VISUALIZING ANDROID APP SIMILARITY
Cybercriminals pose a serious threat to mobile software systems

Most techniques are machine learning based
Cybercriminals pose a serious threat to mobile software systems

Most techniques are machine learning based
BACKGROUND & MOTIVATION

- Cybercriminals pose a serious threat to mobile software systems
- Most techniques are machine learning based
Cybercriminals pose a serious threat to mobile software systems

Q: How do we select benign samples?
A: Most malware studies = randomly but...
PROBLEM

- Random benign can produce separable, vulnerable patterns [1]
- Malware adopt benign behaviors to evade detection
- Select benign samples similar to malware → mitigate vulnerability

PROBLEM

- Random benign can produce separable, vulnerable patterns [1]
- Malware adopt benign behaviors to evade detection
- Select benign samples similar to malware \(\rightarrow\) mitigate vulnerability

GOAL

Help researchers identify similar benign samples w.r.t a malware set

OUR DATASETS

Data Source:
- Google Play
- VirusTotal

Data Amount:
- Benign: 50000+
- Malware: 10000

Data Time Range:
- Benign: 2016 ~ 2019
- Malware: 2016 ~ 2019
OUR DATASETS

Data Source: Google Play Store

Data Amount: 50000+

Data Time Range: 2016 ~ 2019
OUR PROPOSED VISUALIZATION

Heatmap to reveal the similarity between samples

Similarity can be calculated as:
- Distance between samples feature vectors
- Cosine similarity between sample feature vectors
- Etc.
THANK YOU!
BACKUPS
DATA ATTRIBUTES

Feature vector can be extracted through static / dynamic analysis

- Permission
- API Calls
- Components
- Etc.

- Traffic Network
- Battery Usage
- Information Flow
- Etc.

Static Analysis

Dynamic Analysis

\[ x = \begin{bmatrix} x_1 \\ x_2 \\ x_d \end{bmatrix} \]

Feature Vector
Possible interaction

- Allow user to select modify the set of features for similarity calculation
- Allow user to select a subset of samples:
  - Automatically identify the features contribute to similarity/difference the most → the set of interesting features
  - Show the distribution of samples over the interesting features