Lightweight, Obfuscation-Resilient Detection and Family Identification of Android Malware

Joshua Garcia, Mahmoud Hammad, and Sam Malek
Department of Informatics
University of California, Irvine
Irvine, California, USA
{joshug4, hammadm, malek}@uci.edu

ABSTRACT

The number of malicious Android apps has been and continues to increase rapidly. These malware can damage or alter other files or settings, install additional applications, obfuscate their behaviors, propagate quickly, and so on. To identify and handle such malware, a security analyst can significantly benefit from identifying the family to which a malicious app belongs rather than only detecting if an app is malicious. To address these challenges, we present a novel machine learning-based Android malware detection and family-identification approach, RevealDroid, that operates without the need to perform complex program analyses or extract large sets of features. RevealDroid's selected features leverage categorized Android API usage, reflection-based features, and features from native binaries of apps. We assess RevealDroid for accuracy, efficiency, and obfuscation resilience using a large dataset consisting of more than 54,000 malicious and benign apps. Our experiments show that RevealDroid achieves an accuracy of 98% in detection of malware and an accuracy of 95% in determination of their families. We further demonstrate RevealDroid's superiority against state-of-the-art approaches. [URL of original paper: https://dl.acm.org/citation.cfm?id=3162625]

KEYWORDS

Android malware, obfuscation, machine learning, lightweight, native code, reflection

1 EXTENDED ABSTRACT

The number of malicious Android apps is increasing rapidly. Android malware can damage or alter other files or settings, install additional applications, etc. To determine such behaviors, a security analyst can significantly benefit from identifying the family to which an Android malware belongs, rather than only detecting if an app is malicious. Techniques for detecting Android malware, and determining their families, lack the ability to handle certain obfuscations that aim to thwart detection. Moreover, some prior techniques face scalability issues, preventing them from detecting malware in a timely manner.

In this paper, we introduce RevealDroid, a lightweight machine learning-based approach for detecting malicious Android apps and identifying their families. RevealDroid leverages a set of features selected to achieve obfuscation resiliency, efficiency of analysis, and accuracy. It does not require complex program analyses (e.g., data-flow analysis) or large sets of features (e.g., hundreds of thousands of features), which can lead to scalability problems. More specifically, our selected machine-learning features are based on Android-API usage, including resolution of APIs invoked using reflection, and function calls (e.g., system calls) made by native binaries within an Android app. No previous work has included native-code feature extraction to detect malware. Including features based on reflection and native code significantly aids RevealDroid with achieving obfuscation resiliency.

RevealDroid is capable of accurately detecting malicious apps with a 98% accuracy, and identifying their families with a 95% accuracy, in under 90 seconds on average. RevealDroid can maintain high accuracy even for obfuscated apps. We evaluate RevealDroid's detection and family identification accuracy by comparing its ability to correctly identify malware and classify its family on a dataset of over 24,600 benign apps and over 30,000 malicious apps from two different malware repositories. We further compare RevealDroid's detection and family-identification accuracy against state-of-the-research approaches: Adagio, Drebin, and MUDFLOW, all of which are approaches for malware detection; and Dendroid, an approach for malware-family identification. RevealDroid has an overall greater accuracy by about 11%-25% and mislabels 25%-54% fewer benign apps as malicious than MUDFLOW; RevealDroid achieves up to 23% greater accuracy than Adagio and up to 60% greater accuracy than Drebin. Additionally, RevealDroid achieves a 24%-70% higher classification rate than Dendroid.

This paper makes the following contributions:

- RevealDroid demonstrates that highly lightweight analyses that extract API-based features—including those based on reflection—and native code features combined with machine learning, can achieve high accuracy, scalability, and obfuscation resiliency.
- We construct an updated dataset of over 27,900 malware apps labeled with their 447 malware families and assess RevealDroid's family-identification accuracy on that dataset. We make this updated dataset available online for researchers and practitioners.
- To evaluate RevealDroid's obfuscation resiliency, we apply several transformations to malware apps in order to obfuscate them and assess its ability to detect and identify families of those transformed apps. Using these transformed apps, we compare RevealDroid's accuracy for detection against Adagio, Drebin, and MUDFLOW, and for family identification against Dendroid. We also make the transformed dataset available online.
- We assess the efficiency of RevealDroid's feature extraction and machine-learning classification. We show that RevealDroid's features can be extracted, on average, in under 90 seconds—while still exhibiting obfuscation resiliency and accuracy. We further demonstrate that RevealDroid can produce classifiers efficiently, as compared to other state-of-the-research tools.

ICSE '18, May 27-June 3, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5638-1/18/05.
https://doi.org/10.1145/3180155.3182551