Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification

Arash Habibi Lashkari, Andi Fitriah A. Kadir, Laya Taheri and Ali A. Ghorbani

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

ABSTRACT

Problem

✓ Machine learning methods proposed in previous work typically reported high detection performance and fast prediction times on fixed and defective datasets

✓ Based on some shortcomings most of datasets are not suitable for real-world deployment

Goal

✓ Propose a systematic approach to generate Android malware dataset using real smartphones instead of emulators

✓ Develop a new dataset namely CICAndMal2017, which covers all the shortcoming and limitation of previous datasets

Evaluation

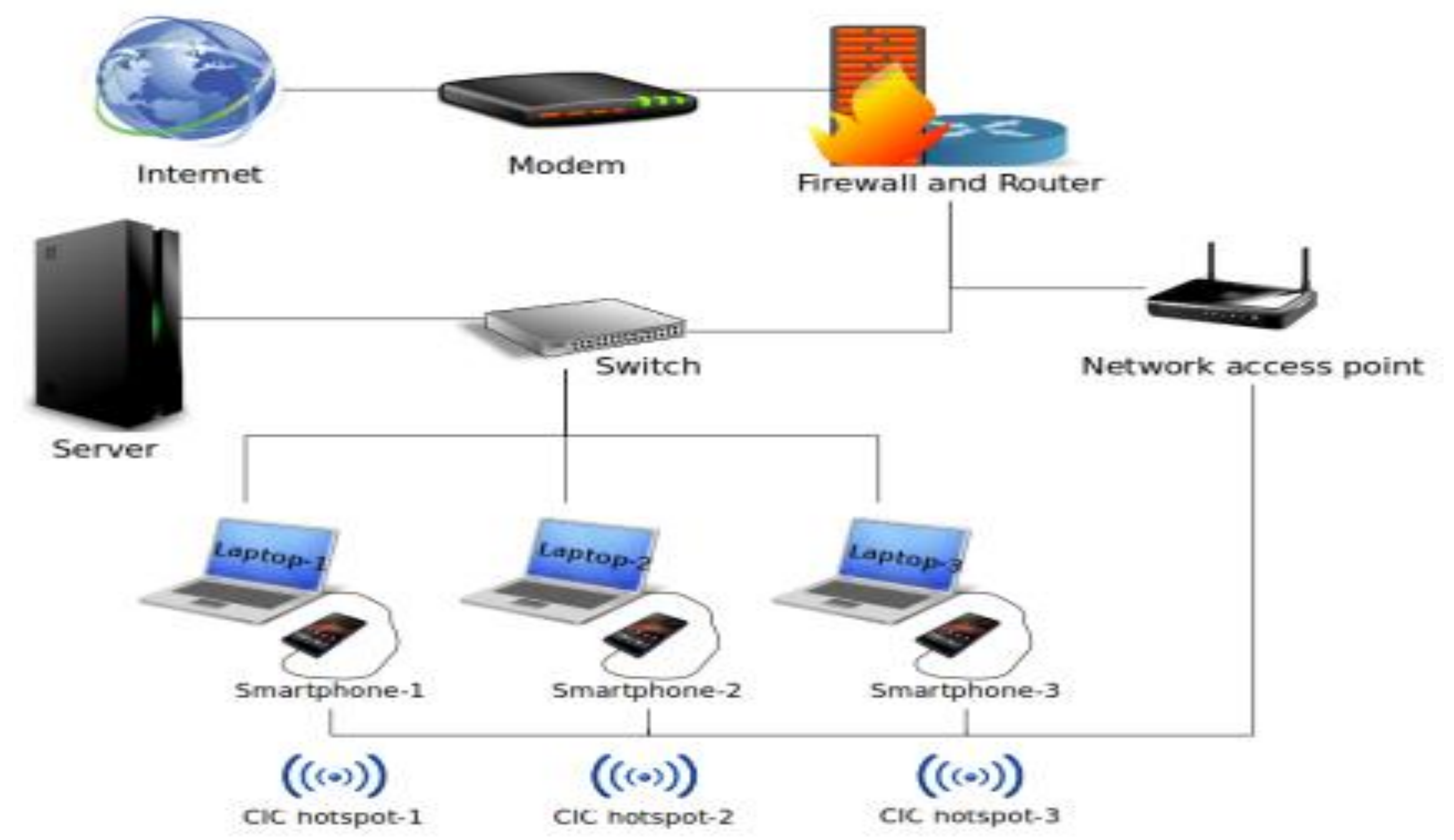
✓ Offer 80 network traffic features to select the best features set

✓ Showed the average precision 85% and recall 88% for three classifiers namely Random Forest(RF), K-Nearest Neighbor (KNN), and Decision Tree (DT)

Previous Available Datasets

Year	Dataset Title	Type	Captured Behavioral Features	Number of Samples	Shortcomings
2012	GENOME Project	Static	Studied components of the malicious source code, tracked API calls and studied permission lists	1260 malware	Lack of dynamic features, Installation
2014	DREBIN	Static	Studied malicious source code and manifest file features such as permission lists and API calls	5560 malware - 123,453 benign	Lack of dynamic features, Installation
2017	AMD	Static	Studied malicious components of code	405 malware	Static analysis
Our proposed Dataset	*CICAndMal2017*	*Static & Dynamic*	*is completely labelled and includes network traffic, logs, API/SYS calls, phone statistics, and memory dumps of 42 malware families.*	*Installed 429 malware - 5,065 benign*	*Address previous Shortcomings*

The Network Architecture



User-Interaction Scenarios

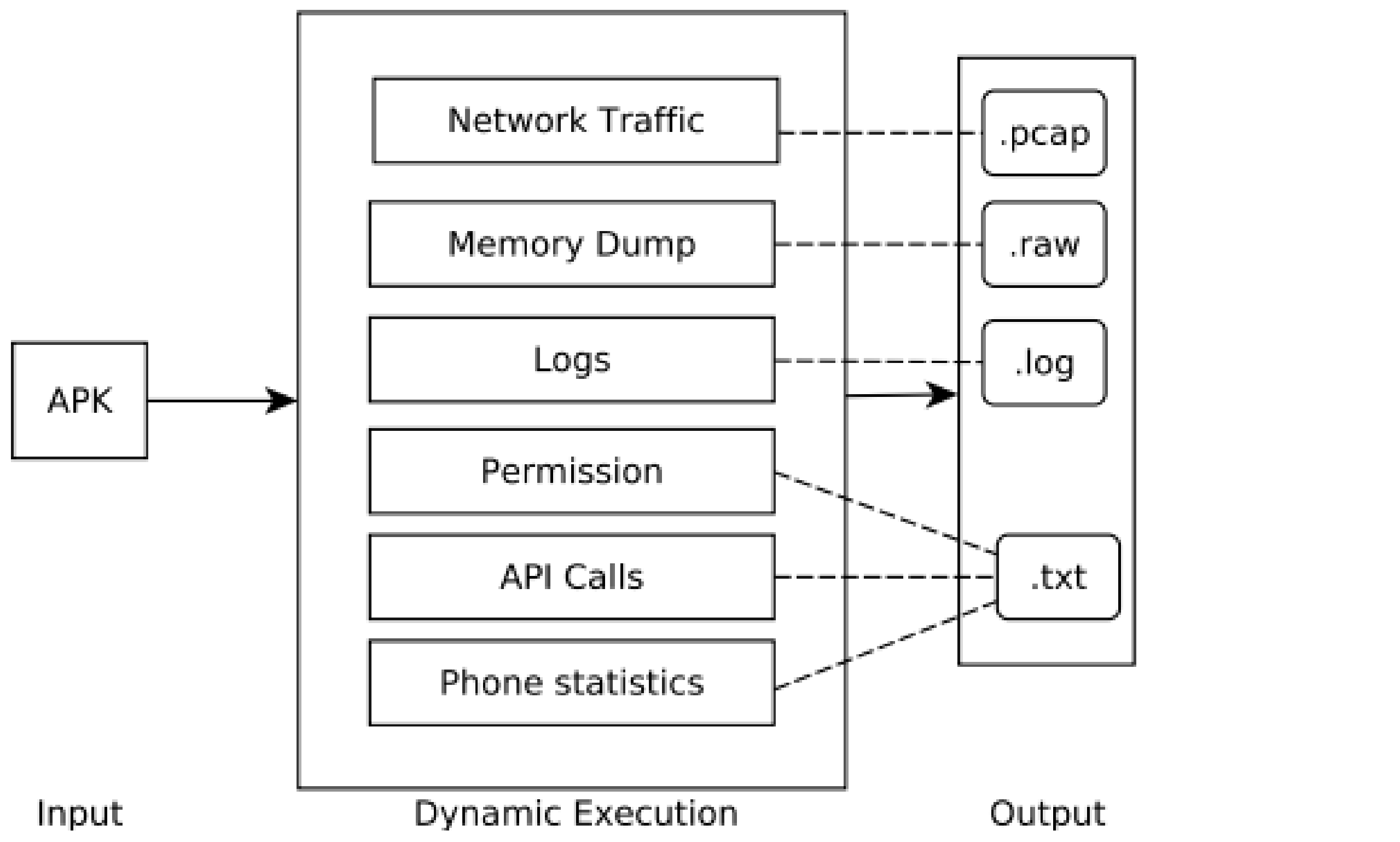
Category	Scenario	Rem.
Benign	- Send Message - Make Call - Enable GPS - Browse Internet	SIM card disable
Adware	- Send Message - Make Call - Enable GPS - Browse Internet	SIM card disable
Scareware	- Send Message - Make Call - Enable GPS - Browse Internet - Click/follow popup	SIM card disable
Ransom	- Send Message - Make Call - Enable GPS - Browse Internet - Click/follow popup - Set the four-digit PIN and lock the phone - Click/interact with any popup message - Save more than 10 contacts in the contact list - Save the following documents in both internal and external storage: jpeg, jpg, png, bmp, gif, pdf, doc, docx, txt, avi, mkv, 3gp, mp4 (size more than 10 KB)	SIM card disable
SMS mal-ware	- Send Message and SMS - Make Call - Enable GPS - Browse Internet - Install AV (AVG, Avast, BitDefender) - Save more than 10 contacts in the contact list	SIM card enable

Taxonomy of Malware Behaviors

	Family	Year	AV Labelled	Total Collected	Total Captured	Attack												Communication			
						A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	C1	C2	C3	C4
ADWARE	Dowgin	2016	Gdata	50	10		✓														
	Ewind	2017	Koodous	50	10	✓		✓		✓		✓						✓	✓		
	Feiwo	2016	Fortinet	100	15	✓															
	Gooligan	2016	Fortinet	43	14	✓				✓	✓	✓						✓		✓	
	Kemoge	2015	Lookout	35	11							✓									
	koodous	2017	Koodous	50	10																
	Mobidash	2015	Enet32	32	10				✓									✓	✓	✓	
	Selfmite	2014	AntiVirus	6	4		✓	✓				✓									
	Shuanet	2015	Lookout	24	10					✓	✓										
	Youmi	2015	Gdata	50	10	✓						✓									

20 types of attacks (A1-A20) and 4 types of C&C communications (C1-C4)

Captured & Monitored Data Sources



States of Data Capturing: Installation, Before restart, After restart

Conclusion and Future Works

- Reviewed serious drawbacks of available previous datasets
- Show actual malicious behavior by installing on real devices
- Importance of User-interaction scenarios for malware activation
- Using real smartphones instead of emulators
- Design different activation scenarios to trigger different families
- Focused on the network traffic
- Extract more than 80 network traffic features
- Future work: Extract the useful features from other data

Network Traffic Analysis Results

Dataset:	Training (10-fold cross validation)									Evaluation (Testing set)								
Scenario:	A (Malware Binary)			B (Malware Category)			C (Malware Families)			A (Malware Binary)			B (Malware Category)			C (Malware Families)		
Algorithm:	RF	KNN	DT	RF	KNN	DT	RF	KNN	DT	RF	KNN	DT	RF	KNN	DT	RF	KNN	DT
Precision (%):	84.00	83.60	85.10	46.50	45.70	46.50	22	21.50	21.00	85.80	85.40	85.10	49.90	49.50	47.80	27	27.24	26.66
Recall (%):	87.50	87.30	88.00	45.50	44.80	44.70	21.50	21.60	21.40	88.30	88.10	88.00	48.50	48.00	45.90	25.50	23.74	20.06