IMPROVİNG THE MALWARE DETECTİON RATİO USİNG DATA MİNİNG TECHNİQUES

Abhay Pratap Singh

Department of CSE,, Manav Rachana İnternational University, Faridabad (India)

ABSTRACT

In today's word computer security is become essential part of internet security. Attacker may execute malware or buffer overflow attacks to gain access system malware is also big concern in computer security, now a days malware analysis is big task for the cyber security expert because over the last two to three years attacker become so smart they are used various kinds of new techniques like armoring, tunneling etc. With the help of these techniques they can easily bypass malware in good anti-virus software. in this paper, we are presenting a data mining approach which improves the malware detection ratio with high precision, whereas current anti-virus detection technologies which is based on signature, code emulation, anomaly based, these are the technologies are failed to detect malware, although some of the technologies are good, but in terms of money they are expensive and one important thing sometimes they also give the false positive results, which is our primary concern in this paper.

Keywords: Computer Security;Advance Escaping Techniques;Malware Detection;Data Mining.

I. INTRODUCTİON

The contineous growth of malwares become serious threat for computer security, it can affect the Integrity, control flow, and the functionality of a system. Therefore, their detection is a major concern within the cyber security as well as user community. As malicious code can affect the data and control flow of a program, static flow analysis may naturally be helpful as part of the detection process. Still various approaches and antivirus scanners are used to detect the malicious code, but all these techniques are signature, anomaly, and code emulation based approach, still these are approaches ineffective to cope with new malware, although some method are reliable but they have also vulnerabilities. To avoid detection by the traditional based approaches, a number of stealth techniques and obfuscation techniques are available. Malware writers may use various obfuscation techniques to hide themselves from the various anti-viruses. Simple obfuscation deals with inserting NOP (no operation) instructions, swapping registers, and reordering independent instructions. Malware can be obfuscated using two techniques polymorphic techniques and metamorphic techniques. Obfuscation attempts to hide the true intentions of malicious code without extending the behaviors exhibited by the malware. Behavior addition/modification effectively creates new malware, although the essence of the malware may not have change.

II. RELATED WORK

In this section we discuss about various techniques used for malicious code detection and role of data mining in cyber security.

In [1] author used data mining techniques for extracting variable length instruction sequences that can identify Trojans from clean programs. The analysis is facilitated by the program control flow information contained in the instruction sequences.

In [2] author presents cw sandbox, which executes malware samples in a simulated environment, monitors all system calls, and automatically generates a detailed report to simplify and automate the malware analyst's task. It monitors all the executed functionality.

In [3] the author has focused on deobfuscation of actual obfuscated code in order to reveal true intent of that piece of code.

In [4] author proposed a statistic-based metamorphic virus detection technique and proves that detection based on statistics is a useful approach in detecting self-mutated malwares. Six statistic feature that include percentage of NOP instruction at the end of subroutines, percentage of NOP instruction in random.

In [5], a behavior based detection approach is proposed to address malware detection. The behaviors of interest are defined as static system call sequences and they are derived by statically analyzing binary code.

In [6], the author talks about role of data mining in cyber security and it is being applied to problem such as intrusion detection and auditing and it also addressed surveillance problem.

III. ADVANCED TECHNIQUES FOR ESCAPING FROM ANTIVIRUSES

Since current antivirus scanner are fail to cope with new malware just because of malware writers apply some obfuscation techniques, some important techniques are listed below:

A. Code Obfuscation

Obfuscation techniques are those that are used by malware writers to avoid detection and analysis becomes so complex ex.-inserting some bunch of junk instruction (NOP).

B. Oligomorphism

These viruses use multiple decryption routines to avoid giving a signature for the antivirus software. The decryption routine is chosen randomly on infection. But, if the antivirus software have signatures for all of the decryption routine, detection is possible.

C. Polymorphism

These viruses change the look of the virus code every time it infects a new file. This is achieved by changing the decryption routine and is much harder to detect using signatures.

D. Metamorphism

These viruses change the virus body instead of appearance. This is possible by using equivalent and unneeded functions or by changing the sequence of statements in the code slightly (as long as the logic remains relevant).This way every specimen looks different and generation of a signature is harder.

E. Armouring

An armoring virus is virus that makes analysis very difficult. These kinds of viruses use various anti-debugging, anti-heuristics, anti-goat, anti-VM (virtual machine detection) techniques.

F. Tunneling

Tunneling is mainly used to hide behavior blocking antivirus software. These capture operating system interrupts. So, whenever these interrupts are made, the virus executes first and after that control is passed to the original destination. This way they are at a much deeper level in the operating system than the antivirus software and may avoid detection by it.

G. Stealth

A stealth virus is a type of virus that tries to remain undiscovered by hiding the infection events from everyone, instead of trying to obfuscate its code. It achieves this by restoring certain original properties of the host file. Example: Timestamps

IV PROPOSED METHODLOGY

Our research has been performed by some basic steps:

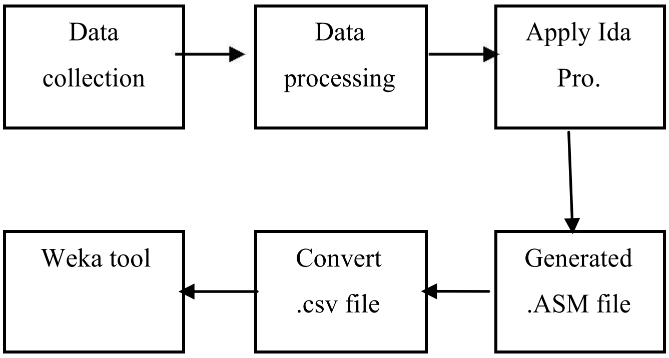
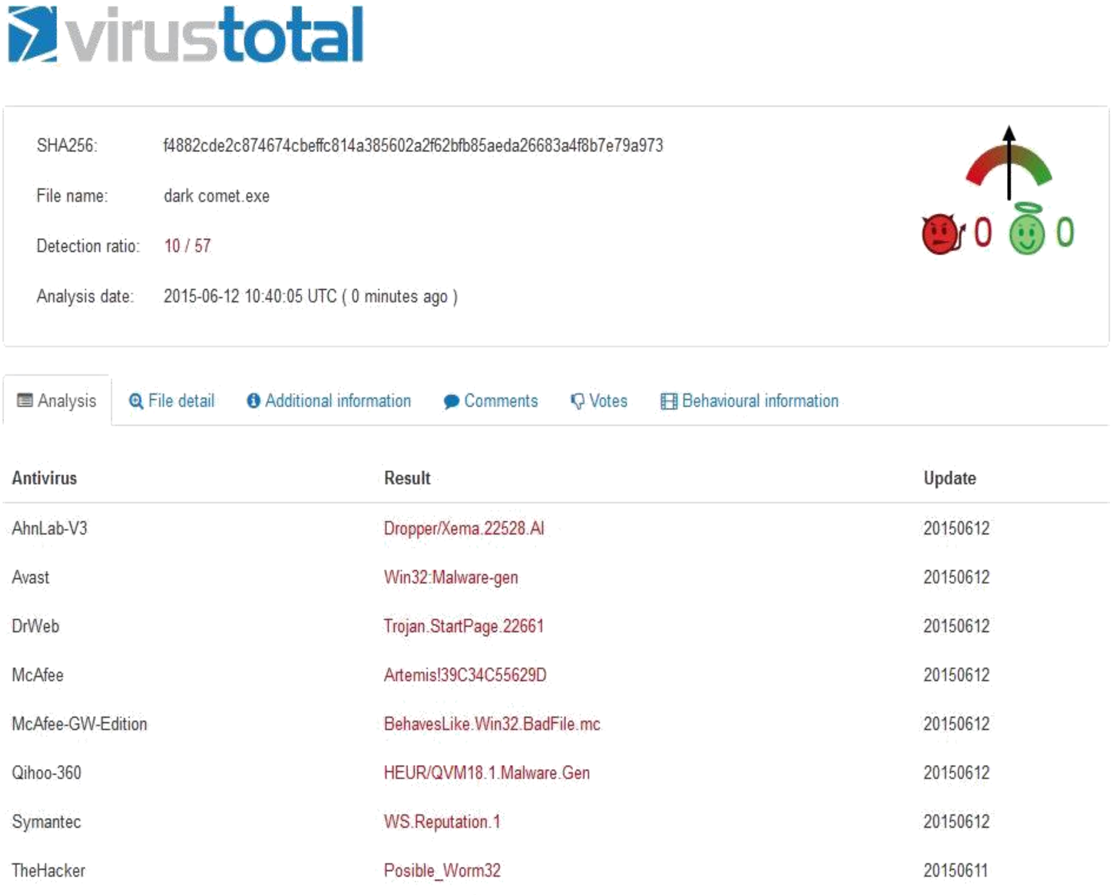


Fig. 1 Basic Steps of Proposed Methodology

- **Data collection:** First we have downloaded a malicious dark comet Trojan virus from offensive computing, after that we configured a virus, we selected this malware because in this study our main aim to detection of Trojan virus.
- **Data processing and preparation:** In this section we deal with data processing using reverse engineering tool like Ida pro as well as PEiD anti-packing tool. First we process the PEiD tool to check whether the file is compressed or packed, encrypted if file is compressed it will not open in Ida pro, so first we have to check file is encrypted or not, afterwards, we give the file as input to disassembler and they get the assembly code of these fields and return the called system functions list from these assembly codes. We just insert some garbage instruction like NOP instruction it will not alter the functionality of code but will increase the size of code.
- **Analysis of results:**In this section we analyze the malware code, IDA pro generated ASM file (automatic storage management), we can easily see that where we modified the code. Our aim to analyze and detect malware by examining the shared pattern using machine learning techniques, after that we convert this .asm file into .csv file, later on we applied this file on weka data mining tool. The advantage of this method include its high success rate in malware detection because it is directly in contact with malware binary codes and also there is no need to run them and we can understand whether it is malware or not using their code and obtaining the shared sequence of called system functions.

V. EXPERİMENT RESULT

Our proposed methodology can detect the malware along with obfuscation when compared with existing antivirus scanner. The interesting fact our study revealed that when we tested Trojan file in total virus site there is only 10 antivirus scanner detect and all scanner are failed to detect them, but when we applied weka data mining tool and performed j48 algorithm on .csv file then we will get 80% detection ratio which is remarkable rate, fig. shows a graph of data mining operation it clearly indicates that data mining method is far better than traditional based scanners.



SHA256: f4882cde2c874674cbeffc814a385602a2f62bfb85aeda26683a4f8b7e79a973

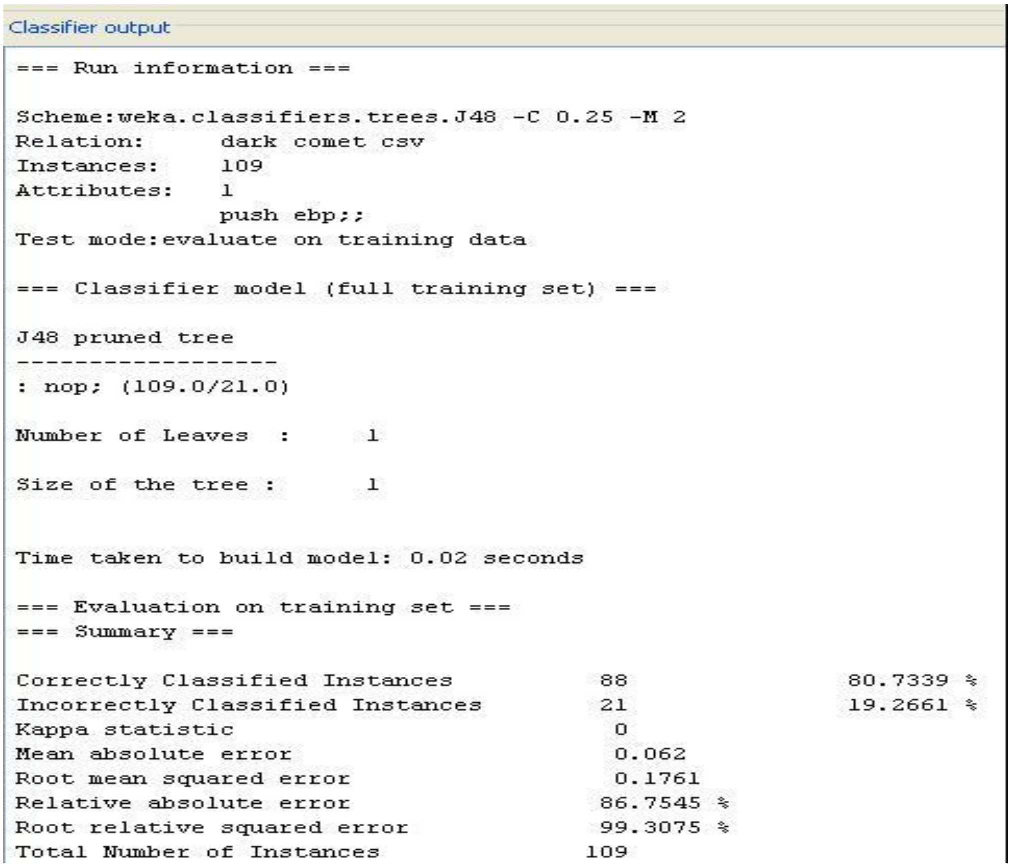
File name: dark comet.exe

Detection ratio: 10 / 57

Analysis date: 2015-06-12 10:40:05 UTC (0 minutes ago)

Antivirus	Result	Update
AhnLab-V3	Dropper/Xema.22528.AI	20150612
Avast	Win32:Malware-gen	20150612
DrWeb	Trojan.StartPage.22661	20150612
McAfee	Artemis!39C34C55629D	20150612
McAfee-GW-Edition	BehavesLike.Win32.BadFile.mc	20150612
Qihoo-360	HEUR/QVM18.1.Malware.Gen	20150612
Symantec	WS.Reputation.1	20150612
TheHacker	Posible_Worm32	20150611

Fig. 2.Results produced when the System is tested on Virus Total website against various antiviruses



```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     dark comet csv
Instances:    109
Attributes:   1
              push ebp;;
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------
: nop; (109.0/21.0)

Number of Leaves  :     1

Size of the tree :      1


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          88               80.7339 %
Incorrectly Classified Instances        21               19.2661 %
Kappa statistic                          0
Mean absolute error                      0.062
Root mean squared error                  0.1761
Relative absolute error                 86.7545 %
Root relative squared error             99.3075 %
Total Number of Instances              109
```

Fig.3. Results Produced by Weka

VI. CONCLUSIONS

In this paper we provided an introduction to the malware research using data mining techniques, malwares are becoming widespread and more complex every day. As example of their complexity, we can note the need of using polymorphism techniques; transformation and encryption, the traditional method such as matching some code strings of malwares signatures don't enough efficiency. Data mining framework which easily detect undetectable computer virus rather than compare to traditional based antivirus.

Table I.Comparison of Proposed and Existing Work

Malware	Traditional based detection	Proposed method
Obfuscated virus files	Cannot detect	Can detect
Malicious files without obfuscation	Can detect	Can detect.

REFERENCES

[1] D.M.A. Hussain et al. (Eds.): "Detecting Trojans Using Data Mining Techniques", CCIS 20, pp. 400–411, 2008.Springer-Verlag Berlin Heidelberg 2008.

[2] Carsten Willems, Thorsten Holz, Felix Freiling: "Toward Automated Dynamic Malware Analysis Using CWSandbox", IEEE Security and Privacy, vol. 5, no. 2, pp. 32-39, Mar/Apr, 2007.

[3] A. Lakhotia, E. U. Kumar, M. Vennable, "A Method for Detecting Obfuscated Calls in Malicious Binaries", IEEE transactions on Software Engineering, Vol 31, No 11, November (2006).

[4] Govindaraju. A, Faculty, Department of Computer Science, Master Thesis, "Exhaustive Statistical Analysis for Detection of Metamorphic Malware". San Jose State University, San Jose, CA (2010).

[5] Ding Yuxin*, Yuan Xuebing, Zhou Di, Dong Li, An Zhancha," Feature representation and selection in malicious code detection methods based on static system calls"Computers & Security (2011) ,article in press, science direct journal.

[6] Sudha nagesh research scholar, HOD- computer science, Navkies junior residential college, Nelamangala "Role of data mining in cyber security" published in journal of exclusive Management science – May 2013-vol 2 issue 5 –ISSN 2277 -5684.