Android Financial Malware Detection Through Network Flow Analysis

Andi Fitriah A. Kadir, Natalia Stakhanova, Ali A. Ghorbani

{andi.fitriah, natalia.stakhanova, ghorbani}@unb.ca

Canadian Institute for Cybersecurity (CIC), University of New Brunswick (UNB)

THE OVERVIEW

"Android financial malware exists because information has value now" - *Anonymous*

Current Problems:

- ✓ What constitutes Android Financial Malware (AFM) is still ambigous
- ✓ Current solution focused more on malware binary detection
- ✓ Most of the available datasets are crafted for static analysis
- ✓ More sophisticated techniques to thwart malware detection

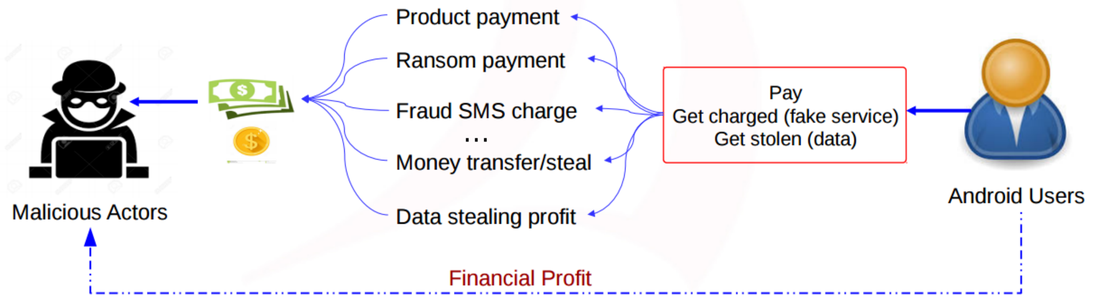
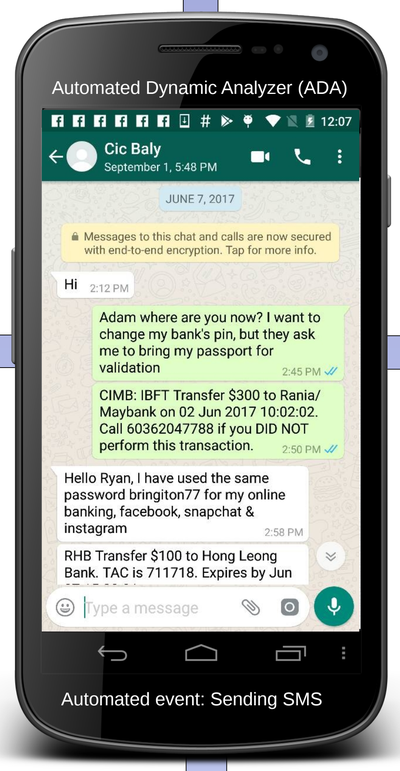


Fig. 1: Android Financial Malware (AFM)

Our Definition of AFM:

- ✓ Direct financial profit or money exchange to the fraudsters
- ✓ Financial transaction includes reselling or direct transactions
- ✓ Without the user's knowledge or consent



THE SOLUTION

"For every one second, there is a new malware" - *Ralf Benzmüller*

Goal:

We focus on detecting malware and categorizing its type based on the defined taxonomy.

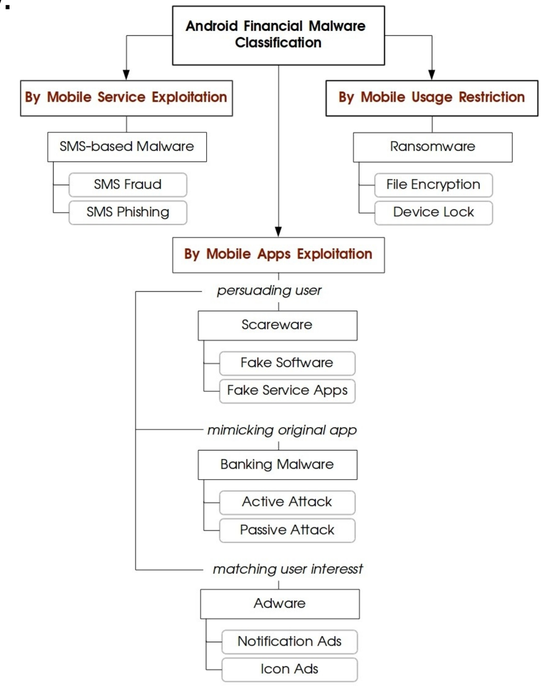


Fig. 2: Proposed Taxonomy

THE EXPERIMENTS

Dataset:

- ✓ 500 Malware and 5k Benign samples (2010 - 2017)
- ✓ 5 categories; 10 sub-categories (Fig 2 of Taxonomy)
- ✓ Run analysis on real environment (ADA on smartphone)

Table I: Dataset splitting for training and evaluation

Dataset	Dataset Ratio	Training	Evaluation
1	50:50	500 (250 M, 250 B)	500 (250 M, 250 B)
2	60:40	750 (300 M 450 B)	500 (200 M, 300 B)
3	70:30	1200 (350 M, 850 B)	500 (150 M, 350 B)
4	80:20	2000 (400 M, 1600 B)	500 (100 M, 400 B)
5	90:10	4500 (450 M, 4050 B)	500 (50 M, 450 B)

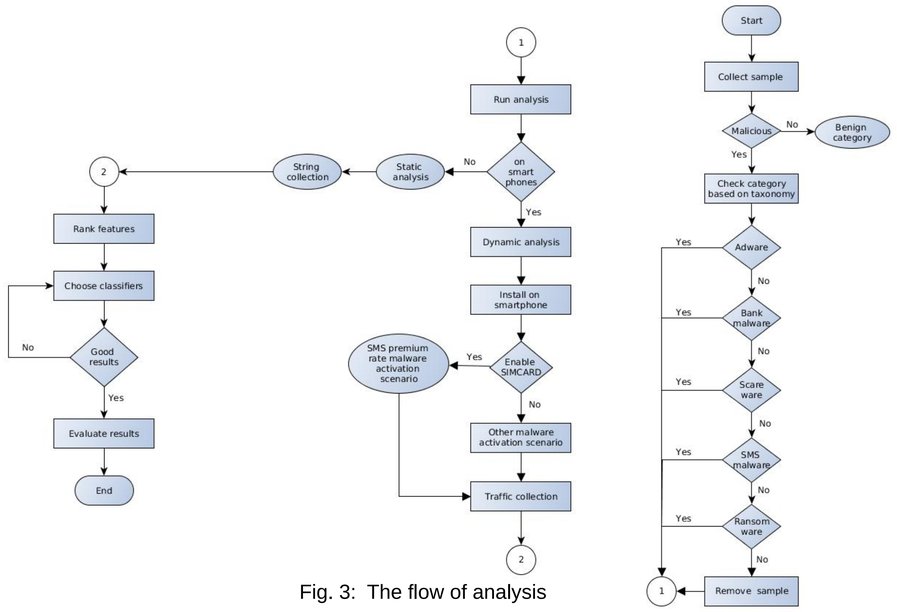


Fig. 3: The flow of analysis

THE RESULTS

Scenario A: Malware Detection

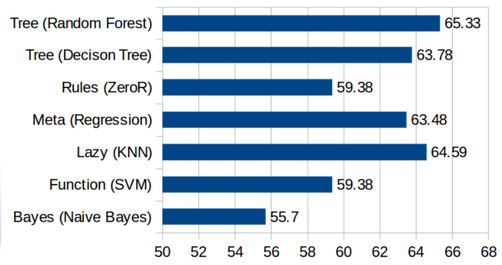


Fig. 4: Best classifier (Random Forest)

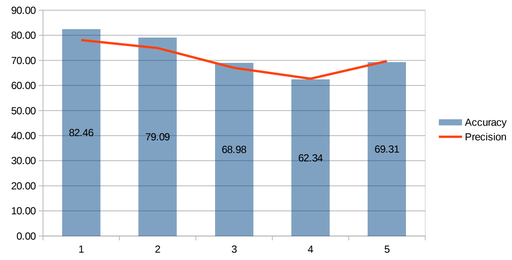


Fig. 5: Malware vs Benign (Random Forest)

Scenario B: Malware Categorization

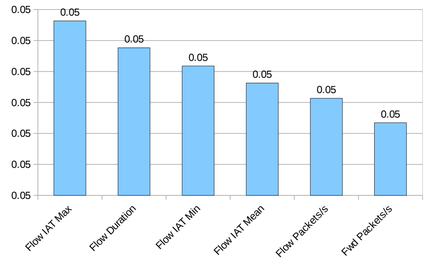


Fig. 6: Feature ranks via Random Forest Importance

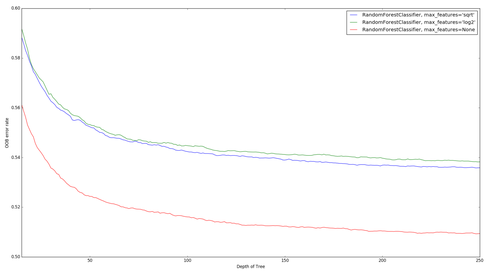


Fig. 7: Sub-category: Out-of-bag error of Random Forest

Table II: Sub-category results (10 classes) with Random Forest

Dataset:	Set-1	Set-2	Set-3
ROC:	87.5	83.5	82.4
F1-Score:	55.5	46.6	45.6
FP Rate:	0.065	0.066	0.074