Research Exhibition in Mathematics and Computer Sciences (REMACS 6.0)/
ISBN: 978-629-97440-5-4

MALWARE DATA COLLECTION USING CUCKOO SANDBOX

Ahmad Fikri Muhammad, Mohd Faris Mohd Fuzi and Hafizah Hajimia
College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA Perlis Branch
afmuhammad00@gmail.com, farisfuzi@uitm.edu.my and hafizah.hajimia@uitm.edu.my

ABSTRACT – As the threat landscape continues to evolve, the need for effective malware analysis and detection techniques becomes increasingly crucial. Cuckoo Sandbox is an open-source automated malware analysis system that allows for the execution of suspicious files and the collection of comprehensive data on their behaviour. Cuckoo Sandbox able to run malware samples for analysis, running them in a controlled environment, and monitoring their activities. Furthermore, the objectives of this project is to presents the diverse range of data collected by Cuckoo Sandbox during the analysis process. This includes system call traces, network traffic, registry modifications, file system changes, and screenshots, among other valuable information. The results of the analysis was successfully analysed and can be used for malware analyst and researcher. It emphasizes the significance of this rich dataset in understanding the behaviour and capabilities of malware. It highlighted the importance of robust data collection techniques in combating the ever-growing threat of malware in today's digital landscape.

Keywords: Data collection, Cuckoo Sandbox, behaviour

1. INTRODUCTION

The objective of this project is to gather malware data collection using Cuckoo Sandbox. This will involve collecting a large sample of dataset from various sites and files to gather as much as possible of malware behaviour. Once the dataset has been collected, Cuckoo Sandbox will be used to analyze the data and produce a information to show the behaviour of the data either malware or benign by predicting the score, out of 10. Moreover, the used of Cuckoo Sandbox also able to evaluate other performance metric to analyse the behaviour thoroughly. After that, the sample of unknown files either malicious or benign will be tested and analysed to create own sample of data that can be used by other researcher.

2. METHODOLOGY

This project will begin by do some literature review to gain a better understanding of the project goals and objectives. Relevant topics will be identified by researching various sources. Then, proceed to the project development where it need to install a VirtualBox to keep the host machine safe. Linux operating system will be used in this project which is Ubuntu 18.04. After that, the implementation of Cuckoo Sandbox inside the Ubuntu. After all the installation is done, the next step will be to prepare the data needed to do the analysing. Multiple sources such as Kaggle, VirusBay, Das Malwerk will be searched for gather information about malware behavioural. These data from multiple sources are using different type of files such as .exe, .csv, .zip, and other more. URL's also will be used in this project to do the analysis. Cuckoo Sandbox will run a hybrid analysis to gain information and metrics value needed. The final data will be compiled in CSV file format for further use.

3. RESULTS AND DISCUSSION

The collected data after the analysis is complete will be compiled in one file. It contained all the signatures, md5 hash value, sha256 hash value, score, file name, and file type. All the information is automatically gained from the Cuckoo Sandbox that has been configured. In this project, only 40 files and URL are being use to analysed and gather all the information to create a sample of dataset that can be used to other researcher and malware analyst.

4. NOVELTY OF RESEARCH / PRODUCT

Previous research on malware analysis and data collection has been done by a few researchers. One of the previous works was A Framework for Collecting and Analysis PE Malware Using Modern Honey Network by Muhamad Malik Matin & Rahardjo, in 2020. Their research is to identify malware PE file type formats and to develop a honeypot. There are 1222 malware has been collected during the research and 77% is PE file format and 23% is other files format. Next project is by Lu, Cai, and Tang in 2022 about Research on the Construction of Malware Variant Datasets and Their Detection Method. In this research shows that malware samples and API sequences are difficult to obtains. Their objectives was to enhance the ability of detection even under obfuscation and variants, and to create a dataset of obfuscated and unobfuscated malware variants. Last but not least, Mal-warehouse: A data collection as a service of mobile malware behavioural patterns by Kouliaridis, Barmpatsalou and Kambourakis. This research is to develop an open-source tool performing data collection-as-a-service for Android malware behavioral patterns.

5. CONCLUSION

In conclusion, malware data collection using Cuckoo Sandbox for this project was only a prototype to gain information, behaviour and signature of malware in different types files format. This research had its own limitations during the development and implementation process. There are several ideas and recommendations to improve this research project for future work.

REFERENCES

Lu, F., Cai, Z., Lin, Z., Bao, Y., & Tang, M. (2022). Research on the Construction of Malware Variant Datasets and Their Detection Method. *Applied Sciences (Switzerland)*, *12*(15). https://doi.org/10.3390/app12157546

Muhamad Malik Matin, I., & Rahardjo, B. (2020, October 23). A Framework for Collecting and Analysis PE Malware Using Modern Honey Network (MHN). *2020 8th International Conference on Cyber and IT Service Management, CITSM 2020*. https://doi.org/10.1109/CITSM50537.2020.9268810

Kouliaridis, V., Barmpatsalou, K., Kambourakis, G., & Wang, G. (2018). Mal-Warehouse: A Data Collection-as-a-Service of Mobile Malware Behavioral Patterns. *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*,1503–1508. https://doi.org/10.1109/SmartWorld.2018.00260