Noble Feature Extraction of Malware from Contents of File

[1]Hemant J. Chaudhari, Prof. M. S. Mahindrakar[2]
Shri Guru Gobind Singhji
Institute of Engineering and Technology
Vishnupuri, Nanded

Abstract—*Malware family identification is a critical process involving extraction of distinctive property from a set of malware samples. Now a day several malware authors use various techniques to prevent the identification of unique property of their programs, such as, encryption and obfuscation. In this paper, we present features extraction of malware from contents of the file. First of all we scanning sample dataset or executable file through the virus total online tool[4] then disassemble given file by using IDA pro tool[1]; Convert given file into N-gram sequential pattern by using KfNgram tool[2]; Measurement of used symbols, sections, metadata and finally calculate the entropy.*

Our goal in this research is to introduce a noble set of features to understood malware features.

Keywords—*Feature Extraction of Malware, N-gram, Sequential pattern, Malware features, Set of attributes, Metadata, Malicious Symbols, Sections, Entropy.*

I. INTRODUCTION

Malicious software, commonly known as Malware, is essentially software that is intended to infiltrate a computer system without the consent of the system's owner. It is an instance of malicious code with intention to harm a computer or network. It covers a range of threats like virus, Trojans, adware's, spywares, etc. They replicate themselves and enter into the system in different ways; either multiple media or through the most popular way of getting downloaded into the system as the genuine application. Since different malware detection system has been introduced till date to circumvent the attacks caused by malwares.

There are two different techniques static and dynamic to analyze malware infected files. Dynamic analysis also known as behavioral analysis, in dynamic analysis detection of malware relies on information that is collected from the operating system at run-time (i.e., during the execution of the program) such as system calls, network access and files.

This approach has several disadvantages.

First, it is difficult to simulate appropriate conditions for malicious functions of a program, such as the vulnerable applications that the malware will be activated.

Second, it is not clear what the required period of time is needed to observe the appearance of the malicious activity of a program. In static analysis, information about a program or its expected behaviors employs explicit and implicit observations in its binary code.

The main advantage of static analysis is its ability to examine a suspected file without actually executing it and thereby provides rapid classification.

Our goal in this research is to introduce a noble set of features to understand the properties and at-tributes of malware files. First of all we scanning sample dataset or executable file through the virus total online tool[4] then disassemble given file by using IDA pro tool[1]; Convert given file into N-gram sequential pattern by using KfNgram tool[2]; Measurement of used symbols, sections, metadata and finally calculate the entropy. Our goal is to introduce a noble set of features to understood malware features and Experimental evaluations on a standard malware data collection are performed to evaluate the proposed technique.

II. RELATED WORK

Several possible techniques have been implemented in the past for malware detection. Chatchai Liang-boonprakong Ohm Sornil, Bangkok, Thailand Classification of Malware Families Based on N-grams Sequential Pattern Features [8], in this paper, they have proposed n-grams sequential pattern features for classifying malware into 10 families. N-grams are created from the binary content of files; n-gram sequential patterns are formed; and patterns are reduced to a minimal set by sequential floating forward selection procedure. Mansour Ahmadi Dmitry Ulyanov, University of Cagliari, Italy Novel Feature Extraction, Selection and Fusion for Effective Malware Family Classification [9].They have presented a malware classification system characterized by a limited complexity both in feature design and in the classification mechanism employed.

To attain this goal, a number of novel features to represent in a compact way some discriminate characteristics between different families. They both have shown general techniques of feature extraction of malware and classify them through the malware families.

In [11], Smita Ranveer and Swapnaja Hiray, Comparative Analysis of Feature Extraction Methods of Malware Detection, This paper gives an overview of malware detection techniques based on static, dynamic and hybrid analysis of executable. They have been presented a comparative assessment of features and illuminated their effect on performance of the system. They have found that, high accuracy and TPR can be achieved by selecting an appropriate feature extraction method. Although op-code and PE features enhanced the speed and accuracy of malware detection system, they give rise to false positives.

III. PROPOSED ARCHITECTURE

As this paper we focuses on noble feature ex-traction of malware from the contents of file , the most relevant issue is related to the choice of the features(Attributes)that will be used to represent each malware sample for selection of noble features. Our approach is totally based on the feature extraction of malware, so we should integrate different types of feature.

A. Features Extraction

Feature extraction is the process of defining a set of features. That is efficiently or meaningfully represents the information for analysis and classification.

There are 3 major modules of our feature extraction method.

1) File Disassembling
2) N-Gram Conversion
3) Detection and measurement Phase

1) File Disassembling

In the stage-1 we are performing the file disassembling task on the given datasets or executable files. Here, we are using IDA pro freeware version for file disassembling. The Interactive Disassembler (IDA) tool is one of the most popular recursive traversal Disassembler, which performs automatic code analysis on binary files using cross-references between code sections, knowledge of parameters of API calls, and other information. We have shown basic information of IDA pro tool.

2) N-Gram Conversion

An N-gram is a contiguous sequence of n items from a given sequence. N-gram is intensively used for characterizing sequences in different areas. An n-gram is an n-character slice of a longer string. First, IDA-Pro, a tool that Disassembler files, is used to extract content of a file into a long string of hexadecimals. The string is then processed into a set of overlapping n-grams. We explore n-grams of several different lengths. The kfngram tool is employed to generate n-gram slices. In the experiments our tests are run with n=1, n=2, n=3 and n=4.

3) Detection and Measurement Phase

In this stage we are detect the malicious code like predefined symbols, registers and operation codes. Here, we are compute the size of files, the number of lines in the file the measure the use of frequencies of the predefined symbols like +,-,*./,,-, =, @,?, ,, [,] etc.

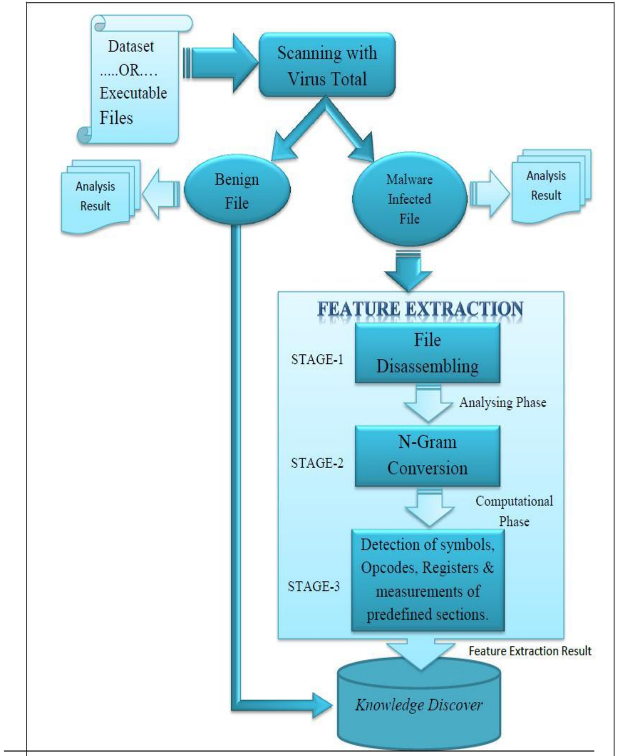


Figure 1. Proposed Architecture for Feature Extraction of malware.

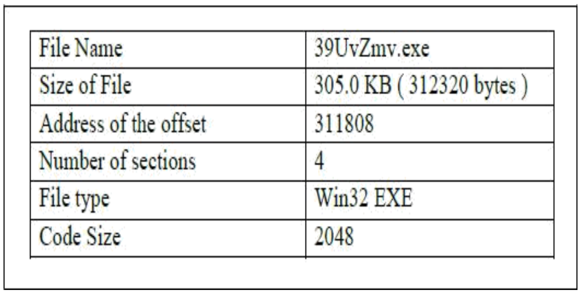
IV. EXTRACTED FEATURES FROM THE DISASSEMBLED FILE

We have extracted the following features of malware from contents of file.

1. Metadata 2. Symbols 3. Section 4. Entropy

1) Metadata

Metadata is data that describes other data. Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier. For example, author, date created and date modified and file size are examples of very basic document metadata. Having the ability to filter through that metadata makes it much easier for someone to locate a specific document. In our research we have extracted the metadata features like size of the file, address of the offset, Number of sections, type of file and the size of code.



File Name	39UvZmv.exe
Size of File	305.0 KB (312320 bytes)
Address of the offset	311808
Number of sections	4
File type	Win32 EXE
Code Size	2048

Table 1. Metadata Report.

Dropper Code: It is a program that when run will install a virus, Trojan horse or worm onto a hard drive, floppy disk or other memory media. The dropper itself is not a virus, it does not replicate; instead, it's more like a Trojan horse in that it carries the malicious code with it and is not detected by virus scanning software because it is not an infected file but carries the code to "drop" a virus into a system. Droppers are uncommon.

2) Symbols:

The frequencies of the following set of symbols (SYM), -, +, *,], [, ?, @, are taken into account as a high frequency of these characters is typical of code that has been designed to evade detection, for example by resorting to indirect calls, or dynamic library loading. Here we are using Kfngram tool to convert suspected file into the n-grams for measurement of used symbols. N-gram is a contiguous sequence of n items from a given sequence. N-gram is intensively used for characterizing sequences in different areas, e.g. computational linguistics, and DNA sequencing.

3) Section:

A PE consists of some predefined sections like .text, .data, .bss, .rdata, .edata, .idata, .rsrc, .tls, and .reloc. Because of evasion techniques like packing, the default sections can be modified, reordered, and new sections can be created. We extract different characteristics from sections (SEC).

4) Entropy

Entropy (ENT) is a measure of the amount of disorder, and can be used to detect the possible presence of obfuscation. Entropy is computed on the byte-level representation of each malware sample and the goal is to measure the disorder of the distribution of bytes in the byte-code as a value between 0 (Order) and 8 (Randomness).

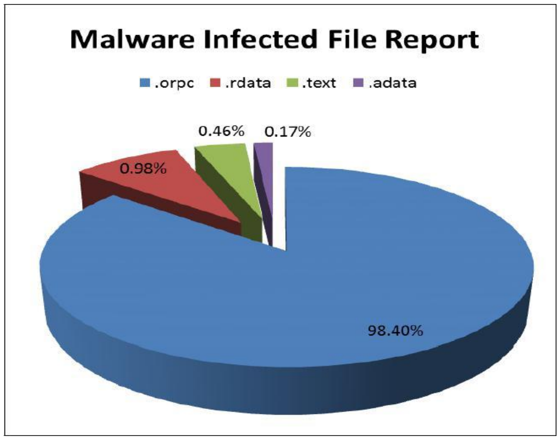


Figure 2. Section wise report of malware file.

V. CONCLUSION

We presented a feature extraction of Malware from the contents of file; we have extracted four features of malware file from their contents. To attain this goal, we used online tools namely as Virus Total, Hex to ASCII Text converter, Count Number of lines in Text and count symbols. The tools used in the project are freely available, and the overall feature extraction process is significantly depends on the analysis report of virus total and IDA pro tools. IDA Pro is freeware commercial software/tool. The used numbers of samples are different for testing and comparison. Malware family identification is a complex process but feature extraction report helps to identify malware type. From the different experiments considering feature extraction, we observed that a number of features are involved into the creation of malware file by the hackers.

We have extract the features namely as Metadata, Symbols, Section and Entropy. Recreating the experiment is generally feasible for the most part. Both Virus Total and PEframe are free to use, and so is the software required to carry out the experiments itself. Another significant challenge for recreating the experiment is the availability of the data set itself, which, at the time of writing, is not publicly available. All the tools are free and open-source, except from Virus Total. Virus Total offers an API, which one can connect to perform analysis for a range of files. The public API is limited, both in terms of scans per minute and scans per day.

REFERENCES

[1] IDA-Pro tool, available at http:// www.hex-rays.com

[2] KfNgram tool available at http://www.kwicfinder.com/kfNgram

[3] HxD Tool, available at https://mhnexus.de/en/downloads.php?product =HxD

[4] Virus-Total online tool available at, https://www.virustotal.com

[5] Hex-to-text Converter online tool, available at http://www.rapidtables.com/convert/number/ hex-to-ascii.html

[6] Count number of lines online tool, available at https://www.tools4noobs.com/onlinetools

[7] VXheavens Website for Datasets http://vx.netlux.org

[8] Chatchai Liangboonprakong Ohm Sornil, Bangkok, Thailand Classification of Malware Families Based on N-grams Sequential Pattern Features IEEE 2013.

[9] Mansour Ahmadi Dmitry Ulyanov, University of Cagliari, Italy Novel Feature Extraction, Selection and Fusion for Effective Malware Family Classification CODASPY 16, March 09-11, 2016, New Orleans, LA, USA.

[10] Smita Ranveer Swapnaja Hiray,Sinhgad College of Engineering, Pune Comparative Analysis of Feature Extraction, Methods of Malware Detection International Journal of Computer Applications (0975 8887) Volume 120 No. 5, June 2015.

[11] ROBERT LYDA, Sparta JAMES HAMROCK, McDonald Bradley Using Entropy Analysis to Find Encrypted and Packed Malware in 1540-7993/07 2007 IEEE SECURITY PRIVACY.