Multidimensional Investigation of Source Port 0 Probing

By

Elias Bou-Harb, Nour-Eddine Lakhdari, Hamad Binsalleeh and Mourad Debbabi

From the proceedings of

The Digital Forensic Research Conference

DFRWS 2014 USA

Denver, CO (Aug 3rd - 6th)

Digital Investigation 11 (2014) S114–S123

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Multidimensional investigation of source port 0 probing

Elias Bou-Harb*, Nour-Eddine Lakhdari, Hamad Binsalleeh, Mourad Debbabi

The National Cyber-Forensics and Training Alliance (NCFTA) Canada, Concordia University, Concordia Institute for Information Systems Engineering, Montreal, Quebec, Canada H3G1M8

Keywords:
Port 0 probing
Darknet
Malware
Passive DNS
Data correlation

A B S T R A C T

During November 2013, the operational cyber/network security community reported an unprecedented increase of traffic originating from source port 0. This event was deemed as malicious although its core aim and mechanism were obscured. This paper investigates that event using a multifaceted approach that leverages three real network security feeds that we receive on a daily basis, namely, darknet, passive DNS and malware data. The goal is to analyze such event from the perspectives of those feeds in order to generate significant insights and inferences that could contribute to disclosing the inner details of that incident. The approach extracts and subsequently fingerprints such malicious traffic from the received darknet data. By executing unsupervised machine learning techniques on the extracted traffic, we disclose clusters of activities that share similar machinery. Further, by employing a set of statistical-based behavioral analytics, we capture the mechanisms of those clusters, including their strategies, techniques and nature. We consequently correlate the sources with passive DNS in order to investigate their maliciousness. Moreover, to examine if the sources are malware contaminated, we execute a correlation mechanism between the darknet data and the malware feeds. The outcome reveals that such traffic indeed is reconnaissance/probing activities originating from three different horizontal scans utilizing packets with a TCP header length of 0 or packets with odd flag combinations. The results as well demonstrate that 28% of the scanning sources host malicious/blacklisted domains as they are often used for spamming, phishing and other fraud activities. Additionally, the outcome portrays that the bot probing sources are infected by 'Virus.Win32.Sality'. By correlating various evidence, we confirm that such malware specimen is in fact responsible for part of the source port 0 probing event. We concur that this work is a first attempt ever to comprehend the machinery of such unique event and we hope that the community could consider it as a building block for auxiliary analysis and investigation.

Introduction

Probing is often defined as the task of scanning enterprise networks or Internet wide services in an attempt to search for vulnerabilities or ways to infiltrate IT assets. It is typically considered a significant cyber security concern due to the fact that it is commonly the primary stage of an intrusion attempt that enables an attacker to remotely locate, target, and subsequently exploit vulnerable systems. For instance, hackers have employed probing techniques to identify numerous misconfigured proxy servers at the New York Times to access a sensitive database that disclosed more than 3000 social security numbers (New York Times internal network hacked). Further, the United States Computer Emergency Readiness Team (US-CERT) revealed that attackers had performed coordinated probing activities to

* Corresponding author. Tel.: +1 5146495049.
E-mail address: e_bouh@encs.concordia.ca (E. Bou-Harb).

http://dx.doi.org/10.1016/j.diin.2014.05.012
1742-2876/

fingerprint WordPress sites and consequently launched their targeted attacks (WordPress sites targeted). Moreover, it was disclosed that hackers had leveraged sophisticated scanning events to orchestrate multiple breaches of Sony's PlayStation Network taking it offline for 24 days and costing the company an estimated $171 million (PlayStation network outage caused by 'external intrusion'). More alarming, a recent incident reported that attackers had escalated a series of "surveillance missions" against cyber-physical infrastructure operating various US energy firms that permitted the hackers to infiltrate the control-system software and subsequently manipulate oil and gas pipelines (Iran hacks energy firms). Although, on one hand, Panjwani et al. (2005) concluded that a momentous 50% of attacks against cyber systems are preceded by some form of probing activity, however, on the other hand, such observed activities might simply reflect the "background radiation" of various Internet-scale random scanning activities, remnants of past worm/virus outbreaks, or other malware activities on the global Internet at large (Moore et al., 2004). Indeed, it is known in the cyber security and digital forensics communities that it is a daunting task to infer and uncover the intention, mechanism and the nature of the perceived probing activities (Jin et al., 2007).

Background

On November 2nd, 2013, security researchers at Cisco Systems reported that their worldwide deployed sensors have detected a massive increase in TCP source port 0 traffic.[1] They further elaborated that the magnitude observed by the sensors was elevated by 20 times than typical traffic originating from the same port and transport protocol on other days. According to the researchers, such event renders the largest spike in network traffic originating from TCP source port 0 in the last decade. In a follow-up discussion,[2] the researchers noted that such port, according to its RFC, is engineered to be reserved, and that such traffic could be used to fingerprint various operating systems. Additionally, the security researchers speculated about the aim, mechanism and source of that traffic by stating that such rare event could be some sort of a research project, a malware infected probing botnet, a targeted reconnaissance event aiming to launch an immediate or a prolonged malicious task, or even a broken embedded device or a new piece of malware with a bug in its scanning code.

The event was interestingly also observed by DShield/Internet Storm Center (ISC).[3] ISC data comprises of millions of intrusion detection log entries gathered daily from sensors covering more than 500 thousand Internet Protocol (IP) addresses in over 50 countries. As shown in Fig. 1, the event was apparent on four days, namely, November 2nd,[4] November 21st, November 24th and November 25th, 2013. The latter fact is particularly demonstrated by the peaks of the TCP ratio of port 0 on those specific days.

[1] http://tinyurl.com/pds443n.
[2] http://tinyurl.com/n8j58hs.
[3] http://www.dshield.org/port.html.
[4] Coinciding with Cisco reports although not quite as significant.

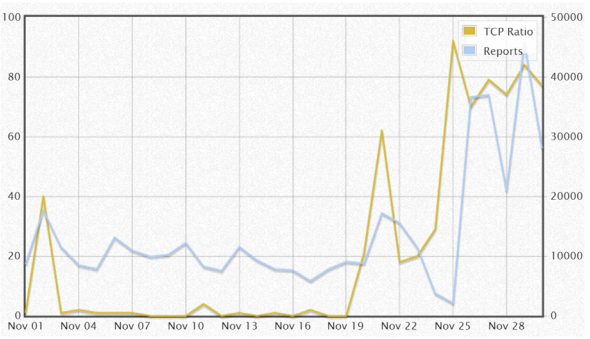

Fig. 1. The source port 0 event as observed by DShield/Internet Storm Center.

Contributions

Motivated by the requirement to shed the light on that incident in order to generate inferences and insights that could contribute in disclosing the inner details of such an unprecedented event, we frame the paper's contributions as follows:

- Proposing a multifaceted approach that leverages three real network security feeds. The approach exploits darknet data (i.e., Internet traffic destined to half a million routable yet unallocated IP addresses) to extract, analyze and uncover the machinery of such traffic. Further, the approach employs correlation between the latter and passive DNS (i.e., Internet-wide authoritative DNS responses) to study the maliciousness of the such traffic. Moreover, the proposed approach correlates darknet-extracted traffic with malware feeds to answer questions related to contamination and attribution. To the best of our knowledge, 1) the proposed approach that correlates those three feeds in an effort to understand a cyber event has never been attempted before and 2) the yielded outcome from adopting such an approach related to this specific event is unique in the literature.
- Employing 1) machine learning data clustering techniques to partition the port 0 traffic according to similar machinery and 2) a set of novel behavioral analytics that scrutinize such traffic to capture the behavior of the sources.
- Evaluating the proposed approach using 30 GB of real darknet traffic, 1.4 billion DNS records and 30 million malware analysis reports.

Organization

The remaining of this paper is organized as follows. In the following section, we present the proposed approach. Specifically, we elaborate on how source port 0 traffic is extracted and fingerprinted from darknet traffic in addition to presenting the analytics that are used to disclose the machinery of such traffic. Further, we discuss the goals and

Table 1
Features description.

Features		
Data link layer	1	Delta time with previous packet
	2	Packet length
	3	Frame length
	4	Capture length
	5	The flag 'frame' is marked
Network layer	6	IP header length
	7	IP flags
	8	IP flags: reversed bit
	9	IP flags: do not fragment bit
	10	IP flags: more fragments bit
	11	IP fragment offset
	12	IP time to live
	13	IP protocol
Transport layer	14	TCP segment length
	15	TCP sequence number
	16	TCP next sequence number
	17	TCP acknowledgment number
	18	TCP header length
	19	TCP flags
	20	TCP flags: congestion window
	21	TCP flags: ECN-echo
	22	TCP flags: urgent
	23	TCP flags: acknowledgment
	24	TCP flags: push
	25	TCP flags: reset
	26	TCP flags: syn
	27	TCP flags: fin
	28	TCP window size
	29	UDP length

the correlation mechanisms with passive DNS and malware data. In Section 3, we empirically evaluate the proposed approach, which yields numerous insights related to the source port 0 event, including, the traffic's machinery, the maliciousness of the sources and their corresponding malware attribution. In Section 4, we review the related work on various concerned topics. Finally, concluding remarks and future work are stated in Section 5.

Proposed approach

In this section, we present the proposed approach that is composed of three mechanisms, namely, darknet analysis, passive DNS correlation and malware correlation.

Darknet analysis

We possess real darknet data that we are receiving on a daily basis from a trusted third party. The data is around 12 GB per day. Such traffic originates from the Internet and is destined to numerous network sensors. The data mostly consists of unsolicited TCP, UDP and ICMP traffic. It might contain as well some DNS traffic. In a nutshell, darknet traffic is Internet traffic destined to routable but unused Internet addresses (i.e., dark sensors). Since these addresses are unallocated, any traffic targeting them is deemed as suspicious. Darknet traffic is typically composed of three types of traffic, namely, probing, backscattered and misconfiguration. Probing arises from bots, worms and tools (or binaries) while backscattered traffic commonly refers to unsolicited traffic that is the result of responses to Denial of Service (i.e., DoS) attacks with spoofed source IP addresses. On the other hand, misconfiguration traffic is due to network/routing or hardware/software faults causing such traffic to be sent to the darknet sensors. Darknet analysis has shown to be an effective method to generate Internet-scale cyber threat intelligence (Michael et al., 2005).

Although the monitored darknet space is relatively large (i.e., /13), we were unable to notice the existence of the source port 0 event on November 2nd or November 21st. However, we were able to retrieve around 30 GB of darknet data that encompasses the event from November 24th and 25th. We base our darknet analysis approach on such data.

Traffic extraction

In order to retrieve the packets of the source port 0 event, we created a simplistic TCPdump filter[5] that captures any darknet traffic that is utilizing TCP as the transport protocol and 0 as the source port. We applied the filter upon the 30 GB darknet data. We further refined the output by filtering out any darknet misconfiguration that could exist. To accomplish this, we adopt a metric that records the average number of sources per destination darknet address. This metric should be significantly larger for misconfiguration than probing traffic. However, although it differentiates misconfiguration from scanning, it could include as well backscattered traffic as they also can possess a large average number of sources per destination (i.e, in case of a DoS). To cope with this issue, we observe, per the technique in Wustrow et al. (2010), flags in packet headers, such as TCP SYN + ACK, RST, RST + ACK, ACK, etc., that resemble backscattered traffic (Wustrow et al., 2010). Subsequently, we filter out flows that lack that observation, deeming them as misconfiguration. The remaining output is rendered as the generated traffic from the source port 0 event, which is saved in a packet capture (i.e., pcap) format for further analysis.

Traffic fingerprinting

To identify the nature of the source port 0 event, we implemented the technique from Bou-Harb et al. (2014). The latter approach specifically operates on darknet data and possesses the capability to distinguish between probing and DoS sessions. To accomplish this, the technique leverages the detrended fluctuation analysis statistical method and assigns a certain unique correlation value depending on the nature of each session. Readers who are interested in the inner details of such technique are kindly referred to Bou-Harb et al. (2014). By subjecting the source port 0 event traffic to the technique, the outcome revealed that 97% of the sessions are related to probing activities. We manually inspected the remaining 3%, which demonstrated that they are misconfiguration traffic that apparently, were not previously filtered as expected. We also confirmed such probing results by exposing the source port 0 event traffic to Snort's[6] probing engine, the sfPortscan pre-processor,[7] which yielded a similar result.

[5] http://www.danielmiessler.com/study/tcpdump/.
[6] http://www.snort.org/.
[7] http://manual.snort.org/node78.html.

2.1.3. Traffic clustering

For the purpose of disclosing the inner mechanisms of the source port 0 event, we refer to machine learning techniques. Such techniques allow us to efficiently, effectively and automatically uncover clusters of activities sharing similar machinery within the global event, without relying on strenuous manual analysis.

When the data observations are not pre-labeled into defined numerical or categorical classes, as in our case, two standard widely deployed algorithms for data clustering using unsupervised learning could be employed. These are the *k*-means (MacQueen et al., 1967) and the EM (Dempster et al., 1977) algorithms. On one hand, the *k*-means algorithm finds *k* clusters by choosing *n* data points at random as initial cluster centers. Each data point is then assigned to the cluster with the center that is closest to that point. Each cluster center is then replaced by the mean of all the data points that have been assigned to that cluster. Note that, the *k*-means algorithm operates by minimizing the sum of squared Euclidean distances between data records in a cluster and the clusters mean vector. This process is iterated until no data point is reassigned to a different cluster. On the other hand, the EM algorithm views the data clustering problem as a framework for data density estimation using a probability density function. An effective representation of the probability density function is the *mixture model*, which asserts that the data is a combination of *k* individual component densities corresponding to *k* clusters. The EM problem can be summarized as follows: given a set of data observations, identify a set of *k* populations in the data and provide a density distribution model of each of the populations. Readers who are interested in the details of the EM are kindly referred to Dempster et al. (1977).

We proceed by going back to the source port 0 event traffic pcap file that we have previously isolated. Subsequently, we extracted from it a total of 29 data link, network and transport layer packet features as summarized in Table 1. The latter features have been shown to produce distinguishing characteristics when applied on network data (Alshammari and Nur Zincir-Heywood, 2011). This feature extraction procedure was achieved using the open source jNetPcap API.[8] We consequently compiled the extracted features into a unified data file and applied the *k*-means and the EM algorithms, leveraging MATLAB's default clustering functionality and the WEKA data mining tool,[9] respectively.

Behavioral analytics

In an attempt to capture the machinery of the probing sources/clusters, we present the following set of novel behavioral analytics. Such proposed approach takes as input the previously extracted probing sessions (recall Section 2.1.2) and outputs a series of behavioral characteristics related to the probing sources. In what follows, we pinpoint the concerned questions and subsequently present the undertaken approach in an attempt to answer those.

Is the probing traffic random or does it follow a certain pattern? When sources generate their probing traffic, it is significant to capture the fashion in which they accomplish that. To achieve this task, we proceed as follows. For each distinct pair of hosts retrieved from the probing sessions (probing source to target), we test for randomness in the generated traffic using the non-parametric Wald–Wolfowitz statistic test. If the result is positive, we record it for that specific probing source and apply the test for the remaining probing sessions. If the outcome is negative, we infer that the generated traffic follows a certain pattern. To capture the specific employed pattern, we model the probing traffic as a Poisson process and retrieve the maximum likelihood estimate intervals (at a 95% confidence level) for the Poisson parameter λ that corresponds to that traffic. The choice to model the traffic as a Poisson distribution is motivated by Li et al. (2011), where the authors observed that probe arrivals is coherent with that distribution. After the test has executed for all the probing sources, we apply the CLUstEring based on local Shrinking (CLUES) algorithm on the generated patterns. CLUES allows non-parametric clustering without having to select an initial number of clusters. The outcome of that operation is a set of specific λ intervals. The aim of this is to map each probing source that was shown to employ a pattern to a certain λ interval by removing overlapping values that could have existed within the initially generated λ intervals.

How are the targets being probed? As revealed in Dainotti et al. (2012), coordinated probing sources employ various strategies when probing their targets. These strategies could include IP-sequential, reverse IP-sequential, uniform permutation or other types of permutations. In an attempt to capture the probing strategies, we execute the following. For each probing source, we extract its corresponding distribution of target IPs. To differentiate between sequential and permutation probing, we apply the Mann–Kendall statistic test, a non-parametric hypothesis testing approach, to check for monotonicity in those distributions. The rational behind the monotonicity test is that sequential probing will indeed induce a monotonic signal in the distribution of target IPs while permutation probing will not. Further, in this work, we set the significance level to 0.5% since a higher value could introduce false positives. To differentiate between (forward) IP-sequential and reverse IP-sequential, for those distributions that tested positive for monotonicity, we also record the slope of the distribution; a positive slope defines a forward IP-sequential strategy while a negative one renders a reverse IP-sequential strategy. For those distributions that tested negative for monotonicity (i.e., not a sequential strategy), we leverage the chi-square goodness-of-fit statistic test. The latter insight will inform us whether or not the employed strategy is a uniform permutation; if the test fails, then the employed strategy will be deemed as a permutation; uniform permutation otherwise.

What is the nature of the probing source? It is significant as well to infer the nature of the probing source; is

[8] http://jnetpcap.com/.
[9] http://www.cs.waikato.ac.nz/ml/weka/.

it a probing tool or a probing bot. From the two previous questions, we can infer those probing events that are random and monotonic. It is known that monotonic probing is a behavior of probing tools in which the latter sequentially scan their targets (IPs and ports). Furthermore, for random events, the monotonic trend checking can help filter out traffic caused by the non-bot scanners (Li et al., 2011). Thus, we deem a probing source as leveraging a probing tool if their traffic is randomly generated and if they adopt a sequential probing strategy (i.e., including reverse IP-sequential); a bot otherwise.

Is the probing targeted or dispersed? When sources probe their targets, it would be interesting to infer whether their probing traffic is targeted toward a small set of IPs or dispersed. To answer this, for each probing source b, we denote GF(b) as the collection of flows generated by that specific source that target the dark space. The destination target IPs used by the flows in GF(b) induce an empirical distribution. Subsequently, we borrow the concept of relative uncertainty, an information theoretical metric and apply it on those distributions. The latter index is a decisive metric of variety, randomness or uniformity in a distribution, regardless of the sample size. An outcome that is close to 0 defines that the probing source is using a targeted approach while an outcome value close to 1 means that its corresponding probing traffic is dispersed.

It is evident that the latter set of behavioral analytics significantly depend on numerous statistical tests and methods to capture the behavior of the probing sources. We assert that such approach is arguably more sound than heuristics or randomly set thresholds. It is also worthy to mention that all the employed statistical tests assume that the data is drawn from the same distribution. Since the approach operates on one type of data, namely, darknet data, we can safely presume that the values follow and are in fact drawn from the same distribution.

Passive DNS correlation

We are also receiving on a daily basis around 1.3 million Domain Name System (DNS) messages from Farsight Security Inc. (Security Information Exchange). Such data is collected by observing DNS traffic between recursive DNS resolvers on the Internet. The latter is often dubbed as passive DNS data, which constitutes the successful translations and associations between domains and IP addresses. Passive DNS analysis has shown to be an effective approach to generate cyber threat intelligence (Bilge et al., 2011). We amalgamate a database that contains the last three-year period of such traffic ($\approx$1.4 billion records) in order to investigate the source port 0 probing event.

The rationale of employing passive DNS correlation is rendered by the requirement to contribute in investigating and perhaps attributing the probing sources of such an unprecedented event to certain malicious entities (i.e., malicious domains, for instance). Particularly, we aim to extract the following information about the suspicious IP addresses (previously retrieved from darknet analysis) that have participated in the probing event.

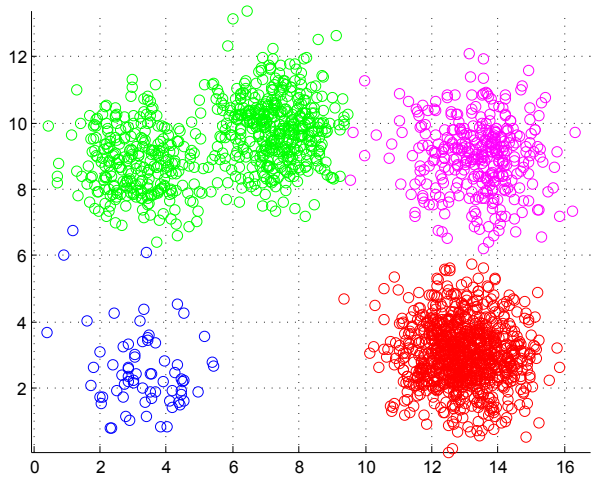

Fig. 2. Port 0 event traffic clusters.

- Hosting capability: Malicious entities typically host a significant number of services and malicious domains on a limited number of IP addresses for cost-effectiveness reasons. Thus, by inferring domains that are resolved to a specific IP address, we can pinpoint and analyze these hosted domains to reveal the level of maliciousness of that IP address.
- Intensity: By computing the number of DNS messages that utilize a certain IP address, we can deduce the levels of accessibility and involvement of that IP address in malicious activities.
- Aliveness: By recording the first and last timestamp that a specific IP address has been observed in DNS traffic (i.e., resolved to certain domain(s)), we can infer the participation period and aim of that IP address. Further, we can refine the analyzed passive DNS interval for the purpose of investigating and attributing that IP address with a certain cyber event.

Malware correlation

We possess as well dynamic malware analysis reports (i.e., XML reports) of malware binaries for the last four years. We are receiving such feed on a daily basis with an average of 30 thousand XML malware reports from ThreatTrack Security[10]. Up to the event date in November 2013, we have accumulated more than 30 million malware analysis reports since January 2010. The XML reports are produced by analyzing the malware binaries in a controlled environment. Each XML report corresponds to only one malware sample. It is worthy to mention that these reports contain the executed activities by the malware samples at the network and system levels. On one hand, the network level activities refer to the connections and the exchanged packets, including IP addresses, port numbers, urls, visited domains and the actual payload data that has been sent. On

[10] http://www.threattracksecurity.com/.

the other hand, the system level activities constitute the list of Dynamic-link Library (DLL) files that are utilized by the malware, the key registry changes, and the memory usage. We leverage such XML reports to investigate the source port 0 probing event. Specifically, we attempt to answer the following two questions by presenting their corresponding approaches:

Which malware has infected the probing machines before or during the occurrence of the event? In order to infer which malware has infected the probing machines, we present Algorithm 1. Simplistically, the Algorithm parses the XML reports mining for those malware that utilize the probing machine IP addresses (previously retrieved from darknet analysis) as per the destination IP address criterion. The Algorithm further refines the output by only considering those malware that connect to the probing IPs in November 2013.

Algorithm 1. Extracting malware samples that have infected the probing machines.

```
Input: Dynamic malware analysis reports: XMLs;
List of the scanning source IPs: ProbingIPs;
Output: List of malware samples that infected the
probing machines: MalwareList
for xml in XMLs do
    if xml.getMalware().getdestIP() in ProbingIPs then
        if xml.getMalware().getTime()==Nov, 2013 then
            MalwareList.add(xml.getMalware().getName());
        end
    end
end
```

Which malware generated the probing traffic? In an attempt to attribute the probing machines to a certain malware, we perform the following. We filter the entire set of malware XML reports by focusing on those samples that execute traffic from TCP source port 0. We subsequently match the outcome from the latter procedure with the list of malware that infected the probing machines that was derived from Algorithm 1. The rational behind this approach states that if a certain machine has been infected by a specific malware sample, in which it was derived that such machine is generating TCP source port 0 traffic, then it is highly probable that this specific malware is causing such traffic.

Empirical evaluation

This section abides with the proposed approach that was previously discussed to disclose various inferences from the perspective of the three data feeds.

Darknet inferences

The source port 0 event traffic was rendered by more than 1 million probing packets originating from TCP source port 0 destined to the monitored darknet space. It is significant to note that we typically observe, on other days, less than 1000 packets per darknet day originating from TCP source port 0. The traffic originates from 27 unique hosts, 17 distinct countries, 24 diverse Internet Service Providers (ISPs) and from within 25 operational organizations. We refrain from publishing statistics and rigorous information related to the latter due to sensitivity and legal issues. We next investigated some characteristics related to those packets. We noticed that the Time to Live (TTL) values of the packets change with the source IP addresses. This advocates that IP spoofing is less likely or non-existent (Templeton and Levitt, 2003). The fact that the source IP addresses are not spoofed corroborates that such packets are indeed related to scanning/probing activities (so the actual scanner can essentially receive back the probing results), as it was inferred in Section 2.1.2. We also noticed that the packets arrival rate is slow, averaging around 3 packets per second. This is compliant with slow scanning activities, which are known to be difficult to be detected (Jaeyeon et al., 2004). Upon a closer investigation of the packet headers, we observed that the majority of the packets either include a TCP header length of 0 or are malformed. Further, almost all the packets contain odd flag combinations (i.e., FIN, SYN, RST, PSH, ACK, FIN, URG, PSH, Reserved). Typically, probing by employing the latter flags is considered as performing 'stealthy' scanning activities as they are engineered to evade firewall detection by only sending a single frame to a TCP port without any TCP handshaking or any additional packet transfers (Bou-Harb et al., 2013).

We proceed by executing the clustering approach in coherence with Section 2.1.3. Recall, that the aim is to disclose traffic clusters that share similar machinery. The output of the EM algorithm is depicted in Fig. 2; we omit the output of the *k*-means since it revealed a similar result.

Such outcome provides evidence that the traffic originates from 4 different classes. To further test the validity of this result, we produced a silhouette graph of the EM clusters as shown in Fig. 3. Commonly, a silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. A value of 1 indicates that the points are very distant from neighboring clusters, a value of 0 informs that the points are not distant from other clusters while a negative value indicates that the points are erroneously placed in that cluster. From Fig. 3, it is shown that a significant amount of points in all the 4 classes have a large silhouette value, greater than 0.6 (Kaufman and Rousseeuw, 2009), indicating that the clusters are separated from neighboring clusters. This provides incentives to validate the quality of the formed EM clusters. By closely investigating each cluster, we determined that the first cluster represents a horizontal scan focused on destination port 0 from a single IP address located in Germany targeting around 800 thousand unique destination addresses. The use of destination port 0 is a frequently employed technique by scanners to fingerprint the operating systems of the targeted destinations in order to tailor future attacks based on that retrieved information. Further, the second cluster discloses a probing campaign targeting more than 60 thousand destination ports and originating from a single Dutch IP address. The latter insight was also observed and confirmed by Cisco.[11] The third cluster renders another

[11] http://tinyurl.com/kndnj82.

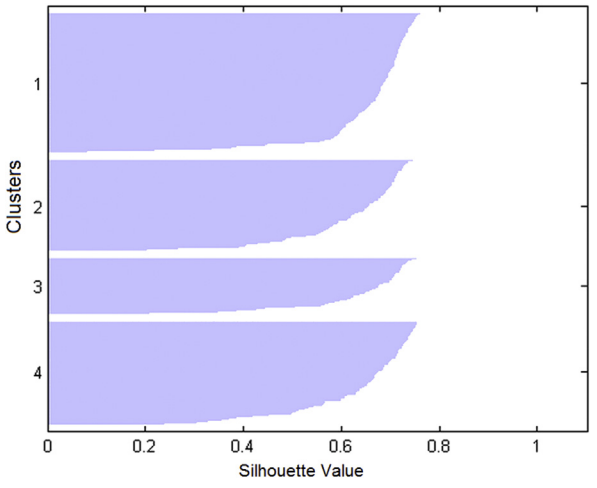

Fig. 3. A silhouette plot of the EM clusters.

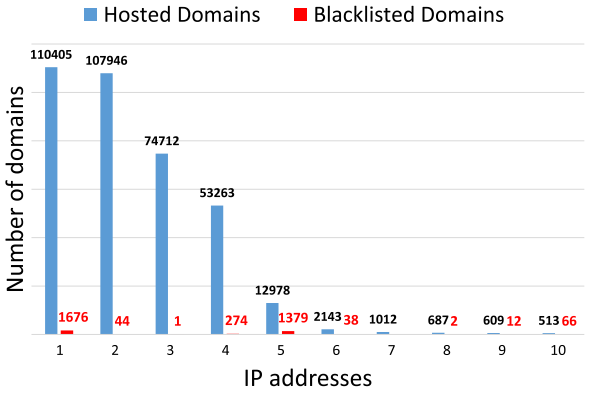

Fig. 4. Hosted and blacklisted domains of the probing sources.

horizontal scan that specifically targeted three destination ports, namely, TCP ports 445, 22 and 3389, which respectively represent the Microsoft directory, the secure shell (i.e., ssh) and the remote desktop protocol services. The latter are known to suffer from various vulnerabilities and are often exploited.[12] It is worthy to mention that this horizontal scan targeted 9 thousand destinations on port 445, 7 thousand destinations on port 22 and around 5.5 thousand destinations on port 3389. Recall, that all the probing activities in the three previous clusters originate from TCP source port 0. The last minor cluster captured darknet misconfiguration traffic advocating the obtained result of Section 2.1.2.

To further investigate the mechanisms of the probing sources, we invoked the behavioral analytics that were presented in Section 2.1.4. It was revealed that 62% of the probing sources used certain patterns when generating their probing traffic. Concerning the employed probing strategy, it is shown that 57% of the probing sources leveraged a permutation while the remaining adopted a

[12] http://tinyurl.com/kkfs6pq,http://tinyurl.com/m5684jo.

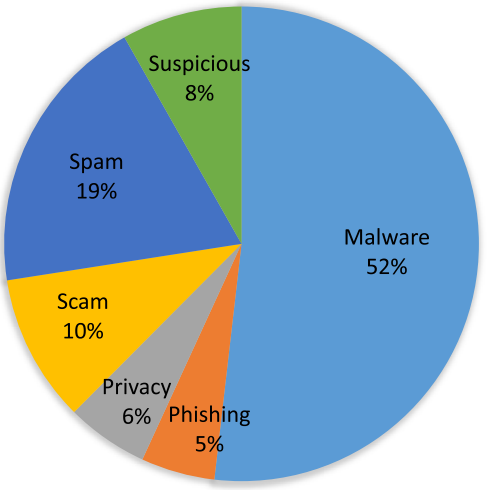

Fig. 5. The nature of the blacklisted domains.

sequential strategy when probing their targets. Of those that employed a permutation, 76% used a uniform permutation while 24% adopted other types of permutations. The majority (≈98%) of those that employed a sequential strategy were found to adopt a forward IP-sequential strategy while only 2% adopted a reverse IP-sequential strategy. It is noteworthy to mention that in Leonard and Loguinov (2010), the researchers dismissed the possible use of this strategy since, as they noted, the strategy is difficult to be used to extrapolate certain metrics from especially when dealing with partial probes. Further, the analytics disclosed that ≈55% of the sources were probes from bots while the remaining were generated from probing tools. Moreover, it was inferred that all the probing sources were generating probing that is dispersed as opposed to targeting a small set of IPs. To the best of our knowledge, the previously generated inferences represent the first comprehensive empirical results of probing behaviors. In the context of the probing clusters that were disclosed in the previous section, it was shown that clusters

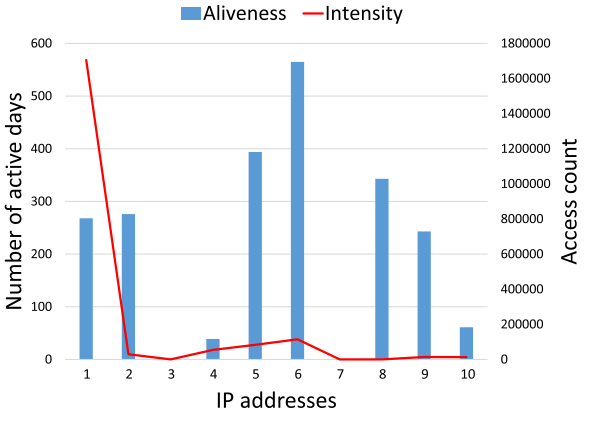

Fig. 6. Investigating the aliveness and access count of the malicious domains.

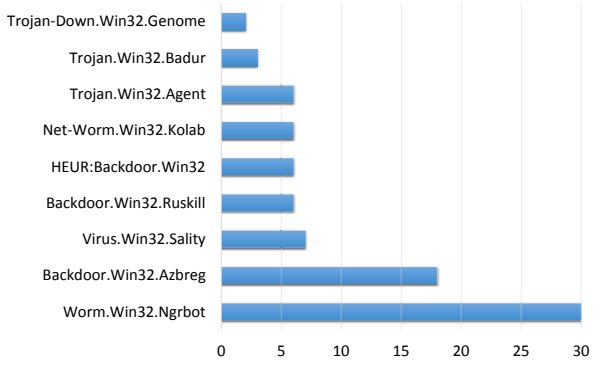

Fig. 7. Malware and their corresponding number of connections.

one and two were generating random probing traffic as opposed to using a certain pattern, employed a sequential strategy when probing their targets and were found to be leveraging a probing tool. The latter insight proposes that such probing activities have been executed by the same 'initiator/author' since they are adopting the same mechanism and probing characteristics, which could reveal that they both intended to achieve the same desired goal. Conversely, in the third cluster, the majority ($\approx$92%) were found to adopt certain patterns when generating their probing traffic, employed a permutation when probing their targets and were inferred to be bots. This suggests that such probing activities could be malware-orchestrated (Dainotti et al., 2012) and thus there is a momentous need to explore and investigate their malevolence.

Passive DNS inferences

We employ the approach of Section 2.2 to investigate the maliciousness of all the probing sources. Fig. 4 reveals the top 10 probing IP addresses that were shown to host (i.e., resolve to) the most domains. It could be inferred that the most significant number of resolved domains ranges from around 13 thousand domains to peaking at around 110 thousand domains per the probing IPs. We also investigated a subset of those domains that are related to malicious activities. To achieve this task, we correlated the extracted domains with publically available datasets and resources, namely, the Malware Domain List,[13] Zeus Tracker[14] and McAfee's siteAdvisor.[15] The outcome is also depicted in Fig. 4, which reveals the number of blacklisted domains. The fact that 28% of the probing IP addresses were shown to host numerous blacklisted/malicious domains provides an alarming signal that the source port 0 probing traffic could be originating from a malicious entity with malevolent goals. To identify the nature of those blacklisted domains, we leveraged the Web of Trust reputation system.[16] The outcome from such a procedure is depicted in Fig. 5. The results demonstrate that more than half of the malicious domains could be attributed to malware; this advocates our decision to leverage malware data to investigate the event in question. It was also shown that those malicious domains are blacklisted as they are often used for spamming, phishing and other fraud activities. We further investigated those malicious domains by analyzing their aliveness and intensity as discussed in Section 2.2 and exposed in Fig. 6. One can notice, on one hand, that some IPs with their corresponding blacklisted resolved domains retain less active days but possess high accessibility, which render them extremely effective in their maliciousness. On the other hand, some domains have a prolonged online presence yet possess low access counts. We envision that such malicious domains are intentionally not intended to be accessed but are rather playing a hosting or a supporting (i.e., back-end) role for other malicious tasks and services.

Table 2
Malware samples generating TCP source port 0 traffic.

Email-Worm.Win32.Mydoom
Worm.Win32.AutoRun
Virus.Win32.Sality
Virus.Win32.Expiro
Backdoor.Win32.Xtoober
Trojan-Downloader.Win32.Agent
Trojan-Dropper.Win32.Small
Trojan.Win32.Pincav
Trojan.Win32.Jorik
Trojan-Downloader.Win32.Delf
Trojan-Downloader.Win32.Genome
Backdoor.Win32.Gbot
Backdoor.Win32.Popwin
Email-Worm.Win32.Rays
Email-Worm.Win32.Runouce
Packed.JS.Agent
Trojan-Banker.Win32.Banker
Trojan-Downloader.Win32.FlyStudio
Backdoor.Win32.Banito
Backdoor.Win32.VB
HackTool.Win32.Injecter

Malware inferences

Motivated by the fact that the sources of the 3rd probing cluster, as revealed in Section 3.1, were inferred to be bots coupled with the conclusion that more than half of the probing sources resolve to malware-infected domains as demonstrated in Fig. 5, in this section, we investigate the source port 0 probing event from the malware feed perspective, in accordance with the proposed approach of Section 2.3.

Fig. 7 depicts the malware samples and their corresponding number of connections to the probing machines. It could be noted that the malware specimens, namely, Sality and Ngrbot, indeed refer to bot families.[17] Further, Table 2 reveals the malware samples that generated TCP source port 0 traffic.

[13] http://www.malwaredomainlist.com/.
[14] https://zeustracker.abuse.ch/.
[15] http://www.siteadvisor.ca/.
[16] https://www.mywot.com/.
[17] http://www.symantec.com/connect/blogs/all-one-malware-overview-sality.

By correlating Fig. 7 and Table 2, we can notice that 'Virus.Win32.Sality' is the common factor. In other words, such malware sample has infected some of the probing machines and is simultaneously generating TCP source port 0 traffic. It is noteworthy to mention that such sample has been previously attributed to malicious activities; Dainotti et al. (2012) had documented a large-scale probing campaign that was able to probe the entire IPv4 address space in 12 days. Interestingly, the authors pinpointed that the malware responsible for such campaign was in fact the Sality malware. Thus, from all the extracted insights and by leveraging the three data feeds, we strongly postulate that 'Virus.Win32.Sality' is responsible for part of the TCP source port 0 event.

Related work

In this section, we review the literature related to various concerned topics.

Analyzing Probing Events: The authors of Yu et al. (2007a,b) studied probing activities toward a large campus network using netflow data. Their goal was to infer the probing strategies of scanners and thereby assess the harmfulness of their actions. They introduced the notion of gray IP space, developed techniques to identify potential scanners, and subsequently studied their scanning behaviors. In another work, the authors of Li et al. (2009, 2011) presented an analysis that drew upon extensive honeynet data to explore the prevalence of different types of scanning. Additionally, they designed mathematical and observational schemes to extrapolate the global properties of scanning events including total population and target scope.

Probing Measurement Studies: In addition to Dainotti et al. (2012) and Internet census (2012), Benoit and Trudel (2007) presented the world's first Web census while Heidemann et al. (2008) were among the first to survey edge hosts in the visible Internet. Further, Pryadkin et al. (2004) offered an empirical evaluation of IP address space occupancy whereas Cui and Stolfo (2010) presented a quantitative analysis of the insecurity of embedded network devices obtained from a wide-area scan. In a slightly different work, Leonard and Loguinov (2010) demonstrated IRLscanner, a tool which aimed at maximizing politeness yet provided scanning rates that achieved coverage of the Internet in minutes.

Malware and Probing Correlation: Nakao et al. (2009) were among the first to exploit the idea of correlating malware and probing activities to detect zero-day attacks. The authors leveraged the nicter framework (Inoue et al., 2008) to study the inter-relations between those two activities. They developed scan profiles by observing the dark space and correlated them with malware profiles that had been generated in a controlled environment. In an another closely related work, Song et al. (2011) carried out correlation analysis between 10 spamming botnets and malware-infected hosts as observed by honeypots. They disclosed that the majority of the spamming botnets have been infected by at least four different malware. The authors as well developed methods to identify which exact malware type/family has been the cause of contamination. In a marginally diverse effort, Eto et al. (2009) proposed a malware distinction method based on scan patterns by employing spectrum analysis. The authors stated that by observing certain probing patterns, one can recognize the similarities and dissimilarities between different types of malware. The authors noted that the latter could be used as a fingerprint to effectively infer infection.

Passive DNS Correlation: Correlating DNS activity with IP traffic has been employed to detect malware scanning activities. Since probing techniques do not typically use DNS messages, Whyte et al. (2005) proposed an approach to monitor network traffic and correlate it with any observed DNS messages. Similarly, in Whyte et al. (2006), the same authors addressed the problem of mass-mailing worm activities by utilizing DNS MX messages and SMTP server operations. In order to track scam infrastructure, Konte et al. (2009) correlated some known spam URLs with passive DNS traffic to understand the dynamics of hosting scam campaigns. In alternative works, passive DNS traffic has been used to address the problem of domain blacklisting. For instance, Notos (Antonakakis et al., 2010) and EXPOSURE (Bilge et al., 2011) utilized some behavioral features that are extracted from passive DNS traffic to detect malicious domain names. In a slightly differed work, DNS patterns from the upper DNS hierarchy has been analyzed and deployed in a system dubbed as Kopis (Antonakakis et al., 2011) to extend and detect new related malicious domains. Last but not least, Pleiades (Antonakakis et al., 2012) used passive DNS to tackle the problem of Domain Generation Algorithms by observing NXDOMAIN DNS response messages.

It is indeed evident that the presented work is unique by employing and correlating three real data feeds, namely, darknet, passive DNS and malware data, to analyze, understand and extract inferences related to an unprecedented, previously unanalyzed cyber event, explicitly, the source port 0 probing event.

Conclusion

This paper investigated a rare cyber event that was rendered by excessive traffic originating from source port 0. The goal was to shed the light on the inner details of that event to uncover its mechanism and nature of its sources. To achieve those goals, we proposed a multifaceted approach that exploited and correlated a significant amount of real network security data, including, darknet, passive DNS and malware information. The outcome revealed three probing clusters, in which one of them was shown to be originating from infected bots. By analyzing the maliciousness of the probing sources, the approach uncovered that 28% of those are related to malicious domains. Further, by correlating darknet and malware data, the approach was capable of attributing part of the event to the Sality malware. We envision that such approach could be applicable to analyze other cyber events with similar nature. As for future work, we are working on extending the proposed approach by investigating similarity mechanisms between extracted malicious darknet traffic and generated malware traffic to fortify the attribution evidence.

Acknowledgment

The authors are grateful for NCFTA Canada, Concordia University and the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting this work. The first author is supported by the Alexander Graham Bell Canada Graduate Scholarship (CGS) from NSERC.

References

Alshammari Riyad, Nur Zincir-Heywood A. Can encrypted traffic be identified without port numbers, ip addresses and payload inspection? Comput Netw 2011;55(6):1326–50. ISSN 1389-1286. Doi: 10.1016/j.comnet.2010.12.002. http://www.sciencedirect.com/science/article/pii/S1389128610003695.

Antonakakis Manos, et al. Building a dynamic reputation system for dns. In: Proceedings of the 19th USENIX conference on security; 2010.

Antonakakis Manos, et al. Detecting malware domains at the upper DNS hierarchy. In: USENIX security symposium; 2011.

Antonakakis Manos, et al. From throw-away traffic to bots: detecting the rise of DGA-based malware. In: Proceedings of the 21st USENIX security symposium; 2012.

Benoit Darcy, Trudel André. World's first web census. Int J Web Inf Syst; 2007.

Bilge Leyla, et al. Exposure finding malicious domains using passive DNS analysis. In: NDSS symposium 2011; 2011.

Bou-Harb E, Debbabi M, Assi C. Cyber scanning: a comprehensive survey. Commun Surv Tutor IEEE PP 2013;99:1–24.

Bou-Harb Elias, Debbabi Mourad, Assi Chadi. On fingerprinting probing activities. Comput Secur 2014;43(0):35–48.

Cui Ang, Stolfo Salvatore J. A quantitative analysis of the insecurity of embedded network devices: results of a wide-area scan. In: The 26th annual computer security applications conference. New York, NY, USA: ACM; 2010.

Dainotti Alberto, et al. Analysis of a "/0" stealth scan from a botnet. In: The 2012 ACM conference on internet measurement conference, IMC '12. ACM; 2012.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B Methodol; 1977:1–38.

Eto Masashi, et al. A proposal of malware distinction method based on scan patterns using spectrum analysis. In: Leung Chi Sing, Lee Minho, Chan Jonathan H, editors. Neural information processing. Springer Berlin Heidelberg; 2009. pp. 565–72.

Heidemann John, et al. Census and survey of the visible internet. In: The 8th ACM SIGCOMM conference on internet measurement. New York, NY, USA: ACM; 2008.

Inoue D, et al. nicter: An incident analysis system toward binding network monitoring with malware analysis. In: Information security threats data collection and sharing, 2008. WISTDCS '08; 2008.

Internet census 2012 – port scanning/0 using insecure embedded devices. http://tinyurl.com/c8af8lt.

Iran hacks energy firms. http://tinyurl.com/opjw79c.

Jaeyeon Jung, et al. Fast portscan detection using sequential hypothesis testing. In: 2004 IEEE S&P. IEEE; 2004.

Jin Yu, Simon Gyorgy, Xu Kuai, Zhang Zhi-Li, Kumar Vipin. Gray's anatomy: dissecting scanning activities using ip gray space analysis. SysML07; 2007.

Kaufman Leonard, Rousseeuw Peter J. Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley & Sons; 2009.

Konte Maria, Feamster Nick, Jung Jaeyeon. Dynamics of online scam hosting infrastructure. In: Passive and active network measurement. Springer; 2009. pp. 219–28.

Leonard Derek, Loguinov Dmitri. Demystifying service discovery: implementing an internet-wide scanner. In: The 10th ACM SIGCOMM conference on internet measurement. New York, NY, USA: ACM; 2010.

Li Zhichun, Goyal Anup, Chen Yan, Paxson Vern. Automating analysis of large-scale botnet probing events. In: The 4th international symposium on information, computer, and communications security, ASIACCS '09. New York, NY, USA: ACM; 2009.

Li Zhichun, et al. Towards situational awareness of large-scale botnet probing events. IEEE Trans Inf Forensics Secur; 2011.

MacQueen J, et al. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1; 1967. p. 14. California, USA.

Michael Bailey, et al. The internet motion sensor-a distributed blackhole monitoring system. In: NDSS; 2005.

Moore David, Shannon Colleen, Voelker Geoffrey M, Savage Stefan. Network telescopes: technical report. San Diego: Department of Computer Science and Engineering, University of California; 2004.

Nakao Koji, et al. Practical correlation analysis between scan and malware profiles against zero-day attacks based on darknet monitoring. IEICE Trans Inf Syst; May 2009.

New York Times internal network hacked. http://tinyurl.com/cvnrsac.

Panjwani S, et al. An experimental evaluation to determine if port scans are precursors to an attack. In: The international conference on dependable systems and networks. DSN 2005; 2005. pp. 602–11.

PlayStation network outage caused by 'external intrusion'. http://tinyurl.com/6cbcldv.

Pryadkin Y, Lindell R, Bannister J, Govindan R. An empirical evaluation of IP address space occupancy. USC/ISI technical report ISI-TR; 2004.

Security Information Exchange (SIE). Farsight Security Inc. https://www.farsightsecurity.com.

Song Jungsuk, et al. Correlation analysis between spamming botnets and malware infected hosts. In: 2012 IEEE/IPSJ 12th international symposium on applications and the internet; 2011.

Templeton Steven J, Levitt Karl E. Detecting spoofed packets. In: DARPA information survivability conference and exposition, 2003. Proceedings, vol. 1. IEEE; 2003. pp. 164–75.

Whyte David, Kranakis Evangelos, van Oorschot Paul C. DNS-based detection of scanning worms in an enterprise network. In: NDSS; 2005.

Whyte David, van Oorschot Paul C, Kranakis Evangelos. Addressing SMTP-based mass-mailing activity within enterprise networks. In: Computer security applications conference, 2006. ACSAC'06. 22nd annual. IEEE; 2006. pp. 393–402.

WordPress sites targeted by mass brute-force attack. http://tinyurl.com/cxmgjax.

Wustrow Eric, et al. Internet background radiation revisited. In: Proceedings of the 10th annual conference on internet measurement. ACM; 2010. pp. 62–74.

Yu Jin, et al. Gray's anatomy: dissecting scanning activities using IP gray space analysis. Usenix SysML07; 2007.

Yu Jin, et al. Identifying and tracking suspicious activities through IP gray space analysis. In: The 3rd annual ACM workshop on mining network data. New York, NY, USA: ACM; 2007.