Digital Investigation 20 (2017) S47–S55

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2017 Europe — Proceedings of the Fourth Annual DFRWS Europe

Behavioral Service Graphs: A formal data-driven approach for prompt investigation of enterprise and internet-wide infections

Elias Bou-Harb [a, *], Mark Scanlon [b]

[a] Cyber Threat Intelligence Laboratory, Florida Atlantic University, USA
[b] School of Computer Science, University College Dublin, Ireland

ARTICLE INFO

Article history:
Received 31 January 2017
Accepted 31 January 2017

Keywords:
Probing
Infections
Graphs
Threat modeling
Data analytics
Network forensics

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 (i.e., campaigns). 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 data behavioral analytics in conjunction with formal graph theoretic concepts. We evaluate the proposed approach in two deployment scenarios, namely, as an enterprise edge engine and as a global capability in a security operations center model. The empirical evaluation that employs 10 GB of real botnet traffic and 80 GB of real darknet traffic indeed demonstrates the accuracy, effectiveness and simplicity of the generated network-based evidence.

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 (Garcia-Teodoro et al., 2009), 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 time consuming (Pilli et al., 2010; Adeyemi et al., 2013). This makes the generated evidence relatively obsolete to be acted upon in a timely manner and most certainly 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.

Indeed, local and Internet-scale networks have been increasingly getting abused by various modernized attacks, including, distributed denial of service attacks (Fu et al., 2012), amplification

* Corresponding author.
E-mail address: ebouharb@fau.edu (E. Bou-Harb).

http://dx.doi.org/10.1016/j.diin.2017.02.002

attempts (Kührer et al., 2014), spamming (Xie et al., 2008) and advanced persistent threats (Daly, 2009). Such attacks are almost always being generated by groups of infected and distributed machines controlled by an external entity (Silva et al., 2013). In this paper, we refer to the latter orchestrated malicious groups as 'campaigns'. An effective approach to generate network forensic insights and inferences related to those campaigns is to analyze their generated probing activities. Such activities refer to reconnaissance techniques that are typically employed by those campaigns to obtain information about their targets prior to launching their targeted attacks (Allman et al., 2007). In fact, Panjwani et al. (2008) concluded that around 50% of attacks are indeed preceded by some form of probing activity. Additionally, such activity has been reported in numerous occasions as a concrete evidence of infection (Wang et al., 2014a; Whyte et al., 2006).

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 campaigns?*

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

- Proposing a set of data behavioral analytics that scrutinize perceived probing activities to capture their various obscured features (i.e., machinery, strategies, natures, etc.). The analytics uniquely employ numerous statistical and entropy-based techniques to effectively generate feature vectors related to the infected probing sources or hosts.
- 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 within a constructed campaign. 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 correlate such distributed groups of infected machines. The approach is prompt since (1) it exploits probing activities to rapidly infer infections and (2) the inferred group of infected machines possesses the minimum number of members to formally claim that such group is indeed a malicious campaign. The latter idea is especially imperative as this will allow actionable thwarting of campaigns as soon as there exists evidence of their construction.
- Empirically evaluating the proposed approach using two real and significant datasets under two different deployments scenarios. The output concurs 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 focused analysis.

The road-map of this paper is as follows. In the next section, we elaborate on the proposed approach. Specifically, we disclose the data preprocessing step, the employment of the data behavioral analytics, the rationale and construction of Behavioral Service Graphs and detail how they can be exploited to achieve the intended goals. In Section Empirical evaluation, we empirically evaluate the proposed approach and verity its accuracy and insights. We provide a discussion related to the approach, its limitations and possible improvements in Section Proposed approach: limitations & possible improvements. In Section Related work, we survey the related work on various concerned topics. Finally, Section Concluding remarks summarizes the goals, the methods and the results of the proposed approach and paves the way for future work that aims at providing extended network-based evidence to further support investigations.

Proposed approach

In this section, we describe and detail the rationale and employed steps of the proposed approach. In a nutshell, 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.

Fingerprinting probing activities

Motivated by the fact that probing activities precede a plethora of attacks (Allman et al., 2007; Panjwani et al., 2008) coupled with the rationale that such activities are the very first signs of any infection (Wang et al., 2014a; Whyte et al., 2006), the proposed approach leverages the latter to extract probing activities generated from infected machines. The intrusion detection system community provides extensive techniques on how to accomplish this task (Bhuyan et al., 2011). In this work, to successfully and accurately fingerprint probing activities, we leverage the work by Staniford et al. (2002) and cross match the output, for validation purposes, by using two open-source detection systems, namely, Snort (Roesch et al, 1999) and Bro (Paxson, 1999). We have selected to employ the latter three approaches as they are the de-facto standards when it comes to probing detection, possess the capability to operate in real-time, and have been extensively and repetitively evaluated and validated. The output of this step is probing traffic, generated from unique sources, coupled with their network sessions that have been saved in packet capture (.pcap) format for further analysis.

Data behavioral analytics

In order to capture the behaviors of the inferred probing sources, we propose to employ the following set of behavioral analytics. This aims at generating the feature vectors of the infected probing machines to be employed as input for the subsequent steps. 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 execute their probing traffic, it is imperative to infer and capture the fashion in which they achieve their goal. To realize this task, we proceed as follows. For each unique pair of hosts extracted from the probing sessions (probing source to target), we test for randomness of their inter-arrival times in the traffic using the non-parametric Wald-Wolfowitz statistic test. If the outcome is positive, we record it for that precise probing source and apply the test for the remaining probing sessions. If the result is negative, we conclude that the generated traffic follows a certain pattern. To deduce the particular employed pattern, we model the probing traffic as a Poisson distribution and capture the maximum likelihood estimates for the Poisson parameter λ that corresponds to that traffic, at a 95% confidence level. The choice to model the traffic as a Poisson process is motivated by our previous work (Bou-