ANALYSING MALICIOUS PE FILES WITH INSTRUCTION N-GRAMS USING MACHINE LEARNING TECHNIQUES

SANDEEP R MODI,

PG STUDENT, DEPT. OF COMPUTER SCIENCE & ENGINEERING, MEWAR UNIVERSITY, GANGRAR, CHITTORGARH, RAJASTHAN

sandeepmodi50@yahoo.com

ABSTRACT : *Malicious software also known as malware/malcode fulflls evil intention(s) of its author(s). Now a days, modern malware is facilitating crimeware and ransomware syndicate by exhibiting dynamic, stealthy behaviour with self mutation. They avail administrative rights to control the victim computer. Malware writers depend on evasion techniques like code obfuscation, code packing, encryption, polymorphism to avoid detection by Anti-Virus (AV) scanners as AV use unique byte pattern known as signature to detect malware. AV scanners need frequent update of known signatures to ag malicious code.*

INTRODUCTION:- Advent of Internet has increased the appearance of malware in the digital world. Use of Internet based applications like paying phone and electric bills, money transfer among accounts and financial transactions is growing rapidly. Increased use of Internet by native users increase the risk of password break due to simplicity, stealing transaction information from cookies stored on personal computers, or exploiting the vulnerabilities in operating system and software programs.

Malicious software refers to all software performing nefarious activities. Malware is known by different names like Virus, Worm, Trojan, Backdoor, Rootkit, Software Robot to name a few. Malicious software replicate own self and exploit vulnerabilities of the existing system. Microsoft Windows is one of the most prevalent propriety operating system used across the globe and Portable Executable (PE32) file format is the target of the malware writers.

OBJECTIVE:- Malware possess less functional diversity compared to normal programs due to their prime motive of malicious intent. Hence detection schemes can be successful against malware attacks. Our focus of research is Windows Portable executable (PE) as it is the most targeted file system . Our focus of research is Windows Portable executable (PE) as it is the most targeted file system .

THE OBJECTIVE OF THIS THESIS ARE:

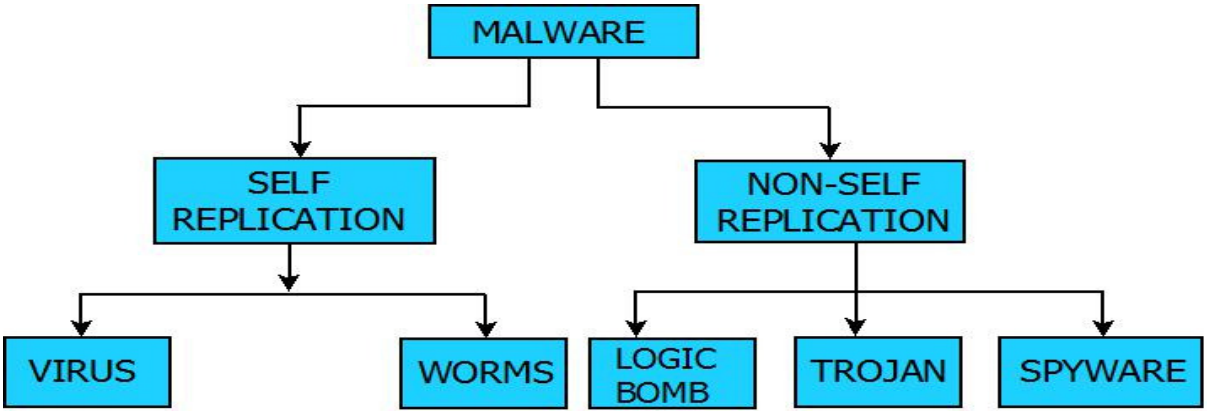
- Identify benign and malicious code, or variation of same program, using non--signatured Instruction n-gram method
- Devise non-signatured method to identify malicious files from benign (non-viral) executables using machine learning technique.
- Propose identification method to detect malicious PE executables based on Structural opcode features of PE.
- Discrimination of Malware and Native executables.
- Use pattern recognition techniques in form of classification algorithms to identify zero day / unknown malware.

Work according to Machine learning Techniqes. Those major points are shown below.

LITERATURE SURVEY:-

- In this , Malware and its categories is discussed. Diffierent ways of infection and propagation used by malicious software is explained. Detection methods and their approach is discussed later in the chapter.
- Covers related work in areas of structural analysis using diffierent parameters of PE files. In this dissertation, structural opcode instructions features are used to diterentiate native and malicious executables. Our approach identifies Packed files with good accuracy.

So finally covered all types of Malware and how to detect their malware so putting detection methods as well as their related w



TYPES OF MALWARE:-

- Virus
- Worms
- Trojans
- Logic Bombs
- Polymophic Malware etc….

DETECTION METHODS:-

- Static
- Dynamic
- Hybrid

RELATED WORKS:-

- Signature based detection
- Non-Signature based detection
  - Important/Prominent mnemonic *n*-grams are extracted using Class-wise document frequency (CDF) [1].
  - CDF for an *n*-gram is given by:

where:

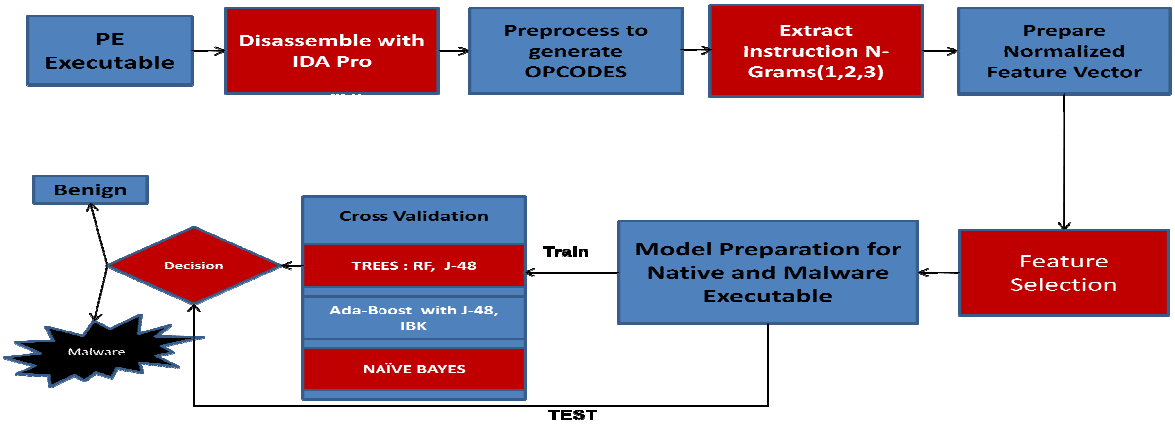
$$F(\eta) = \sum_{i} \sum_{C\varepsilon\{M,B\}} p_i(v,C) \log \frac{p_i(v,C)}{p_i(v)\,p_i(C)}$$

C*:** Class which is either Malware (M***) of Benign (***B***)

v*:** ***v (1 or 0), indicates the presence or absence of a ***n***-gram in a class (***C***).

-

METHODOLOGY OVERVIEW



THESIS ORGANIZATION:

- Malware and its categories is discussed. Different ways of infection and propagation used by malicious software is explained. Detection methods and their approach is discussed later in the chapter.
- covers related work in areas of structural analysis using different parameters of PE files.

CONCLUSION:-Malware signatures are increasing exponentially every year. Malware analysis and detection techniques need to keep pace with the ever increasing threat of new techniques employed by malware writers to thwart detection. Traditional exact string matching approach is insufficent to fight the threat of obfuscation of existing samples. Non signature based detection methods are gaining ground against packed and metamorphic malware. We have addressed issues in code packing, dealt with a non signatured approach of instruction n - - grams to detect malicious code among a pool of benign programs.

FUTURE SCOPE:- In future we would like to implement hybrid approach to deploy an effective malicious file analysis and detection mechanism.

REFERENCE /BIBLIOGRAPHY :-

[1] L. Breiman. Random Forests. Machine Learning, 2001.
[2] Windows XP Service Pack{3.:" http://windows.microsoft.com/en-IN/ windows/products/windows-xp.
[3] Microsoft Portable Executable and Common Object File Format Speci_caton, 1999.
[4] R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[5] P.-N. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining. Pearson.
[6] WEKA.:"Data Mining with Open Source Machine Learning Software. http:// www.cs.waikato.ac.nz/ml/weka, Last Accessed March 2012.
[7] Virus Collections.:" VXHeavens). http://vl.netlux.org/vl.php/, Last Accessed February 2012.
[8] TPR TNR.:"http://en.wikipedia.org/wiki/TypeI~and~TypeIIerrors, Last Accessed February 2012.