IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. VIII (May-Jun. 2014), PP 43-46
www.iosrjournals.org

Identification of common parameters for classifying malwares with decision tree

Aparna Verma, Dr. M.S.Rao, Dr. A.K.Gupta
Sam Higinbottom Institute of Agriculture, Technology & Sciences, Allahabad, India

Abstract: *Malware analysis is a very tedious and time consuming process. It is the process of determining the behavior and purpose of a given malware sample created by the hacker. Every day new malware sample is being created and released on a large scale by malware authors. Anti-virus companies also requires its database to be updated to protect the victims and organizations. So, this clearly states each one of us to have the tools to analyse the malware. The purpose of this paper is to help Information security professionals, students and peoples, forensic investigator to understand the nature of malwares, and classify them according to the parameters developed by our model with the help of Matlab.*
Keywords: *malware, anti-virus, matlab*

I. INTRODUCTION

The term malware is defined as a malicious software which cause damage to the user system without letting them to know. Day to day anti virus company releases new updated version of their products but still consumers fall victim. Whenever new malware is released in the market its signature is not updated recently in the database,moreover the knowledge about the functionality of malware is very important for their removal. It is a very challenging task for the forensic investigator to perform behavior analysis of malwares which is a time consuming process. So, in this paper we propose a decision tree [1] created by Matlab where various malwares has been analysed both by behavior & automated analysis where a huge number of parameters has been gathered with the help of four tools they are: Regshot, Process Monitor,Comodo,Anubis & buster sandbox analyser. Such parameters has been minimized by our model to save time so that by looking into those parameters investigator can easily classify the malwares.

II. METHODOLOGY

Secure environment

Before start the analysis, establish an environment so that the machine did not connect to remote machine, detach the machine from the internet[2]. Install the operating system most malware and malware analysis tools run on windows, so installed windows XP, as it's still the most popular operating system and the target for most malware. Install the tools for analysis. After loading the tools, record MD5 hashes of all tools used to ensure that the malware does not install a rootkit. Next install Deep Freeze by Faronics is a solution that prevents permanent changes to a computers file system.

Static, dynamic and automated analysis were choosen to monitor the initial execution of the malwares. For this research total 70 samples. Used the following tools to analyze the malwares to gather information about the files it creates, modifies and reads and whether it can bypass the firewall detection or not is as follows:

1) **Virus Total:** Is a free malware**,** URL online scanning service.[3]
2) **Regshot:** It is an open source registry compare tool. It allows comparing to take snapshot of registry and compare it with second one taken after some changes done in the system.[4]
3) **Process Monitor:** monitor, registry, file and network changes
4) **PEiD:** Help to identify packed executables packer.
5) **EXE:** gives information whether malware is packed or unpacked.
6) **Buster Sandbox Analyzer:** Is a tool that has been designed to analyze the behavior of processes and the changes made to the system.
7) **Comodo**: Automated malware analyzer.
8) **Anubis:** An automated tool for analyzing malware.[5]

III. RESULTS & DISCUSSION

Total 70 samples were analyzed using different tools such as Regshot, Process Monitor, Buster Sandbox Analyzer, comodo and Anubis. Both static and dynamic analysis was performed. PEiD and EXE showed what type of compilers were used for compiling it. Some of the malwares were packed while others

were unpacked. Different parameters and values were collected after analyzing the various samples by several tools. Consequently, a Decision Tree [6]is created using the engineering software called MATLAB.

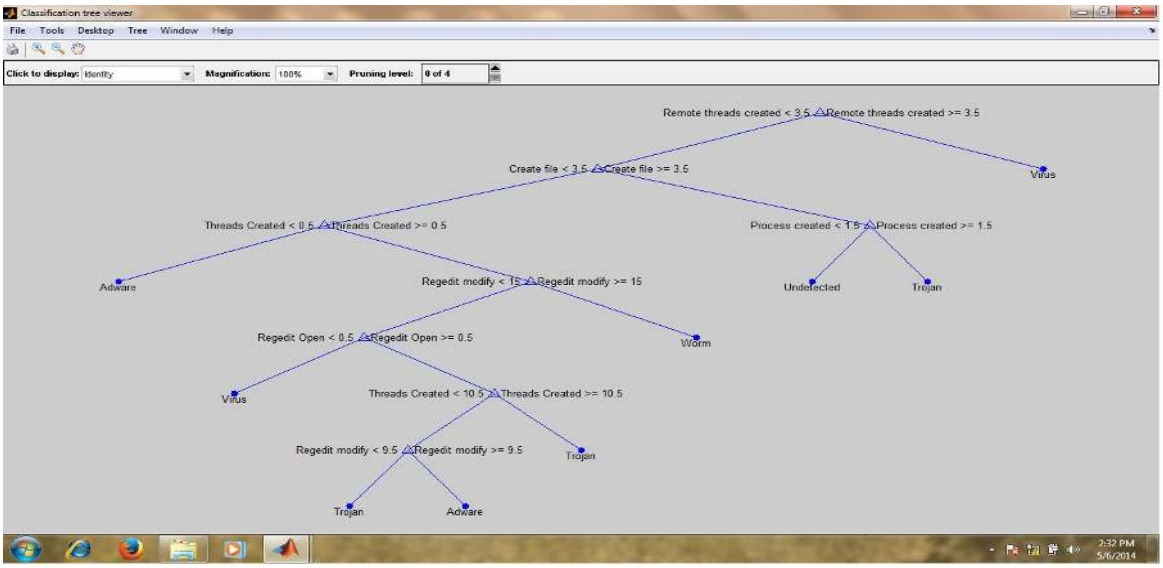

Figure .1. Decision tree of different malwares

After obtaining the different values and parameters of the 70 odd samples run on different analysis tools; a decision tree was created as shown above highlighting the 06 significants obtained, based on ID3 algorithm(MATLAB software) certain values assigned by different parameters.

Now, we take 50 new samples. These 50 new samples are run on the following analyzing tools (Regshot, Process Monitor, Buster Sandbox Analyzer,.PEid, EXE, Anubis, and Comodo). We obtain different values and parameters as shown below in the table.

The next step. We identify the values and parameters obtained by the concluded Significants as in the above shown Decision Tree.

Thereafter, we select the same values and parameters as identified by the Significants in the decision tree for the 50 new samples.

Henceforth, we get a set of values and parameters from the above Decision Tree and a new set of values and parameters from the 50 new samples run.

Ultimately, we compare these sets obtained comprising of values and parameters of earlier samples and 50 new sample.

	Remote threads created	**Create file**	**Process created**	**Threads created**	**Regedit modify**	**Regedit open**	**Threads created**	**regedit modify**	**Result by decision tree**
Virus	5	1	1	13	2	1	13	2	Virus
Virus	5	1	3	14	2	1	14	2	Virus
Virus	0	0	0	12	2	1	12	2	Trojan
Virus	5	4	4	12	2	1	12	2	Virus
Virus	5	2	0	13	2	1	13	2	Virus
Trojan	0	0	0	1	1	1	1	1	Trojan
Trojan	0	0	0	0	0	1	0	0	Adware
Virus	5	2	0	13	2	1	13	2	Virus
Trojan	2	0	0	10	2	1	10	2	Trojan
Trojan	0	0	0	2	1	1	2	1	Trojan
Virus	5	5	1	11	2	1	11	2	Virus
Virus	0	0	0	11	2	1	11	2	Trojan
Trojan	0	0	0	11	2	1	11	2	Trojan
Trojan	0	0	0	6	2	1	6	2	Trojan
Trojan	0	2	0	1	2	1	1	2	Trojan
Trojan	0	0	0	10	2	1	10	2	Trojan
Virus	5	24	0	10	2	1	10	2	Virus
Virus	1	5	2	13	0	1	13	0	Trojan
Trojan	1	13	2	16	0	1	16	0	Trojan
Adware	1	3	1	0	0	1	0	0	Adware
Trojan	1	1	1	15	1	1	15	1	Trojan

Trojan	0	0	0	1	0	1	1	0	Trojan
Virus	5	3	0	12	2	1	12	2	Virus
Virus	5	2	0	9	2	1	9	2	Virus
Undetected	1	52	3	4	0	1	4	0	Trojan
Virus	5	3	2	11	2	1	11	2	Virus
Trojan	1	3	1	10	0	1	10	0	Trojan
Adware	1	12	3	2	0	1	2	0	Trojan
Virus	5	3	0	12	2	1	12	2	Virus
Undetected	1	4	1	16	0	1	16	0	Undetected
Undetected	1	4	1	15	12	1	15	12	Undetected
Adware	0	1	0	1	12	1	1	12	Adware
Virus	5	0	2	12	2	1	12	2	Virus
Virus	5	5	0	11	2	1	11	2	Virus
Trojan	0	0	0	4	1	1	4	1	Trojan
Virus	5	5	0	11	2	1	11	2	Virus
Trojan	1	3	1	10	0	1	10	0	Trojan
Trojan	1	3	1	9	0	1	9	0	Trojan
Trojan	0	2	0	1	0	1	1	0	Trojan
Virus	5	2	1	12	2	1	12	2	Virus
Virus	5	24	0	12	2	1	12	2	Virus
Virus	5	2	0	13	2	1	13	2	Virus
Virus	5	5	1	14	2	1	14	2	Virus
Virus	5	0	0	12	2	1	12	2	Virus
Trojan	1	1	3	13	14	1	13	14	Trojan
Virus	5	8	1	15	2	1	15	2	Virus
Virus	1	14	2	4	6	1	4	6	Trojan
Virus	5	4	0	13	2	1	13	2	Virus
Virus	5	1	0	1	2	1	1	2	Virus
Virus	5	2	1	14	2	1	14	2	Virus

Table 1: Comparison of parameters generated by different tools parameters identified by significant as per Decision tree

In the above table, observe that the first sample which is Virus when run on a tool is concluded as Virus when compared to the result obtained by the Decision tree, whereas third sample which is Virus when run on a tool is concluded as Trojan when compared to the result obtained by the Decision tree.
Similarly, the remaining samples when run on a tool, is characterized as various types of malware as shown in the last column of the above table obtained when compared to the result by the decision tree.

✖ On the event of running various tools for analysing a bunch of malwares, the tools generated different and common attributes of given malwares. So in order to run so many tools, which will generate many attributes it is better to develop a model by decision tree to minimize the attributes during investigation.

✖ Therefore a decision tree is created by running the derived attributes on the Matlab software. Consequently, the Matlab software identified the common attributes inheritently existing in the various samples of malwares.

Therefore, by performing the aforesaid procedure, a forensic investigator is able in identifying the common parameters of different types of malwares. The procedure further enables to conclude on a decision based on probability of the common parameters. Thus, a forensic investigator is able in differentiating among the various types of malwares, eg. virus, worm,trojan etc

IV. CONCLUSION

Each individual and an organization is vulnerable from the threat of malwares. Malwares have become an effective instrument to damage, destroy and incur mammoth losses not only restricted to individuals but also to highly e-secured environment of organizations. The exploitation of computer programs is being visualized as the next threat to information storing and sharing. A comprehensive research in detection, analyzing, identification, repairing, removing of malwares is required to explore this undiscovered field. Therefore, cyber crimes needs to be thoroughly and meticulously conducted similar to a murder investigation . The decision tree highlights the parameters to look out for the analysis whenever we are subjected to a cyber attack. Every antivirus is not 100% safe, malware authors are very smart, they make use of crypters & binders [7]to bypass even Antivirus. So, instead of using so many tools, forensic investigator can make use of our model to classify the malwares. Once the investigator has searched those six parameters about the malware during investigation, it becomes easy for them to classify by our model. The identification of the binary structure of a malware helps in knowing its features, characteristics , behavior and composition. Collection of such information yields in

developing its countermeasures depending upon its type (worm, rootkit). The above model is a very simple and helpful tool even to the least computer literate to understand and differentiate among the various types of malware.

ACKNOWLEDGEMENT

I am very thankful to Dr. Vrijendra Singh & Dr. M.S.Rao without their support research would have not been possible who guided me in right track and always encouraged with their ideas.

REFERENCES

[1]. Mitchell T M, Machine learning,Singapore,Mc Graw Hill,1997
[2]. Distler,Malware analysis:An introduction,SANS institute,2007
[3]. Michael Hael Ligh, Steven Adair, Blake Harstein & Matthew Richard, Malware analyst's cookbook & DVD:Tools and techniques for fighting malicious code(Wiley Publishing Inc.,Indianapolis,Indiana,2011) 90-92
[4]. Michael Hael Ligh, Steven Adair, Blake Harstein & Matthew Richard, Malware analyst's cookbook & DVD:Tools and techniques for fighting malicious code(Wiley Publishing Inc.,Indianapolis,Indiana,2011) 288-290
[5]. Egele M,Scholte T,Kirda E,Kruegel C,A survey on automated dynamic malware analysis techniques and tools,ACM computing,Vol 44(2),Article6.2012,42
[6]. Alsabti K, Ranka S, Singh V, Clouds:A decision tree classifier for large datasets,Conference on knowledge discovery & data mining,(KDD-98)1998
[7]. Murray Brand,Analysis Avoidance Techniques of malicious software,Edith Cowan university,30th November 2010