Behavioral Service Graphs: A Big Data Approach for Prompt Investigation of Internet-wide Infections

Elias Bou-Harb[1], Mark Scanlon[2] and Claude Fachkha[3]

[1]Cyber Threat Intelligence Laboratory, Florida Atlantic University, USA
ebouharb@fau.edu
[2]School of Computer Science, University College Dublin, Ireland
mark.scanlon@ucd.ie
[3]Cyber Security Center, New York University, Abu Dhabi
cf92@nyu.edu

***Abstract*—The task of generating network-based evidence to support network forensic investigation is becoming increasingly prominent. Undoubtedly, such evidence is significantly imperative as it not only can be used to diagnose and respond to various network-related issues (i.e., performance bottlenecks, routing issues, etc.) but more importantly, can be leveraged to infer and further investigate network security intrusions and infections. In this context, this paper proposes a proactive approach that aims at generating accurate and actionable network-based evidence related to groups of compromised network machines. The approach is envisioned to guide investigators to promptly pinpoint such malicious groups for possible immediate mitigation as well as empowering network and digital forensic specialists to further examine those machines using auxiliary collected data or extracted digital artifacts. On one hand, the promptness of the approach is successfully achieved by monitoring and correlating perceived probing activities, which are typically the very first signs of an infection or misdemeanors. On the other hand, the generated evidence is accurate as it is based on an anomaly inference that fuses big data behavioral analytics in conjunction with formal graph theoretical concepts. We evaluate the proposed approach as a global capability in a security operations center. The empirical evaluations, which employ 80 GB of real darknet traffic, indeed demonstrates the accuracy, effectiveness and simplicity of the generated network-based evidence.**

***Index Terms*—Probing, Infections, Graphs, Threat modeling, Big data analytics, Network forensics**

I. Introduction

Undeniably, network forensics presents a rich problem space that typically deals with the collection, preservation, analysis and presentation of network-based knowledge. It is often exploited to generate actionable insights and intelligence that could be effectively leveraged by investigators. The latter is especially factual when attempting to fingerprint, assess and mitigate network security intrusions and misdemeanors. However, this attempt is recurrently hindered by various current technical challenges that face network forensics. First, network forensic analysts are significantly overwhelmed by huge amounts of low quality evidence. Such evidence is often generated from intrusion detection systems that are known to suffer from elevated levels of both false positives and negatives [1], rendering the combined task of identifying relevant information and attributing the true malicious entity extremely challenging, if not impossible. Second, most network forensic approaches are passive or reactive, employ manual ad-hoc methods and are strenuously time consuming [2, 3]. This makes the generated evidence relatively obsolete to be acted upon in a timely manner and most certainty decreases its reliability and wastes valuable resources. Third, contemporary cyber attacks are getting more sophisticated than ever and continue to operate in an excessively coordinated and distributed manner. To this end, network forensic science is relatively lagging behind such advancement in the attacks. Further, most current network forensic practices do not support distributed inference, and if they do, they force the analysts to go through an error-prone agonizing process of correlating dispersed unstructured evidence to infer a specific security incident.

In essence, the presented research and development work attempts to answer the following question: How can we design an approach that is able to effectively process, analyze and correlate large volumes of network traffic to generate, in a very prompt manner, formal, highly-accurate and actionable network forensic evidence that could be leveraged to infer infected machines, and simultaneously possesses the capability to practically operate in different deployment scenarios?

This paper attempts to answer the above. Specifically, the major contributions of this paper could be summarized in the following:

- Presenting *Behavioral Service Graphs*, a novel approach that aims at providing investigators/analysts, network administrators and/or security operators with network forensic evidence related to infected machines. The approach models the probing sources that show evidence of infection as graphs. By exploiting ancillary graph theoretic concepts such as the maximum spanning tree and Erdős-Rényi random graphs, the approach is able to infer and

978-1-5090-2914-3/16/$31.00 ©2016 IEEE

correlate distributed groups of infected machines (i.e., campaigns). The approach is prompt since it exploits probing activities to rapidly infer infection. Further, the inferred group of infected machines possesses the minimum number of members to formally claim that such group is a campaign. This idea is especially imperative as this will allow actionable thwarting of campaigns as soon as there exist evidence of their construction.
- Empirically evaluating the proposed approach using a significant real dataset. The output concur that the extracted inferences exhibit noteworthy accuracy and can generate significant, accurate and formal forensic insights that could be used for prompt mitigation and to facilitate further analysis.

The road-map of this paper is as follows. In the next section, we survey the related work on various concerned topics. In Section III, we elaborate on the proposed approach. In Section IV, we empirically evaluate the proposed approach and verity its accuracy and insights. Advantages, limitations and possible improvements related to the proposed approach are highlighted in Section V. Finally, concluding remarks are provided in Section VI.

II. Related Work

In this section, we briefly review the related work on two topics, namely, anomaly detection using graphs and big data forensic approaches.

Anomaly detection using graphs: Wang et al. [4] approached the problem of anomaly detection as a change-point hypothesis constructed on a time series of graphs. The authors proposed a stochastic model that is based on the use of scan statistics; metrics that can extract normal traffic and compare it to anomalous traffic. Their model was evaluated and validated on real email data. In an another work, Brdiczka et al. [5] proposed an approach for proactive detection of insider threats by combining structural anomaly detection from social and information networks, and psychological profiling of individuals. Their approach is specifically tailored to detect anomalies in multi-player online games. In a different work, Hassanzadeh et al. [6] proposed a framework for analyzing the effectiveness of various graph theoretic properties in detecting anomalous users on online social networks. Their empirical evaluations demonstrated that such derived properties are indeed accurate in modeling anomalous behaviors. Further, Ding et al. [7] employed bipartite graph representation of network flow traffic coupled with community detection techniques in an attempt to infer malicious sources. To achieve such a task, the authors further employed hard thresholds and heuristics derived from empirical evaluations.

Big data forensic approaches: The author in [8] explored the challenges of big data as applied to digital investigation. The proposed approach elaborated on how techniques and algorithms that are typically used in big data analysis could possibly be adapted to the unique context of digital forensics. The author discussed various approaches ranging from managing evidence via Map-Reduce to machine learning techniques, and analysis of big forensic disk images and network traffic dumps. In an alternate work, Zhu [9] proposed a big data iterative algorithm for discovering network attack patterns via a feedback mechanism. The author claimed that the algorithm is fully unsupervised as it does not require apriori user-defined thresholds. Simulations were conducted to validate the accuracy of the proposed approach. Moreover, in [10], Garfinkel presented numerous lessons learned from writing digital forensic tools and managing a 30 TB digital evidence corpus. Specifically, the author elaborated on the technical difficulties analyzing such data, the possible hardware and software issues that could be faced, and how to accurately retain the extracted evidence. The author concluded by stating some present issues related to big data forensic approaches, namely, diversity of data that needs to be analyzed, the size of the data sets, and the mismatch between the technical skills of investigators and the difficulty level of the work.

The proposed approach of this work is unlike the above two categories as (1) it tackles a different problem rendered by inferring Internet-wide infections, (2) uniquely scrutinizes probing activities using a set of behavioral analytics to infer infections, (3) employs a new concept of similarity service graphs to infer campaigns of infected machines, (4) exclusively exploits graph theoretic notions such as the maximum spanning tree and Erdős-Rényi random graphs to infer the niche of the infected campaign and (5) it has been empirically evaluated using a real dataset in a unique deployment scenario.

III. Proposed Approach

In this section, we describe and detail the rationale and employed steps of the proposed approach. In summary, the proposed approach (1) fingerprints and extracts probing activities from perceived network traffic, (2) applies the proposed behavioral analytics to generate feature vectors related to the infected probing sources, (3) constructs Behavioral Service Graphs that model those probing machines and (4) manipulates such graphs to infer distributed campaigns possessing minimum members of infected machines. The latter four steps are detailed next.

A. Fingerprinting Probing Activities

Motivated by the fact that probing activities precede the majority of attacks [11] coupled with the rationale that such activities are the very first signs of any infection [12], the proposed approach leverages the latter to extract probing activities generated from infected machines. To achieve this, we leverage our previously proposed approach as detailed in [13].

B. Big Data Behavioral Analytics

In order to capture the behaviors of the inferred probing sources, we exploit our previously proposed big data behav-

ioral analytics as described in [14]. Such analytics take as input the previously extracted probing sessions and outputs a series of behavioral characteristics related to the probing sources. Such characteristics include the nature of the probing sources, their embedded traffic patterns, their techniques used to probe their destinations, in addition to miscellaneous inferences such as their rates, probing ports and destinations' overlap.

C. Behavioral Service Graphs

We model the probing machines that show signs of infection (i.e., those inferred as bots using the behavioral analytics) coupled with their feature vectors using what we refer to as Behavioral Service Graphs. Such graphs are of the form $G = (N, E)$ where N represents the set of infected probing sources/machines (i.e., nodes) and E characterizes the edges between such nodes. It is worthy to mention that G is an undirected complete graph [15], with weights on the edges representing the probability of behavioral similarity (P_{bs}) computed by piecewise comparisons between the previously inferred feature vectors of each of the nodes.

Another feature of such graphs is that they are designed to provide additional forensic evidence related to what service is being probed. The service is rendered by the destination port number inferred from the detected probing packets. Hence, the word 'Service' in Behavioral Service Graphs. Therefore, in essence, each constructed graph is actually modeling infected machines, their behavioral similarity and what specific network service is being probed. This aims at providing the investigator with additional inference about the activities of the current infections and to warn about possible future attacks that could specifically abuse that service.

In summary, Behavioral Service Graphs allow the prompt inference of bot infected machines by solely analyzing their probing activities. Further, they extend such inferences to automate the amalgamation of evidence from distributed entities as well as providing auxiliary valuable insights related to the behaviors of the infected machines and their possible intended actions.

D. Friends of the enemy stay closely connected: Inferring Infected Campaigns

Previous work [16] demonstrated that coordinated bots within a campaign probe their targets in a similar fashion. Indeed, Behavioral Service Graphs were initially engineered to naturally and intuitively support the latter; they cluster the infected machines targeting the same service and they combine their feature vectors (and their similarly probability) for further analysis. The proposed approach executes two steps to retrieve the minimum number of infected machines to deem a group of infected machines as a campaign.

First, given a complete Behavioral Service Graph $G = (N, E)$, the approach extracts a subgraph $G' = (N', E')$ where $N' = N$ and $E' \subseteq E$. This aims at reducing the number of edges while maximizing the behavior probability between the infected machines (i.e., nodes). To achieve this task, we employ the graph theoretic concept of a maximum spanning tree [17] by implementing a slightly modified version of Kruskal's algorithm [18]. Although there exists a plethora of approaches for the creation of maximum spanning trees, this algorithm was the basis of many and is abundantly available in numerous tool sets.

Second, to understand the structure of the subgraph formed by members of a campaign on a Behavioral Service Graph, suppose that there are m bots (i.e., infected machines) in the network, and therefore there are m corresponding nodes on the graph. Let the set $X = \{X_1, X_2, \cdots, X_m\}$ denote these nodes and P_e denote the probability of having an edge between any given X_i and X_j, for $i \neq j$ where $1 \leq i \leq m$ and $1 \leq j \leq m$. Since P_e would exist with an equal and a random probability given any pair of of X_i and X_j, the subgraph formed by the nodes X_1, X_2, $\cdots$, X_m on a Behavioral Service Graph is indeed an Erdős-Rényi random graph [19, 20], where each possible edge in the graph possesses an equal probability of being created.

One interesting property shown by Erdős and Rényi is that, Erdős-Rényi graphs have a sharp threshold of edge probability for graph connectivity [20]. Simplified, if the edge-probability is greater than such threshold, then all of the nodes produced by such a model will be strongly connected. Erdős and Rényi have shown that the sharp connectivity threshold is $th_s = \frac{ln\theta}{\theta}$, where θ is the number of nodes in the graph. The proposed approach exploits this neat graph theoretical property; given the previously extracted maximum spanning tree subgraph, the approach eliminates all nodes/edges whose bot-edge probability (i.e., behavioral similarity P_{bs}) is less than th_s, deeming the rest of the nodes, given such formal forensic evidence, as the minimum number of infected machines forming a campaign.

In conclusion, according to the random peer selection model, the niche members of the same infected campaign are expected to be closely connected to each other on a subgraph extracted from Behavioral Service Graphs.

IV. Empirical Evaluation

In this section, we port the approach to a global scale and elaborate on how it can be employed to monitor, infer and distribute Internet-scale forensic intelligence. Thus, in this scenario, the approach is envisioned to operate in a model similar to what is dubbed as a global Security Operation Center (SOC). Typically, such operational centers have access to significant various real-time and raw data streams from around the globe. They often exploit such data for analysis, correlation and generation of intelligence that would be distributed to concerned parties for alert and mitigation purposes. Such centers were initially formed as global independent entities to combat an increasing trend of

external (in contrary to internal) threats and attacks.

Thus, in this scenario, Behavioral Service Graphs are postulated to be deployed as an additional forensic capability in one of those SOC centers. In this context, we operate the scheme by investigating darknet data. In a nutshell, a darknet (also commonly referred to as a network telescope) is a set of routable and allocated yet unused IP addresses [21]. It represents a partial view of the entire Internet address space. From a design perspective, a darknet is transparent and indistinguishable compared with the rest of the Internet space. From a deployment perspective, it is rendered by network sensors that are implemented and dispersed on numerous strategic points throughout the Internet. Such sensors are often distributed and are typically hosted by various global entities, including Internet Service Provides (ISPs), academic and research facilities and backbone networks. The aim of a darknet is to provide a lens on Internet-wide malicious traffic; since darknet IP addresses are unused, any traffic targeting them represents anomalous unsolicited traffic. Such traffic (i.e., darknet data) could be leveraged to generate various cyber threat intelligence, including inferences and insights related to probing activities; some of the probes of an infected machine, while probing the Internet space, will also hit the darknet and thus will be subsequently captured. Recall, that the probing machine, while spraying its probes, can not avoid the darknet as it does have any knowledge about its existence. Further, it is extremely rare if not impossible for a probing source to have any capability dedicated to such avoidance [22]. To this end, we utilize real darknet data that is provided by the Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) program.

A. The ground truth

There exists a need to have a concrete knowledge about a ground truth to properly evaluate the proposed scheme. For this purpose, in this scenario, we rely on a reported Internet-scale event related to a large-scale probing campaign. Particularly, on October 10, 2012, the Internet Storm Center (ISC) received a report of a probing campaign targeting Internet SQL servers. This incident was also interestingly corroborated by Dshield. Dshield data comprises of millions of intrusion detection log entries gathered daily from sensors covering more than 500,000 IP addresses in over 50 countries. Further, the ISC report noted that the probing campaign involved more than 9,000 distributed sources which aim at exploiting that service. We rely on the occurrence of such disclosed incident as the ground truth as we proceed in our evaluation.

B. Evaluation

From our darknet data repository, we extract one week of data retaining to the period of October 4^{th} to October 11^{th}, 2012. The aim is to employ the proposed approach on such data to evaluate the scheme's capability and effectiveness in disclosing insights related to that reported campaign.

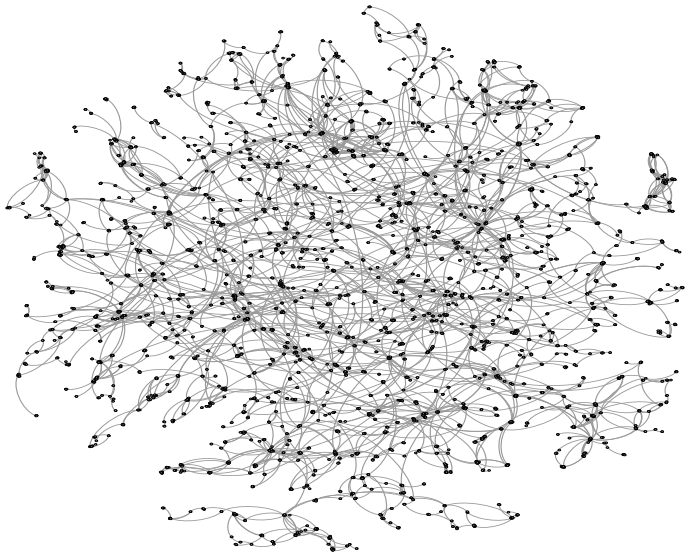

Fig. 1: The proposed approach revealing the bots of the SQL probing campaign

By executing the proposed approach on the extracted probing traffic from our darknet dataset, the outcome demonstrated that one of the Behavioral Service Graphs was indeed able to infer and correlate around 800 unique sources targeting the SQL service. Further, the behavioral analytics (1) showed strong behavioral similarity between those sources and (2) inferred that those sources were indeed bots, thus providing strong evidence that such campaign was triggered from Internet-wide infected machines. The latter inferred bots could be visualized as in Figure 1. Additionally, it might be interesting to mention that the proposed approach deemed 84 bots as the niche of the campaign by leveraging the approach of Section III-D.

Thus, provided with such forensic evidence, SOC analysts can demand an immediate take-down of those 84 bots to limit the expansion of such campaign on the global Internet. Additionally, they can promptly notify concerned parties to employ mitigation approaches against the abuse of SQL servers.

V. Proposed Approach: Advantages, Limitations & Possible Improvements

Other than generating prompt and formal forensic evidence, the proposed approach can also be deemed as being simple. We define simple as being (1) reliable, (2) cost-effective, (3) highly-performant and (4) stackable. Indeed, the proposed approach is reliable as it repetitively yielded accurate results in different deployment scenarios under numerous experimental setups. Further, it demonstrated scalability characteristics and still maintained precision. Further, the approach is cost-effective as it possesses the capability of operating on commodity machines without requiring any additional

hardware or software utilities. This idea is particularly true when the approach is executed as a SOC capability providing forensic intelligence to other Internet enterprises, alleviating the latter from the burden of implementation scenarios and their corresponding supplementary costs. We also deem the proposed approach as highly-performant. In fact, it takes only several minutes to build Behavioral Service Graphs coupled with their corresponding subgraphs, and to infer the niche of the infected campaigns. Last but not least, the approach could be used as a building block (as input) for other approaches or as a complementary scheme to provide auxiliary forensic evidence.

However, it is definitely realistic to acknowledge several limitations of the proposed approach. First, although the approach analyzes probing activities by leveraging behavioral analytics to infer enterprise and Internet-wide bots, there exists a need to further fortify the infection evidence. To this end, we are currently devising an approach that would correlate perceived probing activities from such bots with malware samples to corroborate the infection evidence as well as to attribute the inferred infected machines to a specific malware family. Second, currently, the proposed approach of Section III-D infers the niche of the campaign by heuristically selecting the nodes/edges that possess a similarity behavior above a threshold indicated by the Erdős-Rényi random graphs. It would be interesting to find a formal mathematical computation to infer the number of such nodes as a function of the overall campaign nodes and the similarity behavior. Finally, the proposed approach is still experimental. We are working on rendering it operational in real-time in several deployment scenarios.

VI. Concluding Remarks

We have devised Behavioral Service Graphs, an approach that is able to effectively process, analyze and correlate large volumes of network traffic to generate, in a very prompt manner, formal, highly-accurate and actionable network forensic evidence that could be leveraged by investigators to infer Internet-wide infected machines. Rigorous empirical evaluations with real data under a SOC deployment scenario indeed verified the accuracy and effectiveness of the approach. We hope that the forensic community could consider the approach as a building block for complementary analysis and investigation. As for future work, other than tackling the issues mentioned in Section V, we are working on campaign analysis; the ability to infer what the probing infected bots will eventually execute after finalizing their probing activities. We aim to achieve the latter by correlating the generated inferences from this work with other data sources (i.e., passive DNS, public intrusion and firewall logs, etc.). These future objectives ultimately aim at providing extended network-based evidence to further support investigations.

References

[1] Pedro Garcia-Teodoro, J Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1):18–28, 2009.

[2] Emmanuel S Pilli, Ramesh C Joshi, and Rajdeep Niyogi. Network forensic frameworks: Survey and research challenges. *Digital Investigation*, 7(1):14–27, 2010.

[3] Ikuesan R Adeyemi, Shukor Abd Razak, and Nor Amira Nor Azhan. A review of current research in network forensic analysis. *International Journal of Digital Crime and Forensics (IJDCF)*, 5(1):1–26, 2013.

[4] Heng Wang, Minh Tang, Y. Park, and C.E. Priebe. Locality statistics for anomaly detection in time series of graphs. *Signal Processing, IEEE Transactions on*, 62(3):703–717, Feb 2014.

[5] O et al. Brdiczka. Proactive insider threat detection through graph learning. In *IEEE Symposium on Security and Privacy Workshops*, pages 142–149, May 2012.

[6] Reza et al. Hassanzadeh. Analyzing the effectiveness of graph metrics for anomaly detection in online social networks. In *Web Information Systems Engineering - WISE 2012*, volume 7651, pages 624–630. Springer Berlin Heidelberg, 2012.

[7] Qi et al. Ding. Intrusion as (anti)social communication: Characterization and detection. In *18th ACM SIGKDD*, pages 886–894, 2012.

[8] Alessandro Guarino. Digital forensics as a big data challenge. In Helmut Reimer, Norbert Pohlmann, and Wolfgang Schneider, editors, *ISSE 2013 Securing Electronic Business Processes*, pages 197–203. Springer Fachmedien Wiesbaden, 2013.

[9] Ying Zhu. Attack pattern discovery in forensic investigation of network attacks. *Selected Areas in Communications, IEEE Journal on*, 29(7):1349–1357, August 2011.

[10] Simson Garfinkel. Lessons learned writing digital forensics tools and managing a 30tb digital evidence corpus. *Digital Investigation*, 9, Supplement(0):S80 – S89, 2012. The Proceedings of the Twelfth Annual {DFRWS} Conference 12th Annual Digital Forensics Research Conference.

[11] Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. Cyber scanning: a comprehensive survey. *IEEE Communications Surveys & Tutorials*, 16(3):1496–1519, 2014.

[12] David Whyte, Evangelos Kranakis, and Paul C van Oorschot. Dns-based detection of scanning worms in an enterprise network. In *NDSS*, 2006.

[13] Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. On fingerprinting probing activities. *Computers & Security*, 43:35–48, 2014.

[14] Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. Behavioral analytics for inferring large-scale orchestrated probing events. In *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*, pages 506–511. IEEE, 2014.

[15] Josep Díaz, Jordi Petit, and Maria Serna. A survey of graph layout problems. *ACM Computing Surveys (CSUR)*, 34(3):313–356, 2002.

[16] Moheeb Abu Rajab, Jay Zarfoss, Fabian Monrose, and Andreas Terzis. A multifaceted approach to understanding the botnet phenomenon. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 41–52. ACM, 2006.

[17] Kenta Ozeki and Tomoki Yamashita. Spanning trees: A survey. *Graphs and Combinatorics*, 27(1):1–26, 2011.

[18] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

[19] Paul Erdős and Alfréd Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[20] Paul Erd6s and A Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[21] David Moore, Colleen Shannon, Geoffrey M Voelker, and Stefan Savage. *Network telescopes: Technical report*. Department of Computer Science and Engineering, University of California, San Diego, 2004.

[22] 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.