Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope

Philipp Richter
MIT / Akamai
richterp@csail.mit.edu

Arthur Berger
Akamai / MIT
arthur@akamai.com

ABSTRACT

Scanning of hosts on the Internet to identify vulnerable devices and services is a key component in many of today's cyberattacks. Tracking this scanning activity, in turn, provides an excellent signal to assess the current state-of-affairs for many vulnerabilities and their exploitation. So far, studies tracking scanning activity have relied on unsolicited traffic captured in darknets, focusing on random scans of the address space. In this work, we track scanning activity through the lens of unsolicited traffic captured at the firewalls of some 89,000 hosts of a major Content Distribution Network (CDN). Our vantage point has two distinguishing features compared to darknets: *(i)* it is distributed across some 1,300 networks, and *(ii)* its servers are *live*, offering services and thus emitting traffic. While all servers receive a baseline level of probing from Internet-wide scans, i.e., scans targeting random subsets of or the entire IPv4 space, we show that some 30% of all logged scan traffic is the result of localized scans. We find that localized scanning campaigns often target narrow regions in the address space, and that their characteristics in terms of target selection strategy and scanned services differ vastly from the more widely known Internet-wide scans. Our observations imply that conventional darknets can only partially illuminate scanning activity, and may severely underestimate widespread attempts to scan and exploit individual services in specific prefixes or networks. Our methods can be adapted for individual network operators to assess if they are subjected to targeted scanning activity.

CCS CONCEPTS

• **Networks** → **Network measurement**; **Network security**;

KEYWORDS

Internet scanning, Internet security, network telescope, unsolicited traffic.

ACM Reference Format:
Philipp Richter and Arthur Berger. 2019. Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope. In *Internet Measurement Conference (IMC '19), October 21–23, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, Article 4, 14 pages. https://doi.org/10.1145/3355369.3355595

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
IMC '19, October 21–23, 2019, Amsterdam, Netherlands
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6948-0/19/10…$15.00
https://doi.org/10.1145/3355369.3355595

1 INTRODUCTION

Scanning of hosts on the Internet for vulnerable services is a key component for cyberattacks ranging from exploitation of end devices or servers up to massive botnets capable of carrying out DDoS attacks exceeding several Tbps. Given that scanning is a key component in many attack vectors, tracking scanning activity can illuminate the current status of botnets (e.g, Mirai), can show which vulnerabilities are targeted, who originates scanning traffic, and which networks are particularly targeted by scanners. More generally, scanning activity provides us with an excellent signal for the current state-of-affairs for many vulnerabilities and their exploitation. As such, scanning activity might well be suitable to indicate potential future cyberattacks.

While the existence of widespread port scanning activity is widely known [7, 8, 10, 18], comparably little research has been devoted to study and understand scanners and their strategies in detail. The lack of high quality data stems mostly from the absence of data sources that can illuminate scanning activity at scale. The few recent studies on scanning base their analysis on traffic arriving at darknets: a portion of routed address space, such as a /8 IPv4 prefix or subsets thereof, that does not emit any traffic, and only records incoming packets. Darknets face two fundamental shortcomings: Firstly, they are isolated within the address space, often announced by universities or research networks. Secondly, since darknets do not emit any traffic, their only attraction of traffic is their routed address space. Thus, darknets will only be able to detect scanning activity that targets either the entire IPv4 space, a sufficiently large random subset, or the unlikely case of scans targeting the darknet itself [10, 18]. They cannot provide insights into scans targeting specific regions of the address space, e.g., prefixes or networks with known clients, servers or other types of "live" hosts. Scans targeting individual networks, however, could indicate the intent of attackers to exploit or attack those particular networks, posing a potentially much greater risk to those networks than a random scan of the entire IPv4 space.

In this work, we leverage a unique dataset that allows us to overcome the limitations of network telescopes based on darknets, and to illuminate the prevalence, and individual strategies of Internet scanning that have not previously been documented. Our telescope is based on the logs of unsolicited packets blocked at the firewalls of servers of a major CDN, which are *distributed* over more than 1,300 networks, and are *live*, in the sense of offering services to end users and thus emitting traffic.

Our key contributions and findings are as follows:

- We provide a detailed study of our distributed vantage point, 89,000 CDN servers, its suitability as a network telescope, and

the unsolicited traffic logged. We find that all CDN servers receive a consistent baseline number of unsolicited packets, *baseline radiation*, but also show evidence of local concentrations of unsolicited traffic. We quantify the additional unsolicited traffic attracted due to exposure of CDN IP addresses in forward DNS responses, and develop tools to isolate scanning activity from other traffic components. We find that some 87% of logged traffic is the result of scanning activity.
- We develop tools and metrics to categorize scans into individual *scanning strategies*. While Internet-wide scans of the full IPv4 space and of random subsets thereof are the majority of overall scan traffic, we find that localized scanning campaigns constitute some 30% of scan traffic, in terms of number of scans, number of packets in scans, and number of sources initiating the scans. We find that localized scans often target addresses in narrow areas of the address space (e.g., only a small number of routed prefixes), and that these scans show significantly different characteristics when compared to more widely-known Internet-wide scans both in terms of services targeted, repeated stateful scanning behavior, as well as scanner origins. Many of these characteristics only become visible after isolating these localized events, since the volume of Internet-wide scanning campaigns can easily mask characteristics of localized scans.
- We compare our visibility against a /8 darknet and leverage our ability to classify individual scans to separate background radiation into *baseline scanning* and targeted scanning. We show that IP addresses of darknets receive baseline scanning activity, but little in terms of targeted scanning, in stark contrast to the IP addresses of our distributed telescope. We find a three-fold increase in baseline radiation over the last 3 years.

To the best of our knowledge, our work is the first to document widespread localized scanning activity in today's Internet. Our findings have relevance for the research community, as we show that darknets, and derived statistics and inferences, are biased towards Internet-wide scans and might underestimate exploitation attempts of specific services in specific areas of the address space. As our classification methods can be adapted to individual networks, our findings are of practical use for network operators who want to determine if their hosts and infrastructure are subjected to targeted scans. The remainder of this paper is organized as follows: We review related work in § 2 and introduce our vantage point and dataset in § 3. We scrutinize background radiation logged in our dataset and introduce our scan detection mechanism in § 4. We study target-selection strategies of our identified scans in § 5 and drill into further properties of scans in § 6. We compare the visibility of scans in the broader Internet in § 7 and conclude with our implications and future work in § 8.

2 RELATED WORK

Scanning the address space is a key element leveraged by many botnets and worms [8, 23, 31]. While key for malicious actors, scanning the IPv4 space also became more relevant for measurement studies finding vulnerable host populations, patching strategies, address activity, (e.g., [17, 19, 20, 24]), fueled by the arrival of tools enabling fast scanning of large parts of the IPv4 space in short time periods [6, 21]. Unsolicited traffic received in darknets, Internet Background Radiation, has been widely used to study spread and activity of botnets or exploitation attempts of vulnerabilities [8, 12, 16]. Other works studied both general characteristics of Background Radiation and how they can be used for network analysis, inference, and debugging [10, 13–15, 22, 28, 32, 33].

To the best of our knowledge, only one recent work presents broad and detailed characteristics of widespread scanning behavior in the Internet [18]. Using a darknet telescope, the authors focus solely on Internet-wide scans, i.e., scans that probe a random subset, or the entirety, of the IPv4 space. In this context, they make the inference that if a source probes a given percentage of the addresses of the darknet, then that source is likely probing that percentage of addresses of the public IPv4 address space. Benson et al. [10] studied scan visibility in two darknets: one is the same as in [18] and the other darknet is another, different /8 prefix. One of their findings is that scan traffic arrives with equal probability in the two different (appropriately scaled) darknets, which is a negative indication of localized scanning at the level of these two vantage points. Our work complements and extends previous work. We find widespread evidence of localized scanning activity and illuminate a more complex picture of scanning activity in today's Internet.

3 A DISTRIBUTED NETWORK TELESCOPE

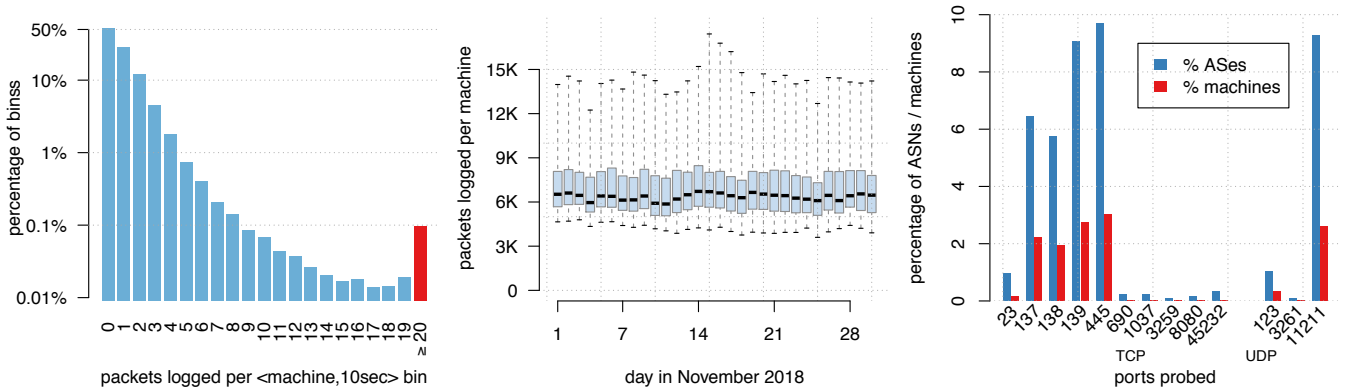
In this section, we introduce our vantage point, relevant properties of our data collection mechanism, and a first-order characterization of the traffic arriving at our telescope.

3.1 Data Collection & Sampling

We base this work on logs of unsolicited packets collected at the firewall of a subset of the servers operated by a major Content Distribution Network.[1] The subset we examine consists of 89,000 servers, and where each server has two publicly facing IPv4 addresses, and both addresses are in the same /24 address block.[2] Although this set of 178,000 addresses is small relative to a darknet of, say, a /8 IPv4 prefix of 16.8 million addresses, these 178,000 addresses are located in 2,800 routed BGP prefixes originated by 1,347 Autonomous Systems in 156 countries, spread across 172 different /8 prefixes.

Client-facing and operations IP address: Of the two public IPv4 addresses on each server, in the same /24 prefix, only one is ever returned in forward DNS queries for domain names hosted by the CDN (when clients access content hosted on the CDN). Herein, we call this address the *client-facing IP*, and the other address we call the *operations IP*. The operations IP is used solely for CDN-internal communication [29]. The distinction between these two addresses that is relevant for this study is that the operations IP is never exposed in replies to forward DNS queries. Both addresses respond to ICMP pings, and both have PTR records set in the DNS. Each CDN machine runs services on some port numbers (most prominently port 80/443 for HTTP(S), and several services using non-standard port numbers for internal communication and customer services). All traffic *not* destined to any of the ports running an active service is dropped by a firewall, and, as described in the next paragraph, a

[1] For more information on the spread and visibility of the CDN, we refer to [30].
[2] The servers also have IPv6 connectivity, but the unsolicited traffic over IPv6 is a tiny fraction of the total, and we do not report on it.



(a) Number of packets logged per 10 second - machine bin. Less than 0.1% of 10-second bins logged more than 20 packets.

(b) Daily packets logged per machine. Lines show median, box shows 25th and 75th percentile, whiskers show 5th and 95th percentile of machines. Each machine has two public IPv4 addresses.

(c) Inbound packets blocked by networks hosting the telescope machines and percentage of telescope machines in these ASes. Up to 10% of networks filter specific ports on their ingress, yet the effect on our data collection remains small.

Figure 1: Statistics on packets received per machine.

total		UDP		TCP		TCP SYN	
pkts	bytes	pkts	bytes	pkts	bytes	pkts	bytes
19.4B	1.2T	2.3B	482G	17.1B	719G	16.8B	702G
100%	100%	11.9%	40.0%	88.0%	59.6%	86.3%	58.2%

Table 1: Total packets logged in November 2018.

portion is logged. We refer to these packets as *unsolicited traffic*. The logs are initially stored locally on the server, and then collected into a centralized repository via a distributed data collection framework.

Token bucket sampling: During heavy bursts of unsolicited traffic, such as DDoS attacks against the CDN machines,[3] the logging of unsolicited packets is controlled via a *token bucket*, which thus maximizes the performance of the server under attack. The sample machines are each configured to have a local token bucket of capacity 10. Each logged packet consumes 1 token. Each second, 2 new tokens are added to the token bucket, until the bucket is at max capacity 10. Thus, the token bucket limits prolonged traffic bursts to only 2 packets per second. Instances of sporadic traffic, on the other hand, will be fully logged and not undergo any sampling. We study the effect of this sampling on our dataset in the following section.

3.2 Dataset Characteristics

Table 1 shows totals of logged packets and bytes from 89,000 machines in the month of November 2018, our primary measurement window. TCP packets with the SYN flag set make up the vast majority of logged packets, and UDP only accounts for some 12%, but some 40% of the bytes (recall that TCP SYN packets typically do not carry payload). The high percentage of TCP SYN packets suggests that the majority of the logged data are actual connection attempts to our servers, and not the result of backscatter traffic, i.e., traffic that third-party servers send in response to DDoS attacks using spoofed source IP addresses, which would show up with a set ACK or RST flag [11, 32]. We note that Wustrow et al. already reported an overall increase of the percentage of Internet Background Radiation with only the SYN flag set between 2006 and 2010 and found that packets with the SYN flag set comprised some 94% of TCP traffic they received in a darknet in 2010 [32]. We find that, as of today, even more than 98% of the TCP traffic we log has only the SYN flag set, a first hint towards widespread scanning activity.

Non-burst vs. burst state: We next assess how often the token buckets are in a burst state, i.e., subjected to large amounts of unsolicited traffic, since such traffic bursts trigger performance controls which cause sampled logging of unsolicited packets. Figure 1a partitions the logging into (machine, 10 second) bins, and shows the number of packets that each individual machine logged within each 10 second timeframe in our measurement window. We see that in more than ≈ 50% of (machine, 10 second) bins, the respective machine did not log any unsolicited traffic. In another 22% of bins, one single packet was received, and in another 10% of bins, two packets were received. Only in less than 0.1% of bins, machines logged at least 20 packets, indicating that the token bucket was in a burst state. In November 2018, "burst packets", i.e., the sum of all packets logged by a machine that logged 20 or more packets in the particular 10 second bin, make up some 2.3% of our total dataset (while an unknown number of unsolicited packets were just blocked and not logged). The more important takeaway, however, is that in almost 99.9% of (machine, 10 second) bins, the machine was *not* in a burst state and hence logged all unsolicited traffic. Thus, our dataset provides excellent visibility into sporadic traffic, as caused, e.g., by scanning. To put these findings into perspective over time, Figure 1b shows, for each day in November 2018, the distribution of packets received per individual machine. Recall that the machines have two public IPv4 addresses. Note that the majority of machines

[3] While a majority of these attacks target services running on the CDN's servers (e.g., Web) and are thus not visible in our dataset, others target non-service ports and are reflected in our dataset.

(the box shows the 25/75th percentile) receive a comparable number of packets, and that few machines log significantly less packets than that. The 5th percentile of machines (the whiskers are the 5th and 95th percentile), still log some 5K packets per day, close to the median of around 6.5K packets. On the other end of the spectrum, we see that some 5-10% of machines log significantly more packets.

3.3 Assessing Presence of Filtering

The CDN servers are hosted in more than 1,300 individual networks, and we do not have control over potential filtering that happens, e.g., at the AS boundary. Port-based filtering at network boundaries is common, e.g., to prevent exploitation of well-known services in local networks (e.g., Windows file sharing, see [1]), and has the potential to affect the visibility of parts of our telescope. To assess the prevalence of port-based network filtering, we sent traffic on benign and on commonly filtered ports from a host in a major cloud hoster to 5 of our machines within each AS. Figure 1c shows, for tested ports and transport protocols, the percentage of ASes in which no machine received our probing packets. In these cases, we infer that the hosting AS deploys port-based filtering. We note that an upwards of 10% of ASes (blue bars in Figure 1c) filter specific ports at the AS boundary; the most common being port 445 (Windows Remote Desktop) and Windows NetBIOS (137,138,139) service. We also observe filtering on UDP ports, e.g., port 11211, which has been used for amplification attacks recently. We observe virtually zero filtering on more benign port numbers, both for TCP as well as UDP. The red bars in Figure 1c show the percentage of affected *machines* (which are a subset of the CDN machines) when taking their distribution across ASes into account, i.e., we tag ASes as filtering/non-filtering on a given port and then tag all servers in this AS consistently. With the tested port numbers, we find that less than 3% of our servers are affected by port filtering by the hosting AS. This observation makes us confident that disparate deployment of network filtering—while clearly present—does not severely affect our inferences of scanning behavior.[4]

3.4 Baseline Traffic and local Concentrations

Figure 1b suggests a mostly consistent distribution of packets over machines. Recall, however, that each machine has two publicly routed IPv4 addresses in the same /24, a *client-facing IP* and an *operations IP*. To get a better intuition of how this difference in *surface* of our telescope plays out, i.e., if and to what degree the two types of telescope IPs receive different amounts of traffic, we next study logged traffic of our machines on each of these interfaces. Figure 2 shows the daily traffic on client-facing and operations interface for 4 example machines. Both top machines show a steady number of packets received on both client-facing and operations interfaces, and we note that a large number machines follow this pattern most of the time. Traffic is balanced over both IP addresses, steadily at ≈3,000 packets per day and IP address. We refer to this phenomenon as *baseline radiation*, which we will further examine in § 7. The bottom-left machine shows some clear traffic spikes, however we see that the spikes are exactly correlated over both IP addresses. We

[4]A notable exception would be scans that are executed from *within* a network that deploys filtering at its network boundary, which would result in highly localized visibility of these scans.

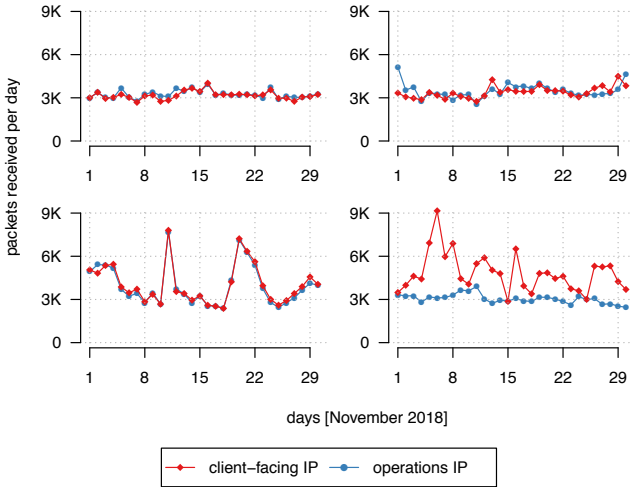


Figure 2: Four example CDN machines and daily packets received in November 2018. The top examples show machines that received only baseline radiation. Bottom-left example shows traffic spikes correlate over both machine IP addresses (*CDN-agnostic pattern*). Bottom-right example shows no change in baseline radiation for the operations IP address, but shows spikes for the client-facing IP (*CDN-targeted pattern*).

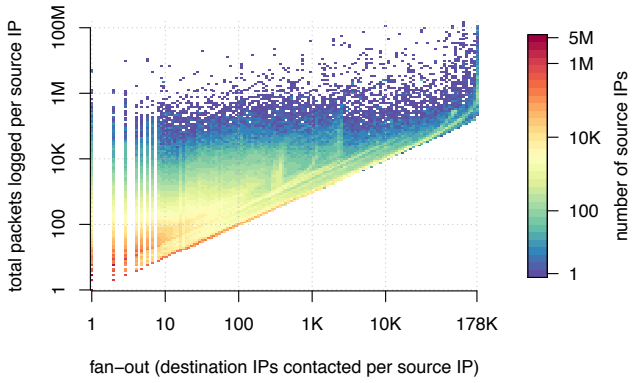
refer to such spikes as *CDN-agnostic*, since we do not see evidence that hosts targeted traffic specifically at the client-facing IP of CDN servers, but rather against entire address blocks and/or networks. Lastly, we show an example of a machine, where the operations interface receives only baseline radiation, but we observe spikes of traffic targeting the client-facing IP. We refer to such spikes as *CDN-targeted*, since these packets were clearly *not* destined at an entire range or network, but towards IP addresses exposed via forward DNS. We only show examples here, but want to highlight that all machines we manually inspected (several thousands), while sometimes showing vastly different traffic levels and "amplitudes", all fall into one of the 3 shown behavioral patterns, making us confident that we cover the significant cases. This distinction between baseline radiation, CDN-agnostic and CDN-targeted traffic is vital for our upcoming characterization of unsolicited traffic and scanning activity, since it highlights the different ways that our machines attract unsolicited traffic.

4 SCRUTINIZING SOURCES

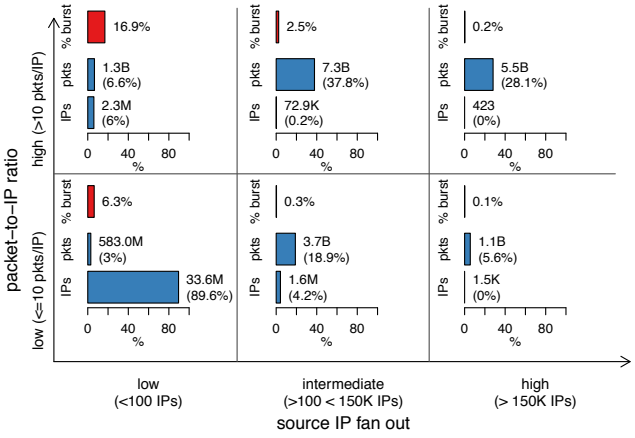
Next, we shift our perspective and present a source-centric perspective of unsolicited traffic arriving at our telescope. We first show macroscopic properties of the activity of individual source addresses and then proceed to identify and dissect scanning activity in our dataset.

4.1 Fan-out and Packet Ratio

Figure 3a shows a heatmap of every source IP address that was seen in the month of November 2018 (N=37.5M), where we bin source IPs by the number of destination IPs it contacted (x-axis), which



(a) Per source IP (N=37.5M): destination IPs contacted (x-axis) vs. total packets logged (y-axis) for November 2018.



(b) Distribution of packets, number of unique source IP addresses, and percent bursts, when partitioning source IPs by fan-out (x-axis) and packet-to-IP ratio (y-axis). In each tile, we show contributions to totals of source IPs and packets, and the percentage of packets within each tile that were received as part of packet bursts.

Figure 3: Source IP address statistics.

we call the "fan-out," as well as the total number of packets we logged from this source across all machines (y-axis). Towards the very bottom-left, we see source IPs that contacted only a very small number of destination IP addresses and sent only a tiny number of packets. Towards the very right of the figure, we find source IPs that contacted most or all IP addresses of our telescope, and some of them sent on the order of ≈100M packets within one month, hinting towards multiple full scans of the IPv4 space carried out by single IP addresses.

Partitioning source IPs: To get a better understanding of the proportions of both IP addresses and traffic, we next partition source IP addresses according to *(i)* their fan-out, the number of contacted IP addresses, and *(ii)* their packet-to-IP ratio, i.e., the number of packets divided by the fan-out. Figure 3b shows percentages of unique source IP addresses, total packets, as well as the share of packets received as part of *bursts* (recall our sampling from § 3.1) for our entire dataset. At a high-level, we realized it was illuminating

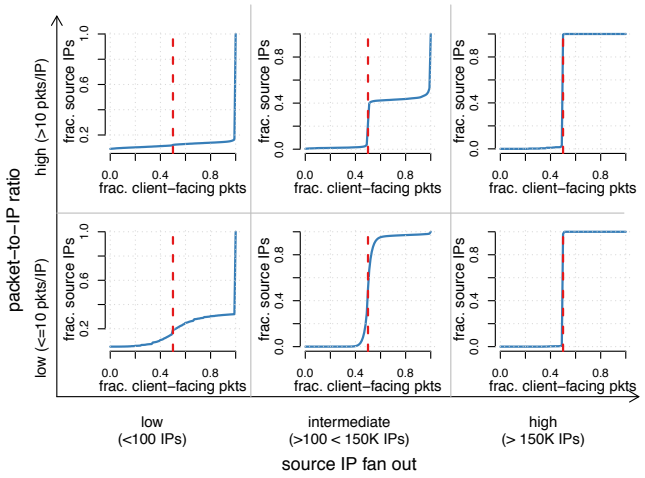


Figure 4: Per source IP: Fraction of packets targeting client-facing IPs, i.e., addresses exposed via forward DNS to serve CDN content. The red dashed line indicates that packets from a source are balanced over client-facing and operations IP addresses.

to partition source addresses by their fan-out into three regimes: "few," "in-between," and "almost all" destination IPs, and to partition the packet-to-IP ratio into two regimes: "few," and "more than few." We experimented with various cutoff points that would capture our intent, and found that it did not matter w.r.t. the subsequent results, and thus we chose values that would convey their ballpark nature. We split source IP addresses by a fan-out (x-axis) exceeding 100 IPs, and exceeding 150K IPs. And we picked 10 for the cutoff for the packet-to-IP ratio (y-axis).

Distribution of IP and packet counts: Figure 3b shows the number of source IP addresses and total packets sampled for each individual tile. We find that the vast majority of source IP addresses falls in the lower left tile, contacting only a small number of machines (39% of source IPs in this tile hit a single telescope IP) and send a small number of packets (41% of source IPs in this tile sent fewer than 5 packets). While this lower left tile comprises the majority of source IPs, these IPs are unlikely to be scanning the address space and account only for 3% of packets. On the other end of the spectrum (top right tile), we see some 423 source IP addresses (about 0.001% of source IPs), yet accounting for some 28% of the overall packets. We note that only a tiny fraction of packets from sources contacting more than 100 telescope IPs are logged in a burst state. Most burst packets are sent by sources hitting a small number of machines with a large number of packets (top left tile), consistent with DDoS attacks to flood individual machines.

4.2 CDN-targeted vs. CDN-agnostic

From the example machines shown in Figure 2, we know that some traffic components target exclusively the client-facing IP, as exposed via DNS to clients, while others target client-facing and operations IP addresses equally. We next leverage this observation to classify the behavior of source IP addresses. To this end, we

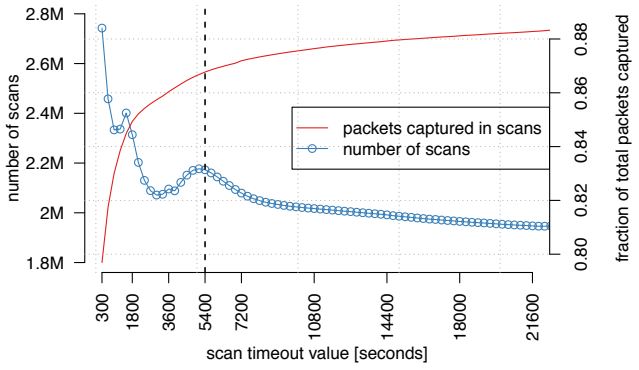


Figure 5: Number of detected scans and fraction of all packets captured in scans for different timeout values.

calculate for each source IP address the ratio of number of packets sent to telescope addresses that are client-facing IPs divided by the total number of packets from that source address. Figure 4 shows, for each of our 6 partitions of source IP addresses, a CDF of this ratio. Source IP addresses that target only a small number of machines almost exclusively direct traffic at the client-facing IPs, the two left-hand tiles. Furthermore, in the top left-hand tile, where a comparatively large fraction of packets were logged when the token-bank was in the burst state (recall § 4.1), some of the sources are likely involved in Denial-of-Service attacks, nominally directed against a customer of the CDN, i.e., attacks targeted at a domain name. Sources that target a large number of the telescope addresses, the two right-hand tiles, are exclusively CDN-agnostic: The traffic is balanced across client-facing and operations IP addresses. The middle, top tile is an intriguing mixture of the left and righthand tiles: the sources, which hit a subset of machines and with many packets per IP address, are evenly divided into CDN-targeted and CDN-agnostic source behavior. Within this middle tile, CDN-agnostic sources typically hit more destination IPs than the CDN-targeted sources. In particular, for sources in this tile that hit more than 1,000 telescope IPs, more than 92% show CDN-agnostic behavior. We will further scrutinize the behavior of these addresses in Section 5.

Share of CDN-targeted traffic: Source IPs that send at least 99% of their packets to client-facing IPs account for some 9.9% of the overall logged traffic, and comprise some 64% of all source IP addresses seen. This traffic was attracted to our telescope solely due to the telescope containing addresses that are the A record in replies to forward DNS requests. Given that most packets logged during *burst* state are targeted at client-facing IP addresses, we point out that the actual share of this CDN-targeted traffic is likely much higher, but is not logged due to the token-bucket sampling. In terms of our overall dataset, however, we note than some 90% of logged traffic is *CDN-agnostic*, i.e., not targeted at the CDN's customers.

4.3 Identifying Scans

So far, we considered all packets sent by a source throughout our observation period, November 2018. Next, we seek to isolate individual *scan* events.

Scan definition: In our work, a scan, by a given source address, consists of sequence of probes that hit at least n distinct destination IP addresses, and the interarrival times of the probes to any address of the telescope are less than a given *timeout* interval.[5] In this work, we choose $n = 100$, since we introduce metrics in the following sections that require a certain number of packets to yield distinctive results, and we found 100 destination IPs to be a good compromise between capturing small-scale scans, as well as providing enough packets for characterizing individual scans. Having n fixed, we next need to decide on a *timeout* threshold.

Timeout settings: Figure 5 shows the number of identified scans in our dataset (blue dotted line), as well as the fraction of total packets that are captured within scans (red line) for alternative values of the timeout threshold. A larger timeout value necessarily aggregates a larger number of packets into scans, hence the monotonic increase of packets that we classify to belong to scans (red line). Overall, we note that a timeout value of 300 seconds already groups more than 80% of all packets into scans and that increasing the timeout value has only a comparably small effect on the fraction of packets considered as scans. The relationship between timeout value and *number* of scans (blue dotted line) is determined by two phenomena: A larger timeout value aggregates individual scans together, resulting in a decreasing number of individual scans. At the same time, since we require at least 100 packets in a scan, a larger timeout value yields more scans from source IPs that send packets at an overall low rate, i.e., they would not be considered a scan for shorter timeout values. We see the number of scan events decreasing rather rapidly for aggregation values between 300s and 3600s (one hour) and note a slight increase in identified scans at 5400 seconds (1.5 hours). We note that this value is dependent on our choice of $n = 100$ destination IPs, and not a natural property of the dataset. In the following, we choose 5400 seconds as our timeout threshold, as a reasonable compromise in obtaining both a high number of scans and a high fraction of probes classified into scans.

Detectable scanning rate: To put this 5400 second timeout in perspective; suppose a source is scanning the "full" IPv4 space in random order. If the source sends probes slowly, then our telescope has less of a chance of detecting this activity as a scan. Assuming a full scan means all IPv4 ranges that are publicly routable, which is 3.7 billion addresses, and given our telescope of 178K addresses (and some simplifying assumptions), then, if the source sends probes at 30 packets per second, then the telescope will classify the activity as a scan with probability 0.95; if 50 pps, then the probability is 0.9998. Durumeric et al. [18] found that 95% of scans they detected were conducted at rates of at least 100 pps, making us confident that our timeout settings capture a large majority of scanning activity, when assuming random scanning order.

Total identified scans: In total, we identified 2.2M scans, and which contain 87% of all logged traffic in our dataset. We note that less than 1% of packets classified belonging to scans were logged when the token bank was in the *burst* state, § 3.2. Seen on a per-scan basis, 71% of scans had 0 packets as part of bursts, 98% of scans had less than 1% of their packets received in burst state, and 99.7% of

[5] We do not require a scan to be on a fixed port number, and study port prevalence and distributions in Section 6.2.

	scans	packets	source IPs
all scans	2.17M	16.87B	1.14M
Internet-wide full	2.8K (0.1%)	27.6%	1.3K (0.1%)
Internet-wide partial	1.4M (66.1%)	39.3%	845K (73.8%)
localized	693.5K (31.9%)	29.0%	331.6K (29.0%)
CDN-targeted	40.6K (1.9%)	4.1%	19.7K (1.7%)

Table 2: Identified scans and their target selection strategies.

scans had less than 10% of their packets received in a burst state. We are thus confident that we do not falsely identify attacks targeted at the CDN machines as scans. In the remainder of this paper, we focus on traffic identified as scans, unless otherwise noted.

5 SCAN TARGET SELECTION STRATEGIES

In this section, we introduce tools and analysis to classify individual scans into different *target selection* categories. The results of our classification are shown in Table 2.

5.1 Internet-wide full IPv4 Scans

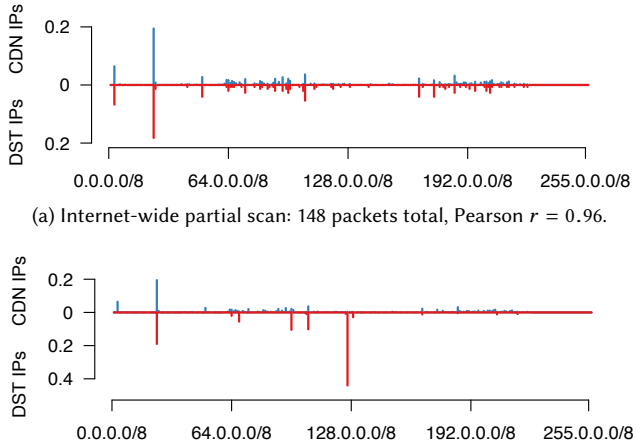
The first, and most straight-forward to detect, target selection strategy are full scans of the IPv4 address space. Recent advancements in scanning tools make scanning of the full space possible, even within hours. We detect full scans by checking if a scan targeted at least 150K of the 178K telescope addresses. This includes both *client-facing* and *operations* IP addresses. Here, we leave some leeway to account for both packet loss, network-specific filtering [9], as well as for servers that might be in a maintenance mode, such as for a system software update. In our dataset, we find a total of 2.8K full scans, originating from 1.3K source IP addresses. While this number is comparably small, we highlight that full scans of the IPv4 space account for more than 27% of all the scanning traffic.

5.2 CDN-targeted Scans (Domain Scans)

The second target selection category are scans that exclusively or primarily target the client-facing IP addresses of the CDN. While we find that the vast majority of scans are equally distributed across client-facing and operations IP addresses, we do find some 1.9% of scans, where the fraction of probes to the client-facing IP is close to 1. Since only the client-facing IP is returned on forward DNS lookups (and never the operations IP), these scans are likely the result of hosts scanning *domain names* (e.g., a host resolving the Alexa top list) We call such scans CDN-targeted, or domain scans.

Note that since the DNS resolution of a given domain name will change over time, and different domain names have different resolutions, scans of a large number of domains can and will result in the CDN resolving requests to different servers, which results in sources hitting different destination IPs. Also, a large number of client-facing addresses could be gathered by scanners that resolve domain names from different locations. Thus, it is reasonable from such activity to reach the threshold of 100 telescope IPs. We set our threshold for a scan to be CDN-targeted if the client-facing-to-total traffic ratio exceeds 0.8,[6] and identify some 40.6K CDN-targeted scans from some 19.7K IP addresses.

[6] As per Figure 4, we note that the client-facing-to-total ratio is either very close to 0.5 or very close to 0.1. Our threshold of 0.8 for classification thus works well.



(a) Internet-wide partial scan: 148 packets total, Pearson $r = 0.96$.

(b) Localized scan: 827 packets total, Pearson $r = 0.38$.

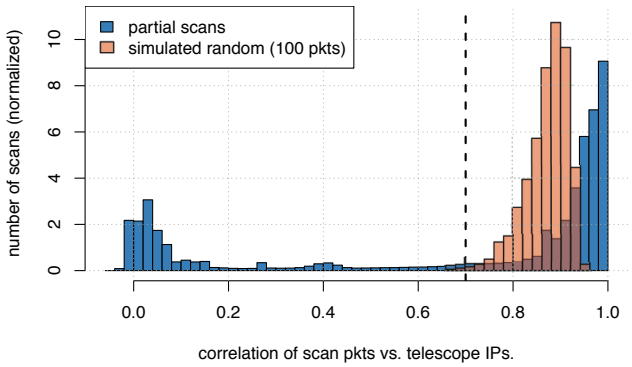
Figure 6: Determining if scans target a random set of destinations: We correlate the fraction of telescope addresses per /8 (positive y-axis) against the fraction of packets received per source per /8 (negative y-axis). Internet-wide full or partial scans of the IPv4 space result in a high correlation.

5.3 Internet-wide partial Scans

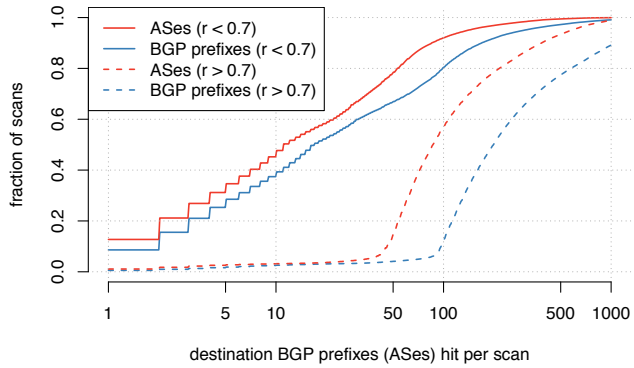
As a complement to full scans, partial scans are scans that probe fewer than 150K telescope addresses. And, with CDN-targeted scans in their own target-selection category, we partition the remaining CDN-agnostic partial scans into two other categories: *Internet-wide partial scans* and *localized scans*, based on whether they are, or are not, consistent with being a randomly selected subset of addresses of the full IPv4 address space.

Dissecting Internet-wide from localized scans: To distinguish Internet-wide from localized scanning strategies, we leverage the insight that an Internet-wide scan targeting a random subset of the IPv4 space will, necessarily, also hit a random subset of individual server IP addresses of our telescope. We first partition our CDN server addresses into /8 bins, and compute the fraction of addresses in each bin. Then we assess—for each scan—how well the distribution of scan packets over /8 bins correlates with the distribution of our server addresses. Figure 6 shows this by example. Here, the source IP in Figure 6a sends 148 packets, and the distribution of these packets across /8 bins correlates well with the distribution of the telescope IPs resulting in a high Pearson correlation of $r = 0.96$. In contrast, the source IP in Figure 6b sends some 827 packets in a scan, and most of these packets are destined to machines in two /8s (see red negative spikes). The correlation for this scan is much lower at $r = 0.38$. This destination pattern is caused by selecting a non-random subset of the IPv4 space.

Figure 7a shows the histogram of Pearson correlation, r, for all partial scans in our dataset (blue bars). We see that the distribution of correlations is strictly bi-modal: Either scans have a high correlation, close to 1, or a very low one. To show the validity of our correlation-based approach and to find a sensible cutoff point, we simulated 1000 iterations of an Internet-wide scan targeting 100 randomly chosen destination IPs (red bars). We chose 100, since



(a) Pearson correlation between scan packets and telescope IPs for partial scans. Random scans have high correlation $r > 0.7$, as tested with simulated random scans with 100 pkts. 32% of partial scans have a low correlation. (Both histograms normalized so that the total area below bars equals 1).



(b) Number of BGP prefixes and routed ASes hit by partial scans, comparing Internet-wide ($r > 0.7$, dotted lines) against localized ($r < 0.7$).

Figure 7: Dissecting Internet-wide from localized scans.

this is the minimum number of destination IPs in our definition of a scan. We find that more than 99% of our simulated random scans have correlation above 0.7. Hence we set 0.7 as our cutoff point for scans to be considered Internet-wide, consistent with a random subset of the IPv4 space and hence our telescope IP addresses.[7]

In total, we identify 1.4M Internet-wide partial scans (Pearson r greater 0.7), 66% of all detected scans and 39% of all scanning packets, originating from 845K source IP addresses.

5.4 Localized Scans

We classify scans as *localized* if they are partial and have a Pearson correlation r lower than 0.7. Figure 6b shows such an example. These scans do not target a random subset of the IPv4 space, but use some other strategy for target selection. We refer to these scans as localized, since the visibility of these scan is dependent on the position of the vantage point in the IPv4 address space. We note that the classification of a scan as localized, per our definition, does not necessarily imply that the scanner targets a tightly confined region of the address space, but only that the scanner does *not* target a random subset of the IPv4 space.

To assess the scope of the address space that is targeted by different localized scans, we show in Figure 7b the number of routed BGP prefixes (as well as ASes) that individual scans hit, contrasting Internet-wide partial scans against localized scans. Here, we can see that visibility of localized scans is often confined to particular regions of the address space. Some 34% of these scans target at most 10 routed prefixes, and only 20% hit more than 100 routed prefixes. This is in stark contrast to partial Internet-wide scans (dashed lines), where over 90% hit more than 100 routed prefixes (recall that our telescope is distributed across some 2,800 routed prefixes). We also aggregated scans by the number of unique ASes hit (i.e., hitting any routed prefix originated by an AS), and which show a slightly higher concentration when compared to routed prefixes (some 42% of scans hit 10 or less ASes). We also tried simple covering prefixes (/16 and /8 prefixes, not shown) as alternative viewpoints of spatial target locality of scans, but found them to not yield better aggregation results as compared to BGP routed prefix and AS aggregation. We find that many localized scans share the commonality of targeting narrow ranges in the address space, and we will further study the properties of these scans in the following sections.

In total, we identify some 693K localized scans, contributing some 29% to all scan traffic, originated from some 332K source IP addresses.

6 SCAN PROPERTIES

With our scan classification in hand, we next drill into the properties of our identified scans. In particular, we study timing of scans, stateful behavior of repeated scans, services targeted, origins of scanner IP addresses, as well as presence of sharded scans.

6.1 Timing Aspects and repeated Scans

Scan timing: Figure 8 shows, for all identified scans, the duration (x-axis), versus the total packets (y-axis) of the scan. We can see that most scans are short in duration, while other scanners are active for the entire period of November 2018, contributing a large number of packets in each scan period. The horizontal structures in the plot show scans at different scanning rates. Scans with higher scanning rates are more likely to finish within a shorter period of time, while we see some low-rate scanning activities that spread across our time window. Notably, we observe concentrations of scan duration at fixed intervals of single days, and multiple days, such as one week. This richness in pattern motivates us to study repeated scans of a given source.

Repeated scans per IP address: We next assess if sources carried out multiple scans in our time window and study the similarity of these scans. Table 9a shows the number of scans per source IP address, where we group source IP addresses into those that only carried out scans of a specific target-selection category (Internet-wide partial, localized, etc.), and source IPs whose scans are in multiple categories (multi). Almost a quarter of all source IPs carried

[7] For statistical tests of sampling distribution equality, we considered the Chi-squared test, and the Kolmogorov Smirnov test. However, we found that for large scans with 10, 000+ or even millions of packets, these test statistics yielded erroneous significant difference, due to the known issue that in very large samples, p-values can approach zero [25]. An alternative was to use Cramer's V measure, which scales the Chi-square statistic, and yielded essentially equivalent results as the Pearson correlation. We chose the latter as we find it simpler and more intuitive.

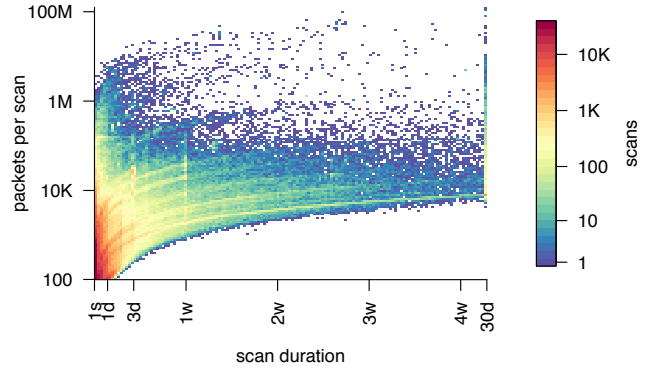
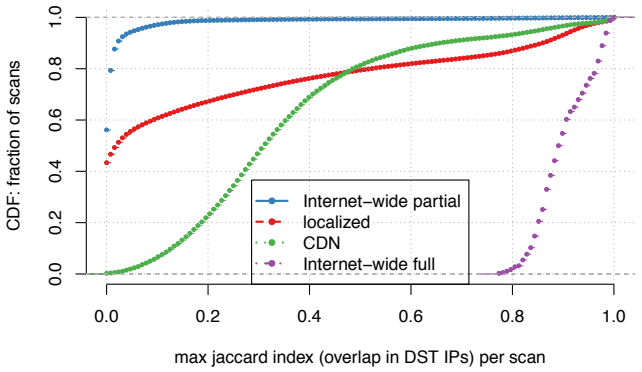


Figure 8: Duration and number of packets for individual *scans*. Horizontal structures represent scans at different rates, vertical structures are scan periods that last for specific durations, e.g., a day, or a week.

source IPs	single scan	2 scans	3 - 10 scans	> 10 scans
ALL	877K (76.7%)	112K (9.7%)	135K (11.8%)	21K (1.8%)
IW full	510 (73.7%)	86 (12.4%)	89 (12.9%)	7 (1.0%)
IW partial	647K (81.5%)	66K (8.3%)	75K (9.4%)	6K (0.8%)
localized	218K (78.0%)	31K (11.1%)	27K (9.7%)	3K (1.2%)
CDN	12K (66.5%)	3K (15.4%)	3K (16.3%)	312 (1.7%)
multi	/	12K (22.9%)	30K (56.5%)	11K (20.5%)

(a) Number of scans detected per source IP and scan type.



(b) Maximum similarity of targeted destination IP addresses per scan, per target-selection category.

Figure 9: Scans carried out per source IP, and similarity of destinations in different scans by a given source IP.

out more than one scan during our time period, and some 1.8% of source IPs carried out more than 10 scans. We see this behavior relatively uniformly across scan types, with the exception of *multi*; in this category more than 20% carried out more than 10 scans.

Measuring scan target similarity: To assess if repeated scans, by a given source IP, target the same or a similar set of addresses, we calculate for each scan its *maximum similarity*, the largest fraction of destination addresses that a scan shared with any other scan from the same source IP. In particular, for each scan i carried out by the same source S, we calculate the Jaccard indices over the sets of targeted destination IP addresses of all other scans j from source S, where $i \neq j$, and where the Jaccard index is the size of the intersection of the two sets divided by the size of the union. We define the maximum of these Jaccard indices as the *maximum similarity* value of the scan i. A similarity value of 1 indicates that a source scanned the exact same set of destination IPs, while a similarity value of 0 indicates that no single destination IP address was scanned in a different scan.

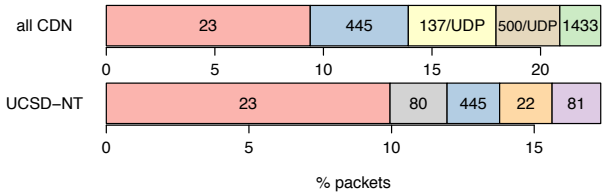
Statefulness in repeated scans: Figure 9b shows the maximum similarity for scans of the different target selection categories. Full scans do show, necessarily, high maximum similarity values, since these scans target a vast majority of telescope's addresses (recall that we define a full scan as > 150K out of the 178K IPs, hence the similarity index is not necessarily 1). Also CDN-targeted scans (which consist of 1.9% of the scans, Table 2) overall have higher similarity values compared to other partial scans. This observation supports our hypothesis that most of these scans indeed target domain names, since the DNS resolutions will sometimes be to the same CDN server address (and sometimes not, and the level of consistency will vary for different domain names). The most interesting observation from Figure 9b, however, is the difference between partial Internet-wide and localized scans, which collectively are 98% of the scans. Internet-wide partial scans show very low similarity values (only 1% have a value higher than 0.2), evidencing that repeated scans by the same source are typically stateless and target "fresh" sets of random IP addresses instead of targeting the same set of destinations, e.g., from a hitlist. Localized scans, however, overall have much higher similarity values: over 30% have a value higher than 0.2 and 19% have a value higher than 0.6. Thus, sources conducting localized scans are much more likely to repeatedly target the same address blocks / destination IP addresses, suggesting the use of either hitlists consisting of addresses or address ranges, or other forms of stateful scanning strategies.

6.2 Services Scanned

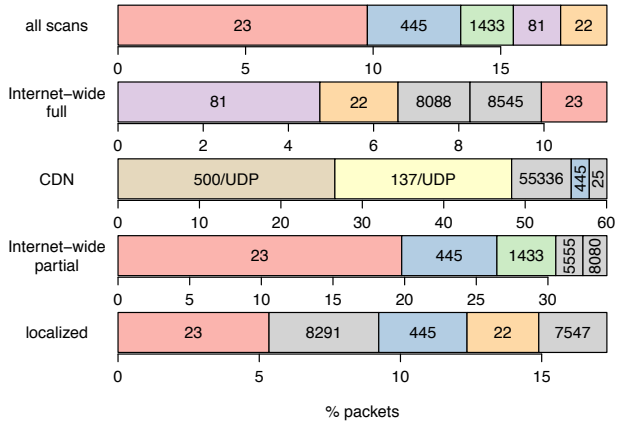
Next, we investigate popular port numbers of our identified scans. We complement our perspective on port prevalence with one month of data collected in November 2018 in a /8 darknet (UCSD-NT), operated by CAIDA/UCSD and available to researchers [2]. We will further describe and compare the visibility of both vantage points in more detail in § 7.

CDN vs. UCSD-NT overall ports: Figure 10a shows the top-5 ports, in terms of packets, for our overall dataset (not filtered for scans), as well as the overall distribution of packets arriving in the UCSD network telescope. Figure 10b shows the port distribution for our identified scans. All barplots show the top-5 port numbers and the percentage of packets contributed. Comparing the overall port distribution in our dataset against UCSD-NT, Figure 10a, we notice that the CDN servers log a much higher fraction of UDP traffic on port 137 and 500 (which we will discuss below). We note, however, that the UCSD-NT port distribution closely matches the port distribution of our overall detected scans (top bar in Figure 10b). Here, 4 of the 5 top ports overlap, with the exception of port 80 (which the CDN telescope does not collect) and port 1433.

Contribution of Internet-wide vs. localized: Note that port 23 (Telnet) and 445 (Remote Desktop Protocol) dominate the overall



(a) Top-5 port numbers seen in all our dataset, as well as in the UCSD darknet.



(b) Top-5 port numbers seen for *scans*, where we show all scans as well as the contributions of individual scan types.

Figure 10: Top-5 port numbers in our dataset, UCSD network telescope, and contribution by individual scan types. All ports are TCP unless tagged as UDP.

port distribution, and these are the most popular ports for partial Internet-wide scans. This makes sense given that these two ports are frequently targeted by worms and botnets, which commonly exhibit random scanning behavior [8]. Localized scans, on the other hand, contribute much less to this class of ports, but show significant activity on ports 8291 and 7547, both ports related to critical vulnerabilities in home routers [26, 27]. In contrast, port 8291 is the 273th most targeted port in the UCSD darknet, and port 7547 the 48th most targeted port by packets. This hints towards actors scanning particular ranges for these ports (recall that many of the CDN servers are located within end-user ISPs, making these ASes prime candidates for scanning for home router vulnerabilities). Generally, we note that localized scans show a higher port diversity (top 5 ports only account for some 17% of all scan traffic), whereas Internet-wide partial scans are more concentrated on a smaller number of ports (top 5 ports account for some 33% of all scanning traffic). We note that when looking at aggregate port statistics, potentially dangerous, localized scans such as on ports 8291 and 7547, would not stand out, since they are masked by the sheer volume of Internet-wide scans on popular port numbers.

CDN-targeted scans: We note that CDN-targeted scans show wildly different behavior, with the two port numbers being 500/UDP (IPsec) and 137/UDP (NetBIOS name resolution). These two port numbers also show up with much higher frequency in our overall dataset, as compared with the UCSD telescope (Figure 10a). Since

rank	Internet-wide full		Internet-wide partial		localized	
1	Ukraine	41.0%	Ukraine	19.0%	Netherlands	18.7%
2	U.S.	14.0%	China	13.9%	Bulgaria	10.6%
3	Netherlands	9.0%	Netherlands	11.5%	U.S.	8.3%
4	China	5.2%	U.S.	11.3%	Russia	7.8%
5	U.K.	4.2%	Russia	7.4%	China	5.6%

(a) Scan packets.

rank	Internet-wide full		Internet-wide partial		localized	
1	U.S.	38.2%	China	20.3%	Brazil	16.2%
2	Netherlands	8.7%	Egypt	8.3%	Russia	12.7%
3	U.K.	7.3%	Russia	8.2%	India	10.3%
4	China	6.9%	Brazil	7.3%	China	8.5%
5	Japan	6.8%	India	3.5%	Taiwan	5.9%

(b) Scan source IP addresses.

Table 3: Top origin countries of scan traffic and sources.

some machines, unsuccessfully, try to establish an IPsec connection and/or NetBIOS name resolution upon establishing a TCP connection (see, e.g., reports [4, 5]), we believe that the majority of these packets do not resemble actual scans or exploits of these port numbers. A more likely explanation is that these packets are a connection artifact of hosts that scan/scrape actual websites (e.g., Alexa Top 1M), hence accessing other services on our machines, most likely Web.[8]

Port co-dependency: We do not restrict our definition of scans by port number, i.e., a scan can target multiple ports. For 53% of detected scans, we logged packets on more than a single port number. 66% of Internet-wide partial scans send packets on multiple port numbers and the most frequent combinations are two-port tuple 23 and 2323 (TCP) as well as 23 and 8080 (TCP), together accounting for 64% of all Internet-wide multi-port scans. These port combinations can be attributed to incarnations of the Mirai botnet [8]. In contrast, only 26% of localized scans probe multiple port numbers, with the most common combination of ports being the 4-tuple of 23, 8080, 7547, 8291 (TCP), accounting for only 19% of all localized multi-port scans.

6.3 Scanner Origins

In this section, we inspect top contributors of scan traffic and source addresses, and assess the effect of sharded scans on our inferences.

Scanner origin countries: Table 3 shows the top-5 origin countries of scans.[9] While Internet-wide full scans are highly concentrated, with Ukraine accounting for some 41% of scan packets, followed by the US with some 14%, we see that partial Internet-wide scans are more spread out, and find that localized scans are even more spread across different origin countries, with the Netherlands accounting for some 19% of localized scan traffic. When looking at unique scan source IP addresses (Table 3b), we see similar distributions for Internet-wide scans, but notice that localized scanner IP addresses show a much higher concentration in Brazil, Russia, and India; countries that show less-pronounced scanning activity when considering only Internet-wide scans. We note that a large

[8] Such connection artifacts, e.g., IPsec, are frequently reported to appear in firewall logs in production networks [4].

[9] We leverage the CDN's proprietary geolocation database to map scanner IP addresses to countries.

rank	Internet-wide full		Internet-wide partial		localized	
1	Hoster (UA)	40.8%	Hoster (UA)	18.5%	Hoster (NL)	8.3%
2	Hoster (US)	5.7%	Hoster (MD)	4.9%	Hoster (BG)	4.2%
3	Hoster (NL)	3.6%	ISP (CN)	4.6%	Hoster (UA)	4.1%
4	Research (US)	3.0%	Hoster (BG)	3.3%	Hoster (NL)	3.4%
5	Hoster (NL)	2.4%	ISP (RU)	3.2%	Hoster (NL)	3.1%

(a) Scan packets.

rank	Internet-wide full		Internet-wide partial		localized	
1	Hoster (NL)	26.7%	ISP (CN)	8.5%	ISP (BR)	9.5%
2	Hoster (US)	5.3%	ISP (CN)	8.4%	ISP (RU)	7.9%
3	Research (US)	5.0%	ISP (EG)	4.2%	ISP (IN)	6.1%
4	Hoster (UK)	3.8%	ISP (RU)	5.4%	ISP (TW)	5.5%
5	Hoster (IT)	2.9%	ISP (TW)	2.7%	ISP (CN)	4.3%

(b) Scan source IP addresses.

Table 4: Top origin ASes of scan traffic and sources.

fraction of localized scans in Brazil corresponds to scans on port number 8219, and point out that Brazil was reported to be a key contributor of Internet-wide scan traffic on port 8219 caused by infections by the Hajime botnet [3, 23]. While Brazil still emits significant Internet-wide partial scans (accounting for some 7% of source addresses), we now see much more pronounced localized scanning activity. We plan to investigate potential links between these scanning behaviors and related events in future work.

Scanner origin networks: In Table 4, we show the types of the top-5 origin ASes of scan traffic, qualified by their respective country code. Indeed, we find that almost all scan traffic from Ukraine can be mapped to a single hoster in this country, accounting for more than 40% of all full scan packets. Eyeballing localized scans, we find them to be much less concentrated on individual ASes, with the top-5 only accounting for some 23% of scan packets, compared to more than 50% for full scans, and 32% for Internet-wide partial scans. The observation that most scanning traffic originates from (bulletproof) hosting ASes is largely consistent with findings in [18]. However, when counting by the number of scanner source IP addresses (Table 4b), the picture changes: ISPs connecting end users to the Internet host a majority of partial as well as localized scanner IP addresses, hinting towards scans originating from, e.g., infected devices as opposed to large-scale measurement campaigns.

Sharded scans: Some scanning tools, such as ZMap, allow *sharding*, i.e., to distribute and execute scans using multiple machines and, more importantly, multiple IP addresses. To assess if our Internet-wide partial and localized scans are in fact sharded scans of, e.g., the full IPv4 space, we re-computed our scan target selection strategies, but aggregate scan traffic from source IP addresses, if *(i)* they are in the same /24 prefix, and *(ii)* they carry out only Internet-wide partial or only localized scans. We find that the overall proportion of *scans* only marginally changes, but the fraction of *packets* categorized as Internet-wide full scans increases from 27.6% to 37.7%, and the fraction of packets categorized as Internet-wide partial scans decreases from 39.3% down to 30.3%. This finding suggests that some of the large Internet-wide partial scans are in fact sharded scans of the full IPv4 space. Localized scans barely aggregate, with the fraction of packets categorized as localized decreasing only marginally from some 29.0% down to 27.9%, suggesting that the vast majority of localized scans are not artifacts of sharded scans.

7 ON SCANNING VISIBILITY

In this section, we compare the visibility into scanning activity as seen from our distributed telescope, and the UCSD network telescope, a centralized /8 darknet. We assess baseline activity in both datasets, and proceed to dissect baseline activity in our dataset.

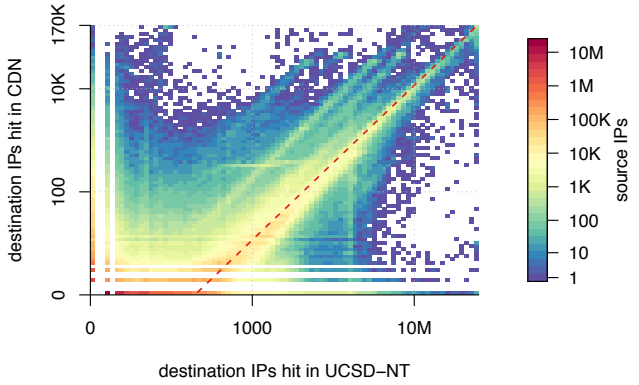
7.1 Comparison with a Darknet

In this section, we compare source visibility and background radiation seen in our CDN logs with a major darknet operated by CAIDA/UC San Diego, who make their data available to the research community [2].

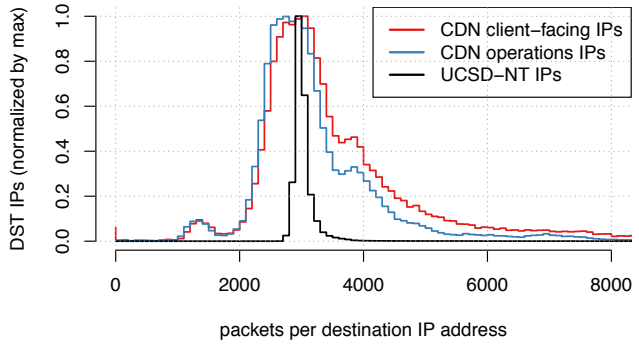
CDN telescope vs. UCSD-NT: The UCSD darknet consists of an entire routed /8 prefix (with a small number of subdelegated blocks). I.e., the *surface* of UCSD-NT is some 15.3M addresses, about 86 times larger than our telescope, and thus a random probe is 86 times more likely to hit the UCSD telescope. We compare the first week of November 2018. During this time period, our telescope logged some 4.5B packets, while UCSD-NT logged some 345B packets, which is 77 times (i.e. less than 86 times) the packets we log. Thus, on average, the CDN telescope logs more packets per IP address when compared to UCSD-NT.

Per source comparison: Figure 11a shows for all source IP addresses (here we show the entire dataset) the number of destination IPs hit in both telescopes. We can segment this plot in three groups: *(i)* source IP addresses hitting destinations in both telescopes proportionally, see concentration on diagonal line. These are source IP addresses that show up with high Pearson correlation in our dataset, and we expect this pattern, since a random scan of the IPv4 space will hit, on average, about 86 times more IP addresses in the UCSD telescope, when compared to the CDN. The dashed red line shows source IP addresses that hit a ratio of 1:86 IPv4 addresses in CDN/UCSD-NT. *(ii)* A second group of IP addresses that show up proportionally more frequently in the UCSD telescope, see area below the diagonal. Recall that sources that probe on port 80 or 443 or ICMP are not logged by the CDN telescope, and given the size difference of the telescope, a large number of small-scale scans will only be visible in this much larger telescope, *(iii)* Lastly, we see a concentration of source IP addresses that hit exclusively (i.e. along the y-axis), or predominantly the CDN telescope. We expect these source IP addresses to carry out localized scans, and will follow up on this observation in the next section.

Per destination comparison: The picture sharpens when studying the number of packets per destination IP in either dataset. Figure 11b shows a histogram of the number of packets per telescope destination IP address for the 1st day of November 2018 (other days look similar). We group the addresses into CDN client-facing IPs (N=86K), CDN operations IPs (N=86K), and UCSD-NT IPs (N=15.3M). To allow comparison given the vastly different counts, we normalize the histogram by showing the largest bin as 1 for all datasets. We see that the UCSD-NT IP addresses show a very pronounced concentration at $x \approx 3000$ packets per day (median = 3091 pkts), and that only very few IP addresses in this darknet receive significantly more packets on a daily basis (90^{th} percentile = 3292 pkts, 95^{th} percentile = 4552 pkts). This confirms our earlier observation of a *baseline* of roughly 3000 packets that our machines receive on a daily basis (recall Figure 2). The IP addresses of our



(a) Per source IP: Destination IP addresses hit in the CDN telescope and UCSD telescope in the first week of November 2018. We see random and full scans hitting both telescope proportionally (red dashed line), and traffic components only visible in either of the two telescopes.



(b) Packets per destination IP on November 1st for CDN telescope and UCSD-NT. The CDN IP addresses have a heavier tail, while UCSD telescope IPs are heavily concentrated at the baseline value of around 3000 packets.

Figure 11: Visibility of CDN telescope vs. UCSD-NT darknet.

CDN telescope also show the peak at the baseline (median operations IPs = 3092 pkts, median client-facing IPs = 3331 pkts). Some IP addresses receive less than a typical UCSD-NT IP address, and we recall that some of our servers are located in networks that filter traffic on specific port numbers, and that we do not log traffic on port 80, 443, nor ICMP traffic. More interestingly, however, we note that CDN IP addresses show a much heavier tail (operations IP 90th percentile = 4509 pkts, 95th percentile = 5595 pkts) which is even more pronounced for client-facing IP addresses (90th percentile = 6150 pkts, 95th percentile = 8187 pkts). Recall that we identified some 87% of all logged traffic to be related to scanning activity (Section 4.3). Thus, the additional traffic that the CDN telescope logs is mostly related to scans — the result from localized scans, as we will show in the next section.

7.2 Baseline Composition

Impact of scanning on baseline radiation: Our findings from both our telescope as well as from the UCSD darknet show that virtually *all* IPv4 addresses receive a baseline of roughly 3000 packets on a daily basis, as of November 2018. With our tools to detect and

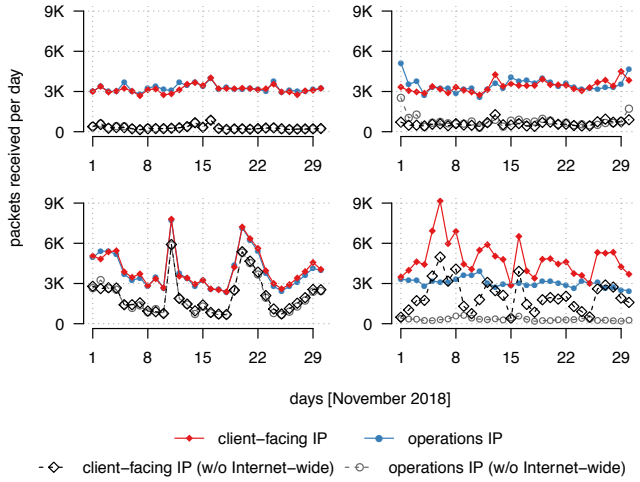


Figure 12: Effect of Internet-wide scanning activity on background radiation: The baseline of unsolicited traffic reduces by some ≈90% when removing packets resulting from Internet-wide scans, while localized spikes persist.

characterize scanning activity in hand, we now re-visit this observation. Figure 12 shows our four example CDN machines featured in Section 3. Here, we show both the packet totals on client-facing and on operations IP address, as well as after filtering out source IP addresses that carry out full or partial Internet-wide scans (we exclude source IP addresses that have a Pearson $r > 0.7$ over the entire time period). Here, we can see that the traffic logged by the two machines on top (only baseline radiation) reduces by more than 90%, leaving only some 10% of logged packets that can not be attributed to Internet-wide scanners. The bottom-left example shows CDN-agnostic spikes, and we note that removing random scan activity shifts the number of logged packets by some ≈ 2.8K packets, but that the spikes remain visible; they are not part of random scans. The last example shows that the baseline radiation received on the operations interface mostly cancels out when removing scanners, while CDN-targeted packets remain visible, yet again shifted by some ≈ 2.8K packets.

Baseline and position in the address space: We conclude with the following observations: *(i)* Routed IPv4 addresses receive a baseline of some 3000 packets on a daily basis as a result of Internet-wide scans of the IPv4 space. Thus, when seen on a per-IP basis, a large majority of background radiation in today's Internet is a result of such scanning activity, and the positioning of the vantage point in the IPv4 space does not affect this baseline. *(ii)* We see localized scanning activity that we can clearly separate from the baseline caused by Internet-wide scans. The visibility of localized scanning activity is heavily dependent on the location of the telescope IP addresses in the IPv4 space.

7.3 Long-term Baseline Evolution

In this study, we focus on the month of November 2018. While our findings of the baseline scan radiation are remarkably stable over the course of this month, we next ask if this observation holds on longer timescales and/or if there are long-term trends when it

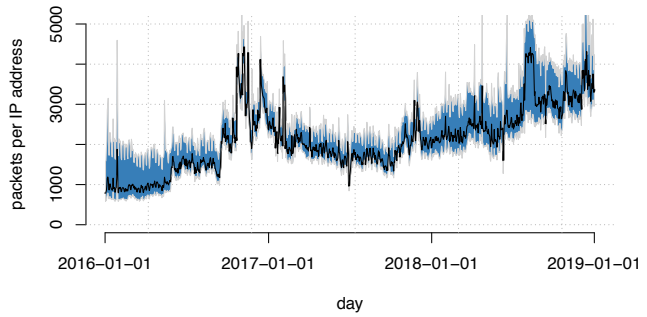


Figure 13: Median and 25th / 75th percentile of daily logged packets per IP address over 3 year period.

comes to baseline radiation. Figure 13 shows, for three years, the daily average number of packets logged at the firewall of each CDN server on each public-facing IPv4 address. Since we are interested in a robust estimate of baseline radiation, we show the median, as well as the 25^{th} and 75^{th} percentile, removing machines that received significantly more or less traffic than the average case. Over the course of three years, we see a 3-fold increase in baseline activity, starting from $\approx$ 1000 packets in early 2016, to some $\approx$ 3100 in December 2018. There is a pronounced spike in late 2016, early 2017. This spike coincides precisely with the widespread infections of the Mirai botnet, which resulted in increased scanning activity, evidenced in another network telescope [8]. Mirai activity leveled off in 2017, but we see a steady increasing trend of baseline radiation over the course of the next 1 1/2 years. If trends of the last years continue, we can expect overall scanning activity, and the resulting baseline radiation, to continue to increase in the foreseeable future.

8 DISCUSSION

In this section, we first discuss pertinent implications of our work for researchers and network operators, and introduce avenues for future work.

8.1 Implications

Our work has implications both for the research community, as well as practical implications for network operators.

Interpretation of scan data: The perhaps most important finding of our work is the evidence of widespread *localized* scanning activity in today's Internet. Such scans pose a potentially greater threat, since these scanning actors may target individual networks and hosts, as opposed to Internet-wide scans, which is an important consideration, e.g., when leveraging darknets to track phenomena such as Botnets or exploitation of vulnerabilities (e.g., [8, 12, 20]). While darknets, especially large ones, provide excellent visibility into random scanning activity, we find that they severely underestimate the number and volume of localized scans carried out, and may miss the sources behind these scans entirely. Even when leveraging *live* vantage points for monitoring scanning traffic, such as our CDN telescope, potentially dangerous scans such as on ports 8291 and 7547 (recall § 6.2) would not stand out in aggregated statistics, but only become visible once we identify and isolate localized scans as a distinct category. Our introduced tools and metrics to identify, isolate, and characterize scanning activity allow for separation of these scanning types. Our metrics, e.g., to track randomness of scans, are general and could be leveraged to dissect background radiation in other vantage points as well.

Threat identification: Our finding that most background radiation a typical IPv4 address receives relates to scanning activity, and that there is a relatively steady level of baseline scanning activity has practical value both for researchers as well as network operators. We find that virtually any routed IPv4 address can expect to receive baseline scan radiation day-in-day-out. We believe that current levels of baseline scan radiation can easily be determined, be it from telescope datasets available to the research community [2], or—given its largely even distribution across the address space—even from background radiation gathered from a smaller number of machines. Once a baseline radiation level is established, network operators can readily determine if their individual hosts or infrastructure receive significantly higher levels of radiation and scan activity, indicating that they might indeed be the target of localized scans. An operator could adapt our method and partition addresses into buckets (not necessarily /8s, but could be, e.g., routed prefixes, or addresses of the network's infrastructure versus those of clients, or of peering or customer networks), and then leverage our method to assess randomness, and definitively detect scans that are focused on subsets of their network.

8.2 Future Work

Understanding localized scans: Our finding of widespread localized scanning activity begs the question of what are the root causes of this activity. Our distributed vantage point clearly shows that many of these scans target narrow regions of the address space, when compared to Internet-wide scans, and yet we do not have definite knowledge of the full extent of selected targets of these scans across the entire IPv4 space. While proportionality is a reasonable assumption to assess coverage for partial Internet-wide scans; a broad range of activity is plausible for localized scans, and open for further research. So far, we have classified locality of scan target selection by routed prefixes and ASes. Leveraging other external data, such as IP hitlists, and data gathered from honeypots, could further illuminate localized and stateful target selection strategies. Further, we propose to study long-term characteristics (e.g., months) of individual source IP addresses. We have shown widespread evidence of repeated, and sometimes changing, activity of scan sources. Long-term behavioral analysis could shed further light on root causes for scan activity (e.g., botnet infections vs. repeated targeted scanning campaigns). Unsolicited IPv6 traffic, currently a small fraction of the probing traffic, is more likely due to responses from forward DNS queries (as comprehensive scanning of the IPv6 address space is not practical), and could become significant with the growth of IPv6 connectivity, which we plan to assess.

Correlating scans and cyberattacks: Our vantage point has the attribute that it both elucidates scanning activity, but is also subject to cyberattacks day-in-day-out. Visibility into both activities could provide the rare opportunity to track both the scanning of IPv4 space for newly discovered vulnerabilities, as well as subsequent cyberattacks carried out by infected devices.

Acknowledgments

We thank our shepherd Alex Halderman and the anonymous reviewers for their thoughtful feedback. Our gratitude goes to the Custom Analytics group in Akamai for their continuous support, in particular from Kelli Brown, Richard Weber, and Jon Thompson. We thank Mobin Javed for her help with HiveQL and David Clark and Steve Bauer for fruitful discussions. Our gratitude goes to CAIDA for providing us and the broader research community with access to their network telescope. This work was partially supported by the MIT Internet Policy Research Initiative, William and Flora Hewlett Foundation grant 2014-1601.

REFERENCES

[1] Best Practices and Considerations in Egress Filtering. https://insights.sei.cmu.edu/sei_blog/2018/04/best-practices-and-considerations-in-egress-filtering.html.
[2] CAIDA UCSD Real-time Network Telescope Data. Available via IMPACT, dataset ID DS-0206. http://www.caida.org/data/passive/telescope-near-real-time_dataset.xml.
[3] Netlab 360: Quick summary about the Port 8291 scan. https://blog.netlab.360.com/quick-summary-port-8291-scan-en/.
[4] SANS ISC InfoSec Forums: IPSEC / ISAKMP Vulnerability wrapup. https://isc.sans.edu/forums/diary/IPSEC+ISAKMP+Vulnerability+wrapup/852.
[5] Why would a Windows machine scan for port 137? https://superuser.com/questions/1306406/why-would-a-windows-machine-scan-for-port-137.
[6] D. Adrian, Z. Durumeric, G. Singh, and A. Halderman. Zippier Zmap: Wnternet-wide Scanning at 10 Gbps. In *USENIX WOOT*, 2014.
[7] M. Allman, V. Paxson, and J. Terrell. A Brief History of Scanning. In *ACM IMC*, 2007.
[8] 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.
[9] S. Bano, P. Richter, M. Javed, S. Sundaresan, Z. Durumeric, S. Murdoch, R. Mortier, and V. Paxson. Scanning the Internet for Liveness. *ACM CCR*, 48(2), 2018.
[10] K. Benson, A. Dainotti, kc claffy, A. Snoeren, and M. Kallitsis. Leveraging Internet Background Radiation for Opportunistic Network Analysis. In *ACM IMC*, 2015.
[11] N. Blenn, V. Ghiëtte, and C. Doerr. Quantifying the Spectrum of Denial-of-Service Attacks Through Internet Backscatter. In *Proceedings of the 12th International Conference on Availability, Reliability and Security*, ARES '17, 2017.
[12] J. Czyz, M. Kallitsis, M. Gharaibeh, C. Papadopoulos, M. Bailey, and M. Karir. Taming the 800 Pound Gorilla: The Rise and Decline of NTP DDoS Attacks. In *ACM IMC*, 2014.
[13] J. Czyz, K. Lady, S. Miller, M. Bailey, M. Kallitsis, and M. Karir. Understanding ipv6 internet background radiation. In *ACM IMC*, 2013.
[14] A. Dainotti, K. Benson, A. King, k. claffy, M. Kallitsis, E. Glatz, and X. Dimitropoulos. Estimating Internet address space usage through passive measurements. *ACM CCR*, 44(1):42–49, 2014.
[15] A. Dainotti, K. Benson, A. King, B. Huffaker, E. Glatz, X. Dimitropoulos, P. Richter, A. Finamore, and A. Snoeren. Lost in Space: Improving Inference of IPv4 Address Space Utilization. *IEEE J. on Sel. Areas in Comm.*, 34(6):1862–1876, Jun 2016.
[16] A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescapé. Analysis of a "/0" stealth scan from a botnet. *IEEE/ACM Trans. Netw.*, 23(2):341–354, Apr 2015.
[17] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and A. Halderman. A Search Engine Backed by Internet-Wide Scanning. In *ACM CCS*, 2015.
[18] Z. Durumeric, M. Bailey, and A. Halderman. An Internet-Wide View of Internet-Wide Scanning. In *USENIX Security Symposium*, 2014.
[19] Z. Durumeric, J. Kasten, M. Bailey, and A. Halderman. Analysis of the HTTPS Certificate Ecosystem. In *ACM IMC*, 2013.
[20] Z. Durumeric, F. Li, J. Kasten, J. Amann, J. Beekman, M. Payer, N. Weaver, D. Adrian, V. Paxson, M. Bailey, and A. Halderman. The Matter of Heartbleed. In *ACM IMC*, 2014.
[21] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-Wide Scanning and its Security Applications. In *USENIX Security Symposium*, 2013.
[22] E. Glatz and X. Dimitropoulos. Classifying Internet One-way Traffic. In *ACM IMC*, 2012.
[23] S. Herwig, K. Harvey, G. Hughey, R. Roberts, and D. Levin. Measurement and Analysis of Hajime, a Peer-to-peer IoT Botnet. In *NDSS*, 2019.
[24] M. Kührer, T. Hupperich, C. Rossow, and T. Holz. Exit from Hell? Reducing the Impact of Amplification DDoS Attacks. In *USENIX Security Symposium*, 2014.
[25] M. Lin, H. Lucas, and G. Shmueli. Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24, December 2013.
[26] NIST. CVE-2016-10372 Detail. https://nvd.nist.gov/vuln/detail/CVE-2016-10372.
[27] NIST. CVE-2018-14847 Detail. https://nvd.nist.gov/vuln/detail/CVE-2018-14847.
[28] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson. Characteristics of Internet Background Radiation. In *ACM IMC*, 2004.
[29] E. Pujol, P. Richter, B. Chandrasekaran, G. Smaragdakis, A. Feldmann, B. Maggs, and K. C. Ng. Back-Office Web Traffic on The Internet. In *ACM IMC*, 2014.
[30] P. Richter, G. Smaragdakis, D. Plonka, and A. Berger. Beyond Counting: New Perspectives on the Active IPv4 Address Space. In *ACM IMC*, 2016.
[31] A. Wang, W. Chang, S. Chen, and A. Mohaisen. Delving into internet DDoS attacks by botnets: characterization and analysis. *IEEE/ACM Trans. Networking*, 26(6), 2018.
[32] E. Wustrow, M. Karir, M. Bailey, F. Jahanian, and G. Huston. Internet Background Radiation Revisited. In *ACM IMC*, 2010.
[33] V. Yegneswaran, P. Barford, and D. Plonka. On the Design and Use of Internet Sinks for Network Abuse Monitoring. In *Recent Advances in Intrusion Detection*, 2004.