Computer Networks 94 (2016) 327–343

Contents lists available at ScienceDirect

Computer Networks

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

A novel cyber security capability: Inferring Internet-scale infections by correlating malware and probing activities

Elias Bou-Harb*, Mourad Debbabi, Chadi Assi

The National Cyber Forensics and Training Alliance (NCFTA) Canada & Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, Canada

ARTICLE INFO

Article history:
Received 17 December 2014
Revised 24 October 2015
Accepted 5 November 2015
Available online 10 November 2015

Keywords:
Probing
Malware
Darknet preprocessing
Big data correlation
Cyber security
Cyber intelligence

ABSTRACT

This paper presents a new approach to infer worldwide 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. 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 initially filters out misconfiguration traffic targeting such space using a probabilistic model. Subsequently, the approach employs statistical methods to infer large-scale probing activities as perceived by the dark space. Consequently, such activities are correlated with malware samples by leveraging fuzzy hashing and entropy based techniques. The proposed approach is empirically evaluated using a recent 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, which were validated using publically available data resources, demonstrate that the extracted inferences exhibit noteworthy accuracy and can generate significant cyber security insights that could be used for effective mitigation.

1. 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 protect the cyberspace from diverse debilitating, intimidating and disrupting cyber threats, such space continues 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

* Corresponding author. Tel.: +1 5146495049.
E-mail address: e_bouh@encs.concordia.ca (E. Bou-Harb).

http://dx.doi.org/10.1016/j.comnet.2015.11.004
1389-1286/

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, this last 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]. This paper, which extends our previous work [10], elaborates on such a cyber security capability that satisfies the mentioned requirements. Specifically, we frame the paper's contributions as follows:

- Proposing a new probabilistic model to preprocess telescope/darknet data to prepare it for effective use. The aim is to fingerprint darknet misconfiguration traffic and subsequently filter it out. The model is advantageous as it does not rely on arbitrary cut-off thresholds, provide separate likelihood models to distinguish between misconfiguration and other malicious darknet traffic and is independent from the nature of the source of the traffic.
- 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 IP addresses, which permits the observation and identification of worldwide probing activities and thus 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 a recent 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 3, 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 4. In Section 5, we pinpoint some limitations of the proposed approach. Finally, concluding remarks and some future work are disclosed in Section 6.

2. 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. Additionally, we pinpoint several proposed methods for inferring worm infections. Finally, we present several research efforts in the areas of active detection of malicious machines as well as malware signature generation.

Nakao et al. [11] 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 [12] 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. [13] 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