POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXXI cycle

Supervisor

Giovanni Squillero

Machine Learning and Other Computational Intelligence Techniques for Security Applications

PhD Candidate: *Andrea Marcelli*

1. Introduction

Malware is a big business [1]. With hundreds of thousands of malware delivered every day, manual analysis in not an option. New automated approaches have to be designed to effectively detect new threats.

2. Objectives

- Scalability (20k – 30k new APKs/day)
- Reduce malware exposure time
- 100% recall and very high precision
- Automate very repetitive tasks to save a considerable amount of experts' time and resources.

3. Proposed method

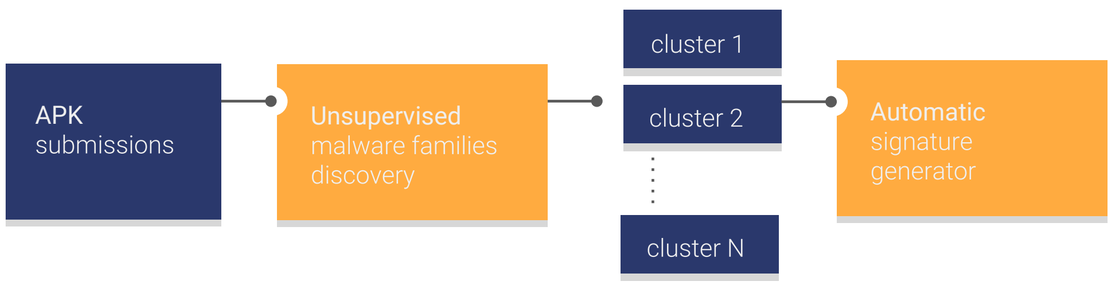

Figure 1: the proposed workflow

Fig. 1 shows the workflow proposed in [2]. New applications are periodically analyzed through an unsupervised procedure. Clusters are then classified in 7 categories: according to their composition, some may require a manual approval, then new signatures are created to detect new variants.

The problem of generating a signature is reduced to a variant of the "Set Cover Problem" (Fig. 2): several heuristics, a greedy, and an evolutionary algorithm are used to create a ruleset that perfectly balances the "generality" with "specificity".

The algorithm was implemented in a tool, YaYaGen [3], presented in [4]. Since Jan. 2018, it is used in Koodous, a real-word mobile AV.

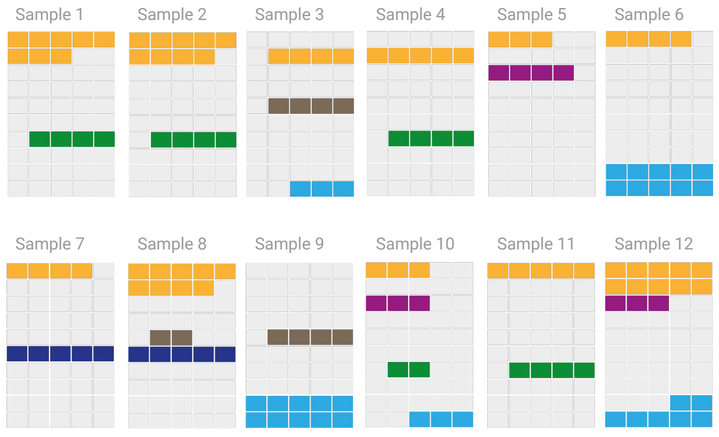

Figure 2: the "Set Cover Problem"

Figure 3: YaYaGen logo

4. Results

Rule Name	Original	YaYaGen	Improvement
SMSSENDER	539	1,004	+86.3%
SYRINGE	220	315	+43.2%
HUMMINGBAD2	136	257	+89.0%
MARCHER2	559	652	+16.6%
SMSREG	159	172	+8.2%
VOLCMANDROPPER	186	430	+131.2%

Figure 4: comparison between manual and automatically generated rules

Extensive tests were carried on a huge dataset of 1.5 millions Android applications. Fig. 4 shows that automatically generated rules increase the number of malware detected, ranging from the 8% to the 131%, while reducing the number false positives.

5. Conclusions

A new scalable semi-supervised approach to dig into massive dataset of applications is presented. It allows to automatically identify malware families and generate YARA rules.

6. References

1. "The Rise of Android Banking Trojans", Andrea Atzeni, Fernando Diaz, Francisco Lopez, Andrea Marcelli, Antonio Sanchez, and Giovanni Squillero. IEEE Potentials, under review
2. "Countering Android Malware: a Scalable Semi-Supervised Approach for Family-Signature Generation". Andrea Atzeni, Fernando Diaz, Andrea Marcelli, Antonio Sanchez, Giovanni Squillero, and Alberto Tonda. IEEE Access, 2018
3. https://github.com/jimmy-sonny/YaYaGen
4. "Looking for the perfect signature: an automatic YARA rules generation algorithm in the AI-era", Andrea Marcelli. BSidesLV 2018 and DEF CON 26