DroidKin: Lightweight Detection of Android Apps Similarity

Hugo Gonzalez, Natalia Stakhanova, Ali A. Ghorbani

Faculty of Computer Science, University of New Brunswick

ISCX Information Security Centre of Excellence

UNB

Our work

- DroidKin, a robust approach for detection of Android apps similarity. Based on a set of characteristics derived from binary and meta data accompanying apk files, DroidKin is able detect similarity among applications under various levels of obfuscation. DroidKin performs analysis pinpointing similarities between applications and identifying their relationships.

Contributions

- Develop a lightweight and robust approach to app similarity detection
- Investigate the impact of lightweight features: meta-data, meta-information and content, on detection of plagiarized apps

Features

- **Meta-data** that accompanies each .apk file. The following features were extracted: serial number, developer's name, size of an .apk file. This descriptive information serves as a reference for analyzed applications.
- **Meta-information** that characterizes .apk file contents. These features include timestamps of the content files; number of internal .apk, .zip, java, .jar, images, libraries and binary files found within a container .apk file; size of .dex file; hash value of .dex file (md5); number of files in .apk file; list of their names with corresponding hash values (md5); number of unique timestamps from the files.
- **N-grams** characterizing the .dex file, using opcodes only, opcodes and operands and bytecode.

Datasets

- Created a dataset with different levels and types of malware app obfuscation (720 apk files)

- Android Malware Genome Project (1260 apps).

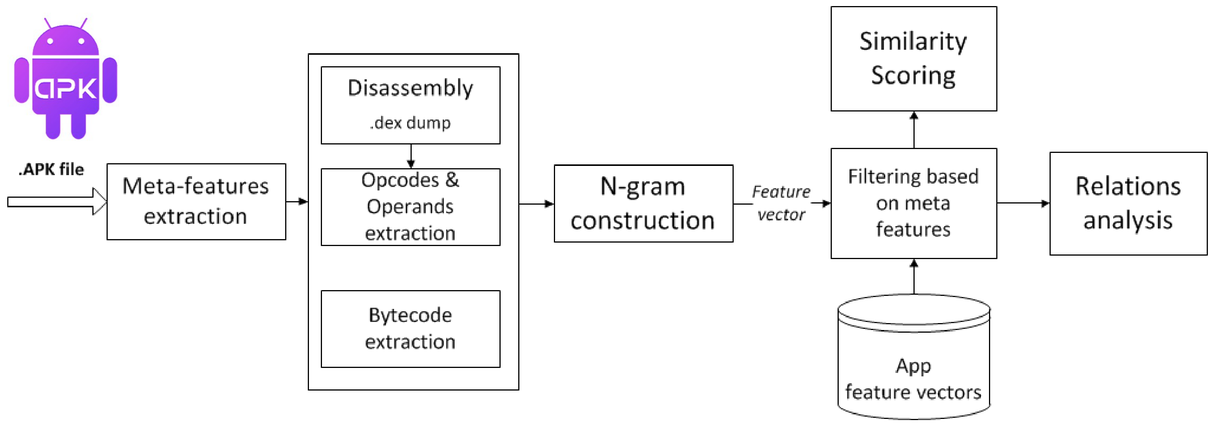
DREBIN

- dataset. (5560 apps).

Scoring

- The pair-wise similarity between related apps is calculated using a variation of a similarity measure Simplified Profile Intersection (SPI) The metric has been proposed for evaluation of source code author profiles based on frequency analysis. Given Simplified Profiles SPi and SPj , the similarity distance, called Similarity Profile Intersection SPI is the size of intersection between these profiles. We use meta-data, meta-information and n-grams to calculate this.

System Design

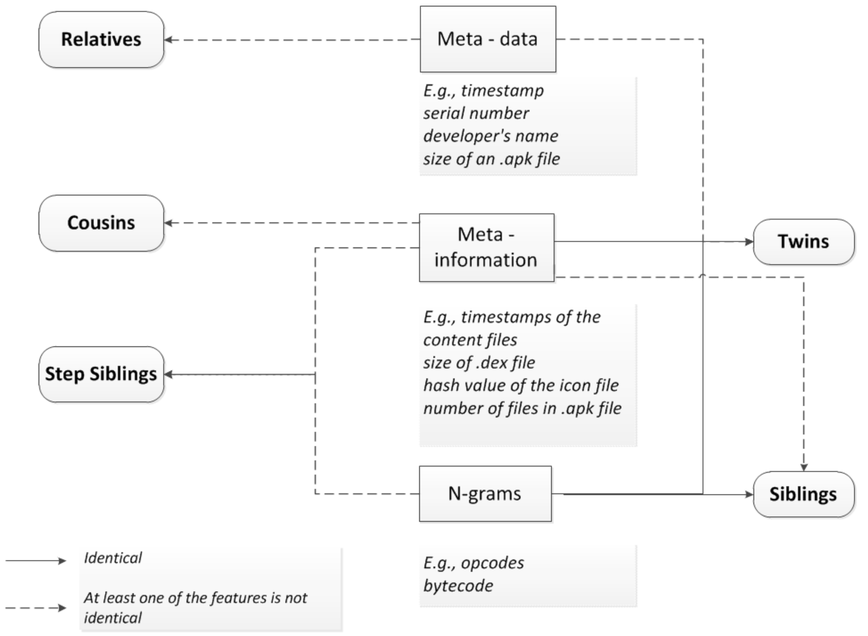


$$SPI_n = |SP_i \bigcap SP_j|$$

- **SPIn =** SPI for the frequency of the ngrams extracted from dex file.
- **SPIc = .6 *** SPI for the content of the files + .4 * SPI for the names of the files.

Relations

- **Twin:**
  SPIn = 100% & SPI > 95%
- **Sibling:**
  SPIn = 100% & SPI > 60%
- **Step-sibling:**
  SPIn > 60% & SPI > 60%
- **Cousin**:
  SPIn <60% & SPI > 60%
- **False step-sibling:**
- SPIn >60% & SPI< 60%



Results