Accepted Manuscript

Title: A pragmatic android malware detection procedure

Author: Paolo Palumbo, Luiza Sayfullina, Dmitriy Komashinskiy, Emil Eirola, Juha Karhunen

PII: S0167-4048(17)30154-2
DOI: http://dx.doi.org/doi: 10.1016/j.cose.2017.07.013
Reference: COSE 1179

To appear in: *Computers & Security*

Received date: 12-9-2016
Revised date: 4-7-2017
Accepted date: 27-7-2017

Please cite this article as: Paolo Palumbo, Luiza Sayfullina, Dmitriy Komashinskiy, Emil Eirola, Juha Karhunen, A pragmatic android malware detection procedure, *Computers & Security* (2017), http://dx.doi.org/doi: 10.1016/j.cose.2017.07.013.

A Pragmatic Android Malware Detection Procedure

Paolo Palumbo[1], Luiza Sayfullina[2], Dmitriy Komashinskiy[1], Emil Eirola[3], and Juha Karhunen[2]

[1] F-Secure Corporation, Helsinki, Finland

[2] Department of Information and Computer Science, Aalto University, Finland

[3] Arcada University of Applied Sciences, Helsinki, Finland

Abstract. The academic security research community has studied the Android malware detection problem extensively. Machine learning methods proposed in previous work typically achieve high reported detection performance on fixed datasets. Some of them also report reasonably fast prediction times. However, most of them are not suitable for real-world deployment because requirements for malware detection go beyond these figures of merit.

In this paper, we introduce several important requirements for deploying Android malware detection systems in the real world. One such requirement is that candidate approaches should be tested against a stream of continuously evolving data. Such streams of evolving data represent the continuous flow of unknown file objects received for categorization, and provide more reliable and realistic estimate of detection performance once deployed in a production environment.

As a case study we designed and implemented an ensemble approach for automatic Android malware detection that meets the real-world requirements we identified. Atomic Naive Bayes classifiers used as inputs for the Support Vector Machine ensemble are based on different APK feature categories, providing fast speed and additional reliability against the attackers due to diversification. Our case study with several malware families showed that different families are detected by different atomic classifiers. To the best of our knowledge, our work contains the first publicly available results generated against evolving data streams of nearly 1 million samples with a model trained over a massive sample set of 120,000 samples.

Keywords: android, malware detection, static analysis, machine learning, classification, ensemble learning, feature selection

1 Introduction

The importance of the Android platform in the mobile operating system space is a well-known fact. During the first half of 2015, the Android platform represented 49.47% of the total mobile operating system market according to data provided by NetMarketShare [1], making it the most