Profiling Internet Scanners: Spatiotemporal Structures and Measurement Ethics

Johan Mazel
NII/ANSSI/JFLI
Email: johan.mazel@ssi.gouv.fr

Romain Fontugne
IIJ Research Lab
Email: romain@iij.ad.jp

Kensuke Fukuda
NII/Sokendai
Email: kensuke@nii.ac.jp

***Abstract*—Scanning is ubiquitous on the Internet. It assists administrators to troubleshoot their own network, researchers to survey the Internet, and malicious actors to assess the attack surface of targeted networks. As users requirements vary, scans in the wild exhibit very diverse characteristics. For example, the coverage, stealthiness and probing speed are drastically varying from one scanning IP to another. In this paper, we study 15 years of backbone traffic to understand the evolution of mass-scanning tool usage, scanning pattern and the concentration of scanning IPs (also called scanners) in small networks. We also propose a new method to classify scanning IPs' spatial and temporal structure into three profiles that reveal vastly different intent. In particular, we find that 33% of scanners repeatedly target the same set of hosts. If unsolicited, identifying this behavior provides good insights on the malicious intent of scanners. In the case of innocuous scanners, publicly documenting scanning activities and giving right to opt out are common ethical practices. Our study shows that documented scanning IPs behave differently from the vast majority of scanners. Furthermore, only 39% of these entities follow online documentation best practices.**

I. Introduction

Scanning IPs (also called scanners) send packets to numerous destinations and analyze corresponding answers, or lack thereof, to acquire knowledge on remote hosts or networks. Consequently they can catalog online hosts and services with OS [1] and network equipment [2] fingerprints. Probing may be defensive, to acquire knowledge on one's own network, or offensive, to assess attack surface of a targeted network [3]–[5].

Mass-scanning tools [6], [7] allow researchers to perform extremely fast probing. These tools may thus cause many alarms although the scans are benign. Understanding scanners intent is paramount to assess existing threats. Our analysis investigates the use of mass-scanning tools and probing patterns along time, as well as, temporal and spatial structure of scans. These aspects provide new insights into scanner sophistication and intent, and complement the rich literature on Internet scanning [8]–[15]. As stated above, high speed probing performed with mass-scanning tools, such as ZMap [6] or Masscan [7], may increase network administrators workload by generating many alarms. Probing entities thus have moral obligations to document their activities so that administrators can easily understand that detected probing is innocuous and give them the right to opt-out of the scans [16]–[18]. Up to our knowledge, this particular facet of scanning ethics has not yet been studied. In this paper, we address all these aspects. We study recent scanning trends, profile scanning IP behaviors and thus reveal potentially malicious intent, and, identify security research-related probing and assess how well scanning entities document their activities.

The contribution of this paper is three-fold. First, we expose scanning recent evolution and show that the use of Internet-wide scanning tools and random probing pattern are increasing. Second, we present a scanner behavior profiling method. This method defines three profiles: scanners either contact unrelated small network prefixes (41% of scanners), or randomly probe the same prefix for a long duration (8% of scanners), or repeatedly scan the same set of hosts (33% of scanners). This last behavior signals a specific probing interest. This provides strong evidence to administrators regarding the malicious intent of scanners that may attack the probed network. Third, we analyze security researchers' scanning (which represents 0.1% of all scanners) and show that: 1) their behavior is distinct from the others scanners in terms of both targeted services and behavior profiles, and, 2) most identified scanning entities partially follow online documentation best practices.

II. Related work

A great deal of attention has been dedicated to scan detection and analysis. Previous works identify scanners as hosts with a high rate of unsuccessful connection [19], or map existing services in a network and consider hosts reaching unavailable services as scanners [20]. For a more complete account of scan detection techniques, we refer the reader to [21].

Several works provide analysis of scans observed in the wild. Some works use stub network datasets: 12 years of IDS logs from June 1994 until December 2006 [8], or, data from the Dshield repository: one month in 2001 and three months in 2002 [9], and the first fifteen days of 2005 [10]. Other studies provide simple description of scans in backbone traffic [11], [12] and darknet traffic [13]–[15]. To evade detection, scanning mechanisms have evolved from simple sequential probing of the IP space to more complex probing schemes [6], [7], [22], [23].

The introduction of mass-scanning tools allow researchers to perform extremely fast probing that can potentially raise many alarms on targeted networks. This increases network

administrators workload. Existing work on network measurement ethics [16], [17] states that researchers should limit measurements' impact on regular users. In the scanning context, this means that scanners must reduce their aggressiveness and appropriately document their activities to ease the work of network administrators [18]. Moreover, legal aspects have also been considered [24]. It appears that appropriate documentation can demonstrate good faith in case of a lawsuit. Up to our knowledge, there is no study of scanning ethics in the wild.

III. Dataset

We analyze network traffic traces from the MAWI repository [25], which is a collection of daily traces measured from 14:00 to 14:15 JST since January 2001 at a backbone link connecting Japanese universities and research institutions to the Internet. It mainly consists of international traffic between universities and commercial ISPs. The Autonomous System (AS) where monitoring is performed announces 8 prefixes through BGP that add up to a /14. Customers ASes' prefixes add up to a /13. Although the duration of each MAWI trace (i.e. 15 minutes) limits our study to a fraction of the daily traffic, the MAWI repository enables us to inspect scanning trends over 15 years. We name this repository "MAWI longitudinal". We also use several multi-day long traces captured on the same measurement point during the Day In The Life of Internet (DITL) events in 2012, 2013, 2014 and 2015. The respective duration of these traces is 63, 72, 24 and 48 hours. We refer to this repository as "MAWI DITL". "MAWI longitudinal" is used in Section IV and "MAWI DITL" is used Section V. Both repositories are leveraged in Section VI. Unlike the publicly available MAWI traces where IP addresses are anonymized, our dataset contains original IP addresses to cross-reference our data with other datasets (DNS, Censys [26], BGP). Less than 2.5% of hosts are behind NAT [27]. We are thus confident that scans from the same scanner, spread in time, originate from the same host.

Abnormal events appearing in the MAWI repository are reported in the MAWILab [11] database then classified and annotated with a taxonomy designed for backbone network anomalies [12]. In this paper, we make use of these results and study the characteristics of traffic annotated with *network scan* labels (i.e. labels with the prefix: $network\_scan$). This ensures that corresponding traffic has a single source and a high number of destinations (> 20). Protocol header information (SYN, ACK, FIN flags for TCP and ICMP type Echo request, Netmask request and Timestamp request for ICMP) is also used to identify different types of network scans. Although the taxonomy identifies UDP scans, we analyze only TCP scans (56% of all scans) because flag-based signatures reduce false positives. To assess the reliability of MAWILab events, we compare the source IP address of events annotated as network scans in the MAWI traces with the IP addresses reported by the SANS Internet Storm Center (ISC) [28] from November 2014 to March 2015. 55% of MAWILab network scans are also present in ISC's suspicious domains. This shows that the majority of IP addresses labeled as scans are also detected by the firewalls participating in the DShield project. SANS ISC and MAWI use very different data source: MAWI uses a single measurement point on Japanese backbone while SANS ISC leverages the DShield sensor, a collection of firewall logs from across the world. This explains why the overlap between MAWI and SANS ISC is not complete.

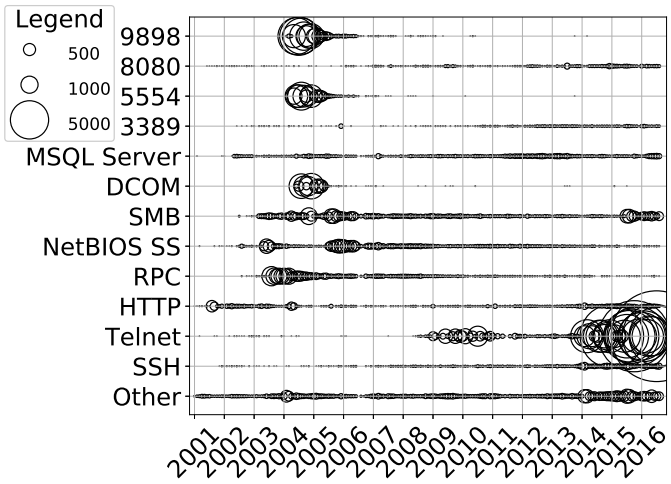


Figure 1: Port targeted by TCP network scans (radius represents the monthly number of scans).

IV. Macroscopic trends and recent evolution

Our study starts with the evolution of scanning characteristics in the 15 year-long "MAWI longitudinal" dataset.

A. Destination port

Figure 1 depicts TCP scans along the 15 years of analyzed traffic. For each scan, we retrieve the dominant destination port. We observe two types of trends. In the first case, some ports or services quickly arise and then slowly decay. The most noticeable example of this is ports linked to worm like ports 9898 (Dabber), or, 1023 and 5554 (Sasser) in 2004. Other services such as RPC (port 135 linked to Blaster worm) experience similar surge but decrease slower. As noted by [8], the decay is likely due to disinfection. A sudden surge in Telnet scans occurs in March 2014. These scans targets Telnet-enabled Internet-of-Things (IoT) devices such as cameras [29]. Contrary to all previous sudden surges, this one does not show any sign of decrease. We hypothesize that this scanning increase is due to the absence of security updates on IoT devices and the regular addition of vulnerable devices on Internet. The second main trend is constituted of classic application or destination ports that were already present 15 years ago and that remain in use today. They are thus constantly scanned during the whole duration of our study. We can here quote SSH (port 22), HTTP (port 80), SMB (port 445), MS SQL Server (port 1433), and HTTP alternative (port 8080). Although not shown here, FTP (port 21) and HTTPS (port 443) exhibit the same behavior.

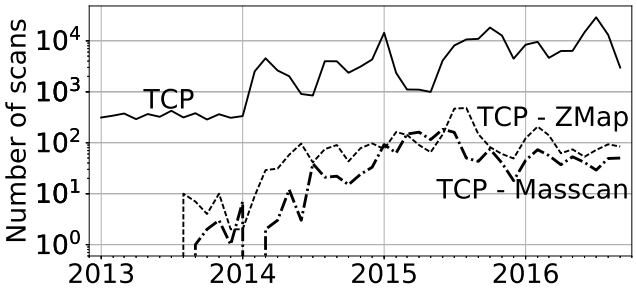


Figure 2: Monthly number of scans exhibiting Zmap and Masscan fingerprints in MAWI longitudinal.

These observations are qualitatively consistent with past literature [8]. Furthermore, recent Telnet scanning surge shown on Figure 1 motivates the constant attention that researchers should maintain on network scans.

B. Mass-scanning tools

Leonard and Loguinov [30] proposed the first Internet-wide scanning tool and showed that their scanning pattern is as polite as possible [22]. Open source scanning tools ZMap [6] and Masscan [7] were then released in, respectively, August 2013 and September 2013. They are able to perform a wide variety of scans using TCP, UDP and ICMP protocols, and implement specific packet fingerprints in the ID field in the IP header [15] that allow easy identification. Figure 2 displays the total number of network scans along with the number of ZMap and Masscan scans. If 95% of packets of a scan match a tool's fingerprint, the scan is considered as having been performed with the considered tool. Following the release of both tools, the number of associated scans immediately arises but then almost disappears. This might be due to initial curiosity. The number of fingerprinted scans then re-increases in the beginning of 2014. Overall, ZMap is more prevalent than Masscan.

Durumeric et al. [15] observed ZMap and Masscan usage in darknet. Their results are difficult to compare to ours because they use a different network scan definition: scans need to reach more than 100 destinations with at least 10 packets per second. Furthermore, it is easy for a malicious actor to remove ZMap and Masscan fingerprints because they are open-source. We may thus underestimate these tools' usage.

C. Scanning patterns

Naive network scans use (incremental or decremental) sequential pattern to reach all addresses in a targeted prefix. Randomizing the successive scanned IP addresses reduces the traffic received by every targeted subnetwork in a certain time-window thus reducing the odds of detection by network administrators. It thus reduces detection odds. A SIP scan [23] using byte-reverse order permutation has previously been observed. Compared to naive sequential scanning, this pattern spread the probing load over the complete targeted prefix at any point in time. The use of this pattern indicates that the attacker, the Sality botnet [31], wanted to avoid detection. Studying scanning patterns provide insights into scanners' sophistication and intent. We here test the monotonicity of probed destination IP addresses using the Mann–Kendall test, a nonparametric trend test [32]. We use a significance level of 0.5% to avoid false positive as suggested by Li et al. [32]. Scans that do not exhibit any trend are considered as random.

Figure 3 displays the number of scans and proportion of random ones. We here consider Telnet scans separately because they constitute the overwhelming majority of scans after March 2014. The overall tendency for non-Telnet scans is an increase in random pattern use. Proportion values are high for the early years but those values are not reliable due to the small number of scans. The increase in random scanning starts at the beginning of 2012. Telnet scans however remain massively non-random across the dataset. As stated above, the purpose of random scanning is to spread the probe load uniformly across the targeted IP range. This makes detection from stub networks much more difficult because of the small number of packets that probe each subnetwork in a given time window [22]. Our results thus show that scanners increasingly use scanning patterns that aim at avoiding detection.

Existing probing patterns analyses in darknet snapshot data support our results. Bou-Harb et al. [33] find that 57% of scans in 2013 are random. Similarly, Fukuda et al. showed that, in November 2006, 10-15% are randomly behaved [34].

D. Scanners proximity in /24 network prefixes

ZMap allows users to probe from a single machine using an IP range. In order to analyze potential occurrence of such scanning, we check whether scanners are located in the same /24 prefix. In the remainder of the paper, scanners that are located in the same /24 prefix of another scanner within a one-month time window are called "co-located". Scanners that are not located in the same /24 as any another scanners are called "lonely"

Figure 4 displays the total number of scanners along with the proportion of lonely scanners. We here notice that in 2004 and, from 2014 to 2016, the proportion of lonely non-Telnet scanners are decreasing. We then compare lonely and co-located scanners characteristics. Co-located scanners tend to uses ZMap and Masscan more than lonely scanners. Less than 12% (respectively 4%) of co-located scanners uses ZMap (resp. Masscan) from 2013 to 2016. Those percentages are equal to 5% for lonely scanners. By analyzing non-Telnet scanners' AS, we identify nine ASes providing hosting or colocation services whose occurrences increase in 2014 and 2015. Among these ASes, we observe Quasi Networks (ex-Ecatel) which is also reported in [15]. Three of these ASes are not visible as sources of scans before 2014. From the start of 2015 to 2016, 69% of co-located scanners and 69 % of associated scans, originate from these nine ASes. Co-located scanners originate from 12911 distinct /24 prefixes with 6447 of them containing non-telnet scanners. We hypothesize that

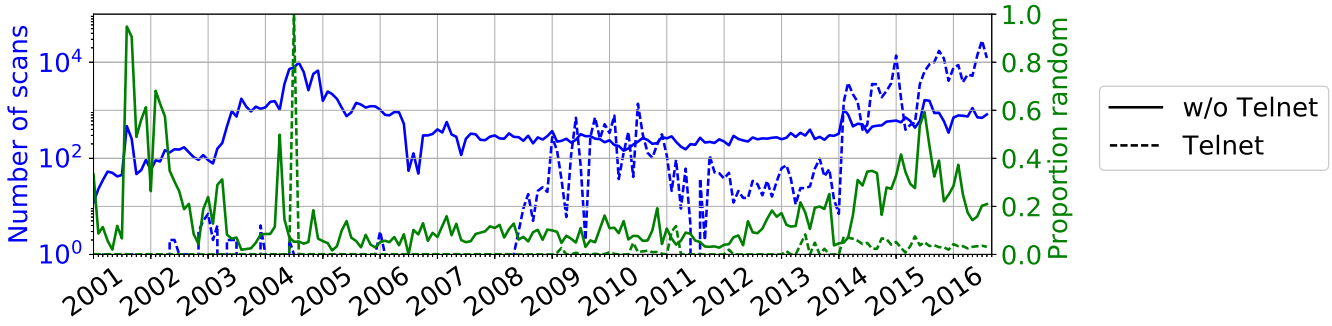


Figure 3: Scanning patterns in MAWI longitudinal: monthly number and percentage of scanners that use random patterns.

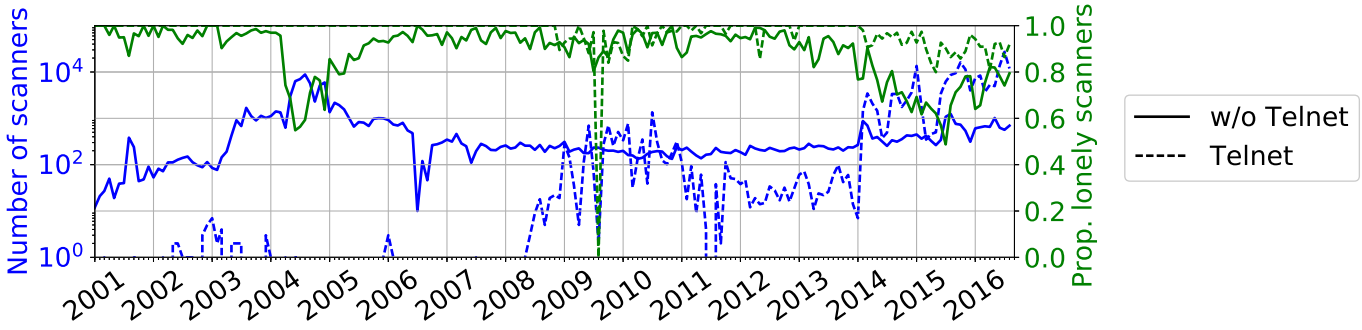


Figure 4: Monthly number of scanning IPs (scanners) and proportion of lonely scanner (alone in /24 network prefix) in MAWI longitudinal.

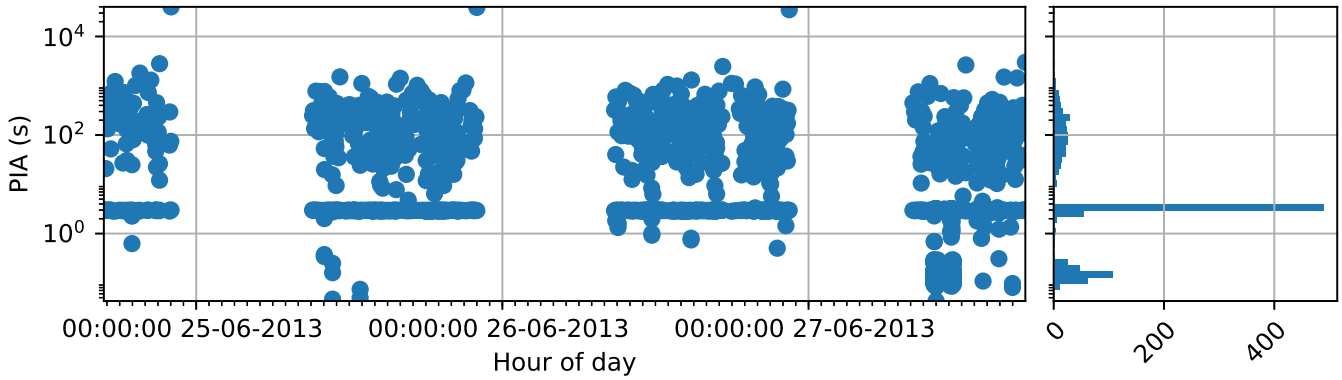


Figure 5: Packet Inter-Arrival (PIA) for a single scanning IP (scanner) observed in the MAWI DITL 2013 that targets port 445 (SMB). This scanner performs 4 scans separated by ten hour gaps that constitute a single *activity period*.

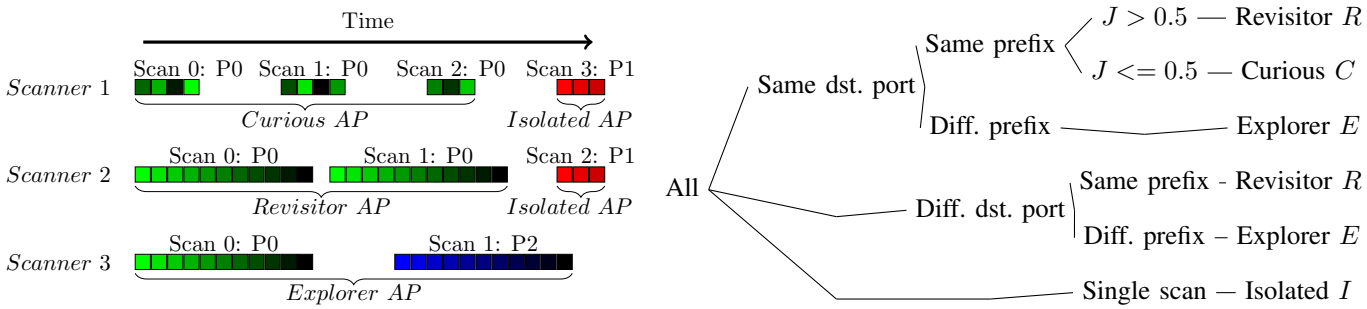
observed co-located scanners are mostly unrelated and only located in the same /24 prefix because they use the same services. Over 15 years, co-located scanners represent less than 11% of scanners on average.

V. Profiling Scanners

Previous section observes recent scanning trend regarding mass-scanning tools, pattern and proximity of scanner in the IP address space that are visible in our 15 year-long dataset, and reveals that scanning sophistication is increasing. This sophistication increases however does not necessarily signal nefarious intentions. For example, random scanning reduces alarms on targeted network, but this is both interesting for malicious actors that want to avoid detection and security researchers that do not want to increase administrators' workload. This section inspects the temporal and spatial patterns of scanners in order to infer scanners' intent.

A. Temporal and spatial structure example

Figure 5 depicts the Packet Inter-Arrival (PIA) times of the packets sent by a scanner identified by MAWILab in the 3-day long 2013 DITL trace. By analyzing the destination IP



(a) Activity period examples

(b) Classification of activity period: J: Jaccard index-based overlap

Figure 6: Scanners' activity period examples (a) and classification (b). Scanner 1 executes three scans that always target the same prefix P0 but never the same addresses inside P0. This is a curious activity period. This scanner then performs a single scan (targeting P1) which constitutes an isolated activity period. Scanner 2 exhibits a revisitor activity period made of two scans that repeatedly target the same addresses in P0. Similarly to scanner 1, scanner 2 also exhibits an isolated activity period. Scanner 3 performs two scans that target different prefixes (P0 and P2) that constitutes an explorer activity period.

address, we observe that all probes target all our monitored prefix which are spread in a /1 network prefix. All probes reach the destination port 445. This scanner sends two probes to each destination but successive pairs of packets do not target adjacent IP addresses. Namely, the pattern does not exhibit any trend and is thus random (see Section IV-C). This scanner exhibits two specific PIA values. The small PIA value (2-3 seconds) separates two probes sent to the same destination address. The high PIA values (between 5s and 1 hour) are due to inactive periods between pairs of probes. This scanner performs 4 scans that are separated by a ten hour pause in a periodic pattern. This example suggests that the considered scanner performs several scans that share some characteristics. We name this group of related scans an *activity period*.

Activity periods are sequence of scans originating from the same scanning IP that share some characteristics. An activity period's start and end times are the start time of its first scan and the end time of its last scan. We do not use any timeout and only rely on MAWILab alarms' timestamps. Figure 6a presents three examples of scanner that contain different type of activity periods. Similarly to the example provided in section V-A, Scanner 1 performs three scans that always target the same prefix P0 but never the same addresses inside P0. It gradually increases its knowledge on P0 scan after scan. These three scans constitute a "curious" activity period. Scanner 2 performs two scans that repeatedly target the same addresses in P0: this is a "revisitor" activity period. This scanner executes several scans on P0 and thus captures its dynamic. Both scanner 1 and 2 also perform a single scan targeting P2. This is an "isolated" activity period. Scanner 3 performs two scans that target different prefixes (P0 and P2) that constitutes an "explorer" activity period.

B. Activity period classification

We now present the classification method that we use to extract the activity period types presented above. Examples presented on Figure 6a do not take into account the targeted destination port. We thus add this criterion to improve the granularity of activity period classification. First, we discard scanners that perform only one scan. Then, we iteratively classify scanner activity periods into the categories presented in Figure 6b from the top to the bottom using the destination port and targeted network prefix as criteria. We thus first check if successive scans target the same destination port and reach hosts located in the same network prefix (see Figures 5 and 6a). For each scanner, if successive scans target similar network prefix, they are grouped in an activity period. The targeted network prefix of a single scan is defined as the CIDR prefix that contains all the destination addresses of the considered scan. Two prefixes are considered as similar if they are equal, or if one is included in the other and the difference of the prefixes length is not greater than 1. Activity periods with scans that target similar prefixes using the same destination port are labeled as "revisitor" or "curious". Scans in "revisitor" activity periods visit the same set of destination hosts several times, hence the "revisitor" label (see Scanner 2 on figure 6a). This repeated behavior intends to capture the dynamic of the probed hosts. Scans in "curious" activity periods do not exhibit any spatial overlap but instead incrementally acquire knowledge on the IP address space (see Scanner 1 on figure 6a). They thus do not target a set of hosts but instead target a specific characteristic (e.g. vulnerability) on a large number of hosts. To separate "revisitor" activity periods from curious ones, we define the Jaccard index $J(A)$, a standard set similarity index, of an activity period A that contains n scans as: $J(A) = \frac{|\bigcap(S_i)|}{|\bigcup(S_i)|}, i \in 1 \dots n$ where S_i is the set of destination addresses of the i^{th} scan in A. A is labeled as "revisitor" if $J(A) > t_J$, "curious" otherwise. t_J is empirically fixed at 0.5, in between in the two extremes values: 0 and 1. Less than 22% of activity periods targeting the same prefix have J between 0.2 and 0.8. Successive scans

Table I: Number and percentage of IPs that perform each type of activity periods in MAWI DITL. All percentages are relative to the number of scanners for the considered dataset (each year or total). IPs that perform a single scan are displayed on "1 scan". IPs that perform several scans are accounted in the "Multiple scans" part of the table. Some of these IPs only contain 1 AP type; this means that their APs are either E or C or R. The sum of I, E, C and R is not equal to 100 because some scanners perform APs of different types.

Year		2012	2013	2014 w/o Telnet	2015 w/o Telnet	2014 Telnet	2015 Telnet	Total
Labels								
All		1416	1483	1654	11860	21293	63882	101362
1 scan		20%	27%	22%	59%	23%	29%	31%
Mutl. scans	1 AP type	44%	46%	40%	30%	42%	42%	41%
	Iso. (I)	29%	22%	36%	9%	26%	22%	22%
	Expl. (E)	62%	38%	66%	20%	52%	40%	41%
	Cur. (C)	3%	17%	17%	20%	1%	8%	8%
	Rev. (R)	34%	22%	7%	4%	37%	38%	33%

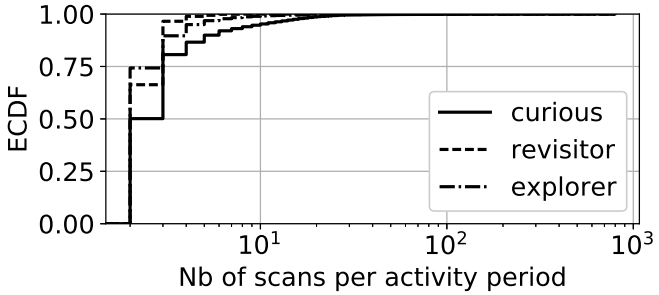


Figure 7: ECDF of the number of scans in activity periods for explorer (solid line), curious (dashed line) and revisitor (dash-dotted line) in MAWI DITL.

targeting network prefixes that are not similar (according to the CIDR and prefix length-based definition above) on the same destination port are labeled "explorer" (see Scanner 3 on figure 6a). We then follow a similar procedure to build activity periods out of scans that target distinct destination ports. Successive scans that always target the same hosts are classified as "revisitor". We did not observe scans that target the same prefix using distinct destination ports without IP address overlap, and thus did not introduce a label similar to "curious". Then, successive scans targeting prefixes that are not similar and distinct destination ports are labeled as "explorer". After extracting these three types of activity periods, remaining scans are labeled as "isolated" (see Scanner 1 and 2 on figure 6a).

Overall, "curious" activity periods exhibit a greater focus than "explorer" ones because they target a single prefix. "Revisitor" activity periods exhibit an even stronger intent by repeatedly probing the same IP addresses. This means that they are interested in acquiring information on a very specific target. Their occurrence can thus help network administrators anticipate security threats.

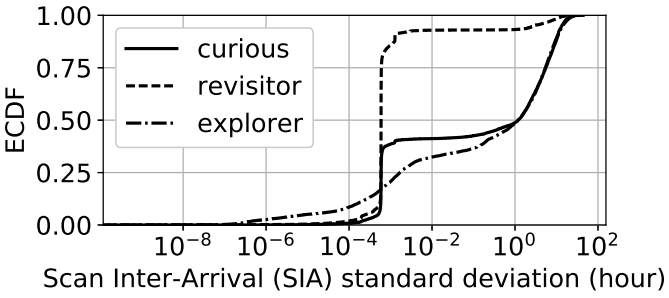


Figure 8: ECDF of standard deviation of Scan Inter-Arrival (SIA) time in activity periods for explorer (solid line), curious (dashed line) and revisitor (dash-dotted line) in MAWI DITL.

C. Classification results

We apply this method to the scans documented in MAWILab and found in MAWI DITL dataset. Results are displayed on Table I. Due to the Telnet scanning surge since 2014, we consider Telnet and non-Telnet probing separately. Overall, 69% of scanners perform more than one scan. 44% of scanners perform a single type of activity periods. 41% have explorer activity periods, 8% contain curious ones and 33% perform revisitor activity periods. 98% of revisitors and 34% of explorer target a single destination port. The sum of isolated, explorer, curious and revisitors is not equal to 100 because some scanners perform APs of different types. We first consider non-Telnet scanners. Revisitor activity periods are steadily decreasing while curious ones are increasing. We hypothesize that this increase is due to the rise of mass-scanning tools (see Section IV-B). Explorer activity periods do not exhibit any specific trend. Telnet scanners contain less curious activity periods and more explorer and revisitor than non-Telnet scanners. This is due to the fact that many Telnet scans target several /24 prefixes. Obtained percentages are very close to those obtained by removing scans performed by co-located scanners (see Section IV-D).

D. Activity period characteristics

We then compare the characteristics of the three types of activity period: explorer, curious, and revisitor. Figure 7 depicts the Empirical Cumulative Distribution Functions (ECDF) of the number of scans per activity periods. Curious activity periods contain a higher number of scans than revisitor and explorer activity periods. We then focus on scan periodicity by analyzing the standard deviation of Scan Inter-Arrival time. Figure 8 displays the ECDF of this metric for activity periods that contain at least three scans. Scans in revisitor activity periods occur in more regular interval. This is consistent with an automated periodic probing of a specific set of hosts.

By further analyzing activity period characteristics, we note that revisitors activity periods and scans have a shorter duration than explorer ones which are in turn shorter than curious. The same ranking applies to network prefix size

Table II: Entities activity in MAWI traces. Entities short names: TUM: Technische Universität München, ISP: Internet Scanning Project, NS Alliance: Network Security Alliance, PS: Proxy Scan, SBA: SBA Research. Dataset: M: MAWI (ex M14-16 → MAWI 2014-2016), D: DITL (ex: D14 → DITL 2014). # scans and # IP indicate the number of scans and the number of distinct IP from this entity. Prefix length range describes the size of the networks targeted by the considered entity.

Entity type	Entity name	Dataset	# scans	# IP	Prefix len. range	# destination port					Tool			Activity periods			
						22	80	443	8080	Other	Other	ZMap	Masscan	Isolated	Explorer	Curious	Revisitor
Academic	Berkeley	M15	3	1	1-18		3					3		-	-	-	-
		D15	6	1	1		6					6		1			
	Cambridge	M15	3	1	1-18		3					3		-	-	-	-
		D15	12	1	1-18		12					12		2			
	Michigan	M14-16	25	24	1-18	6	3	5		7	4	25		-	-	-	-
		D15	416	170	1-18	32	36	81		251	19	413		3		91	34
	TUM	M13-15	10	2	1-18	1	1	4		3	9	2		-	-	-	-
Company	Cymru	M13-15	6	1	1			6			6			-	-	-	-
		D15	6	1	1			6								1	
	Eddie	M14	3	2	1			3			3			-	-	-	-
	Cornejo	D14	8	1	1			8			8					1	
	Errata	M13-16	8	4	1		4	1	2	1			8	-	-	-	-
	Security	D14	16	1	1	8	8						16	10		2	
	IPredator	M14	2	1	1		2						2	-	-	-	-
	ISP	M14	2	1	1		2						2	-	-	-	-
	Labs	M14-16	61	27	1	1				60		61		-	-	-	-
	Rapid7	D14	12	11	1		12					12				1	
	NS Alliance	M13-16	19	2	1-8				1	18		1	18	-	-	-	-
	PLC	M14-16	18	1	1					18		18		-	-	-	-
	Scan	D14	33	1	1					33		33				1	
		D15	2	1	1					2		2				1	
	Project 25499	M14-15	17	5	1		2	5		10		17		-	-	-	-
	Proxy Scan	M15	7	1	1				7	7		7		-	-	-	-
	SBA	M15-16	1	1	1			1	1				1	-	-	-	-
	360.cn	M14-16	655	7	1	29	7	5	9	598	21	144	504	-	-	-	-
		D14	1135	3	1-2	87	31	29		980	8	1135		2		25	
		D15	19	4	1	1				18			19	3		4	1
	Shadow	M13-16	462	135	1-24		3	186		273	13	449		-	-	-	-
	Server	D13	72	1	1-24			72			72			10	3	21	
		D14	279	31	1			279				279				31	
		D15	3374	139	1-18		1251	425		1698		3374		125	34	685	4
	Shodan	M13-16	80	8	1	1	19	12	1	38	71			-	-	-	-
		D13	5	2	1			1		4	5					2	
		D14	35	1	1			17			53						2
		D15	803	10	1	5	43	53	73	627	805			9		208	4

of activity periods. Scans in revisitors and explorer activity periods exhibit similar network prefix coverage (i.e. they reach the same proportion of destination addresses inside their targeted prefix). Scans in curious activity periods have a much lower coverage. Similarly, scans of revisitor and explorer activity periods have a higher packet rate than curious ones. These observations are consistent with their activity periods roles: curious activity periods perform slow incremental scans and acquire knowledge one scan at a time, while revisitor and explorer ones quickly gather information on a smaller scope.

From a general point of view, curious scanners seem to target large prefixes in a random manner. Their goal is likely to scan a wide part of the Internet for a specific purpose. However, revisitor scanners exhibit an intent to acquire knowledge on a specific target. Our method thus helps administrators to understand the actual intent of scanners, and thus, provide a clear picture of the active threats towards their network.

VI. Publicly documented scanning entities

By analyzing DNS name of scanning IPs using ZMap and Masscan, we noticed several security researchers, from both universities and companies, such as University of Michigan or Shadow Server. We extended this analysis on our whole dataset, 15 years of MAWI longitudinal and DITL, using DNS and Censys [26], and identified 18 entities.

A. General results

We present our results in Table II. These entities' scans (resp. scanning IPs) represent 0.44% (resp. 0.1%) of all the observed scans (resp. IPs), 22% (resp. 18%) of ZMap ones and 32% (resp. 3.7%) of Masscan ones. Some entities perform many scans during a long period of time (e.g. Shadow Server)

Table III: Scanning entity identifiability for famous scanners. "Part." means not directly (but documented in blog posts).

Entity type	Entity name	# IP	IP DNS PTR prop	IP webpage prop	PTR webpage prop	Webpage information: Email contact	Optout	IPs/ Prefix avail.
Academic	Cambridge	1	1.0	1.0	1.0	Yes	Yes	Yes
	Michigan	181	1.0	1.0	0.92	Yes	Yes	Yes
	Berkeley	1	1.0	1.0	1.0	Yes	Yes	No
	TUM	2	1.0	1.0	1.0	Yes	Yes	Yes
Company	Cymru	1	1.0	1.0	1.0	Yes	Yes	Yes
	Eddie Cornejo	3	1.0	0.0	0.0	-	-	-
	Errata Security	4	0.75	0.5	0.5	Part.	Part.	Part.
	IPredator	1	1.0	0.0	0.0	-	-	-
	IS Project	1	1.0	1.0	1.0	Yes	Yes	No
	Labs Rapid7	27	0.93	0.93	0.93	Yes	Yes	Yes
	NS Alliance	2	0.0	0.5	1.0	No	Yes	No
	PLC Scan	1	1.0	0.0	1.0	Yes	Yes	No
	Project 25499	5	1.0	1.0	1.0	Yes	Yes	Yes
	Proxy Scan	1	1.0	1.0	1.0	No	Yes	No
	SBA Research	2	1.0	1.0	1.0	Yes	Yes	Yes
	360.cn	7	1.0	0.0	0.0	-	-	-
	ShadowServer	140	0.98	0.97	0.97	Yes	No	No
	Shodan	11	1.0	0.0	0.5	No	No	No

while others are only active punctually (e.g. SBA research). 55% of entities scan all our monitored prefixes which are spread in a /1 network prefix, while other entities target some of our monitored prefix spread in a prefix of lengths between 1 and 24. We only display the number of scans that target SSH, HTTP, HTTPS and POP. Some entities such as Michigan, 360.cn, Shadow Server or Shodan actually probe many other ports. The proportion of Telnet scans from identified entities is much smaller than that of all observed scans (see Figure 3). Entities mainly use ZMap and Masscan although ZMap is more prevalent. The increase between DITL 2013 and 2014 clearly emphasizes the rise of both tools (see Section IV-B): except Shodan, all identified entities now use ZMap or Masscan. Using the classification presented in Section V-B, we observe that until 2013, these entities' activity periods are isolated or explorer or curious. As they switch to use ZMap and Masscan, their activity periods almost completely become curious. Curious activity periods contain scans that target the same prefix but different IP addresses which is consistent with the probing behavior of ZMap and Masscan. As curious represents only 8% of all scanners (see Table I), identified entities are thus behaving differently from other scanners. This further shows that the method proposed in Section V provides accurate insights on scanner behavior.

B. Deployed online documentation infrastructure

Scanning can have detrimental effects on Internet users. Users sharing an Internet access with an aggressive scanner may experience reduced bandwidth. Operators of scanned network may also see a workload increase due to scan detection alarms. Scanning entities thus need to ensure than they minimize harm to other users [16], [17]. To this end, ZMap [6] documentation proposes several guidelines [18]. It especially emphasizes three aspects that are relevant for network administrators. First, researchers must state the benign nature of the traffic through DNS PTR record and webpages. Second, probing purpose must be explained. Third, one must provide contact information and the possibility to opt-out from probing. Beyond the obvious interest of informing administrators, appropriate documentation may also demonstrate good faith in case of lawsuit [24]. We here analyze how entities document their scanning and whether they propose opt-out mechanisms and contact information. Table III presents our results. We gather IP-reachable webpage using Censys.io [26] and automate crawling of scanning IPs' DNS PTR record and associated webpage.

All entities, except Network Security Alliance, provide PTR records for the majority of their scanning IPs. Some entities do not have any webpages accessible with either IP or PTR record that document their scanning: Eddie Cornejo, 360.cn and IPredator. Shodan redirects some PTRs to their homepage. Errata Security sets up webpages but only lists previous scanning results. Errata Security however documents its probing through blog posts [35]–[38] but this makes it difficult for operators to understand that the activity they investigate is innocuous scanning. We then manually analyzed every setup webpage. All entities provide contact email address except Network Security Alliance and Proxy Scan. These entities both scan port 8080 and use a form for opt-out request. We hypothesize that these are cooperating entities. Shadow Server webpages do not propose opt-out. Finally, many entities do not provide the IP addresses or prefixes that they use to perform scans: Berkeley, Internet Scanning Project, Network Security Alliance, Proxy Scan and Shadow Server.

Some entities that do not scan anymore or change their scanning IP along time may have removed DNS PTR record and/or webpages. We thus may miss some infrastructure that may have been setup in the past. We contacted entities that do not follow guidelines and asked them about the state of their infrastructure in the past. We chose to update Table III accordingly despite the fact that we cannot verify their statements. From a general point of view, only 39% of entities completely follow ZMap guidelines. Existing online scanning documentation is thus not sufficient.

VII. Concluding remarks

Scanning is a pervasive component of network traffic. We provide new insights into recent research such as the rise of Telnet scanning, and the increase of mass-scanning tool and random scanning patterns usage. We propose a new method that profiles scanners' behavior, and discover that 33% of scanners repeatedly target the same hosts along time. These scanners show intent to acquire knowledge on a specific target and update this knowledge along time. Their occurrence can alert administrators that their network is under scrutiny from an attacker. Publicly documented scanners behave differently from other scanners. For example, Telnet scanning is recently on the rise but documented scanners only marginally probe Telnet. Furthermore, they mainly perform spread random scans. This further shows that our profiling method is efficient

to discriminate scanners' behavior. Finally, only 39% of these scanners follow online documentation best practices.

ACKNOWLEDGMENT

This research has been supported by the Japanese Society for Promotion of Science under grant 15H02699.

REFERENCES

[1] G. F. Lyon, "Remote os detection via TCP/IP stack fingerprinting," in *Phrack*, vol. 8, no. 54.

[2] Y. Vanaubel, J.-J. Pansiot, P. Mérindol, and B. Donnet, "Network fingerprinting: TTL-based router signatures," in *Proc. of IMC'13*, 2013, pp. 369–376.

[3] Z. Durumeric, J. Kasten, M. Bailey, and J. A. Halderman, "Analysis of the HTTPS certificate ecosystem," in *Proc. of IMC'13*, 2013, pp. 291–304.

[4] 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 *Proc. of IMC'14*, 2014, pp. 435–448.

[5] Z. Durumeric, J. Kasten, D. Adrian, J. A. Halderman, M. Bailey, F. Li, N. Weaver, J. Amann, J. Beekman, M. Payer, and V. Paxson, "The matter of Heartbleed," in *Proc. of IMC'14*, 2014, pp. 475–488.

[6] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast internet-wide scanning and its security applications," in *Proc. of USENIX Security'13*, 2013, pp. 605–620.

[7] R. D. Graham, "MASSCAN: Mass ip port scanner," https://github.com/robertdavidgraham/masscan, accessed: 2017-01-31.

[8] M. Allman, V. Paxson, and J. Terrell, "A brief history of scanning," in *Proc. of IMC'07*, 2007, pp. 77–82.

[9] V. Yegneswaran, P. Barford, and J. Ullrich, "Internet intrusions: Global characteristics and prevalence," in *Proc. of SIGMETRICS'03*, 2003, pp. 138–147.

[10] A. Wahid, C. Leckie, and C. Zhou, "Characterising the evolution in scanning activity of suspicious hosts," in *Proc. of NSS'09*, 2009, pp. 344–350.

[11] R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda, "MAWILab: combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking," in *Proc. of CoNEXT'10*, 2010, pp. 1–12.

[12] J. Mazel, R. Fontugne, and K. Fukuda, "Taxonomy of anomalies in backbone network traffic," in *Proc. of TRAC'14*, 2014, pp. 30–36.

[13] E. Glatz and X. Dimitropoulos, "Classifying internet one-way traffic," in *Proc. of IMC'12*, 2012, pp. 37–50.

[14] N. Brownlee, "One-way traffic monitoring with iatmon," in *Proc. of PAM 2012*, 2012, pp. 179–188.

[15] Z. Durumeric, M. Bailey, and J. A. Halderman, "An internet-wide view of internet-wide scanning," in *Proc. of USENIX Security'14*, 2014, pp. 65–78.

[16] M. Bailey, A. Burstein, K. Claffy, S. Clayman, D. Dittrich, J. Heidemann, E. Kenneally, D. Maughan, J. McNeill, P. Neumann, C. Scheper, L. Tien, C. Papadopoulos, W. Visscher, and J. Westby, "The menlo report: Ethical principles guiding information and communication technology research," *US Department of Homeland Security*, 2011.

[17] C. Partridge and M. Allman, "Addressing ethical considerations in network measurement papers," in *Proc. of NS Ethics'15*, 2015, pp. 33–33.

[18] "ZMap documentation - scanning best practices," https://zmap.io/documentation.html#bestpractices, accessed: 2017-01-31.

[19] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan, "Fast portscan detection using sequential hypothesis testing," in *Proc. of SP 2004*, 2004, pp. 211–225.

[20] M. Alsaleh and P. C. van Oorschot, "Network scan detection with LQS: A lightweight, quick and stateful algorithm," in *Proc. of AsiaCCS'11*, 2011, pp. 102–113.

[21] M. H. Bhuyan, D. K. Bhattacharyya, and J. Kalita, "Surveying port scans and their detection methodologies," *Computer Journal*, vol. 54, no. 10, pp. 1565–1581, 2011.

[22] D. Leonard, Z. Yao, X. Wang, and D. Loguinov, "Stochastic analysis of horizontal IP scanning," in *Proc. of INFOCOM'12*, 2012, pp. 2077–2085.

[23] A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescapè, "Analysis of a /0 stealth scan from a botnet," *Transactions on Networking*, vol. 23, no. 2, pp. 341–354, 2015.

[24] M. Hoffman, "Legal considerations for widespread scanning," https://community.rapid7.com/community/infosec/sonar/blog/2013/10/30/legal-considerations-for-widespread-scanning, accessed: 2017-01-31.

[25] "MAWI," http://mawi.wide.ad.jp/mawi/, accessed: 2017-01-31.

[26] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman, "A search engine backed by Internet-wide scanning," in *Proc. of CCS'15*, 2015, pp. 542–553.

[27] S. Mongkolluksamee, K. Fukuda, and P. Pongpaibool, "Counting NATted hosts by observing TCP/IP field behaviors," in *Proc. of ICC'12*, 2012, pp. 1265–1270.

[28] "SANS internet storm center," https://isc.sans.edu/feeds/daily_sources.

[29] Y. M. P. Pa, S. Suzuki, K. Yoshioka, T. Matsumoto, T. Kasama, and C. Rossow, "IoTPOT: Analysing the rise of IoT compromises," in *Proc. of WOOT'15*, 2015.

[30] D. Leonard and D. Loguinov, "Demystifying internet-wide service discovery," *Transactions on Networking*, vol. 21, no. 6, pp. 1760–1773, 2013.

[31] N. Falliere, "A distributed cracker for VoIP," http://www.symantec.com/connect/blogs/distributed-cracker-voip, 2011, accessed: 2017-01-31.

[32] Z. Li, A. Goyal, Y. Chen, and V. Paxson, "Towards situational awareness of large-scale botnet probing events," *Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 175–188, 2011.

[33] E. Bou-Harb, M. Debbabi, and C. Assi, "On fingerprinting probing activities," *Computers & Security*, vol. 43, pp. 35 – 48, 2014.

[34] K. Fukuda and R. Fontugne, "Estimating speed of scanning activities with a hough transform," in *Proc. of ICC'10*, 2010, pp. 1–5.

[35] ErrataSecurity, http://blog.erratasec.com/2011/10/scanning-internet.html.

[36] ——, http://blog.erratasec.com/2013/07/scanning-internet.html.

[37] ——, http://blog.erratasec.com/2013/09/we-scanned-internet-for-port-22.html.

[38] ——, http://blog.erratasec.com/2016/05/doing-full-scan-of-internet-right-now.html.