Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

# Screening smartphone applications using malware family signatures

*Jehyun Lee, Suyeon Lee, Heejo Lee*[*]

*Department of Computer Science and Engineering, Korea University, Seoul 136-713, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 1 November 2014
Received in revised form
15 January 2015
Accepted 10 February 2015
Available online xxx

*Keywords:*
Smartphone security
Android
Malware
Variant detection
Static analysis
Family signature

ABSTRACT

The sharp increase in smartphone malware has become one of the most serious security problems. Since the Android platform has taken the dominant position in smartphone popularity, the number of Android malware has grown correspondingly and represents critical threat to the smartphone users. This rise in malware is primarily attributable to the occurrence of variants of existing malware. A set of variants stem from one malware can be considered as one malware family, and malware families cover more than half of the Android malware population. A conventional technique for defeating malware is the use of signature matching which is efficient from a time perspective but not very practical because of its lack of robustness against the malware variants. As a counter approach for handling the issue of variants behavior analysis techniques have been proposed but require extensive time and resources. In this paper, we propose an Android malware detection mechanism that uses automated family signature extraction and family signature matching. Key concept of the mechanism is to extract a set of family representative binary patterns from evaluated family members as a signature and to classify each set of variants into a malware family via an estimation of similarity to the signatures. The proposed family signature and detection mechanism offers more flexible variant detection than does the legacy signature matching, which is strictly dependent on the presence of a specific string. Furthermore, compared with the previous behavior analysis techniques considering family detection, the proposed family signature has higher detection accuracy without the need for the significant overhead of data and control flow analysis. Using the proposed signature, we can detect new variants of known malware efficiently and accurately by static matching. We evaluated our mechanism with 5846 real world Android malware samples belonging to 48 families collected in April 2014 at an anti-virus company; experimental results showed that; our mechanism achieved greater than 97% accuracy in detection of variants. We also demonstrated that the mechanism has a linear time complexity with the number of target applications.

## 1. Introduction

Smart devices are currently facing a serious threat posed by the surge in malware. The smartphone has become the most popular target for malware writers since it contains a great deal of user information and has mobile billing capability. The Android platform in particular, occupying the dominant position in smartphone market share (Mawston, 2014), accounted for 97% of all mobile malware in 2013, as reported by F-Secure (Aquilino et al., 2014). Recently, Android malware has increasingly

[*] *Corresponding author*. Tel.: +82 2 3290 3638.
E-mail addresses: arondit@korea.ac.kr (J. Lee), suyeonl@korea.ac.kr (S. Lee), heejo@korea.ac.kr (H. Lee).

http://dx.doi.org/10.1016/j.cose.2015.02.003
0167-4048/© 2015 Elsevier Ltd. All rights reserved.

adopted several obfuscation techniques such as metamorphism and repackaging in order to avoid detection or recognition by the user. This trend is confirmed in a 2014 report by Symantec (Wood et al., 2014) in 2014, which observes that Android malware authors focus more of their efforts on improving existing malware than on creating new malware. Indeed, the top ten Android malware families made up 76% of all Android malware reported during the first quarter of 2014 (F-Secure Labs, 2014). Compared with desktop malware, smartphone malware can cause a more direct invasion of privacy and greater potential for economic damage to users. However, the flood of Android malware variants on smartphones hampers development of efficient strategies for dealing with malware attack. Accordingly, a mechanism that prevents malware by efficiently filtering variants of known malware, is needed to retain smartphone security and user privacy.

Previous approaches for variant detection based on behavior analysis are not suitable for identifying the malware family to which a detected malware variant belongs. These approaches detect variants by assessing the similarity of behaviors such as the frequency or sequence of application programming interface (API) calls (Kwon and Lee, 2012; Aafer et al., 2013; Deshotels et al., 2014), code semantics (Crussell et al., 2013; Suarez-Tangil et al., 2014), and commonly shared byte or string patterns and strings (Faruki et al., 2013a, 2013b; Sanz et al., 2013a, 2013b, 2014), to those of known malware. Extracting and comparing behaviors from large numbers of target executables requires heavy computing overhead. Detection based on behavior similarity is a useful method for covering unknown malware, but it requires disassembling, behavior feature modeling, and complex clustering algorithms to the every inspection target applications. The behavior analysis approaches which use the dynamic analysis methods (Enck et al., 2010; Gilbert et al., 2011; Yan and Yin, 2012) cover more sophisticated variants which are hard to detect by the static analysis methods, but the number of inspection targets need to be reduced by a complement method before the heavy analysis.

The alternative approach often employed by vendors of antivirus (AV) software using a representative signature is effective in defining and detecting malware families. In contrast to the behavior-based approaches, it is also efficient in terms of time and space complexity. However, the signatures not only have narrow detection coverage of a malware family due to strict decision conditions and naive evidence such as Android application package (APK) names and single class and method name, but also are easily defeated by malware that adopts code obfuscation such as repackaging and metamorphism. Hence, we conclude that a reinvestigation of the overall code for behavior analysis and the construction of an additional signature for a slight modulation of malware are each inefficient ways to improve malware detection when considered against the small effort that is consumed in making a variant.

In this paper, we propose an Android malware detection mechanism that screens new variants of a known malware family out. Most Android malware is a manipulated version of existing malware, and in malware belonging to the same family large portions of code and resources remain unchanged. Exploiting this feature, we detect variants efficiently and accurately by analyzing the representative parts of a family. The proposed mechanism uses a family signature that is common to the family and excludes other families via a weighting factor. The proposed signature structure consists of four parts extracted from a *Dalvik executable* (*DEX*) file, which is the executable file within an APK. The signature consists of the names classes, methods, character strings, and method bodies. A signature has multiple entries in each part, and each entity has an associated weight, according to how well the signature entity represents the identity of the malware family. In other words, the class and method names, hard-coded character strings, and reused codes that appear commonly in family members and rarely in other families have a higher weight.

In experimental evaluation, our mechanism showed high detection performance and low time consumption for variant detection. We evaluated our mechanism using 14,120 Android malware samples collected at an anti-virus company in 2014. These included 5846 family malware belonging to 48 families and 8274 samples of individual malware that were not part of any family or were part of a small family. For our evaluation, we preprocessed and refined malware families and their memberships since different AV vendors that had investigated the malware had given them different family labels. In the experiments using family signatures, our mechanism showed **97% detection accuracy**, with greater than **97% of recall performance** in the Monte Carlo validation. For individual malware detection performance, we compared the family signatures with the individual malware samples; our mechanism detected **1820 (22%)** of the malware samples out of the 8274 individuals in the malware set. This result shows that our mechanism can detect the malware manipulated in various ways such as the modifying package name, class name and part of codes, and code reordering while conventional signature-based approaches are not ordinarily able to detect such variants. Finally, in the scalability evaluation, the family signatures needed only 20 MB to cover the approximately 8000 Android malware samples. In terms of time consumption, the hashed signature matching process took only 10 s on average to screen a thousand applications against a million signature entries in a desktop PC.

Our contribution is twofold:

- We propose a type of Android malware family signature that can be used for accurately and efficiently detecting variants of known malware families. A family signature is a flexible signature for a malware family sharing the class name, method name, character strings, and code bodies of the original malware. It solves an existing malware detection issue by multiplexing decision conditions with multiple signature entries along with their representative weights. This contribution makes it possible to detect malware including the variants, even variants that have adopted an evasion technique such as metamorphism or code modification.
- We have reduced the number of signatures needed. The family signature represents a malware family by a single signature set covering a large number of family members, including newly appeared variants. The family signature consists of binary patterns and character strings shared by members of the same malware family. By estimating similarity to the various known families, our mechanism detects and classifies malware families to a practical