International Journal of Computer Applications Technology and Research
Volume 7–Issue 03, 129-132, 2018, ISSN:-2319–8656

Malware Family Detection Approach using Image Processing Techniques: Visualization Technique

Poonam Parmuval
Department of Computer Engineering
BVM Engineering College,
V.V.Nagar, India

Mosin Hasan
Department of Computer Engineering
BVM Engineering College,
V.V.Nagar, India

Samip Patel
Department of Computer Engineering
BVM Engineering College,
V.V.Nagar, India

Abstract: - The risk of malicious software has increased a lot since last decade as the use of internet has increased drastically. According to Avast Test report, 22,000 to 25,000 new malware have been reported every day. Even though huge malwares having different structures are introduced every day, their nature of working is almost similar to old malwares. The malwares with the similar functionalities are considered to be the member of the same family. Classification and detection of malware family are important to design its signature of anti-malware software. In this article, we represent the concise study carried out on detecting various malware family. This paper mainly focus on visualization technique for classifying malware family. Visualization technique uses image processing approaches to classify the malwares. The malware executable binary files are transformed into image and this images are used to detect the family of malware.

Keywords: - Malwares, Malware Family, Malware Visualization, Malware Detection, Malware family classification

1. INTRODUCTION

Nowadays it becomes difficult to live without mobile, internet, computers. As the digital world grows today the security and protection of computer system have become biggest concern. Malware: Malware is malevolent software which is designed to breach the security of the system or to harm the computer's operating system [1] [2]. Harms caused by malwares can be stealing personal information, locking file system, password stealing, showing unwanted content, etc. This malwares are classified into different types.

Adware: Adware is an advertisement-focused application that installs themselves on systems [2].

Spyware: It spies on activities performed by victims and tracks the internet activities to send an advertisement to the system [2].

Virus: It is contagious code that link itself to another software and then regenerates itself [2] [3].

Worm: Worms are the self-replicating code that deletes or corrupts the files on the computer. It works to eat operating system files and data files [2].

Trojan: It arrives as useful to the user and tries to enter into victims system. It discovers personal information (financial).

Ransomware: Ransomware is introduced to lock the data of victim. Then attacker demands to pay for unlocking data [2].

Rootkit: Attacker would gain root permissions and install various applications and utilities (maybe malicious), called "kit," on the victim's system [4].

Key loggers: It note every key press on the keyboard and gains the important information like username, password, and email content, etc.

1.1 Malware Family

Many approaches have been made to detect and prevent malwares but the attacker advances their technique and develops many new malwares. This makes the traditional anti-virus software difficult to resist the violation of malicious codes. According to AV-Test Report approx. 250,000 new malwares have been reported every day [5]. The new malwares are generated from the previous malwares with the help of techniques like encryption, obfuscation, mutation, etc. [1]. This have been proved by researchers working on malware. The variants can also be generated using executable packers. Packer is a utility that applies compression and encryption on executables to make them undetectable by malware scanners [6]. The attackers generate many new variants of malware by modifying the code or by using the packers to evade the detection by current anti-malware software. Though the variants seem new to anti-malware, their functionalities remain similar to old malwares. The malwares with similar functionalities are considered to be the member of the same family [1].

The following are few family of malwares [1].

Agent: Agent is family that most of its variant download and install adware or malware on the attacked system. It may also change the configuration properties for Windows [7].

Allaple: Win32/Allaple is network worm family that spread to other devices connected to a LAN and perform denial-of-service (DoS) attacks [8].

Fakerean: This family of security program pretends to examine your system against malware, and generate a report that shows lots of malwares. The program will demand money to scan deeply [8].

Rbot: Rbot is a family of backdoor malwares who allows attackers to control victim's computers [9].

C2LOP: It is a Trojan family that changes browser settings, a bookmark to advertisements, show advertisements [9].

Once the malware family is detected, it becomes easy to know the vulnerabilities of the malware and hence easy to prevent it. To prevent such malicious attacks many defensive techniques have been developed. But the group of attackers comes up with the new solutions to evade these defence techniques and generate thousands of new variants. So it continuously required to identify all new malwares and find their solutions.

1.2 Malware Detection Techniques

The techniques for malware detection are broadly classified into '*static analysis'* and '*dynamic analysis'.*

Static Analysis:

Analysing malware without running them are considered as Static analysis [10]. This approach includes Signature-based, Permission-based, and Component-based analysis. The Signature-based method extracts the semantic patterns from malware and generates a unique signature to match particular malware [3]. It won't detect the variant or unknown malware. The Permission-based method identifies dangerous permission requested by malware to detect malware [10]. The Component-based method disassembles the malware to extract and analyse the important components (i.e. opcodes, activities, services, receivers etc.), for identification of the vulnerable attacks. The major limitation of the static analysis strategy is that these techniques fail to identify malicious behaviour obfuscated malware [10].

Dynamic Analysis:

In dynamic analysis, the behaviour of malicious code is observed and noted by running the malware executable in a virtual environment or emulator [10] [11]. It monitors the system level calls and the attributes. These strategies give more accurate result than static analysis but it is the time-consuming process as it requires several executions in a virtual machine [11].

2. VISUALIZATION TECHNIQUE

Dynamic analysis can identify malwares more accurately but it is very time consuming process. L. Nataraj [1] has proposed a technique which depend on image processing methods to classify malware families. This technique is called Visualization technique. The malware executable binary files or PE files are converted into image and this image is used to identify the type of malware.

Visualization technique has two main strands. One focuses in Dataset and dataset generation technique and second focuses on image processing aspect.

2.1 Data Set for Malware family:

Maligm dataset comprises of 9339 malware samples distributed in 25 malware families with the varying number of variants per family [1]. It contains malwares in form of grayscale images.

Table 1 Malimg Dataset families

No.	Class	Family Name	No. of Variants
1	Worm	Allaple.L	1591
2	Worm	Allaple.A	2949
3	Worm	Yuner.A	800
4	PWS	Lolyda.AA 1	213
5	PWS	Lolyda.AA 2	184
6	PWS	Lolyda.AA 3	123
7	Trojan	C2Lop.P	146
8	Trojan	C2Lop.gen!G	200
9	Dialer	Instant access	431
10	Trojan-Downloader	Swizzor.gen!I	132
11	Trojan-Downloader	Swizzor.gen!E	128
12	Worm	VB.AT	408
13	Rogue	Fakerean	381
14	Trojan	Alueron.gen!J	198
15	Trojan	Malex.gen!J	136
16	PWS	Lolyda.AT	159
17	Dialer	Adialer.C	125
18	Trojan-Downloader	Wintrim.BX	97
19	Dialer	Dialplatform.B	177
20	Trojan-Downloader	Dontovo.A	162
21	Trojan-Downloader	Obfuscator.AD	142
22	Backdoor	Agent.FYI	116
23	Worm:AutoIT	Autorun.K	106
24	Backdoor	Rbot!gen	158
25	Trojan	Skintrim.N	80

2.2 Image Processing

This strand is divided in three phases that are Malware image generation, Feature extraction and classification.

2.2.1 Malware Image Generation

The malware binaries are grouped into 8-bit vectors which represent hex value from 00 to FF. These vectors are

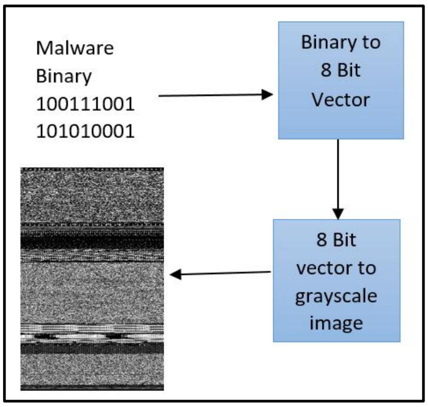

Fig 1. Malware Binary to image [1]

represented as pixel values i.e. intensity of grayscale image

ranging from 0-255 [1]. The width of the images are predefined but the height of images are allowed to vary based on size. By visualizing malware as image one can notice that the malware variants that are the member of the same family show structural and visual similarity [1]. Along with that, it is also noticed that the malware variant from different family shows structural and visual dissimilarities.

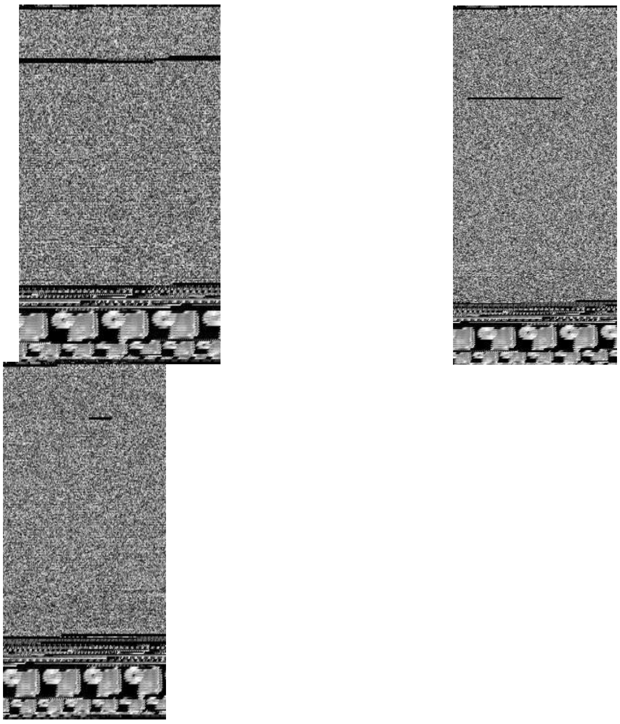

Fig 2. Variants of Fakrean family

2.2.2 Feature Extraction

The features of images are used to classify the malwares into their family. The features can be texture pattern, frequencies in image, intensity, colour feature, etc. These features are gathered by computing mean, standard deviation, Euclidean Distance, etc [6]. For the image. Many algorithms like CLD (Colour Layout Descriptor), HTD (Homogeneous Texture Descriptor), GIST are available to fetch different features of the image and generate the feature vector.

2.2.3 Classification

This feature vector can be applied to any classifier to identify the family of malware. The classifier can be SVM (Support vector machine), K-mean classifier, neural network, etc.

The main reason for using visualization method is, that execution of binary is not required [1]. It is independent of operating system.

3. RELATED WORK

Nataraj, Jacob, et al. [1] describe visualization technique for malware classification which is based on image processing. Greyscale image called malware image is generated from executable binary files. This shows that there exist structural similarities between malware of same family [1]. It uses GIST algorithm for obtaining feature vector from malware image. The feature vector is provided to K-mean classifier for classification. The author obtains accuracy of 0.9718 by using this strategy that works without executing any the malware binary.

Nataraj, Yegneswaran et al. [11] shows comparative study between dynamic analysis of malware and malware image analysis. Their experiments prove that the image based method is more efficient and useful than dynamic analysis.

Nataraj, Manjunath, et al. [6] developed a system called SARVAM. It is content based system for searching retrieving matching images from large databases. The content of a query object is used to find similar objects in a larger database [6]. During the initial phase, first, it generates the fingerprint of a large set of malware image samples using GIST. It obtains the Antivirus (AV) labels that are used to describe nature of malware from Virustotal, and uses Nearest Neighbours (NN) algorithm to extract similar fingerprints. To increase the efficiency of Nearest Neighbour method Balltree structure is used. During the second phase i.e. querying, it compute the fingerprints of the new samples and match it with the existing fingerprints in the database to retrieve the top matches [6].

Hasan [3] describe malware, impacts of malware, various malware and their identification and prevention techniques like signature-based and heuristic method and limitations of this techniques. The author concludes that traditional malware identification techniques and anti-malware techniques are not sufficient. New techniques have to be developed for obfuscated malwares [3].

LIU et al. [12] described an efficient static method to detect and classify malware variants. It works in two stages: first is feature extraction; the other is classification. For extracting features the executable files are translated into controlled disassembly files and then mapped into the grayscale images (0-255 pixel range) by dividing a file into 8-bit blocks. Then, a local mean method is applied to compact the gray-scale images to improve the efficiency. Each pixels of this image represent a feature of that malware [12]. Finally, they uses K-mean and the diversity selection based novel ensemble learning to classify malware. Ensemble learning is a method which assemble multiple weak classifiers to build single strong classifier [12].

Kancherla et al. [13] presented a visualization based method for malware detection. The binary executable file is converted to 8-bit 1-dimensional vector. One vector represents the intensity of one pixel of the image. The image width is kept fixed on basis of the size of the file. Later, it extracts low-level features. They have extracted three different sets of features: Intensity-based (average intensity, no. of pixels with same intensity), Wavelet-based (Horizontal, Diagonal and Vertical coefficients) and Gabor based features (specific frequency content) [13]. Then apply those features to SVM (support vector machine) algorithm for malware detection and classification. This method does not need to unpacking or decryption.

Zainudeen et al. [14] shows a new dynamic analysis technique that work by highlighting the behaviour of malware for malware visualization. This technique represents the malware behaviour in the images (called behaviour image). It starts with monitoring API calls i.e. malware behaviour by executing malware in a VM [14]. Behaviour- to- colour map is generated to represent malicious features of malware. Behaviour – to – colour mapping is done by grouping and sorting APIs based on the level of the maliciousness. Here hot- to- cold colour ramp is used to assign colours (RGB Colour model) to APIs [14]. Hot colours (e.g. red, orange, etc.) represents malicious APIs and the cold colours (e.g. cyan,) represent APIs that are non-malicious. Using this map generate the behaviour image by assigning colours to each captured behaviour i.e. APIs. They noted that variants of a family have recognizable similar pattern even if they have different size and hashes [14].

Zhang et al. [15] developed a technique to classify malware using opcode. They disassemble executable files into opcodes sequences and then converts the opcodes into images. First, they decompile the unpacked binaries to extract their opcodes sequences. Now for each executable, it generate an opcode profile, where each profile contains a list of the opcode sequences with length 2 and their frequencies [15]. Each pixel is a multiplication with the probability of the opcodes sequence and its information gain. To make images easier to recognized and classified it histogram normalization, erosion and dilation are applied. Finally, uses the convolutional neural network (CNN) for identification and classification of malware images.

Mohanaiah et al. [16] present an application of GLCM to get texture based features. GLCM stand for “grey level co-occurrence matrix”. They compute features like ‘Angular second moment’ (ASM), ‘Inverse difference moment’ (IDM), correlation, and Entropy.

4. CONCLUSION

Based on the above study it can be concluded that the malwares are growing continuously so it is required to develop or to improve current techniques to handle malware attacks. The visualization based technique is proved to be the current trend for malware detection as in this technique there is no requirement of executing malware. It eliminates need of emulator and efficient for new malware detection.

5. FUTURE WORK

The current research is made on grey scale image. It is possible to elaborate the work towards the coloured image. The research can be extended towards reducing time and space by compressing the feature vector size and finding a specific highly matching region on the image.

6. REFERENCES

[1] L. Nataraj, S. Karthikeyan, J. Gregoire and B. S. Manjunath, "Malware Images: Visualization and Automatic Classification," in *International Symposium on Visualization for Cyber Security*, pittsburgh, usa, 2011.

[2] N. DuPaul, "Common Malware Types: " ,12 October 2012. [Online]. Available: https://www.veracode.com/blog/2012/10/common-malware-types-cybersecurity-101. [Accessed 20 December 2017].

[3] M. Hasan, "Boyer Moore Algorithm Application in Malware Detection," *International Journal of Scientific Research in Engineering*, vol. I, pp. 69-77, 2017.

[4] Wikipedia, "Rootkit," 12 December 2017. [Online]. Available: https://en.wikipedia.org/wiki/Rootkit. [Accessed 20 December 2017].

[5] The AV-TEST Institute, "www.av-test.org," 10 January 2018. [Online]. Available: https://www.av-test.org/en/statistics/malware/. [Accessed 20 January 2018].

[6] L. Nataraj, B. Manjunathan, D. Kirat and Giova, "SARVAM: Search And RetrieVAl of Malware," in *Annual Computer Security Applications Conference (ACSAC)*, 2013.

[7] F-secure LAb, "Win32/Agent," [Online]. Available: https://www.f-secure.com/v-descs/agent.shtml. [Accessed December 2017].

[8] F- Secure Lab, "Virus," 18 November 2011. [Online]. Available: https://www.f-secure.com/v-descs/. [Accessed December 2017].

[9] "Win32/Rbot," 16 April 2011. [Online]. Available: https://www.microsoft.com/en-us/wdsi/threats/malware-encyclopedia-description?Name=Win32/Rbot. [Accessed 21 December 2017].

[10] Wikipedia, "Malware analysis," 24 September 2017. [Online]. Available: https://en.wikipedia.org/wiki/Malware_analysis. [Accessed 22 December 2017].

[11] L. Nataraj, V. Yegneswaran and P. Porras, "A Comparative Assessment of Malware Classification using Binary Texture Analysis and Dynamic Analysis," in *Workshop on Artificial Intelligence and Security (AISec)*, Chicago, 2011.

[12] L. LIU and B. WANG, "Malware Classification Using Gray-Scale Image and Ensemble Learning," in *International Conference on System and Informatics*, 2016.

[13] K. Kancherla and S. Mukkamala, "Image Visualization based Malware Detection," in *Symposium on Computational Intelligence in Cyber Security (CICS)*, 2013.

[14] S. Zainudeen and M. A. Maarof, "Malware Behavior Image for Malware Variant," in *International Symposium on Biometric and Security Technologies (ISBAST)*, 2014.

[15] J. Zhang, Z. Qin, H. Yin, L. Ou and Y. Hu, "IRMD: Malware variant Detection using opcode," in *International Conference on Parallel and Distributed Systems*, 2016.

[16] Mohanaiah , Sathyanarayana and GuruKumar, *International Journal of Scientific and Research Publications*, vol. 3, no. 5, p. Image Texture Feature Extraction Using GLCM, 2013.