International Journal of Engineering, Science and Mathematics
Vol. 7 Issue 1, January 2018,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

DETECTING MALWARE BY DATA MINING

Shafiqul Abidin
HMR Institute of Technology & Management (GGSIP University)
Hamidpur, Delhi, India

Rajeev Kumar
Department of Computer Science
Kalka Institute for Research and Advanced Studies (GGSIP University)
Alaknanda, New Delhi, India

Varun Tiwari
Comm- IT Career Academy (GGSIP University)
New Delhi, India

ABSTRACT

The exponentially growth of malware has created number of security threats in IT industry. A large number of viruses are developed and millions of applications are infected and suffered on daily basis. Trojan is one of the fatal and deadly types of malware. But it is often said as legitimate software. They hide themselves within harmless programs. Trojan survived by going unnoticed. They look like just about anything like the computer game as downloaded from different websites. Sometimes even a popup advertisement might try to install something on our computer. Trojan can trick you into using them. In this paper, data mining technique is being proposed to detect Trojan. The technique is based on Naive Bayes – this technique is simple to put into practice and we achieve amazing results in large number of cases. But practically, dependencies exist among variables.

KEYWORDS: Trojan Detection; Data Mining; Decision Tree; Naive Bayesian Network; Naïve Classification Technique**.**

1. INTRODUCTION

Malware can be said as the collective term for virus, Trojan horse and other malicious that can infect the system. Since, many years these harmful items have evolved and affected smart phones and tablets as well. Malware is sometimes known as computer contaminant, as in the legal codes of several U.S states [1]. Malware comprised of damaging function that is called Payload which has various effects. It bears the quality to be unnoticed. This unnoticed nature is achieved by actively hiding and showing no presence to users. The generic term malware comes from "malicious software", where malicious describes any code in any part of a software system that is intended to cause damage to a system. The types of malicious codes are Virus, Worms, Trojan and other malicious.

But what is the difference between a virus and a worm? What is the difference between these two and Trojan? Does antivirus apply against Worms, Trojan, Virus and other malicious codes? All these questions come from one source and it's the complex and complicated world of destructive codes [2].

Several types of malicious codes have some kind of behavior which are described below-

Virus

A code which get attached itself to a host program and propagates whenever that infected program executes.

Worms

Unlike virus, a worm does not attach itself to an infected executable program but it spreads itself by transferring via network which includes some connected computers.

Trojan Horse

This includes a hidden program component, which are in a form of pieces of software code which opens a backdoor into the affected computer and thereby allow almost full access to the user noticing.

Trojan horse often referred as Trojans. In 1986, the first Trojan was ‘PC-Write’. Trojan is derived from the Ancient Greek story of the wooden horse that used to protect the city of Troy. Trojans are totally different from virus and worms they do not introduce themselves or disseminate themselves into other files, it just represents itself as useful or interesting that tempt a user to install it. Trojans are classified according to the type of actions they can perform on a computer.

Backdoor

This gives malicious users control to do anything they wish on the infected computer, which includes sending, receiving, deleting, launching files, display and rebooting also.

Exploit

This is a piece of data or a sequence of commands that take advantage and attacks within application software running on the system.

Rootkit

This is designed in order to prevent malicious programs being detected. It is difficult to detect because it is activated each time system boots up.

Trojan-Downloadert

This can easily download and install different types of new malicious program into a system.

There are also other types of Trojan too like Trojan Banker, Trojan DDOS and many.Trojan can do a lot of harm to a system like - destruction to the system; corrupt data or delete; modify data; spy; use computer resource; infect other connected device etc. In short short it can do a lot of harm to the system.

As signature method is a traditional and usual method to detect malicious program. They are created manually; it matches with at least one byte code pattern of the software. As, researchers have tried to present more reliable methods for malware detection.

2. RELATED WORK

The process of identifying malware is called analyzing, which are roughly divided into static and dynamic analysis.

STATIC ANALYSIS

In this program code is checked. But in actual program code does not execute. It investigates and detects coding flaws, back doors and potentially unwanted codes. In static method, binary codes are checked and detected according to the binary codes given.

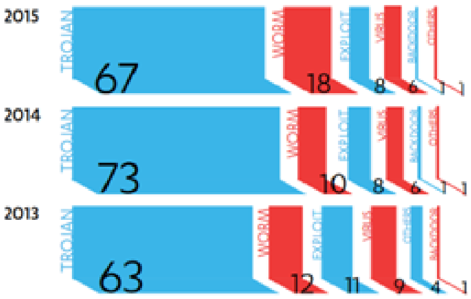

Figure 1. Percentage of Malware Detection Reported

DYNAMIC ANALYSIS

Dynamic Analysis is evaluated in a runtime environment. Its key objective is to find bugs in a program, during run time instead of repeated code examining. Actually they analyze what is being happened behind the scene. Sometimes static and dynamic analysis is considered as glass-box testing. In this article an effort has been made to get ascertain static analysis method by implementing a data mining technique to detect & clean Trojan.

3. LITERATURE SURVEY

Earlier malware was detected using signature methods, and then researchers found a number of classifiers for analysis and detection of malware. Many classifiers used n-gram i.e. a series of bits in some order and extracted from hex dump like mentioned in an International Journal of Intelligent Information Systems paper, presented on “Malware detection using data mining techniques” that has a higher success rate, as it finds whether there is a malware or not using the binary codes and as per rootkit detection success rate is over 97% [3].

In another paper decision tree and Naïve Bayes data mining techniques are used to detect virus. That consists of more than three thousand malicious and more than one thousand benign programs were there, where firstly op-code is used as vector and secondly, op-code as well as first operand were taken where benign and virus programs were mixed which affects the effectiveness of both the classifiers [4]. Proposed surveillance spyware detection system (SSDS), where features were considered as both static and dynamic using information gain method 76 static and 14 dynamic were discovered. According, to their research SSDS is a better performer than other known antivirus like Norton, Kaspersky, etc. A signature based method called as SAVE (Static Analysis of Vicious Executable), represented as API calls and as well as used Euclidean distance to compare signature with API calls [5].

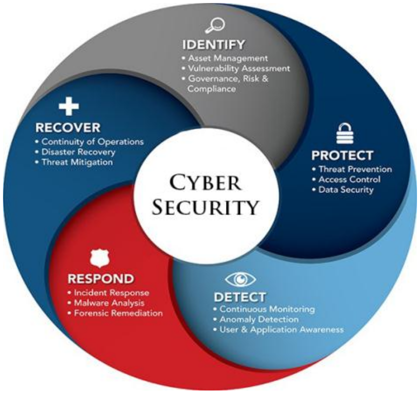

Figure 2. Cyber Security Model

So, besides data mining method other techniques are also used to detect malware as malware detection is a very important part in security. Data preparation is needed for data mining process, where data needs to be collected then its feature needs to be extracted and model should be classified to get the result [6] [11].

4. NAÏVE BAYES

This Naïve Bayes is based on probability and works on independent assumptions. This method is used both for multi class classification and byte sequence data Naive Bayes has been studied. Since, 1950s and is popular for data classification, where document is assigned to one or more categories (can be text, spam, image or music, etc.) [7].

Applications of Naïve Bayes Classification include:

- Text Classification
- Spam Filtering
- Hybrid Recommender System
- Online Applications – Simple Emotion Modeling.

In 2003, Virus detection using data mining method is was published i.e. Multi Naive Bayes. They are quite well in complex problems. Naive Bayes are simple, fast and highly scalable that only requires small amount of data to estimate. It is based on conditional probabilities i.e. calculates a probability by counting the frequency of values and combination of a given data [8].
In this method, we want to compute a certain given text document, as it states-

$P(x/y) = P(y/x). P(x)/P(y)$ (i)
Where,
$P(x/y)$ = posterior probability
$P(x)$ = prior probability

Priori: probability of an event before the evidence is observed.
Posterior: probability of an event after the evidence is observed

Here, x is a vector x=(x1, x2… xn).To use Naive Bayes technique, we assume features which occurs independently. Suppose feature is F, then

$F=(F1,F2\ldots Fn).P(x, F1, F2\ldots Fn)=P(x).P(y1\ldots\ldots., yn/x)$
$=P(x).P(y1/x).P(y2/x\ldots\ldots yn/x,y1)$
$=P(x).P(y1/x).P(y2/x).P(y3\ldots\ldots.yn/x,y1,y)$
$=P(x).P(y1/x).P(y2/x,x1)\ldots\ldots.P(yn/x,y1,y2,y3\ldots\ldots yn-1)$
As, $i \neq j$
$P(x/F) = \Pi ni=1 \; P (Fi/x)*P(x)/ \Pi nj=1 \; P (Fj)$ (ii)

Since, denominator is same for all the classes. So, we take maximum as computed in (ii) equation, we get
$P(x/y1\ldots\ldots\ldots yn)=max(P(x)\Pi ni=1P(yi/x)$

We first, collected data then feature extractionand then we applied the equation for the program.

5. PROPOSED APPROACH

Analysis of program can be carried out in each step of our method like data collection, data preprocessing, feature extraction, and feature selection. At last decision tree and naive Bayesian network algorithms have been suggested and practical are carried to find the effectiveness of proposed technique.

6. DATA COLLECTION AND PROCESSING

We have downloaded collection of Trojan codes at http://vxheaven.org/vl.php and benign files were collected from a PC running windows XP includes operating system files and various windows application. Dataset consist of 4722 PE files, where 3000 are Trojan and 1722 benign programs.

The goal was to gather useful features and extract useful features from PEiD that distinguish between malicious and benign files where distribution of different packed, not packed, Trojans and benign programs are there.

FEATURE SELECTION AND EXTRACTION

In this, decision trees are divided into subsets and concurrently a connected decision tree is developed incrementally. Decision tree construction is to find attributes. These attributes return the highest information gain (i.e., the most homogeneous branches). Here, data set one comprises 890 Trojan codes and 150 benign programs. The expected results using the using the equation:

= - (|benign|/|X| log2 |benign|/|X|+|Trojan|/|X| log2 |Trojan|/|X|)
=-150/(890+150)log2(150/890+150)+(-890)/890+150log2 (890/890+150).

The attribute with the largest information gain is chosen the decision node and negligible information gain can be discarded to reduce number of features to speed up the classification. ID3 algorithm is run recursively, until all data are classified.

CLASSIFICATION AND MODEL TRAINING

Each data set was then fed to Naive Bayes technique; experiments were repeated several times using random sub-sampling holdout method, to obtain the accuracy from the iteration method. So, these results can be obtained [9].

TABLE 1.Results from Iteration Method

Naive	1 byte	76.1	41.2	73.3
Bayesian	2 byte	80.7	41.2	77.1
Decision	1 byte	93.2	29.4	89.5
Tree	2 byte	94.3	23.5	91.4

7. RESULTS

Results have been obtained after testing the data using the obtained data set to evaluate the correctness of the classification model for Trojan detection. The four estimates define the member. True Positive (TP): Number of programs correctly identified as Trojan codes. False Positive (FP): Number of benign programs incorrectly identified as Trojan codes. True Negative (TN): Number of programs correctly identified as benign programs.

False Negative (FN): Number of Trojan codes incorrectly identified as Trojan codes.

The action of every classifier was evaluated using false alarm, overall accuracy and detection rate: Detection Rate (DR): Percentage of correctly identified malicious programs.

Detection Rate = TP/(TP+FN)

False Alarm Rate (FAR) or False Positive Rate (FPR): Percentage of wrongly identified benign Programs –

False Alarm Rate = FP/(TN+FP)

Overall Accuracy: Percentage of correctly identified

Programs Overall Accuracy = TP+TN/(TP+TN+FP+FN)

This data set was experimented. The unknown Trojan detection rates 93.2 per cent and 76.1 per cent with accuracies of 89.5 per cent and 73.3 per cent were obtained in first experiments. Each element consists of only the op-code whereas unknown Trojan detection rates are 94.3 per cent and 80.7 per cent and accuracies rise to 91.4 per cent and 77.1 per cent. More information is surfaced in each iteration and therefore Naïve Bayes classifier performs more accurately [10].

8. CONCLUSION

Naïve Bayes technique is simple to put into practice and we achieve amazing results in large number of cases. But practically, dependencies exist among variables. This article examined Trojan Detection using Naive Bayes technique. As Trojan detection is one of the major measures for security. This technique automatically extracts Trojan qualities from Trojan programs. Further, these qualities are used for classification. The obtained results and outcomes indicate that the rate of detection - the Decision Tree and Naive Bayes classifiers computed as 94.3 per cent and 80.7 per cent and the accuracy 91.4 per cent and 77.1 per cent respectively. This shows that Decision tree performs well if compared with Naive Bayes classifier. We need to put into effect suitable policies and checkup the legal aspects and need to undertake privacy from all directions for the security purpose.

REFERENCES

[1] Matthew G. Schultz,Eleazar Eskin,Erez Ado and Salvatore J. Stolon"Data Mining Methods for Detection of New Malicious Executable".

[2] Muezzin Ahmed Siddiqui, Morgan Wang, "Detecting Trojans Using Data Mining Techniques", January 2008.

[3] Tzu-Yen Wang, Shi-Jinn Horn, Ming-Yang Su, Chin-Suing Wu, Pang-Chu Wang and Wei-Zen Su ," A Surveillance Spyware Detection System Based on Data Mining Methods",2004.

[4] Yuen Kou, Chang-Tine Lu, Sir rat Sinvongwattana You-Ping Huang " Survey of Fraud Detection Techniques", March, 2004.

[5] Jay-Hwang WANG, Peter S. DENG, Yi-Sheen FAN, Li-Jing JAW,Yu-Ching LIU,"VIRUS DETECTION USING DATA MINING TECHINQUES",2004.

[6] Sung, A.H., Xu,J. ,Chavez,P., Mukkamala, S. "Static analyzer of vicious executables", 20th Annual Computer Security Applications Conference,2004.

[7] Karta Mathura, Sari Hiranwal, "A Survey on Techniques in Detection and Analyzing Malware Executable", April 2013.

[8] DauberKauri R. Chakra" Feature selection and clustering for malicious and benign software characterization" August, 2014.

[9] Sara Najari, Iman Lotfi, "Malware Detection using Data Mining Techniques", International Journal of Intelligent Information Systems, October 20, 2014.

[10] Mittal A. Saeed, Ali Selma, Ali M. A. Abuagoub," A Survey on Malware and Malware Detection System" International Journal of Computer Applications, April 2013.

[11] Shafiqul Abidin, Mohd Izhar , "Attacks on Wireless and its Limitations" , International Journal of Computer Science and Engineering, Volume 5, Issue 11, November 2017.