Analysis and Characterisation of Botnet Scan Traffic

Angelos K. Marnerides*, Andreas U. Mauthe†
* School of Computing & Mathematical Sciences, Liverpool John Moores University, UK
a.marnerides@ljmu.ac.uk
†InfoLab21, School of Computing & Communications, Lancaster University, UK
a.mauthe@lancaster.ac.uk

Abstract—**Botnets compose a major source of malicious activity over a network and their early identification and detection is considered as a top priority by security experts. The majority of botmasters rely heavily on a scan procedure in order to detect vulnerable hosts and establish their botnets via a command and control (C&C) server. In this paper we examine the statistical characteristics of the scan process invoked by the Mariposa and Zeus botnets and demonstrate the applicability of conditional entropy as a robust metric for profiling it using real pre-captured operational data. Our analysis conducted on real datasets demonstrates that the distributional behaviour of conditional entropy for Mariposa and Zeus-related scan flows differs significantly from flows manifested by the commonly used NMAP scans. In contrast with the typically used by attackers Stealth and Connect NMAP scans, we show that consecutive scanning flows initiated by the C&C servers of the examined botnets exhibit a high dependency between themselves in regards of their conditional entropy. Thus, we argue that the observation of such scan flows under our proposed scheme can sufficiently aid network security experts towards the adequate profiling and early identification of botnet activity.**

Index Terms—**Botnets, Mariposa, Zeus, NMAP Stealth, NMAP Connect, Scan Traffic, Conditional Entropy**

I. Introduction

The term *botnet* relates to a network of compromised machines (i.e. bots) that without their knowledge are controlled remotely by a botmaster(s) after they were infected by malicious software. In recent years, botnets have shown to be the basic platform for the execution of a wide variety of attacks such as Distributed Denial of Service (DDoS), spamming, phishing and identity theft that in several cases were targeting critical socioeconomic infrastructures, governmental services or commercial organizations [1], [2], [3]. Hence, nowadays botnets are for the most part related to the growth of cyber crimes as well as cyber warfare and their behaviour has attracted a considerable interest by the research community [1], [2], [3], [5]. In parallel, their immediate detection is seen as a challenging task for network security experts within industrial organisations [1], [2].

Throughout the years, the capabilities disclosed within the functionality of botnets have significantly increased leading to the point where traditional categories of malicious software (i.e. malware) may not be easily identified. As mentioned in [1], there is a plethora of reports of a large number of botnets composed by thousands of systems where the total estimated number of compromised machines used by botnets today is in millions. In contrast with the past, the motivation behind the malicious activity of botnets has significantly changed due to the diversification and ubiquitous properties of the Internet that acts as the hub to several socioeconomical infrastructures. Hence, attackers do not primarily seek recognition from the hackers community as in the early stages of the Internet but rather prefer to obtain financial gain or even infiltrate and destroy governmental services [1], [3].

A core procedure within the construction of botnets as well as their expansion by further malware and virus propagation relates with the efficiency of the scan procedure invoked by the botmaster via its Command & Control (C&C) server. The scanning process employed by a botnet is a large scaled coordinated event and usually involves a large number of bots [2]. Mainly, the objective of this process is to provide knowledge to the botmaster regarding vulnerable hosts and services running over particular network domains. Thus, the majority of botmasters construct sophisticated scan procedures in order to avoid common port scanning detection schemes which are currently employed by ISPs(e.g. with tools like NMAP [1]), to detect a high number of vulnerable hosts and speed up the expansion of their botnet. As evidently shown in many cases [4], the scanning procedure is considered as a fundamental element within a botnet and its explicit characterisation would beneficially aid the operations of security components within an organisation in proactively identifying the spread or early establishment of a botnet.

The contribution of our work resides within the statistical characterisation and profiling of scanning activity as triggered by botnets. In particular we provide an insight regarding the intrinsic properties of scan traffic manifested by the well known *Mariposa* [5] and *Zeus* botnets [7] and we profile them under the conditional entropy metric using real backbone and access network packet traces. In contrast with previous studies on both botnets [5], [6], [7] that have been mainly concerned with the system-wise properties and effects of the malware forged by this botnet we elaborate in detail on their scanning procedure. Through our analysis we show that their initiated scan process is between them similar and differs significantly from other scan activities. In order to demonstrate this property we have compared their flow scan characteristics with commonly practised scan processes initiated by the well known NMAP tool. Hence, we have gathered and analyzed packet traces of the *NMAP Connect* and *NMAP Stealth* scans

[1] Nmap : http://nmap.org/

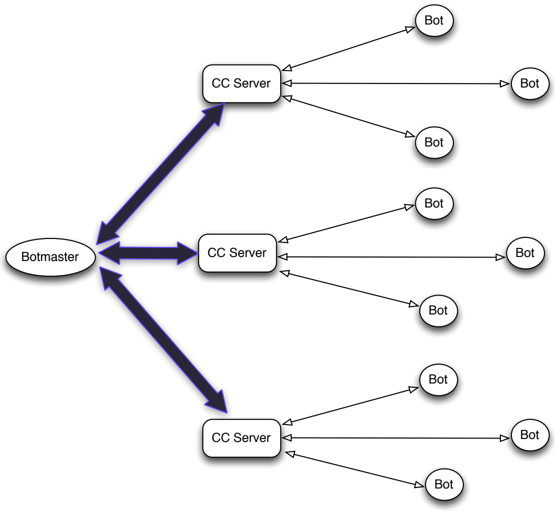

Fig. 1: A conceptual view of the structure for both Mariposa and the Zeus botnet in the examined datasets.

and further indicate their explicit differences with Mariposa and Zeus-related scans. Overall, we argue, that the study documented here may sufficiently contribute to the future studies of botnets and can significantly contribute towards the manifestation of early botnet detection schemes.

The remainder of this paper is structured as follows: Section II provides a brief description of the Mariposa and Zeus botnets as well as the NMAP Stealth and Connect types of scans. Subsequently, Section III introduces the datasets and the methodology employed in this work whereas Section IV demonstrates and discusses the results obtained in our evaluation. Finally, Section V concludes and summarises this paper.

II. Botnets & Scans

A. The Mariposa Botnet

The Mariposa botnet was first discovered by security experts in December 2008 and its malicious operations were mainly dependent on the exploitations of the "Butterfly bot" malware developed by the DDP Spanish hacking group [5]. As documented in the technical report provided in [5], this particular botnet was in a position to compromise more than 10 million machines worldwide and fork several Denial of Service (DoS) attacks as well as numerous bulk e-mail spamming. The expansion of this botnet was stopped in late 2009 by the immediate acts performed by the Mariposa Working Group [5]. However, as we indicate in this work, the Mariposa malware still exists since our initial investigation through a *de facto* IDS such as Snort [2] identified several flows that carried the bot's signature.

As illustrated in Fig. 1, the activities of any botnet (including Mariposa and Zeus that we describe next) are initiated by several command and control (i.e. C&C) servers which are controlled by a botmaster(s). Its capabilities may be considered as threatening since the botmaster may infinitely extend the functionality of the malicious software beyond the initial compromise [5] due to the fact that the botnet has the ability to download and execute arbitrary executable programs. As also identified in this paper, the majority of the C&C servers receive and transmit instructions and data through encrypted UDP datagrams. The instructions given to the C&C or from the CC to the compromised machines include download instructions for a malware update, introduction of new control domain names as well as the renaming of ASCII commands. Under a centralised C&C server scheme [3], a botmaster requires to obtain knowledge regarding a plethora of DNS servers and further identify vulnerabilities that would aid the establishment of new C&C servers that will eventually spread malicious software to new compromised machines. Naturally, the knowledge regarding vulnerable DNS domains is obtained by simply scanning nearby networks through the already settled C&C servers. Hence, the botmaster sends several scan requests to its compromised C&C servers that would subsequently convey these requests to the bots under their control. The detailed analyses provided in [5], [6] indicate that the majority of the Mariposa C&C servers communicate via UDP datagrams and particularly via customized commands of the traceroute UNIX utility [4] that perform DNS lookups. Hence the greatest number of scan traffic would be embodied within the UDP data traffic. As we describe in section IV our investigation came to the same conclusion but with the difference that a small portion of scan traffic was also classified under ICMP and not UDP since the traceroute program can also be initiated under ICMP.

B. The Zeus Botnet

The Zeus botnet is known by many names (e.g. ZBOT, WSNPoem, PRG) and it was firstly tracked by the US Federal Bureau of Investigation (FBI) in 2010 where the botnet was observed to infect more than 3.6 millions of machines worldwide [7]. Despite the fact that it was identified in 2010, recent investigations from the FBI have witnessed new variants [8] of the botnet that infected more than 10 millions of machines that participated in financial cybercrimes such as bank fraud and money laundering transactions. The reason behind the rise-up of several Zeus variants relates with the ease of their deployments by the widely known and freely available Zeus crimeware toolkit [7] where users of such software may easily "craft" a botnet with varying characteristics.

One of the core modifications behind each new Zeus version is the change of the communication protocol between the "zombie" bots with the C&C server(s) that is controlled by the botmaster. Hence, there has been a number of Zeus botnets that employed a decentralised C&C server(s) communication scheme. Moreover, the main operations with respect to the propagation and establishment of the Zeus malware through the networked bots are changed based on modifications on the packet payload signatures.

[2] Snort IDS: http://www.snort.org/

[3] Despite the fact that both Mariposa and Zeus botnets may appear under a P2P-based and decentralised C&C architecture, the examined datasets in this piece of work have revealed a centralised C&C architecture.

[4] Traceroute : http://www.traceroute.org/

TABLE I: Captured Operational Traces

Set	Source	Year	Link Type	Packets	Bytes	Scan-related flows
SamplePoint-F	WIDE Mawai [10]	2012	Backbone	30.3G	65G	2.4M
Zeus	SNORT VRT Labs[11]	2014	Access	8.21K	6.2M	1.2K
UCSD Telescope	CAIDA [12]	2008	Access	16.3M	1.6G	218K

Nonetheless, in the majority of cases where a Zeus botnet was identified and analysed it was revealed that under a centralised C&C communication scheme this particular botnet relies heavily on the HTTP protocol. In particular, the botnet utilises a pull method for synchronising the C&C communication scheme and all the exchanged HTTP messages are encrypted. Moreover, every Zeus malware instance that already exists in a compromised machine initiates DNS lookups as well as *block* port scanning procedures which are mainly triggered by customised processes designed by the botmaster. Block scanning combines both *vertical* and *horizontal* scanning where an attacker scans several destination ports on a given host (i.e. vertical) but also multiple hosts on the same destination port number (i.e. horizontal).However, as we show in this paper and discuss in Section IV the examined Zeus sample indicated a strong *horizontal* port scan procedure.

C. NMAP Scans

This work aims to compare the port scan traffic of the two aforementioned botnets with standardised port scan practises that have been excessively employed by network operators as well as attackers. Therefore, we have chosen to investigate and compare the behaviours of two types of port scans which are easily achieved with the use of the NMAP tool. The NMAP tool provides the capabilities to manifest a range of port and address space scans but the mostly known types are considered to be the Connect and Stealth scans which we have extracted from a real pre-captured packet trace as we explain in the next section.

The NMAP Connect Scan is also known as the TCP SYN connect() scan and is considered as the most typical form of scan performed when using the NMAP facility. This particular approach utilizes the local Operating System's (OS) connect() function in order to initiate a TCP connection to a remote device by starting with a TCP SYN packet with an IP Don't Fragment flag set to 1. As soon as it completes the 3-way TCP handshake on a given port it tears down the connection with a RST/ACK packet.

On the other hand, the NMAP Stealth Scan behaves as a half-open TCP SYN scan since in contrast with the Connect Scan it does not complete the 3-way TCP handshake. Similarly with the Connect Scan it utilizes the connect() OS function and initially sends a TCP SYN packet but with a different flag on the IP Don't Fragment option (i.e. set to 0) in order to receive a SYN/ACK packet response from the remote port. Subsequently, a RST/ACK packet is sent in order to terminate the incomplete TCP handshake procedure. In comparison with the Connect Scan, the Stealth Scan is naturally completed faster and information regarding open ports is derived quicker.

III. Data Description & Methodology

A. Data Description & Pre-processing

As Table I shows, this work has examined three different datasets. The Samplepoint-F is the dataset where the Mariposa botnet was detected and consists of anonymised packet traces captured on a 150Mbps US-Japan TransPacific backbone link and captured by the WIDE Mawi working group in March 2012 [10]. The Zeus dataset is composed of three merged packet traces which were filtered out of a larger packet trace by the researchers in the VRT labs [11] that was captured on an access network in 2014. Finally the third sample we use from the USCD Telescope dataset was captured by CAIDA [12] between December 2008 and January 2009 within an experimental globally routed $/8$ network that filters out legitimate traffic and keeps anomalous unidirectional IP flows which are caused by a number of events such as router misconfigurations, address space and port scanning and DDoS attacks[5]. Overall, this dataset is in a position to provide a snapshot of the anomalous background occurring on the $1/256^{th}$ of all public IPv4 destination addresses on the Internet. However, for the purpose of this work and in order to extract some representative NMAP scan samples we only used a small subset of this dataset.

Apart from the already provided Zeus-related flows, this work first had to isolate scan traffic flows for the Mariposa and NMAP scans. Thus, an initial step was to employ deep packet inspection (DPI) on the captured packets traces and detect any Mariposa-related payload signatures using the Snort IDS on the Samplepoint-F traffic trace. The NMAP Stealth and Connect scans were easily extracted from the UCSD Telescope dataset using custom Wireshark [6] filters.

The scan flows for all 4 scan processes of the Mariposa, Zeus, NMAP Connect and NMAP Stealth were statistically characterised under the conditional metric derived by the distribution of 17 selected flow features for each flow. The 17 "raw" per flow statistical features that form the basis of our work are the following:

- Src port
- Dst port
- Count of packets
- Count of bytes
- Flow duration
- Mean packet size
- Mean packet inter-arrival time
- Sizes of the first 10 packets for each unidirectional flow

Given the above pre-processing step, we subsequently constructed the per-flow feature vectors and then computed the

[5] For the purpose of this paper we have particularly assessed the NMAP Stealth and NMAP Connect scans captured on December 2008 [12].

[6] Wireshark Traffic Analyzer: https://www.wireshark.org/

conditional entropy between them as we describe next.

B. Theoretical Methodology

The foundational element within our methodology lies with the computation of conditional entropy between the aforementioned feature vectors for each consecutive unidirectional scan flow. Essentially, the outcome of this computation is to measure the amount of information for a scan flow X with features $[x_1, x_2, \cdots, x_n]$ and a probability distribution $P_X(x)$ by knowing the occurrence of a scan flow Y with features $[y_1, y_2, \cdots, y_m]$ and a probability distribution $P_Y(y)$. In simple words, the conditional entropy metric provides the level of dependence between two random and independent flows and can pinpoint on whether several transmissions within the network are linearly dependent on each other with respect to the information they carry.

The conditional probability of x given y, $P_{X|Y}(x|y)$, aids to construct the conditional entropy $H(X|Y)$ which is defined by:

$$H(X|Y) = \sum_y P_Y(y) - \sum_x P_{X|Y}(x|y) log(P_{X|Y}(x|y)) \quad (1)$$

As we following describe, the resulted conditional entropy vectors were compared using the two sample Kolmogorov-Smirnov (KS) test in order to measure the level of similarity between the botnet-based and the NMAP-based scan processes.

C. Two Sample Kolmogorov-Smirnov (KS) Test

In order to validate our hypothesis on whether the conditional entropy derived for all the scan flows of the Mariposa, Zeus and NMAP scans would exhibit a level of similarity we have employed the Kolmogorov-Smirnov (KS) test that we following describe.

Let $T_m = t_1, \ldots, t_m$, $T_z = t_1, \ldots, t_n$, $T_n c = t_1, \ldots, t_c$ and $T_n s = t_1, \ldots, t_s$ represent the resulting conditional entropy vectors gathered for each transport layer network flow where T_m, T_z, $T_n c$ and $T_n s$ denote the vectors for the Mariposa, Zeus, NMAP Connect and NMAP Stealth scan respectively. However, for the sake of clarity we elaborate on the two sample KS test by taking as examples the T_m and T_z vectors.

The vector T_m of size m has the cumulative distribution function (i.e. c.d.f) $F(x)$ and T_z of size n a c.d.f with $G(x)$ and their corresponding *empirical* c.d.fs as $F_m(x)$ and $G_n(x)$ respectively. Under these terms, the KS test holds two hypotheses, the null hypothesis $H_0 : F = G$ and the rejection of the null hypothesis $H_1 : F \neq G$. In order to validate the null hypothesis via measuring the statistical (in)significance between $F_m(x)$ and $G_n(x)$ it is required to compute the Kolmogorov-Smirnov statistic D_{mn} defined as:

$$D_{mn} = \left(\frac{mn}{m+n}\right)^{1/2} \sup_n |F_m(x) - G_n(x)| \quad (2)$$

The null hypothesis is rejected in the statistical significance level a if

$$\sqrt{\frac{mn}{m+n}} D_{mn} > K_\alpha \quad (3)$$

where K_α of statistical significance level α can be found from the relationship of the Kolmogorov distribution K denoted as:

$$K = \sup_{t \in [0,1]} |B(t)| \quad (4)$$

and has the following relationship [7]

$$Pr\left(K \leq K_\alpha\right) = 1 - \alpha \quad (5)$$

The statistical significance level α is the most critical parameter that actually determines the sensitivity at which the KS test will reject the null hypothesis or otherwise. Thus, after experimentation we kept it to hold the value of 0.05 since while tuning this parameter with higher values there was a biased result in favor of the null hypothesis.

In order to thoroughly determine and exhibit the resulting outcomes of our hypothesis validation we have computed the resulting *p-value*. In practise, the p-value denotes the probability on whether the KS test statistic derived by equation 2 and examined in equation 3 is extreme or more extreme than the observed value under the null hypothesis. As we show in section IV-B if the resulted *p-value* is much lower than the significance level a then we reject the null hypothesis and the opposite if greater.

IV. Results

A. Analysis of Volume-based Features

The evaluation conducted on the network traces had first to comprehend the network-wise properties of the scan flows triggered by the CC servers of Mariposa and Zeus as well as the NMAP-related scans. Therefore, we visualised and manually inspected the volume-wise activity with respect to their destination and source IP ports.

Figures 2, 3, 4 and 5 provide the behaviour of bytes and packets with respect to the source and destination (src/dst) IP ports per each unidirectional scan flow. Overall, it is clearly evidenced that both botnets exhibit a much higher level of consistency in their scan activities since there are clear patterns with respect to the outgoing and incoming volume-wise distribution (i.e. counts of bytes and packets) of their associated scan flows on specific IP src/dst ports. On the other hand the flows triggered by the two NMAP scans behave in a random fashion for an extremely large range of IP src/dst ports. However, the explicit case of the Mariposa botnet indicates a *block scan* property where both horizontal and vertical scans are visible.

Mariposa Scan: A closer look in our datasets revealed that the scan activities of the Mariposa C&C server were not restricted at only using UDP datagrams as reported in [5] but also to the transmission of ICMP packets. In particular, the compromised C&C server used the scanning capabilities of the traceroute utility and was arbitrarily aiming to detect vulnerable ports on several domains. The manual inspection on the actual packet traces revealed some instances where

[7] $B(t)$ in equation 4 denotes the Brownian bridge of a continuous stochastic process.

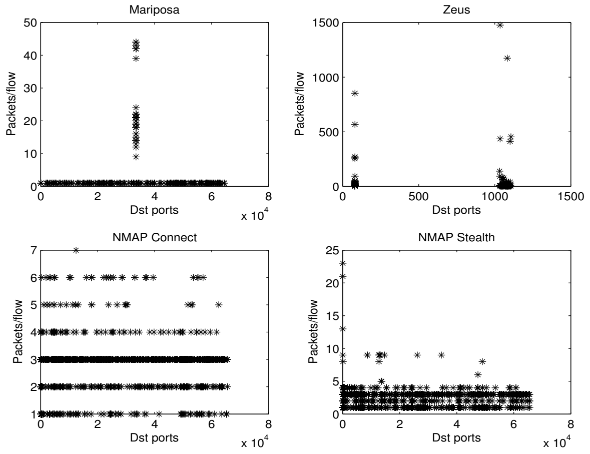

Fig. 2: Packet activity for all types of scans with respect to IP destination ports.

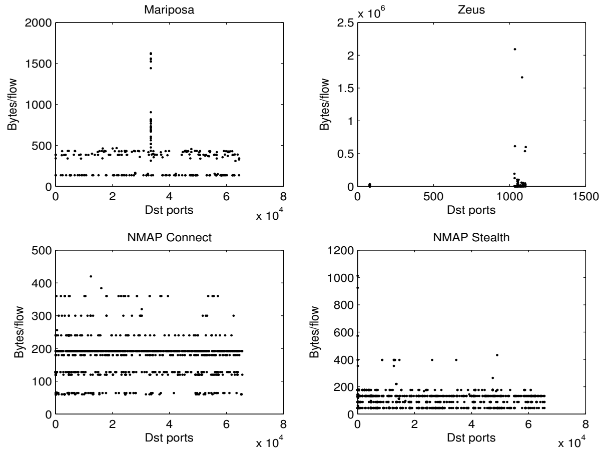

Fig. 3: Byte activity for all types of scans with respect to IP destination ports.

the traceroute utility was transmitting a sequence of three ICMP echo request packets on a given destination host with increasing Time-To-Live (TTL) values. Moreover, the C&C server performed scans to open hosts on a range of DNS domains under the traceroute utility by the transmission of UDP datagrams. From both a packet and byte-wise perspective on Figures 2 and 3, it is fairly obvious to pinpoint that the scan activity is being due to the vertical shape of the packets and bytes on the UDP port 33434 which is a standard traceroute port. Several flows triggered by the C&C server were not only initiated for scanning purposes but also contained reply data from the traceroute service from each remote host that naturally important network topology information was present regarding the next hop of their adjacent network. These flows are easily identifiable in both figures due to the high byte and packet-wise consumption they exposed. Finally, low volume flows (i.e. low packet and byte counts per flow) reside on the horizontal shape(s) in both Fig. 2 and Fig. 3 for the Mariposa-related plots and they mainly relate with the aforementioned random DNS lookups.

The above findings are also verified while observing the volume distribution with respect to the initiated source ports in all the Mariposa-related scan flows. As demonstrated explicitly for the Mariposa plots via Fig. 4 and Fig. 5, there is a crisp visualisation of the volume-based behaviour in regards of the sender source ports under two vertical shapes. The most volume-wise intensive vertical shape denotes all the flows related with any traceroute activity on the UDP destination port 33434 which were initiated within the range 52995-53586 of UDP source ports. We argue that these ports were used since they are unregistered by IANA as well as arbitrary, thus any signature-based IDS or scan detector would not be in a position to flag them easily[8]. Apart from these highly visible source ports, the second most frequent source port is the UDP port 53 which justifies the fact of random and low-volume DNS lookups and surely complies with the horizontal shapes of the Mariposa plots in Figures. 2 and 3. The majority of the DNS lookup flows were single-packet flows with an average of 135 bytes.

Zeus Scan: The Zeus botnet scan activity exhibited a far more distinguishable persona than the one initiated by the Mariposa botnet. From both packet and byte-wise distributions on the targeted IP destination ports, it is evidenced that there was the initiation of a typical *horizontal scan* procedure. This behaviour is justified by the visualisations in Figures 2 and 3 where an immense amount of bytes and packets are destined towards the TCP ports 1032,1033 and 1035 as well as the TCP port 80 (i.e. vertical shapes). Hence, the scanning activities of Zeus were mainly dedicated on the HTTP protocol and the objective was to scan for vulnerable web servers on these particular ports on a number of different hosts on several domains. At the same time, the requests sent to the TCP destination ports 1032,1033 and 1035 are mapped as the actual probing, scanning and further deployment of the Zeus trojan on remote hosts. In particular, Windows-based machines are prone on this particular ports due to the insecure initiation of the Message Application Programming Interface (MAPI) protocol on the Outlook e-mail clients. Hence, this finding verifies the fact that Zeus-related Trojans as initiated by their botmaster hold the objective to infiltrate mail applications in order to enforce their propagation through spamming and further establish strong foundations towards phishing attacks.

Nevertheless, apart from pure informational scanning it is also noticeable via Fig. 2 and Fig. 3 the intention of the scan flows from the Zeus C&C server to flood these particular ports with unidirectional flows that contain large numbers of packets and extremely large byte sizes that go beyond the Maximum Transmission Unit (MTU) threshold of 1500 bytes per packet. Such behaviours are also depicted and justified in the Zeus-related plots provided in Figures 4 and 5. In particular, the revealed vertical shapes demonstrate that the mostly active TCP source port was the the HTTP port 80 as well as those

[8] Around 60% of these flows were not detected initially by Snort since most of the pre-defined Zeus-related rules were only observing the packet payload. Thus we had to refine Snort-specific rulesets alongside customized tcpdump filters in order to extract them.

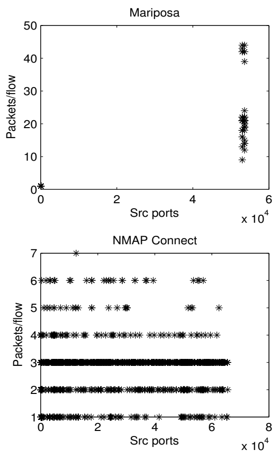

Fig. 4: Packet activity for all types of scans with respect to IP source ports.

wthin the range of $1051-1099$. The greatest majority of byte-wise (with a lesser number of packets) intensive scan flows were sourced by TCP port 80 and they were probing multiple hosts on the same destination port for open web servers. On the other hand the scanning and flooding flows destined randomly to the TCP ports 1032,1033 and 1035 were sent from TCP source ports ranging between $1051-1099$. Each of this flow was characterised under large numbers of packets but each holding an extremely low byte size.

NMAP Scans: As anticipated by our initial hypothesis, both NMAP scans have shown an utter random behaviour where a vast range of TCP src/dst ports was used in order to perform the extracted portscans. Hence, the generated plots that represent the packet and byte-wise behaviour for both the NMAP Connect and NMAP Stealth indicate that their scan process is not volume-wise intensive and their characteristics comply with their description provided earlier (Section II-C). In contrast with the volume-wise scan profiles of the Zeus and Mariposa botnets, the NMAP-based port scans cannot be classified either as horizontal nor vertical since random TCP destination ports were requested on random hosts.

In particular and as exhibited by Figures 2, 3, 4 and 5, all the unidirectional flows related with the NMAP Connect scan had a maximum number of 7 packets with a maximum size of 430 bytes. The randomness behind all the scans are easily distinguishable by the shape of all the plots related with this particular scan where all possible TCP src/dst ports within the range of $0-9999$ were used.

Under a similar fashion, the NMAP Stealth scan exhibited a random scan behaviour where all the possible combinations of TCP src/dst port pairs were used. The flows related with this scan process have demonstrated similar volume utilization with those of the NMAP Connect scan but with some minor cases of higher packet and byte count. As visually depicted by Figures 2, 3, 4 and 5 the largest number of unidirectional flows had less than 5 packets with an average size of 350 bytes. However, there were around 30 flows that exhibited a greater number of packets within the range of $7-23$ packets with

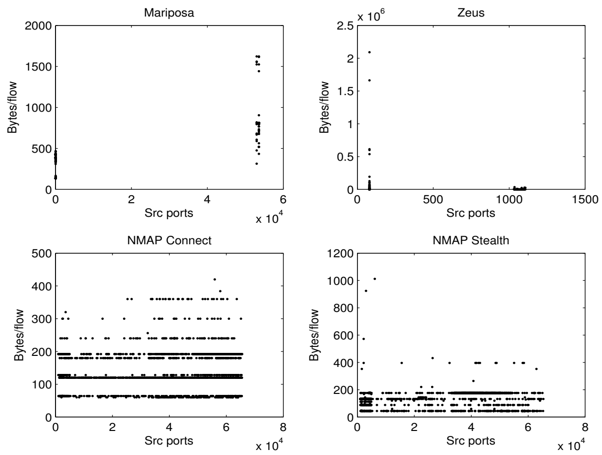

Fig. 5: Byte activity for all types of scans with respect to IP source ports.

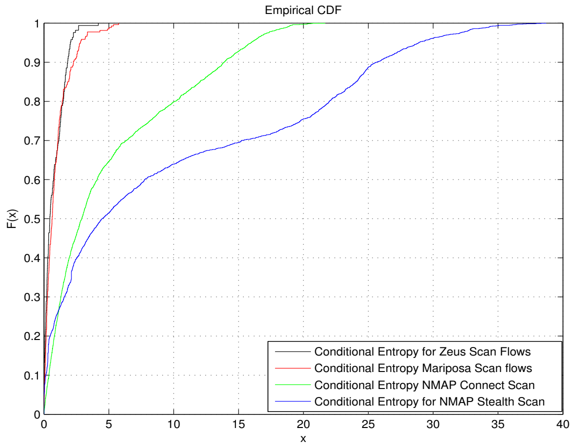

Fig. 6: Empirical CDFs of the resulting conditional entropy vectors for each scan process.

higher byte utilization between 400 and 1100. Nonetheless, none of these flows appeared with a packet that reaches or overpasses the MTU threshold as it happenned with the botnet-related scan packets discussed earlier.

B. Scan Traffic Comparison

By following the methodology presented in Section III we were able to compute the conditional entropy vectors after treating each scan flow as a vector composed by the 17 flow-features presented earlier in Section III-B. A subsequent step was to compute the empirical cummulative distribution function (CDF) for the newly composed conditional entropy vectors and finally compare them with the two sample KS test that we have also described earlier (Section III-C).

Fig. 6 illustrates the resulting empirical CDFs of the conditional entropy vectors associated for each type of scan. Undoubtedly , it is evident that the CDFs corresponding to NMAP-based scans differ significantly from the CDFs that represent the conditional entropy between the scan flows triggerred by the Zeus and Mariposa botnet. In parallel, the

TABLE II: The asymptotic p-values between the various types of scans under the two-sample KS test. $H(0)$ denotes the null hypothesis of similarity between the examined distributions where the significance level $\alpha = 0.05$.

Scan Comparison	**p-value**	$H(0)$
Zeus vs. Mariposa	0.2033 > α	accepted
Zeus vs. NMAP Stealth	3.5405e-13 < α	rejected
Zeus vs. NMAP Connect	3.6905e-12 < α	rejected
Mariposa vs. NMAP Stealth	8.9434e-21 < α	rejected
Mariposa vs. NMAP Connect	9.9434e-19 < α	rejected
NMAP Stealth vs. NMAP Connect	6.8722e-61 <α	rejected

generated CDFs for the two botnets have a visual similarity and they establish a leverage towards the acceptance of the null hypothesis in the KS test. Hence it can be hypothetized that their scan characteristics could comply with the same distribution. This hypothesis is validated under the KS-test where all the resulting p-values are computed as depicted in Table II.The computed p-values surely justify our hypothesis on whether the scan flows produced by the two botnets hold a significant statistical similarity between them with respect to the conditional entropy. At the same time the comparison of both botnet conditional entropy vectors with the NMAP-related scan traffic conditional entropy vectors have appeared to reject the null hypothesis since the resulted p-values where extremely lower than the significance level α. Similarly, the comparison between the two different scans composed by the same tool have also indicated a completely different persona with respect to their resulting conditional entropy, thus the generated p-value led to reject the null hypothesis. Thus, they do not have common characteristics from a statistical viewpoint with respect to their flow-based feature distributions.

Overall, this simple comparison between the 4 different scan traffic profiles is leading to the conclusion that coordinated and carefully-designed botnets manifest intrinsic scan properties which are not easily identifiable and separable by typical port scans initiated by widely-used tools such as NMAP. Thus, their identification and further grouping would have to be expressed in terms of meaningful statistical features.

V. Conclusions

This paper provides a thorough analysis and insight regarding the scan processes initiated by the Mariposa and Zeus botnets as identified on real backbone and access network traces since both botnets have been excessively reported as two of the most dangerous botnets in the wild with millions of machines compromised. Hence, we initially examine their scan procedures with respect to their volume-based characteristics and we further compare them under their resulting conditional entropy metric with two commonly practised by attackers NMAP-based scans (i.e. NMAP Connect and NMAP Stealth scans) that we have extracted from real pre-captured operational network data.

The detailed volume-based analysis revealed that the Mariposa botnet's C&C server employed a *block* types of port scans where both horizontal and vertical types of scans are accommodated using the UDP and ICMP protocol. In parallel, the Zeus botnet demonstrates a distinctive horizontal port scan scan where scan-related flows are initiated on the same TCP destination port on multiple hosts for different domains. On the other hand, scan flows initiated by NMAP Stealth and NMAP Connect scans appeared to operate arbitrarily using random src/dst TCP ports. However, despite the volume-wise differences between the scanning procedures between the Mariposa and the Zeus botnets our comparison under the conditional entropy metric as composed by 17 per-flow features demonstrated that both scanning processes hold an extremely significant statistical similarity. In parallel, both botnet scan processes do not hold any similar characteristics with any of the two examined NMAP-based scan processes. Hence, we have shown that carefully crafted botnets do aim to go beyond standardised scan procedures as implemented by commonly used tools such as NMAP. In parallel, we demonstrate that regardless of the differing botnet volume-wise scan behaviour, a characterisation under conditional entropy may adequately profile them. Therefore, the outcomes of this work demonstrate the capability of our proposed scheme for profiling differing botnet scan traffic. We have provided a simple botnet scan traffic profiling methodology under the conditional entropy metric that goes beyond rule-based schemes and can sufficiently aid network security experts while composing network diagnostics. Overall we argue that the study and method reported herein can significantly contribute and constitute as a vital component within the design of holistic early botnet detection tools.

Acknowledgments

The authors would like to thank the UK EPSRC funded "Situation-Aware Information Infrastructure" (SAI^2) project, EP/L026015/1, that has kindly supported this work.

References

[1] Barford, P., Blodgett, M., Toward Botnet Mesocosms. In Proceedings of the USENIX First Workshop on Hot Topics in Understanding Botnets (HotBots I), April, 2007

[2] Karasaridis, A., Rexroad, B., Hoeflin, D., Wide-scale Botnet Detection and characterisation, in Proceedings of the USENIX First Workshop on Hot Topics in Understanding Botnets, HotBots' 07, 2007

[3] Li, Z., Goyal, A., Chen, Y., , Honeynet-based Botnet Scan Traffic Analysis, in Journal Advances in Information Security, Springer, 2008

[4] Panjwani, S.; Tan, S.; Jarrin, K.M.; Cukier, Michel, "An experimental evaluation to determine if port scans are precursors to an attack," Dependable Systems and Networks, 2005. DSN 2005

[5] Defence Intelligence LTD, Technical Report : Mariposa Botnet Analysis , http://defintel.com/docs/Mariposa_Analysis.pdf

[6] Sinha, P., Boukhtouta, A., Belarde, V., H., Debbabi, M., Insights from the Analysis of the Mariposa Botnet, in Proceedings of the 5^{th} International Conference on Risks and Security of Internet and Systems (CRiSIS) 2010

[7] Binsalleeh, H., Ormerod, T., Boukhtouta, A., Sinha, P., Youssef, A., Debbabi, M., Wang, L., "On the analysis of the Zeus botnet crimeware toolkit," Privacy Security and Trust (PST), 2010 Eighth Annual International Conference on , vol., no., pp.31,38, 17-19 Aug. 2010

[8] FBI Report on the GameOver Zeus botnet, FBI.gov, http://www.fbi.gov/news/stories/2014/june/gameover-zeus-botnet-disrupted, 2014

[9] Lu, C. and Brooks, R., Botnet traffic detection using hidden Markov models. In the 7^{th} Annual CSIIRW, Oak Ridge, TN, USA, 2011, Article 31,

[10] The Mawi working group : http://mawi.wide.ad.jp/mawi/

[11] SNORT Sourcefire Vulnerability Research Labs: http://labs.snort.org/

[12] The CAIDA UCSD Network Telescope "Patch Tuesday" Dataset 21-11-2008: http://www.caida.org/data/passive/telescope-patch-tuesday_dataset.xml