Who is Knocking on the Telnet Port: A Large-Scale Empirical Study of Network Scanning

Hwanjo Heo
KAIST
ETRI
hwanjo@kaist.ac.kr

Seungwon Shin
KAIST
claude@kaist.ac.kr

ABSTRACT

Network scanning is the primary procedure preceding many network attacks. Until recently, network scanning has been widely studied to report a continued growth in volume and Internet-wide trends including the underpinning of distributed scannings by lingering Internet worms. It is, nevertheless, imperative to keep us informed with the current state of network scanning, for factual and comprehensive understanding of the security threats we are facing, and new trends to serve as the presage of imminent threats.

In this paper, we analyze the up-to-date connection-level log data of a large-scale campus network to study the recent scanning trends in breadth. We find, most importantly, the scanning landscape is greatly shifted, predominantly by an unprecedented rise in Telnet service scannings. Furthermore, not only are the scan sources comprehensively identified in terms of targeted services and geographical/network locations, but also their characteristics, such as being responsible in scanning and their connection-level behavior, are studied.

CCS CONCEPTS

• **Security and privacy → Network security**;

KEYWORDS

Network Scanning; Security; Network Measurement

ACM Reference Format:
Hwanjo Heo and Seungwon Shin. 2018. Who is Knocking on the Telnet Port: A Large-Scale Empirical Study of Network Scanning. In *ASIA CCS '18: 2018 ACM Asia Conference on Computer and Communications Security, June 4–8, 2018, Incheon, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3196494.3196537

1 INTRODUCTION

The first step for a successful cyber attack will be finding valuable victims, and this step will be commonly instantiated by sending reconnaissance packets (known as network scanning) to random or targeted remote hosts. Probing remote hosts lets attackers know who (i.e., an accessible host through networks) serves what (i.e., network services, such as a web service or a DNS service); knowing this information (i.e., who serves what) is absolutely imperative for attackers. Hence, attackers aggressively gather this information by directly searching target hosts or even employing already deployed malware (e.g., botnet) for efficiency.

As such, since this network scanning is an indispensable process for cyber attacks, attention should still be paid to it, even though it has been studied, investigated, and monitored for a long time. Indeed, researchers and practitioners have already deeply surveyed, analyzed, and measured this behavior [3, 8, 26, 34, 35]. However, it should be kept in mind that the characteristics of network scanning (e.g., main target services and scan origins) are quite sensitive to the trends of popular network services and popular malware, and thus they have been actively changed to match these trends. Internet worms, from CodeRed in 2001 to Conficker until recently, have incessantly populated scanning traffic at large [8, 35]. Public disclosure of vulnerabilities, such as OpenSSL Heartbleed [7] and the Linksys router backdoor [10], has also triggered scanning surges looking for susceptible Internet hosts world-wide [8].

Recently, IoT botnets have become a major security concern. Record-breaking DDoS attacks were launched from a large number of Mirai [18] infected IoT devices in the last quarter of 2016 [16, 20]. There are an increasing number of IoT botnets, used for purposes ranging from launching massive DDoS attacks [6] to mining cryptocurrencies [14]. This implies that the current characteristics of network scanning are likely to be different from those of past years. This situation motivates us to investigate the characteristics of past and recent network scanning trials.

In this paper, we study the current network scanning trends from the 31-day-long connection logs acquired from deployed firewalls of a large-scale campus network. We identify scanning attempts from a set of several billion connection metadata, comprising both allowed connections and connections blocked by firewalls. We further compare our results to the previous scanning analysis results in a number of perspectives, to report a remarkably shifted scanning landscape due to the rise of Telnet scannings seemingly from IoT botnets.

Our contributions are summarized as follows. First, we provide the quantitative results with respect to the recent trends of network scanning from a large-scale enterprise production network measurement. With scanning traffic being inflated by at least fourfold in only two years, we observe a pronounced shift of the most frequently targeted service away from MS-DS[1] – often exploited by Internet worms such as Conficker [27] – towards Telnet, which is known to be prominently targeted by IoT botnets such as Mirai; almost two-thirds of scanning probes now target the Telnet service

Permission to make digital or hard copies of all or part 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 components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ASIA CCS '18, June 4–8, 2018, Incheon, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5576-6/18/06…$15.00
https://doi.org/10.1145/3196494.3196537

[1] Microsoft Directory Service. This service binds to TCP port 445. Some studies refer to it as NetBios or SAMBA.

while MS-DS scans account for only 0.57%. We measured a large-scale *production* enterprise network with real-world applications and security instrumentation; to this end, we have established a hybrid scanner detection scheme, not only incorporating firewall-blocked connections, but also reducing false-positives induced by real-world application traffic.

Second, spatial and temporal variabilities of scanning traffic are studied to find that quantitative scanning results of target services that are often exploited by Internet worms or botnets can vary by measured networks, possibly due to their own target IP generation logic or implementations. Furthermore, the temporal variability (in our 31-day measurement duration) of the scanning traffic is mostly imposed by the variability of the Telnet scanning.

Finally, we have analyzed the scan sources and their traffic composition in various perspectives such as identification of responsible scanners who admittedly perform scans for research or information gathering purposes, being listed in IP blocklists or being located in dynamically allocated IP address spaces, and their connection-level scanning behavior. To this end, we attribute the pervasive Telnet scanners to IoT botnets with compelling evidences.

2 DATA COLLECTION AND METHODOLOGY

2.1 Data Set

To analyze the characteristics of network scan trials, we have used connection log information over 31-day (from June 2016 to July 2016) from two firewalls deployed on our campus network. This network consists of one /16 and two /20 networks and it operates two firewalls in an active-active HA clustering mode; they are directly connected to the enterprise access router (outward) and additional middleboxes (inward). A dedicated log server records all traces[2] of both allowed and denied connections – at the time of arrival – including timestamp, five-tuple information, and TCP flags from both firewalls; it also records the number of total transferred pkts/bytes, and observed TCP flags if explicitly terminated (by TCP FIN) or timeout occurs.

Table 1: Connection population of our data set

Protocol	Inbound		Outbound	
	Allowed	Denied	Allowed	Denied
TCP	2.09B	4.18B	4.68B	1.16B
UDP	4.38B	1.07B	2.84B	0.87B
Other	0.01B	0.09B	0.00B	0.39B

Table 1 shows the number of inbound and outbound connections (over 21 billion in total) that are allowed and denied by the firewalls during the entire measurement period (i.e., 31 days). As shown here, there are a large number of TCP and UDP connections, and it implies that this data will provide a clear view of diverse scan trials (both TCP based and UDP based). In addition, firewalls deny many connections, and it is a difficult process to identify scanners without knowing connection success probability. To handle this case, information on denied connections are carefully investigated and a way of sifting out non-scanning connections is suggested (this is detailed in the following section).

Our scanning measurement has a number of advantages over the previous ones with respect to the data set. First, many previous measurement studies [8, 26, 34] have collected packets destined to *unused* IP address spaces. It is known that the target selection of distributed scanning by Internet worms is often *not* random intentionally [30] (for efficient infection) or unexpectedly [21] (for implementation issues). Targeted network scannings are also conceivably expected by purposeful attackers. We have measured a large-scale *production* enterprise network accommodating several tens of thousands of regular users.

Second, scanning measurement studies based on *used* IP spaces are *incomplete* in that they are missing UDP connections [3] or in that they are comprising either rejected connections by firewalls or IDS detected portscans of allowed connections [35]. Our measurement provides a *complete* view of TCP and UDP network port scannings identified from connections that were both allowed and rejected by firewalls.

One limitation of our data set is the exclusion of protocols other than TCP and UDP. Even though they are small in volume (see Table 1), a large number of ICMP packets are observed in some network in 2004 [26]. However, it is reported that the volume has become very small in 2010 [34] and 2014 [8]. Also, it is often recommended to block or rate-limit ICMP messages [12].

2.2 Scanning Classification Methodology

Previously, researchers have tried to identify network scan attempts from network traces or connection log information. They consider packets heading to unused IP address spaces (i.e., darknet) as network scan trials [8, 26, 34] or employ classification algorithms (e.g., measuring a connection success probability) to statistically figure out network scan (or suspicious) attempts [3, 8, 19, 35]. Unfortunately, those methods cannot be directly applied to our collected data to reveal scan trails. Thus, we devise a hybrid approach with a few refinements to the existing solutions.

Connection requests allowed by the firewall are responded to if the requested service port is open at the target host. On the other hand, denied connection requests never go through the firewall and those services or hosts blocked by the policy should be considered unavailable to the remote hosts. However, misconfiguration or ignorance of such firewall policy in the applications could proliferate such blocked connection requests not made for scanning purposes. To this end, connection success probability based classification of scanning traffic was applied for allowed connections and a filtering scheme of non-scanning connections was applied for denied connections. More information of this classification will be explained as follows.

2.2.1 Scanning classification for allowed connections. We consider the case for determining if a *remote* (from our measured network) host is a *scanner* by observing the outcomes of its connection attempts to *local* (to our network) hosts. For a given remote host r, let $T(r)$ be the number of distinct (local IP, port) pairs to which r makes connection attempts; likewise, let $H(r)$ be the number of distinct local IP addresses and $V(r)$ be the number of distinct port

[2]The firewall systems denote that no events (connection trials) are missed, and thus we assume that they have captured all connection trials without loss.

numbers. We define Y_i as a random variable that indicates the outcome of the connection attempt by r to the i_{th} distinct (local IP, port) pair where $0 \leq i < T(r)$. $Y_i = 1$ if the connection request is responded to, $Y_i = 0$ otherwise. In the case of UDP which is connectionless, we utilize the firewall's UDP session state machine which groups bidirectional unique five-tuple UDP packets into a pseudo-connection with a 10-second timeout. We treat the first UDP packet of each pseudo-connection as a connection request and determine that it is *responded to* if any packet, with a reverse five-tuple, is replied to before the timeout occurs.

We determine that the remote host r is a scanner if the following conditions are satisfied:

(1) $Pr[Y{=}1] < \theta_0$ for the connection success probability threshold θ_0.
(2) $H(r)/T(r) < \theta_1$ or $V(r)/T(r) < \theta_1$ for the spread threshold θ_1.
(3) $T(r) \geq \theta_2$ for the fanout threshold θ_2.

As to the first condition, θ_0 depends on the monitored network's available host and open port density; θ_0 limits the probability of successful connection attempts by a scanner who does not have *a priori* knowledge regarding the distribution of available hosts or open ports. This threshold is set to 0.2 in [19] and 1/3 in [3]. We set $\theta_0 = 0.3$ as our monitored network has a /24 subnet in which 27% of the hosts responded on TCP/80, which is the most popular service port at the time we scanned.

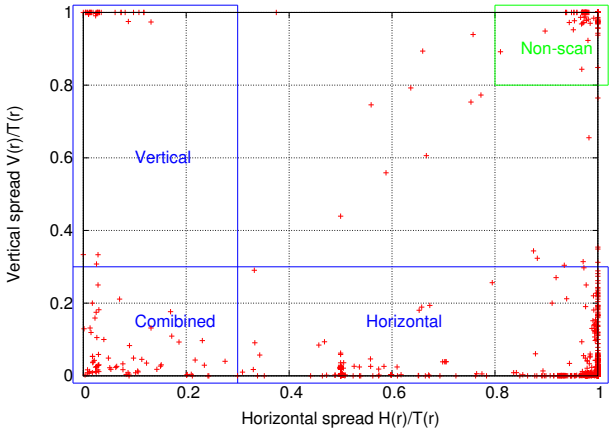


Figure 1: Horizontal and vertical spread of remote hosts with $Pr[Y{=}1] < 0.3$

The second condition is devised to filter out the recent P2P client implementations that bind to randomly chosen ports from our scanner classification. The horizontal spread $H(r)/T(r)$ captures the host fanout normalized to the number of connection attempts to distinct (local IP, port) pairs. For example, $H(r)/T(r) = 1$ means every connection attempt of r is targeting different hosts. The vertical spread $V(r)/T(r)$ captures the normalized service (port) fanout.

Figure 1 shows the remote hosts with $Pr[Y{=}1] < 0.3$ and $T(r) \geq 32$ in terms of horizontal and vertical spread from a single-day-long connection log. We further annotate the (blue) regions that our second condition captures into horizontal, vertical, and combined scanners as a coarse classification of scanning behavior with $\theta_1 = 0.3$.[3] The (green) area in which both spread variables exceed 0.8 is labeled as non-scan. We manually inspected the connection attempts made by the remote hosts who are labeled as non-scan to find that they indeed appear to be P2P applications who connect to different hosts with different ephemeral destination ports.

Finally, *the third condition* filters out the remote hosts with small fanout. If a remote host makes only a small number of connection attempts, we cannot reasonably conclude that it is a scanner from its connection success probability. Figure 2 shows the population of each annotated group as a function of the fanout threshold θ_2. It is noticeable that a majority of the population is either horizontal scanning or non-scan; remote hosts with large fanouts are mostly horizontal scanners while those with small fanout exhibit the non-scan (i.e., P2P-like) behavior. The unclassified population hampers our classification scheme with a small fanout threshold $\theta_2 < 8$ due to having small denominators in calculating $Pr[Y{=}1]$. We have chosen $\theta_2 = 8$ to avoid this normalization effect and thus ignored very small possible scanners who only probe less than eight targets. Lastly, our scanner classification is performed separately for each single-day-long connection trace – meaning an IP address determined to be a scanner in one day (by utilizing the connection success probability of that day) is not necessarily a scanner on the other days, for the reason discussed in [3].

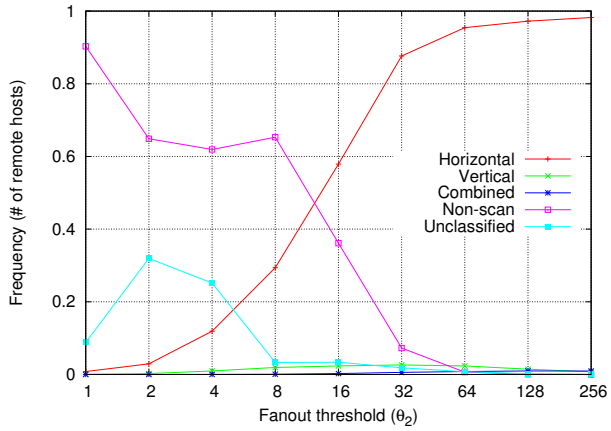


Figure 2: Remote host classification as a function of fanout threshold θ_2

2.2.2 Filtering of non-scanning connections for denied connection attempts. It is difficult to identify scanning attempts from rejected connections (by a firewall). First, we cannot perform the connection success probability based classification since dropped connection requests always result in connection failures; $Pr[Y{=}1] = 0$ if all connections are denied. One may claim that every denied connection is an artifact of scanning; however, it is not always the case since applications, who are unaware of firewall rules, may attempt to establish connections to blocked destinations. To see if we can reasonably filter out such cases, we have analyzed the denied connections to find that the majority of them can be grouped by their

[3]We have empirically chosen this threshold value as the proportion of unclassified remote hosts that have only a small difference, as long as $0.2 < \theta_1 < 0.8$.

destination ports to reveal their application; more than 85% of denied connections of TCP are either port 6881 (BitTorrent default port) or port 23 (Telnet) connection and more than 80% of denied connections of UDP are destined for port 6881.

Furthermore, the number of IP addresses to which connection attempts are made daily is distinctly distributed according to their target ports. For example, TCP/6881 connection attempts (1.9B in a month) are targeted to only 2014 IP addresses while TCP/23 connection attempts (1.7B) are targeted to 73,628 IP addresses [4] out of the total of 73,728 IP addresses to which the network is assigned. The horizontal scanners of TCP/23 sweep through the entire IP address space while some BitTorrent clients which bind to the default port are connection-attempted but blocked by the firewall policy.

We have expanded our observation to the other service ports, i.e., the top 400 TCP and the top 200 UDP ports that account for more than 98% of all denied connections, to find that the number of targeted IP addresses to which connections are attempted is indeed bimodally distributed. Figure 3 shows the distribution of the number of target IP addresses to which blocked connections are attempted by the top 400 TCP and top 200 UDP service ports. We choose the threshold (lower bound) of 40,000 to filter out non-scanning blocked connection attempts that target only a small number of local IP addresses.

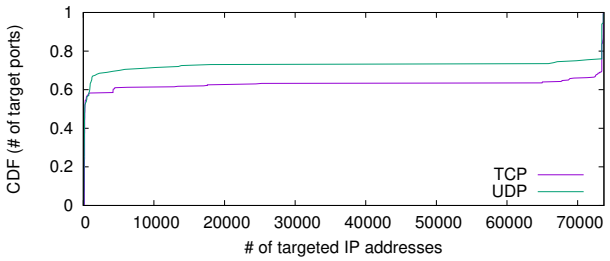


Figure 3: Cumulative distribution of the number of target ports over the number of targeted IP addresses to which connection attempts are made.

Overall, we have identified 2.65 billion scan probes from 3.78 million scanners[5] throughout our measurement period. In this paper, we mainly analyzed *inbound* connection logs to identify and study *inbound* network scanning trials; however, *outbound* connection logs were also utilized for manual investigations such as tracing communications between local hosts and informed Mirai C&C server IP addresses in Section 4. Although we reported an unprecedented amount of inbound network scanning attempts, we were able to identify only a handful of outbound scanning episodes at most in a single day; we put the outbound network scanning out of scope not only because they are not our main interest in studying Internet scanners but also they are not abundant in our data set.

[4]Technically, there are 100 local IP addresses, comprising four small contiguous subnets, that are never recorded as destination IPs at the firewall. We believe that those subnets are routed to a different physical location or blocked by the access router's ACL for some reason.

[5]Each scanner is uniquely identifiable by its IP address. There should be source IP address spoofings which we were not able to discern.

3 OVERALL RESULTS

3.1 Targeted Services

Table 2 shows the 20 most frequently targeted services of inbound scanning from our measurement. Each entry shows the name of service (protocol/port number), the average number of scan probes per day (in million), and its relative proportion. Multiple retries of SYN packets in 10s timeouts are counted as a single scan probe.[6]

The most interesting result is that the Telnet (TCP/23) scanning, which recorded over 53 million scan probes a day, accounts for 62.98% of the entire scanning probes. It indicates a great shift in the scanning trends from recent results which unarguably report that MS-DS (TCP/445) accounts for a significant portion of traffic due to wide-spread worms such as Conficker; close to half of all scan traffic targeted the MS-DS service in 2014 [8] and 83.1% of TCP packets targeted TCP/445 in 2010 [34]. A long term study of scanning trends also indicates that Telnet has never been a major contributor of scan traffic [3]. We suspect that the rise of Telnet scanning is attributed to the recent surge of IoT botnets such as Mirai [18]. The botnet is designed to make infected bots scan TCP/23 of randomly generated IP addresses to spread themselves. We further analyze the TCP/23 scanning in Section 4.

While other services account for no more than 5% each in contrast to the dominance of Telnet scan traffic, UDP/53413 is unusually notable; it is reported that Netis routers have a backdoor which is exploitable at UDP port 53413 [36]. Responsible scanners such as Shadowserver [11] indeed regularly scan this service. Finally, note that only 0.57% of scan traffic is MS-DS, which is regarded as the most frequently exploited service by Internet worms.

3.2 Scan Dynamics

Overall, we observed 2.65 billion scan probes targeting 73,628 IP addresses of one /16 network and two /20 networks during a 31-day measurement period in June and July of 2016; the average number of scan probes received by a single IP per day is 1159.

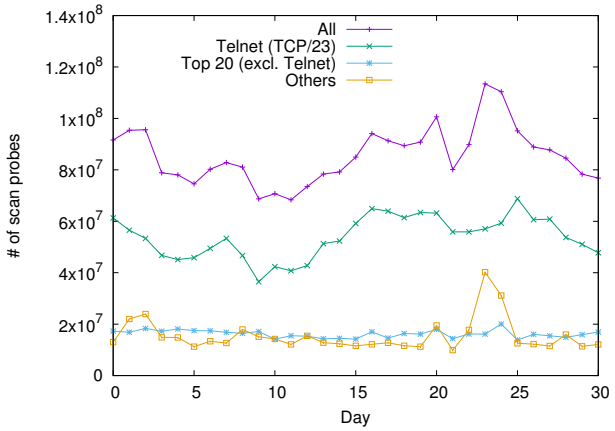


Figure 4: Daily scan traffic with three targeted service groups

[6]We find a number of HTTP/HTTP-alternative scanners who send burst SYN packets (up to 120) to every scanning target. We do not want an amplification of HTTP/HTTP-alternative scan traffic by those eccentric scanners.

Table 2: Frequently targeted services of inbound scanning

Service	Scan probes	Service	Scan probes	Service	Scan probes
Telnet (TCP/23)	53.91M(62.98%)	HTTPS (TCP/443)	0.67M (0.78%)	HTTP-alt (TCP/8888)	0.31M (0.36%)
MS-SQL (TCP/1433)	3.60M (4.21%)	HTTP-alt (TCP/8080)	0.56M (0.66%)	HTTP-alt (TCP/81)	0.30M (0.35%)
Netis-vuln (UDP/53413)	2.89M (3.38%)	MS-DS (TCP/445)	0.49M (0.57%)	MS-RPC (TCP/135)	0.30M (0.35%)
SSH (TCP/22)	1.57M (1.83%)	SIP (UDP/5060)	0.43M (0.51%)	SUN-RPC (UDP/111)	0.27M (0.32%)
HTTP (TCP/80)	1.29M (1.51%)	Web-proxy (TCP/3128)	0.37M (0.44%)	NetFone (TCP/10200)	0.26M (0.30%)
MySQL (TCP/3306)	1.01M (1.18%)	DNS (UDP/53)	0.35M (0.41%)	HTTP-alt (TCP/8081)	0.26M (0.30%)
MS-RDP (TCP/3389)	0.93M (1.08%)	SMTP (TCP/25)	0.35M (0.40%)	Other	15.46M(18.07%)

Figure 4 shows the daily scan traffic in terms of the number of scan probes. We further categorize scan traffic to the three exclusive groups: (i) Telnet scan traffic, (ii) scan traffic belonging to the top 20 frequently targeted services other than Telnet, (see Table 2) and (iii) all remaining scan traffic. Note that the temporal variability of scan traffic is mostly dictated by Telnet scanning. The scan traffic belonging to the other 19 frequently targeted services show little variability in the meantime. The remaining scan traffic exhibits two short periods during days 1-2 and days 23-24 with a burst of scan traffic; it turns out that each of these episodes is initiated by an onset of vertical scanning to selected local targets by only a few remote hosts.

The overall scanning volume and scanning frequency with respect to target services are far different from what is reported previously. Table 3 summarizes the previous reports in comparison to ours. The normalized scan traffic is defined as the number of daily average scan probes per IP address; in other words, the total number of scan probes is divided by the product of measurement network size and measurement duration (in days). The two target services, i.e., MS-DS and Telnet, are compared in terms of the proportion of scan probes. [7]

Although each study has its own unique settings and some values of normalized scan traffic are calculated from estimated quantities, we clearly observe an increasing trend of scan traffic. It is especially notable that a single IP received 255 scan probes per day in 2014 – a daily average of 1.4 billion packets were received at a darknet comprising 5.5 million addresses [8] – and 1159 scan probes a day in 2016. Note that the number of scan probes have increased by at least four times in the past two years; this comparison is conservative since each packet targeting darknet IP addresses is counted as a single scan probe[8] by [8]. Furthermore, MS-DS scanning traffic, showing a significant proportion of traffic from mid 2000 to 2014, accounts for only 0.57% now. On the other hand, Telnet scanning accounts for roughly two-thirds of the entire scan traffic.

One interesting observation we have made from our temporal analysis is that scanners are not frequently recurring, possibly due to the shifted scanning landscape. Figure 5 shows the average number of unique scanner IP addresses observed by a single local IP address during the measurement period. The number of scanning IPs are further grouped by the number of recurring days. For example, '0 day' means the scanner is only observed on that particular day (not recurring). On average, there are 142 remote IPs that scan a single local IP for a day. 41.94% of them are not recurring; slightly more than half of them scan the same IP at least once more during the same month. One implication of this lack of recurrence is the ineffectiveness of using IP blacklisting in blocking scanners.

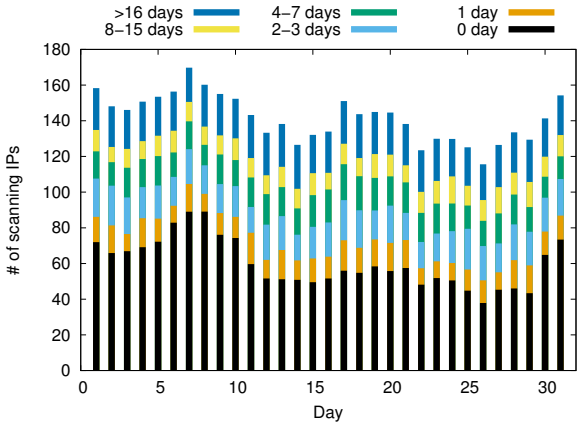


Figure 5: The average number of scan sources, grouped by the number of recurring days, for each local host

3.3 Scan Source Distribution

We have identified approximately 3.78 million unique scanner IP addresses. We have used MaxMind GeoLite2 [24] to retrieve the geographical location of scanner IP addresses and Routeviews [2] for IP-to-AS mapping. Table 4 and Table 5 show the top 15 countries and the top 20 ASes contributing the most scan probes, respectively. While China and the United States contribute the most as previously reported [8], this dominance of the top originating countries is now lessening; the top five countries accounted for 76% of scan traffic in 2014 [8], while they contribute slightly more than half in our result.

A recent scanning study claims that scans now mostly come from bullet-proof hosting providers based on a result showing that 49 out of the top 100 contributing ASes are dedicated hosting services [8]. However, we similarly categorize the top 100 contributing ASes to find that only 16 are hosting providers, while 75 are ISPs. This discrepancy can be attributed to the fact that the previous result is drawn for large scanners – targeting >10% of IPv4 address spaces, thus excluding less aggressive scans from worms and botnets – or

[7] Some values are not quantitatively precise since only visual comparison of scanning traffic is possible [3] or only a small number of frequently targeted services are reported [26, 35].
[8] Darknet packet traces have non-scan traffic such as backscatter traffic [34].

Table 3: Summary of previous scanning measurement results and ours

Year	Authors [Ref]	Normalized scan traffic	MS-DS (TCP/445) scanning	Telnet (TCP/23) scanning
1994 ~ 2006	Allman et al. [3]	0.08 ~ 76	one of the most frequently targeted services	nominal
2001 ~ 2002	Yegneswaran et al. [35]	3.84 ~ 5.64	not listed in top 10	not listed in top 10
2004	Pang et al. [26]	61	19.7%	not listed in top 8
2006 ~ 2010	Wustrow et al. [34]	14 ~ 132	23.1% ~ 83.1%	1.7% (at a few blocks)
2014	Durumeric et al. [8]	255	41.7%	2.8% (<1% IPv4 scans)
2016	our work	**1159**	**0.57%**	**62.98%**

the scan sources for the Telnet service, which account for more than half of all scans in our results, are mostly located in ISP ASes; we find that 98 out of the top 100 contributing ASes of Telnet scan probes are ISP ASes. We further study the scan sources we have identified in Section 4.

Table 4: Scans by originating country

Country	Scan probes	Country	Scan probes
China	19.18%	Columbia	2.99%
United States	11.91%	India	2.97%
Brazil	7.65%	Philippines	2.58%
Taiwan	6.54%	Germany	2.41%
Vietnam	6.19%	Seychelles	1.87%
Korea, Republic	6.09%	Mexico	1.35%
Turkey	3.58%	Argentina	1.21%
Russia	3.39%	Other	20.09%

Table 5: Scans by originating AS

ASN	AS name	Country	Scan probes
4134	CHINANET-BACKBONE	China	7.74%
3462	HINET	Taiwan	5.52%
4837	CHINA169-BACKBONE	China	4.67%
4766	KIXS-AS-KR	Korea, Republic	3.04%
9121	TTNET	Turkey	2.68%
45899	VNPT-AS-VN	Vietnam	2.61%
28573	CLARO	Brazil	2.45%
9299	IPG-AS-AP	Phillipines	2.38%
29073	QUASINETWORKS	Netherlands	1.80%
6939	HURRICANE	United States	1.64%
18881	TELEFÔNICA	Brazil	1.62%
3816	COLOMBIA	Colombia	1.62%
7552	VIETEL-AS-AP	Vietnam	1.46%
10439	CARINET	United States	1.33%
23650	CHINANET-JS-AS-AP	China	1.30%
18403	FPT-AS-AP	Vietnam	1.17%
30633	LEASEWEB-USA-WDC-01	United States	1.14%
27699	TELEFÔNICA	Brazil	1.02%
9829	BSNL-NIB	India	0.91%
51167	CONTABO	Germany	0.91%
	Other		52.18%

3.4 Scanning Fanout

We sift the remote hosts with the scanning fanout condition $T(r) \geq 8$, due to the limitation of reliably determining connection success probability with small fanout in Section 2.2. Although it is inherently difficult to identify these very small scanners – thus, they are often excluded [3, 8, 35] – it is important to know how large and small scanners, in terms of scanning fanout, differ by observing their composition as a function of scanning fanout.

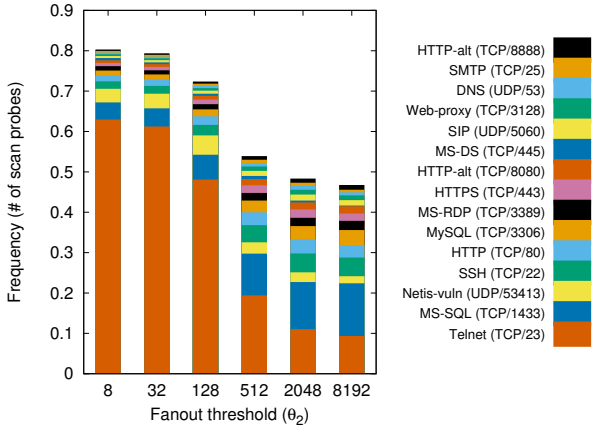


Figure 6: Top 15 frequently targeted service composition as a function of fanout threshold

Figure 6 shows the relative (scan probe) frequency of the top 15 frequently targeted services as a function of the fanout threshold θ_2. While there is little difference in both small threshold range of $\theta_2 < 32$ and large threshold range of $\theta_2 > 2048$, the relative proportion of Telnet scanning greatly fades out in the range $128 < \theta_2 < 512$; the other 14 services and unlabeled services have inflated their relative proportions because of the reduced proportion of Telnet scanning.

To understand this discrepancy and to find plausible reasons, we have investigated the IoT botnets. The publicly released source code of the recent IoT botnet Mirai [18] has a scanner component which generates random target IP addresses with a hardcoded scanning rate of 160 pps (packets per second). Assuming uniform (and random) IP address generation, Mirai scans 13.824 million target IP addresses in a single day. Considering the number of our measurement IP addresses, which is $2^{16} + 2^{12} + 2^{12}$, we expected to observe scanning probes to approximately 237 IP addresses by a single Mirai infected remote device in a day. As the fanout $T(r)$ of a remote host r is computed by single-day connection traces (see Section 2.2), the fact that the proportion of Telnet scanning is greatly reduced in the scanning fanout range $128 < \theta_2 < 512$ is

consistent with the expected number of scanning targets by Mirai, i.e., 237, with its hardcoded scanning rate of 160 pps.

4 CASE STUDIES

4.1 Responsible Scanners

It is known that there are a number of (groups of) scanners which operate on a regular basis for research, security, and information gathering purposes. While their contribution to scanning traffic has yet to be quantitatively evaluated, several of them have surfaced previously [8]. We define a scan server with an IP address *sIP* to be a responsible scanner [9] if at least one of the following conditions is satisfied:

(1) An HTTP GET request to *sIP* at TCP/80 yields a response with information which explicitly informs the intention of scanning. All such scanners we queried also provide the purpose of scanning and information regarding how to request an exclusion.
(2) The subdomain label of the domain name associated to *sIP* contains 'scan' as a substring. (e.g., research-scan.cis.upenn.edu, proxyscanner.underworld.no)
(3) The second-level label of the domain name indicates scanning related research or search engines. (e.g., census*.shodan.io, icsresearch*.plcscan.org)
(4) The owner information of the associated AS indicates intentional scanning. The only AS which we find with this condition is AS42570 KS-ASN1.

It is arguable whether one can qualify some scanners satisfying a subset of the above conditions *responsible*; for example, Shodan [31] scanners can be identified by domain names while no information regarding the intention of scanning or exclusion request is provided by serving web pages on the scanner IPs or putting such information on the AS owner description. We use the term *responsible scanner* in a broad sense.

We have collected approximately 3.78 million scanner IP addresses which sent at least one scan probe during our measurement period. For each IP address, we have retrieved the domain name via reverse DNS lookup, associated AS information [2], and the HTTP GET response page, if available.

Table 6 shows the list of identified responsible scanners who contribute at least 0.0001% of the scanning traffic. Scan probes are shown as the average of daily scan probes.

Overall, the scan probes originating from responsible scanners account for approximately 5% of scanning traffic. Roughly 700 out of 3.78 million scanner IP addresses are identified as responsible scanners. Shodan and Shadowserver foundation [11] are contributing two-thirds of all scan probes originating from responsible scanners. Shodan scans a wide range of UDP ports for vulnerability scan. Shadowserver scans a comprehensive set of well-known services and vulnerable ports. [10] Several universities conduct regularly scheduled scannings for security research. While scanning servers hosted by the majority of research projects and security companies look legitimate, we find some scanners, such as Internet Research Project [11], are suspicious in the sense that they seem to be disguising the scanning purpose as research.

Table 6: Responsible scanners

Responsible scanner	# IPs	# Ports	Scan probes
Shodan	15	28348	1,444,017(1.6870%)
Shadowserver	209	279	1,398,331(1.6336%)
U of Michigan	176	23	450,308(0.5261%)
Internet Research Project	1	73	254,121(0.2969%)
UC Berkeley	2	2	152,308(0.1779%)
UC San Diego	1	1	130,982(0.1530%)
ipip.net	3	7	119,028(0.1391%)
NetworkSecurityAlliance Proxy Scan	2	27	109,189(0.1276%)
Nagra Port Scan	242	12	86,012(0.1005%)
Project Sonar by Rapid7	28	22	73,444(0.0858%)
PLC Global Census Scan	1	20	63,332(0.0740%)
Ruhr-Universität	1	4	31,013(0.0362%)
U of Pennsylvania	1	1	7,361(0.0086%)
SBA Research	1	1	4,075(0.0048%)
Project 25499	2	2	2,814(0.0033%)
ProbeTheNet	1	1	2,477(0.0029%)
RWTH Aachen University	1	1	2,365(0.0028%)
Underworld	1	41	382(0.0004%)
P2P-NET	1	137	216(0.0003%)
EsperNet	1	53	178(0.0002%)
Sum	690		4,331,936(5.0610%)

Figure 7 shows major responsible scanner's contribution to scanning traffic in the top 50 frequently targeted services. Shodan is the most comprehensive – it scans 34 out of the top 50 services – although the proportion of scan traffic is small. Shadowserver's contribution is noticeably large for the selected target services it purposefully scans. Lastly, the target services that are frequently exploited by botnets and worms, e.g., Telnet (TCP/23) and MS-DS (TCP/445), have only a small proportion of scan traffic originating from responsible scanners.

4.2 IP Blocklists and Dynamic IP Addresses

Spamhaus XEN blocklist [22] is a composite database of three different IP address lists: the Spamhaus Block List (SBL) is a list of IP addresses which are involved in sending unsolicited bulk e-mails. The eXploits Block List (XBL) consists of IP addresses of hijacked PCs infected by various exploits. The Policy Block List (PBL) is a database of dynamic IP address prefixes; while acceptance of unauthenticated SMTP messages from hosts in PBL is not advised, PBL does not imply any malicious activity from the IP addresses that belong to the list. The *abuseat.org* CBL (Composite Blocking List) [1] lists IP addresses exhibiting open-proxy or dedicated spam bot behaviors which have been abused in sending spam e-mails. it is also claimed that the list includes some portion of botnet infrastructures not directly involved in spam activities.

Table 7 summarizes the proportion of scanner IP addresses and originating scan probes belonging to the Spamhaus XEN and *abuseat.org* blocklists. As for the Spamhaus XEN, note that the blocklists – SBL, XBL, and PBL – are not exclusive. It is not very

[9] We use the term in a broad sense. Some scanners in our classification do not explicitly inform their purpose or accept exclusion requests.
[10] We omit detailed information regarding the target ports due to limited space. The list of Shadowserver's target ports are available online [11].
[11] The scanner IP address 208.100.26.228 serves a web page describing the purpose of scanning and how to opt out.

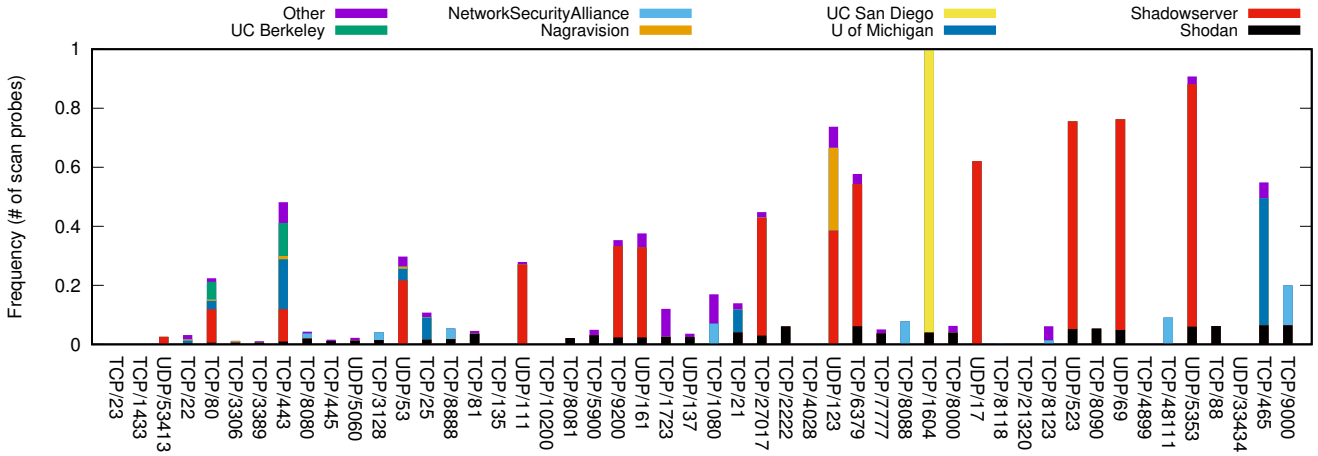


Figure 7: Scan probes contributed by major responsible scanners in the top 50 frequently targeted services

surprising that IP blocklists are not very effective in blocking scanners, as approximately 5% of scan probes originated from IPs listed in either SBL or XBL; of the remainder of the scan probes (95%), roughly half originated from IP addresses in PBL and the other half was not listed. Interestingly, PBL listed IPs account for 85.46% of the entire scanner IP address set while unlisted IPs account for only 14.11% of the entire set; however, they both contribute a similar number of scan probes. In other words, static scanners generate roughly six times more scan probes than dynamic scanners whose IP addresses are dynamically assigned. Considering TCP/23 alone, approximately 60% of scan traffic originated from the PBL listed (i.e., dynamically allocated) IP addresses while less than 20% of scan traffic originated from the PBL for the five well-known services: SSH, FTP, HTTP(s), and DNS.

Table 7: Scanner IPs and scan probes by Spamhaus XEN and *abuseat.org* blocklists

Provider	Blocklist	Scanner IPs	Scan probes
Spamhaus XEN	SBL	0.34%	1.32%
	XBL	3.53%	4.14%
	PBL (Dynamic IP address)	85.46%	49.57%
	Not listed	14.11%	47.00%
abuseat.org	CBL	4.37%	4.44%
	Not listed	95.63%	95.56%

On the other hand, the *abuseat.org* CBL (Composite Blocking List) also provides a number of C&C server IP addresses of Mirai botnet. We have searched our entire firewall logs to find connections incoming from or outgoing to those – 72 as of today – IP addresses. We identify only a single matching IP address which have sent TCP SYN probes to 78 different enterprise IP addresses at port 23; however, all of them are blocked by our firewall as the firewall policy is set to block inbound TCP/23 connections targeted to most of the enterprise IP address ranges.

4.3 IP Address Space Variability

Many scanning measurement studies follow the network telescope approach [25] to extrapolate local measurement of unused IP address ranges, i.e., darknet, to the Internet scale; it is based on an assumption of uniformly distributed target selection in which the probability of a single probe packet being detected is modeled by a geometric distribution [8]. On the other hand, information gain by aggregating measurement from different subnets is studied in an information theoretic approach [35]. As our measurement is performed at three different IP address ranges – one /16 network and two /20 networks – that are spatially apart, we compare our scanning results of the three different IP address ranges we have measured.

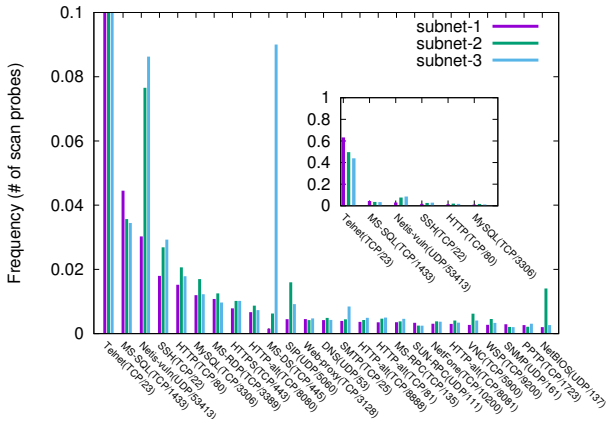


Figure 8: Scan volume per targeted services across subnets

Figure 8 depicts the relative proportion (by measurement subnets) of scan volumes of the top 25 most frequently targeted services. *subnet-1* is the /16 network. *subnet-2* and *subnet-3* are the two /20 networks. While the overall trends of Telnet scanning being the majority is consistent over the subnets, relative scan volumes differ

for a few relevant services. Telnet scanning accounts for more than 60% in *subnet-1* while it does around 40% in *subnet-3*. Netis-vuln (UDP/53413) scan volume of *subnet-1* contribute less than half of the other IP address ranges. Also, MS-DS (TCP/445) scanning is concentrated in *subnet-3*, contributing more than 60 times more than other subnets.

To address this discrepancy, we further divide the /16 network, i.e., *subnet-1*, to 16 equally sized /20 subnets and evaluated if this variability can be similarly observed in contiguous smaller (and same sized) /20 subdivisions of *subnet-1*. Figure 9 shows scan volumes of the three services – Telnet, Netis-vuln (UDP/53413), and MS-DS – for the 16 /20 subdivisions of *subnet-1* together with *subnet-2* and *subnet-3*.

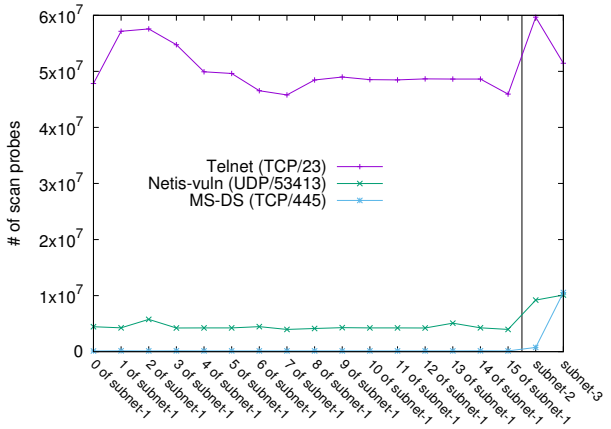
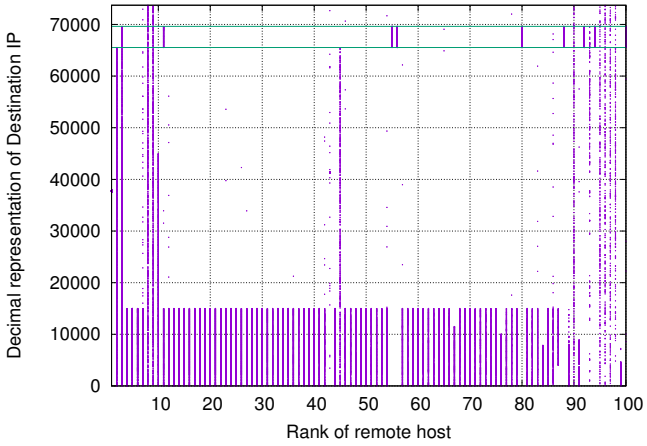


Figure 9: Scan traffic across subnets and subdivisions

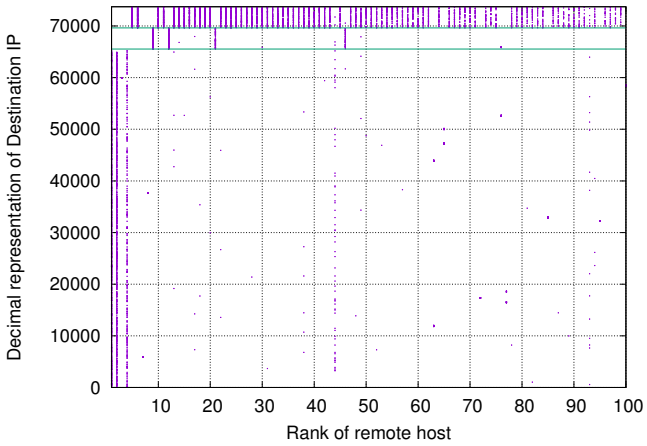
A mixed observation by the services is seen: for Netis-vuln (UDP/53413) and MS-DS, the scan volumes of *subnet-2* and *subnet-3* are noticeably higher than those of the subdivisions of *subnet-1*. In the meantime, the initial subdivisions of *subnet-1* and *subnet-2* have higher Telnet scan traffic.To address this discrepancy, we have observed the scanning behavior of heavy scanners to find that many of them are *not uniform* in their target IP selection.

Figure 10 shows the destination IP address trajectory of the top 100 scanners of the service Telnet and MS-DS from a single day measurement. The destination IP addresses are converted to a decimal representation which has 0 ~ 65535 for *subnet-1*, 65536 ~ 65536+4096 for *subnet-2*, and 65536+4096 ~ 65536+8192 for *subnet-3*. As for Telnet, the initial 15,000 IP addresses of *subnet-1* and the entire set of *subnet-2* addresses are far more frequently scanned than the other IP address ranges. [12] On the other hand, *subnet-3* is far more frequently targeted than the other subnets by MS-DS scanners. We speculate that IP generation logics of some botnets and worms are *not uniformly random*. Telnet service scanners target the low address space of our /16 IP address range more frequently. Also, *subnet-2* is a /20 network with the third digit of IP address *16* (thus, it belongs to the low address space inside the 15,000 IP addresses of the projected /16 network) while *subnet-3* has the third digit

[12] We manually picked up the TCP/23 scanners with this eccentric behavior – scanning only the first 15,000 IP addresses – to find that 40% of them belong to AS 3816 Columbia Telecommunicaciones, CO while the other 60% are spread worldwide.



(a) Telnet (TCP/23)



(b) MS-DS (TCP/445)

Figure 10: Scan destination IP trajectory by the top 100 remote hosts

number *128* (thus, it is above the beginning 15,000 IP addresses of the /16 network).

Although the implementation of IP generation logics of several worms and botnets are publicly available in source codes [18, 33] and analysis reports [9, 27], none of them appears to have a matching logic to our observation. For example, Mirai [18] and Conficker [27] generate random IP addresses with exclusion of some address ranges; they do not have any biased IP generation logic by design, [13] except that some versions of Conficker have a scanning technique to infect nearby subnets [30]. However, we have not found any statistically significant bias of scan sources' IP address distribution being nearby IP addresses of *subnet-3*. There are also possibilities of system-dependent factors; e.g., low entropy in Linux random

[13] Although the released Mirai source code does not implement the observed bias in its scanning target IP generation, a preliminary version of it may have such bug, e.g., a bug in writing logical expressions to implement the exclusion list [18], to show the observed behavior.

number generation [13] and *Witty* worm's skewed PRNG [21], have incurred such biased target IP address generation.

4.4 Connection-level Behavior of Scanning

The connection-level behavior of non-worm scanners exhibits Zipf's law, like many other phenomena on the Internet [35]. We have examined the distribution of scanner's scan volumes for each service, not only to validate if the claim holds for scanners targeting non-worm service ports, but also to understand the connection-level behavior of frequently targeted – yet uncharted by previous measurement – services.

Figure 11 shows the complementary cumulative distribution of scan volumes of scanners for seven representative services: Telnet, DNS, Netis-vuln, MS-DS, HTTPS, HTTP and SSH. Note that the well-known services with a large proportion of scan traffic that originated from responsible scanners (i.e., HTTP, DNS, and HTTPS) or that are known to be frequently scanned by large scanners from dedicated hosting services [8] (i.e., SSH) show a similar degree of tail weight, exhibiting a heavier tail than the other services (i.e., Telnet, Netis-vuln, and MS-DS) that are depicted. We also calculate the tail exponent α, where $1 < \alpha < 2$ for heavy-tailed distributions, using Hill's estimator [15] to find that $\alpha = 1.70$ for HTTP (TCP/80), 1.57 for DNS (UDP/53), and 2.00 for HTTPS (TCP/443). Netis-vuln (UDP/53413) scanners, whose $\alpha = 3.39$, show a discontinuous pattern due to a mixed composition of a short-tailed majority and only a small number of very large scanners (they are indeed mostly Shadowserver scanners as depicted in Figure 7.) [14] MS-DS (TCP/445), which is likely to be worm scanning, has a cut-off that is shorter by roughly one order of magnitude, indicating that it doesn't have very large scanners. Finally, Telnet (TCP/23) scanning clearly exhibits a light-tailed distribution with $\alpha = 2.82$; we speculate that this implies worm-like behavior (possibly by IoT botnets) of Telnet (TCP/23) scanning that is far different from the non-worm scanning behavior where the "worst offenders" contribute a significant portion of the scanning traffic [35].

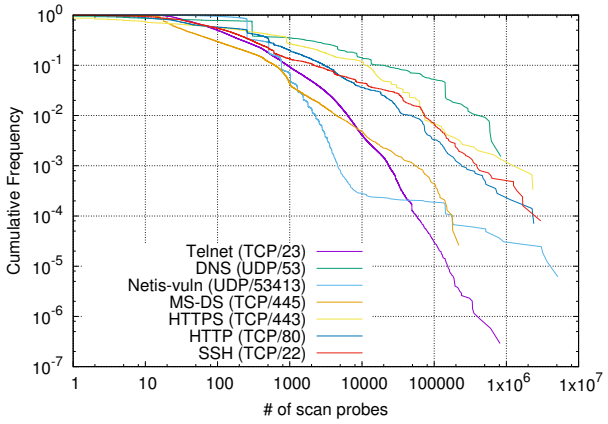


Figure 11: CCDF of scanner's scan volume

[14]To the best of our knowledge, it is unknown how the Netis vulnerability is widely exploited in the security community. We suspect that it is exploited in the form of self-spreading worms or botnet-like infection models to reveal its connection-level behavior appearing as the short-tailed component in the figure.

5 DISCUSSION

5.1 Telnet Scanning and IoT Botnets

In unveiling the identity of Telnet scanners, we have encountered two limitations: the firewall policy is written to block most inbound TCP/23 connections. In consequence, no botnet infections or C&C communications were observed. Also, we were unable to analyze the payload of packets since our data set only comprised connection metadata. We believe that it is ultimately required to instrument honeypots or active responders [26] with packet-level measurement to taxologically identify and study the behavior of Telnet scanners.

Nevertheless, we have studied several aspects of Telnet scanners to come up with some compelling evidence supporting the idea that they are seemingly IoT botnets. First, they were not responsible scanners (Section 4.1). Second, a large portion of Telnet scanning originated from dynamically allocated IP addresses (Section 4.2) of ISP ASes (Section 3.3). Third, Telnet scanning traffic was more pronounced in scannings with small fanout than in large-scale scannings, such as Internet-wide scannings; we also point out that the observed scanning fanout is consistent with Mirai's hardcoded scanning rate (Section 3.4). Lastly, they do not exhibit the connection-level behavior of non-worm scanners (Section 4.4).

Table 8: Telnet and other scan probes by originating country

Telnet scan probes		Other scan probes	
China	11.10%	China	41.12%
Brazil	10.86%	United States	22.18%
Vietnam	9.15%	Seychelles	6.21%
Taiwan	7.83%	Korea, Republic	5.98%
Korea, Republic	5.99%	Taiwan	5.37%
Turkey	5.27%	Japan	2.56%
Columbia	4.62%	Netherlands	1.40%

Finally, Antonakakis et al. have claimed that the Mirai initiated its infection on August 1st, 2016 by identifying Mirai scan probes with their eccentric TCP sequence number behavior [4]. The report also points out that the scan origins are geographically biased to South America and Southest Asia, i.e., Brazil, Colombia, and Vietnam. An earlier Mirai report by Incapsula also made an observation that Vietnam and Brazil were the top two Mirai origin countries [17]. The scan origins (in Table 4) have been divided between Telnet (TCP/23) and other scans in Table 8 to find that Brazil and Vietnam belong to the top three origins for Telnet scanning only. Other scan probes are mostly contributed by the usual suspects, such as China and the United States. [15] We believe that China is still the most prolific country for Telnet scans, since our categorization of Telnet scanners should have included non-botnet scanners.

An emergence of Telnet traffic in June of 2016 – two months before the alleged Mirai outbreak in August – was also noticed from the port activity dataset of the Internet Storm Center [5]. Figure 12 shows the number of Internet sources (per day) of TCP/23 port events which have been reported by distributed firewalls and IDS.

[15]It turns out that the geographical location, i.e., country, of the IP addresses belonging to the Dutch bullet-proof hosting service provider Quasi Networks LTD is identified as Seychelles, the third most contributing country of non-Telnet scans, by Maxmind [24].

Note that the number of Internet sources of Telnet port events suddenly increased around June prior subsequent surges around August and mid-September of 2016. [16] Nonetheless, ISC's data set cannot be utilized to provide quantitative scanning reports like ours. We further discuss this limitation in Section 5.3.

5.2 IP Space Variability

Network telescope [25] is a widely adopted approach in studying Internet-scale phenomena. As we have shown in Section 4.3, discrepancies with respect to target services across measured networks, or the IP address spaces they are assigned to, hamper quantitatively precise understanding. It not only affects measurement accuracy, but also has implications on security-related instrumentations such as honeypot deployment and middlebox load-balancing; honeypots in one network location might not be as effective as honeypots in another location. Also, the static load-balancing of security middleboxes, whose performance is proportional to incoming threats, would result in inefficient operations if one only considered the size of network. Finally, this problem should be more pronounced with the far more sparse and expanded address space of IPv6.

5.3 Defense

An immediate and effective defense of scanning attacks from existing IoT botnets is blocking TCP/23 at vantage network locations; our enterprise fortunately applied this policy in the first place, except for only a small number of IP addresses, to show no outbound Telnet scanning during the entire measurement period.[17] However, the vulnerability exploitable via the Telnet service port is only a tip of thousands of vulnerabilities we have to deal with in future [37]. Network security of IoT devices and related services should be fundamentally revisited in this sense. Although the connection success probability-based detection of scanners, e.g., TRW [19], should be able to block such scanning attempts, prevalent P2P application traffic will be innocently blocked by such detection scheme in an ordinary enterprise setup with firewalls or DHCP (see Section 2.2.1). We have suggested a way to discern scanners from such innocent P2P peers.

We collected our data two months prior to when Mirai was unveiled in August 2016 [23]. A massive DDoS attack on a security journalist's web site [20] and the Dyn cyberattack [16] followed in September and October 2016, respectively. The unprecedented increase of Telnet scannings we reported would have served as an early signal which one may have scrutinized in order to come up with a defense if there had been an Internet-scale scanning measurement system. To the best of our knowledge, Internet Storm Center [5] is the most similar project to this kind. It collects voluntary submissions of firewall logs of denied connections and portscan logs of NIDS to generate an aggregate report of Internet intrusions worldwide. However, it has limitations in providing comprehensive scanning reports; for example, firewall rejection and portscan events are very different in nature but inseparable in the report. Also, the system in need requires collecting all connection logs – including those that are *allowed* by firewalls or *unalerted* connection logs of NIDS – for the construction of a comprehensive scanning report as the ultimate goal of such a system is to have us alerted with unknown imminent threats (not only readily identifiable or blocked ones).

[16] The Mirai scanning traffic suddenly increased along with newly initiated TCP/2323 scanning traffic. It explains why TCP/2323 scan traffic is rarely observed in our measurement data (Table 2), while the disclosed Mirai source code [18] indeed contains a scanning module which scans TCP/2323.

[17] We only observe outbound Telnet scanning probes as part of several vertical scanning activities.

5.4 Limitations of Data Set

Scanning trends keep changing due to the shifts of popular services and vulnerabilities exploited by popular network attacks; we have indeed shown that the scanning landscape is greatly shifted from what was observed only two years ago. At the time of publication of this paper, it is possible that the scanning landscape will be shifted disparately from what is reported by us, since our data set is almost two years old. However, ISC dashboard shows that the Telnet port is still most frequently alerted as of March 2018 [5].

Another limitation of our data set is the possibility of source IP spoofings in scan probes. Scanner information, such as geographical location and identity as a responsible scanner, in our results are retrieved from the source IP addresses of our firewall connection logs. As the scanning attempts are characterized by high probability of connection failure (see Section 2.2), the origin (scanner) IP addresses of such probes cannot be reliably validated by a successful handshake. In other words, we cannot rule out the possibility of source IP spoofings in our results. This problem also holds for the darknet measurement studies since connections are never established. We believe estimating the degree of source IP spoofings in connection attempt logs is a challenging problem which requires a non-trivial network instrumentation.

6 RELATED WORK

Network scanning has often been studied by analyzing packet traces destined to unused IP address spaces (or darknet): Pang et al. studied *Internet background radiation* which comprises network scanning traffic and other nonproductive traffic such as backscatter or benign traffic by misconfiguration [26]. It has been revisited to find significant changes in the types of malicious activities [34]. Durumeric et al. have analyzed darknet traffic to study Internet-wide scanning activities [8]. On the other hand, connection-level information – usually collected at firewalls or NIDS – is also utilized to study global characteristics of Internet intrusions [35] and a 12-year-long historical trends of TCP scanning activities [3].

As for large-scale studies of network threat sources, Ramachandran and Feamster analyzed the network-level distribution of spammers and the effectiveness of reputation-based blacklisting [29]. Shin and Gu reported empirical results on Conficker infected hosts and their distribution in anticipation of source-based containment [30]. We similarly study scan sources in additional contexts of responsible scanners and IP-based blocklists. A number of responsible scanners are first surfaced by Durumeric et al. [8]. Honeypots [28] are also instrumented to further study scan sources [4].

Scanner classification schemes from connection-level packet metadata are devised for fast and accurate online detection: Staniford et al. proposed a stealthy portscan detection scheme by anomaly sensing and correlation [32]. Jung et al. developed TRW, which identifies scanning hosts using sequential hypothesis testing [19]. Finally, Antonakakis et al. have performed the first in-depth study of the Mirai botnet from a number of measurement perspectives [4].

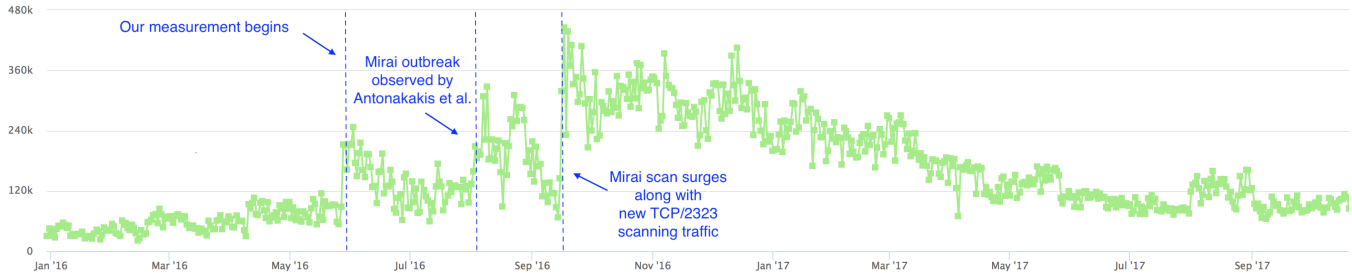


Figure 12: TCP/23 port event sources collected by ISC

7 CONCLUSION

In this paper, we presented the current network scanning landscape that has markedly shifted from previous studies, through a large-scale enterprise network measurement and scanning identification process. Our results show inflated scanning traffic by an unprecedented rise in Telnet scannings. We also studied the identified scanners and their behavior in various perspectives to report interesting findings including that the Telnet scanners are seemingly IoT botnets.

ACKNOWLEDGMENTS

This work is supported by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2015-0-00189 Development of Core Technologies for SDN-based Moving Target Defense and No. 2016-0-00078 Cloud based Security Intelligence Technology Development for the Customized Security Service Provisioning)

REFERENCES

[1] abuseat.org. 2016. The CBL. (2016). http://www.abuseat.org
[2] University of Oregon Advanced Network Technology Center. 2005. University of Oregon Route View Project. (2005). http://www.routeviews.org
[3] Mark Allman, Vern Paxson, and Jeff Terrell. 2007. A brief history of scanning. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 77–82.
[4] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Luca Invernizzi, Michalis Kallitsis, and others. 2017. Understanding the Mirai Botnet. (2017).
[5] SANS Internet Storm Center. n.d.. Internet Storm Center. (n.d.). https://secure.dshield.org
[6] Catalin Cimpanu. 2016. There's a 120,000-Strong IoT DDoS Botnet Lurking Around. (2016). http://news.softpedia.com/news/there-s-a-120-000-strong-iot-ddos-botnet-lurking-around-507773.shtml
[7] Codenomicon. 2014. The Heartbleed Bug. (2014). http://heartbleed.com
[8] Zakir Durumeric, Michael Bailey, and J Alex Halderman. 2014. An Internet-Wide View of Internet-Wide Scanning.. In *USENIX Security*. 65–78.
[9] Sam Edwards. 2016. Hajime: Analysis of a decentralized internet worm for IoT devices. https://security.rapiditynetworks.com/publications/2016-10-16/hajime.pdf, (2016).
[10] elvanderb. 2015. TCP-32764: some codes and notes about the backdoor listening on TCP-32764 in linksys WAG200G. (2015). https://github.com/elvanderb/TCP-32764
[11] Shadowserver Foundation. 2017. Shadowserver. (2017). https://www.shadowserver.org/wiki/
[12] Fernando Gont, Guillermo Gont, and Carlos Pignataro. 2013. *Recommendations for filtering ICMP messages*. Internet-Draft. IETF Secretariat. http://www.ietf.org/internet-drafts/draft-ietf-opsec-icmp-filtering-04.txt
[13] Zvi Gutterman, Benny Pinkas, and Tzachy Reinman. 2006. Analysis of the linux random number generator. In *Security and Privacy, 2006 IEEE Symposium on*. IEEE, 15–pp.
[14] Kaoru Hayashi. 2014. IoT Worm Used to Mine Cryptocurrency. (2014). https://www.symantec.com/connect/blogs/iot-worm-used-mine-cryptocurrency
[15] Bruce M Hill and others. 1975. A simple general approach to inference about the tail of a distribution. *The annals of statistics* 3, 5 (1975), 1163–1174.
[16] Scott Hilton. 2016. Dyn Analysis Summary Of Friday October 21 Attack. (2016). http://dyn.com/blog/dyn-analysis-summary-of-friday-october-21-attack/
[17] Incapsula. 2016. Breaking Down Mirai: An IoT DDoS Botnet Analysis. (2016). https://www.incapsula.com/blog/malware-analysis-mirai-ddos-botnet.html
[18] jgamblin. 2016. Mirai Source Code. (2016). https://github.com/jgamblin/Mirai-Source-Code
[19] Jaeyeon Jung, Vern Paxson, Arthur W Berger, and Hari Balakrishnan. 2004. Fast portscan detection using sequential hypothesis testing. In *Security and Privacy, 2004. Proceedings. 2004 IEEE Symposium on*. IEEE, 211–225.
[20] Brian Krebs. 2016. KrebsOnSecurity Hit With Record DDoS. (2016). https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos/
[21] Abhishek Kumar, Vern Paxson, and Nicholas Weaver. 2005. Exploiting underlying structure for detailed reconstruction of an internet-scale event. In *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*. USENIX Association, 33–33.
[22] The Smaphaus Project Ltd. 2017. The Spamhaus Project - ZEN. (2017). https://www.spamhaus.org/zen/
[23] MalwareMustDie. 2016. MMD-0056-2016-Linux/Mirai, how an old ELD malcode is recycled. (2016). http://blog.malwaremustdie.org/2016/08/mmd-0056-2016-linuxmirai-just.html
[24] Inc. MaxMind. 2017. GeoLite2 Free Downloadable Databases. (2017). http://dev.maxmind.com/geoip/geoip2/geolite2/
[25] David Moore, Colleen Shannon, Geoffrey M Voelker, and Stefan Savage. 2004. *Network telescopes: Technical report*. Department of Computer Science and Engineering, University of California, San Diego.
[26] Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry Peterson. 2004. Characteristics of internet background radiation. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. ACM, 27–40.
[27] Phillip A Porras, Hassen Saïdi, and Vinod Yegneswaran. 2009. A Foray into Conficker's Logic and Rendezvous Points.. In *LEET*.
[28] Niels Provos and others. 2004. A Virtual Honeypot Framework.. In *USENIX Security Symposium*, Vol. 173. 1–14.
[29] Anirudh Ramachandran and Nick Feamster. 2006. Understanding the network-level behavior of spammers. In *ACM SIGCOMM Computer Communication Review*, Vol. 36. ACM, 291–302.
[30] Seungwon Shin and Guofei Gu. 2010. Conficker and beyond: a large-scale empirical study. In *Proceedings of the 26th Annual Computer Security Applications Conference*. ACM, 151–160.
[31] Shodan. 2017. Shodan. (2017). https://www.shodan.io
[32] Stuart Staniford, James A Hoagland, and Joseph M McAlerney. 2002. Practical automated detection of stealthy portscans. *Journal of Computer Security* 10, 1-2 (2002), 105–136.
[33] Visgean. 2011. Zeus Trojan Horse Source Code. (2011). https://github.com/Visgean/Zeus
[34] Eric Wustrow, Manish Karir, Michael Bailey, Farnam Jahanian, and Geoff Huston. 2010. Internet background radiation revisited. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 62–74.
[35] Vinod Yegneswaran, Paul Barford, and Johannes Ullrich. 2003. Internet intrusions: Global characteristics and prevalence. *ACM SIGMETRICS Performance Evaluation Review* 31, 1 (2003), 138–147.
[36] Tim Yeh. 2014. Netis Routers Leave Wide Open Backdoor. http://blog.trendmicro.com/trendlabs-security-intelligence/ netis-routers-leave-wide-open-backdoor/. (2014).
[37] Tianlong Yu, Vyas Sekar, Srinivasan Seshan, Yuvraj Agarwal, and Chenren Xu. 2015. Handling a trillion (unfixable) flaws on a billion devices: Rethinking network security for the Internet-of-Things. In *Proceedings of the 14th ACM Workshop on Hot Topics in Networks*. ACM, 5.