Predictive Modeling and Classification of DDoS Incidents Using Machine Learning

GUMMADI TIRUMALA1, KATAMREDDY MAHENDRA2

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

#2 Assistant Professor, Department of CSE-IoT, PBR Visvodaya Institute of Technology and Science, Kavali

ABSTRACT_ The introduction of rogue apps poses a significant threat to the Android platform. Most types of network interfaces rely on integrated functions to steal users' personal information and launch attack activities. In this research, we offer a method for detecting malware that is both effective and automatic, based on text semantics in network traffic. In particular, we treat each HTTP transaction generated by mobile apps as a text document that can be analysed using natural language processing to extract text-level information. Later, network traffic is used to develop an effective malware detection model. We investigate the traffic flow header with the N-gram method from natural language processing (NLP). Then, we offer an automatic feature selection approach that uses the chi-square test to discover relevant characteristics. It is used to see if there is a substantial relationship between the two variables. We offer a unique technique for malware detection that use NLP methods and treats mobile communications as documents. We use an artificial feature selection approach based on N-gram sequences to extract relevant features from traffic flow semantics. Our approaches identify some malware that can evade detection by antiviral scanners. In addition, we provide a detection system that directs traffic to your institutional enterprise network, home network, and 3G / 4G mobile networks. Integrating the system connected to the computer to detect questionable network behaviour.

1.INTRODUCTION

Typically, distributed denial-of-service attacks are used to describe distributed network attacks. The Android platform faces a serious threat from malicious apps. The majority of network interfaces with integrated functions steal personal information from users and initiate attacks. Using the text semantics of network traffic, we propose a method for automatic

malware detection in this paper. Particularly, we think of each mobile app-generated HTTP flow as a text document that can be processed by natural language processing to extract features at the text level. A useful malware detection model is later created by utilizing network traffic. The natural language processing (NLP) N-gram method is used to examine the traffic flow header. Then, at that point, we propose a programmed highlight determination calculation in light of chi-square test to distinguish significant elements. It is used to determine whether the two variables have a significant connection. Treating mobile traffic as documents is our novel approach to malware detection using NLP techniques. To extract meaningful features from the semantics of traffic flows, we employ an N-gram sequence-based automatic feature selection algorithm. Some malware that can't be detected by antiviral scanners is discovered by our methods. We also create a detection system that directs traffic to your home network, 3G/4G mobile network, and institution's enterprise network. integrating the computer-connected system to identify suspicious network behaviors

2.LITERATURE SURVEY

2.1 TITLE:

An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection

AUTHOR:

Monowar H.Bhuyan[a]D.K.Bhattacharyya[b]J.K.Kalita[c]

CONTENT:

Distributed Denial of Service (DDoS) attacks represent a major threat to uninterrupted and efficient Internet service. In this paper, we empirically evaluate several major information metrics, namely, Hartley entropy, Shannon entropy, Renyi's entropy, generalized entropy, Kullback–Leibler divergence and generalized information distance measure in their ability to detect both low-rate and high-rate DDoS attacks. These metrics can be used to describe characteristics of network traffic data and an appropriate metric facilitates building an effective model to detect both low-rate and high-rate DDoS attacks. We use MIT Lincoln Laboratory, CAIDA and TUIDS DDoS datasets to illustrate the efficiency and effectiveness of each metric for DDoS detection.

2.2 TITLE:

Defending against flooding-based distributed denial-of-service attacks: a tutorial

AUTHOR:

Rocky K. C. Chang

CONTENT:

Flooding-based distributed denial-of-service (DDoS) attack presents a very serious threat to the stability of the Internet. In a typical DDoS attack, a large number of compromised hosts are amassed to send useless packets to jam a victim, or its Internet connection, or both. In the last two years, it was discovered that DDoS attack methods and tools are becoming more sophisticated, effective, and also more difficult to trace to the real attackers. On the defense side, current technologies are still unable to withstand large-scale attacks. The main purpose of this article is therefore twofold. The first one is to describe various DDoS attack methods, and to present a systematic review and evaluation of the existing defense mechanisms. The second is to discuss a longer-term solution, dubbed the Internet-firewall approach,that attempts to intercept attack packets in the Internet core, well before reaching the victim

3.PROPOSED SYSTEM

The method we use to find the DDoS attack is described in this section. The method that follows is characterized by the five-step application process of data mining techniques in network systems that was discussed in. The proposed approach's main goal is to reduce irrelevant and noisy network traffic data before the preprocessing and classification stages of DDoS detection while maintaining high accuracy, low false positive rate, fast running time, and low resource consumption. The estimation of the FSD features' entropy over a time-based sliding window is the first step in our strategy. The co-clustering algorithm divides the received network traffic into three clusters whenever the average entropy of a given time window exceeds either its lower or upper thresholds. DDoS attacks frequently result in abrupt shifts in the distribution of incoming network traffic, which can be detected by entropy estimation over time sliding windows. DDoS traffic is thought to be in incoming network traffic during the time windows with abnormal entropy values. The emphasis just on the thought time windows permits to channel significant measure of organization traffic information, accordingly just pertinent information is chosen for the excess strides of the proposed approach. Additionally, when there is no abnormal entropy, significant resources are saved. We estimate the information gain ratio between each of the obtained clusters and the average entropy of the FSD features between the received network traffic data during the current time window and the

normal cluster. During a DDoS period, the amount of attack traffic generated is significantly greater than the normal traffic, as discussed in the preceding section. As a result, the two clusters that contain only normal traffic and those that preserve more information about the DDoS attack can be distinguished by estimating the information gain ratio using FSD features. Therefore, excluding some noisy normal instances of the network traffic data for classification in the form of a cluster improves classification accuracy and false positive rates. Assuming that one cluster contains only normal traffic following the network traffic clustering, another contains only DDoS traffic, and the third contains both DDoS and normal traffic

3.1 IMPLEMENTATION

There are three modules can be divided here for this project they are listed as below

- User Apps
- DDOS Attack Deduction
- Classifications of DDOS attack
- Graphical analysis

From the above four modules, project is implemented. Bag of discriminative words are achieved

1. **User Apps**

User handling for some various times of smart phones ,desktops laptops and tablets .If any kind of devices attacks for some unauthorized Malwaresoftwares.In this Malwareon threats for user personal dates includes for personal contact, bank account numbers and any kind of personal documents are hacking in possible.

1. **DDOS Attack Deduction**

User search the any link Notably, not all network traffic data generated by malicious apps correspond to malicious traffic. Many malwaretake the form of repackaged benign apps; thus, Malware can also contain the basic functions of a benign app.Subsequently, the network traffic they generate can be characterized by mixed benign and malicious network traffic.We examine the traffic flow header using Co-clustering algorithm from the natural language processing (NLP).

3. **Classifications of DDOS Attack:**

Here, we compare the classification performance of Co-clustering algorithm with other popular machine learning algorithms. We have selected several popular classification algorithms. For all algorithms, we attempt to use multiple sets of parameters to maximize the performance of each algorithm. Using Co-clustering algorithm algorithms classification for malwarebag-of-words

weightage.

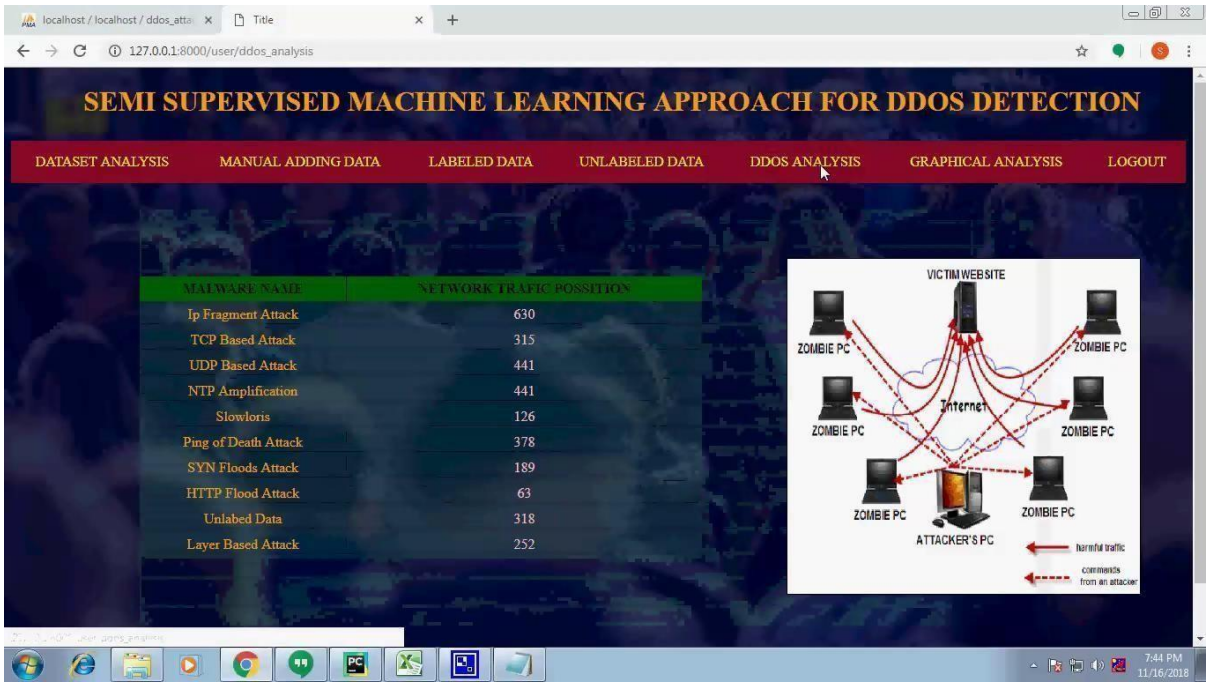
4. Graphical analysis

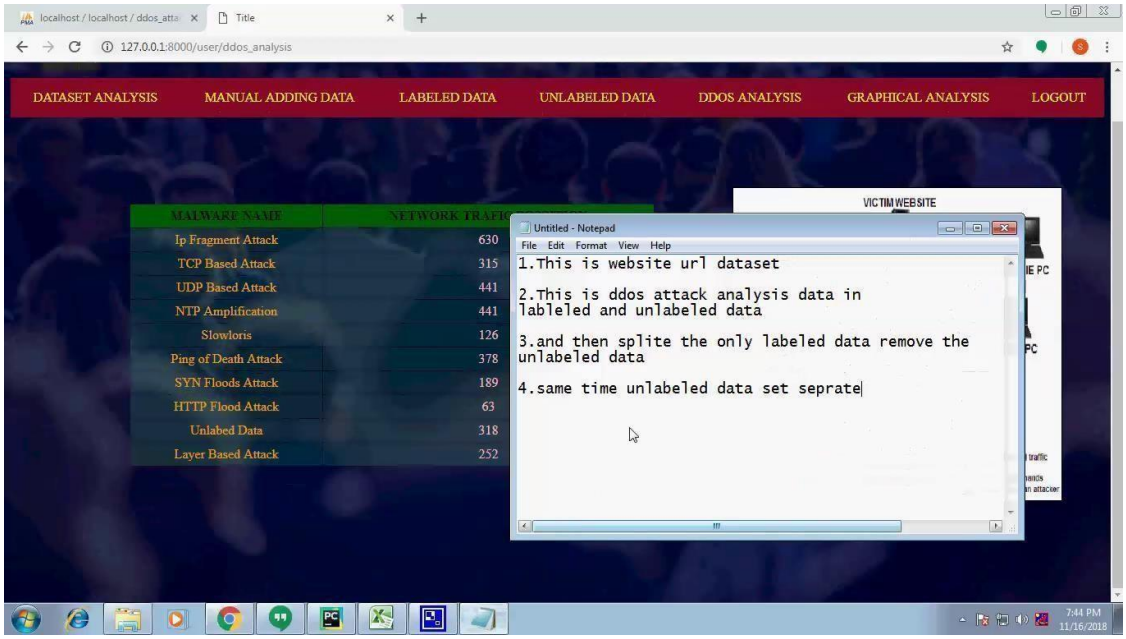
The graph analysis is done by the values taken from the result analysis part and it can be analyzed by the graphical representations. Such as pie chart, pyramid chart and funnel chart here in this project.

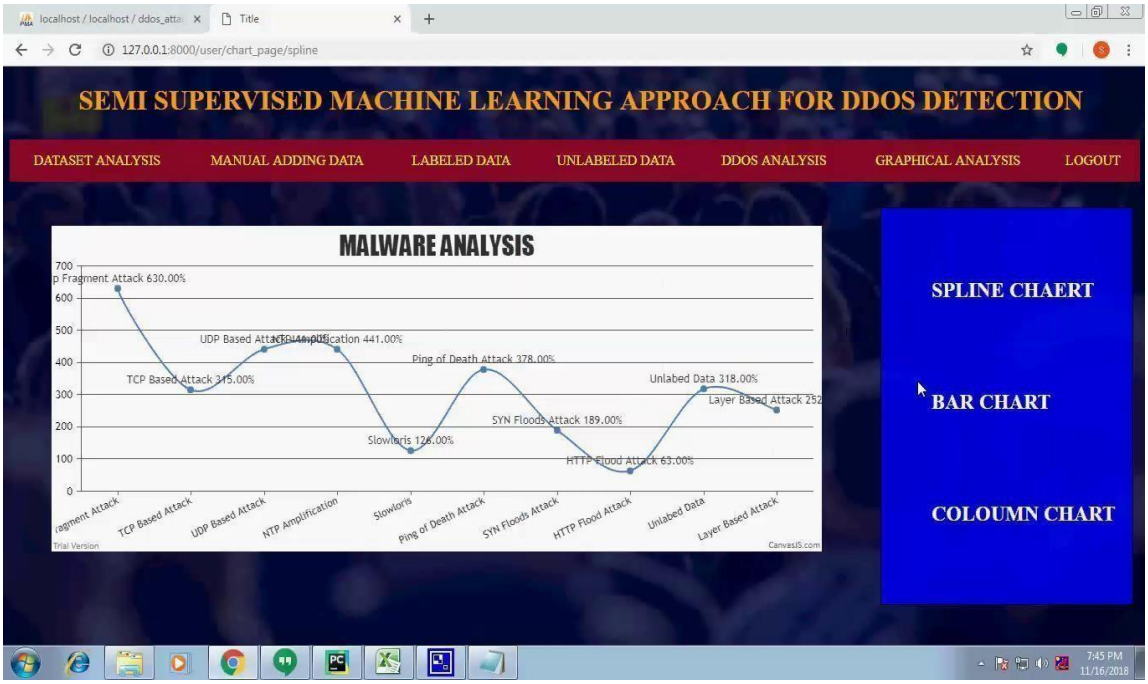
3.1.1 ALGORITHM

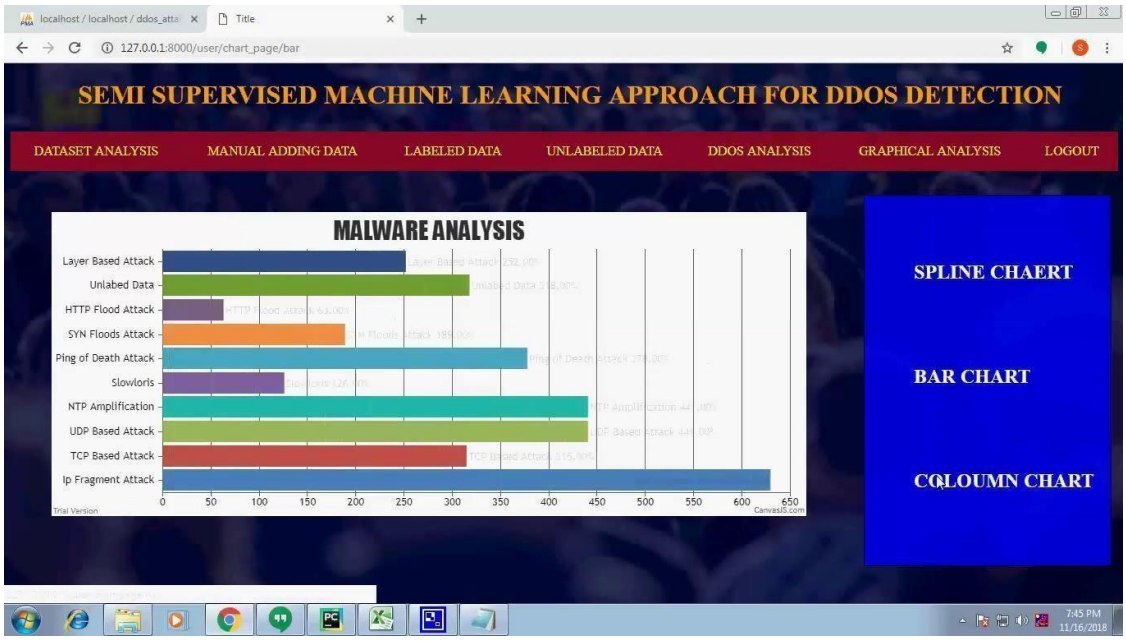
Co-clustering algorithm performs a simultaneous clustering of rows and columns of a data matrix based on a specific criterion . It produces clusters of rows and columns which represent sub-matrices of the original data matrix with some desired properties. Clustering simultaneously rows and columns of a data matrix yields three major benefits: Dimensionality reduction, as each cluster is createdbased on a subset of the original features. More compressed data representation with preservation of information in the original data. Significant reduction of the clustering computational complexity. The co-clustering computational complexity is O(mkl+ nkl) which is much smaller than that of the traditional Kmeans algorithm O(mnk) . Wherem is the number of rows, n is the number of columns, kis the number of clusters and l is the number of column cluster

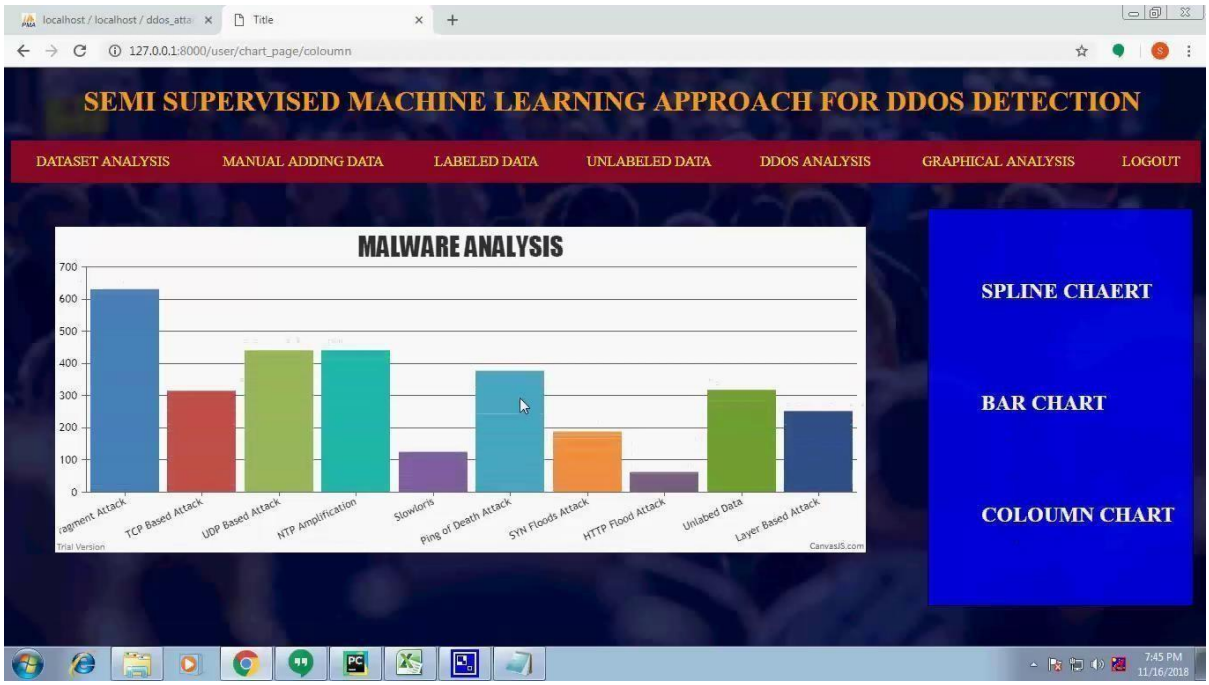
4.RESULTS AND DISCUSSION

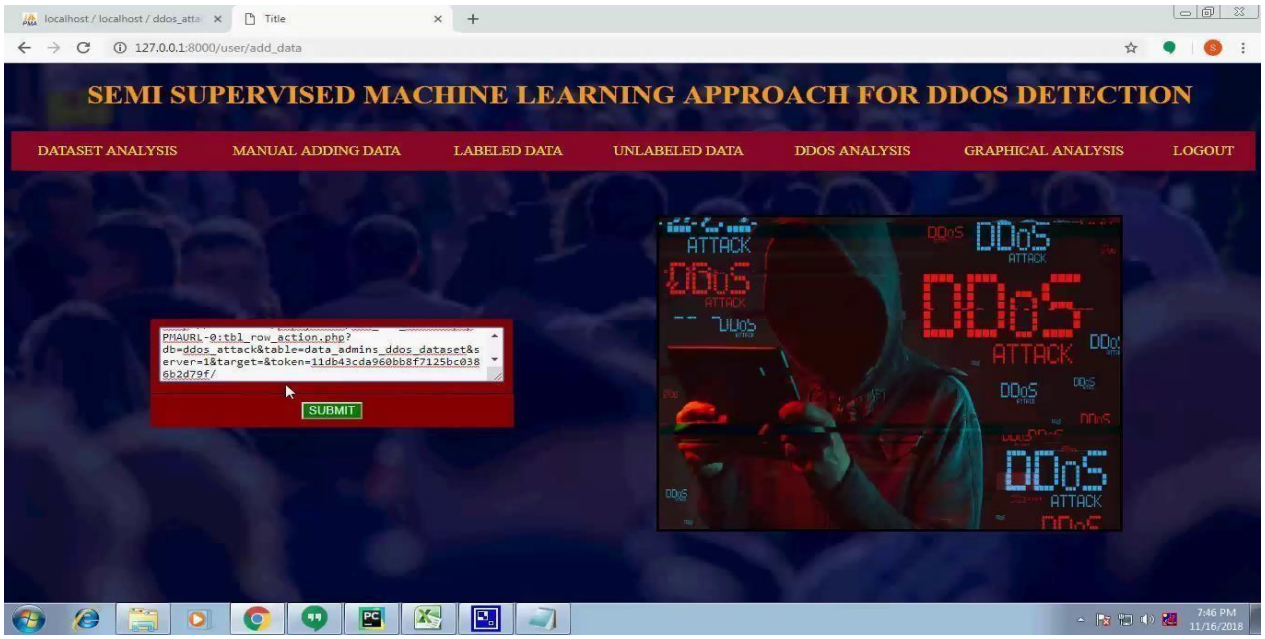



DATASET ANALYSIS
MANUAL ADDING DATA
LABELED DATA
UNLABELED DATA
DDOS ANALYSIS
GRAPHICAL ANALYSIS
LOGOUT
MALWARE NAME
NETWORK TRAFFIC
Ip Fragment Attack 630
TCP Based Attack 315
UDP Based Attack 441
NTP Amplification 441
Slowloris 126
Ping of Death Attack 378
SYN Floods Attack 189
HTTP Flood Attack 63
Unlabed Data 318
Layer Based Attack 252
VICTIM WEBSITE
Untitled - Notepad
File Edit Format View Help
1.This is website url dataset
2.This is ddos attack analysis data in lableled and unlabeled data
3.and then splite the only labeled data remove the unlabeled data
4.same time unlabeled data set seprate


SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION
DATASET ANALYSIS
MANUAL ADDING DATA
LABELED DATA
UNLABELED DATA
DDOS ANALYSIS
GRAPHICAL ANALYSIS
LOGOUT
MALWARE ANALYSIS
SPLINE CHAERT
BAR CHART
COLOUMN CHART


SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION
DATASET ANALYSIS
MANUAL ADDING DATA
LABELED DATA
UNLABELED DATA
DDOS ANALYSIS
GRAPHICAL ANALYSIS
LOGOUT
MALWARE ANALYSIS
Layer Based Attack
Unlabed Data
HTTP Flood Attack
SYN Floods Attack
Ping of Death Attack
Slowloris
NTP Amplification
UDP Based Attack
TCP Based Attack
Ip Fragment Attack
SPLINE CHAERT
BAR CHART
COLOUMN CHART


SEMI SUPERVISED MACHINE LEARNING APPROACH FOR DDOS DETECTION
DATASET ANALYSIS
MANUAL ADDING DATA
LABELED DATA
UNLABELED DATA
DDOS ANALYSIS
GRAPHICAL ANALYSIS
LOGOUT
MALWARE ANALYSIS
TCP Based Attack
UDP Based Attack
NTP Amplification
Slowloris
Ping of Death Attack
SYN Floods Attack
HTTP Flood Attack
Unlabed Data
Layer Based Attack
SPLINE CHAERT
BAR CHART
COLOUMN CHART



5.CONCLUSION

Android is a brand-new and fastest-growing malware threat. Currently, the expanding size and variety of mobile malware pose no threat to numerous research methods or antivirus scanners. We present a method for mobile malware detection based on network traffic flows that makes the assumption that each HTTP flow is a document and employs NLP string analysis to analyze HTTP flow requests. A useful malware detection model is created using the N-Gram line generation, feature selection, and SVM algorithms. Our evaluation demonstrates this solution's efficacy, and our trained model outperforms previous methods in significantly identifying malicious leaks and false warnings. The false rate for harmful traffic is 0.45%, while the harmful traffic detection rate is 99.15 percent. Utilizing the newly discovered malware further validates the proposed system's performance. The sample outperforms other popular anti-virus scanners in its ability to identify 54.81 percent of potentially harmful applications when applied in actual environments. The test demonstrates that our model can be detected by malware models, and that this does not prevent other virus scanners from being detected. It is also possible to obtain basically brand-new VirusTotal detection reports for malicious models. Added: We will kindly retrain, refresh, and update the new malware as soon as new tablets are added to training samples.

REFERENCES

1. Bhuyan MH, Bhattacharyya DK, Kalita JK (2015) An empirical evaluation of information metrics for low-rate and high-rate ddos attack detection. Pattern Recogn Lett 51:1–7

Article Google Scholar

2. Lin S-C, Tseng S-S (2004) Constructing detection knowledge for ddos intrusion tolerance. Exp Syst Appl 27(3):379–390

Article Google Scholar

3. Chang RKC (2002) Defending against flooding-based distributed denial-of-service attacks: a tutorial. IEEE Commun Mag 40(10):42–51

Article Google Scholar

4. Yu S (2014)

Distributed denial of

service attack and

defense. Springer, Berlin

Book Google Scholar

5. Wikipedia (2016) https://en.wikipedia.org/wiki/2016_Dyn_cyberattack. (Online; accessed 10 Apr 2017)

6. theguardian (2016) Ddos attack that disrupted internet was largest of its kind in history, experts say. https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet. (Online; accessed 10 Apr 2017)

7. Kalegele K, Sasai K, Takahashi H, Kitagata G, Kinoshita T (2015) Four decades of data mining in network and systems management. IEEE Trans Knowl Data Eng 27(10):2700– 2716

Article Google Scholar

8. Han J, Pei J, Kamber M (2006) What is data mining. Data mining: concepts and techniques. Morgan Kaufinann

9. Berkhin P (2006) A survey of clustering data mining techniques. In: Grouping multidimensional data. Springer, pp 25–71

10. Mori T (2002) Information gain ratio as term weight: the case of summarization of ir results. In: Proceedings of the 19th international conference on computational linguistics, vol
1. Association for Computational Linguistics, pp 1–7

2016 dyn cyberattack.

11. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3– 42

Article MATH Google Scholar

12. Tavallaee M, Bagheri E, Lu W, Ghorbani A-A (2009) A detailed analysis of the kdd cup 99 data set. In: Proceedings of the second IEEE symposium on computational intelligence for security and defence applications 2009

13. Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA (2012) Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Comput Secur 31:357–37