Behavioral Analytics for Inferring Large-Scale Orchestrated Probing Events

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

***Abstract*—The significant dependence on cyberspace has indeed brought new risks that often compromise, exploit and damage invaluable data and systems. Thus, the capability to proactively infer malicious activities is of paramount importance. In this context, inferring probing events, which are commonly the first stage of any cyber attack, render a promising tactic to achieve that task. We have been receiving for the past three years 12 GB of daily malicious real darknet data (i.e., Internet traffic destined to half a million routable yet unallocated IP addresses) from more than 12 countries. This paper exploits such data to propose a novel approach that aims at capturing the behavior of the probing sources in an attempt to infer their orchestration (i.e., coordination) pattern. The latter defines a recently discovered characteristic of a new phenomenon of probing events that could be ominously leveraged to cause drastic Internet-wide and enterprise impacts as precursors of various cyber attacks. To accomplish its goals, the proposed approach leverages various signal and statistical techniques, information theoretical metrics, fuzzy approaches with real malware traffic and data mining methods. The approach is validated through one use case that arguably proves that a previously analyzed orchestrated probing event from last year is indeed still active, yet operating in a stealthy, very low rate mode. We envision that the proposed approach that is tailored towards darknet data, which is frequently, abundantly and effectively used to generate cyber threat intelligence, could be used by network security analysts, emergency response teams and/or observers of cyber events to infer large-scale orchestrated probing events for early cyber attack warning and notification.**

I. Introduction

The ever increasing growth and embracing of the Internet has been one of the greatest social and technological change of our decade. It is indeed a massive force which aims at driving economical development, reducing trade barriers, and allowing individuals, industries and governments across the world to communicate and cooperate. However, the significant dependence on cyberspace has brought new risks that often compromise, exploit and damage invaluable data and systems in ways that are extremely complex and difficult to detect or defend against. In fact, recent events demonstrated that cyberspace could be subjected to amplified, debilitating and disrupting cyber attacks leading to drastic impacts on provided network and Internet services. For instance, Google has recently been targeted by a cyber attack in which 7 of its services, including, maps, news and translator, were hacked and defaced. Further, Canada's Brandon University has lately announced that a critical online database containing students and professors sensitive information, including social insurance numbers, addresses, and salaries, has been leaked to an anonymous party. Another academic institute, namely, Michigan State University, disclosed that several high-ranked employees' direct-deposit earnings have been hijacked and redirected to unknown and untraceable bank accounts. Moreover, the African Petroleum Producers' Associate, an oil and gaz organization, suffered from a devastating cyber attack that hit its website and its operations. Another example would be numerous governmental websites of the United States, Russia, Finland, Pakistan, and Armenia that were also recently deemed as victims of cyber crime. Despite efforts to protect the cyberspace, the latest reports from senior government officials [1] highlighted that only limited progress has been made in improving the cyber security of crucial networks.

Probing, the task of scanning enterprise networks or Internet wide services, searching for vulnerabilities or ways to infiltrate IT assets, is a significant cyber security concern. The latter is due to the fact that probing is commonly the primary stage of an intrusion attempt that enables an attacker to remotely locate, target, and subsequently exploit vulnerable systems. It is basically a core technique and a facilitating factor of the above mentioned and others cyber attacks. 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 3,000 social security numbers. Further, the United States Computer Emergency Readiness Team (US-CERT) revealed that attackers had performed coordinated probing activities to fingerprint WordPress sites and consequently launched their targeted attacks. 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. 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. Thus, it is not surprising that Panjwani et al. [2] concluded that a momentous 50% of attacks against cyber systems are preceded by some form of probing activity.

Recently, there has been a noteworthy shift towards a new phenomenon of probing events. These are distinguished from previous probing incidents as (1) the population of the participating bots is several orders of magnitude larger, (2) the target scope is generally the entire Internet Protocol (IP) address space, and (3) the bots adopt well-orchestrated, often botmaster-coordinated, stealth scan strategies that maximize

978-1-4799-3088-3/14/$31.00 ©2014 IEEE

targets' coverage while minimizing redundancy and overlap [3]. Lately, Dainotti et al. [3] from the Cooperative Association for Internet Data Analysis (CAIDA) presented a pioneering measurement and analysis study of a 12-day Internet-wide orchestrated probing event targeting VoIP (SIP) servers. In a follow-up work [4], the same authors admitted that they have detected the reported event including the malware responsible for its actions (i.e., Sality malware) "serendipitously" (i.e., luckily and accidentally) while analyzing a totally unrelated phenomenon. They also stated that since *currently there exist no cyber security capability to discover such large-scale orchestrated probing events*, other similar events targeting diverse Internet and organizational infrastructure are going undetected. In another inquisitive, well executed work, an "anonymous" presented and published online [5] what they dubbed as the "Carna Botnet". The author exploited poorly protected Internet devices, developed and distributed a custom binary, to generate one of the largest and most comprehensive IPv4 census ever.

The aforementioned two events differ on various key observations. The work by Dainotti et al. disclosed that the bots were recruited into the probing botnet by means of a new-generation malware while the Carna Botnet was augmented using a custom code binary. Moreover, Dainotti et al. discovered that the bots were coordinated by a botmaster in a Command-and-Control (C&C) infrastructure where the bots used a reverse IP-sequential strategy to perform their probing, while the Carna Botnet was C&C-less and its bots used an interleaving permutation method to scan its targets. Further, the work by Dainotti et al. documented a horizontal scan that targeted world-wide SIP servers, while the Carna Botnet did not focus on one specific service but rather attempted to retrieve any available information that was associated with any host and/or service. Readers that are interested in more details related to the discussed events are kindly referred to [3, 5].

In this paper, we propose a generic approach that aims at identifying orchestrated probing activities, similar to the two previously mentioned events. The approach achieves its goal by capturing the behavior of the probing sources by employing a set of novel behavioral analytics that scrutinize darknet data. Subsequently, the sources with similar behaviors are automatically clustered to infer such orchestrated incidents. Indeed, the capability to infer such events could be leveraged for early cyber attack warning and notification.

The remaining of this paper is organized as follows. The next section elaborates on the proposed approach. Specifically, it pinpoints (1) how probing events are extracted from darknet data, (2) the set of behavioral analytics that capture the machinery of the probing sources and (3) how the automatic clustering and event inference is accomplished. In Section III, we present and discuss the use case that validates the proposed approach. Section IV briefly highlights the related work. Finally, Section V concludes this paper and paves the way for future work.

II. Proposed Approach

We possess real darknet data that we are receiving on a daily basis since three years from a trusted third party. The data is around 12 GB per day. Such traffic originates from the Internet and is destined to numerous /13 network sensors. The darknet sensors cover more than 12 countries and monitor around half a million dark IPs. 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 serivice 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 [6]. In the subsequent, we elaborate on the proposed approach that scrutinize such darknet data in an attempt to infer large-scale orchestrated probing events. The approach is divided into three main parts, namely, probing extraction, probing behavioral modeling and event inference.

A. Probing Extraction

In [7], we proposed a new approach to fingerprint probing activities from darknet data. 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. 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 [7]. 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.

B. Probing Behavioral Modeling

One of the main contributions of this paper is rendered by the following set of behavioral analytics that aim at capturing the machinery of the probing sources. The proposed approach takes as input the previously extracted probing sessions 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 [8], 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 [3, 5], 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 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 [8]. 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 towards 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.

Are the probing sources infected? On one hand, the authors of [3] pinpointed that the probing bots were infected by the Sality malware and were coordinated in a C&C infrastructure. On the other hand, the author of [5] did not employ any specific malware to recruit the probing bots. Thus, it is of momentous importance to infer whether or not the probing sources are infected by a malware and if they are, which exact malware type/family/variant is causing the probing. The latter insight would be an added-value inference, for this work, for the purpose of correlating the sources into an orchestrated event, as well as, for future work, for the analysis of the pinpointed malware binary for the sake of understanding its inner workings and perhaps inferring its C&C servers. In an attempt to answer this question, we proceed as follows. We do receive, on a daily basis, malware samples from ThreatTrack Security (formerly GFI). We operate a dynamic malware analysis module that is based on the GFI sandbox environment (i.e., controlled environment). After receiving the 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 60 thousand unique and recent (June 2012 to September 2013) malware samples as packet captures (pcaps). The pcaps contain one-way 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. To investigate if the probing sessions that were initially fingerprinted as such demonstrate any signs of malware infection, 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 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; 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 relative entropy, between the given probing session and the malware pcaps. 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 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 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 between those and the given probing session. 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 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. In summary, the outcome of the aforementioned approach is whether or not the probing sources demonstrate signs of malware infection and if they do, which exact (or probable) malware type/family/variant is responsible for their probing.

Miscellaneous Inferences: For each probing source, we also record its rate (packets/second), its ratio of destination overlaps defined as $r = \frac{nc}{nt}$ where nc defines the number of common sessions between all the sources and nt is total number of all probing sessions, and its target ports.

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.

C. Event Inference

Previous works [9] suggested that coordinated bots within a botnet campaign probe their targets in a similar fashion. The approach exploits this idea by automatically building patterns that consist of similar probing behavioral characteristics. The latter aim at identifying and correlating the probing sources into an orchestrated probing event. Currently, the approach considers the criteria (i.e., features) that are summarized in Table I. The authors of [3] pinpointed that the probing event covered 86.6% of their monitored dark IP space. The approach considers the % of dark IP space coverage. The latter condition is a configurable criterion where its rationale states that if the correlated bots are part of an orchestrated event, then they might cover a certain percentage of the dark IP space. Inferred from Section II-A, the employed probing technique is a significant behavior; [3] demonstrated that all the probing bots used UDP scanning. Further, the approach consumes all the previously inferred probing machinery that is derived from the behavioral analytics. It considers the fashion of the generated probing traffic; whether random or not, and which specific pattern has been adopted if not random; which probing strategy has been employed; whether the probing source is a probing tool or a bot; whether the probing is targeted or dispersed; whether or not the probing sources demonstrate any signs of malware infection and which specific malware type/family/variant if they do. Additionally, bots/sources in an orchestrated probing event are postulated to possess similar probing rates and ratios of destination overlaps; we consider a 90% confidence interval as being similar. Finally, the approach considers the target port as a significant criterion; [3] disclosed that all the probing bots used port 5060.

% of dark IP space coverage
Employed probing technique
Probing traffic (Random Vs Patterns)
Employed pattern
Adopted probing strategy
Nature of probing source
Type of probing (Targeted Vs Dispersed)
Signs of malware infection
Exact malware type/variant
Probing rate
Ratio of destination overlaps
Target port

TABLE I: Criteria adopted by the Proposed Approach

Another main contribution of this paper is to automatically infer orchestrated probing events. To achieve that, the approach leverages all the previously extracted inferences and insights related to the probing sessions/sources to build and parse a Frequent Pattern (FP) tree. In such a tree, each node after the root represents a feature extracted from the probing sessions, which is shared by the sub-trees beneath. Each path in the tree represents sets of features that co-occur in the sessions, in non-increasing order of frequency of occurrences. Thus, two sessions that have several frequent features in common and are different just on infrequent features will share a common path in the tree. The proposed approach also employs the FP tree based mining method, FP-growth, for mining the complete set of generated frequent patterns. As an outcome, the generated patterns represent frequent and similar probing behavioral characteristics that correlate the probing sources into orchestrated probing events. We should emphasize that such an approach possess the following advantages. First, the generated patterns are not defined *a priori;* they are naturally and automatically detected. This permits the inference of novel orchestrated probing events, without requiring previous knowledge about their specifications. The latter is a crucial feature, especially with the continuous evolution of such

events and their employed strategies. Second, the FP-Tree not only provides the capability for the system to correlate the probing sources into orchestrated events, but also semantically describes how the probing sessions have been constructed. Third, by engineering parsing algorithms that traverse the FP-Tree in various ways, the system can infer horizontal probing campaigns, similar to the probing event in [3], as well as orchestrated probing events that do not focus on one port but rather probe multiple targets on various ports, similar to the event in [5]. Fourth, from a performance perceptive, the employed approach is scalable since probing sessions are not compared pairwise, which could lead to a quadratic complexity. In fact, the cost of the algorithm is the cost of inserting probing session features in the FP-Tree, which is linear.

III. Evaluation and Validation

We evaluate the proposed approach using 330 GB of darknet data extracted from the month of June, 2013. We execute the proposed approach in coherence with the details of Sections II-A, II-B and II-C. We visualize the outcome of the behavioral analytics as depicted in Figure 1. Such 'flower-based' result intuitively and creatively illustrate how the FP-tree is constructed. Recall, that the tree depicts frequent probing features that co-occur in the probing sessions, which are generated by the probing behavioral analytics from Section II-B. One can notice several groupings or clusters that depict probing events sharing various common machinery. For the sake of this work, we have devised a parsing algorithm that automatically build patterns from the FP-tree that aim at capturing orchestrated probing events that probe horizontally; probe all IPs by focusing on specific ports.

A. Case Study

The proposed approach automatically inferred the pattern that is summarized in Table II. The pattern permitted the detection, identification and correlation of 846 unique probing bots into a well-defined orchestrated probing event that targeted the VoIP (SIP) service. It is shown that this event, that was initiated on the 17^{th} of June, adopted UDP scanning, probed around 65% of the monitored dark space (i.e., 300 thousand dark IPs) where all its bots did not follow a certain pattern when generating their probing traffic. Further, the results demonstrate that the bots employed a reverse IP-sequential probing strategy when probing their targets. Moreover, the malware responsible for this event was shown to be attributed to the Sality malware.

B. Validation

Since currently there exist no cyber security capability to discover such large-scale orchestrated probing events (i.e., lack of ground truth), we are unable to directly compare the inferred event with other systems or approaches. However, in an attempt to validate our results, we resort to publicly accessible data that is provided by DShield/Internet Storm Center (ISC). ISC data comprises of millions of intrusion detection log entries that is gathered daily from sensors covering more than 500 thousand IP addresses in over 50 countries. From such data, we extract Figure 2 that depicts probe counts generating probing activities towards UDP port 5060 during the month of June, 2013. Figure 2 indeed reveals a significant peak on the 17^{th} of June consisting of an increase number of probe counts targeting the SIP service; this is the same day where the orchestrated event, that was previously inferred by the proposed approach, was detected to be targeting the SIP service. The latter strongly advocate that the proposed approach was indeed accurately successful in inferring that unusual event. Note that, this inferred event went undetected and unreported in the cyber security community until now.

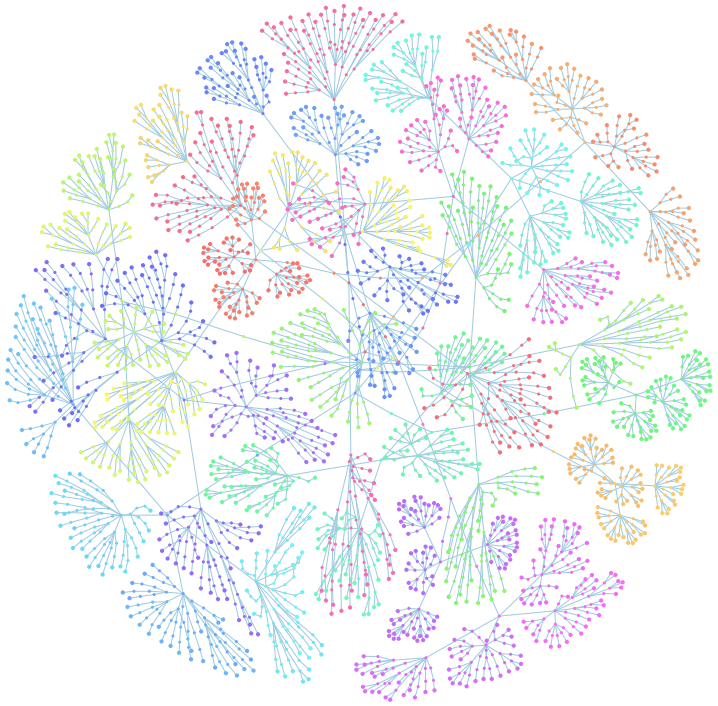

Fig. 1: Visualization of the outcome of the probing behavioral analytics in the FP-tree

Employed probing technique: **UDP**
Probing traffic (Random Vs Patterns): **Random**
Employed pattern: **Null**
Adopted probing strategy: **Reverse IP-sequential**
Nature of probing source: **Bot**
Type of probing (Targeted Vs Dispersed): **Dispersed**
Signs of malware infection: **Yes**
Exact malware type/variant:**Virus.Win32.Sality.bh**
Probing rate: **12 pps**
Target port: **5060**

TABLE II: The automatically inferred pattern capturing a large-scale orchestrated probing event

C. Discussion

The aforementioned event is indeed interesting; as elaborated at the beginning of this paper, in [3], CAIDA presented a measurement and analysis study of an orchestrated probing event that targeted VoIP (SIP) servers. According to CAIDA, the event occurred from January 31^{st}, 2011 till February 12^{th}, 2012. The event that the proposed approach was able to

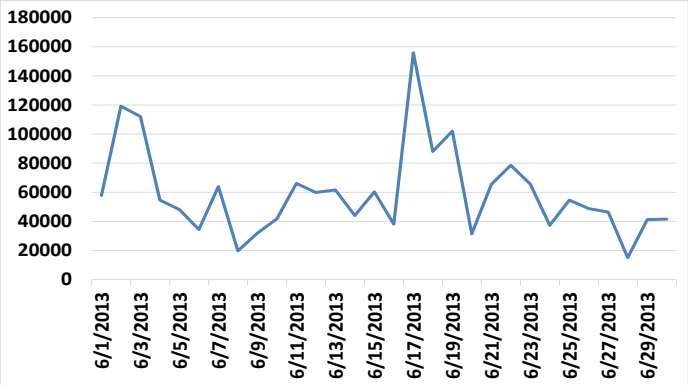

Fig. 2: Probe Counts extracted from DShield/ISC Data (June 2013): Probes Targeting the SIP Service

automatically infer possesses the exact characteristics of that CAIDA event. They both were attributed to the Sality malware, generated UDP scanning, targeted SIP servers on port 5060, and used a reverse IP-sequential probing strategy. The latter is predominantly stimulating because that strategy is known to be extremely under-employed [10]. However, one distinguishing feature between those two events is that the probing bots of the event that the proposed approach was able to infer has considerably lower probing rate than those of CAIDA's event; on average, the bots of the inferred event recorded 12 packets per seconds while those of CAIDA measured around 60 packets per second. Such information (1) arguably proves that CAIDA's event from last year is indeed still active, yet operating in a stealthy, very low rate mode in an attempt to achieve its reconnaissance task without being detected or (2) a new instance of the same orchestrated event commanded by the same C&C took place in June without any cyber security body reporting it. In either cases, we find the latter information motivating and to a certain degree puzzling. Thus, we aim in the near future work to track that event to verify and elaborate on its inner details.

IV. Related Work

In this section, we briefly highlight on some related work in various concerned topics. In the area of extracting probing events, Li et al. [8] extracted such events from darknet traffic using time series analysis. They further executed manual analysis and visualization techniques to extract the rough boundaries of such events. In the context of analyzing probing events, the authors of [8] presented an analysis that drew upon extensive honeynet data to explore the prevalence of different types of scanning activities. Additionally, they designed mathematical and observational schemes to extrapolate the global properties of scanning events including total population and target scope. In the area of probing measurement studies, in addition to [3, 5], Benoit et al. [11] presented the world's first Web census while Heidemann et al. [12] were among the first to survey edge hosts in the visible Internet. Further, Pryadkin et al. [13] offered an empirical evaluation of IP address space occupancy whereas Cui and Stolfo [14] presented a quantitative analysis of the insecurity of embedded network devices obtained from a wide-area scan. Last but not least, a number of botnet detection systems have been proposed in the literature such as [15]. Some of those investigates specific channels, others might require deep packet inspection or training periods, while the majority depends on malware infections and/or attack life-cycles. To the best of our knowledge and as stated in [4], the capability to infer large-scale orchestrated probing events does not exist, rendering the proposed approach as a novel contribution.

V. Conclusion

This paper investigates a new phenomena of probing events that is rendered by their orchestration characteristic. To tackle this, we elaborated, evaluated and validated a new approach that exploits significant amount of darknet data to infer such malicious activities. At the core of this approach is a set of behavioral analytics that scrutinize such darknet data to capture the machinery the probing sources. The case study highlighted on one event that took place earlier this year that was never reported in the operational cyber security community nor in the literature. We are currently continuing to develop the proposed approach to make it operational in a real-time fashion. We envision that this approach could be easily leveraged by any security operator or observer of cyber events that deal with darknet data, such as CAIDA for instance, for early cyber attack warning and notification as well as for simplified analysis and tracking of such events.

References

[1] Action Plan 2010-2015 for Canada's Cyber Security Strategy. http://tinyurl.com/plxmua8.
[2] Panjwani, S et al. An experimental evaluation to determine if port scans are precursors to an attack. In *DSN*, 2005.
[3] Dainotti, A et al. Analysis of a "/0" Stealth Scan from a Botnet. In *IMC*, Nov 2012.
[4] Dainotti, A et al. Analysis of internet-wide probing using darknets. In *ACM BADGERS*, 2012.
[5] Internet Census 2012-Port scanning /0 using insecure embedded devices. http://tinyurl.com/c8af8lt.
[6] Bailey, M. et al. The internet motion sensor: A distributed blackhole monitoring system. In *SNDSS*, 2005.
[7] Bou-Harb, E. et al. A statistical approach for fingerprinting probing activities. In *IEEE ARES*, 2013.
[8] Li, Zhichun et al. Towards situational awareness of large-scale botnet probing events. *IEEE TIFS*, 2011.
[9] Abu Rajab et al. A multifaceted approach to understanding the botnet phenomenon. In *IMC*, 2006.
[10] Derek Leonard and Dmitri Loguinov. Demystifying service discovery: implementing an internet-wide scanner. IMC '10. ACM, 2010.
[11] Benoit, Darcy et al. World's first web census. *International Journal of Web Information Systems*, 2007.
[12] Heidemann, John et al. Census and survey of the visible internet. In *IMC*, 2008.
[13] Pryadkin, Y et al. An empirical evaluation of ip address space occupancy. *USC/ISI-TR*, 2004.
[14] Cui, Ang et al. A quantitative analysis of the insecurity of embedded network devices: results of a wide-area scan. In *ACSAC*, 2010.
[15] Gu, Guofei et al. Bothunter: detecting malware infection through ids-driven dialog correlation. In *USENIX Security Symposium*, 2007.