Inferring Internet-scale Infections by Correlating Malware and Probing Activities

Elias Bou-Harb, Claude Fachkha, Mourad Debbabi, Chadi Assi
NCFTA & Concordia University
Montreal, Quebec, Canada
{e_bouh, c_fachkh, debbabi, assi}@encs.concordia.ca

***Abstract*—This paper presents a new approach to infer malware-infected machines by solely analyzing their generated probing activities. In contrary to other adopted methods, the proposed approach does not rely on symptoms of infection to detect compromised machines. This allows the inference of malware infection at very early stages of contamination. The approach aims at detecting whether the machines are infected or not as well as pinpointing the exact malware type/family, if the machines were found to be compromised. The latter insights allow network security operators of diverse organizations, Internet service providers and backbone networks to promptly detect their clients' compromised machines in addition to effectively providing them with tailored anti-malware/patch solutions. To achieve the intended goals, the proposed approach exploits the darknet Internet space and employs statistical methods to infer large-scale probing activities. Subsequently, such activities are correlated with malware samples by leveraging fuzzy hashing and entropy based techniques. The proposed approach is empirically evaluated using 60 GB of real darknet traffic and 65 thousand real malware samples. The results concur that the rationale of exploiting probing activities for worldwide early malware infection detection is indeed very promising. Further, the results demonstrate that the extracted inferences exhibit noteworthy accuracy and can generate significant cyber security insights that could be used for effective mitigation.**

I. Introduction

Today, the safety and security of our society is significantly dependent on having a secure infrastructure. This infrastructure is largely controlled and operated using cyberspace. Although tremendous efforts have been carried out to immune the cyberspace from diverse debilitating, intimidating and disrupting cyber threats, such space remains to host highly sophisticated malicious entities. The latter could be ominously leveraged to cause drastic Internet-wide and enterprise impacts by means of large-scale probing campaigns [1], distributed denial of service attacks [2], advanced persistent threats [3] and spamming botnets [4]. According to Panda Security, a staggering 33% of worldwide Internet machines are infected by malware [5]. Moreover, McAfee records over 100 thousand new malware samples every day; a momentous 69 threats every minute or around one new threat every second [6].

Network security operators of private and governmental organizations, Internet Service Provides (ISPs) and content delivery networks as well as backbone networks face, on a daily basis, the crucial challenge of dealing with their clients' malware-infected machines. The latter not only hinders the clients' overall experience and productivity but also jeopardizes the entire cyber security of the provider (i.e., causing vulnerabilities or opening backdoors in the internal network). Further, it significantly degrades the provided quality of service since the compromised machines will most often cause excessive increase in bandwidth that could be rendered by extreme Peer to Peer (P2P) usage, spamming, command-and-control communications and malicious Internet downloads. Additionally, if providers' networks were used to trigger, for instance, a malware-orchestrated spamming campaign, then such providers could as well encounter serious legal issues for misusing their infrastructure (i.e., for example, under the Canadian House Government Bill C-28 Act [7]). Consequently, this will immensely adversely affect the operators' business, reliability and reputation.

Thus, network security operators are interested in possessing a cyber security capability that generates inferences and insights related to their clients' malware-infected machines. It is significant for them to be able to pinpoint such machines in addition to extract intelligence related to the exact malware type/family. The latter will facilitate the distribution of suitable and tailored anti-malware solutions to those compromised clients. Indeed, this cyber security capability should possess the following requirements. First, it should be prompt; it must possess the ability to detect the infection as early as possible in an attempt to thwart the creation of botnets and to limit the sustained possible collateral damage and any symptoms of infection. Second, it should be cost-effective; the approach should not overburden the provider with implementation scenarios and their corresponding supplementary costs. In fact, the latter point is extremely imperative and decisive; ISPs are frequently accused of ignoring their clients' malware infections because the task to detect and disinfect them is tedious, prolonged and undoubtedly expensive [8, 9]. In this paper, we elaborate on such a cyber security capability that satisfies the mentioned requirements. Specifically, we frame the paper's contributions as follows:

- Proposing a new approach to infer Internet-scale malware-compromised machines. The approach aims at detecting such machines as well as identifying their exact infection type. The approach achieves its aims without recording or analyzing the symptoms of infection (i.e., spamming, excessive bandwidth usage, etc.), which renders it efficient from both space and processing perspectives. Further, it exploits probing activities to attain early detection of contamination incidents in addition to requiring no implementation at the providers' premises, eliminating the cost burden.

- Leveraging the darknet Internet space, around half a million routable but unallocated IPs, which permits the observation and identification of worldwide malware-infected machines, without requiring any providers' aid or information.
- Correlating malware and probing network activities to achieve the intended goals by employing numerous statistical, fuzzy hashing and entropy based techniques.
- Evaluating the proposed approach using 60 GB of real darknet traffic and 65 thousand real malware samples.

The remainder of this paper is organized as follows. In the next section, we review the related work. In Section III, we elaborate on the proposed approach. Specifically, we explain its rationale, describe its components and present the leveraged mechanism, methods and techniques. We empirically evaluate the approach in Section IV. In Section V, we pinpoint some limitations of the proposed approach. Finally, concluding remarks and some future work are disclosed in Section VI.

II. Related Work

In this section, we review some literature work related to malware and probing correlation analysis. Further, we briefly highlight two approaches that are adopted in the industry for the purpose of detecting malware-infected machines.

Nakao et al. [10] 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 [11] 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. Their work seems limited in a number of points. First, the authors did not validate the accuracy of the extracted probing activities from the dark space. Second, the extracted profiles were based on few textual network and transport-layer features, where the actual correlation engine's mechanism was obscured. Third, their experiments were based on only one malware sample. In contrary, in this work, we first apply a validated statistical approach to accurately extract probing activities from darknet traffic. Second, in a first attempt ever, we correlate probing and malware activities by applying fuzzy hashing and information theoretical based techniques on the entire network traffic that was generated by those activities. Third, our experiments involve around 65 thousand malware samples. Fourth, the aim of this work differs as it is rendered by the capability to provide network security operators, worldwide, with the ability to rapidly and cost-effectively detect their clients' infections, without requiring the providers to maintain an implementation nor provide any aid or disclose any sensitive network related information. In an another closely related work, Song et al. [12] 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. Our work differs from this work as we are correlating probing activities rather than spamming for early infection detection. Further, we are leveraging the darknet space instead of honeypots to extract Internet-scale cyber security intelligence. In a slightly different work, Eto et el. [13] 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. The authors however, did not perform any correlation but rather limited their work to observation and analysis.

The industry has also developed approaches to identify malware-compromised machines. For instance, True Internet, one of Thailand's largest ISPs, had adopted a behavioral approach to infer its clients infected machines. Their approach monitors the symptoms of infection, including but not limited to, spamming, excessive P2P usage and Denial of Service (DoS) attempts, and subsequently triggers an alert towards a controller, which then automatically quarantine the client. Although such approaches might be effective, they are typically late in detection, which might cause serious vulnerabilities within the provider. Moreover, they are usually not cost-effective as the provider ought to purchase and maintain other detection systems. Another example would be NetCologne, an ISP and cable provider in Germany, that took a different approach to automate how it deals with subscribers that are infected with malware. NetCologne setup and maintained a honeypot; infected machines often attempt to attack other computers on the same network and hence the honeypot is an accessible target that allows the identification of compromised machines. While this approach seems practical in detecting infected machines, it is neither cost-effective since the provider needs to implement and maintain the honeypot nor it is able to identify the exact malware type/family that had contaminated those machines. Moreover, honeypot evasion approaches are known to be effective [14] and are often adopted by sophisticated malware.

III. Proposed Approach

In this section, we elaborate on the proposed approach as depicted in Figure 1. Specifically, we discuss its rationale, describe its components and present its employed mechanism, methods and techniques.

The rationale behind the proposed approach stems from the need to detect the infection at early stages of contamination. In this context, probing or scanning activities are known to be the very first symptoms of infection [15, 16]. On the other hand, the Internet dark space, a set of unallocated yet routable IP addresses (i.e., dark sensors), has shown to be an effective method to generate Internet-scale cyber threat intelligence [17, 18]. Thus, in a nutshell, the proposed approach aims at extracting probing activities as received by a darknet and subsequently correlating them with malware samples. By leveraging geo-location information, the approach strives to generate insights related to worldwide compromised machines in addition to identifying their exact infected malware type/family. Subsequently, the extracted inferences are distributed to concerned providers. In the sequel, we elaborate on each component of the proposed approach.

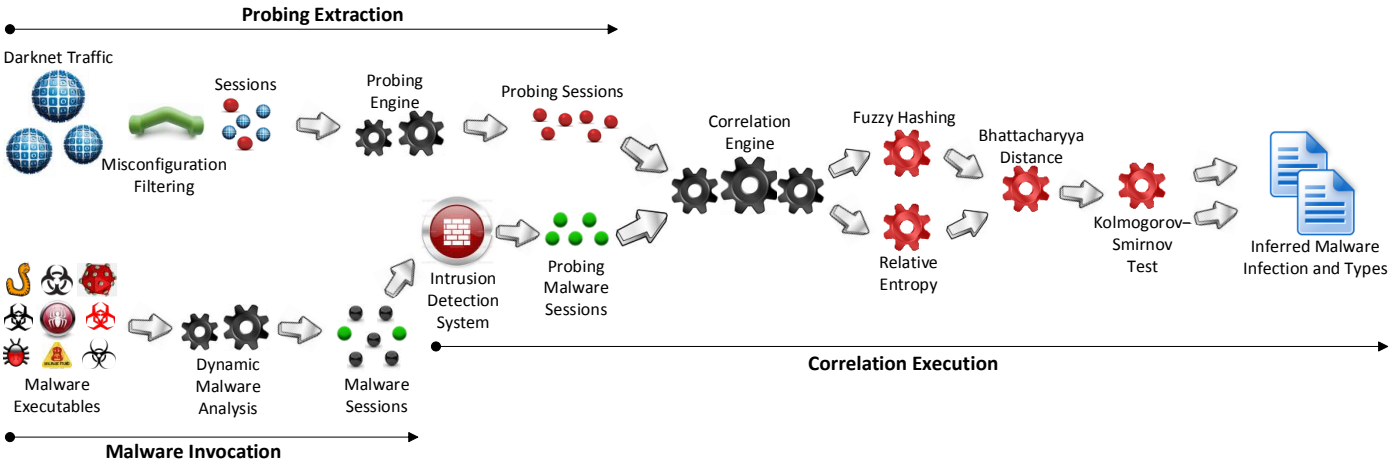

Fig. 1: The Components of the Proposed Approach

A. *Probing Extraction*

In [16][1], we proposed a new approach to fingerprint probing activities. The approach aimed at detecting the probing activity and identifying the exact technique that was employed in the activity. The approach is advantageous in comparison with other methods as it does not rely on identifying the scanning source and is independent from the scanning strategy (remote to local, local to remote, local to local), the scanning aim (wide range or target specific) and the scanning method (single source or distributed). When empirically evaluated using a significant amount of real darknet data, the approach yielded 0 false negative in compassion with the leading network intrusion detection system, Snort[2]. To achieve its aims, the approach uniquely employs the detrended fluctuation analysis technique coupled with numerous diverse statistical methods. Readers that are interested in more details related to the approach are kindly referred to [16]. In this work, and to successfully extract probing activities from darknet traffic, we adopt and leverage the previously proposed approach. The outcome of this procedure are accurate and validated probing sessions in packet capture format (i.e., pcaps).

B. *Malware Invocation*

We operate a dynamic malware analysis module that is based on ThreatTrack Security sandbox environment[3] (i.e., controlled environment). After receiving daily malware samples from ThreatTrack feeds, they are interactively sent to the sandbox, where they are executed by client machines. The clients could be virtual or real and possess the capability to run Windows or Unix, depending on the malware type under execution. The behavior of each malware is monitored and all its corresponding activities (i.e., created files, processes, network traffic, etc.) are recorded. For the sake of this work, we extract the network traffic generated by approximately 65 thousand unique and recent malware samples as pcaps. The pcaps contain communication traffic generated from the malware to other internal or external hosts. Those malware samples belong to diverse malware types including, Trojan, Virus, Worm, Backdoor, and AdWare coupled with their corresponding families and variants. We rely on Kaspersky for a uniform malware naming convention[4].

C. *Correlation Execution*

We formulate the problem of correlating probing and malware sessions as follows. Given a probing session that is extracted from darknet traffic, 1) investigate whether or not the session originates from a malware and 2) identify the exact or probable malware type/family that is generating such probing, if it was shown that the session is malware-related. In an attempt to address this problem, we perform the following. We leverage Snort's probing engine, the sfPortscan pre-processor, to detect which malware pcaps possess any signs of probing activity. We omit those malware pcaps that demonstrate a negative output. To attribute a specific malware to a probing session, we adopt a two-step procedure. First, we apply the notion of fuzzy hashing [19] between the probing session and the remaining malware pcaps. Fuzzy hashing is advantageous in comparison with typical hashing as it can provide a percentage of similarity between two samples rather than producing a null value if the samples are different. This popular technique is derived from the digital forensics research field and is typically applied on files or images [19, 20]; to the best of our knowledge, our approach is among the first to explore the capabilities of this technique on cyber security data. We further apply an information theoretical metric, relative entropy, as proposed by Lee and Xiang [21], between the given probing session and the malware pcaps. Relative entropy is a measure of the distance of the regularities between two datasets. If the relative entropy is $= 0$, this indicates that the two datasets have the same regularity. At this point, we 1) omit the probing sessions that demonstrate less than 5% similarity using both tests[5] and 2) select the top 10% malware pcaps that were found to minimize the entropy and maximize the fuzzy hashing percentage. The rational behind the latter approach stems from the need to filter out the malware pcaps

[1]This work won the best paper award at IEEE ARES 2013.
[2]http://www.snort.org/
[3]http://www.threattracksecurity.com/
[4]http://securelist.com/en/threats/detect?chapter=136
[5]These sessions indicate that they do not possess any malware-related behavior.

that do not possess probing signs similar to the probing session. Second, using the remaining 10% malware pcaps, we extract their probing sessions as pinpointed by sfPortscan. For each of the malware probing sessions, we apply the Bhattacharyya distance [22] between those and the given probing session. The latter statistic test is an established and an effective metric to determine the overlap of two sample distributions. By selecting 1% of malware pcaps that were shown to reduce the Bhattacharyya distance, we further significantly reduce the possible malware pcaps that the given probing session could be similar to. Finally, to exactly attribute the given probing session to a specific malware, we employ the two sample Kolmogorov-Smirnov statistic test [23] between the remaining malware probing sessions and the given probing session. The test will output 0 if a positive match occur; 1 otherwise. If a positive match occurs, this indicates that the probing session has been generated from the inferred exact malware. Otherwise, we refer back to the output of the Bhattacharyya distance and select a set of probable malware pcaps that were shown to be relatively close to the given probing session.

IV. Empirical Evaluation

We possess real darknet data that we are receiving on a daily basis from a trusted third party. Such traffic originates from the Internet and is destined to numerous /13 network sensors (i.e., around half a million IPs). In a nutshell, darknet traffic is Internet traffic destined to routable but unused Internet addresses. Since these addresses are unallocated, any traffic targeting them is deemed as suspicious. We use one week of data (60 GB) that was collected during the duration of July 25th to August 1st 2013, to empirically evaluate our approach.

Darknet traffic is typically composed of three types of traffic, namely, scanning, backscattered and misconfiguration [24]. Scanning arises from bots and worms while backscattered traffic commonly refers to unsolicited traffic that is the result of responses to denial of service 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.

The first aim is to filter out misconfiguration data. We use a simple metric that records the average number of sources per destination darknet address. This metric should be significantly larger for misconfiguration than scanning traffic. However, although it differentiates misconfiguration from scanning traffic, 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 [24], flags in packet headers, such as TCP SYN+ACK, RST, RST+ACK, ACK, etc., that resemble backscattered traffic [24]. Subsequently, we filter out flows that lack that observation, deeming them as misconfiguration.

Second, we aggregate the connections into sessions using an approach similar to the first step algorithm by Kannan et al. [25]. We consider all those connections within T_{aggreg} of each other as part of the same session for a given pair of hosts. We used the same proposed threshold, T_{aggreg} = 100 seconds, and found that this seems to correctly group the majority of connections between any given pair of hosts. We further execute the probing fingerprinting approach that was briefly highlighted in Section III-A to extract 200 probing sessions. It is noteworthy to mention, that each of those probing sessions is being generated by a unique source. Keeping in mind that most probing activity is generated from non-spoofed Internet Protocol (IP) addresses (so the actual scanner can essentially receive back the probing results), the extracted probing sessions indeed resemble real machines.

We proceed by investigating any signs of probing activities within the 65 thousand malware pcaps. The output disclosed that 13,105 malware pcaps possessed such activities. The latter corroborates that a significant amount of malware samples in fact generate probing activities; leveraging such activities for early infection detection might be a viable and a promising approach. The distribution of the types of those probing activities within the identified malware pcaps is depicted in Figure 2. It is disclosed that UDP probing is the most employed

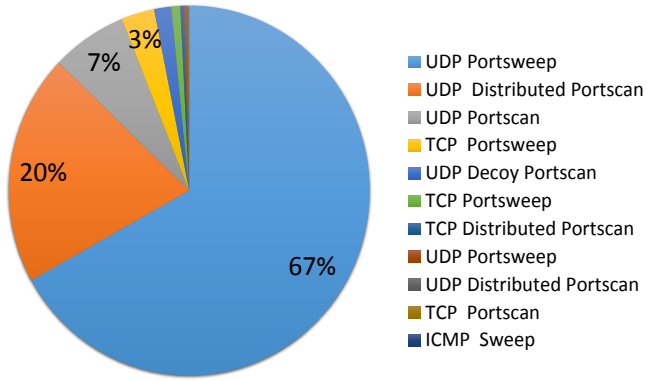

Fig. 2: Types of Probing Activities generated by Malware

probing technique; such result is coherent with other malware studies [26], where the authors revealed that UDP is the most used transport-layer protocol for malicious command-and-control communications. Further, Figure 3 illustrates the top 10 malware types that were found to trigger such probing activities. One interesting observation that could be extracted from such result is related to 'Virus.Win32.Sality.bh'; Dainotti et al. [27] have recently documented a large-scale probing campaign that was able to probe VoIP (SIP) servers of the entire IPv4 address space in 12 days. The authors pinpointed that the malware responsible for such campaign was in fact the Sality malware; the same malware that we found, using our data set, to be generating the majority of the probing activities.

We proceed in an attempt to reduce the number of malware pcaps that could be eventually attributed to the darknet probing sessions. In accordance with Section III-C, we executed the fuzzy hashing and relative entropy approach. To accomplish the former, we leveraged deeptoad[6], a fuzzy hashing implementation, while we employed matlab[7] to accomplish the latter. Subsequently, we select the top 10% (1,310) malware pcaps that were found to minimize the entropy and maximize the fuzzy hashing percentage, in comparison with the darknet probing sessions. At this point, 129 probing sessions were

[6]https://code.google.com/p/deeptoad/

[7]http://www.mathworks.com/matlabcentral/fileexchange/35625-information-theory-toolbox/content/relativeEntropy.m

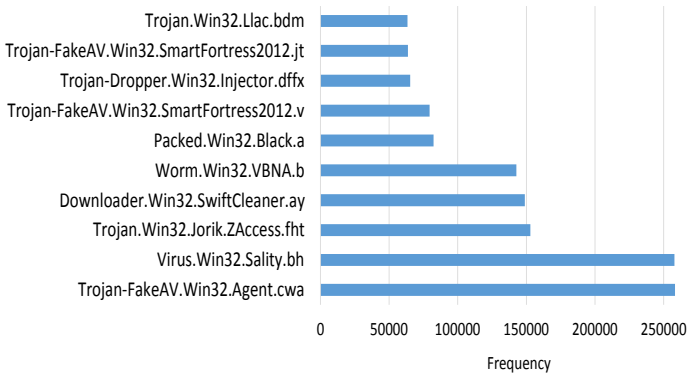

Fig. 3: Distribution of Probing Malware Types/Families

filtered out, indicating that they do not possess any malware-related behavior.

For the purpose of attributing the remaining probing sessions to a manageable and a probable set of malware pcaps, in coherence with the proposed approach of Section III-C, we proceed by executing the Bhattacharyya distance and selecting the 1% of malware pcaps (13 pcaps) that are shown to be statistically close to each of the probing sessions. Table I provides a specimen, for space limitations, that couples 5 probing sources with few of their corresponding possible set of malware infections. Intuitively, we record the complete malware outcome for all the probing sources.

Probing Source 1	Trojan-Downloader.Win32.KiayksayRen.b Trojan-FakeAV.Win32.SmartFortress2012.il Trojan.Win32.VBKrypt.hadj Trojan-Dropper.Win32.Agent.dtki
Probing Source 2	Trojan-Downloader.Win32.KiayksayRen.b Trojan-Dropper.Win32.Dapato.aflm Trojan-Dropper.Win32.Injector.dknf Trojan.Win32.Jorik.Shakblades.foc
Probing Source 3	Trojan-Downloader.Win32.KiayksayRen.b Worm.Win32.AutoIt.xl Trojan.Win32.Scar.furz Packed.Win32.PolyCrypt.d
Probing Source 4	DTrojan-Downloader.Win32.KiayksayRen.b Trojan-Downloader.Win32.Dapato.gje Trojan.Win32.Agent.btmu Virus.Win32.Sality.bh Trojan-FakeAV.Win32.SmartFortress2012.v
Probing Source 5	Trojan-Downloader.Win32.KiayksayRen.b Trojan-Spy.Win32.SpyEyes.acxb Trojan.Win32.FakeAV.lete Packed.Win32.PolyCrypt.d

TABLE I: Probing Sources coupled with their probable malware samples

Note that 'Trojan-Downloader.Win32.KiayksayRen.b', that frequently appears in the above result, has been confirmed by numerous anti-malware engines and services to be a significant sign of machine exploitation[8], particularly those running the Windows operating system. Thus, the proposed approach seems accurate and practical in pinpointing compromised machines in addition to disclosing the probable malware types that caused their contamination.

To exactly identify which malware sample is responsible for the darknet extracted probings sessions, we proceed by employing the Kolmogorov-Smirnov statistic test, as instructed in Section III-C. Table II shows the extracted insights for a sample of 10 probing sessions while Figure 4 visualizes the worldwide map of the fingerprinted infections.

Probing Source a	Trojan-Spy.Win32.VB.gt
Probing Source b	Virus.Win32.Sality.s
Probing Source c	Trojan-FakeAV.Win32.SmartFortress2012.jv
Probing Source d	Trojan-Dropper.Win32.Injector.dpdj
Probing Source e	Backdoor.Win32.Bifrose.fur
Probing Source f	Virus.Win32.Cabanas.MsgBox
Probing Source g	Trojan.Win32.Jorik.Downloader.akr
Probing Source h	Trojan-FakeAV.Win32.Agent.cwa
Probing Source i	Trojan.Win32.Swisyn.bahq
Probing Source j	Worm.Win32.Juched.djh

TABLE II: A Sample of Inferred Infections

Note that we also generate supplementary material related to the infections including geo-location information per real source (i.e., hostname), organization, ISP, city, region and country. Although, we refrain from publishing those due to sensitivity/legal issues, we can note that the infections originate from 67 diverse operational providers, 56 distinct ISPs and 33 different countries. Such results, which we postulate to be communicated to concerned providers, concur that the proposed approach possesses the capability to infer compromised machines in addition to pinpointing the exact malware type/family that was responsible for their contamination.

Although we are unable to validate the existence of every single obtained inference related to the extracted worldwide infections due to legal and logistic constraints, we have observed, from the obtained results, a number of events that advocate the accuracy and completeness of the proposed approach. First, we inferred that the majority of the infections that are related to the previously mentioned 'Virus.Win32.Sality.bh' are originating from Thailand; in [27], the authors disclosed that the bots that contributed to the large-scale VoIP probing campaign that were found to be infected by the same Sality malware, were in fact attributed to Thailand. Second, we noticed that Chinese ISPs lead in the number of generated infections. According to our results, one of the top extracted malware infections that is generated from those ISPs is the 'Trojan-Banker.Win32.Banker.adx'. This malware is a data stealing program that captures banking credentials such as account numbers and passwords from infected users. The latter insights were confirmed by McAfee in which they further concurred that China is in fact responsible from more than 45% of such contamination[9]. Third, from our results, we deduced that the malware 'Backdoor:Win32/Bifrose' was originating from a specific middle-Eastern country. The latter malware allows an external attacker to access the compromised machine to perform various malicious actions. McAfee also confirmed our findings by revealing that the same country is indeed the most contributor to such an infection[10].

[8]http://tinyurl.com/ktlqp4r

[9]http://www.mcafee.com/threat-intelligence/malware/default.aspx?id=853515

[10]http://www.mcafee.com/threat-intelligence/malware/default.aspx?id=1594628

Fig. 4: Inferred Worldwide Infections by Correlating Malware and Probing Activities

V. Approach Limitations

It is realistic to acknowledge a number of limitations in the proposed approach. First, the approach leverages the dark space to infer Internet-scale probing activities. Although the monitored space is relatively large (i.e., /13), we are unable to monitor events that do not target such space. Subsequently, the approach will be unable to correlate those "unseen" activities with malware samples, and thus will fail to detect and identify their corresponding malware infections. Second, the approach relies on malware samples that actually execute probing activities. Although, from our experiments, the number of those malware seems to be significant, the approach will not be able to detect malware that do not probe. In this case, our correlation engine could be used in conjunction with already deployed approaches, similar to those that rely on honeypots to accomplish the detection. Third, the approach is still experimental; its development is ongoing for the purpose of making it operational in an automated and a real-time fashion.

VI. Conclusion

This paper presented a new approach to infer malware-infected machines. The approach aims at providing network operators with a cyber security capability to detect their clients' compromised machines in addition to pinpointing the exact malware type that caused their contamination. The approach is efficient as it does not record or analyze the symptoms of infection. Further, it is prompt as it exploits probing activities, which are the very first indications of contamination. Moreover, the proposed approach is cost-effective as it does not require any implementation or maintenance costs at the providers' sides. To accomplish its goals, the proposed approach exploits the dark space to infer and validate Internet-scale probing activities. Consequently, it correlates such activities with malware samples by uniquely employing various statistical, fuzzy hashing and information theoretical metrics. The approach was empirically evaluated using a significant amount of real darknet and malware samples. The extracted inferences and insights revealed promising accuracy in addition to concurring that the rationale of exploiting probing activities for worldwide early malware infection detection is indeed practically viable. As for future work, we strive to leverage this work coupled with clustering mechanisms based on probing behavioral analysis in an attempt to infer malware-orchestrated campaigns.

Acknowledgment

The authors are grateful for 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

[1] A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescap. Analysis of a "/0" Stealth Scan from a Botnet. In *Internet Measurement Conference (IMC)*, Nov 2012.

[2] Jelena Mirkovic and Peter Reiher. A taxonomy of ddos attack and ddos defense mechanisms. *ACM SIGCOMM Computer Communication Review*, 34(2):39–53, 2004.

[3] M Daly. Advanced persistent threat. *Usenix, Nov*, 4, 2009.

[4] Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, and Ivan Osipkov. Spamming botnets: signatures and characteristics. In *ACM SIGCOMM Computer Communication Review*, volume 38, pages 171–182. ACM, 2008.

[5] Panda Security. Worldwide infected machines. http://tinyurl.com/o24ky8t.

[6] ScMagazine-McAfee. The state of malware in 2013. http://tinyurl.com/ohjprsc.

[7] Parliament of Canada. BILL C-28. http://tinyurl.com/avh9vzv.

[8] Shuaibu Hassan Usman. A review of responsibilities of internet service providers toward their customers' network security. *Journal of Theoretical and Applied Information Technology*, 49(1), 2013.

[9] ZDNet. ISPs accused of ignoring botnet invasion. http://tinyurl.com/lt48jzl.

[10] Koji NAKAO, Daisuke INOUE, Masashi ETO, and Katsunari YOSHIOKA. Practical correlation analysis between scan and malware profiles against zero-day attacks based on darknet monitoring. *IEICE Transactions on Information and Systems*, 92(5):787–798, may 2009.

[11] D. Inoue, M. Eto, K. Yoshioka, S. Baba, K. Suzuki, J. Nakazato, K. Ohtaka, and K. Nakao. nicter: An incident analysis system toward binding network monitoring with malware analysis. In *Information Security Threats Data Collection and Sharing, 2008. WISTDCS '08.*, pages 58–66, 2008.

[12] Jungsuk Song, Jumpei Shimamura, Masashi Eto, Daisuke Inoue, and Koji Nakao. Correlation analysis between spamming botnets and malware infected hosts. *2012 IEEE/IPSJ 12th International Symposium on Applications and the Internet*, 0:372–375, 2011.

[13] Masashi Eto, Kotaro Sonoda, Daisuke Inoue, Katsunari Yoshioka, and Koji Nakao. A proposal of malware distinction method based on scan patterns using spectrum analysis. In ChiSing Leung, Minho Lee, and JonathanH. Chan, editors, *Neural Information Processing*, volume 5864 of *Lecture Notes in Computer Science*, pages 565–572. Springer Berlin Heidelberg, 2009.

[14] Evan Cooke, Michael Bailey, Farnam Jahanian, and Richard Mortier. The dark oracle: Perspective-aware unused and unreachable address discovery. In *NSDI*, volume 6, pages 8–8, 2006.

[15] E. Bou-Harb, M. Debbabi, and C. Assi. Cyber scanning: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, PP(99):1–24, 2013.

[16] E. Bou-Harb, M. Debbabi, and C. Assi. A statistical approach for fingerprinting probing activities. In *2013 Eighth International Conference on Availability, Reliability and Security (ARES),*, pages 21–30, Sept 2013.

[17] M. Bailey, E. Cooke, F. Jahanian, J. Nazario, D. Watson, et al. The internet motion sensor: A distributed blackhole monitoring system. In *Proceedings of the 12th ISOC Symposium on Network and Distributed Systems Security (SNDSS)*, pages 167–179, 2005.

[18] C. Fachkha, E. Bou-Harb, A. Boukhtouta, S. Dinh, F. Iqbal, and M. Debbabi. Investigating the dark cyberspace: Profiling, threat-based analysis and correlation. In *2012 7th International Conference on Risk and Security of Internet and Systems (CRiSIS),*, pages 1–8, Oct 2012.

[19] Jesse Kornblum. Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation*, 3, Supplement(0):91 – 97, 2006. The Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS '06).

[20] Yo-Ping Huang, Tsun-Wei Chang, and F.-E. Sandnes. An efficient fuzzy hashing model for image retrieval. In *Fuzzy Information Processing Society, 2006. NAFIPS 2006. Annual meeting of the North American*, pages 223–228, 2006.

[21] Wenke Lee and Dong Xiang. Information-theoretic measures for anomaly detection. In *Security and Privacy, 2001. S P 2001. Proceedings. 2001 IEEE Symposium on*, pages 130–143, 2001.

[22] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology,*, 15(1):52–60, 1967.

[23] Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.

[24] Eric Wustrow, Manish Karir, Michael Bailey, Farnam Jahanian, and Geoff Huston. Internet background radiation revisited. In *Proceedings of the 10th annual conference on Internet measurement*, pages 62–74. ACM, 2010.

[25] Jayanthkumar Kannan, Jaeyeon Jung, Vern Paxson, and Can Emre Koksal. Semi-automated discovery of application session structure. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 119–132. ACM, 2006.

[26] Thomas Karagiannis, Andre Broido, Michalis Faloutsos, and Kc claffy. Transport layer identification of p2p traffic. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, IMC '04, pages 121–134, New York, NY, USA, 2004. ACM.

[27] A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescap. Analysis of a "/0" Stealth Scan from a Botnet. In *Internet Measurement Conference (IMC)*, Nov 2012.