

Overview of Malware Analysis and Detection

Aziz Makandar
Professor
Karnataka State Women's University
Vijayapura

Anita Patrot
Research scholar
Karnataka State Women's University
Vijayapura

ABSTRACT

Identify a malicious data in a several types of files is a challenging task. Malware is a computer virus this is also a name given to a group of malicious data like umbrella to all types of malicious data like virus, worm, Trojan and so on. Several methods have been devised to smooth the progress of malware analysis and one of them is through visualization techniques. The visualization technique is used to visualize the features of malware or variants in a gray scale image of malware. The malware is visualized as an image with the concepts of image processing techniques it will identify the malware. The malware behaviors are identified in one of them such as encrypted malware, polymorphic malware, metamorphic malware, and obfuscated which have the ability to change their code as they propagate. In this paper different types of malware are discusses briefly with their categorization of malware families. The different techniques are used to identify and classify malware. Which is motivated especially on behaviors of malware samples which are similar in texture and some extent through this we can classify the malware data. This paper provides an overview of existing malware detection techniques with the different types of malware family descriptions.

General Terms

Image Processing, Pattern Recognition

Keywords

Malware, Static Analysis, Dynamic Analysis, Detection, Classification, Visualization.

1. INTRODUCTION

The security issues are more challenging task now days. The malicious behavior identify with different available tools online is a temporary to detect and identify different types of malware because of increasing malware day to day. The increasing variants of malware are a challenging task to the antivirus vendors. Today along with the development of the internet in the all fields which motivates that try to implement such a method or tool to detect and identify different incoming malware automatically. Then uses are increasing day by day uncontrollable without any authority of a particular author. The number of malware distributes more especially for financial profits are increase and malware is created by various automated tools and methods in online. Malware types and families are rapidly increasing in the interconnection of emerging information and technology systems as well as the number of malware attackers is also increasing more.

The effect of malicious data affect the various computer networks, infrastructures, services, file sharing, online social networking, and Bluetooth wireless networks. In existing systems the first malware issue is dedicated journal to study about the malware in 1988. It analyzes the proposed data from the hackers then they start to develop new features for identify

the malware variants and design new strategies to control and manage the plan activities of disturbing process. The traditional approach to detect malicious data is a signature based automatic detection frame work which is used to detect and classify the malware. The signatures look like group of footprints of malware due to the rapid increase in malicious data in large number of malware [1].

The approaches towards the malware identification and detection are a recent issue as well as challenging to the security issue. The novel approach is proposed by the author's one who used the characteristics and analyze the malware at broader level. Which is used a executable files are converted into a binary string 0's and 1's by combining these binary values a vector is represented later these vectors are converted into an image [2]. The overcome from the traditional signature based method malware analysis techniques are used in two broad categories. The static and dynamic analyses of malwares are used from several decades the malware progress is uncontrollable and increased various types of malware families. The detection and identifying malware is done by antivirus software, dynamic analysis is for estimation performance of malware still not good enough for detecting malware. Hackers are going to get more valued information and data from online banking, online transactions, online shopping, etc. for protecting private data and information is a big challenge to detect malicious data through network [6] and virtual machine environment [26-27]. A Malware image analysis [29] technique helps the analyst to understand the risks and intensions associated with malicious code.

The security specialists use all possible techniques and methods to stop and remove the threats while the malware developers utilize new types of malwares that pass implemented security features by applying pre-processing [30]. In this study looking into malwares to understand the definition, categories of malwares as shown in the figure.1, the variants of malwares, detecting techniques and classifiers in order to contribute to the process of protection and security enhancement[31]. Malicious data has been designed to achieve some targets such as collecting sensitive data accessing private computer systems sometimes harming the systems. The malware can reach the systems in different ways and through multiple media, the most common way is the downloading process from the internet, once the malware finds its way to the systems based on various variants of malware. The malware hides itself in the case of spyware, these hidden malware sends critical information about the computer to the source, based on above challenges it is critical to carry out an in depth analysis to understand the malware for superior detection and removal chances. The direction of malware detection is a challenging task to identify new variants of newly created malware even though the researchers find out the solutions to detect malicious data through various techniques. The visualization technique is also used for identification of malware variants more effective

compare to other technique [33] through the classification and detection of malware by using image processing techniques. Therefore, our research is put forth in developing an efficient algorithm for malware detection and classification in further research work. This paper highlights the brief introduction of malware and its techniques to identify as well as classify malware samples into particular family.

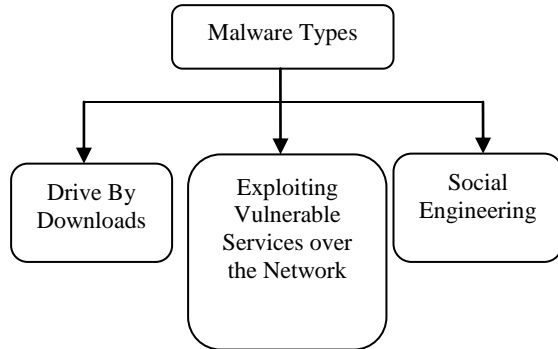


Figure.1 Malware Behaviour

2. MALWARE ANALYSIS

Purpose of malware and functionality of a given samples in gray scale images such as sample virus, Trojan horse, etc. This is important prerequisite for the development of removal tools that can thoroughly detect and remove malicious data from the infected files. Traditionally the malware analysis process follows the manual process for the malware is tedious as well as time-intensive for the increasing malwares on network for detectors it is difficult to zero attack from the hackers. The malware analysis is broadly categorized static analysis and dynamic analysis as shown below figure2.

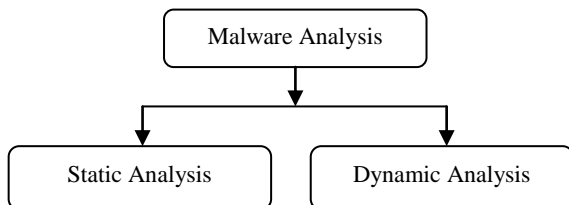


Figure.2. Malware Analysis

2.1 Static Analysis

Analysing software without executing it is called static analysis this analysis techniques can applied on different representations of binaries of malicious data it helps in the finding of memory corruption as well as prove the correctness of the system. Static analysis can be used on the basis of binary representation. The various techniques are used for static malware analysis to detect malicious data from the binaries.

2.2 Dynamic Analysis

The analysis of behaviour of a malware code which interaction with the system while it is being executed in a controlled environment such as virtual machine and sandbox etc. The malware is debugged while running using a debugger such as GDB or Win DGB to watch the behaviour of the malware step by step while its instructions are being processed by the processor and their live effects on RAM. Several online automated tools exist for dynamic analysis of malwares example Anubis [5], TT Analyser [6], and Norman Sandbox [19]. The analysis system is required to have an appropriate representation of malwares which is used for

classification based on the similarity measures or features vectors. This is done by watching and logging the behaviour of the malware while running on the host in virtual machines and sandboxes are extensively used for this type of analysis [25].

3. EXISTING WORK

A large number of new malware samples detected at Antivirus vendors every day requires an automated approach so as to limit the number of samples that require close human analysis several artificial intelligence techniques particularly machine learning based techniques have been used in literature for automated malware analysis and classification such as k-nearest neighbour[26], Association Rule (AR), Support Vector Machine(SVM), Decision Tree(DT), Random Forest(RF), Naive Bays (NB)[27] and Clustering have been proposed for detecting and classifying unknown samples in to their known malware families. A few of these used in the literature are discussed in this section. Schultz.et.al. [1] were the first to introduce the concept of data mining for detecting malwares they used three different static features for malware classification, portable executable, strings and byte sequences. They used dataset consisted of 4266 files including 3265 malicious and 1001 benign programs, A rule indication algorithm called ripper [8]. Natraj.et al. [10] proposed a simple and fast method for visualizing and classifying malwares using image processing techniques which visualize malware binaries as gray scale images. A k-nearest neighbour technique with Euclidean distance method is used for malware classification through it is very fast method compared to other malware analysis methods, the limitations of this method is an attacker can adopt counter measures to beat the system because this method uses global image based features. Natraj et al. [2] explored the advantages of static analysis method. In their work, they introduced a simple and effective method for visualizing and classifying malware using image processing techniques this technique also exhibits resilience to popular obfuscation techniques such as encryption. Further, they proposed SigMal framework based on signal processing techniques to process large amounts of binary samples.

Natraj.et.al [11], the author compared binary texture based analysis based on image processing techniques with that of dynamic analysis. They observed that classification using this method is faster as well as scalable and is comparable to dynamic analysis in terms of accuracy they also proved that this approach can robustly classify large number of malwares into two groups such as packed and unpacked. The limitation is that this method is vulnerable to knowledgeable adversaries who can obfuscate their malicious code to defeat texture analysis. Natraj.et.al [12], proposed a SigMal framework based on signal processing techniques which is operate on large amounts of binary samples and it has been observed that many samples are received by such systems are variants of previously seen malware also they retain some similarities at the binary level. This system improves the state-of-the-art by leveraging techniques borrowed from signature from the samples. This system uses an efficient-neighbour search technique which is also scalable. Kong.et.al. [13] they presented by using technique, Function call graph and extracted the fine grained features based on function call graph for each malware samples are then observed the similarity between two malware programs by applying discriminate distance matrices learning which clusters the malware samples belonging to the same family while keeping the different clusters separate by a marginal distance.

Kyoung Soo Han. Jae Hyun. Boojoong kang [32], in this they are converting the binary samples into gray scale images then applying the entropy graph and threshold for each image then based on that they are classifying the malware images. This gives the effectively distinguish malware families. Syed Zainudeen Mohd Shaïd et.al [33], they highlights the findings in visualizing malware behavior and various variants that shows the potential benefits for malware classification this also identify the malware with high accuracy. They applied the color map which is Cold and Hot method for coloring the gray scale image according to its malicious behavior and they identified malware based on color code which is red, if it is red color that part of image is malicious otherwise blue color, which is non malicious while comparing this with Natraj paper they got high accuracy upto 99.33%. It's very easy to identify malware variants through color.

It is discussed in previous sections related to malware analysis followed to collect data from the various research papers based on their existing study of malware detection and propagation as well as behavior of malware. The malware ratios are shown in figure 3. Then grouped the concepts, topics and ideas from the papers group them in specific structure as required as shown in the above table 1. A malware detection technique includes the methods such as static, dynamic and hybrid analysis. The existing methods, techniques, and approaches are grouped together to get detail information regarding their contribution and the resultant accuracy of individual authors.

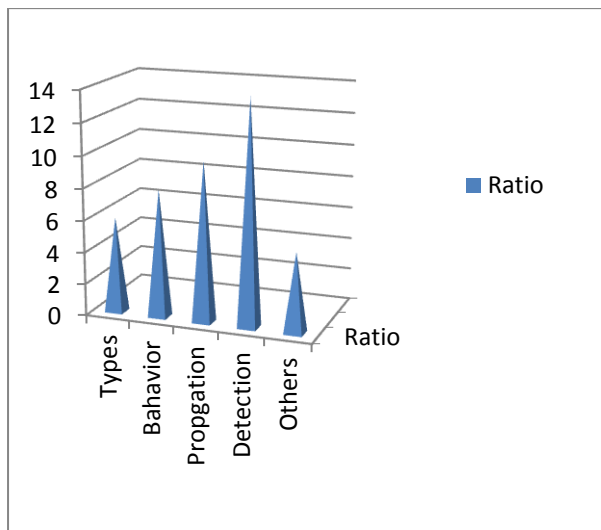


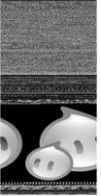


Figure 3: Categorization of Malicious Data

Syed Zainudeen Mohd Shaïd et.al [32], they highlights the findings in visualizing malware behavior and various variants that shows the potential benefits for malware classification this also identify the malware with high accuracy. They applied the color map which is Cold and Hot method for coloring the gray scale image according to its malicious behavior and they identified malware based on color code which is red, if it is red color that part of image is malicious otherwise blue color, which is non malicious while comparing this with Natraj paper they got high accuracy upto 99.33%. Its very easy to identify malware variants through color. Kyoung Soo Han. Jae Hyun. Boojoong kang [33], in this they are converting the binary samples into gray scale images then applying the entropy graph and threshold for each image then based on that they are classifying the malware images. This gives the effectively distinguish malware families.

Table.1 Literature Survey

Author	Literature Survey		
	Techniques	Results	Contribution
Natraj[10]	Image Processing techniques, KNN and Euclidean Distance Method.		They proposed an algorithm is fastest result.
Sayed [33]	Cold and Hot color Map Done on Gray scale image		Malware behaviour visualization by using colour map
Kuyong Soo Ha[32]	BMP Conversion, Entropy graph and classification		Conversion of binary samples into gray scale image and then entropy graph

4. VISUALIZATION OF MALWARE

The malware samples are visualized in 2 dimensional that is images as shown in the figure 4. These images are having similarity between them for those external behavior or global features of these images can be used further for classification to perform the texture analysis of these images for better classification results. Malware can be used to represent binaries [28]. The following figure 3, shows the conversion of malware binaries samples into gray scale images by applying 8 bit vector for a each pixel and resultant texture image is a gray scale image of binaries. The comparative factor of individual malware family occurred during the access of network for many reasons. As discussed earlier about the various types of malware families such as virus, worm, logic bomb, bootnet, spam, sniffers, Trojan horse, trapdoor, cookies, adware, and spyware. Each individual malware families are compared with the factors such as, creation techniques, execution environment, propagation media, and negative impacts. The malware images are formed by using 2D array this can be visualize gray scale images as shown in figure 4 the range between 0 to 255 which indicates 0 as black and 255 as white.

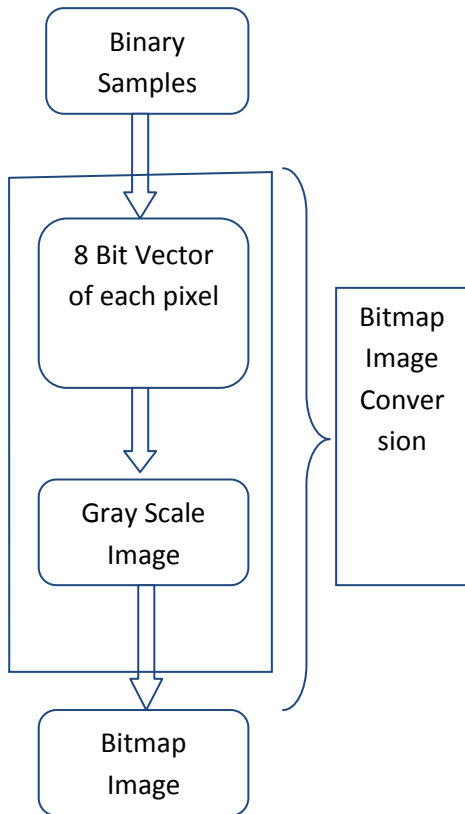


Figure .4 Bitmap Conversions

Table.2 Data set of Malware

Dataset	Dataset with number of samples			
	No of Family	No of Samples	Types of Sample	Global Behaviors
Manhuer Dataset	24	3131	Grayscale image	Texture
Microsoft Dataset	100	5162	Grayscale image	Texture

The malware dataset consist of 3131 samples and 5162 samples from the Microsoft Dataset which is in a gray scale images and every sample look like a complete texture images each portion of the malware having different texture patterns. We can observe in set of different malware family in figure 5. Such as instances, obfuscation, Malex!J, Azero,etc.

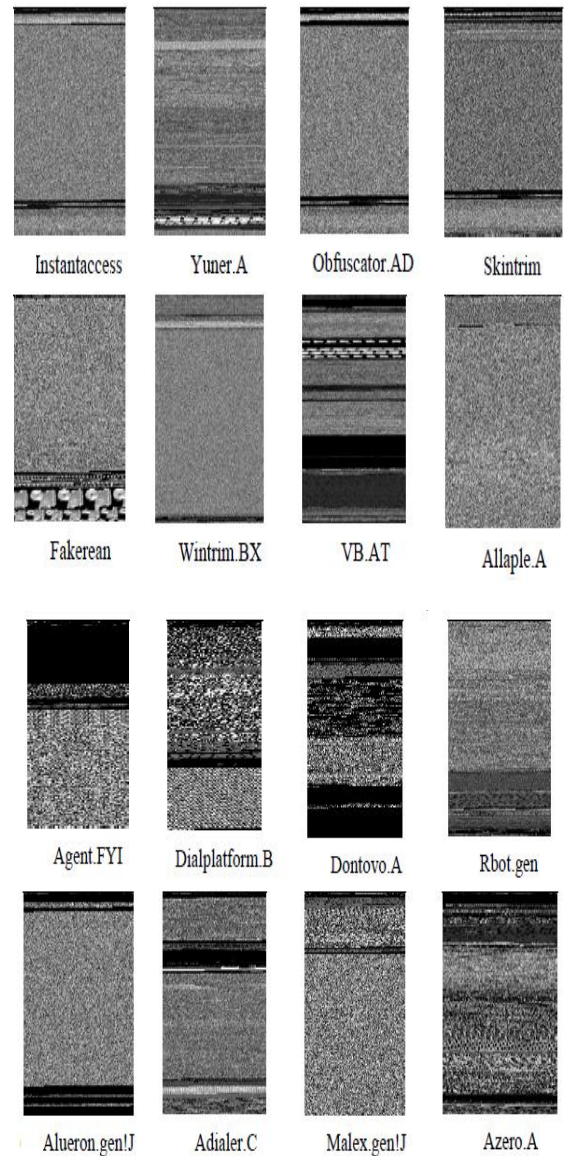


Figure.5 Various Malware Families

5. CONCLUSION AND FUTURE WORK

This paper provides the brief introduction of malware and its type which are in the form of variants the concept of malware analysis, detection and classification techniques. The developer tries to find out the new techniques to detect malicious data affected files and software. The detectors analyse malware behaviour continuously and try to resist the techniques and strategies based on types of variants. Malware detection has a still challenging task to detect new malicious data. The increasing developments in information technology also arises lots of hackers to steal the privacy data by applying malwares in various forms. Detection of malware is challenging task to identify these behaviours and patterns is that targets related areas image and signal processing, computer vision techniques and so on. The machine learning techniques that are used to detect and classify malwares are not adequate to handle challenges arising from huge amount of data.

6. ACKNOWLEDGMENTS

This research work is funded by UGC under Rajiv Gandhi National Fellowship (RGNF) UGC Letter No: F1-17.1/2014-15/RGNF-2014-15-SC-KAR-69608, February, 2015.

7. REFERENCES

- [1] Schultz, M., Eskin, E., Zadok, F. and Stolfo, S. "Data Mining Methods for Detection of New Malicious Executables". Proceedings of 2001 IEEE Symposium on Security and Privacy, pp. 38-49, 2001.
- [2] Bayer, U., Moser, A., Kruegel, C. and Kirda, E. "Dynamic Analysis of Malicious Code". Journal in Computer Virology, pp.67-77, 2006.
- [3] Infographic: The State of Malware, 2013.
- [4] Anubis. <http://anubis.isecslab.org/>
- [5] Bayer, U., Kruegel, C. and Kirda, E. TT, "Analyze: A Tool for Analyzing Malware". Proceedings of the 15th European Institute for Computer Antivirus Research Annual Conference, 2006.
- [6] Ahmed, M. "NIDS: A Network Based Approach to Intrusion Detection and Prevention". Computer Science and Information Technology - Spring Conference, 2009.
- [7] D. L. Donoho, "De-noising by soft-thresholding," IEEE Trans. Information Theory, 1995.
- [8] Cohen, W. "Fast Effective Rule Induction". Proceedings of 12th International Conference on Machine Learning, San Francisco, pp. 115-123, 1995.
- [9] Kolter, J. and Maloof, M. "Learning to Detect Malicious Executables in the Wild". Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2721,2744, 2006.
- [10] Nataraj, L., Karthikeyan, S., Jacob, G. and Manjunath, B. "Malware Images: Visualization and Automatic Classification". Proceedings of the 8th International Symposium on Visualization for Cyber Security, Article No. 4. 2011
- [11] Nataraj, L., Yegneswaran, V., Porras, P. and Zhang, J. "A Comparative Assessment of Malware Classification Using Binary Texture Analysis and Dynamic Analysis". Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, pp. 21-30, 2011.
- [12] Nataraj, L. "SigMal: A Static Signal Processing Based Malware Triage". 2013
- [13] Kong, D. and Yan, G. Discriminant. "Malware Distance Learning on Structural Information for Automated Malware Classification". Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems, pp. 347-348, 2013.
- [14] IDAPro. https://www.hex-rays.com/products/ida/support/download_freeware.shtml
- [15] OllyDbg. <http://www.ollydbg.de/>
- [16] Indyk, P. and Motwani, R. "Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality". Proceedings of 30th Annual ACM Symposium on Theory of Computing, Dallas, pp.604-613, 1998.
- [17] Tian, R., Batten, L. And Versteeg, S. "Function Length as a tool for malware classification". Proceedings of the 3rd international conference on malicious and unwanted software, Fairfax, pp. 57-64, October 2008
- [18] Zolkipli, M.F. and Jantan, A. "An approach for malware behavioural identification and classification". Proceedings of the 3rd international conference on Computer research and Development, Shanghai,, 11-13 pp.191-194, 2011.
- [19] Rieck, k., Trinius, P., Willems, C. And Holz, T, "Automatic Analysis of malware behaviour using machine learning". Journal of Computer Security, 19,639-668, 2011.
- [20] Anderson, B., Quist, D., Neil, J., Storlie, C. And Lane, T. "Graph based malware detection using dynamic analysis". Journal in computer virology, pp. 247-258. 2011.
- [21] Bayer, U., Comparetti, P.M., Hlauschek, C. And Kruegel, C. "Scalable, Behaviour-based Malware Clustering". Proceedings of the 16th Annual Network and Distributed System Security Symposium, 2009.
- [22] Tian, R., Batten, L. And Versteeg, S. "Function Length as a tool for malware classification". Proceedings of the 3rd international conference on malicious and unwanted software, Fairfax, pp. 57-64, 2008.
- [23] Aziz Makandar, Bhagirathi Halalli, "Image Enhancement Techniques Using Highpass Filter and Lowpass Filters". International Journal of Computer Applications (0975-8887) Volume 109 – No. 14, January 2015.
- [24] Anderson, B., Storlie, C. And Lane, T. "Improving Malware Classification: Bridging the static/dynamic gap". Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence, 2012.
- [25] Garfinkel, T. and M. Rosenblum, "A virtual machine introspection based architecture for intrusion detection". pp. 191—206, 2003.
- [26] Lagar-Cavilla, H.A., "Flexible Computing with Virtual Machines". 2009.
- [27] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, pp. 21–27, 1967.
- [28] Aziz Makandar, Anita Patrot, "Computation of Pre-processing Techniques for Image Restoration". International Journal of Computer Applications (0975-8887) Volume 113 – No. 4, March 2015.
- [29] I. H. Witten, E. Frank, and M. A. Hall, Data "Mining: Practical Machine Learning Tools and Techniques". 3rd ed. Morgan Kaufmann Inc, 2011.

- [30] R. K. Shahzad, S. I. Haider, and N. Lavesson, "Detection of spyware by mining executable files," in Proceedings of the 5th International Conference on Availability, Reliability, and Security. IEEE Computer Society, pp. 295–302, 2010.
- [31] Aziz Makandar, Anita Patrot and Bhagirathi Halalli, "Color Image Analysis and Contrast Stretching using Histogram Equalization". International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682 Vol.27, No.27, July 2014.
- [32] Kyoung Soo Han ,Jae Hyun Lim, Boojoong Kang, Eul Gyu Im, "Malware Analysis Using Entropy Graphs", Springer-Verlag Berlin Heidelberg 2014, Int. J. Inf. Secur. 14:1–14 DOI 10.1007/s10207-014-0242-0, 2015.
- [33] Syed Zainudeen Mohd Shaid, Mohd Aizaini Maarof (2014). Malware Behavior Image for Malware Variant Identification. IEEE, International Symposium on Biometric and Security Technologies (ISBAST).