Open Android - An Android Feature Database for Data Scientists

University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING

UNH Cyber Forensics Research & Education Group
Robert Schmicker, Dr. Frank Breitinger, Dr. Ibrahim Baggili

UNHcFREG

Motivation

With a **market share** of around **80%**, it is needless to mention that the **Android operating system** is and will be a **popular target** for **malware** developers. On the other hand, much **research** is going on in order to find reliable ways to **identify Android malware** / suspicious applications.

Currently there are **three major obstacles** in the Android malware detection community:

- Data Scientists have **difficulty comparing** Android malware detection **algorithms** due to data scientists using **different data sets**
- **Data Scientists** must take the **role of a programmer** and malware security / forensics researcher
- Data Scientists have **difficulty obtaining** Android **malware samples**

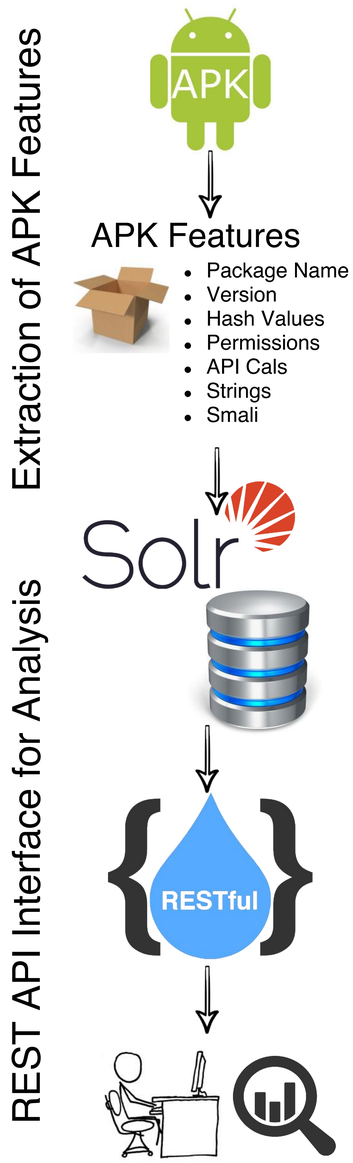
Our work provides:

- Open data set of **7000+** Android **APK's** features
- **REST API** for data scientists to use
- Open source extraction **tool kit** for features

Key Findings

- Android features can be difficult to extract to those who have minimal programming capabilities
- Android malware is difficult to obtain for research purposes
- At this current time, no other feature set exists for Android malware detection by machine learning

Process



Malicious and benign APK samples

JSON files are created from features

JSON files are indexed through Solr for fast searching

Using Solr's built in REST API, data scientists can download their feature sets

Data scientists can now access features without programming

Results

Extraction
- JSON file for every APK
- Time – 4 hours for 7,000 APK's

Solr
- Indexing engine for JSON files
- Allows for extremely fast searching through over 100GB of data

REST API
- Available to the public
- Time – 14 minutes for 7,000 APK's
- Saves data scientists time not having to extract information themselves

Data Scientists
- No longer need programming for features
- Central data set for easier algorithm comparison
- Malware features can now be shared safely

Future Work

- Further evaluation of viability against the manual solution must be examined
- Allowing the source code for the Open Android tool kit to be available to the public

Acknowledgements: We would like to thank Zhaoheng Yang and Xiaolu Zhang for having their tools open source and participating in the creation of the Open Android tool kit.