On the Origin of Scanning: The Impact of Location on Internet-Wide Scans

Gerry Wan
Stanford University

Liz Izhikevich
Stanford University

David Adrian
Censys, Inc.

Katsunari Yoshioka
Yokohama National University

Ralph Holz
University of Twente
University of Sydney

Christian Rossow
CISPA Helmholtz Center for Information Security

Zakir Durumeric
Stanford University

ABSTRACT

Fast IPv4 scanning has enabled researchers to answer a wealth of security and networking questions. Yet, despite widespread use, there has been little validation of the methodology's accuracy, including whether a single scan provides sufficient coverage. In this paper, we analyze how scan origin affects the results of Internet-wide scans by completing three HTTP, HTTPS, and SSH scans from seven geographically and topologically diverse networks. We find that individual origins miss an average 1.6–8.4% of HTTP, 1.5–4.6% of HTTPS, and 8.3–18.2% of SSH hosts. We analyze why origins see different hosts, and show how permanent and temporary blocking, packet loss, geographic biases, and transient outages affect scan results. We discuss the implications for scanning and provide recommendations for future studies.

ACM Reference Format:
Gerry Wan, Liz Izhikevich, David Adrian, Katsunari Yoshioka, Ralph Holz, Christian Rossow, and Zakir Durumeric. 2020. On the Origin of Scanning: The Impact of Location on Internet-Wide Scans. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3419394.3424214

1 INTRODUCTION

Fast IPv4 scanning has become a standard measurement technique for understanding edge host behavior on the Internet. Popularized by tools like ZMap [22] and Masscan [24], Internet scanning has enabled hundreds of papers on service deployment [4, 18, 20, 31, 53, 57], outages [9, 16, 22, 32, 33, 47, 55], host liveness [7, 12, 27, 46, 56], security weaknesses [14, 45, 60], operator behavior [3, 19, 21, 23, 41], botnets [5, 39], and censorship [34, 49, 50], as well as helped uncover new vulnerabilities [6, 8, 13, 29]. Yet, despite the technique's recent popularity, there has been relatively little analysis of its accuracy and completeness.

In this paper, we quantify the coverage provided by single-probe Internet-wide IPv4 scans and investigate how the network used for conducting scans ("scan origin") affects their results. We complete three trials of synchronized HTTP, HTTPS, and SSH ZMap + ZGrab scans from five geographically and topologically diverse academic networks in Australia, Brazil, Germany, Japan, and the United States as well as from Censys [17] and Carinet (a popular cloud provider that permits scanning). We show that origins miss 1.6–8.4% of HTTP, 1.5–4.6% of HTTPS, and 8.3–18.2% of SSH hosts in a single-probe scan—about twice the loss originally estimated by Durumeric et al. [19, 22]. There are a confluence of factors that affect coverage, including regional access restrictions, intentional non-deterministic server behavior, dynamic blocking, extremely lossy links, and short-lived, localized outages.

Most inaccessible HTTP(S) hosts are missed transiently in only a single scan. Transient loss is unpredictable and highly variable, but is not simply due to random packet drop. In almost all cases when one probe is dropped, secondary probes are also lost. Factors like topological distance, peering relationships, and geographic boundaries are poor indicators for the transient inaccessibility that origins experience. Destination networks rarely have a "best" scan origin; in nearly one quarter of destination ASes, the scan origin that had the best coverage in one trial will have the worst coverage in the next, even for major providers like Google and Amazon. While it is typically difficult to explain why origins sometimes experience high lossiness, we uncover evidence of short-lived outages that affect only a subset of origins and account for 14–36% of transient loss. ZMap's retransmission scheme fails to account for most transient loss, but loss is easy to overcome by scanning from 2–3 sufficiently diverse vantage points, which achieves 98–99% coverage of HTTP(S) hosts and minimizes variance ($\sigma = 0.08\%$).

Network policies also bias the hosts that each origin can reach. Censys misses 2–13 times as many HTTP(S) hosts as our academic origins, far overshadowing the hosts that are transiently lost. Most of the hosts that Censys misses are in a small handful of large providers, but ISP decisions also block individual scanners from accessing large portions of some countries. For example, Censys is unable to reach 27% of hosts in South Africa and 43% of hosts in Bangladesh. Geographic restrictions also prevent origins from accessing all hosts. Just over 1% of Japanese and 2% of Australian HTTP servers are only accessible from within the country, and more than 100 American networks—primarily belonging to small financial, healthcare, and utility companies—are entirely inaccessible from Brazil. Regional policies do not materially affect global results, but can skew analyses of specific countries and industries.

IMC '20, October 27–29, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8138-3/20/10…$15.00
https://doi.org/10.1145/3419394.3424214

SSH scans miss five times as many hosts as HTTP(S) scans. We trace this discrepancy back to several large providers—most prominently Alibaba—that dynamically detect and block SSH scanners, as well as non-deterministic behavior in OpenSSH where servers will probabilistically drop sessions after detecting multiple unauthenticated connections. This protection prevents initial completion of an SSH handshake with most missing hosts, but can be easily detected and avoided with immediate retries.

Our results indicate that single-probe Internet-wide scans achieve lower global coverage than originally estimated (96.3% vs. 97.9% [22]). This result does not invalidate the methodology, and in most cases, the increased loss will not meaningfully change research results based on Internet scans. However, loss is not simply due to random packet drop, as was previously suggested. The differences in hosts and networks visible from different scan origins can bias studies that focus on specific geographic regions or types of networks, which researchers should consider when designing experiments. Most missing hosts are lost due to transient network problems, which are nearly impossible to predict, but if researchers need improved coverage, they can achieve this by scanning from 2–3 diverse origins, scanning with multiple probes with delay between probes to the same host, or performing multiple independent scans.

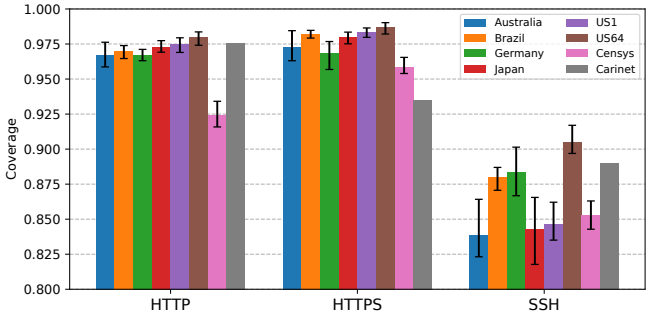


Figure 1: IPv4 host coverage by scan origin (2 probes)—Each origin sees a distinct set of hosts in each scan. On average, origins scanning SSH will see 10% fewer ground truth hosts compared to HTTP(S).

2 METHODOLOGY

To quantify the impact of network origin on Internet scan results, we performed nine synchronized IPv4 scans from seven geographically diverse networks. We specifically completed three trials of HTTP, HTTPS, and SSH ZMap + ZGrab scans of the full IPv4 address space from academic institutions in Australia (University of Sydney), Brazil (Universidade Federal de Minas Gerais), Germany (Max Planck Institute for Informatics), Japan (Yokohama National University), and the United States (Stanford University), as well as Censys [17]. For one trial, we also scanned from Carinet, a commercial cloud provider that Rapid7 uses for Project Sonar [51].

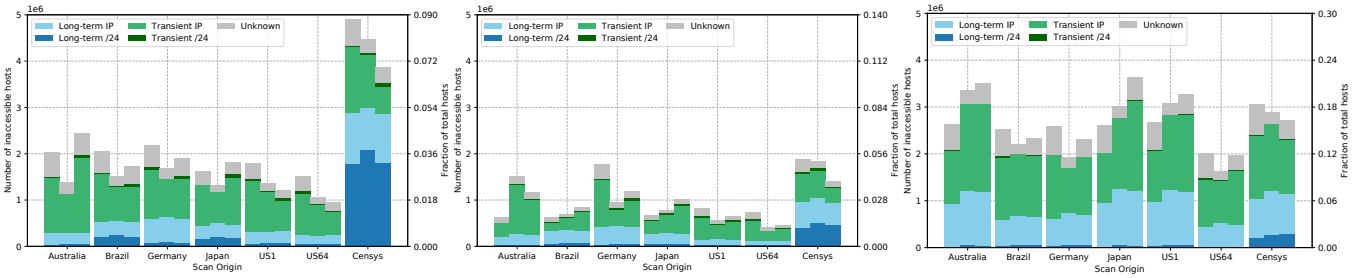
We scanned from a single source IP address from all locations except from Stanford University, where we performed two independent scans, one with 1 IP ("US_1") and one with a contiguous block of 64 IPs ("US_{64}"). We were unable to scan with more than one IP from other origins, but use the 64-IP origin to analyze the impact of multiple source addresses instead of multiple scan origins. These vantage points represent all continents except Africa and Antarctica, as well as academic, commercial, and cloud networks. Because we only scan from Carinet in one trial, we exclude the origin from aggregate statistics unless noted otherwise. We refer to origins by country name for simplicity, but we emphasize that origins are affected by a congruence of factors including geographic location, IP registration country, upstream provider, peering policies, as well as IP and network reputation.

Not all of our origins have clean scanning reputations. The Australian and German IPs have previously been used for individual scans. The U.S. IPs have never been used for scanning, but reside in a /24 network that commonly performs scans. The Japanese and Brazilian IPs, along with their respective /24s, have never been used for scanning. The Censys IP belongs to one of the company's research servers, not an operational server used for daily scanning, but it is part of the company's published IP block. None of the academic IP addresses appeared on any public blocklists. We have no history of the Carinet IP beyond confirming that it was not on any public blocklists prior to our study.

For each trial, we run a ZMap [22] TCP SYN scan and immediately complete a follow-up application layer handshake with L4-responsive hosts (i.e., hosts that respond with a SYN-ACK packet) using ZGrab [17]. On TCP/80, we complete an HTTP `GET /`; on TCP/443, we complete a TLS handshake using the TLS 1.2 cipher suites in modern Chrome; and on SSH, we complete a partial SSH handshake that terminates after the protocol version exchange. We choose these protocols because they are well-known TCP protocols that are frequently studied by researchers [19] and commonly used on the Internet [17]. We start each ZMap scan at the same time across all origins and use the same ZMap seed, which ensures that all scanners scan the same addresses at approximately the same time. Each ZMap scan sends two back-to-back SYN packets to every destination IP address, which prior work estimates will achieve 98.8% coverage of listening hosts [22].

To ensure that scanners do not fall out of sync due to mismatched hardware and to confirm that upstream networks can transit scans, we completed a series of ZMap scans targeting 1% of the IPv4 space in October 2019. These experiments confirmed that all origins can scan at 100K packets per second (pps) and that there is no increased packet drop above minimal scan speeds (i.e., 1,000 pps). We snapshot a routing table from our U.S. scan origin at the start of each trial to determine origin ASes. We use MaxMind GeoIP2 Lite [44] for IP geolocation.

We ran the final full experiments on October 21, November 20, and December 10, 2019. Each trial elapsed approximately 21 hours. The maximum asynchrony we found in L7-responses was 2 hours for HTTP, 15 minutes for HTTPS, and 12 minutes for SSH, which occurred at the end of the trials when our Australia and Brazil scanners fell behind the others. This appears to be due to slight differences in server resources across origins as well as differences in the number of hosts that timed out (and thus missed) for each origin. Because connection timeouts require more time to finish than a normal handshake, scanners that see more timeouts fall behind other scanners.



(a) HTTP (b) HTTPS (c) SSH

Figure 2: Breakdown of missing hosts by scan origin and trial—Censys is long-term inaccessible from the largest number of hosts across all protocols. For other origins, transient loss accounts for the majority of missing hosts.

Limitations. There is no known ground truth for live Internet hosts. We estimate "ground truth" as the set of hosts that successfully complete an application-layer handshake with any scan origin in a given trial. We limit our analyses to hosts that complete an L7 handshake to reduce the impact that firewalls, middleboxes, and DDoS protections have on our results. We acknowledge that we are unable to detect hosts that are inaccessible from all of our scan origins, but we are limited to organizations that allow Internet-scanning and further wish to minimize the number of probes that destination hosts simultaneously receive. We choose to scan from a single academic location on each continent, as well as Censys and Carinet because the research community relies on their published data. As we show in Section 6, even this limited number of probes may cause some hosts to drop or reset connections.

Our experiments are limited to three trials of each protocol, spread over eight weeks, which may amplify noise caused by temporal churn. This also precludes certain longitudinal statistical analyses, but we believe our current data sufficiently demonstrates the biases that arise from single origin scans. In addition, because we do not have access to edge hosts nor the exact paths taken by our probes, we are often only able to hypothesize the precise root cause for inaccessible hosts. There are also inherent biases to scanning from academic networks, but we focus this work on examining the differences that arise from diverse scan origins that resemble what researchers are likely to use in their own studies. Most cloud providers do not allow scanning and many researchers use their academic networks when conducting experiments [19].

Ethical Considerations. We take several steps to minimize the impact of our experiments, as well as follow the best practices set forth by Durumeric et al. [22]. We limit scans to a single perspective on each continent and limit additional origins to those that researchers commonly rely on. We focus on a small number of protocols that researchers frequently study, and we use scanning tools that have been tested and repeatedly used by prior studies. In all cases, our scanners follow protocol specifications, and we immediately close connections once a handshake completes. We configured an HTTP page on scan hosts to redirect to a single website that explained our study. Rather than scanning at full speed, we limited each scan to 100K pps from each origin and spread our experiments over several weeks. We also synchronized blocklists by combining the IP ranges that previously requested exclusion from any scan origin. This resulted in the exclusion of 17.8M IPv4 addresses (0.5% of public IPv4) from the study. During the course of our study, we received exclusion requests from 9 organizations, which we immediately honored and removed from analysis. The data used in this paper will be posted to the `Scans.io` Internet-Wide Scan Data Repository.

3 RESULTS SUMMARY

Every scan origin discovers a distinct set of hosts, as can be seen in Figure 1. The six academic origins each see an average 97.2% of HTTP(S) hosts while Censys sees only 92.5% of HTTP(S) hosts. Surprisingly, Censys sees about the same number of SSH hosts as Australia, Japan, and US_1. No single origin consistently has the best coverage across all trials and no single origin achieves greater coverage than 98% of HTTP, 99% of HTTPS, or 92% of SSH hosts in any trial. We show the detailed breakdown of results by trial in Appendix A.

To verify that there is a meaningful difference between scan origins, we compare the number of hosts seen (and not seen) by each pair of origins per protocol using McNemar's test and find statistically significant differences ($p < 0.001$) between all pairs of scan origins in all trials. We choose multiple paired tests over Cochran's Q test (the k-group extension of McNemar's) since a single differing origin can produce a statistically significant result in the latter. We apply a Bonferroni correction to account for multiple analyses on the same data set. The visible variation across protocols and the clear differences between Censys, US_{64}, and the academic scanners in Figure 1 further suggest that there are a multitude of reasons beyond random packet loss that contribute to the observed differences. As we will show in the next three sections, these include blocking and firewalls, geographic routing policies, and transient burst outages.

To better understand the factors that affect scan results, we separate missing hosts across two dimensions: long-term versus transient and host versus network behavior. Long-term inaccessible hosts and networks are likely lost due to firewalls or other filtering behavior (e.g., networks that have blocked an origin or limits access to specific geographic regions), or to a persistent lack of connectivity between the origin and destination network. Hosts and networks may be transiently inaccessible due to packet loss, temporary routing issues, real-time scan blocking, or other transient network outages. We analyze these categories separately because they have different root causes and impacts on our results.

We consider a host transiently inaccessible from a scan origin when (1) the host was inaccessible from the origin but accessible (i.e., successfully completes an application-layer handshake) from a different origin in the same trial, and (2) the host was accessible from the original scan origin in another trial. Hosts inaccessible from a scan origin for all three trials are long-term inaccessible. We label hosts present in only one trial as unknown since it is unclear whether this is due to a transient issue, a long-term change, or if the host went offline.

We further split missing hosts into networks and individual IPs. We aggregate ground truth IPs by /24 and calculate the fraction of hosts in each /24 that are accessible, transiently inaccessible, long-term inaccessible, or unknown for each origin. We require that a /24 have at least two ground truth hosts with consistent behavior to be considered as a single unit in order to avoid attributing issues that affect a sporadic host to those that affect an entire network. We choose /24s as the unit to analyze as they are the smallest publicly routable network and are often administered by the same entity [27]. We acknowledge that our methodology does not capture whether a policy is enforced within the network or on all edge hosts on the network, but we argue the policy remains a network-level decision in either case.

Transient issues account for just over half (51.6%) of missing hosts and nearly always affect individual hosts rather than entire networks (49.7% vs. 1.9%), as shown in Figure 2. One third of missing hosts are missing long-term; the remainder are unknown. By definition, the number of long-term inaccessible hosts remains relatively stable across trials. Small variations arise due to hosts not being seen by any origin in a trial. For transiently missing hosts, not only do the missing hosts themselves differ across trials, but transient loss rates also differ across both origins and trials. The largest temporal change occurs between HTTPS trials 1 and 2 for Australia (+275%). We discuss transient differences in Section 5.

4 LONG-TERM INACCESSIBILITY

A significant fraction of the differences in coverage between origins are due to long-term inaccessible hosts: 4M HTTP (6.8%), 1.7M HTTPS (4.1%), and 3.1M SSH (16%) hosts are inaccessible in all three trials from at least one origin. 92% and 34% of long-term inaccessible HTTP(S) and SSH hosts are unresponsive at Layer 4. Much of this is due to Censys, which sees five times more long-term HTTP(S) inaccessibility than the other origins. While blocking is undoubtedly a major component, there appear to be other factors as well. The two origins that have never conducted prior Internet scans (Brazil and Japan) have nearly double the long-term inaccessible HTTP(S) hosts than three origins that regularly perform scans from their subnet (Australia, US_1, and US_{64}).

Excluding Censys, about half (47%) of long-term inaccessible hosts are inaccessible from only one origin (Figure 3), but there are significant differences between academic origins. For example, Germany exclusively misses over three times as many HTTP(S) hosts as the other academic origins (Table 1). We also find that 5–10% of inaccessible hosts are exclusively accessible from a single origin. Australia and Japan each see more than twice as many exclusively accessible HTTP hosts as other origins. In this section, we focus on describing why different origins can only access a subset of Internet hosts on HTTP and HTTPS, and highlight SSH behavior separately in Section 6.

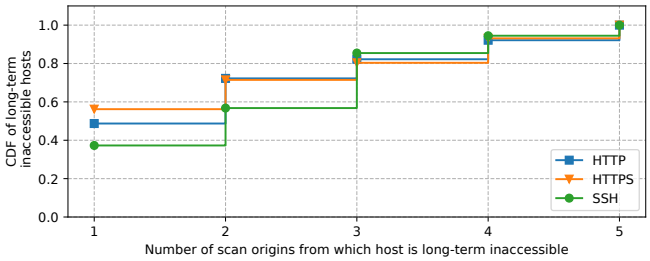


Figure 3: Long-term inaccessibility among origins—Excluding Censys, nearly half of long-term inaccessible hosts are inaccessible from only one origin.

	AU	BR	DE	JP	US_1	US_{64}	CEN
Acc. HTTP%	23.6	10.2	6.8	20.3	1.8	33.8	3.6
Acc. HTTPS%	11.7	13.7	18.1	18.1	0.3	30.8	7.3
Acc. SSH%	10.0	4.4	8.0	8.2	1.0	64.4	4.1
Inacc. HTTP%	1.2	2.9	8.9	2.1	0.9	0.6	83.4
Inacc. HTTPS%	2.1	8.0	15.9	3.7	0.9	0.4	68.9
Inacc. SSH%	10.9	7.9	14.4	10.6	12.9	6.6	36.7

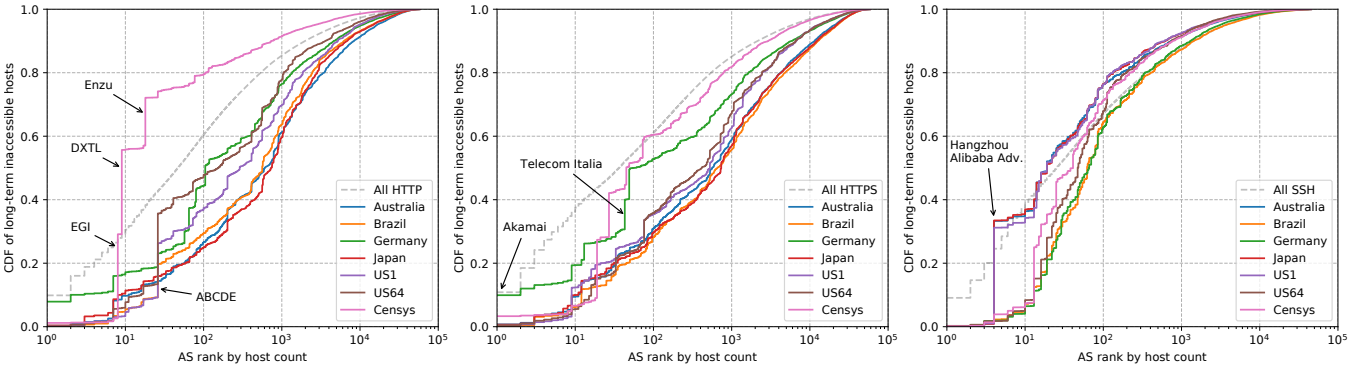
Table 1: Breakdown of origins responsible for hosts exclusively (in)accessible from a single origin—US_{64} sees the most exclusively accessible hosts while Censys has the most exclusively inaccessible hosts across all protocols.

4.1 Censys Blocking

Hosts that are inaccessible from Censys account for 83% of HTTP, 69% of HTTPS, and 37% of SSH hosts that are long-term inaccessible from a single scan origin (Table 1). Censys is also the only origin where long-term missing IPs primarily belong to fully inaccessible networks rather than individual hosts. This is not surprising—Censys scans *significantly* more than the other origins (at least 106 times more frequently in the past 6 months), and we expect some operators to block Censys. In total, 2.9M HTTP (5%), 1M HTTPS (2.4%), and 1.1M (5.6%) SSH hosts are long-term inaccessible to Censys.

The bulk of Censys' long-term inaccessible hosts belong to a handful of ASes (Figure 4). For HTTP, three hosting providers (DXTL Tseung Kwan O Service, EGI, and Enzu) account for 67% of inaccessible hosts but less than 4% of global HTTP hosts. HTTPS is similar with 38% of long-term inaccessible hosts belonging to the same three ASes, despite accounting for 1.0% of all HTTPS hosts. More than 99.99% of hosts from DXTL Tseung Kwan O Service and Enzu were inaccessible from Censys in all trials, while 90% of EGI hosts were long-term inaccessible in trial 1, but became completely and exclusively inaccessible by the third trial. Excluding these top three ASes, Censys still persistently misses 1.5 times as many HTTP hosts as the second-worst origin (Germany), and 1.4 times as many HTTPS hosts.

While this is significantly more blocking than Durumeric et al. previously estimated [19], Censys performs continuous scanning, and the difference could be attributable to the business decisions of



(a) HTTP (b) HTTPS (c) SSH

Figure 4: Distribution of long-term inaccessible hosts by AS relative to ground truth—Three ASes account for 67% of the long-term inaccessible HTTP hosts from Censys. In general, long-term inaccessible hosts are more evenly distributed across ASes for other origins, with several exceptions labeled.

only a handful of providers. It does, however, show that IP reputation and previous behavior can have a significant impact on scan results, overshadowing the differences due to transient loss that we discuss in Section 5. Since the time of our initial study, Censys has changed and increased the IP ranges that they use for scanning.

4.2 Academic Visibility

The single-IP academic origins in our study consistently miss an average 0.68% of HTTP(S) hosts and 4.4% of SSH hosts. Germany misses 1.1–3.6 times as many HTTP(S) hosts as the other academic origins; about 40% of these are exclusively inaccessible to Germany and belong to Telecom Italia, Telecom Italia Sparkle, and Akamai (Figure 4b). Though less than 1% of Akamai IPs are inaccessible to Germany, 36% of Telecom Italia and 46% of Telecom Italia Sparkle are long-term inaccessible, 85% of which are exclusively inaccessible. Using the packet loss metric described later in Section 5.2, we discover *extremely* high packet loss rates (over 40%), which suggests that Germany experiences a persistent lack of connectivity to these destination networks rather than explicit blocking.

Japan and Brazil are long-term inaccessible from more HTTP(S) hosts than US_1 and Australia, despite never having conducted full Internet-wide scans before (Figure 2). Surprisingly, Brazil and Japan are more than twice as likely to both miss the same /24 than either of the individual origins alone. About 70% of the hosts that are a part of the /24s inaccessible from both Brazil and Japan geolocate to Eastern Europe, resulting in 1.4% of Russia, 12.2% of Estonia, and 3% of Ukraine and Romania being long-term inaccessible from either origin. The Eastern-European ASes responsible appear to all be hosting companies or ISPs (e.g., SantaPlus). It is unclear why both countries are blocked by these providers.

Brazil loses the most entire ASes: nearly 1.4 times as many ASes as Censys and 6.5 times as many as US_1 (Figure 5). About half of the networks that block only Brazil are American health or financial companies. This contrasts Censys, where 40% of blocked networks are government owned and 22% are consumer businesses such as Jack-in-the-Box (AS 46603). American businesses may block Brazil because of the high number of Mirai infections in the country [5].

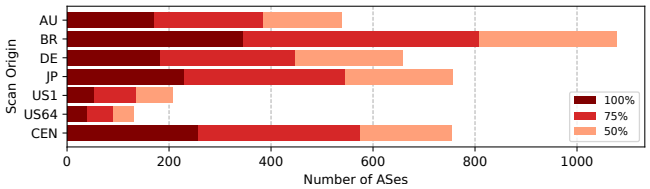


Figure 5: Long-term inaccessible ASes—Brazil suffers from the largest number of completely (100%) inaccessible ASes. We also show the number of ASes that are at least 75% and 50% inaccessible.

We find no obvious relationship (e.g., shared upstream peers, owners) among the ASes which only block Brazil. However, we do find that 14% of networks that block all non-US origins are owned by Tegna Inc., an American digital media company.

US_1 and US_{64} are particularly affected by ABCDE Group Company Limited (AS 133201, a large cloud provider in Hong Kong), in which 56K hosts account for 17% and 22% of long-term inaccessible HTTP hosts for US_1 and US_{64}. Interestingly, the same 56K hosts are also inaccessible from Brazil and Censys, albeit at proportionally smaller fractions of their inaccessible hosts. However, in general, long-term inaccessible hosts for academic origins are more evenly spread among ASes than for Censys.

4.3 Increased US_{64} Visibility

US_{64} consistently has the lowest number of long-term inaccessible hosts because several networks block all other scan origins. It is clear from Table 1 that at equal scan rates, scanning with 64 IPs has advantages over using just one. US_{64} exclusively sees 1.5 times more HTTP(S) and 6 times more SSH hosts than single-IP origins do, and consistently has the lowest number of long-term inaccessible hosts on all protocols. We manually investigate the largest ASes with hosts only accessible from US_{64} and uncover evidence of intrusion detection systems that detect and block IPs with high scan rates. For example, hosts from Ruhr-Universität Bochum (AS 29484) were accessible from all origins for the first 2 hours of the trial 1 HTTPS scan (the first full scan we conducted), but afterwards only US_{64} had visibility into the network in all of our later scans. We confirmed this behavior with network administrators at the institution. We

	>1M Hosts					>100K Hosts					>10K Hosts					>1K Hosts				
	HK	US	GB	CN	RU	ZA	AR	IT	AT	VE	BD	EC	AM	EE	AL	BF	LY	MN	MW	SD
AU	0.3	0.5	0.4	0.3	0.4	0.4	0.5	1.3	0.8	0.4	0.5	1.0	0.6	0.1	0.6	4.6	1.0	0.2	7.8	2.8
BR	2.3	0.8	0.8	0.4	1.8	0.7	0.2	1.3	7.9	0.3	0.4	0.7	0.1	12.2	9.9	4.5	0.2	0.1	7.2	1.1
DE	0.9	0.8	0.7	0.4	0.3	0.7	9.9	9.7	0.7	7.7	3.0	10.2	12.5	0.0	1.1	3.4	34.1	0.1	1.6	26.9
JP	0.2	0.7	0.9	1.0	1.9	0.3	0.6	1.1	7.7	0.4	3.2	3.4	0.2	12.2	10.0	37.9	0.7	0.4	28.6	0.6
US_1	2.3	0.4	0.2	0.5	0.3	0.5	0.7	0.8	0.6	0.9	1.3	7.7	0.2	0.1	1.1	38.0	3.2	0.4	28.6	1.9
US_{64}	2.2	0.3	0.3	0.4	0.2	0.4	0.3	0.3	0.4	0.8	0.2	5.2	0.1	0.1	0.4	1.7	3.0	0.1	1.9	2.0
CEN	9.8	7.3	2.6	2.3	1.7	27.0	5.2	6.2	1.6	2.9	42.9	17.3	0.3	0.4	5.9	37.7	16.1	30.4	28.7	13.4

Table 2: Countries with the most long-term inaccessible HTTP hosts—Coverage of countries can be greatly influenced by scan origin, but a significant fraction of missing hosts are often due to a handful of major ASes; red indicates one AS accounts for the majority of inaccessible hosts, orange two, and yellow at least three.

also observe similar behavior from SK Broadband (AS 9318), which accounts for over half of the SSH hosts that are exclusively accessible from US_{64}. In Section 7, we discuss whether scanners should use multiple source IPs.

4.4 Geographic Biases

The countries with the greatest number of hosts broadly account for the largest portion of long-term inaccessible hosts from any origin, simply due to their raw host count. We observe a high Spearman's rank correlation (ρ=0.92, p<0.001) between the total number of hosts and the number of inaccessible hosts in each country (and therefore also the percentage of total inaccessible hosts). Despite this overall trend, coverage of individual countries can be greatly influenced by scan origin, especially for countries with fewer hosts. In 50 countries (or dependent territories), more than 10% of their HTTP, HTTPS, or SSH hosts are long-term inaccessible from a scan origin, and in 19 countries, more than 25% of their hosts are inaccessible. Nearly all countries where a scan origin misses a significant fraction of hosts are composed of only a single or small handful of major ASes (Table 2). Indeed, there is only one country (Libya) where more than 30% of hosts are inaccessible and the majority of hosts are not hosted by a single ISP. In the most severe cases, 43% of hosts in Bangladesh and 27% of hosts in South Africa are consistently inaccessible from Censys. In both countries, this is primarily due to *DXTL Tseung Kwan O Service* blocking Censys.

While there is no clear pattern between the origins that have the best coverage of destination countries, we do find that origins typically have better coverage of hosts within the same country than external origins do, albeit by an arguably insignificant amount relative to the number of global hosts. We exclude US_{64} and combine US_1 and Censys to analyze the effects of geographic proximity on exclusive access of hosts in Figure 6. In Japan, about 1.1% of all HTTP hosts are only accessible from within the country. 40% are located in Bekkoame Internet (Figure 7), a Japanese hosting provider with 0.9% of all HTTP hosts. NTT accounts for the second most with 29% of the exclusively accessible hosts and 11% of all HTTP hosts in Japan. The long tail of other ASes is composed of various Japanese cloud/hosting providers and university networks. Interestingly, the United States contains the second highest number of hosts exclusively accessible from Japan; however, 40% of these hosts belong to Gateway Inc. (AS 132827), a hosting provider registered in Japan. This suggests regional restrictions that only allow access from specific locations.

About 2% of Australian HTTP hosts are only accessible from within Australia. Just over 80% of those are served by WebCentral (AS 7496), a Sydney-based digital agency that is the ninth largest AS in Australia by HTTP host count. While the vast majority of hosts that are exclusively accessible from Japan geolocate to Japan, only half of the hosts exclusively accessible to Australia geolocate there (85% vs. 48%). Just under half geolocate to the U.S., Germany, Great Britain, Netherlands, and France. We suspect that these may be geolocation inaccuracies: 92% of the hosts exclusively accessible to Australia but geolocate to a different country belong to Cloudflare, which confirmed that the IPs are advertised via anycast. Cloudflare also confirmed that these hosts accessible to only Australia are a misconfiguration, which was resolved after our report. Curiously, most hosts accessible exclusively from Brazil are in the United States. About two-thirds of these belong to *WA K-20 Telecommunications Network*, an educational ISP in Washington State. The hosts serve Brazil an HTTP page titled "Blocked Site" but consistently drop connections from other scan origins.

We see no regional correlation beyond countries (e.g., Japan does not see more hosts in Asia than other origins). While exclusive accessibility from another origin can account for up to 20% of the long-term inaccessible hosts for any specific origin, this does not significantly impact the results of global scans. On average, only 0.17% of all HTTP hosts are exclusively accessible from a single scan origin. We find similar regional access limits for HTTPS and SSH, although to a lesser extent than HTTP. The analogs of Table 2 and Figure 6 can be found in Appendix B.

4.5 Summary

One third of missing hosts are long-term inaccessible. Much of this is due to Censys, which experiences five times as much long-term HTTP(S) inaccessibility as other perspectives. For Censys, blocking overshadows transiently missed hosts, but for other single-IP perspectives, long-term inaccessibility is a relatively minor problem that affects only an average 0.68% of all HTTP(S) hosts. However, we note that while only a small fraction of global hosts are missed, hosts are not uniformly inaccessible. The decisions of a small number of ISPs can cause scan origins to meaningfully lose coverage of entire countries (e.g., 43% of hosts in Bangladesh are inaccessible

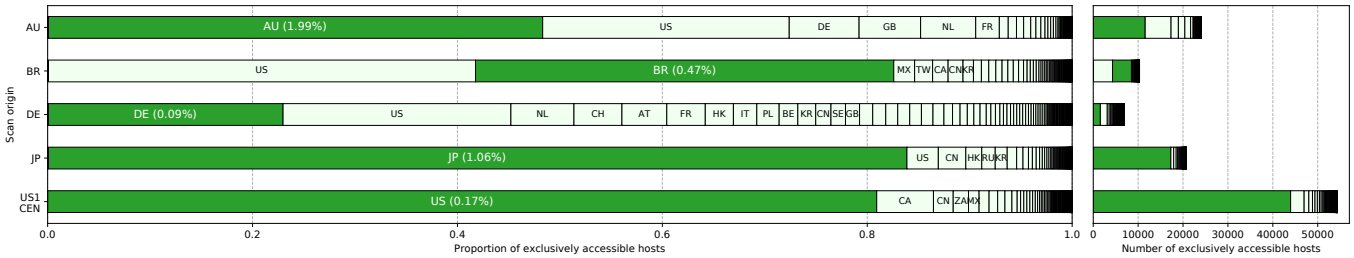


Figure 6: Exclusively accessible HTTP hosts by country—Origins within a country typically have better accessibility than external origins do. Dark green indicates hosts that are only accessible by scanning from within the country. For these, we additionally show the fraction of that country's total hosts that are exclusively accessible from within the country.

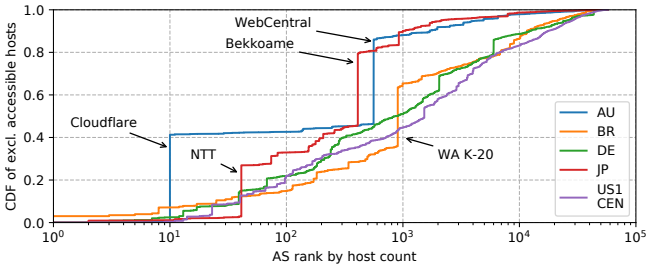


Figure 7: AS Distribution of exclusively accessible HTTP hosts—We identify the ASes that account for the largest fraction of hosts that are exclusively accessible from a single origin country.

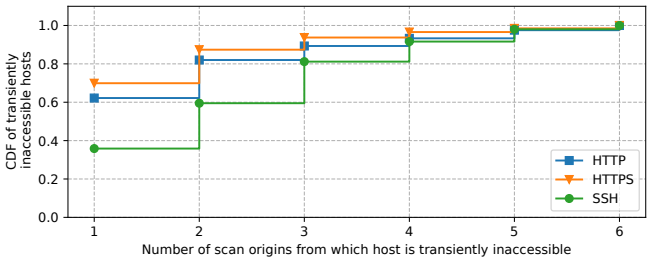


Figure 8: Transient inaccessibility among origins—Nearly half of transiently inaccessible HTTP(S) hosts are missed by only one origin. SSH hosts are more likely to be missed by more than one.

from Censys), and in some regions, certain sites are only accessible from within the country. US_{64} consistently misses the fewest hosts because its lower scan rate per IP address prevents it from being automatically blocked by destination networks, and may be a technique that helps researchers maintain visibility.

5 TRANSIENT INACCESSIBILITY

The majority of missing hosts are lost transiently (i.e., in some but not all trials). This short-term loss results in origins missing an average 1.4% of all HTTP(S) and 7% of all SSH hosts for double probe scans. We simulate scanning with one probe by requiring successful responses to both of our ZMap probes, and estimate that origins miss 2.7% of HTTP(S) and 8.3% of SSH hosts in single probe scans. As can be seen in Figure 2, there is considerable variance in transient loss across both trials and perspectives. For example, Germany sees 5.3 times more transient loss than Brazil in HTTPS trial 1, and Australia sees a 2.75 times increase in HTTPS loss between trials 1 and 2.

Two thirds of transiently inaccessible HTTP(S) hosts are missed by only one scan origin (Figure 8). For about 40% of destination ASes, the difference in host coverage between any two origins is greater than 1%, and for 16–25% of ASes, depending on protocol, the difference is greater than 10% (Figure 9). This loss affects all sizes of networks and nearly 25% of transiently missed hosts are from the 200 largest networks where some origins miss tens of thousands of hosts in a single AS. We show the ASes with the greatest transient differences across origins in Table 3; all are within the top 100 ASes by host count. Beyond the top five ASes, a notable fraction of affected networks are Chinese, which is consistent with prior work that has shown that packet loss on paths to China is unusually high and unstable [63].

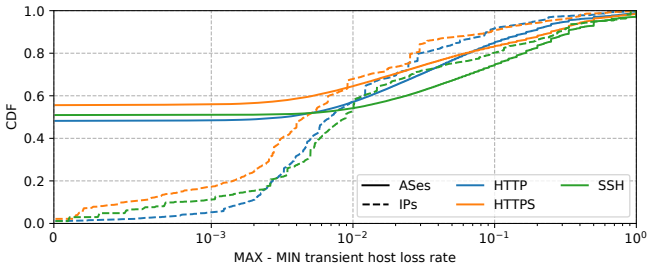


Figure 9: Distribution of differences in transient loss rate among origins—Transient loss rates from different origins are identical for half of all destination ASes, while loss rates can differ by more than 10% for about 20% of ASes. The dashed line shows the CDF weighted by AS size.

No scan origin consistently has the most or least transiently-missed hosts, and 96% of transient inaccessibility is due to missing individual hosts rather than entire /24 networks. There is no single best or worst scan origin, but we do find that origins have distinct characteristics in how and where they transiently lose hosts. Packet loss alone does not account for the variability we observe. We also find that real-time scan detection and blocking, probabilistic blocking, burst outages, and other transient connectivity problems also affect scan results. We note that a significantly larger proportion of SSH hosts are transiently missed, which we discuss separately in Section 6.

AS	Δ(%)	Diff	Ratio
ABCDE Group Co. (HK)	62.1	144K	136
HZ Alibaba Adv. (CN)	7.8	128K	8.1
Akamai (US)	2.2	126K	15.0
Psychz Networks (US)	15.6	71K	21.2
Telecom IT. Sparkle (IT)	77.0	58K	2,167
Telecom IT. (IT)	13.3	53K	60.7

(a) HTTP

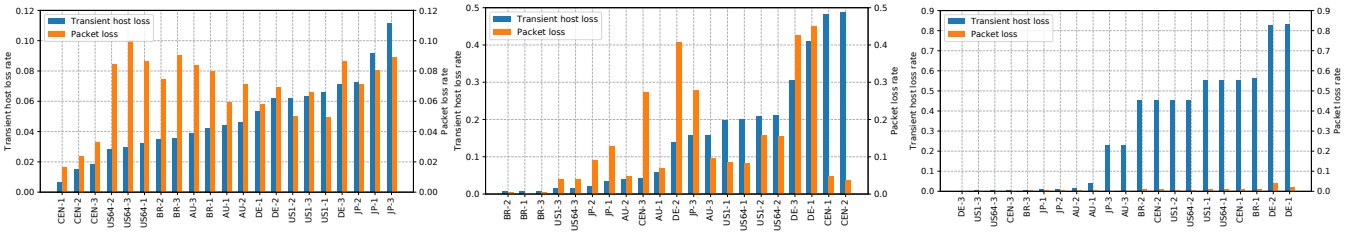
AS	Δ(%)	Diff	Ratio
HZ Alibaba Adv. (CN)	20.5	145K	68
Akamai (US)	2.1	97K	37.8
Telecom IT. (IT)	53.7	57K	137
Telecom IT. Sparkle (IT)	66.7	51K	2,929
Tencent (CN)	25.9	43K	13.4
China Telecom (CN)	18.0	40K	8.8

(b) HTTPS

AS	Δ(%)	Diff	Ratio
HZ Alibaba Adv. (CN)	20.5	145K	68
Akamai (US)	2.1	97K	37.8
Telecom IT. (IT)	53.7	57K	137
Telecom IT. Sparkle (IT)	66.7	51K	2,929
Tencent (CN)	25.9	43K	13.4
China Telecom (CN)	18.0	40K	8.8

(c) SSH

Table 3: ASes with the largest range of transient host loss rates—Large ASes in China and Italy are the most likely to cause different scanning origins to perceive significantly different transient host loss.



(a) Hangzhou Alibaba Adv. (AS 37963, CN) **(b) Telecom Italia (AS 3269, IT)** **(c) ABCDE Group (AS 133201, HK)**

Figure 10: Transient host vs. packet loss—The fraction of transiently missed HTTP hosts and the estimated packet loss rates are not correlated in ASes for which there is a wide range of transient host loss perceived by different origins.

5.1 Origin-Stability of Transient Loss

The origin with the best or the worst coverage of a destination AS is highly variable and changes between trials. For about 23% of ASes, the worst scanning origin in one trial will become the best scanning origin in another trial, or vice versa. ASes with dramatic changes are not limited to small organizations where a few dropped packets can have a disproportionate impact. The three largest ASes where the best origin drops to worst for HTTP belong to Amazon, Digital Ocean, and Google. We show ABCDE Group (a large hosting provider in Hong Kong and the fifth largest AS where the best origin flips to worst) in Figure 10c.

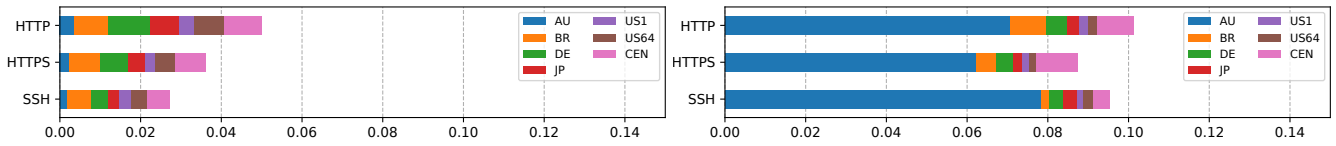
Fewer than 5% of ASes have a consistent best origin across all trials. There is no consistent geographic relationship between the origins that consistently provide the best coverage. There is similarly little correlation between the origins with the least packet drop and best coverage, likely because packet loss rates between origins and best-origin-consistent ASes tend to be very low (<0.5%) and random noise can change rankings. It is nearly impossible to predict which origin will have the best transient coverage of any destination network.

On the other hand, 10% of ASes have a consistently worst scan origin. Australia is the worst origin for 72% of ASes that have a consistent worst origin, with most lost hosts geolocating to Russia and the United States (Figure 11b). Surprisingly, about half of all Russian hosts belong to networks where Australia consistently sees the least. For HTTP(S), over 90% of Kazakh hosts are also consistently the worst seen from Australia. We emphasize that this does not describe all hosts in the country that are inaccessible from Australia; rather, they represent the fraction of hosts in the country that are consistently the most likely to be missed by Australia. In the ASes where Australia consistently had the highest transient loss, the average packet loss rate was more than ten times larger than the second worst origin. A similar pattern emerges for the countries that consistently have the worst coverage from Australia. For example, Australia saw an average 4.1% packet drop rate to affected Russian ASes while the next most lossy origin saw only 0.44% drop; Kazakhstan saw 4.6% (Australia) vs. 0.39% (second worst). This could be caused by a consistently congested path between Australia and these networks, but we are unable to pinpoint where in the path.

5.2 Impact of Packet Loss

Fast Internet-wide scanning is inordinately affected by packet drop since scanners like ZMap [22] cannot distinguish between unresponsive hosts and dropped probe packets. We estimate random packet drop by counting the number of hosts that receive one versus two of the ZMap probes. To reduce the effects of middleboxes and hosts that deviate from the TCP protocol, we exclude RST packets, ignore duplicate responses, and restrict our analysis to hosts that complete an L7 handshake with at least one origin during the trial. This provides only a lower bound on packet drop because it excludes cases where both probes are lost, but we cannot reliably determine whether this is due to packet drop.

Globally, we observe packet drop rates between 0.44–1.6%, depending on trial and origin. Australia has the highest packet loss, which is unsurprising given that it is the origin with consistently worst connectivity to the largest number of ASes (Figure 11b). However, there is only a weak correlation between the ASes with high packet drop and the ASes with high transient loss within each origin (Spearman's $\rho = 0.40–0.52$, $p < 0.001$). In 55% of ASes with HTTP hosts and 43% of ASes with HTTPS and/or SSH hosts, there is no statistically significant relationship within the AS between origins that experience the most packet loss and transient host loss. Because



(a) Best-origin-consistent **(b) Worst-origin-consistent**

Figure 11: Consistent best and worst scan origins relative to destination ASes — Less than 5% of ASes have a consistent scan origin with minimum transient loss. Australia is most often the scan origin with consistent maximum transient loss.

we only estimate a lower bound on drop rates and cannot assume independence of drops [48], we are unable to give a reasonable estimate for the proportion of transiently missed hosts caused by packet loss without severely underestimating its true value. While random packet loss undoubtedly contributes to transient host loss, it is not sufficient to explain the variation in missed hosts among origins except in extreme cases where origin(s) see unusually high or low drop rates.

We observe significant packet loss to China from all origins, consistent with prior work (3–14% vs. 5–15% [63]), but it is not the only cause of transient inaccessibility. For example, there is a stable rank ordering of origins with the best visibility of Alibaba, but there is no meaningful correlation between those origins and the origins with the lowest packet drop (Spearman's $\rho = 0.18$, $p = 0.44$, Figure 10a). Contrary to prior work [63], Japan does not have lower packet loss rates despite its proximity to China. There is also significant packet loss to Telecom Italia (AS 3269) from all origins except from Brazil (μ = 16% vs. 0.3%, Figure 10b). Telecom Italia and Telecom Italia Sparkle are the two largest ASes where Brazil has consistent best coverage, likely because TIM Brasil is a subsidiary of Telecom Italia [62]. Germany has exceptionally high loss rates to these two ASes. Censys has high transient host loss and low packet loss in the first two trials, but flips to low host loss and high packet loss in the third trial.

5.3 Burst Outages

Beyond random packet drop, localized temporary outages also cause transient loss. To quantify how many hosts are lost due to burst events, we analyze the number of transiently lost hosts per hour for every origin–destination AS pair and look for bursts of inaccessibility. We choose an hour granularity as it is the smallest logical time frame in which we would expect to see an average sized AS (≈1,000 hosts) experiencing random uniform packet loss to lose more than one host per hour. We identify statistically significant bursts of transiently missing hosts by searching for outliers in the noise-component of the time series that are two standard deviations away from the average expected noise. To extract the noise component, we subtract the smoothed time series — obtained by a rolling window, which on average minimizes the average mean square error (i.e., 4 hours) — from the original time series.

We find that 14–36% of transient loss, depending upon the protocol, trial, and origin, coincides with a burst outage. Across protocols and trials, there is no consistent origin which experiences the largest or smallest fraction of hosts lost due to transient bursts. An example of significant bursty loss occurs for Brazil during HTTPS trial 3, in which 8% of all transiently-missing hosts are lost in a single hour, affecting 39% of scanned ASes, including Akamai and Amazon. In general, 45% of destination ASes which contain at least one transiently missing host across all protocols and trials experience at least one transient burst loss that can be detected at the hour granularity. The majority (roughly 60% for all protocols) of transient bursts within a destination AS at a given hour occur for just one origin and at least 91% of transient bursts occur simultaneously for three origins or fewer. Across protocols, Australia is always the most likely to be the single scan origin that experiences a burst loss event, accounting for 30–40% of single origin burst outages. There is no temporal pattern of when these occur.

We also analyzed whether scan origins see variable coverage based on the local time that scans were performed (e.g., do any origins see decreased packet drop or increased coverage at night?). We did not observe any consistent pattern for any of our origins.

5.4 Summary

Most missing hosts are lost transiently in a subset of trials. Transient inaccessibility is inconsistent and unpredictable, shifting dramatically between trials, even for large providers like Google and Amazon. It is nearly impossible to predict which origin will have the best coverage of a destination network — scanning closer to a network does not improve visibility. While few destination networks have a consistent best origin, when there is a consistent worst origin, this is nearly always Australia and is due to extreme packet drop. Broadly, however, transient loss is not entirely attributable to simple random packet drop — except in extreme cases — and the networks with the worst visibility often have the lowest random packet drop rates. We discuss the impact of scanning from multiple origins on transient loss in Section 7.

6 SSH BEHAVIOR

While HTTP and HTTPS exhibit similar behavior, SSH has a unique dynamic. Scan origins see 10% fewer SSH hosts than HTTP(S) (Figure 1), experience five times more transient and long-term loss (Figure 2), and are less likely to be the sole origin that misses a particular SSH host (Figure 3, 8). As we discuss in this section, these differences are due to security protections specific to SSH.

Nearly 40% of long-term inaccessible SSH hosts for Australia, Japan, and US_1, and 24% of transiently inaccessible hosts for Censys, Germany, and Brazil are hosted by Alibaba (AS 37963, 45102). Alibaba appears to detect single-IP scans two-thirds of the way into trial 1 and immediately blocks the origins (Figure 12). Alibaba's scan detection is non-deterministic and blocks origins at different times across all three trials. Notably, the network-wide blocking behavior causes SSH hosts to respond with a RST immediately after completing a TCP handshake. Alibaba is the only network that

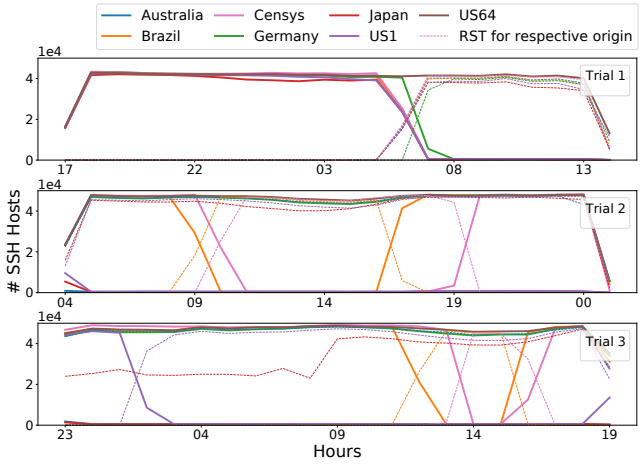


Figure 12: Temporal Blocking by SSH hosts in Alibaba Networks—Across all origins using only one source IP, Alibaba intermittently detects scanning and thus causes all SSH hosts to RST the connection after completing a TCP handshake.

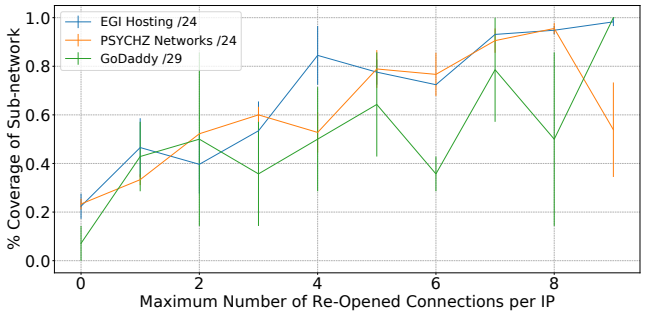


Figure 13: Scanning Probabilistic Temporarily Blocking Hosts—Increasing the number of times to retry a failed TCP connection increases the probability that a probabilistic temporarily blocking IP successfully completes an SSH handshake.

sends RSTs for all hosts in the network when scanning is detected and does so for only SSH.

Excluding Alibaba hosts, 57% of transiently missed SSH hosts explicitly close connections by sending a RST or FIN-ACK packet after the TCP handshake completes, in contrast to 70% of transiently missed HTTP(S) hosts that drop the connection rather than close it. This suggests that SSH hosts are more likely to explicitly deny connection requests, but do not do so consistently. We analyze hosts in the ten ASes that exhibit the most transient SSH hosts. We find that hosts typically close the connection after the TCP connection completes, but occasionally complete the SSH handshake. To further investigate, we conduct an additional experiment from US_1 in which we select a random candidate sub-network from each of the top ten ASes in terms of number of transiently missed SSH hosts. We iteratively scan all hosts in the sub-network while each time increasing the maximum number of times we retry the SSH handshake. By re-trying the SSH handshake, we achieve higher coverage of each network (Figure 13). For example, re-trying the

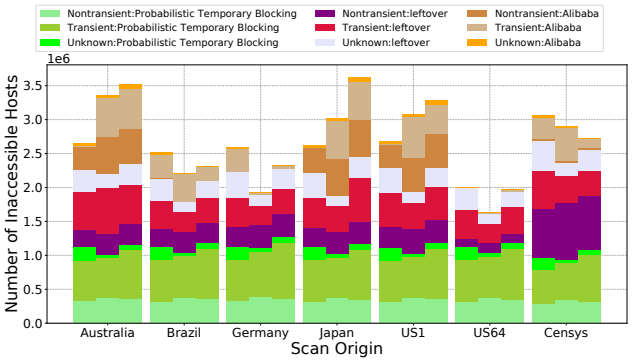


Figure 14: Further Breakdown of missing SSH hosts—Probabilistic temporary blocking and temporal blocking due to Alibaba contribute to over half of the missing SSH hosts. Probabilistic temporary blocking affects all origins relatively equally, while Alibaba only selectively blocks certain origins when scanning is detected.

SSH handshake up to eight times results in successful handshakes with 90% of responding IPs in EGI Hosting and Psychz Networks.

A Psychz Networks article attributes the non-deterministic closure to the OpenSSH `MaxStartups` host setting [52]: a three-tuple that specifies the maximum number of concurrent unauthenticated connections to the SSH daemon, the probability that a new connection is refused once the maximum is reached, and a strict maximum of unauthenticated connections after which all connection attempts are refused [61]. Increasing the number of consecutive SSH handshakes increases the probability that the SSH connection is not refused, as long as the maximum number of unauthenticated connections is not reached. Notably, scanning from multiple origins simultaneously increases the likelihood of an SSH host using the `MaxStartups` property to reject the connection, as the number of concurrent unauthenticated connections collectively increases.

To quantify the number of probabilistic temporarily blocking hosts, we categorize any IP that closes the connection after a TCP handshake with at least one origin and successfully completes an SSH handshake with another origin as due to probabilistic blocking. We estimate that this behavior causes the loss of 1.1M SSH hosts (32–63% of missed SSH hosts across origins and trials), regardless of the number of source IPs used. We further highlight that 30% of all probabilistic temporarily blocking IPs appear to be long-term inaccessible. However, by repeating the experiment described above, we confirm that the long-term inaccessible IPs are also probabilistically blocked and only appear to be non-transient due to the probabilistic nature of this phenomenon. We show a breakdown of the reasons that origins miss SSH hosts in Figure 14. After accounting for probabilistic temporary blocking and Alibaba's scanning detection, the number of missing SSH hosts across all origins becomes 2.2 times smaller than HTTP and 1.1 times larger than HTTPS.

7 DISCUSSION AND LESSONS LEARNED

The median scan origin in our study misses nearly twice the number of hosts as Durumeric et al. originally estimated [22]: 96.3% vs. 97.9% (1 probe) and 97.6% vs. 98.8% (2 probes) coverage for HTTPS. In the

worst case, a single-probe scan from one origin only achieves 91.4% coverage for HTTP and 95.0% coverage for HTTPS. This result hardly invalidates the methodology, and fast Internet scanning has provided sufficient coverage for meaningful contributions in the security and networking communities (e.g., [1, 6, 8, 10, 13–15, 18, 21, 28, 29, 37, 41–43, 50, 58]). In most cases, the increased loss will not meaningfully change the high-level results of Internet-wide scans. However, we emphasize that loss is not simply due to random packet drop. The hosts that origins miss could bias results that focus on specific geographic regions or types of networks, which researchers should consider when designing experiments. Some scan origins also experience more transient loss (e.g., Australia) or long-term loss (e.g., Censys, Germany, Japan, Brazil) than others, and researchers should validate the coverage when scanning from a new location.

Multi-origin scanning. Transiently missed hosts are lost inconsistently and unpredictably. There are no clean results suggesting that scanning topologically or geographically closer to a destination reduces transient loss. Rather, transient loss changes dramatically across trials, and our hypotheses based on topological and regional distance, publicly visible peering relationships, traceroute results, and packet drop rarely panned out when we manually investigated individual networks.

Scanning from two origins helps considerably, increasing the median single probe HTTP coverage to 98.3% and double probe coverage to 98.9% (Figure 15). The variance experienced by pairs of scan origins is dramatically lower than individual origins, which suggests that scanning from any two sufficiently diverse origins significantly improves coverage. At three origins, the median coverage of a 1-probe scan is 99.1% and a 2-probe scan is 99.4% with exceptionally low variance ($\sigma = 0.08\%$). Analogs of Figure 15 for HTTPS and SSH are in Appendix D.

The combination of origins that provide the *best* coverage is difficult to predict. Australia has some of the worst transient loss, but the AU–US_1 pair had the best overall coverage in our study. However, when factoring out long-term inaccessibility, CEN–JP and BR–CEN achieved the best HTTP and HTTPS coverage, respectively. AU–DE–US_1 was the best triad, but the range of coverage between any triad is 0.24%, which suggests that the exact locations may not matter as long as they are sufficiently diverse. We emphasize that the best combination of origins does not necessarily consist of those that achieve the highest individual coverage. Rather, each additional origin in a multi-origin scan should be diverse enough to maximize the number of new hosts that become visible.

Origin Diversity. All of our scan origins were both topologically and geographically diverse, with the exception of US_1 and US_{64}. To determine whether simply using multiple upstream transit providers at the same geographic location provides the same improved coverage, we performed a follow up experiment where we completed simultaneous scans from three Tier-1 ISPs in the same physical data center. In September 2020, we completed two HTTP ZMap + ZGrab scans of the full IPv4 address space from our original Australia, Germany, Japan, US_1, and Censys perspectives as well as from three hosts located in the Chicago Equinix CHI4 data center. Each host peered with one of Hurricane Electric, NTT, and Telia

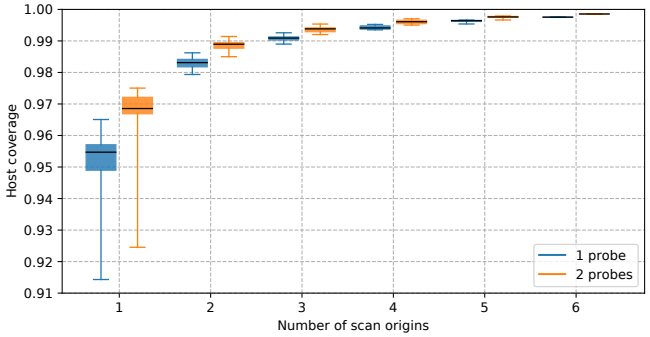


Figure 15: Multi-origin coverage of HTTP hosts — A single origin scan provides a median 95.5% coverage of HTTP hosts. Scanning from two or three origins provides 98% and 99% coverage, with significantly reduced variability. This indicates that scanning from any combination of sufficiently diverse origins provides high coverage of hosts.

Carrier using a unique ASN and unique /24 netblock. The three IP ranges had never previously been used for scanning.

In both trials, Hurricane Electric had the highest coverage of HTTP hosts among all three providers and the five geographic locations (98.1% using 2 probes). The triad of the three Tier-1s in the same location (HE–NTT–TELIA) provided the worst coverage of any three perspectives in both trials ($\mu = 98.7\%$, single probe). It is not inherently surprising that scanning from the same location provides worse coverage than three geographically diverse data centers. Equinix CHI4 is one of the major IXPs in the Midwest, the three providers employ hot-potato routing policies, and they all peer with other major ISPs at the IXP. Traffic to many destination networks may utilize the same paths regardless of the first hop transit provider (e.g., if the transit provider peers with the destination network at the IXP location). We note, however, that variance among all 3-origin scans is low ($\sigma = 0.1\%$) even if the three providers are collocated: HE–NTT–TELIA saw 0.4% fewer hosts than the median triad (Figure 18). The approach provides improved coverage over a single perspective within range of other geographically diverse triads of perspectives and likely at a reduced cost compared to deploying servers in multiple locations.

Multi-probe scanning. Sending two consecutive probes achieves higher coverage than one (96.9% vs. 95.5%), but in almost every case, significantly less coverage than sending one probe from two separate origins. Sending one probe from three origins typically provides better coverage than sending two probes from two origins, and requires less bandwidth. In more than 93% of cases where at least one probe was lost from an origin, both probes were lost, which suggests that multiple consecutive probes do not provide meaningful resilience against packet drop — packet loss is simply not uniform random. This problem can be partially mitigated by delaying the time between probes in a scan as proposed by Bano et al. [7] instead of sending probes consecutively. We encourage organizations and researchers performing a significant amount of scanning to consider using 2–3 vantage points. If researchers only have a single vantage point, we suggest scanning with multiple

probes with delay between probes to the same host, or to perform multiple independent trials of the experiment.

Scanner Blocking. Blocking is more severe than previously estimated [19]. Before changing their IP ranges, Censys persistently misses 5 times more hosts than the academic origins, far overshadowing the number of hosts they miss transiently. While 67% of the hosts that Censys misses belong to only three networks, after excluding these networks Censys still misses nearly 1.5 times more HTTP hosts than the second-worst origin. Since the time of our initial study, Censys has updated their IP ranges used for scanning. We confirmed that scanning with a fresh source IP increased Censys' coverage of HTTP hosts by more than 5.5%. Scanning with multiple source IP addresses also appears to prevent some intrusion detection systems from picking up on scans, and a single probe scan from US_{64} achieves slightly higher HTTP coverage than a 1-probe scan from CEN–DE (98% vs. 97.9%), but slightly lower coverage than the median pair of 1-probe, single-IP origins (98.3%).

Regional Biases. We find evidence of regional blocklists and allowlists. While there is no indication that the Japan and Brazil IPs have scanned before, both exclusively miss tens of thousands of hosts. Over 70% of lost hosts are located in Eastern Europe or are financial/health-related business networks in the United States. We also find that some websites are only accessible from within the same country. While exclusively accessible networks are not large enough to affect our global statistics, they may affect a meaningful number of sites within a country. None of our scan origins are located in countries known for maintaining separate Internet infrastructure, and the problem is likely more pronounced in other regions (e.g., China and Russia).

8 RELATED WORK

There is a significant body of work that develops scanning methods (e.g., [2, 11, 17, 22, 24, 35, 40]) as well as uses the methodology to study Internet behavior (e.g., [1, 6–8, 12, 14, 21, 25, 28, 29, 41, 50, 54]). Several past studies acknowledge differences between perspectives, but do not directly measure the source causes.

In 2008, Heidemann et al. completed an ICMP census of the allocated IPv4 address space from two U.S. locations; the response rate of their two origins were within 5% of each other for 96% of /24 network blocks [27]. Averaged across each pair of our origins, we find that 87% of /24 blocks achieve a response rate within 5%. This may be due to greater geographic or topological diversity among scan origins. In 2012, Durumeric et al. estimated that a 1 packet scan achieves 97.9% coverage and 2 packets 98.8% coverage by performing a series of multi-packet scans from a single origin [22]. We find similar, but slightly lower coverage rates at 96.3% coverage for 1 packet and 97.6% for 2 packet scans from a single origin. Our number is likely lower because the original ZMap work assumes that packet drop is uniform random, which we show is not true. Adrian et al. completed a similar measurement when estimating coverage at 10 gbE [2].

Later, in 2014, Durumeric et al. completed simultaneous scans of TCP/443 from two academic institutions in the United States to measure the impact of operators blocking scan traffic [19]; they estimated that 0.4–0.6% of HTTPS hosts are inaccessible due to blocking. We find dramatically (8.5 times) more blocking of Censys, likely due to their consistent scanning. They also do not explore transient versus long-term host inaccessibility. Guo et al. discuss the prevalence of ICMP rate limiting, which may affect some scans, though all probes in our study are TCP-based [26].

There is also a large body of work that focuses on Internet censorship, which could contribute to the differences between scan origins. Pearce et al. [50] performed DNS queries from geographically distributed resolvers to quantify DNS manipulation in different countries. Khattak et al. analyzed the differential treatment of Tor users and note that the view of the global web seems to change depending on where a scan originates, even for non-Tor control nodes [34]. None of the vantage points we use are located in countries known for censoring access, and we do not find that censorship is a major cause of the differences between origins in our study.

Padmanabhan et al. performed a nine year longitudinal study on the effects of weather conditions on host outages using pings from 10 geographically dispersed PlanetLab nodes [47]. They use multiple vantage points for redundancy but do not analyze the differences between vantage points. However, they recognize that hosts can be unresponsive to all origins, and correlate the probability of such dropout events with various factors. Shavitt et al. measure the impact of vantage point distribution on creating AS topology maps using hundreds of DIMES agents, and use graph convergence techniques to show that it can take up to 40 different vantage points for the Internet topology to converge [59]. The authors also stress the importance of distributed vantage points in active Internet measurement infrastructure. Kliman-Silver et al. studied the impact of geolocation on web search personalization [36]. Kumar et al. considered the difference in coverage of IoT devices between in-home and Internet-wide scanning [38]. Holterbach et al. investigate the similarity between results from topologically different RIPE Atlas nodes and find that probe selection can increase the number of discovered IPs by as much as 25% compared to the default RIPE Atlas probe selection, but did not investigate why this occurs [30].

9 CONCLUSION

In this paper, we investigated how the networks used to conduct Internet-wide scans affect their results. We showed that a single-origin, single-probe scan sees about 96% of HTTP(S) and 84% of SSH hosts globally. This is more than twice the loss originally estimated by Durumeric et al. [22] and is not simply due to uniform random packet drop. Host inaccessibility is caused by both transient and long-term network problems. Transient loss is generally inconsistent across origins, though some origins consistently experience greater transient loss than others. While unpredictable, transient loss can be reliably overcome by scanning from 2–3 diverse origins. Blocking of networks used for regular scanning is also more pronounced than previously believed. We find that regional access limitations can bias results, and that in several countries, the policy decisions of a single ISP can substantially limit a scanner's visibility. To increase coverage, we encourage researchers to consider using 2–3 diverse vantage points, multiple source IP addresses, and/or sending multiple probes with a delay between them. Overall, loss of global coverage from single-origin scans remains low enough that it likely does not change the high-level results of prior work, and

Internet-wide scanning remains a powerful technique. However, researchers should be cognizant of potential bias when scanning from a single location.

ACKNOWLEDGEMENTS

We thank Ítalo Cunha, who provided our Brazilian scan origin, and without whose help this study would not have been possible. The authors also thank Renata Teixeira, Kimberly Ruth, Jeff Cody, Ethan Uberseder, Jessica Sinnott, Wilson Nguyen, Catherine Han, as well as our shepherd, Romain Fontugne, the Stanford University security and networking teams, and the Censys operations team. This work was supported in part by the National Science Foundation under award CNS-1823192, Cisco Systems, Inc., Google., Inc., NSF Graduate Fellowship DGE-1656518, and a Stanford Graduate Fellowship. This work was supported partially by the Australian Government through the Australian Research Council's Discovery Projects Scheme (Project DP180104030). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

REFERENCES

[1] D. Adrian, K. Bhargavan, Z. Durumeric, P. Gaudry, M. Green, J. A. Halderman, N. Heninger, D. Springall, E. Thomé, L. Valenta, et al. Imperfect Forward Secrecy: How Diffie-Hellman Fails in Practice. In *ACM Conference on Computer and Communications Security*, 2015.

[2] D. Adrian, Z. Durumeric, G. Singh, and J. A. Halderman. Zippier ZMap: Internet-Wide Scanning at 10 Gbps. In *8th USENIX Workshop on Offensive Technologies*, 2014.

[3] M. Allman, V. Paxson, and J. Terrell. A Brief History of Scanning. In *Internet Measurement Conference*, 2007.

[4] J. Amann, O. Gasser, Q. Scheitle, L. Brent, G. Carle, and R. Holz. Mission Accomplished? HTTPS Security after DigiNotar. In *ACM Internet Measurement Conference*, 2017.

[5] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou. Understanding the Mirai Botnet. In *USENIX Security Symposium*, 2017.

[6] N. Aviram, S. Schinzel, J. Somorovsky, N. Heninger, M. Dankel, J. Steube, L. Valenta, D. Adrian, J. A. Halderman, V. Dukhovni, et al. DROWN: Breaking TLS using SSLv2. In *25th USENIX Security Symposium*, 2016.

[7] S. Bano, P. Richter, M. Javed, S. Sundaresan, Z. Durumeric, S. Murdoch, R. Mortier, and V. Paxson. Scanning the Internet for Liveness. In *ACM Computer Communication Review*, 2018.

[8] B. Beurdouche, K. Bhargavan, A. Delignat-Lavaud, C. Fournet, M. Kohlweiss, A. Pironti, P.-Y. Strub, and J. K. Zinzindohoue. A Messy State of the Union: Taming the Composite State Machines of TLS. In *IEEE Symposium on Security and Privacy*, 2015.

[9] R. Beverly, M. Luckie, L. Mosley, and k. claffy. Measuring and Characterizing IPv6 Router Availability. In *Passive and Active Network Measurement Workshop (PAM)*, 2015.

[10] C. Brubaker, S. Jana, B. Ray, S. Khurshid, and V. Shmatikov. Using Frankencerts for Automated Adversarial Testing of Certificate Validation in SSL/TLS Implementations. In *IEEE Symposium on Security and Privacy*, 2014.

[11] R. Bush, O. Maennel, M. Roughan, and S. Uhlig. Internet Optometry: Assessing the Broken Glasses in Internet Reachability. In *ACM Internet Measurement Conference*, 2009.

[12] X. Cai and J. Heidemann. Understanding Block-Level Address Usage in the Visible Internet. *ACM SIGCOMM Computer Communication Review*, 2011.

[13] S. Checkoway, R. Niederhagen, A. Everspaugh, M. Green, T. Lange, T. Ristenpart, D. J. Bernstein, J. Maskiewicz, H. Shacham, and M. Fredrikson. On the Practical Exploitability of Dual EC in TLS Implementations. In *23rd USENIX Security Symposium*, 2014.

[14] A. Costin, J. Zaddach, A. Francillon, and D. Balzarotti. A Large-Scale Analysis of the Security of Embedded Firmwares. In *23rd USENIX Security Symposium*, 2014.

[15] J. Czyz, M. Luckie, M. Allman, and M. Bailey. Don't Forget to Lock the Back Door! A Characterization of IPv6 Network Security Policy. In *Symposium on Network and Distributed Systems Security (NDSS)*, 2016.

[16] A. Dhamdhere, R. Teixeira, C. Dovrolis, and C. Diot. NetDiagnoser: Troubleshooting Network Unreachabilities Using End-to-End Probes and Routing Data. In *Proceedings of the 2007 ACM CoNEXT Conference*, 2007.

[17] 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*, 2015.

[18] Z. Durumeric, D. Adrian, A. Mirian, J. Kasten, E. Bursztein, N. Lidzborski, K. Thomas, V. Eranti, M. Bailey, and J. A. Halderman. Neither Snow Nor Rain Nor MITM... An Empirical Analysis of Email Delivery Security. In *ACM Internet Measurement Conference*, 2015.

[19] Z. Durumeric, M. Bailey, and J. A. Halderman. An Internet-Wide View of Internet-Wide Scanning. In *USENIX Security Symposium*, 2014.

[20] Z. Durumeric, J. Kasten, M. Bailey, and J. A. Halderman. Analysis of the HTTPS Certificate Ecosystem. In *ACM Internet Measurement Conference*, 2013.

[21] Z. Durumeric, F. Li, J. Kasten, J. Amann, J. Beekman, M. Payer, N. Weaver, D. Adrian, V. Paxson, M. Bailey, et al. The Matter of Heartbleed. In *ACM Internet Measurement Conference*, 2014.

[22] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-Wide Scanning and its Security Applications. In *USENIX Security Symposium*, 2014.

[23] V. Giotsas, G. Smaragdakis, C. Dietzel, P. Richter, A. Feldmann, and A. Berger. Inferring BGP Blackholing Activity in the Internet. In *ACM Internet Measurement Conference*, 2017.

[24] R. D. Graham. MASSCAN: Mass IP port scanner. *https://github.com/robertdavidgraham/masscan*, 2014.

[25] M. H. Gunes and K. Sarac. Analyzing Router Responsiveness to Active Measurement Probes. In *International Conference on Passive and Active Network Measurement*, 2009.

[26] H. Guo and J. Heidemann. Detecting ICMP Rate Limiting in the Internet. In *Conference on Passive and Active Network Measurement*, 2018.

[27] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and Survey of the Visible Internet. In *ACM Internet Measurement Conference*, 2008.

[28] E. Heilman, A. Kendler, A. Zohar, and S. Goldberg. Eclipse Attacks on Bitcoin's Peer-to-Peer Network. In *USENIX Security Symposium*, 2015.

[29] N. Heninger, Z. Durumeric, E. Wustrow, and J. A. Halderman. Mining Your Ps and Qs: Detection of Widespread Weak Keys in Network Devices. In *USENIX Security Symposium*, 2012.

[30] T. Holterbach, E. Aben, C. Pelsser, R. Bush, and L. Vanbever. Measurement Vantage Point Selection Using a Similarity Metric. In *Applied Networking Research Workshop*, 2017.

[31] R. Holz, J. Amann, O. Mehani, M. Wachs, and M. A. Kaafar. TLS in the Wild: An Internet-Wide Analysis of TLS-based Protocols for Electronic Communication. *Symposium on Network and Distributed System Security (NDSS)*, 2016.

[32] G. Iannaccone, C. Chuah, R. Mortier, S. Bhattacharyya, and C. Diot. Analysis of Link Failures in an IP Backbone. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, 2002.

[33] E. Katz-Bassett, H. V. Madhyastha, J. P. John, A. Krishnamurthy, D. Wetherall, and T. E. Anderson. Studying Black Holes in the Internet with Hubble. In *NSDI*, 2008.

[34] S. Khattak, D. Fifield, S. Afroz, M. Javed, S. Sundaresan, V. Paxson, S. J. Murdoch, and D. McCoy. Do You See What I See? Differential Treatment of Anonymous Users. In *Symposium on Network and Distributed System Security*, 2016.

[35] J. Klick, S. Lau, M. Wählisch, and V. Roth. Towards Better Internet Citizenship: Reducing the Footprint of Internet-Wide Scans by Topology Aware Prefix Selection. In *2016 Internet Measurement Conference*, 2016.

[36] C. Kliman-Silver, A. Hannak, D. Lazer, C. Wilson, and A. Mislove. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Internet Measurement Conference*, 2015.

[37] 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*, 2015.

[38] D. Kumar, K. Shen, B. Case, D. Garg, G. Alperovich, D. Kuznetsov, R. Gupta, and Z. Durumeric. All Things Considered: An Analysis of IoT Devices on Home Networks. In *USENIX Security Symposium*, 2019.

[39] R. Lawshae. Hunting Botnets with ZMap. http://h30499.www3.hp.com/t5/HP-Security-Research-Blog/Hunting-Botnets-with-ZMap/ba-p/6320865#.UvzzgkJdXw1.

[40] D. Leonard and D. Loguinov. Demystifying Service Discovery: Implementing an Internet-Wide Scanner. In *ACM Internet Measurement Conference*, 2010.

[41] F. Li, Z. Durumeric, J. Czyz, M. Karami, M. Bailey, D. McCoy, S. Savage, and V. Paxson. You've Got Vulnerability: Exploring Effective Vulnerability Notifications. In *25th USENIX Security Symposium*, 2016.

[42] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey, and M. Liu. Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents. In *24th USENIX Security Symposium*, 2015.

[43] W. R. Marczak, J. Scott-Railton, M. Marquis-Boire, and V. Paxson. When Governments Hack Opponents: A Look at Actors and Technology. In *23rd USENIX Security Symposium*, 2014.

[44] Maxmind GeoLite2 Database. https://dev.maxmind.com/geoip/geoip2/geolite2/.

[45] A. Mirian, Z. Ma, D. Adrian, M. Tischer, T. Chuenchujit, T. Yardley, R. Berthier, J. Mason, Z. Durumeric, J. A. Halderman, et al. An Internet-Wide View of ICS Devices. In *14th IEEE Conference on Privacy, Security and Trust*, 2016.

[46] R. Padmanabhan, A. Dhamdhere, E. Aben, N. Spring, et al. Reasons Dynamic Addresses Change. In *ACM Internet Measurement Conference*. ACM, 2016.

[47] R. Padmanabhan, A. Schulman, D. Levin, and N. Spring. Residential Links Under the Weather. In *ACM SIGCOMM*, 2019.

[48] V. Paxson. End-to-end Internet Packet Dynamics. In *ACM SIGCOMM Computer Communication Review*, 1997.

[49] P. Pearce, R. Ensafi, F. Li, N. Feamster, and V. Paxson. Augur: Internet-Wide Detection of Connectivity Disruptions. In *IEEE Security and Privacy*, 2017.

[50] P. Pearce, B. Jones, F. Li, R. Ensafi, N. Feamster, N. Weaver, and V. Paxson. Global Measurement of DNS Manipulation. In *USENIX Security Symposium*, 2017.

[51] Project Sonar. https://www.rapid7.com/research/project-sonar.

[52] Psychz Networks Forum - ssh_exchange_identification connection closed by remote host. https://www.psychz.net/client/question/en/sshexchangeidentification-connection-closed-by-remote-host.html.

[53] E. Pujol, P. Richter, B. Chandrasekaran, G. Smaragdakis, A. Feldmann, B. M. Maggs, and K.-C. Ng. Back-office Web Traffic on the Internet. In *ACM Internet Measurement Conference*, 2014.

[54] L. Quan and J. Heidemann. Detecting Internet Outages with Active Probing. Technical report, Citeseer, 2011.

[55] L. Quan, J. Heidemann, and Y. Pradkin. Trinocular: Understanding Internet Reliability through Adaptive Probing. *ACM SIGCOMM Computer Communication Review*, 2013.

[56] L. Quan, J. Heidemann, and Y. Pradkin. When the Internet sleeps: Correlating Diurnal Networks with External Factors. In *ACM Internet Measurement Conference*, 2014.

[57] P. Richter, G. Smaragdakis, D. Plonka, and A. Berger. Beyond Counting: New Perspectives on the Active IPv4 Address Space. In *ACM Internet Measurement Conference*, 2016.

[58] C. Rossow. Amplification Hell: Revisiting Network Protocols for DDoS Abuse. In *Symposium on Network and Distributed Systems Security (NDSS)*, 2014.

[59] Y. Shavitt and U. Weinsberg. Quantifying the Importance of Vantage Point Distribution in Internet Topology Mapping. In *IEEE Journal on Selected Areas in Communications*, 2011.

[60] D. Springall, Z. Durumeric, and J. A. Halderman. FTP: The Forgotten Cloud. In *IEEE/IFIP International Conference on Dependable Systems and Networks*, 2016.

[61] Linux Manual Page - sshd_config. https://linux.die.net/man/5/sshd_config.

[62] TIM Brasil. https://www.tim.com.br/sp/para-voce.

[63] P. Zhu, K. Man, Z. Wang, R. Ensafi, J. A. Halderman, and H. Duan. Characterizing Transnational Internet Performance and the Great Bottleneck of China. In *ACM Sigmetrics*, 2020.

A GROUND-TRUTH COVERAGE

		Australia	Brazil	Germany	Japan	US 1 IP	US 64 IPs	Censys	∩	∪
HTTP	1	96.5%	96.5%	96.3%	97.2%	96.9%	97.4%	91.6%	85.9%	57,829,891
	2	97.6%	97.4%	97.1%	97.7%	97.7%	98.2%	92.4%	87.6%	58,040,919
	3	95.9%	97.1%	96.8%	96.9%	97.9%	98.4%	93.4%	86.6%	58,554,985
	μ	96.7%	97.0%	96.7%	97.3%	97.5%	98.0%	92.5%	86.7%	58,141,932
HTTPS	1	98.5%	98.5%	95.7%	98.3%	98.0%	98.2%	95.4%	90.4%	40,809,122
	2	96.3%	98.3%	97.7%	98.1%	98.6%	99.0%	95.6%	90.5%	41,093,084
	3	97.1%	97.9%	97.1%	97.5%	98.4%	98.9%	96.5%	90.7%	41,098,147
	μ	97.3%	98.2%	96.8%	97.9%	98.3%	98.7%	95.8%	90.5%	41,000,118
SSH	1	86.4%	87.1%	86.7%	86.6%	86.2%	89.7%	84.3%	72.1%	19,457,647
	2	82.8%	88.7%	90.1%	84.6%	84.3%	91.7%	85.2%	70.6%	19,598,041
	3	82.3%	88.3%	88.3%	81.8%	83.5%	90.1%	86.3%	69.0%	19,891,888
	μ	83.8%	88.0%	88.4%	84.3%	84.7%	90.5%	85.3%	70.6%	19,649,192

(a) Fraction of ground truth hosts perceived from each scan origin in all trials (2 probes)—No origin achieves full coverage of hosts, and all origins agree on only 87% of HTTP, 91% of HTTPS, and 71% of SSH hosts. Each trial represents a snapshot of the protocol ecosystem on the day the scan was conducted.

		Australia	Germany	Japan	US 1 IP	Censys	HE	NTT	Telia	∩	∪
HTTP	1	96.5%	96.1%	97.9%	97.8%	97.5%	98.1%	97.9%	97.6%	90.1%	56,094,571
	2	96.6%	96.2%	98.0%	98.0%	97.7%	98.2%	97.8%	97.9%	90.4%	55,934,190
	μ	96.6%	96.2%	97.9%	97.9%	97.6%	98.1%	97.9%	97.8%	90.2%	56,014,381

(b) Fraction of ground truth hosts perceived from each scan origin in follow up HTTP experiment (2 probes)—Hurricane Electric achieves the highest coverage among the three providers and five geographic origins. Censys sees a more than 5% increase in HTTP coverage by scanning with a new IP.

B LONG-TERM INACCESSIBILITY

	>1M Hosts					>100K Hosts					>10K Hosts					>1K Hosts				
	US	GB	CN	FR	NL	ZA	IT	VE	RO	AR	BD	BO	GR	EC	TN	SD	LY	AM	ZW	GU
AU	0.6	0.3	0.2	0.4	0.2	1.0	0.6	0.4	0.6	0.8	0.2	0.3	0.8	0.8	0.2	1.2	0.4	0.5	0.8	0.3
BR	1.1	0.7	0.3	0.6	0.2	1.4	0.4	0.1	1.1	0.2	0.1	0.2	1.1	0.5	0	0.1	0	0.2	0.2	1.3
DE	0.9	0.6	0.5	0.3	0.9	0.7	9.0	7.8	0.2	5.1	0.9	3.6	8.7	7.7	7.8	21.6	21.3	11.7	0.2	0.4
JP	0.8	0.5	0.7	0.3	0.2	0.9	0.3	0.1	0.4	0.5	0.1	0.5	0.8	1	0.1	0.3	0.1	0.2	0.2	0.1
US1	0.3	0.3	0.4	0.2	0.2	0.4	0.2	0.4	0.2	0.3	0	0.2	0.6	3.8	0.3	1.6	1.1	0	0	0.1
US64	0.2	0.2	0.4	0.2	0.1	0.3	0.2	0.5	0.1	0.2	0	0.2	0.5	3.8	0.3	1.4	0.8	0	0.1	0
CEN	2.9	2.2	1.9	1.7	1.4	21.6	4.3	1.9	5.4	2.9	14.3	9.1	4.5	8.4	3.0	10.0	9.8	1	10.4	9.2

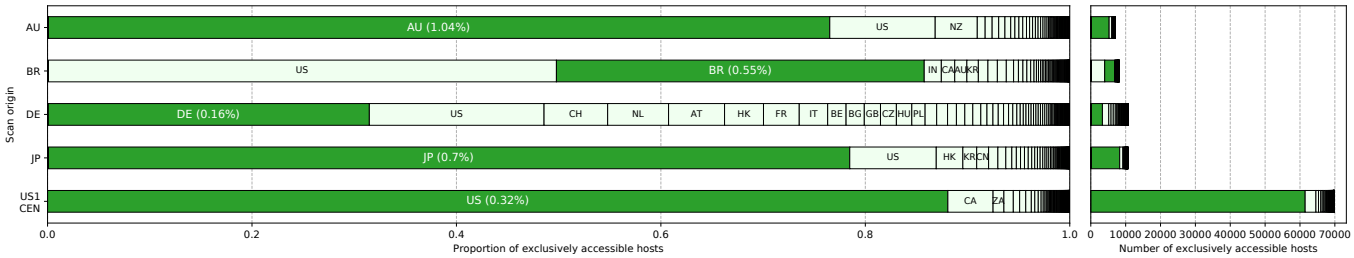
(a) HTTPS

	>1M Hosts					>100K Hosts					>10K Hosts					>1K Hosts				
	CN	US	DE	–	–	KR	IT	PL	HK	AU	BD	ZA	PT	CO	PE	LY	ZW	TN	SD	SN
AU	14.5	4.6	1.4	–	–	14.6	5.7	1.6	7.4	0.4	10.6	5.8	3.7	1.2	1.1	2.0	2.3	1.9	1.3	1.9
BR	1.4	4.6	1.3	–	–	12.8	12.2	1.9	4.5	2.8	5.1	5.5	3.5	7.0	0.4	0.1	1.0	0.4	0.3	0.4
DE	1.3	4.7	2.5	–	–	11.0	10.6	9.0	4.0	2.7	7.9	4.9	3.6	8.6	9.5	33.1	0.5	17.2	14.1	13.2
JP	15.7	4.7	2.9	–	–	12.1	13.3	2.1	6.4	2.3	7.0	4.4	3.8	4.4	0.6	0.1	0.8	3.5	0.4	0.3
US1	14.6	5.0	2.9	–	–	11.9	12.6	1.8	7.3	2.3	4.8	4.1	3.6	9.6	0.6	2.9	1.1	3.4	0.3	0.3
US64	1.3	3.6	2.3	–	–	0.3	0.9	0.7	4.0	2.2	4.2	3.1	0.6	8.6	0.4	2.2	0.8	3.5	0.1	0.3
CEN	3.5	8.5	3.2	–	–	13.2	13.2	1.9	6.4	4.5	37.1	13.1	12.5	1.7	3.4	11.7	17.6	8.8	5.8	3.6

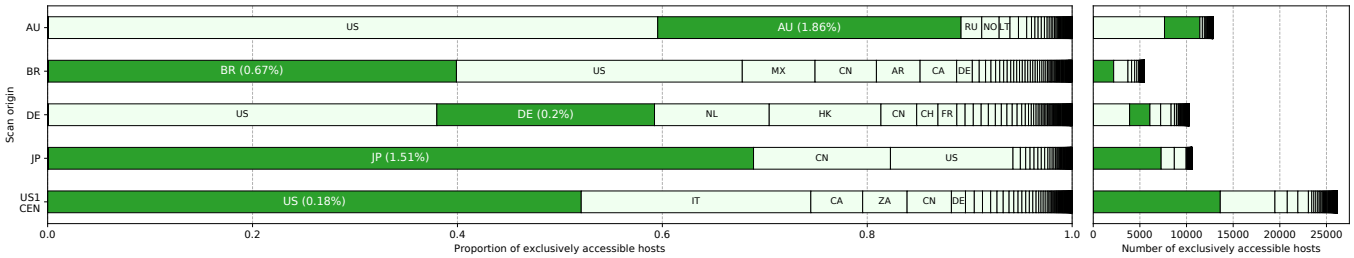
(b) SSH

Table 5: Countries with the most long-term inaccessible hosts—Coverage of countries can be greatly influenced by scan origin, but a significant fraction of missing hosts are often due to a handful of major ASes; red indicates majority inaccessible from one AS, orange two, and yellow for at least three.

C EXCLUSIVE ACCESSIBILITY



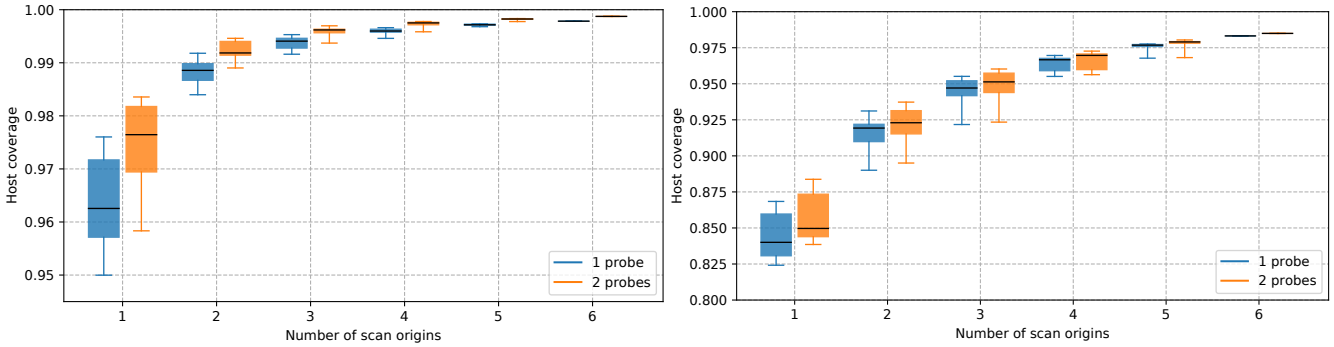
(a) HTTPS



(b) SSH

Figure 16: Exclusively accessible hosts by country—Origins within a country typically have better accessibility than external origins do. Dark green indicates hosts that are only accessible by scanning from within the country. For these, we additionally show the fraction of that country's total hosts that are exclusively accessible from within the country.

D MULTI-ORIGIN COVERAGE



(a) HTTPS **(b) SSH**

Figure 17: Multi-origin coverage—Scanning from three or more origins increases HTTPS coverage by 2–3% over a single origin. SSH requires many more origins to achieve the same coverage, likely due to probabilistic temporary blocking (Section 6).

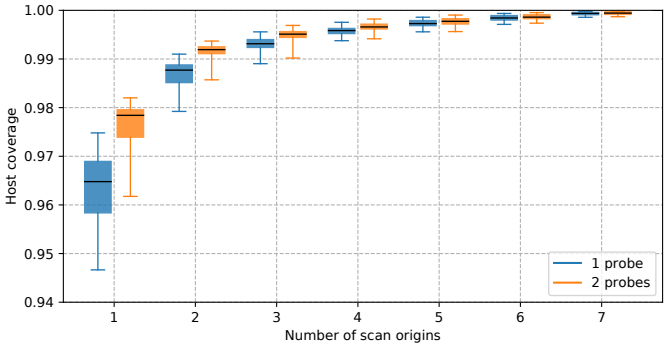


Figure 18: Multi-origin coverage in follow up HTTP experiment—The HE-NTT-TELIA triad, collocated in the same data center, achieves the worst coverage of any three origins.