ClusterMal: Automated Malware Analysis with clustering, anomaly detection and classification of existing and new behavioral analysis

Prabhath Lakmal Rupasinghe [1], Krishnadeva Kesavan[2], Sripa Vimukthi Bannakkotuwa[3], V.V.YY. Wickramanayake[4], M.P.D.H De Silva[5], J.M.D. Fernando[6], K.K.K.K. Sampath[7]

Department of Information Systems Engineering, Sri Lanka Institute of Information Technology (SLIIT), Malabe, Sri Lanka

[1]lakmalr@gmail.lk, [2]deva130989@gmail.com, [3]sripavimukthib@gmail.com, [4]vijiniyw@gmail.com, [5]dinukhasaranga93@gmail.com, [6]malithdinushan@gmail.com, [7]kalhara.sampath@gmail.com

ABSTRACT

ClusterMal is an automated malware analysis system based on the Cuckoo Sandbox technology. Since malware is the Swiss-army knife of cyber criminals and any other adversary to an organization, in these evolving times detecting and removing malware artifacts is not enough: it is vitally important to understand the behavior, context and motivation and goals of the breach. The Cuckoo Sandbox plays a vital role in analyzing malware but has no clustering feature. As a solution, we propose ClusterMal, which is capable of clustering, anomaly detecting and classifying of existing and new behavioral analysis through machine learning. In the proposed method when a suspicious file is thrown at ClusterMal it throws it into several different environments (VMs with different OSs installed) and a detailed result is outlined. If no similar behavior is observed, a new cluster is created along with the report and a signature for future usage. In practical implication, the module is integrated into Cuckoo for command line interactions with providing high reliability and availability.

Keywords — **Malware Analysis, Clustering, Anomaly Detection, Behavioral Analysis, Sandbox Technology, Machine Learning**

1. INTRODUCTION

In this digital age, open source software is becoming the most significant turning point in the software industry. Even though there are a lot of commercial paid licensed software available, users tend to use open source software as they are free, consist high security, and the source code is available for developing better versions, etc. Meantime, malware threats are becoming more organized and hard to detect. In today's technology-dependent world, enormous amounts of data are stored online and offline every minute and can be immediately subject to attacks. IT professionals do their best to protect the warehouses of information, but as they enhance their efforts, so do the attackers.

A significant threat to data security is malware. It is very important to understand malware, their origins, and the behaviors in order to build the most valuable data protection strategy for companies. For this reason, malware researchers and antivirus engineers had to have a way to analyze malware in depth.

Cuckoo Sandbox technology is open source and is used in a lot of commercial products and services around the world. As we described in this paper main goal of ClusterMal is to further develop this cuckoo sandbox with the capability of clustering the generated reports and making a unique signature for clusters, which can be used as an antivirus signature to detect a lot of different but similar behavioral malware with one signature.

The aim of ClusterMal is to provide a good method to the users who trust in open source software, and for the users who have no ability to buy commercial software.

2. RESEARCH OBJECTIVE

The main objective of this research is to implement a new machine learning module in the cuckoo platform. Clustering, anomaly detection, classification of existing behaviors and new behavior analysis is focused here. This research includes developing a module using machine learning in Cuckoo using Scikit-Learn that should be able to cluster all reports according to similar behaviors. When a new sample is analyzed if no similar behavior is observed, a new cluster should be created along with a virus signature for future references.

The importance of this research is there's the ability to identify how malware act in a different environment by referring the cluster detailed report, also this can be used to produce a signature to the

malware which can be helpful to detect it in the future. Currently, the most of the malware analyzing software are not freely available, but a cuckoo is an open-source software. Hence, with this extended version of Cuckoo (ClusterMal) anyone can use this product freely.

3. METHODOLOGY

The solution overview is mainly based on Cuckoo Malware Analysis which is shown in figure 1. This solution is a module for machine learning in Cuckoo framework using Scikit-Learn that should be able to cluster all reports according to similar behaviors. An algorithm can be developed for clustering purposes using machine learning. Once that a clustering exists and a new sample is analyzed, the new report can be assigned to one of the clusters and compared with similar samples. The module should also be able to perform anomaly detection, hence Alternatively, if no similar behavior is observed, a new cluster should be created. It should be possible to choose among several methods to do this. For example, the distance between the clusters could be measured, or a Support Vector Machines (SVMs) could be trained on existing data using the cluster labels. In machine learning, SVMs are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [21]. After the functionality based on stored analysis data from Cuckoo Sandbox is implemented, the module will be integrated into Cuckoo for command line interaction.

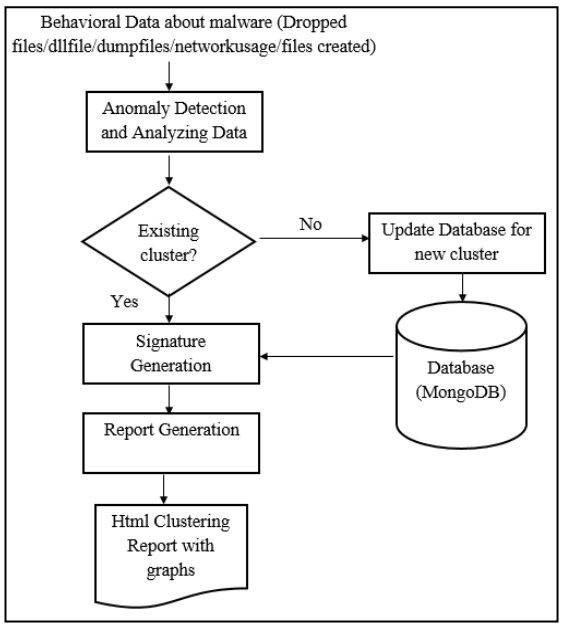

Figure 13: Flow Chart of ClusterMal

ClusterMal is implemented for Linux operating system (Ubuntu 15.0 Wily) based on open-source Cuckoo 2.0 RC1 sandbox. Cuckoo Sandbox is a free open source software that automates the task of analyzing any malicious file under many different environments and produce a report. Python 3.5.1, Yara 3.4.0, g++, MongoDB 3.2, ssdeep and libcap2-bin should be install in the ClusterMal host to run ClusterMal successfully. Few python libraries should be added for several functions as follows,

Python-magic for identifying file formats
Python-dpkt for extracting information from pcaps
Python-mako for rendering html reports and web gui
Python-sqlalchemy
Python jinja2 and Python bottle are necessary for web.py utility

SSDEEP should be installed on ClusterMal host, for calculating fuzzy hashes.libcap2-bin since it is necessary for the cuckoo to run Tcpdump as a non-root user. YARA is used to support identification and classification of malware samples. Python 3.5.1, adobe reader and MS Office should be installed on the VMs. Since ClusterMal has inherited almost all the features of Cuckoo Sandbox, ClusterMal is the new generation of Automated Malware Analysis. Scikit-Learn will be used to develop a module for machine learning in Cuckoo. That would be able to cluster all reports according to similar behaviors.

Research work was carried out to identify available similar applications for malware analysis systems. Linux is used as the operating system and develop using Linux command line. Clustering is used as a method of unsupervised learning, and a common technique for statistical data analysis used in many fields. Here, clustering is the process of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Scikit is used to learn to develop the malware clustering algorithm. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license (a family of permissive free software licenses, imposing minimal restrictions on the redistribution of covered software) and is distributed under many

Linux distributions, encouraging academic and commercial use. There are a lot of benefits in Scikit –learn to develop machine learning algorithms. It provides consistence interface to machine learning models which make is easy to use new models. It also provides many tuning parameters but with the sensible default so that it can be tuned to optimal performance and can use models quickly. It's under active development by active community so it provides the best support. Except for machine learning models Scikit learning also provide a rich set of functionality for companion tasks such as models selection, model evaluation, and data preparation.

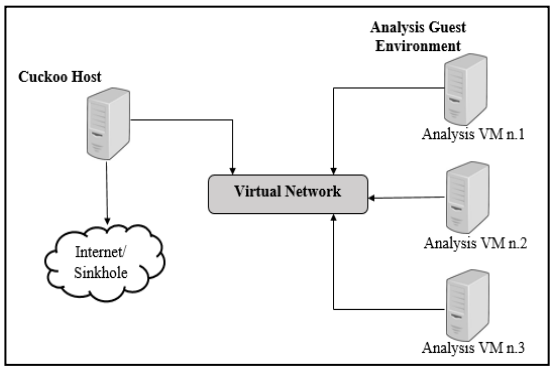

Figure 2: High-level architecture of Cuckoo [15]

Figure 1 and Figure 2 gives the high-level architectures of Cuckoo and ClusterMal where the enhancement of the technology is visible. In this research, a module: ClusterMal is developed for the existing sandbox: Cuckoo and considering the process of extending of the new module in the end to the existing one.

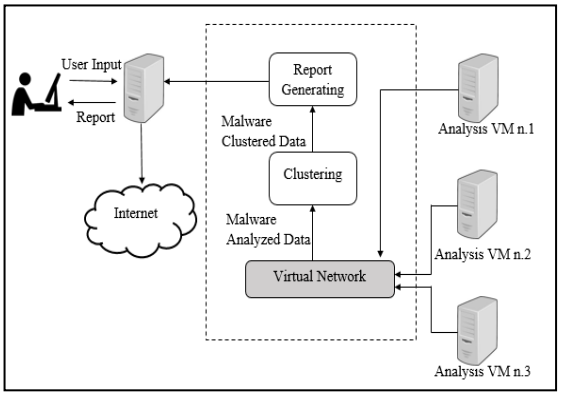

Figure 3: High-level architecture of ClusterMal

The first step is to identify the common behaviors of Malware and Cluster them according to the similar behaviors. After analyzing the data about the behaviors of malware, get the similar behaviors and cluster them accordingly. Then make a cluster of the similar behavior of the malware with the analyzed data. When a cluster exists and a new sample is analyzed, the new report can be assigned to one of the clusters and compared with similar samples. Anomaly detection should be done only if no similar behavior is observed and a new cluster should be created. Also the above process creates a virus signature for each and every clustering groups which is a unique feature and enhances the value of the research. ClusterMal, which is capable of clustering ant it detects anomalies and classify the existing and new behaviors through machine learning.

4. RESULT AND DISCUSSION

Researchers and vendors can get an idea about how the malware acts in a different environment by observing the cluster report. This is an open source approach. Therefore, the solution would help Students who are studying about malware, Antivirus Developers, Malware Analysts and Security Consultants. ClusterMal command line application will only run on Linux Platform. The application will not run on any different operating systems and Linux platforms also should have the recommended software and hardware installed. If anyone needs to install the system first Linux platform is needed.

Under performance requirements, the system will only allow one instance of this application to run on the VM in a given time. Users can not start more than one instance of ClusterMal Application. Multiple users are allowed to use this application from the same VM. The time is taken to generate the report will depends on the size of the malicious file. There will be no specific design constraints to the research. The system will use standard principles of design, and it will maintain the consistency of the user interfaces by using same color and background patterns.

Considering software system attributes high reliability, high availability and high maintainability can be achieved. ClusterMal has less probability of failure since it handle fewer amounts of data but it may fail due to OS failures such as low battery or system crashes. It always provides accurate result because this application contains a machine learning module for clustering. ClusterMal application has high availability. The system installed in user's guest OS will be available at any time.

5. FUTURE WORK

This research can be extended to a level such that it can be more user-friendly. It can be given a new face as an antivirus scanner by using the signatures generated by the ClusterMal. This concept can be further developed as a Security Alert System and send alerts about the currently spreading malware and their behaviors via Email and SMS. ClusterMal can be developed as a web application to enhance the usability. This can be developed for Apple, Android, and Windows platforms as well.

6. CONCLUSION

ClusterMal is an automated malware analysis system using cuckoo sandbox which is capable of clustering, detecting the anomaly and classifying the existing and the new behavioral analysis through machine learning. This is proposed as a solution for the missing of clustering features in the cuckoo sandbox. With the development of this research, there is a facility of identifying similar behaviors of malware, analyzing them and clustering them, when comparing with samples if there are no similar behaviors observed, a new cluster and a signature are created. Using this signature virus guard vendors can implement proper countermeasures for malware in different environments. This research provides a great opportunity to get an idea about how malware acts in a different environment by going through the cluster reports, get the signature details for future detections and have an open source solution for the enhancement in detecting malware.

7. REFERENCES

[1] U. Bayer, I. Habibi, D. Balzarotti, E. Kirda, and C. Kruegel, "A view on current malware behaviors," *Proc. 2nd USENIX Conf. Large-scale Exploit. emergent Threat. botnets, spyware, worms, more*, p. 8, 2009.

[2] U. Bayer, E. Kirda, and C. Kruegel, "Improving the efficiency of dynamic malware analysis," *Proc. 2010 ACM Symp. Appl. Comput. - SAC '10*, p. 1871, 2010.

[3] P. Beaucamps, I. Gnaedig, J. Marion, P. Beaucamps, I. Gnaedig, J. M. Behavior, P. Beaucamps, I. Gnaedig, and J. Marion, "Behavior Abstraction in Malware Analysis - Extended Version To cite this version: Behavior Abstraction in Malware Analysis," 2010.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 4, no. 4. 2006.

[5] F. Channel, S. Area, and A. A. From, "Fibre Channel Storage Area Networks: An Analysis Ins title te Au tho r r eta ins full l rig."

[6] M. Egele, T. Scholte, E. Kirda, and S. Barbara, "A survey on automated dynamic malware analysis techniques and tools," ACM Comput. Surv., vol. 63, no. Ngcit, pp. 41–47, 2011.

[7] S. Gadhiya, "Techniques for Malware Analysis," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 4, pp. 972–975, 2013.

[8] E. Gandotra, D. Bansal, and S. Sofat, "Malware Analysis and Classification: A Survey," J. Inf. Secur., vol. 5, no. April, pp. 56–64, 2014.

[9] D. Keragala, "InfoSec Reading Room," SANS Inst., p. 16, 2014.

[10] J. Hegedus, Y. Miche, A. Ilin, and A. Lendasse, "Methodology for behavioral-based malware analysis and detection using random projections and K-Nearest Neighbors classifiers," in Proceedings - 2011 7th International Conference on Computational Intelligence and Security, CIS 2011, 2011, pp. 1016–1023.

[11] H. De Huang, C. S. Lee, H. Y. Kao, Y. L. Tsai, and J. G. Chang, "Malware behavioral analysis system: TWMAN," IEEE SSCI 2011 - Symp. Ser. Comput. Intell. - IA 2011 2011 IEEE Symp. Intell. Agents, pp. 1–8, 2011.

[12] K. Kendall and C. McMillan, "Practical malware analysis," Black Hat Conf. USA, pp. 1–10, 2007.

[13] K. Mathur and S. Hiranwal, "A Survey on Techniques in Detection and Analyzing Malware Executables," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 4, pp. 422–428, 2013.

[14] V. Notani, “Behavioral Analysis of Malware,” vol. 787, 2009.

[15] D. Oktavianto and I. Muhardianto, Cuckoo Malware Analysis. 2013.

[16] A. Persistence, “InfoSec Reading Room Sleeping Your Way out of the Sandbox.”

[17] K. Rieck, T. Holz, C. Willems, P. Düssel, and P. Laskov, “Learning and classification of malware behavior,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008, vol. 5137 LNCS, pp. 108–125.

[18] Kromer.pl, ‘Choosing the Best Sandbox for Malware Analysis’, 2013. [Online]. Available: malware-analysis/

[19] Cuckoosandbox.org. [Online]. Available: https://cuckoosandbox.org/index.html

[20] Open-source software [Online].Available: https://en.wikipedia.org/wiki/Open-source_software

[21] Support Vector Machine [Online].Available: https://en.wikipedia.org/wiki/Support_vector_machi