RESEARCHERID
THOMSON REUTERS

[Bist* *et al.*, 6(4): April, 2017]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A WIDE SCALE SURVEY ON CLUSTERING TECHNIQUES FOR MALWARE DETECTION

Ankur Singh Bist*
* KIET, Ghaziabad

DOI: 10.5281/zenodo.546351

ABSTRACT

Use of internet has enhanced the problem of malwares. At the same time the advancement in technology has given rise to evolution of portable mobile devices. It has become the danger point of day today activities like mails, mobile banking etc. Our purpose is to trace various clustering techniques for malware detection.

KEYWORDS: Classifiers, Classification Methods, Computer Virus.

INTRODUCTION

Internet has become target of malicious codes due to its increasing use. Malicious codes are executable code and have the capability to replicate. It makes their survival strong. Viruses design and evolution attached with the area of programming. Similar to other computer programs viruses carry functions that are intelligent for providing protection in such a manner that detection remains not easy for virus scanner [1].

Viruses have to take various procedures of intellect for continued existence. That is why they may have complex encrypting and decrypting engines. These are the most frequent methods used by computer viruses in current scenario. They make use of these techniques to mask the antivirus and to adopt the certain environment for their expansion [2].

Figure 1: Assembly code of Virus File

Polymorphic viruses try to hide the decrypting module. More complex methods were developed enabling the virus designers to change the code of one virus file and make multiple morphed copies while maintaining its functionalities. These are the type of viruses which have the ability to mutate itself with the code changed but without changing its functionalities. Metamorphic virus can become a serious threat considering the fact that there can be thousands of variants of one virus file with their signature being totally different.

Metamorphic viruses transform its code in a specific manner very frequently and require to be prohibited. Their analysis will lead to evolve a framework where the overall process of detection will be bounded in specific outcomes of continuing evolving results. It is essential to make a distinction between replicating programs and its similar forms. Reproducing programs will not necessarily damage your system [3] [4] [5]. There is big fight between designers of virus and antivirus. The enhanced knowledge about the certain patterns, specifications can be designed. Various malicious codes can be evolved and incremented in well precise and efficient manner. For

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

perfect identification of a metamorphic virus, identification routines must be written that can generate the essential instruction set of the virus code from the actual occurrence of the infection.

Code obfuscation is one of the important properties adopted by metamorphic viruses. The mutating behavior of metamorphic viruses is due to code obfuscation techniques .There are various code obfuscation techniques.

a. Dead code insertion
b. Variable renaming
c. Break and join transformation
d. Expressing reshaping
e. Statement reordering

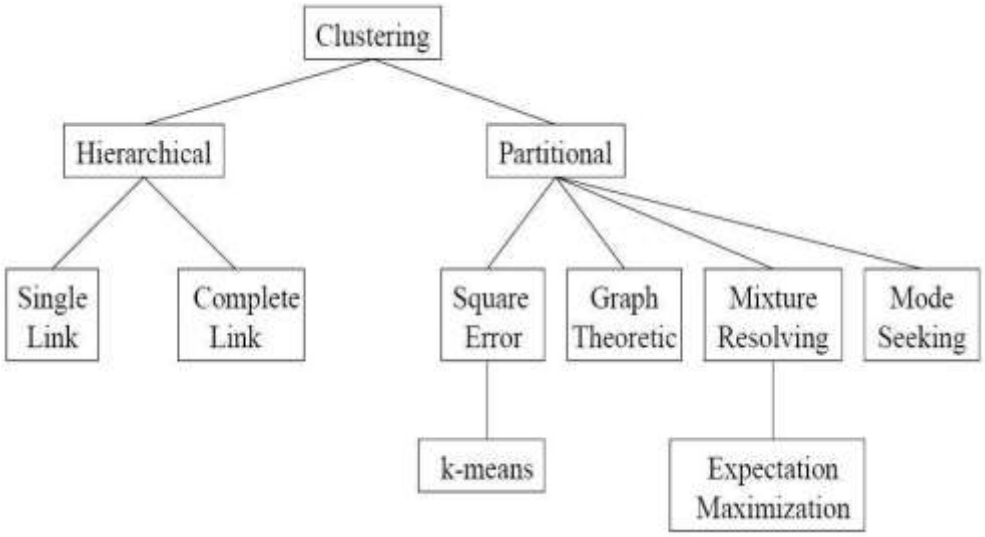


Figure 2: Types of Clustering

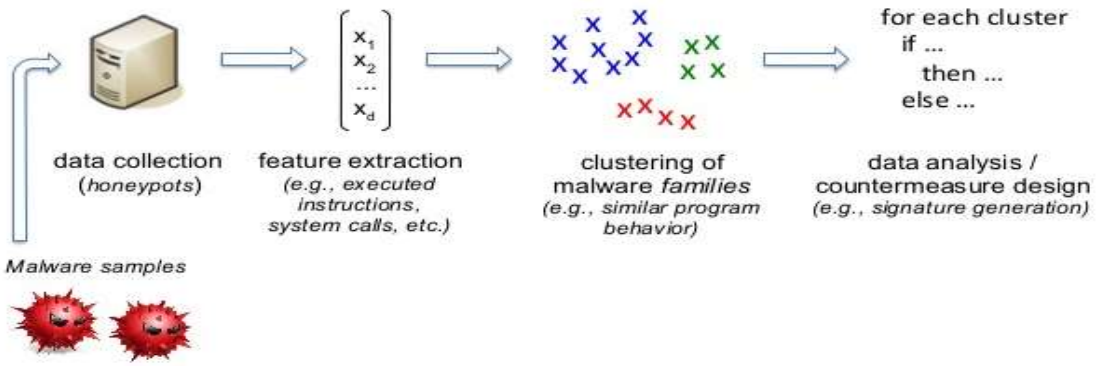


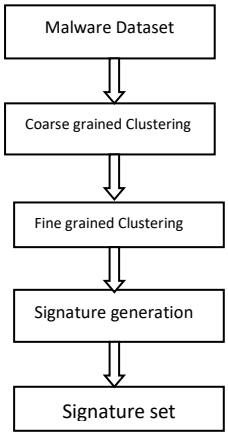
Figure 3: Malware Clustering Process

Based on detailed system events, Bayer proposed a scalable malware clustering technique. This technique was not good for network level behavioral signatures.

Perdisci proposed a method to get malware clusters that can aid the automatic creation of fine quality malware signatures. This process proved very useful to detect botnet command and control. With this processing time is being reduced from several hours to few minutes.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Overview of proposed technique



Authors observed various peculiar points about malware behavioral clustering:-

1. Draw plot between BIRCH radius and separation index
2. Draw plot between BIRCH radius and cohesion index
3. Time required completing the clustering and signature generation process for different values of BIRCH radius.
4. Number of coarse and fine grain clusters generated for different values of the BIRCH radius.

Analysis is done on 65000 different malware samples explains the importance of work that reduced processing time from several hours to few minutes.

Perdisci designed fully automated clustering validity tool for the detection of malware. The results after analysis shows that VAMO outperforms with voting based malware detector. Some of important results are shown here.

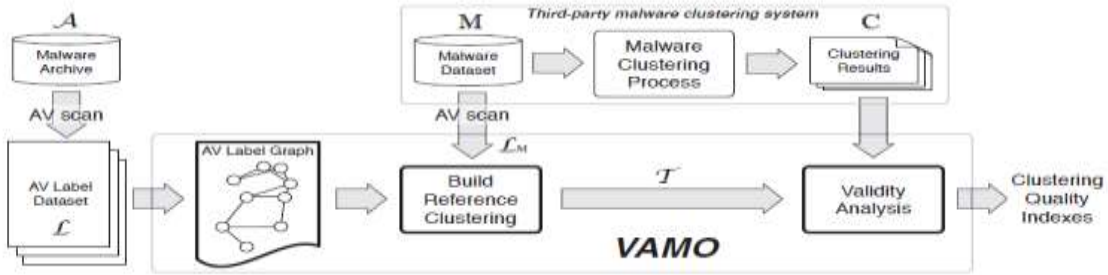


Figure 4: VAMO System Overview

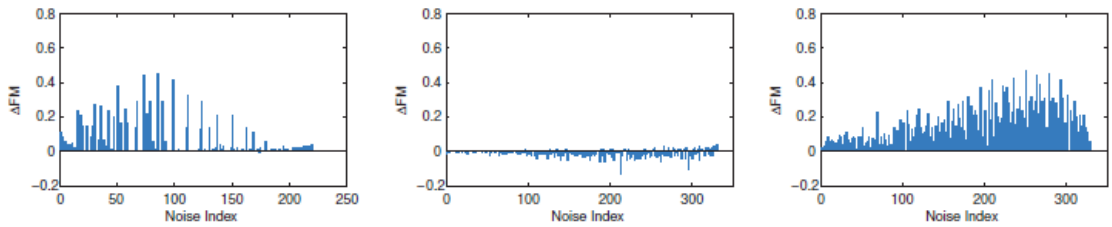


Figure 5: VAMO vs Majority Voting

Ciprian proposed malware clustering using suffix trees. Suffix tree data structure is used to find long enough substring that corresponds to portions of a program's code. Based on common substring clustering is done. Deq distance method is used to calculate distance between opcode sequences. A clustering technique based on Ukkonen's suffix tree is used.

Usha Narra explained clustering technique for malware detection. Author used K-means and Expectation

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Maximization (EM) for malware clustering.

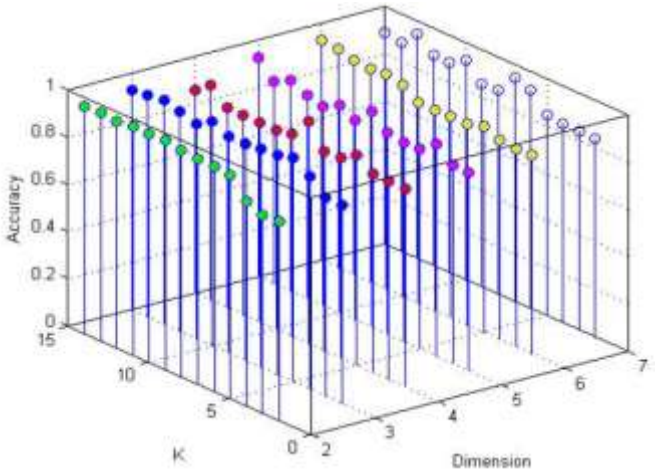


Figure 6: Accuracy measure for EM clustering

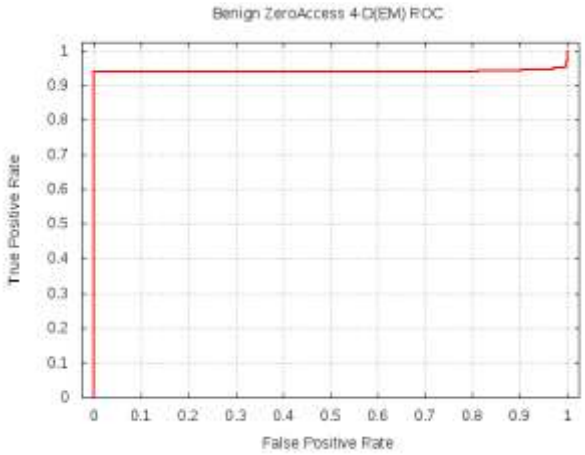


Figure 7: ROC curve of EM Clustering (4D)

EM and K-means algorithms are used to cluster 8000 malware samples. Expectation maximization clustering is an iterative method for finding maximum a posteriori map (MAP) in order to determine parameters in statistical model where the model focuses on unobserved latent variables. K means is a technique of vector quantization and used for cluster analysis.

Matthew Asquith proposed a technique for clustering of malware metadata and use of data structure called aggregation overlay graphs. Mondal and Deshpande introduced the idea of aggregation overlay graph in which

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

virtual nodes are added to bipartite graphs connecting each node of a sub graph to a new virtual node reduces the total number of edges.

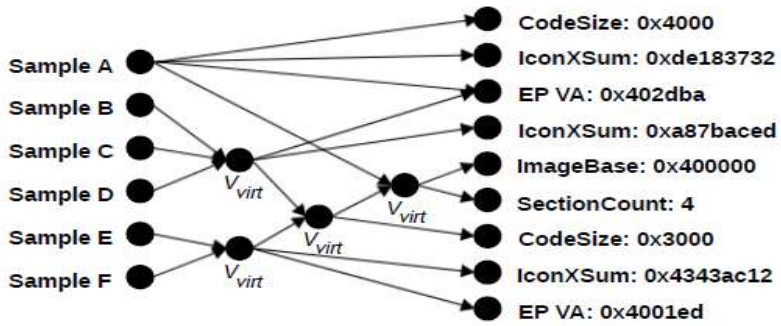


Figure 8: An aggregation overlay graph designed from metadata

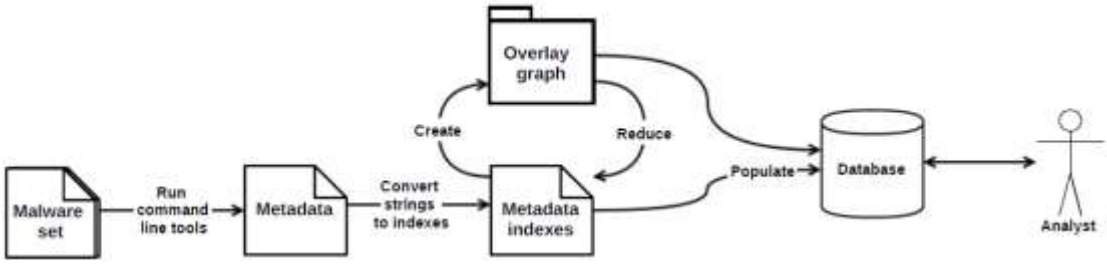


Figure 9: Complete system with aggregation overlay graph

Spectral malware behavior clustering

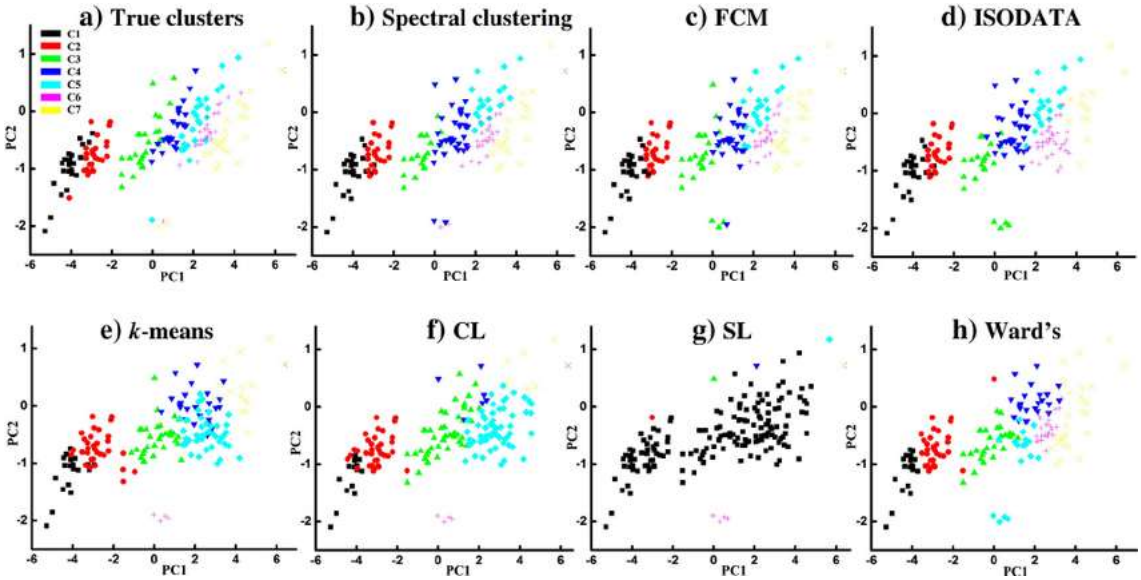


Figure 10: Comparison among different clustering methods for clustering e-nose dataset of adulteration. (a) True clusters, (b) spectral clustering, (c) FCM, (d) ISODATA, (e) k -means, (f) complete linkage, (g) single linkage, and (h) Ward's linkage. [10]

Chris Giannella and Eric Bloedorn proposed a technique using spectral malware behavior clustering better than Hierarchical agglomerative clustering. The run time of proposed algorithm is smaller than Rieck's algorithm.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

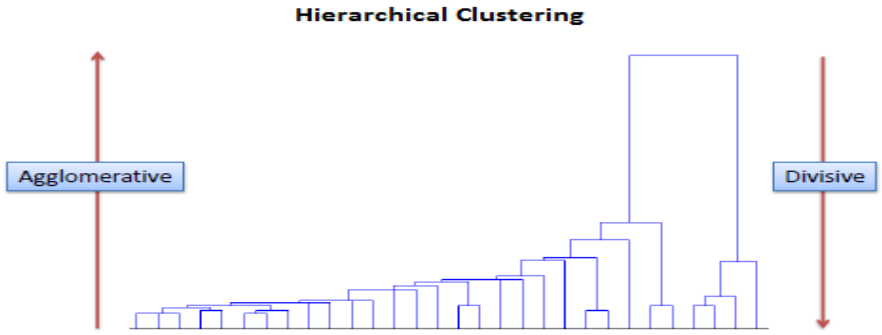


Figure 11: Agglomerative Clustering

Peng Li et al. explained about challenges and scope of malware clustering evaluation. Authors focused on the importance of BCHKK algorithm. The important analysis lies in identifying the capability of plagiarism detectors in malware clustering. Some important results are depicted in following graphs.

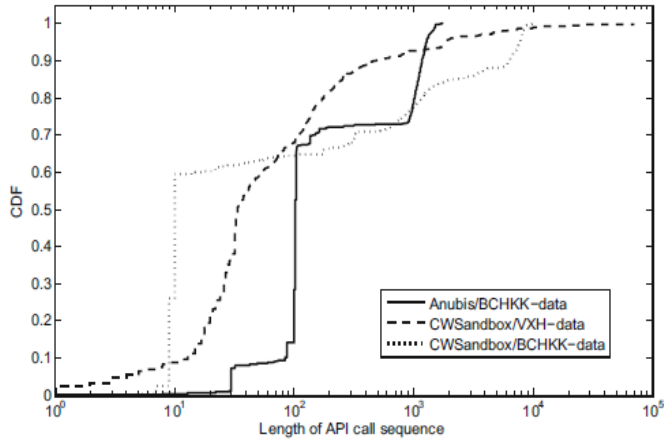


Figure 12 : API sequence call length extracted from BCHKK data using CWSandbox

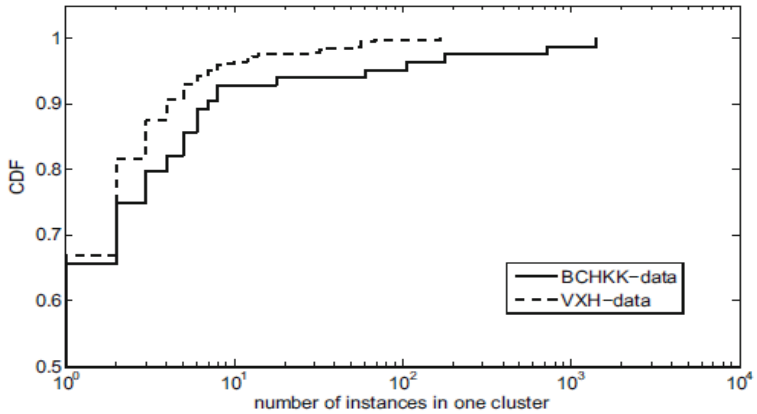


Figure 13: Reference cluster size distribution of BCHKK data and VXH data

CONCLUSIONS

Large number of research papers has been written in the field of malware clustering techniques. Our purpose in this paper is to address the problem of malwares and its solutions using clustering. After analyzing various articles in concerned domain, we found that clustering can be used for real time mitigation of malwares.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

REFERENCES

[1] Ulrich, B., Paolo, M. C., Clemens, H., Christopher, K., & Engin, K. (2009, February). Scalable, behavior-based malware clustering. In *Proceedings of Network and Distributed System Security Symposium*.
[2] Perdisci, Roberto, Wenke Lee, and Nick Feamster. "Behavioral Clustering of HTTP-Based Malware and Signature Generation Using Malicious Network Traces." *NSDI*. 2010.
[3] **Roberto Perdisci**, Davide Ariu, Giorgio Giacinto. "Scalable Fine-Grained Behavioral Clustering of HTTP-Based Malware." ***Computer Networks***, *Special Issue on Botnet Activity: Analysis, Detection and Shutdown*, 57(2):487–500, 2013.
[4] **Roberto Perdisci**, ManChon U. "VAMO: Towards a Fully Automated Malware Clustering Validity Analysis". *28th Annual Computer Security Applications Conference*, **ACSAC 2012**
[5] Guofei Gu, Roberto Perdisci, Junjie Zhang, Wenke Lee. "BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection". *USENIX SecuritySymposium* 2008.
[6] **Roberto Perdisci**, Giorgio Giacinto, Fabio Roli. "Alarm clustering for intrusion detection systems in computer networks". *Engineering Applications of Artificial Intelligence* (**EAAI**), 19(4), 2006, pp. 429-438.
[7] Giorgio Giacinto, **Roberto Perdisci**, and Fabio Roli, "Alarm Clustering for Intrusion Detection Systems in Computer Networks".*International Conference on Machine Learning and Data Mining in Pattern recognition*, **MLDM 2005**.
[8] Oprişa, Ciprian, George Cabău, and Gheorghe Sebestyen Pal. "Malware clustering using suffix trees." *Journal of Computer Virology and Hacking Techniques* 12.1 (2016): 1-10.
[9] Narra, U., Di Troia, F., Corrado, V. A., Austin, T. H., & Stamp, M. (2015). Clustering versus SVM for malware detection. *Journal of Computer Virology and Hacking Techniques*, 1-12.
[10] Hong, Xuezhen, Jun Wang, and Guande Qi. "Comparison of spectral clustering, K-clustering and hierarchical clustering on e-nose datasets: application to the recognition of material freshness, adulteration levels and pretreatment approaches for tomato juices." *Chemometrics and Intelligent Laboratory Systems* 133 (2014): 17-24.
[11] Giannella, Chris, and Eric Bloedorn. "Spectral malware behavior clustering."*Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*. IEEE, 2015.
[12] Li, P., Liu, L., Gao, D., & Reiter, M. K. (2010, September). On challenges in evaluating malware clustering. In *International Workshop on Recent Advances in Intrusion Detection* (pp. 238-255). Springer Berlin Heidelberg.
[13] Bist, Ankur Singh, and Sunita Jalal. "Identification of metamorphic viruses."*Advance Computing Conference (IACC), 2014 IEEE International*. IEEE, 2014.
[14] Bist, Ankur Singh. "Detection of metamorphic viruses: A survey." *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. IEEE, 2014.
[15] Bist, Ankur Singh. "Classification and identification of Malicious codes." IJCSE. 2012.
[16] Bist, Ankur Singh. "Hybrid model for Computer Viruses: an Approach towards Ideal Behavior." *International Journal of Computer Applications* 45 (2012).