-nothing* variation was still the most dominant and held for all the domains tested and caused changes in 156 destinations, and (2) some domains individually exhibited change in observed censorship. Table 3 summarizes these results. We find that the all-or-nothing behavior is once again the dominant behavior. But as we expand the domain set we discover several domains across 4 categories that experience small-scale changes in censorship behavior based on source IP and port selection. Of note, all of the domains we observed to experience small-scale changes were first censored within

Category	Number of Destinations
All-or-Nothing	156
Online Dating	17
LGBTQ+	7
Gambling	9
Hacking Tools	6

Table 3: *Changes Across Domains.* Number of destinations that observed a change in censorship for sensitive domains, by category. All-or-nothing denotes the effect we observe where we see either all "expected" domains censored, or no domains are censored, based on the source IP and port. The experiment was conducted across 492 destinations with 75 domains across 25 categories. While all-or-nothing behavior was most prevalent, several categories had domains that experienced small-scale censorship changes based on source IP and port. The small-scale changes were exclusively among domains first observed to be censored within the last two years.

the last 2 years. This result points to possible inconsistencies in the configuration of the censorship devices, which has been previously theorized [31].

5.5 Individual Case Study

Building on *BreadCrumb*'s DNS traceroute mechanism, in this section we present a case study on an individual example using network graphs. For this case study we selected a single destination for which varying source IPs (in the same /24) exhibited differing censorship behavior for a particular sensitive domain. For each source IP, we performed DNS traceroute with 50 different source ports and combined the traceroute results into a single network graph. Figure 9 shows the two network graphs we obtained. The experiments in the graph on the left always observed censorship and the experiments in the graph on the right never observed any censorship. The node at which censorship occurred is marked in red, the nodes that only appeared on the left or on the right are marked with different colors. We observe that at layer *N=14*, there is a diversion in path with the source IP on the right, taking it through a path that *never* passes through the node that appears to perform censorship even though the path converges back together to reach the destination. This is a clear example showing that the change in path due to the source parameter causes the packet to not pass through the system that is performing censorship, leading to a change in the observed censorship behavior.

6 Discussion and Recommendations

We have shown that source IP (within a subnet) and source port of the header of censorship measurement packets influence packet routes, and in turn result in varying measured censorship. Such changes can be incorrectly attributed to geographical variations or non-determinism in the results due to packet loss. In this section we provide insights on censorship measurement methodology relating to packet construction and other guidance for censorship researchers when performing remote censorship measurement.

We suggest the following considerations when performing censorship measurement:

- **Selecting Input Parameters:** Source IP and port need to be carefully picked when performing the experiment. Since the last 3 bits of the source IP have a direct impact on results, techniques must perform measurements from a diverse (with respect to the lowest 3 bits) set of source IPs. Source port had a lesser impact on result in our particular study, but depending on the network structure of the country being studied, this observation can change. Therefore techniques must perform experiments with several different randomly selected source ports to rule out the influence on routing changes. Where possible, measurements should be spread out over numerous source IPs within a subnet to ensure diversity of paths.
- **Path Reconstruction:** We have shown that paths vary greatly due to packet construction. When reconstructing network paths relating to censorship and attempting to identify censoring nodes, it's critical to ensure the traceroute packets are constructed identically to the measurement packets, paying careful attention to fields known to be leveraged by routers for load balancing. *BreadCrumb* is one such tool to aid in reliable path reconstruction.
- **Picking Destinations:** We have also shown that measurement paths are dependent on the bit pattern of the lowest 3 bits of destination IP. Measurement techniques must pick destination IPs that have a diverse distribution with respect to these patterns (and thus routes), in addition to geographic or network diversity.
- **Packet Loss and Retries:** Packets are frequently dropped in the network. Measurement techniques must perform careful repetition to establish consistency in their results in order to be able to distinguish between changes caused by routing vs packet loss. This repetition must involve spreading measurements out over different input parameters, *not* simply retrying packets without altering construction.

Towards Censorship Evasion. In this work we established that source parameters under the control of a user can influence remote censorship measurement results. A natural extension of this work is to establish the viability of users within countries experiencing censorship leveraging the described routing phenomena to evade censors. We envision such work could take two directions, both aimed at the development of user-facing tooling. First, it could aim to evade censors completely using the phenomena established in this work. Second, it could seek to leverage route variations to split content across multiple routes in an effort to confound censor activity. We leave further exploration of such evasion technology to future work.

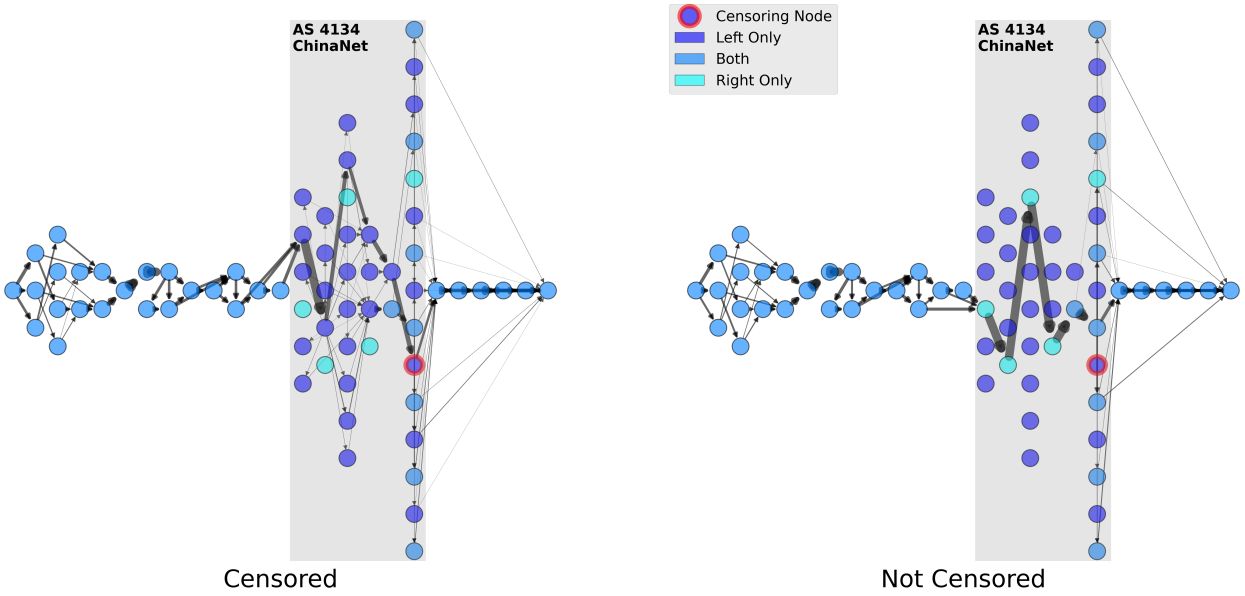


Figure 9: *Route and Censorship Case Study.* This figure shows a concrete example of the path taken by DNS measurement packets between our measurement source and a destination inside China for the censored domain. The shaded region represents nodes in ASN 4134 (ChinaNet), that is known to perform censorship. For this single example destination IP we collect all the (source IP, source port) pairs that resulted in censorship and plot the observed traceroute nodes collected by *BreadCrumb* on the left of the figure. We then group the pairs that *did not* result in censorship on the right of the figure. The only difference between the two path sets is the source IP (within the same /24) and source port. We also manually extract the node we believe is responsible for censorship based on TTL responses, and color it red. We see that there is a distinct set of path differences beginning at N=14 that result in different nodes being seen in the censored vs non-censored routes. The width of the lines represents the number of experiments that flow through the given edge.

7 Conclusion

Despite the development of numerous methods to perform remote longitudinal censorship measurement, few have addressed non-determinism or inconsistencies in censorship results. We built *BreadCrumb*, a measurement methodology and associated tool to understand the prevalence of censorship changes due to router-based load balancing across different source parameters of a DNS probe packet. Using Chinese DNS censorship as a lens, we show that changing the packet parameters used for measurement causes significant changes in the path of the packet which in turn results in changes in measured censorship. *BreadCrumb* shows how routers use certain bits of the source IP and destination IP to make load-based routing decisions that have a direct impact on observed censorship behavior. We provided several insights on selecting input parameters like source IP, source port and destination IP when designing and performing censorship measurement. We also provided guidance for future measurement methods.

8 Acknowledgements

The authors are grateful for the assistance and support of Randy Bush, Zakir Durumeric, and Frank Li. We also thank the anonymous reviewers whose thoughtful feedback improved the work significantly. This work was supported in part by funding from Office of Naval Research (ONR) award N00014-18-1-2662 and DARPA MICE award HR00112190122.

References

[1] Rafael Almeida, Renata Teixeira, Darryl Veitch, Christophe Diot, et al. Classification of load balancing in the internet. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1987–1996. IEEE, 2020.

[2] 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.

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

[4] Brice Augustin, Xavier Cuvellier, Benjamin Orgogozo, Fabien Viger, Timur Friedman, Matthieu Latapy, Clé-

mence Magnien, and Renata Teixeira. Avoiding traceroute anomalies with paris traceroute. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 153–158, 2006.

[5] Brice Augustin, Timur Friedman, and Renata Teixeira. Measuring multipath routing in the internet. *IEEE/ACM Transactions on Networking*, 2010.

[6] Michael Bailey and Craig Labovitz. Censorship and Co-option of the Internet Infrastructure. Technical Report CSE-TR-572-11, University of Michigan, Ann Arbor, MI, USA, July 2011.

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

[8] Kevin Bock, Yair Fax, Kyle Reese, Jasraj Singh, and Dave Levin. Detecting and evading censorship-in-depth: A case study of iran's protocol whitelister. In *10th USENIX Workshop on Free and Open Communications on the Internet, FOCI 20)*, 2020.

[9] Kevin Bock, Gabriel Naval, Kyle Reese, and Dave Levin. Even censors have a backup: Examining china's double https censorship middleboxes. In *Proceedings of the ACM SIGCOMM 2021 Workshop on Free and Open Communications on the Internet*, pages 1–7, 2021.

[10] Abdelberi Chaabane, Terence Chen, Mathieu Cunche, Emiliano De Cristofaro, Arik Friedman, and Mohamed Ali Kaafar. Censorship in the Wild: Analyzing Internet Filtering in Syria. In *ACM Internet Measurement Conference (IMC)*, 2014.

[11] Cisco. Understanding etherchannel load balancing and redundancy on catalyst switches. https://www.cisco.com/c/en/us/support/docs/lan-switching/etherchannel/12023-4.html, 2007.

[12] Cisco. LAG Load Balancing on Cisco 350 and 550 Series Switches. https://www.cisco.com/c/en/us/support/docs/smb/switches/cisco-550/x-series-stackable-managed-switches/1388-LAG-Load-Balancing-on-Cisco-350-and-550-series-Switches.pdf , 2021.

[13] Cisco. Load Balancing in CEF. https://www.ccexpert.us/mpls-network/load-balancing-in-cef.html, 2021.

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

[15] 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.

[16] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J. Alex Halderman. A Search Engine Backed by Internet-Wide Scanning. In *ACM Conference on Computer and Communications Security (CCS)*, 2015.

[17] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. ZMap: Fast Internet-Wide Scanning and its Security Applications. In *USENIX Security Symposium*, 2013.

[18] Roya Ensafi, Philipp Winter, Abdullah Mueen, and Jedidiah R. Crandall. Analyzing the Great Firewall of China Over Space and Time. *Privacy Enhancing Technologies Symposium (PETS)*, 2015.

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

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

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

[22] Devashish Gosain, Anshika Agarwal, Sahil Shekhawat, H. B. Acharya, and Sambuddho Chakravarty. Mending wall: On the implementation of censorship in india. In *Security and Privacy in Communication Networks*, 2018.

[23] Nguyen Phong Hoang, Arian Akhavan Niaki, Jakub Dalek, Jeffrey Knockel, Pellaeon Lin, Bill Marczak, Masashi Crete-Nishihata, Phillipa Gill, and Michalis Polychronakis. How Great is the Great Firewall? Measuring China's DNS Censorship. In *Proceedings of the 30th USENIX Security Symposium*, USENIX Security '21, 2021.

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

[25] Graham Lowe, Patrick Winters, and Michael L. Marcus. The Great DNS Wall of China. Technical report, New York University, 2007.

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

[27] Arian Akhavan Niaki, Nguyen Phong Hoang, Phillipa Gill, Amir Houmansadr, et al. Triplet censors: Demystifying great firewall's DNS censorship behavior. In *10th USENIX Workshop on Free and Open Communications on the Internet, FOCI 20)*, 2020.

[28] Arian Akhavan Niaki, William R. Marczak, Sahand Farhoodi, Andrew McGregor, Phillipa Gill, and Nicholas Weaver. Cache me outside: A new look at DNS cache probing. In *Passive and Active Measurement - 22nd International Conference, PAM 2021, Virtual Event, March 29 - April 1, 2021, Proceedings*, 2021.

[29] OpenNet Initiative. Internet filtering in China in 2004-2005: A country study. https://opennet.net/sites/opennet.net/files/ONI_China_Country_Study.pdf, 2005.

[30] Paul Pearce, Roya Ensafi, Frank Li, Nick Feamster, and Vern Paxson. Augur: Internet-Wide Detection of Connectivity Disruptions. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

[31] Paul Pearce, Ben Jones, Frank Li, Roya Ensafi, Nick Feamster, Nick Weaver, and Vern Paxson. Global Measurement of DNS Manipulation. In *USENIX Security Symposium (USENIX)*, 2017.

[32] Jon Postel. Rfc 792-internet control message protocol, 1981. *URL http://tools. ietf. org/html/rfc792.–Zugriffsdatum*, 19, 2011.

[33] Ram Sundara Raman, Adrian Stoll, Jakub Dalek, Reethika Ramesh, Will Scott, and Roya Ensafi. Measuring the deployment of network censorship filters at global scale. In *NDSS*, 2020.

[34] Reethika Ramesh, Ram Sundara Raman, Matthew Bernhard, Victor Ongkowijaya, Leonid Evdokimov, Anne Edmundson, Steven Sprecher, Muhammad Ikram, and Roya Ensafi. Decentralized control: A case study of russia. In *Network and Distributed Systems Security (NDSS) Symposium 2020*, 2020.

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

[36] Will Scott, Thomas Anderson, Tadayoshi Kohno, and Arvind Krishnamurthy. Satellite: Joint analysis of cdns and network-level interference. USENIX Association, 2016.

[37] Ram Sundara Raman, Prerana Shenoy, Katharina Kohls, and Roya Ensafi. Censored planet: An internet-wide, longitudinal censorship observatory. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 49–66, 2020.

[38] Benjamin VanderSloot, Allison McDonald, Will Scott, J. Alex Halderman, and Roya Ensafi. Quack: Scalable remote measurement of application-layer censorship. In *27th USENIX Security Symposium (USENIX Security 18)*, 2018.

[39] Zhongjie Wang, Yue Cao, Zhiyun Qian, Chengyu Song, and Srikanth V Krishnamurthy. Your state is not mine: a closer look at evading stateful internet censorship. In *Proceedings of the 2017 Internet Measurement Conference*, pages 114–127, 2017.

[40] Zachary Weinberg, Diogo Barradas, and Nicolas Christin. Chinese wall or swiss cheese? keyword filtering in the great firewall of china. In *Proceedings of the Web Conference 2021*, 2021.

[41] Joss Wright. Regional variation in chinese internet filtering. *Information, Communication & Society*, 2014.

[42] Xueyang Xu, Z. Morley Mao, and J. Alex Halderman. Internet Censorship in China: Where Does the Filtering Occur? In *Workshop on Passive and Active Network Measurement (PAM)*, 2011.

A Examples

Table 4 contains examples of destination IPs and source parameter combinations that elicit differences in censorship results.

Destination	Source Parameters	Domain
1.116.30.255	x.x.x.36:12340, x.x.x.37:12340	twitter.com
101.224.206.29	x.x.x.171:41340, x.x.x.170:41340	twitter.com
103.45.149.117	x.x.x.52:12919, x.x.x.53:12919	twitter.com
106.12.43.188	x.x.x.165:20701, x.x.x.164:20701	twitter.com
122.68.118.163	x.x.x.171:24640, x.x.x.172:24640	twitter.com
123.83.136.8	x.x.x.144:36764, x.x.x.165:36764	twitter.com
202.112.57.19	x.x.x.67:54262, x.x.x.68:54262	csis.org
210.45.168.4	x.x.x.163:58461, x.x.x.164:58461	csis.org
36.96.222.132	x.x.x.162:24202, x.x.x.161:24202	bigthink.com
36.109.96.54	x.x.x.47:18532, x.x.x.46:18532	bigthink.com

Table 4: *Examples.* Set of example destination IPs and corresponding source parameters pairs that resulted in censorship differences as of February 2022.