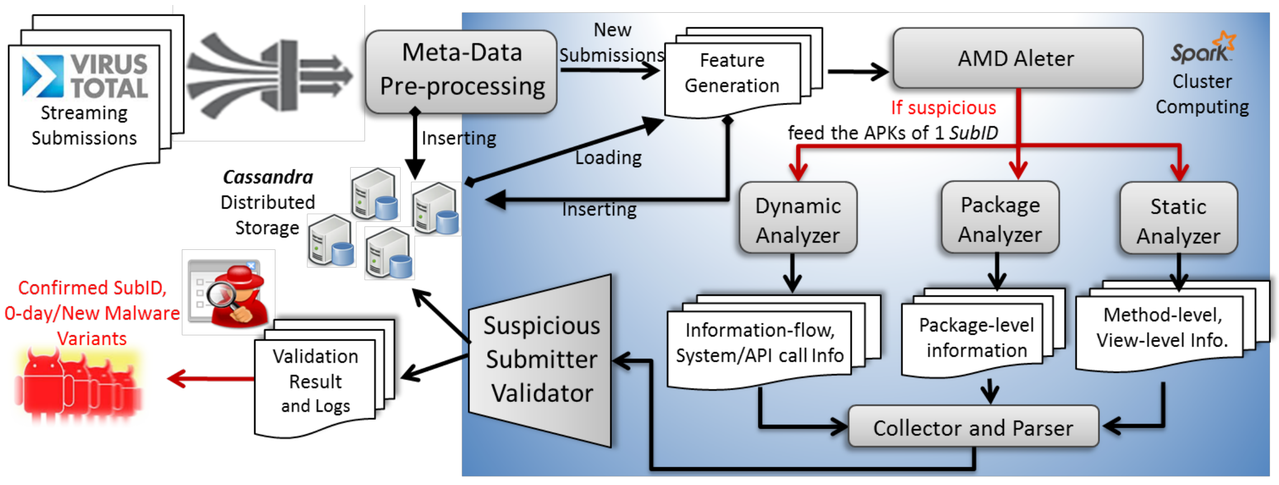
Detecting Android Malware Development on Public Malware Scanning Platforms: *A Large-scale Data-driven Study*

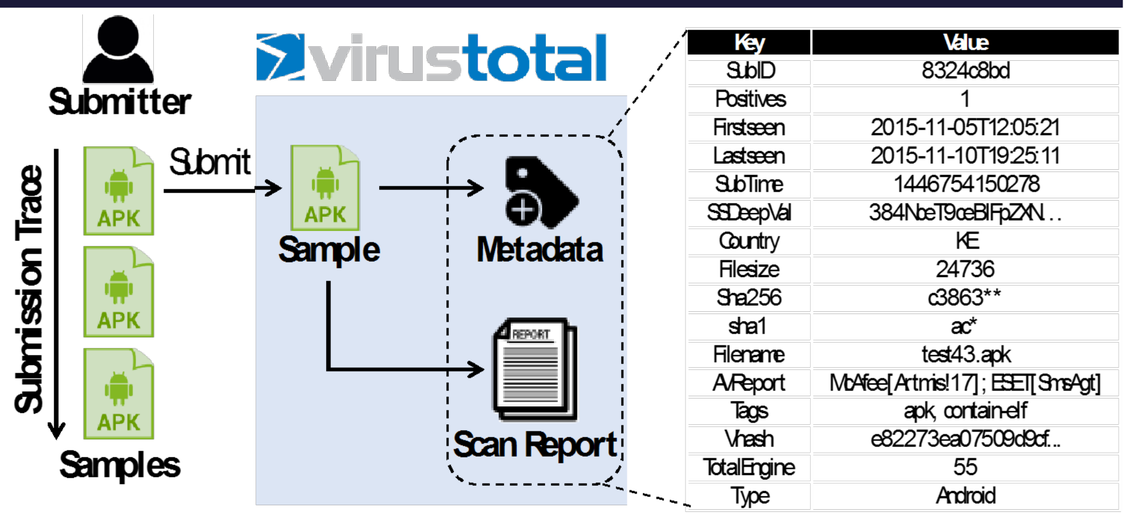
Heqing Huang, Sencun Zhu et al.

Abstract: Android malware scanning services (e.g., VirusTotal) are websites that users submit suspicious Android programs and get an array of malware detection results. With the growing popularity of such websites, we suspect that, these services are not only used by innocent users, but also, malware writers for testing the evasion capability of their malware samples. May this hypothesis be true, it not only provides interesting insight on Android malware development (AMD), but also provides opportunities for important security applications such as zero-day sample detection. In this work, we first validate this hypothesis with massive data; then design a system AMDHunter to hunt for AMDs on VirusTotal that reveals new threats for Android that has never been revealed before. Our study is driven by the large amount of data on VirusTotal— We analyzed 153 million submissions collected on VirusTotal during 102 days.

Android Malware Development Hunter Architecture



Research Challenges



Key	Value
SubID	8324c8bd
Positives	1
Firstseen	2015-11-05T12:05:21
Lastseen	2015-11-10T19:25:11
SubTime	1446754150278
SSDeepVal	384NbeT9ceBIFpZXN…
Country	KE
Filesize	24736
Sha256	c3863**
sha1	ac*
Filename	test43.apk
AVReport	McAfee[Artmis!17]; ESET[SmsAgt]
Tags	apk, contain-elf
Vhash	e82273ea07509d9cf…
TotalEngine	55
Type	Android

VirusTotal (VT): A leading malware submission/scanning platform supported by Google 55+ vendors put the engines with newly updated signatures on VT Used by leading security vendors, IBM security, FireEye, Symantec and normal users.

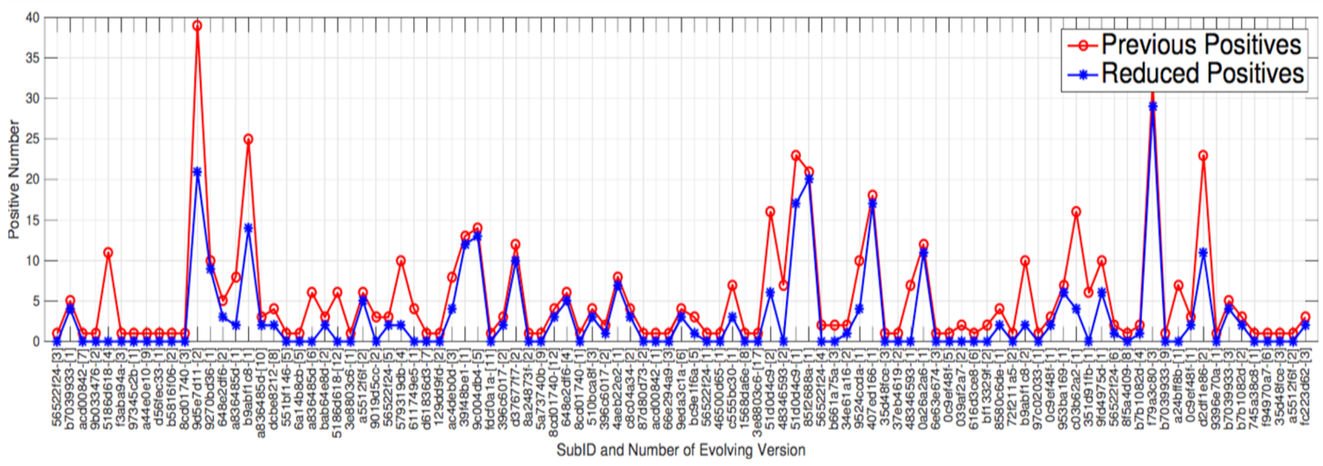
Challenge I: 3Vs (volume, variety and velocity), 1.5 millions of submissions/day, 40 submissions per second, lots of noisy submissions.
Challenge II: limited supporting evidence for validation

- **AMD Alerter**: a pre-filter for AMD submissions based on submitter meta-data centric analysis, which is a **Naïve Bayesian Classifier** built to classify submitters into suspicious or normal groups.
- **AMD Validator**: Design validation logic based on the understanding of Android development process (**package analyzer**) and the nature of AMD cases (**static analyzer** for detecting apps with similar static features, **dynamic analyzer** for detecting obfuscated apps with similar runtime behavior)

Implementation

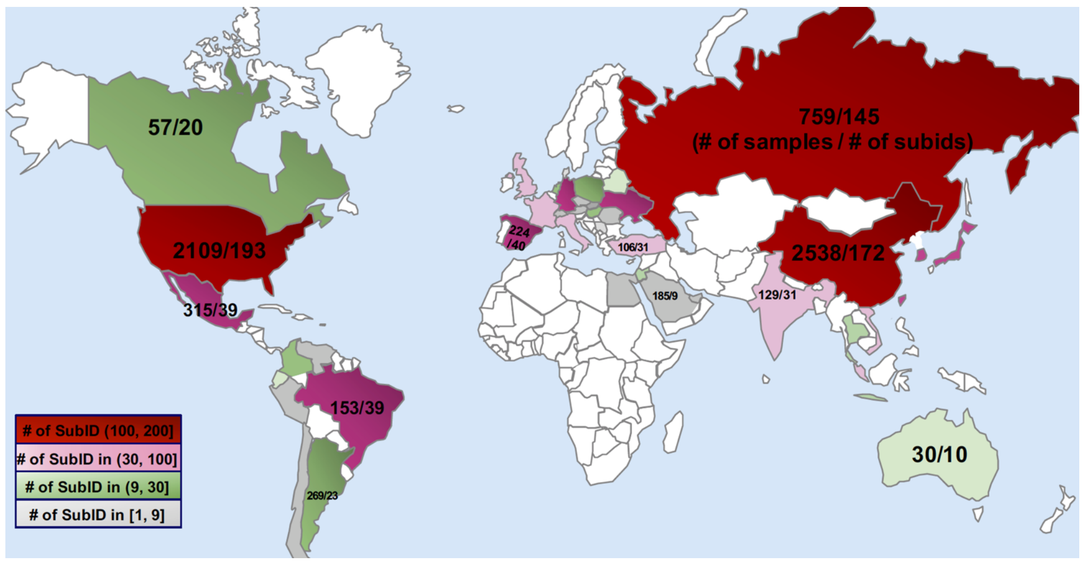
- The current system is built upon a high performance cluster of 20 nodes each with Linux 3.16.0-4-amd 64, 32G RAM, 8 core CPU and uses YARN and Mesos for cluster management.
- Primarily implemented in Scala, the native language that is used to design the Apache Spark, 3,894 lines of Scala code, 2,330 lines of python scripts and 692 lines of shell scripts.
- It leverages the powerful Spark, a distributed computing framework to parse the meta-data and feature generation for each submitter, and Apache Cassandra, an open source distributed database management system to handle data across servers.

Result (1): Evasive Submission Traces



Confirmed 99 Evasive SubIDs 45 reduced to 0 (completely evade all AV engines)

Result (2): Malicious Submitter Distribution



The validated 1,623 submitters were distributed *across **83 countries*** with a total of 13,855 malware samples

Summary of Findings:

- Identified 1,623 AMDs with 13,855 samples from 83 countries, 8,833 of the 13,855 samples were not detected by any AV vendors
- Perform 0-day variants analysis: among 890 randomly selected samples for validation, detect 138 0-day malware (the rest are variants of existing malware).

• Identified many new threats, e.g., the development cases of fake system/banking phishing malware, new rooting exploits, etc.

Related Publications

H. Huang, C. Zheng, J. Zeng, W. Zhou, S. Zhu, P. Liu, S. Chari, C. Zhang. Android Malware Development on Public Malware Scanning Platforms: A Large-scale Data-driven Study. Proceedings of IEEE International Conference on Big Data, 2016