Real-Time Data Based Denial-Of-Service Attack Detection Technique Derived From Statistical Traffic Analysis

Ms. Sanjivani Bakshi
Department of Information Technology
R. M. D. Sinhgad School of Engineering
Pune, India
Email: sanjivani.sumant@gmail.com

Ms. Sweta Kale
Department of Information Technology
R. M. D. Sinhgad School of Engineering
Pune, India
Email: sweta.kale@sinhgad.edu

***Abstract*—Denial of Service (DoS) attacks are one type of aggressive and menacing intrusive behaviour to online servers such as Web servers or Internet Banking servers. These different types of DoS attacks severely degrade the availability of a victim, which can be a host, a server, a router, or an entire network. The attacker flood large amount of useless packets because of which victim is forced out of service from a few minutes to even several days. DoS attacks not only consume the system resources but can also choke the network bandwidth. This causes serious damages to the services running on the victim and requirement of effective DoS attacks detection is essential for the protection of online servers. The DoS attack detection work mainly focuses on the development of network-based detection mechanisms. The detection systems based on these mechanisms monitor traffic transmitting over the protected networks. The projected DoS attack detection system monitors the network traffic to extract the features which are directly associated with DoS attacks. The analysis of TCP packet is discussed in this paper. Based on these features, the statistical traffic analysis like Naive Bayes Classification and Multivariate Correlation models consisting of Geometrical Triangular Area measurements are computed for normal profiles. These models are used as reference model to detect any known as well as unknown DoS attacks in the network with enhanced True positive rate using real time Data.**

Keywords- DoS; TCP; Multivariate Correlation models; Triangular Area measurement; True positive rate

I. INTRODUCTION

The Denial-of-Service (DoS) attacks are one type of aggressive attack in which the attacker attacks on the system mostly on the online servers like Web Servers, DNS servers and Internet Banking Servers [1]. Many large organizations suffers a heavy financial loss because of DoS attacks. In this, attacker generate and flood large number of useless packets and forces the victim out of service for a specific period of time. Because of this the server cannot provide the specified services to the clients. The DoS attack consume the system resources or link bandwidth heavily. There should exist a DoS attack detection technique which will protect the servers from such type of attacks

There are two main types of DoS attack detection techniques, Misuse-based detection technique and Anomaly based detection technique. In Misuse based detection technique, attack can be detected based on the existing signatures. The signatures need to be continuously updated by the administrator. But this technique cannot detect the zero-day DoS attack. Also here is a administrative overhead for the regular signature updates [2, 3].

In Anomaly-based DoS attack detection technique, the incoming network traffic is continuously monitored and checked for the specific deviation from the legitimate traffic. In this system, packet attributes are thoroughly checked for the attack. The system discussed here is the Anomaly-based DoS attack detection system in which statistical traffic analysis is conducted like Bayesian classification and Multivariate Correlation Analysis. The geometrical Triangular Area Measurement is also used to form and calculate the Triangular Area Maps between every two features of the packet. First, the Multivariate Correlation Analysis (MCA) is done for the different packet attributes and a symmetric correlation matrix is generated. From this symmetric correlation matrix, Triangular Area Maps (TAM) are derived which can be further used for DoS attack detection.

The IP and TCP packets headers are keenly studied so that more features attributes can be used for DoS attack detection. Correlation Matrix can be generated by using the quantitative attributes of the packet like, Destination Port, Header length, Payload etc. There are certain useful fields like Destination IP address and flags. The Destination IP address is the dotted decimal address and there are total six flags which can contribute in detection system. These attributes can be utilised by using Bayesian Classification technique. In this certain window of packets is observed and probability of attack is calculated based on same source IP address, same Destination IP address and the flag values. The high probability packets are forwarded to the next module for TAM based MCA generation.

Various statistical methods has been proposed for the DoS attack detection system. Some use Bayesian Classification, Multivariate Correlation analysis or Hidden Markov Model. The previous DoS attack detection system which is based on TAM based MCA technique has high False-Positive rate. Also the system was not tested for Real-Time data. Also Triangular Area Maps are developed for each pair of the packet attributes. By introducing the Bayesian classification prior to TAM based MCA, the number of features can be limited for Triangular Area Map generation. With these approaches,

more precise DoS attack detection system can be developed.

Another drawback is the accuracy of the data supplied to the system. The DoS research community depends extensively on standard datasets (KDD dataset and DARPA dataset) for the purpose of learning and analysis, which are better suited to analysis of intrusion attacks.

II. TECHNICAL OVERVIEW

The statistical traffic analysis techniques are used in this Denial-of-Service attack detection system. The Bayesian Classification, Triangular Area Map based Multivariate Correlation analysis are the main techniques. The ideal solution with respect to DoS attack is to differentiate between good and bad packets. Certain characteristics of the packet can be tapped to differentiate the legitimate packet from attack packet. For this, TCP packet header is keenly studied. The main focus is on IP and TCP packets.

A. Transmission Control Protocol (TCP)

The Transmission Control Protocol (TCP) is a transport layer protocol which is widely used by applications which require reliable transmission. This is provided by TCP using the mechanisms for time-out and retransmission of packets in a sliding window based protocol.

The entities in TCP connection are defined by an IP address and a logical port. For example, TCP ensures connection establishment between two entities by a three way handshake between them, the popular Three-way handshake. TCP maintains handshakes for reliability, error correction, congestion and flow control, etc. The TCP packet header contains fields such as flags and sequence numbers which will determine the state of connectivity between entities [3].

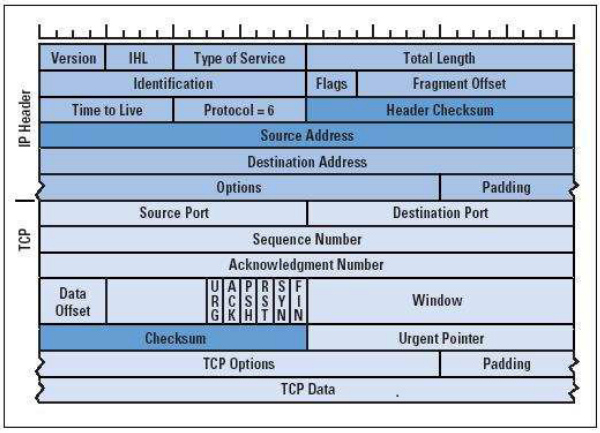

Figure 1. TCP Running over IPV4 Header

The header of the TCP packet over IPv4 is as shown The header of the TCP packet over IPv4 is as shown in Figure 1. Most popular application layer applications run over TCP. This list includes, web services (port 80), mail services (port 25), and file transfer services (port 21). The very fact that TCP accommodates time-outs and retransmissions makes it a hunting ground for malign entities which exploit these properties and cause attacks such as DoS by using these properties [4].

A TCP packet header is used in the detection system. The following parameters were examined in TCP header:

1) TCP flags
2) Payload size
3) Source/Destination Port number and count
4) Source/Destination IP number and count
5) Inter-packet time gaps
6) Total connection time till current packet
7) Number of packets in connection

It is assumed that the overall traffic is streamed for analysis based on destination details like IP address, port number etc. It was experimentally found that TCP flags are best suited among the rest of the attributes.

Hence, in a packet window, modelling is done based on the TCP flags set in the packet. The TCP flags field is a collection of 8-bits, each bit representing one flag. Individual flags or combinations of flags symbolise specific actions in TCP For example, connection establishment, connection closure, requesting data etc.

The different TCP flags are: SYN, ACK, PSH, FIN, URG, RST (standard TCP flags)

B. Windowing

Windowing means splitting input traffic into number of traffic subsets, which fit into logical entities called windows. Window is the number of packets the receiver is willing to receive from a sender at a time. Analysis is done on every single window (traffic subset) to calculate the rate of occurrence of a particular attribute in the window. So better modelling is possible from larger training datasets.

By considering ‘n’ packet in one window, it is possible to find the probability of occurrence of a particular attribute. So by taking into consideration of all the six flag values for the specific Source and Destination IP address, it is possible to calculate the probability of attack at the first stage.

C. Naive Bayes Classiffication

This is the simple probabilistic classifier which is based on ‘Bayes Theorem’. It states that ‘The presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature’. This Bayesian approach is used to find out hidden association. Naive Bayesian classifiers can work with small amounts of training data, and can also accommodate a large number of attributes which makes it a good choice for network modelling for DoS attacks.

A Naive Bayes Classifier is formally defined below:

- Let D be the training set of tuples along with their associated class labels. Each element in the set is an attribute vector $X = \{ x_1, x_2, x_3 , \ldots , x_n \}$.

Let there be m classes, and the set C represent the set of class labels, $C = c_1, c_2, \ldots c_m$.

- Given that an attribute vector X has occurred, classifying X amounts to maximising the value of $P(C_i|X)$, that is, finding out the maximum association of the evidence X with any of the classes in C.

- The value of $P(C_i|X)$ is derived from Bayes theorem

$$P(C_i \mid X) = \frac{P(X \mid Ci) * P(Ci)}{P(X)} \quad (1)$$

where P(Ci) denotes a priori probability.
Naive Bayes Assumption: The attributes $x_1, x_2, x_3, \ldots, x_n$ are independent. Thereby,

$$P(X| C_i) = P(X_1| C_i) * P(X_2| C_i) * \ldots *P(X_n| C_i) \quad (2)$$

- Classifier output $i = \arg \max_{ci} P(C_i \mid X)$.

With this technique, the probability of attack having same_srv_rate , dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate can be found out.

Where X_1 can be the packets having same server as destination, X_2 can be packets having Same Server and hitting to same Port.
With this, it is easy to calculate the above rates which can be further used for MCA and TAM.

D. TAM base MCA

Triangular Area Map generation based Multivariate Correlation Analysis is the main technique used in the system.
It consists of two parts:
1) Multivariate Correlation Analysis
2) Triangular Area Map generation

1)Multivariate Correlation Analysis (MCA) :

Multivariate Correlation Analysis a sophisticated nonpayload-based DoS detection approach. It extracts the correlations between two distinct features within each incoming traffic record. The occurrence of network intrusions cause changes to these correlations so that the changes can be used as indicators to identify the intrusive activities [2, 3].

The following example shows the generated symmeric matrix based on MCA. The input is the NSL-KDD dataset of Normal traffic.

The output of this MCA is a symmetric matrix which has same upper and lower triangles. By selecting upper or lower triangle, Triangular Area Maps can be derived.

For convenience, we have selected Lower Triangular Map Generation Calculation.

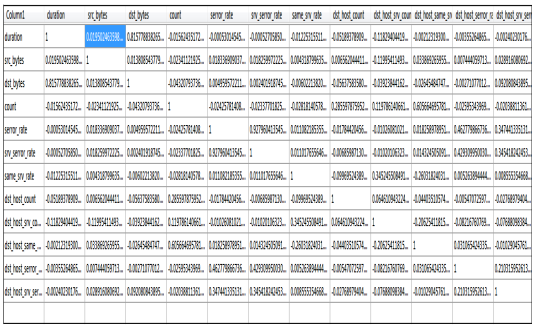

Figure2. Symmetrical Matrix Generated From NSL-KDD Normal traffic Dataset

2) Triangular Area Map (TAM) :

All the extracted correlations are stored in the form of geometrical triangular areas i.e. Triangle Area Maps (TAMs). This is used to differentiate between legitimate and illegitimate traffic records.

From the symmetric matrix, each element in the matrix can be taken which indicates the correlation between the two distinct features. When we know the covariance between two values, Triangular Area Maps can be derived [7].
Assume there is a dataset $X = \{x_1, x_2, \ldots, x_n\}$

- Where $xi = [f_1^i f_2^i \; f_t^i]^T (1 \leq i \leq n)$ represents the i-th n-th dimensional traffic record

- To obtain the triangle formed involving the j-th and p-th features in the i-th observation
- The area of triangle $\Delta f_j^i \, O f_p^i$ is defined by the equation below

$$Tr_{j,p}^i = || (f_j^i, 0) - (0,0) || \; x \; || (0, f_j^i) - (0,0) || / 2 \quad (3)$$

- Repeat the above steps until all possible permutation of any two distinct features in the observation xi are extracted and the corresponding triangle areas are computed, Refer Figure 2 for reference.
Triangle Area Map (TAM) is constructed as

$$TAM^i = [Tr_{j,p}^i]_{txt} \quad (4)$$

- Since TAM^i is a symmetric matrix in which $Tr_{j,p}^i = Tr_{p,j}^i$, the upper or the lower triangle of the map is sufficient to reveal the hidden geometrical correlations

$$TAM_{lower}^t = [Tr_{2,1}^t, Tr_{3,1}^t, \ldots Tr_{t,t-1}^t]^T \quad (5)$$

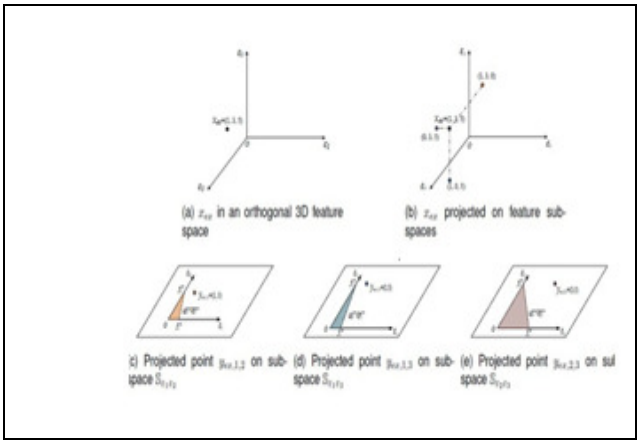

Figure 3. Triangular Area Map Generation

From the Symmetrical Matrix generated from Normal Traffic of NSL-KDD dataset, TAMs Lower are calculated for every attribute. Following figure shows the TAM generated for attribute from Lower Triangle of Covariance Matrix.

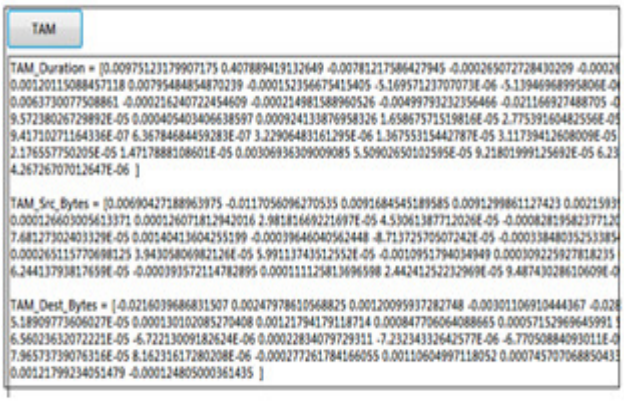

Figure 4. TAM form Lower MCA Matrix

3) Mahalanobis Distance (MD) :

Mahalanobis Distance is the statistical Distance between two points by considering the correlation between the variables [8].

Mahalanobis Distance (MD) is calculated between each TAM and Mean of TAM.

$$MD^{normal} = MD\ (TAM_{lower}^{normal}\ ,\ \overline{TAM_{lower}^{normal}}) \qquad (6)$$

4) Normal Profile Generation :

All the above three methods are used to generate Normal Profile for training data. Normal profile is further stored and used to verified against attack traffic to detect attack.

Normal Profile is validated against each and every observed Mahalanobis Distance. If the observed Mahalanobis Distance is within the range of threshold of Normal Profiles 'μ' and 'σ' then the traffic can considered as 'Normal Traffic', otherwise considered as an 'Attack'

Following is the Normal Profile Generation Algorithm.

1. For i = 1 to g do,
2. $MD^{normal,i} = MD(\ TAM_{lower}^{normal,i}\ ,\ \overline{TAM_{lower}^{normal}})$
3. End For
4. $\mu = \frac{1}{g} \sum MD^{normal,i}$
5. $\sigma^2 = \frac{1}{g-1} \sum MD^{normal,i} - \mu$
6. $Pro = (N\ (\mu, \sigma^2),\ \overline{TAM_{lower}^{normal}})$

Algorithm 1. Algorithm for Normal Profile Generation

III. SYSTEM ARCHITECTURE

In the following section the proposed system, DoS attack detection system architecture is discussed. The complete detection mechanism involves three phases.
The detection mechanism involves four phases.

In phase one basic attributes are selected from ingress network traffic which is designed for the internal servers [6]. Packet Sniffer is designed to capture the inbound traffic. TCP header contains 32 features. The required features are selected. The destination network is monitored and analysed, so that the overhead of the detection is reduced.

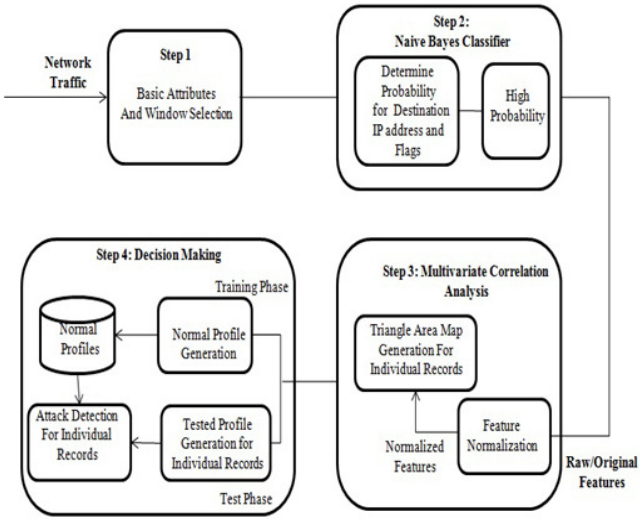

Figure 5. Framework of the proposed denial-of-service attack detection system.

In the second phase, the Naive Bayes Classifier which is a simple probabilistic Classifier based on applying Bayes theorem with Naive independence assumptions. A Naive Bayes Classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. The probability of every packet with the presence of the features like Same Source IP address, Same Destination IP address and TCP header flags is

determined. For example, the traffic rate for a particular Destination addresses from a single or multiple sources on a particular port. When observing flags, for a particular source and destination, numbers of SYN packets are more than number of FIN packets, it can be considered as 'SYN Flood' Attack. If this probability is higher than the packet can be assumed as a attack packet and send a window to TAM based MCA system for further analysis.

In the third phase the multivariate correlation analysis (MCA) is implemented. The quantitative features like destination port, payload, header length etc. are used to generate symmetric covariance matrix. The upper and lower triangle of the symmetric matrix are same so each element in the matrix. Each element in matrix which correlates the two different features of the single observation. The triangle area map (TAM) is generated which is used to extract the correlation between two distinct features within the record which is taken from the first phase. All the triangle area correlations stored in triangle area maps (TAMs) for legitimate traffic records. This provides us with better information to sort out the legitimate and illegitimate traffic records [2].

In phase four, the decision making is done using the anomaly based detection system. The test traffic Mahalanobis Distance is calculated and verified against threshold. If the Mahalanobis Distance is within the threshold limits, then it is considered as a 'Normal Packet' or it can be considered as an 'Attack' [2, 3].

A. Attack Detection

This is a Threshold-based Anomaly detector where 'Normal Profile' is generated by using legitimate traffic records and utilized this for future comparisons with new incoming investigated traffic records. The threshold range is selected and if the incoming traffic record's Mahalanobis Distance is within the Threshold range, then it is treated as 'Normal Traffic Packet' or detected as an 'Attack'.

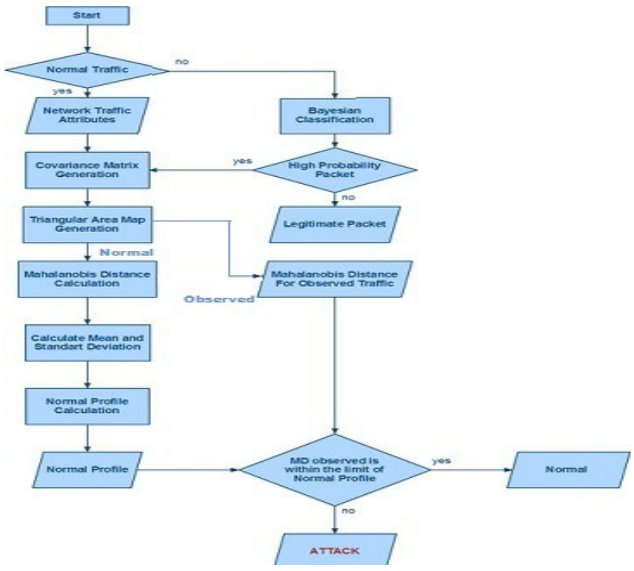

Figure 4. Flowchart of Normal Profile Generation and Attack Detection

So there are three steps for attack detection.

1) Normal profile generation :

Normal Profile is generated based on the Normal Distribution of Mean and standard deviation of Mahalanobis Distance (MD) related to Normal or legitimate traffic, Triangular Area Map of Normal traffic and Covariance matrix of Normal traffic [2, 3].

2) Threshold Selection :

Threshold value is selected using Mean and standard deviation of Mahalanobis Distance related to Normal or legitimate traffic and the value of 'α' [2].

$$\text{Threshold} = \mu + \sigma * \alpha \quad (7)$$

where the value of 'α' ranges from 1 to 3 for Normal Distribution.

3) Attack detection :

To detect DoS attacks, the lower triangle ($TAM_{lower}^{observed}$) of the TAM of an observed record needs to be generated using the proposed Triangle-Area-based MCA approach. Then, the MD between the $TAM_{lower}^{observed}$ and the $\overline{TAM_{lower}^{normal}}$ is calculated $MD^{observed}$ and it is checked against Threshold. If threshold value is more than the predefined range then the packet is considered as 'Attack' [2, 3].

Requirement - $TAM_{lower}^{observed}$, Normal Profile $\overline{TAM_{lower}^{normal}}$

1. Generate $TAM_{lower}^{observed}$ for the observed traffic record.
2. Calculate

   $MD^{observed} = MD\ (TAM_{lower}^{observed}\ ,\ \overline{TAM_{lower}^{normal}})$

3. If $(\mu - \sigma * \alpha) \leq MD^{observed} \leq (\mu + \sigma * \alpha)$
4. Then return **Normal.**
5. Else
6. Return **Attack**
7 End if

Algorithm2. Algorithm for attack detection based on Mahalanobis distance.

IV. PRACTICAL DESIGN CONSIDERATION

When it comes to the testing of the system, there are various challenges. The challenges can be with the availability and authenticity of training data.

1) Training Data Availability :

For the statistical model, large volume of data is required for accurate analysis. But it is not practically possible. So system should able to work with small amount of data.

2) Training Data Authenticity :

It is required that training data should be completely normal. But it can contain abnormal data also. For example, the packet which has both 'SYN' and 'FIN' flag set is considered as abnormal packet.

The evaluation of DoS attack detection system is first conducted using KDD Cup99 data set. KDD Cup99 data set is widely used and readily available dataset. The evaluation of six different types of DoS attacks (Teardrop, Smurf, Pod, Neptune, Land and Back attacks) is available [2, 5].

There are certain drawbacks in this dataset, it contains large number of redundant records. To overcome this, NSL-KDD dataset can be used [5].

V. CONCLUSION

This MCA-based DoS attack detection system which is powered by the triangle-area-based MCA technique and Bayesian Classification technique for anomaly-based detection. The former technique extracts the geometrical correlations hidden in individual pairs of two distinct features within each network traffic record, and offers more accurate characterization for network traffic behaviours. In this technique, the packets entering in TAM based MCA module are reduced due to Bayesian Classification applied in the first filter. So TAMs are not required to generate for each and every attribute. This technique facilitates the system to be able to distinguish both known and unknown DoS attacks from inbound network traffic. Evaluation has been conducted using NSL-KDD data set to verify the effectiveness and performance of the proposed DoS attack detection system.

As a future work, this system can be tested for Real-Time data. The main aim is to enhance the Detection Rate and improve the False-Positive Rate.

ACKNOWLEDGMENT

It is my privilege to express my sincerest regards to my Project guide, Ms. Sweta Kale and Head of Department, Ms. D. T. Kurian, for their valuable inputs, guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of my work. This work is based on the research efforts of Zhiyuan Tan, Aruna Jamdagni, Xiangjian He, Senior Member, IEEE, Priyadarsi Nanda, Member, IEEE, and Ren Ping Liu, Member, IEEE. I am very much indebted to them for their deep insight and inspiring work which is a boost for future researchers. Finally I would like to pay my respect and love to all my family members and especially my daughter for their love and encouragement throughout my career.

REFERENCES

[1] Sanjivani Sumant, Sweta Kale, "Overview of Denial-Of-Service Attack and statistical detection Techniques", IJERT Volume. 3 , Issue. 11 , November - 2014.

[2] Zhiyuan Tan; Jamdagni, A.; Xiangjian He; Nanda, P.; Ren Ping Liu, "A System for Denial-of-Service Attack Detection Based on Multivariate Correlation Analysis," Parallel and Distributed Systems, IEEE Transactions on , vol.25, no.2, pp.447,456, Feb. 2014.

[3] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R.P. Liu, "Triangle-Area-Based Multivariate Correlation Analysis for Effective Denial-of-Service Attack Detection", Proc. IEEE 11th Intl Conf. Trust, Security and Privacy in Computing and Comm., pp. 33-40, 2012.

[4] R Vijayasarathy, A Systems Approach to Network Modeling for DDoS Attack Detection using Naive Bayes Classifier, Thesis for the degree of Master of Science, IIT, Madras, February 2012.

[5] M. Tavallaee, E. Bagheri, L. Wei, and A.A. Ghorbani, "A Detailed Analysis of the KDD Cup 99 Data Set", Proc. IEEE Second Intl Conf. Computational Intelligence for Security and Defense Applications,pp. 1-6, 2009.

[6] S.J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P.K. Chan, "Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project", Proc. DARPA Information Survivability Conf. and Exposition (DISCEX 00), vol. 2, pp. 130-144, 2000.

[7] Author Zhiyuan Tan, "Generation Of Network Behaviour Descriptor Using MCA Based On TAM", Retrieved from http : \\www:kaspersky:com=images=Zhiyuan Tan:pdf

[8]"M-Distance", Retrieved from http:\\classifion:sicyon:com=References=M - distance:pdf