CHAPTER 1

INTRODUCTION

1.1. Introduction

Internet is global network connecting billions of devices like desktop/laptop computers, smart phones, tablets, smart TVs etc. This is used by people to communicate with each other, finding information, sharing resources, accessing all sort of online content, office work, online shopping, banking etc. The use of Internet makes our life easy but it also increases risk of victimization of illegal activity or abuse. Malware intrusions are referred to the attacks on network/host against vulnerable services. These intrusions are carried out to access personal files, destroy sensitive files, steal or gain unauthorized information etc. Malware like viruses, worms, Trojan horses, root kit, botnet, spyware etc. are used for intrusion into computer networks. The Internet and computer networks require security from such malware attacks and intrusions. Security systems like anti viruses, firewalls, Intrusion Detection Systems (IDS) are used to protect computer systems from hackers and crackers.

1.2. Malware Epidemic and Intrusions: Background

Malware (Malicious Software) is used to gain access to computer system, steal information, block services etc. without the knowledge of owner. Malware can be detected by inspecting its payload or by processing network traffic dataset. IDS are also used to detect malware [1]. Payload inspection fails to detect unknown attacks while IDS reacts fast for dynamic behavior malware. Most of the malware are now network based which makes IDS suitable for detection. Unlike anti-viruses, IDS does not require definition updates for malware detection.

1.2.1. Growth of Internet

Communication network was used first time in 1950 for military radar system. Now even houses are connected with computer networks. Internet revolutionized the way
of communication, resource sharing and entertainment of human beings and turned the world into global village. It was used to connect government laboratories mainly for military purposes but now it is used everywhere from individuals to offices, government, social groups, educational institutions, corporate etc. Affordable pricing model and cheap techniques make it possible to increase penetration rate of Internet. According to Internet World Stats [2] around 42.4% of world population has access of Internet in December, 2014. In India around 19.19% of population enjoys access to Internet in June, 2015 [3]. Figure 1.1 shows the growth of Internet in the world.

![Figure 1.1: Growth of Internet in terms of number of users in the World](image)

In 1995, there were only 16 million Internet users while in December 2014 there were 3079 million users. The Internet access percentage of world population has increased from 0.4% in 1995 to 42.4% in 2014 [2]. Figure 1.2 shows the growth of Internet in India from 2000 to 2014 [3]. Internet was introduced in India in 1995 and only 0.53% of India’s population had Internet access in 2000. In the last decade, the growth of Internet has witnessed exponential growth due to easy availability of low cost devices.

1.2.2. Growth of Internet Attacks

Growth of Internet has transformed the way of communication, information sharing, shopping and working of the people. Easy accessibility and connectivity also increases threats to it. According to Kaspersky [4] survey in 2013, 91% of the
companies had at least one external security threat incident. Discovery rate of malware is as low as 6%. It may costs companies $50000 to $2.4 Million.

![Figure 1.2: Growth of Internet in terms of number of users in India](image)

Figure 1.2: Growth of Internet in terms of number of users in India

Figure 1.3 shows the number of incidences of various attacks faced by companies in 2013.

![Figure 1.3: Incidences of various attacks](image)

Figure 1.3: Incidences of various attacks

On average 35% of companies encounter data leakage because of these attacks. 66% of various companies attacked by viruses, worms, spyware and other malicious
programs. 61% of the companies attacked by spam only. Other significant attack consists of Denial of Services (DoS) and phishing attack [5].

According to PandaLab [6] annual report, 34% of the total malware are created in 2014. Figure 1.4 shows the distribution of types of malware created in 2014. Trojans remains leading malware as it contributes 68.84% of newly created malicious software. Viruses and worms are other types of malware which are created in significant numbers.

![Figure 1.4: Type of new malware created in 2014](image)

Figure 1.5 shows the amount of infection created by different type of malware. 65.02% of the infected systems are attacked by Trojans. Trojans is the major contributor of security threats [7]. 45% of the web attacks mitigated in 2013 by Kaspersky products are carried out using malware located in USA and Russia. India continues to be in top 20 countries with higher risk of infection via Internet and ranks 5th in the list of countries where users face highest risk of local infections. 59.26% of users in India are infected by malware in 2013 [7][8].
Figure 1.6 shows the growth rate of new vulnerabilities from the year 2006 to 2014. New vulnerabilities are also known as zero day attacks. These attacks are previously unknown and signature based detection system are not able to detect them.

These vulnerabilities are much more serious as they are detected only after exploited by attackers [9]. From this Figure it is certain that despite so much research and
network security infrastructure development, exploitation by new attacks are not mitigating.

1.3. Pilot Study: A Motivation

A sequence of packets from source to destination device is known as network traffic/packet flow. This flow can contain all the packets transmitted for particular transport connection or set of IP packets in certain time interval. An experiment is setup to capture network traffic dataset for initial investigation. Data is captured using Panjab University - Campus Area Network (PU-CAN). The amount of broadcast packets is analyzed for malicious patterns. This traffic is manually classified and certain alarming results are found despite the use of all security measures by network administrators. This study shows existence of unknown networks as well as suspected broadcast. This study is discussed as below:

1.3.1. Network Traffic Flow Monitoring

Ability to monitor and identification of network traffic flow is critical aspect for security professionals in network management. Monitoring of network traffic is very important for efficient and secure network management. Classifying network traffic can detect hidden patterns of possible threat. This monitoring is crucial for computer networks for detecting and preventing anomalies and attacks. Traffic classification and monitoring techniques have been used to classify anomalies from traffic flow data [10].

1.3.2. Experimental Setup

There are various segregated sub-networks in the PU-CAN. The sub-network which kept on surveillance provides network services to three boy’s hostels with approximately 1200 users. Figure 1.7 shows the network diagram used for setting up of this experiment. The sub-network used for the study consists of network of boys hostel number 4, 5 and 6. In these hostels, the Internet service is provided by wireless access points. In campus, there is a team of network administrators and network
Introduction

engineers to maintain the network. All the access points are connected with Computer Center by high speed optical fiber cable. The access of network services is controlled and monitored by squid proxy server [11]. Firewall is also installed to stop unauthorized access. The network administrator manages a Dynamic Host Configuration Protocol (DHCP) server to dynamically assign Internet Protocol (IP) addresses to the users. The DHCP server is configured to assigns IP addresses in the range of 172.16.40.1 to 172.16.43.254 except some special IPs meant for servers like 172.16.40.1 for Domain Name Server (DNS), 172.16.40.2 for DHCP server and squid proxy server. IP address 172.16.40.11 is used for server which was deployed for capturing of network traffic.

![Network Diagram](image_url)

Figure 1.7: Network Diagram

*Fedora-14 Linux* [12] based server is used along with data capturing and processing utilities. *Tcpdump* [13] is open source Linux based packet analyzer tool used for capturing and analyzing the packet information travelling through the network. It shows the information of TCP/IP and other packets travelling through that network. It is used to view and save packet information as well as to view login IDs, passwords, the URLs and content of websites. It works on *libpcap* which is a C/C++ library for packet capture. Captured data is saved as *pcap* file. Network traffic data is collected
from July 30, 2011 to August 21, 2011. A part of collected data is pre-processed to remove duplicate packets. Open source utility, *wireshark* [14] having easy and seamless Graphical User Interface (GUI), is a packet analyzer used for network analysis, packet capturing, filtering and troubleshooting. Filters can be applied to display packet information as per user's requirements. This software allows plug-ins for integrating new protocols. It is used to convert `.pcap` to `.csv`. Only broadcast data are used for initial investigation.

1.3.3. **Pilot Study: Results**

Figure 1.8 show the difference between total broadcast captured versus genuine official network. It is found that only 33.91% of total broadcast is generated by genuine network of 172.16.40.0/22.

![Figure 1.8: Total broadcast versus genuine broadcast traffic](image)

Figure 1.8 shows the comparative percentage of broadcast traffic by genuine, identified malicious and unidentified malicious networks. It is also found that a network of miss-configured IPs viz 169.254.0.0 subnet exists and generating 3.84% of total broadcast, thereby contributing in network congestion. An unauthorized network of series 192.168.11.0 is also identified during this study. This network is using
campus network infrastructure without official consent and considered as possible threat to network. This unauthorized network may be created by students or some attackers by just configuring another DHCP server to allocate this series of IPs so that genuine user unable to get official IP. It amounts to denial of Internet/network services.

Figure 1.9: Comparison of percentage of broadcast of genuine network, identified malicious and unidentified suspected class in respect of total broadcast

Figure 1.10 shows the broadcast traffic generated by various networks.

Figure 1.10: Broadcast traffic generated by different identified network

This unauthorized network generated 8.90% of total broadcast during the study period. 53.33% of network broadcast traffic is detected from computers like un-
resolved source and destination, zero series network and unresolved IPv6 network. It is found that although signature based anti malware and antivirus systems have been installed in the network, still unauthorized anomalous activities lead the misuse of bandwidth. From this pilot study, it is suggested that anomaly based IDS is required to properly manage the network and find out unauthorized activities operating through that network. A study may be carried out to find the misuse of bandwidth and other resources.

Although firewall and security measures are used in PU-CAN but still there are malicious activities. These activities cannot be detected by signature based systems. It is suggested that IDS should be configured along with firewall. These types of anomalies can be detected only by anomaly based IDS.

1.4. Need of Intrusion Detection System for Malware Detection

In the world of rapidly developing technology, Internet is facing various threats such as malware. It is a generic term which includes viruses, worms, Trojan horses, spyware, adware etc. Hackers uses malware for intrusion into computer networks to steal information like password, files etc. Even internal Local Area Network (LAN) is also seriously struggling with malware. Some malware spread on the whole network and start communication which unnecessarily increases traffic. Due to lesser availability of bandwidth users start suffering degradation in quality of service. This is affecting productivity of internal computer networks in terms of bandwidth and other resources. Some of the malware emphasizes only on bandwidth exhaustion so that user doesn’t get their share of bandwidth. Malware uses advanced features like dynamic ports, IP address spoofing, encrypted payload etc. to avoid detection. These intrusions need to be identified before any type of loss to the organizations. Signature based payload inspection can detect only known attacks. The detection of malware is difficult due to its dynamic behaviors in signature/ payload inspection based systems. Unknown malware attacks can be detected by discovering patterns in network traffic dataset. IDS is designed to defend the network from malicious activities. Anomaly based IDS learn normal behavior from network traffic dataset to detect attacks. This
dataset is huge and imbalanced due to which machine learning based IDS faces problem to process whole dataset. This may also results in over fitting and biased performance. So, it is necessary to identify intrusions created by malware through network traffic behavior. These defense systems also work with other security systems like firewall to provide enhanced security. IDS can be used as a second wall of defense from malware/intrusions and tries to find if someone break through the firewall and access any system.

1.5. Research Gaps

Research gaps in anomaly based malware/intrusion detection by processing of network traffic dataset are summarized as below:

i. The network data is huge, imbalanced and has large numbers of features. So it is practically impossible to run classic soft-computing techniques on whole data. Sampling and feature selection technique may improve the analysis of such huge dataset. But these operations may change overall characteristics of data. Efficient algorithms for feature selection and sampling are desirable for intrusion detection.

ii. Most of the network traffic profiling techniques are port/ packet/ payload/ flow and behavioral based. Port based approaches are no longer valid as malware may keep changing ports. Packet based approach has low accuracy rate and most of them are not in real-time. Payload based approaches are offline, slow and take time for corrective action. Activity and behavioral based approaches are still in initial stage.

iii. Anomaly detection approach with high accuracy, low false alarm rate, less detection time and good feature selection technique is required.

1.6. Research Objectives

i. To Study and analyze various issues such as handling of huge dataset, sampling, feature selection, feature extraction etc. for malware detection and network traffic profiling.
ii. To formulate, design and develop the soft computing based technique for malware detection and network traffic profiling.

iii. Test the proposed technique for malware detection and network traffic profiling.

iv. Validation of the proposed technique.

1.7. Major Contributions of the Thesis

- Undertaken the pilot study of captured Panjab University – Campus Area Network (PU-CAN) traffic dataset emphasizing importance and need of anomaly based detection techniques.
- Analyzed various issues in network traffic dataset like hugeness, imbalances and large feature set.
- Generated synthetic dataset namely Panjab University – Intrusion DataSet (PU-IDS).
- Proposed an intrusion detection technique based on Network Traffic Profiling and Online Sequential Extreme Learning Machine (OS-ELM) and named as Panjab University – $\alpha FST^\beta$ ($PU-\alpha FST^\beta$).
- Developed a technique to handle issues like hugeness and imbalanced nature of various datasets, large feature set, low accuracy and high rate of false alarms.
- Performance evaluation of proposed technique using PU-IDS, Kyoto University and NSL-KDD dataset.
- Minimized time and space complexity by alpha profiling which reduces number of comparisons by 89.64% for PU-IDS, 95.94% for Kyoto and 85.76% for NSL-KDD dataset.
- Optimum selection of features using ensemble of feature selection technique which reduces features by 78.05% for PU-IDS, 31.25% for Kyoto and 48.78% for NSL-KDD dataset.
- Training dataset is reduced by beta profiling. This decrease the sample training size by 93.97% for PU-IDS, 36.71% for Kyoto, 7.83% for binary-class NSL-KDD dataset and 7.66% for multi-class NSL-KDD dataset.
- Achieved accuracy of 99.99% and false positive rate of 0.03% in 1.39 seconds for PU-IDS dataset.
- Yielded accuracy of 96.37% and false positive rate of 5.76% in detection time of 1.22 seconds for Kyoto University dataset.
- Achieved accuracy of 98.66%, false positive rate of 1.74% with lowest time of 2.43 seconds for binary-class NSL-KDD dataset.
- Gained overall accuracy of 97.67%, FP rate of 1.74% in 2.65 seconds of detection time for multi-class NSL-KDD dataset.

1.8. **Thesis Structure**

This thesis presents a proposed intrusion detection technique which considers various issues like imbalances and hugeness of network traffic dataset, large feature set, low accuracy and high rate of false alarms.

Chapter 2 starts with introduction to Intrusion Detection System. Related work of various researchers in the past is also discussed and summarized in this chapter.

Chapter 3 discusses various available network traffic datasets. This chapter also discusses different dataset which have been used in this thesis. The characteristics of these dataset are also explained including the process of synthetic dataset generation. This chapter also discusses various issues like hugeness, imbalances, larger feature set etc. related to network traffic dataset.

Chapter 4 explains proposed technique for malware/intrusion detection. This chapter discusses all three experiments carried out in the thesis along with criteria for performance evaluation.

Chapter 5 presents the results yielded by proposed technique with PU-IDS, Kyoto University and NSL-KDD (binary and multi-class) dataset. Comparative analyses of results with other existing techniques are also carried out.

The final chapter concludes the thesis by summarizing results of research work. Future directions are also discussed in this chapter. From the thesis work, various research papers are published which are attached as an appendix.
1.9. **Summary**

This chapter discusses the growth of Internet and attacks on it. Network traffic dataset is captured from Panjab University campus to analyze the requirement of anomaly based IDS. The network has proxy server, firewall and all other precautions against security threats. Network traffic data is collected during the period of study to find out possible threat. Anomaly based detection method is suggested to detect malware intrusions. This study acts as a motivation of thesis work. The need of IDS to detect malware intrusions is also discussed. This chapter also includes research gaps and objectives of the thesis. The major contributions of the thesis are briefly summarized. The synopsis of all the chapters of the thesis is also presented.