Chapter 1

Introduction

With the advancement of technology and decrease in cost of Personal Computers, access to the Internet has become easier and World Wide Web sites have become more sophisticated and inviting. In 1992 Paul Linder and Mark McCahill released Gopher tool which allowed researchers to retrieve required and specific data from numerous locations (Prezi, 2016a). In 1993 Marc Andreesen (founder of Netscape) developed a web browser at the University of Illinois and World Wide Web became a public domain (Mayo and Newcomb, 2009). In 1994 shopping malls arrived on the Internet, allowing to order pizza from Pizza Hut online or do online banking transactions (Prezi, 2016a). 16 million users were connected to Internet in 1995 and this number reached to 2,937 million in 2014 (Prezi, 2016b). With this rapid growth in number of users connected to internet, all organizations started provided their services over internet.

With rapid growth in the number of services provided over Internet; number of attacks on these services is also increased. Any act of compromising confidentiality, integrity or availability of online or offline computer system is called as Intrusion. 'Vladimir Levin' is the first publicly known Internet bank robber from Russia (Prezi, 2016a).

Albert Gonzalez stole the information of 45.7 million payment cards of US retailer TJX’s customers, which resulted into loss of 256 million dollar (Palermo, 2015). In May 2014 Distributed Denial of Service (DDoS) attack was launched against Bank of China and the Bank of East Asia at a rate of 7.39 Gbps which denied the online services offered by these banks to legitimate customers. On 31st July 2015 customers of Royal Bank of Scotland, NatWest, and Ulster Bank were not able to access online services for fifty minutes due to DDoS attack (Honan, 2015). Attack on any online services results in massive loss in terms of money and reputation of organizations providing Internet based services (i.e. online services).

Figure 1.1 shows the average loss faced by top most companies of seven major countries due to Internet based attacks in 2015. This loss includes the cost of
investigation, detection, recovery and after effects (Pegram, 2016).

![Figure 1.1: Average Loss Due to Cyber Crime in 2015](image)

Firewall is widely used security enforcement tool by all the organizations however, it does not guarantee the complete security and sophisticated attacker may break this barrier. In this situation to know; how the attacker breached the security and what harm he did to the system Intrusion Detection System (IDS) is used by organizations as a second line of defense. Figure 1.2 shows typical placement of firewall and IDS in the organization (Veteranus, 2013).

![Figure 1.2: Placement of Firewall and IDS](image)

### 1.1 Classification of Intrusions

Any sequence of related actions performed by a malicious adversary that results in the compromise of a target system is called as Intrusion. Intrusive activities are broadly
classified into four categories.

i. Probing: It is used for surveillance of network and machines in the network

ii. Remote to Local (R2L): It is an unauthorized access from a remote machine

iii. User to Root (U2R): It is an unauthorized access to local super-user (root) privileges

iv. Denial-of-Service (DOS) and Distributed Denial of Service (DDoS)

Out of these four intrusion activities DoS and DDoS attacks is the most prevalent threat which either exploits vulnerability in computing and communication resources or floods them in order to make the system unavailable for legitimate users. This results in massive loss of data, resources and money.

1.2 Categories of DoS and DDoS Attacks

DoS and DDoS attacks are broadly classified into four categories based on their working mechanism. These are Connection consumption based attacks, Resource consumption based attacks, Vulnerability exploitation attacks and Configuration modification based attacks.

1.2.1 Connection Consumption based Attacks

TCP is a connection oriented protocol. It establishes a connection between server and client before data exchange.

![Figure 1.3: Working of SYN Flood Attack](image_url)
Any server or machine can accept and serve limited number of connection requests. Attacker establishes a huge number of connections with server or target machine so that legitimate users cannot access the service provided by organization. This type of attacks consumes the Operating System’s kernel resources required for connection establishment. SYN Flood attack is one of the widely used attacks which fall under this category. Figure 1.3 shows the working of SYN Flood attack in which the attacker establishes a huge number of half open TCP connections and exhausts the connection pool. Due to its working mechanism a cluster of servers can be slowed down using slow network connections (Geetha K. and Sreenath N., 2014).

### 1.2.2 Bandwidth Consumption based Attacks

Every network has a limited amount of bandwidth. If the volume of network traffic exceeds the bandwidth limit of the network; then it degrades the response time of servers and machines present on the network.

![Figure 1.4: Bandwidth Consumption Attack using Zombie Machines](image)

By exploiting this principal bandwidth computation based DoS and DDoS attacks are launched. Attacker uses preconfigured handler machines to control huge number of preconfigured zombie machine connected to Internet to create a huge flood as shown
in Figure 1.4. UDP flood is widely used attack of this category (Li et al., 2008).

### 1.2.3 Vulnerability Exploitation Attacks

Attacker identifies and exploits the vulnerabilities present in the target system to either crash or slow down it. These types of attacks are very difficult to identify as attacker completely mimics the behavior of legitimate user. In Slow HTTP Request attacks; parts of a HTTP header are sent to HTTP server at very low rate so that time required serving a single request is increased as shown in Figure 1.5. This in turn consumes the available resources for longer time duration.

![Figure 1.5: Working of Slow HTTP Request Attack](image)

In Slow Read attack, attacker sends a request for large file to the server and then announces a small TCP window size. In response to this server send data at low rate to the client and resources are reserved for long duration. Such attacks can slow down a large cluster of servers with machines having slow or dial-up connection (Park et al., 2014).

### 1.3 Categorization of IDS

Figure 1.6 shows different criterion for categorization of IDS. It can be categorized based on eight criterions as detection method, deployment location, data source used...
for attack detection, response time, architecture, number of detection engines used, volume of data processed and hardware/software implementation.

![Figure 1.6: Categorization of IDS](image)

### 1.3.1 Detection Method Categorization

Anomaly detection based IDS (A-IDS) and Signature (Misuse) detection based IDS (S-IDS) are two approaches of detecting intrusive activities. S-IDS look for patterns or signatures of known attacks. If match is found then it generates the alarm. Signature database of known attacks is specified a priori (Shieh and Gligor, 1997). S-IDS can detect known attacks with high accuracy however; it has few drawbacks as below:

i. S-IDS cannot detect novel attacks.

ii. Signature database needs to be updated periodically

On the other hand, A-IDS attempts to estimate the normal behavior of the system to be protected and generate an alarm when the deviation between a current behavior of system and normal behavior exceeds a predefined threshold. A-IDS can detect novel attacks, however, it generates high false alarm rate as it is difficult to perfectly generate normal user behavior profile. Also A-ID needs to be updated periodically as normal user behavior changes over a period of time.

A-IDS can be implemented using supervised as well as unsupervised learning.
approach. In unsupervised learning IDS builds normal user behavior profile by observing attack free traffic for certain period of time; whereas in supervised learning IDS builds normal user behavior profile by using labeled training dataset. Practically it is difficult to get a live network traffic which is attack free for sufficient duration to train IDS and reflects all possible normal user behaviors.

Combination of A-IDS and S-IDS is used to detect known as well as novel attacks with high precision. S-IDS is used to detect known attacks whereas A-IDS is used to detect novel attack. Traffic or Audit Log which does not match with attacks signature as well as normal user behavior is considered as probable intrusive activity.

1.3.2 Deployment Categorization

Based on deployment location of IDS it is categorized as Host based IDS (H-IDS) and Network based IDS (N-IDS). Host based IDS monitors Operating System’s sequence calls, audit logs and event logs of the machine on which it is installed to detect intrusive activities. H-IDS analyze more detailed information as compared to N-IDS and produces less false alarm rate as compared to it. It can also detect novel attacks with high precision as compared to N-IDS however; it is completely dependent on the system on which it is installed. H-IDS need to be installed and maintained on every machine of the organization which needs to be monitored. In large organization where thousands of machines need to be monitored, this becomes very difficult task.

N-IDS monitor the traffic of entire network with the help of sensors placed at various locations of network to detect intrusive activities. It can analyze less detailed information as compared to H-IDS however; it detects the intrusion before it reaches the victim machine as opposed to H-IDS. H-IDS normally detects successful intrusive attempts whereas N-IDS can detect both successful as well as failed intrusive attempts which make it more reliable solution for monitoring critical systems (Magalhaes, 2003).

N-IDS have following limitations:

i. N-IDS are vulnerable to packet spoofing attacks
ii. N-IDS cannot monitor the encrypted traffic in depth
iii. N-IDS cannot detect intrusions in real time under heavy network traffic
iv. N-IDS cannot detect insider attacks with high precision if sensors are not placed at appropriate locations
Therefore, combination of H-IDS and N-IDS is used to improve the detection precision wherein H-IDS is used to detect intrusions which targets specific host in the network like Back, Slow Read attacks, and N-IDS is used to detect resource and bandwidth consumption attacks like ICMP, UDP flood.

1.3.3 Data Source Categorization

Intrusion detection process uses captured network packets or audit logs present in the system(s) being monitored. Audit logs generated by Operating System and other programs running on top of it are used for intrusion detection process. Network Packet Analysis based IDS (NPA-IDS) analyzes either Header or Payload of captured network packer. Network Packet Header analysis based IDS (NPHA-IDS) are light weight as compared to Network Packet Payload analysis based IDS (NPPA-IDS) and can perform better under heavy network traffic. NPPA-IDS can detect intrusions with higher precision, however, NPPA-IDS cannot effectively analyze network traffic in which payload is encrypted (Al-Jarrah et al., 2016).

Combination of NPHA-IDS and NPPA-IDS is used to detect intrusions with higher precision and less computing power. First NPHA-IDS is used to detect possible intrusive packet and then NPPA-IDS is used for detailed analysis of possible intrusive packets.

1.3.4 Response-time Categorization

Based on response time of intrusion detection process IDS is categorized as Real time IDS (R-IDS) and Batch processing (i.e. Off-line) IDS (B-IDS). R-IDS analyzes the data present in data source within a finite and specified time period, B-IDS process the data in batches. R-IDS can detect intrusions as they occurs however, it cannot perform detailed analysis for intrusion detection process under heavy load which degrades it detection performance. B-IDS gives stable performance under heavy load however, due to its high response time under heavy load it is not used for intrusion detection in critical systems like Banking (Pfleeger and Pfleeger, 2003).

1.3.5 Architecture Categorization

Based on architecture IDS are categorized into following four categories:

i. Monolithic IDS (M-IDS)
ii. Hierarchical IDS (Hi-IDS)  
iii. Agent based IDS (AB-IDS)  
iv. Distributed IDS (D-IDS)  

Monolithic IDS (M-IDS) are conventional IDS which perform intrusion detection activity as a single unit and decision making is done at only one level whereas in Hierarchical IDS (Hi-IDS) intrusion alert or aggregated information generated at lower level of IDS is used by higher level for further decision making. This hierarchical nature of information or alert processing improves the detection precision for complex attacks and reduces the false alarm rate. Agent based IDS (AB-IDS) divides the IDS system into small agents programs which are placed at different locations in the network. These agents capture, processes and passes the information for intrusion detection through a predefined chain. These specialized agents places a very little computing overload on a machine on which they are installed which may provide a light weight IDS solution. Distributed IDS (D-IDS) consist of multiple IDS present over a large network communicating with each other through central IDS or distributed agents. D-IDS gives the global view of intrusive activities in the large network and can effectively identify the intruders and intrusions in real time (Sen et al., 2006).

1.3.6 Number of Detection Engines Categorization  
Conventional IDS uses single Intrusion Detection Engine (IDE) to detect intrusive activities. If training set for IDE is balanced then it give high detection precision for all the attacks, however, it is not practically possible to generate balanced training set for intrusion detection process. Conventional single IDE based IDS cannot detect all the attacks with high precision. To solve this problem Ensemble of IDE based IDS (E-IDS) is used. E-IDS uses multiple IDEs simultaneously and then combines their output using majority vote or weighted majority vote approach to detect the intrusion.

1.3.7 Data Size Categorization  
Number of computing machines in an organization increases with a growth of origination, which increases the size of data needs to be analyzed by IDS for intrusion detection. This huge amount of data cannot be processed by conventional IDS servers in real time. To solve this problem Big Data Analysis based IDS (BDA-IDS) are used. If size of data to be analyzed is not big then conventional IDS servers are used.
1.3.8 **Hardware and Software IDS**

IDS can be implemented as Software application as well as Hardware device. Online hardware IDS devices can detect intrusions in real time; however, due to limited memory and computing power they have following lacunas:

i. Under heavy network traffic Online Hardware IDS itself becomes a bottleneck

ii. Hardware IDS need to be upgraded at regular intervals, which is very expensive

On other hand Software IDS requires more time for intrusion detection as compared to Hardware IDS; however, this problem can be solved using Distributed and parallel processing.

1.4 **Motivation**

First Intrusion Detection System was proposed by Denning (1987). She analyzed the System’s audit records using statistical techniques like threshold, standard deviation and multivariate model for Host based Intrusion Detection. Lunt and Jagannathan (1988) have proposed real-time Intrusion-Detection Expert System which learns the normal behavior of user and predicts the attacks. It is based on the assumption that attacks deviates from the normal user behavior.

Meng et al. (2014) have presented a design of signature detection based N-IDS using packet filtering mechanism. In this, payload of incoming packets is analyzed to detect the presence of attack. This process requires huge amount of recourses in terms of memory and processing power to detect attacks in the presence of very heavy network traffic and DOS attacks. Signature based Intrusion Detection Systems can detect known attacks with high accuracy however, lacks in detection of novel attacks.

Meng et al. (2013) have described adaptive character frequency-based exclusive signature matching scheme for design of Signature based Network Intrusion Detection System. This scheme analyzes the network packets payload for signature matching; thus requires huge resources under heavy network traffic.

Paoet al. (2013) have implemented hardware signature detection based N-IDS using memory-based Non-deterministic Finite Automaton (NFA) regular expression match engine. Hardware Intrusion Detection Systems helps to speed up the detection process; and, becomes a bottleneck when, amount of network traffic is increased beyond a limit making the attack more critical.
Zhang et al. (2015) have proposed an Adaptive Stream Projected Outlier deTector (A-SPOT) technique for design of A-IDS. This technique also works effectively for high dimensional dataset. In general, A-IDS can detect novel attacks with high accuracy; however, it generates high false alarm rate. Anomaly detection based IDS perform poor against the attacks which mimics the behavior of normal user (e.g. Back, Slow read, Slow write attacks).

Dangeloa et al. (2015) have presented an A-IDS using uncertainty-managing batch relevance-based approach. Li and feng Xiao (2015) have proposed A-IDS using dual-ant clustering algorithm. Garcia-Teodoro et al. (2015) have proposed application layer A-IDS which automatically generates signature of attacks against HTTP servers. It analyzes the payload of traffic and if detected as anomalous; then payload is used for generating the signatures of attack. Application layer Intrusion Detection Systems requires more processing resources as it has to analyze the complete payload of traffic, also requires other IDS for detecting attacks on lower layers.

Hansen and Salamon (1990) have presented first ensemble of classifiers approach to improve the classification accuracy using majority vote approach. Filippi et al. (1994) argued that, ensemble of classifier approach gives good classification performance even in the presence of imbalanced training dataset.

Zare Moodi et al. (2015) have described design of ensemble of ensemble of one-class classifiers (i.e. ensemble of ensemble of classifiers) to detect known as well as novel attacks. Output of ensemble of classifiers is merged using majority vote approach. In this approach if more than half of the classifiers predicts wrong class then predicted output is also wrong.

Yin et al. (2015) have proposed dynamic creation of ensemble of ensemble of one-class classifiers for designing Intrusion Detection System using weighted majority vote approach. This approach works well for detection of known attacks but performs poor to detect novel attacks which mimics the behavior of normal user or slightly deviates from old attacks. Li et al. (2015) have proposed random ensemble of decision tree approach for deign of IDS. It uses majority vote approach for merging the output of classifiers in ensemble.

Elbasiony et al. (2013) and Kim et al. (2014) have presented design of hybrid Intrusion Detection System using combination of anomaly and signature detection
based IDS. In this approach, if network connection is not classified by S-IDS and detected as outlier by A-IDS, then it is considered as possible attack. This approach performs poor against attacks which mimic the behavior of normal user and detects deviation of normal user from regular profile as possible intrusion.

1.4.1 Limitations of Existing Approaches

Lot of research has been done over two decades for detection of various attacks. However, there is no generalized solution available for detection of attacks as diversity and volume of attacks are increasing day by day. The following section summarizes limitations of the systems suggested by various researchers.

- Signature based IDS is good at detection of known attacks as compared to novel attacks.
- Anomaly based Intrusion Detection System detects novel attacks; however, it generates high false positive rate.
- Hybrid (Anomaly-Signature based) Intrusion Detection System detects known as well as novel attacks; however, deviation from normal user profile is considered as possible intrusion; which is not always true.
- Payload analysis based IDS can detect attacks with high accuracy, however, analysis of entire payload of every packet requires huge memory and processing power. This requirement becomes critical for DoS attacks.
- Application Layer IDS detects attacks against particular service(s) with high accuracy. It requires analysis of entire payload and additional IDS for detecting attacks on lower layers.
- Hardware Intrusion Detection System can provide real time intrusion detection, however, it needs to be upgraded at regular interval, which makes it very expensive solution.
- Majority vote approach for Ensemble of classifiers gives better results for balanced dataset and performs poor against imbalance dataset.
- Weighted vote perform poor if the novel attack behaves like normal user or slightly deviates from old attacks.

1.5 Objectives and Problem Statement

Based on the literature survey and the research carried out by various researchers and the limitations there on, there is still need of IDS system which can address the issues
related to IDS. The main objectives of this research work are summarized as below:

i. To study and implement existing IDS systems

ii. To create real time traffic and develop a new dataset representing attacks which are not present in standard datasets

iii. To design a header analysis based IDS system for detection of Known DoS and DDoS attacks using Adaptive Ensemble of Naive Bayesian Classifiers

iv. To design a header analysis based Hybrid (Anomaly-Signature based) IDS for detection of Known as well as Novel DoS and DDoS attacks using Ensemble of Classifiers

v. To propose a lightweight system to detect Known as well as Novel DoS and DDoS attacks in Real-time using distributed processing.

vi. To propose a system to scale up scale down the required infrastructure using virtual servers for real time Intrusion Detection based on the increase decrease in the volume of network traffic.

**Problem Statement**

The nature of the network traffic is unpredictable. The network traffic datasets are voluminous, complex, heterogeneous, and of varying quality. There is no specific method to predict the intrusion. Different researchers have suggested different techniques to address the problem of intrusion detection.

After thorough understanding of the limitations of the existing systems, following problem has been identified.

“Design an Adaptive Ensemble of Classifiers based Hybrid (Anomaly-Signature detection based) Scalable Network Intrusion Detection System to detect Known and Novel Denial of Service and Distributed Denial of Service attacks in real time”.

**1.6 Major Contributions**

The thesis has described methods and framework for detection of Known and Novel DoS and DDoS attacks. The major contributions of this thesis are:

i. A new dataset is prepared to represent HHTTP Flood, Slow Read and Slow Write attacks which are not covered in standard datasets.

ii. Procedure to create an Adaptive Ensemble of Classifiers using Naive Bayesian classifier to detect Known DoS and DDoS attacks is described.
iii. Combination of Signature detection based IDS and Adaptive Ensemble of Classifiers is proposed to detect Known and Novel DOS and DDoS attacks.

iv. A framework to detection Known and Novel DOS and DDoS attacks in real-time using unutilized CPU cycles of computing machines within the organization is presented.

v. A methodology is proposed to process genuine flood of web user request and detect attacks in real time using virtual servers.

1.7 Organization of Thesis

Chapter 1 outlines the overview of IDS, need of IDS, types of attacks and IDS system, motivation for research work, problem statement, objectives and major contribution of the research work.

Chapter 2 presents the review of contributions made by researchers for design of A-IDS, S-IDS, Anomaly-Signature detection based IDS, H-IDS, N-IDS, Collaborative IDS, D-IDS, Hardware IDS, R-IDS, Big data analysis IDS and Ensemble of Classifiers based IDS.

Chapter 3 describes the analysis and limitations of KDD 99 and CDMC 2012 datasets used by researchers to evaluate the performance of IDS. It also describes the dataset created in laboratory to represent attacks not covered in standard datasets like KDD 99 and CDMC 2012.

Chapter 4 describes the process of creating Adaptive Ensemble of Classifiers for detection of Known DoS and DDoS attacks. It also presents the performance analysis of Adaptive Ensemble of Classifiers based N-IDS using KDD 999, CDMC 2012 and dataset created in laboratory.

Chapter 5 gives the process for combining S-IDS with Adaptive Ensemble of Classifiers based N-IDS to detect Known as well as Novel attacks. The performance of proposed combination is analyzed using KDD 999, CDMC 2012 and dataset created in laboratory.

Chapter 6 presents the framework for detecting DoS and DDoS attacks in real-time using unutilized CPU cycles of computing machines present within the organization. It also describes the experimental environment created in laboratory to evaluate the proposed framework and experimental results.
Chapter 7 describes the process to differentiate genuine flood of web requests from DDoS attacks and handle genuine flood of web requests in real time using virtual servers. It also presents the experimental results and describes the experimental environment created in laboratory using virtual machines to evaluate the proposed system.

Chapter 8 summarizes the major contributions and also presents the overall conclusion of the thesis with suggestions for future work.