Data & Knowledge Engineering xxx (2018) 1–23

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

A heuristics approach to mine behavioural data logs in mobile malware detection system

Giang Nguyen [a,*], Binh Minh Nguyen [b], Dang Tran [b], Ladislav Hluchy [a]

[a] Institute of Informatics, Slovak Academy of Sciences, Dubravska cesta 9, 845 07 Bratislava, Slovakia
[b] School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam

ARTICLE INFO

Keywords:
Mobile security
Situational awareness
Anomaly detection
Incremental machine learning
Natural language processing
Scalable solution design

ABSTRACT

Nowadays, in the era of Internet of Things when everything is connected via the Internet, the number of mobile devices has risen exponentially up to billions around the world. In line with this increase, the volume of data generated is enormous and has attracted malefactors who do ill deeds to others. For hackers, one of the popular threads to mobile devices is to spread malware. These actions are very difficult to prevent because the application installation and configuration rights are set by owners, who usually have very low knowledge or do not care about the security. In this study, our aim is to improve security in the environment of mobile devices by proposing a novel system to detect malware intrusions automatically. Our solution is based on modelling user behaviours and applying the heuristic analysis approach to mobile logs generated during the device operation process. Although behaviours of individual users have a significant impact on the social cyber-security, to achieve the user awareness has still remained one of the major challenges today. For this task, there is proposed a light-weight semantic formalization in the form of physical and logical taxonomy for classifying the collected raw log data. Then a set of techniques is used, like sliding windows, lemmatization, feature selection, term weighting, and so on, to process data. Meanwhile, malware detection tasks are performed based on incremental machine learning mechanisms, because of the potential complexity of this tasks. The solution is developed in the manner to allow the scalability with several blocks that cover pre-processing raw collected logs from mobile devices, automatically creating datasets for machine learning methods, using the best selected model for detecting suspicious activity surrounding malware intrusions, and supporting decision making using a predictive risk factor. We experimented cautiously with the proposal and achieved test results confirm the effectiveness and feasibility of the proposed system in applying to the large-scale mobile environment.

1. Introduction

Today, the popularity of low cost technologies as well as mobile devices coupled with the arrival of social media, social networks, and cloud computing having multiple connectivity possibilities pose a threat for unattended, misconfigured devices and they are even potentially vulnerable for others. We are facing a high increase of new zero-day vulnerabilities, personal records that were stolen or lost, phishing campaigns targeting end-users, malware attacks, fake technical support scams like url-based spreading threats, etc. Parsons et al. [1]; Thread Track Security [2]; and ENISA [3]. Several new effective functions for such threats could be listed here, they

* Corresponding author.
E-mail addresses: giang.ui@savba.sk (G. Nguyen), minhnb@soict.hust.edu.vn (B.M. Nguyen), dangtv18@gmail.com (D. Tran), hluchy.ui@savba.sk (L. Hluchy).

http://dx.doi.org/10.1016/j.datak.2018.03.002
Received 8 July 2017; Received in revised form 5 January 2018; Accepted 8 March 2018
Available online XXX

include anonymization strategies, strong encryption involving HTTPS, flexible key management schemes, and obfuscation methods for the payload detection. However, the truth is that users are sharing more and more information electronically, and cyber-criminals are getting better, and the current anti-phishing solution threat intelligence might not be the sufficient protection.

Although the amount of malware is rising continuously, it is possible to be categorized into three main groups, as follows: back-door (1), security hole (2), and distributed denial of service (DDOS) (3). While the malware of (2) does not generate abnormal logs, and malware of (3) often can be prevented effectively using defined rules (i.e. static analyses), the malware of (1) creates many challenges for computer scientists because it is lurking inside victim's devices for a long time, and it can automatically open communications with outside lacking the user authorization. In this study, we concentrate on settling the security issue related to the back-door malware belonging to the group (1), which is based on mobile operation logs. The critically desired outcome of this work is to detect precisely, as much as possible, the presence of malware intrusion into mobile devices.

At the starting point of the realization we assumed optimal conditions, where the activity posture of all mobile devices operated in a certain network needs to be monitored securely in real-time. Besides, the common and normal situational awareness should be provided to minimize security risks. Regarding the security issue, it can be seen that modelling behaviours of individual users is one of the most important factor. Nevertheless, this task has still remained to be one of the major challenges today in term of the degree of accuracy.

In this direction of making use of technology supporting different aspects of peoples life, one way how to significantly improve the security issue effectiveness, is the incorporation of innovative models and new technologies to understand better the human behaviours on mobile devices. As a result, the consequent trigger action from a detection system should be at least the creation of precise recommendations or warnings Ricci et al. [4]; Tam et al. [5] for users, or the autocratical isolation of previously infected devices from the network in critical situations.

The main goal of our work is to improve the cyber-resilience focused on the user behaviour on mobile devices by adopting the heuristic approach. Thus, our solution enables to automatically detect and create alerts when a malware attacks such devices. There are various difficulties to deal with this problem. First, collected logs belong to the human-generated data classes and they have a high capacity considering the volume, velocity, variety and veracity characteristics. They also contain a huge number of data features and are extremely noisy e.g. over time duplicated information, data with evolving specific characteristics for the detection purpose. Diverse exploratory data analysis techniques are applied to eliminate noise and inaccuracy in the collected logs in order to enhance the data quality and to find patterns that do not conform to the expected normal behaviour. Next, the suspicious activities often have low occurrences that cause imbalanced data for machine learning. So, in working with data processing mechanisms, we also have to select effective methods e.g. support vector machine (SVM), logistic regression (LR) and artificial neural network (NN) to mine data. Choosing the best model for the collected data is another important effort and contribution in our work. Last, the proposed solution offers the scalability and can support the extensive data analysis as well as new model developments with cases of the data increase in the near future.

The paper is organized as follows. In Section 2, related research works are categorized and discussed to highlight the differences and contributions of our work in comparison with the existing efforts. Section 3 describes raw log properties collected from mobile devices. The logs show that there are many challenges to process data because of its complexity. The architecture design for our detection system is presented in Section 4, and it shows the flexibility in the composition through the integrability of block components. We also present the usage of specific mechanisms, like sliding windows, feature processing, lemmatization, term weighting, and so forth, to deal with the raw data. To find out an anomalous situation, we apply different ML methods that are also introduced here. At the end of this section, the scalable deployment is described in detail, where parallel and incremental learning techniques are explored thoroughly. The potential results achieved during experiments and their evaluations are depicted in Section 5 in detail. The conclusion and future works are presented in the last section.

2. Related work

Cyber-security is defined as a set of technologies and processes designed to protect computers, networks, programs and data, against attacks, unauthorized accesses and changes, or destruction Mukkamala et al. [6] and Dua and Du [7]. For example, anti-virus software and intrusion detection systems (IDSs) will help to discover, determine, and identify the unauthorized activity, duplications, alterations, and destructions of the information system. Currently, one of the attractive research topics within the cyber-security field is the malware detection in computing devices. Usually, to recognize threats, data analysis techniques are widely used and also presented in many recent works, such as Buczak and Guven [8] and Tam et al. [5]. Roughly, the techniques which are exploited to detect suspicious software can be divided into the following categories: dynamic, static and heuristic.

The principle of *dynamic analysis techniques* is to test suspect software in a controlled environment (e.g. computers, mobile devices or specific networks) using different tools such as debuggers, process monitors, package sniffers, sandboxes, behaviour and influence monitoring. However, modern threats even have the ability to detect artificial testing environments and become dormant Tam et al. [5] to wait for other harm opportunities.

In *static analysis techniques*, information about programs or their behaviour consists of explicit and implicit observations which are stored in binary/source code. The most well-noted approach is the misuse or signature-based mechanism that represents a common and effective method used by antivirus software. The mechanism relies on the identification of unique signatures, which point clearly at suspicious pieces of software if they occur during operation process. Though this approach offers the fast and effective mechanism, it still has some limitations. Concretely, it is not able to detect a new one, zero-day and obfuscated threats/malware. For example, metamorphic viruses can change their internal structure by the time. This dangerous feature enables to form effective means to avoid