METADroid

Android application classification based on meta-data

Yan Li, Natalia Stakhanova

Faculty of Computer Science, University of New Brunswick

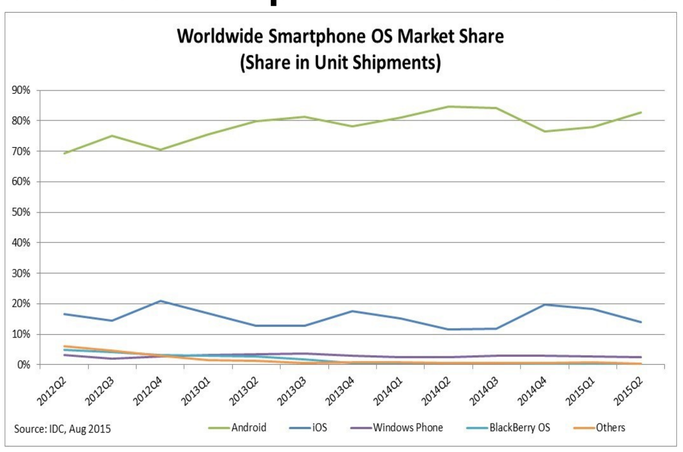
ISCX Information Security Centre of Excellence

UNB

Problem

- Malware targeting Android devices dominates mobile malware market.
- **97%** of malware focuses the Android OS (Pulse Secure, June 2015)
- Classic static and dynamic analysis are expensive and time consuming.

2015 smart phone market share



Objectives

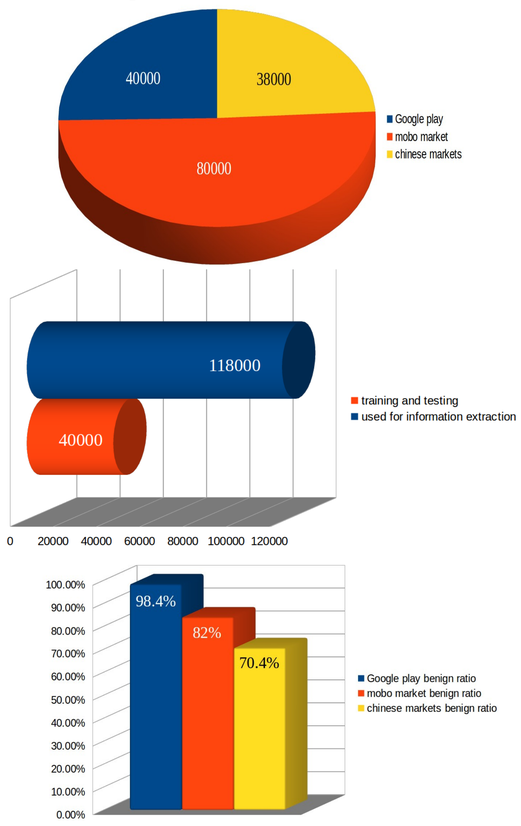
- To build a lightweight Android classification system using meta-data only.
- The whole processing time should be shorter than other Android classification tools.
- As a result of research, we should determine a set of meta-data features that are the most indicative of benign and malware Android apps.

Meta-data feature groups

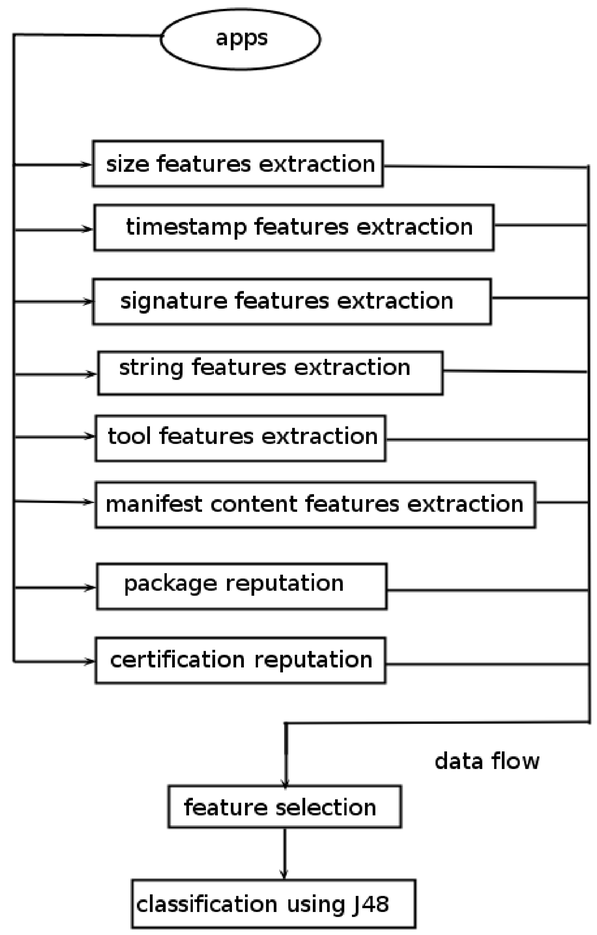
For each meta-data group, there is a list of sub features. For instance, there are more than 50 features in size feature category.

Feature groups	Description
Size	Size of types of files such as: dex, images, xml
Time-stamp	Timestamp patterns of all the files in Android application
Signature	Signature patterns of the current Android app
Android-manifest file	External library names, permissions, activities in the android manifest file
String analysis	Strings in the Android manifest file
Tool analysis	The probably tool used to make the current Android app
Certification reputation	The reputation rate of the signature used to sign the current app
Package name reputation	The reputation rate of the package name of the current app

Experiment Data



Analysis flow



Result

Train and test dataset:
Benign android app rate : 82.7%

Final result:
Correctly Classified Instances 94.3%
Incorrectly Classified Instances 5.7%

Result with each feature separately:

manifest-file analysis
Correctly Classified Instances 93.3719 %
Incorrectly Classified Instances 6.6281 %

reputation package
Correctly Classified Instances 86.8849 %
Incorrectly Classified Instances 13.1151 %

reputation certification
Correctly Classified Instances 86.9957 %
Incorrectly Classified Instances 13.0043 %

size
Correctly Classified Instances 86.963 %
Incorrectly Classified Instances 13.037 %

string analysis
Correctly Classified Instances 91.511 %
Incorrectly Classified Instances 8.489 %

timestamp
Correctly Classified Instances 84.0091 %
Incorrectly Classified Instances 15.9909 %

tool
Correctly Classified Instances 82.7726 %
Incorrectly Classified Instances 17.2274 %