SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT

The STSM applicant submits this report for approval to the STSM coordinator

Action number: CA15140 - Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO)
STSM title: Analysis of Android malware through evolutionary machine learning techniques.
STSM start and end date: 01/11/2017 to 19/01/2018
Grantee name: Andrea Marcelli

PURPOSE OF THE STSM/

(max.500 words)

The purpose of the STSM is to experiment the application of evolutionary machine learning techniques to the field of Android malware analysis and detection, where both EAs are employed in the optimization of machine learning parameters, and ML techniques are applied in the domain of evolutionary computation. Furthermore, the collaboration with the Host Institution aims at developing and releasing a set working tool based on evolutionary machine learning techniques to be integrated in Koodous, a collaborative cloud-based research antivirus platform developed by Hispasec Sistemas (EU FP7-ICT NEMESYS).

Malware is a big business. With hundreds of thousands of malware delivered every day, manual analysis in not an option. Machine learning and Evolutionary computation are powerful tools that achieved incredible results in the most variegate fields. Although the techniques are quite known, their application requires a deep knowledge in the field of usage.

While the current industrial standard for the detection of malicious applications is based on signatures, it is well established that even tiny alterations in the malware code can make them ineffective. In order to tackle such a specificity problem, new signatures are frequently created; however, given the vast amount of new and unidentified malware, this task takes a considerable amount of the experts' time. Hence, it is of utmost practical value automatically identify common malicious patterns, group applications into identifiable families, and reduce the effort required by human experts.

The ultimate goal is to develop an algorithm able to automatically identify malware families and generate the corresponding family signatures, a crucial aid for malware analysts, eventually reducing threat exposure and increasing the quality of the detection.

Koodous is a collaborative cloud-based research antivirus platform developed by Hispasec Sistemas (EU FP7-ICT NEMESYS), focused on detecting fraudulent Android applications employing both malware signatures and community knowledge. The database is one of the largest collection of Android applications ever built, counting more than 20 million of applications, among which 6.7 million have already been identified as malicious. Koodous gives free access to a full set of web-based Android analysis tools, taking advantage of an open community to identify benign or malicious applications.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

(max.500 words)

The work carried out during the STSM focused on the research of new techniques to automatically generate Android malware signatures and identify malware families through behaviour similarities, and apply the proposed methodologies on the field using real large-scale data from the Koodous platform.

Firstly, I developed a tool based to automatically generate a sub-optimal set or rules (using the YARA language) to identify new malware threats with high precision and recall. The tool, named YaYaGen uses a combination of algorithms (EA, ML, Greedy) to optimize each phase of the rule generation. Performances have been evaluated on a massive dataset of millions of applications provided by the Host Institution, showing promising results: in few minutes the routine is able to generate precise ruleset able to extend the detection to future malware variants too. (Some of the rules automatically generated can be seen at: https://koodous.com/analysts/YaYaGen/rulesets)

Secondly, I started to investigate alternative methods to cluster applications according to their behaviour similarities. Among the others, embeddings showed the most promising results: by appling the Doc2Vec algorithm to each analysis report, it is possible to unveil similarities among known and new applications.

Finally, during the time spent at Hispasec Sistemas, I ideated and started developing Koodous Brain, a web platform that aggregates a set of AI tools which provide a substantial aid to Android malware researchers.

Some of the obtained results have already been reported in two journal articles submitted to IJIS and IEEE S&P, while the others will be subject of next submissions. (Further details are available in the following section.)

DESCRIPTION OF THE MAIN RESULTS OBTAINED

(max. 500 words)

Thanks to the period spent at the Host Institution it was possible to better define the challenges and experience the daily needs of mobile malware analysts, apply on the field and on real data the proposed techniques, and finally I had the invaluable opportunity of exchange technical views with experts.

During the STSM, two journal articles describing the results of the research activities have been submitted for review.

The first, “Countering Android Malware: a Scalable Semi-Supervised Approach for Family-Signature Generation”, has been sent to the International Journal of Information Security (IJIS). The article describe a scalable, semi-supervised framework to dig into massive dataset of Android applications and identify new malware families. It is able to automatically cluster applications in families and suggest formal rules (signatures) to identify malware with 100% recall and quite high precision. Although the article includes research results obtained during the last year, during the STSM the proposed method received some decisive improvements, especially about the rule generation. Extensive tests were carried on a huge dataset of 1.5 million Android applications and showed promising results: during the clustering phase new malware were automatically proposed with a minimum precision of the 91%, and automated generated rules improved the number of detections of manual ones ranging from the 8% to the 131%.

The second, titled “The Rise of Android Banking Trojans”, has been sent to the IEEE Security & Privacy. It surveys the issues related to banking Trojans spread: malicious programs written with the purpose of stealing confidential information from user bank accounts and online payment system. Indeed, at the end of 2017, they represent the most dangerous threat in the Android ecosystem. In the article, we explore their evolution, highlighting their increasing capabilities, and how antivirus continuously develop new mechanisms to contrast their diffusion. The collaboration with Hispasec Sistemas was essential, since company researchers

are known to be at the frontline of combating Android bankers Trojans since their first appearance.

In order to provide access to the developed technologies, we ideated and developed “Koodous Brain” a service which aggregates a set of AI tools designed to assist Android malware researchers. The platform eases the process of malware analysis, and it consist of a clustering and recommendation system, an automatic signature generation tool (YaYaGen), and network traffic analysis through graph analytics. It has written using the latest technologies, and scalable solutions (Django, Django Rest Framework, Celery, Redis, Web Sockets). The development is toward the end and it will be soon released to the public.

FUTURE COLLABORATIONS (if applicable)

(max.500 words)

Thanks to the period spent at Hispasec Sistemas, the collaboration with the research team at the Host Institution is stronger then ever, and several interesting research projects are ongoing.

Koodous Brain, the service born to aggregate several artificial intelligence tools to assist Android malware analysis, is still under development: currently only the YaYaGen service, the one in charged of automatically create malware family signatures, is fully functional, while the others are toward the end. The platform is planned to be released within the end of February and the code open sourced.

The research on embedding, a way of represent each app as a vector in a multidimensional space and easily identify similar ones, showed very promising results and I expect to soon describe the methodology and report the results in a conference article.

Finally, I am about to start a new research using evolutionary algorithms and reinforcement learning techniques to challenge the Malware OpenAI Gym (https://github.com/endgameinc/gym-malware), a toolkit that provides several manipulating primitives for executable files, and which assigns a reward based on specific actions taken. As the toolkit only works with Windows executables, an extensions to Android will be proposed.