A Survey of Anomaly Detection Techniques and Hidden Markov Model

Hemlata Sukhwani
M. Tech. Scholar (CSE)
Oriental Institute of Science and Technology, Bhopal

Vikas Sharma
Assistant Professor (CSE)
Oriental Institute of Science and Technology, Bhopal

Sanjay Sharma
Assistant Professor (CSE)
Oriental Institute of Science and Technology, Bhopal

ABSTRACT
An Intrusion detection System is software that is used for the malicious activities performed in the network whether in wired or in wireless. Although there are various techniques implemented for the detection of intrusions but still various techniques are yet to be implemented for the accurate detection of intrusion such that the false positive rate can be minimized. Hidden Markov model is a technique which consists of number of states having initial transition of data and at each transition from one state to another a probability is calculated, this technique can be considered for the detection of intrusions. Here in this paper a complete survey of all the technique implemented for the intrusion detection and their various advantages and disadvantages are discussed such that a new technique can be implemented in future.

Keywords
IDS, Hidden Markov Model, Malicious Activity, Behavioral Distance

1. INTRODUCTION
Protecting networks from computer security attacks is a vital apprehension of computer security. Inclusive compilation and truthful explanation of traffic information are core problems in network traffic anomaly detection. As network traffic may lead to variety of information exchange and sensitive data transfer. Although it is also well known that the dependency of network are also emerging rapidly. Due to this the network condition are very crucial now a days and it will become more complicated in forthcoming time. This traffic may lead to massive damage of network system and its related resources. To analyze network behaviour is comes under Anomaly detection. Many host-based anomaly detection systems have been proposed to detect server compromises to detect intrusions by monitoring the execution of a program to see if its behavior conforms to a model that describes its normal behavior [1].

Nowadays, cyber-terrorism is a potential threat to organizations and countries that have become more dependent on cyber-space Securing cyber-space is a challenging task which requires innovative solutions to deal with cyber-terrorism in all its manifestations and forms. One of the appearances of cyber-terrorism is illegal intrusion into the computer resources of a society. This illegitimate access has the purpose of extracting, modifying or damaging susceptible information. Detecting this threat and responding consequently are the main tasks of intrusion detection systems. There are two main approaches to developing intrusion detection systems: misuse detection and anomaly detection [2]. The misuse detection approach uses patterns (called signatures) to detect the presence of known attacks. A signature can be, for instance, a pattern of behavior, a block of code or a sequence of system calls. The anomaly detection approach builds a model of normal behavior of the system. Any system behavior that does not match this model reported as an anomaly.

Regardless of the approach used, the intrusion detection problem (IDP) has been formulated to classify system behavior patterns into two categories: normal and abnormal. But, is the IDP well formulated? Everyday around the word, information about computer system vulnerabilities is released, automatic tools that exploit such vulnerabilities are developed, and fresh kinds of intrusions or attacks are created. Moreover, computer systems are continually upgraded, new user accounts are opened whereas others are removed or disabled. The dynamic behavior of computer systems does not allow a precise definition of normalcy. Each approach has merits and demerits. Even though a misuse detection scheme is effective and efficient in detecting identified attacks, it infrequently detects new attacks. Alternatively, an anomaly detection approach is very good in detecting unknown attacks; but, it may produce a high number of false alarms because it can report unknown normal behavior as abnormal. An ultimate intrusion detection system will combine the advantages of each approach to generate a high detection rate while maintaining a low number of false alarms.

Deviation from the behavior prescribed by a program is feature of, e.g., code-injection attacks utilizing buffer overflow or format-string vulnerabilities, and so should be investigated. A central research challenge is constructing the model to which the process behavior is compared. This is especially challenging in light of mimicry attacks [3], [4] on virtually every such models, in which an opponent injects code that executes its attacks using behaviors that the model does not distinguish from normal. Assuming their diversity renders these processes vulnerable only to unusual exploits, a victorious attack on one of them should induce a detectable increase in the "distance" between the behaviors of the two processes. In principle, this would make mimicry substantially more difficult, since to avoid detection, the behavior of the compromised process must be close to the simultaneous behavior of the uncompromised one [2].

As explained earlier, detectors need models or rules for detecting intrusions. These models can be built off-line on the basis of earlier network traffic data gathered by agents. Once the model has been built, the task of detecting and stopping intrusions can be performed online. One of the weaknesses of this approach is that it is not adaptive. This is because small changes in traffic affect the model globally. Some approaches

to anomaly detection perform the model construction and anomaly detection simultaneously on-line. In some of these approaches clustering has been used. One of the advantages of online modeling is that it is less time consuming because it does not require a separate training phase. Furthermore, the model reflects the current nature of network traffic. The problem with this approach is that it can lead to inaccurate models. This happens because this approach fails to detect attacks performed systematically over a long period of time. These types of attacks can only be detected by analyzing network traffic gathered over a long period of time. The clusters obtained by clustering network traffic data off-line can be used for either anomaly detection or misuse detection. For anomaly detection, it is the clusters formed by the normal data that are relevant for model construction. For misuse detection, it is the different attack clusters that are used for model construction.

2. RELATED WORK

Debin Gao and Michael K. Reiter had proposed a new and efficient technique of detecting anomalies in the network or in the packets using behavioral distance based Hidden Markov Model. In the paper a behavioral distance between two states in the network is computed and if it is greater than the threshold value than the detection is performed. This new approach based on HMM detects intrusions with substantially greater accuracy than existing schemes [1].

Debin Gao, Michael K. Reiter, and Dawn Song also given the detection of network anomalies by the use of hidden markov models based on the behavioral distances. The behavioral distance between two processes is a measure of the deviation of their behaviors. Behavioral distance has been offered for detecting the compromise of a process, by calculating its behavioral distance from another process executed on the same input. Given that the two processes are miscellaneous and so unlikely to fall prey to the identical attacks, a rise in behavioral distance. In the technique mentioned in [2] implemented a new way of detecting anomalous behavior by including behavioral distance in hidden markov model. The proposed technique implemented detects intrusions with substantially greater accuracy and with performance overhead comparable to that of prior proposals [2].

Kymie Tan and John McHugh proposed a new way of detecting normal or abnormal behavior of the network. The technique implemented here is used for the prediction of anomalous behavior of the packets. The technique uses the concept of information hiding where the attacks possible should be made hide from the normal behavior [3].

David Wagner and Paolo Soto has proposed and implemented mimicry attacks that are based on host-based intrusion detection systems. Although there are various host based intrusion detection techniques that provides security from various attacks in the network. First, they introduce the notion of a mimicry attack that permits a sophisticated attacker to cloak their intrusion to avoid detection by the IDS. Then, they develop a theoretical framework for evaluating the security of IDS against mimicry attacks. They show how to break the security of one published IDS with these methods, and then experimentally confirm the power of mimicry attacks by giving a worked example of an attack on a concrete IDS implementation [4].

Andreas Wespi, Marc Dacier, and Herv´e Debar in 2000 proposed a new methodology for the intrusion detection [6] using the concept of variable length of audit trail patterns. A novel technique is used to build a table of variable-length patterns. This method is based on Teiresias, an algorithm initially developed for discovering rigid patterns in unaligned biological sequences. Here evaluate the quality of technique in a test bed environment, and compare it with the intrusion-detection system proposed by Forrest et al. [7].

R. Sekar and M. Bendre implemented an efficient and fast automaton based method for anomaly detection. The approach builds a compact FSA in a fully automatic and proficient approach, without necessitate access to source code for programs. The space requirements for the FSA are low—of the order of a few kilobytes for distinctive programs. The FSA utilizes only a constant time per system call during the learning as well as detection period. This feature guides to low overheads for intrusion detection. Dissimilar many of the existing techniques, this FSA-technique can capture both short term and long term temporal relationships among system calls, and therefore execute more accurate detection [8].

Debin Gao and Michael K. Reiter has proposed anomaly detection using Gray box of the execution of Graphs. In this paper has introduced a new model of system call behavior, called an execution graph. The execution graph is the first such model that both requires no static analysis of the program source or binary, and be conventional to the control flow graph of the program. After exercised as the model in an anomaly detection system monitoring system calls, it offers two strong properties: (i) it accepts only system call sequences that are consistent with the control flow graph of the program; (ii) it is maximal given a set of training data, importance that any extensions to the implementation graph could permit some intrusions to go undetected [9].

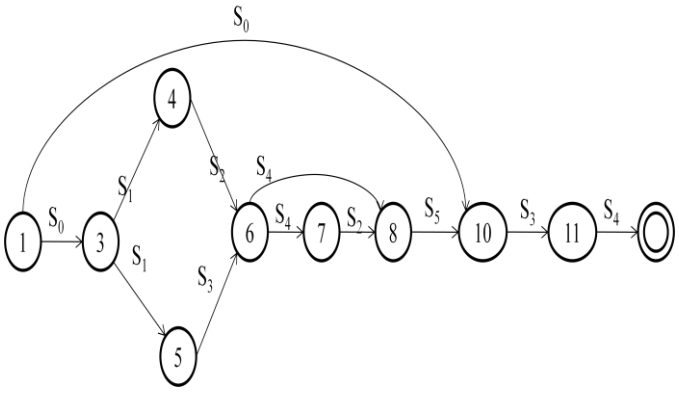

Figure 1: Automaton learning's by the algorithm [8].

Henry Hanping Feng, Oleg M. Kolesnikov implemented stack based call information for the detection of anomalies in the network. The basic idea is to extract return addresses from the call stack, and produce abstract execution path among two program execution points. Experiments show that the method can detect some attacks that cannot be detected by additional methods, while its convergence and false positive performance is comparable to or better than the other approaches [10].

Wassim El-Hajj et al [12] offered On Detecting Port

Scanning using Fuzzy Based Intrusion Detection System. They update Snort by integrating it with a customized Fuzzy Logic controller. They call the new system "Fuzzy Based Snort (FB-Snort)". The aim behind this merge is to better detect port scanning and to reduce the false negative and false positive alarms. Here alternative for using Fuzzy Logic was

based on two main reasons: (i) No clear boundaries exist between normal and uncharacteristic events, (ii) fuzzy logic rules help in smoothing the abrupt separation of normality and abnormality or anomaly. This strategy starts by finding the normal traffic from abnormal traffic using Snort. Then pass some chosen parameters (section IV) to the Fuzzy Logic controller to get one exclusive parameter. This parameter chooses whether an attack presents or not. According to result, FB-Snort reduces the false positive and the false negative alarms [12].

Leonardo et al [13] proposed Inter-Domain Stealthy Port Scan Detection through Complex Event Processing. They address the general problem of detecting inter-domain port scanning activities originating from single sources. The detection is carried out in a cooperative fashion by correlating network traffic data coming from geographically distributed enterprise nodes. To this end, they have designed an IDS architecture that can (i) easily deal with the evolution of the monitored system. In a large scale enterprise network, additional domains may be added; the architecture is able to extend its deployment in order to monitor these new parts of the system; and (ii) devise an easy way for updating the detection logic which also has low cost of ownership and high flexibility. As new security mechanisms are put in practice, malicious attackers enable new ways for circumventing them. In order to cope with this evolving scenario it is required that the architecture can promptly deploy new techniques for facing these brand-new threats. The proposed solution employs so-called Gateway components, i.e., software sensors located at each enterprise geographically dispersed domain that is to be monitored. Gateways send captured network traffic data to a Complex Event Processing (CEP) engine. The engine is responsible for correlating the data and thus discovers spatial and/or temporal relationships among apparently uncorrelated data that would have been undetected by in-house IDSs [13].

Xiaobin Tan and Hongsheng has proposed anomaly detection technique using hidden semi-markov models. Hidden semi Markov model (HSMM) is introduced into intrusion detection. Hidden Markov model (HMM) has been applied in intrusion detection systems many years, but it has a foremost weakness: the inbuilt duration probability density of a state in HMM is exponential that may be unsuitable for the modeling of audit data of computer systems. They can handle this problem well by developing an HSMM for perfect normal processes of computer systems. Based on HSMM, an algorithm of anomaly detection is presented in this that calculates the distance among the processes monitored by intrusion detection system and the perfect normal processes. In this algorithm they employ the average information entropy (AIE) of fixed-length observed sequence as the anomaly detection metric based on maximum entropy principle (MEP). To improve accuracy, the segmental K-means algorithm is applied as training algorithm for the HSMM. By evaluating the correct rate with the experimental results of existing research, it shows that this method can perform a more accurate detection [14].

Zhu Lin and Zhu- Can- Shi proposed Research into the Network Security Model Blended of Data Stream Mining and Intrusion Detection System [15]. They present a network security model built on the integration of data stream mining and intrusion detection system. Data collection module is mainly responsible for the lossless capture of network packets, and meanwhile in charge of some simple packet inspection as well as filtration of error messages. The data the data collection module submits to the pretreatment layer are basically the original data packets. Data module includes training data collation and isolated point exclusion. It is mainly used in the course of data collection, during which the information contains some common processing operations, but with intrusion information excluded as a premise. As an important technique of data mining, cluster analysis enjoys a broad application. Cluster analysis divides the concentrated data objects into a number of groups, making the similarity of data in each group as high as possible while making the similarity among groups as low as possible. Their Data mining algorithm is used to extract security-related attributes of systematic characteristics, and then to generate classification models of security incidents in accordance with these attributes so as to effectively reduce the uncertainty caused by human factors in analyzing intrusion patterns and extracting characteristics, thus achieving an automated screening of security incidents [15].

Kyung Choi et al [16] proposed the data attributes for the SYN flood attack and the buffer overflow attack, and the recognition procedure to find proficient data mining methods for those attacks. According to result obtained, in case of SYN flood attack, a total of 64 mining algorithms are executed with the selected key attributes. Thirty algorithms show 99.833% detection rate and 54 algorithms show more than 90% exposure rate. Next, the decision tree methods with the best detection rate are preferred, and those algorithms show the decision tree as an effect. Sixty four algorithms are achieved with selected key attributes using result of decision tree. Three algorithms show a 100% detection rate, 29 algorithms show 99.833% [16], [17], [18], [19].

3. HIDDEN MARKOV MODEL

A Hidden Markov Model contains five tupples:

N – is the number of states in the model Q {Q1, Q2, Q3,….}.

M – is the number of observation symbols V {V1, V2, V3…..}.

A – State transition Probabilities.

B – is the distribution of each of the states.

Π – is the initial state distribution.

1. The initial transition probability from one state Q1 to another state Q2 at a particular instance of time t+1 depends on the state at time t according to the assumption of markov i.e.

$$a_{ij} = p(q_{t+1} = s_j | q_t = s_i)$$

2. The probabilities of the transition of the states is independent of the actual time where the transition takes place according to the assumption of stationary i.e.

$$p(q_{t+1} = s_j | q_{t1} = s_i) = p(q_{t2+1} = s_j | q_{t2})$$

3. Lets ‘n’ is the number of packets ‘pkt’ send at a particular transition at a particular instance of time.

4. Calculate each step of the transition the state which is most probable $\hat{q}_i$, $1 \le i \le T$ for the observation z_i , $1 \le i \le T$, probability of state transition δt can be computed using viterbi algorithm.

5. After each step of the transition calculate the general probability of the packet to be transmitted at each step Q.

6. The average probability can be computed using

$$\delta_{avg} = \sum_{k=1}^{T} \delta_k^{(i)} \Big/ T$$

7. The condition is checked i.e. if the average probability is less than the threshold value then the intrusion is detected in the packet.

$$\delta_{avg} < L\ (initial\ threshold\ value)$$

4. PROPOSED METHODOLOGY

The various parameters used in HMM such as:

1. 'N' represents the no. of states in the model.
2. There are various individual states in the model as S={S1,S2,S3…..Sn}.
3. State a particular instant of time't' is q_t.
4. 'M' is the number of distinct observation symbols per state; these observation symbols correspond to the physical output of the system being modeled.
5. Various individual symbols are denoted as V={V1,V2,…..Vm}.
6. 'A' is represented as the probability of distribution during the transition of states.
7. 'B' is represented as probability of distribution of the observational symbols. It can be represented as:

$$B = b_{jk}$$

8. $'\pi'$ can be represented as probability distribution of initial state of transition and can be represented as :

$$\pi = \{\pi_i\}n$$

9. The various sequences of the state's OO=OO1, OO2, OO3….OOT', known as indirect observation of the hidden states and T' can be represented as the number of total number of observations taken.

The proposed methodology based on hidden markov model using behavioral distance contains the following parameters as

$$\gamma = \{A, B, \pi\}$$

Here N can be represented as the hidden states let us take it as 5. Here M can be represented as observations to be taken. if the hidden states are taken as SS1, SS2, and SS3 and for SS4, and SS5 the value is 2.

The probability distribution of the state transition is

$$A = \{a_{ij}\}$$

Where

$a_{ij} = P\left[q_{t+1} = S_j \middle| q_t = S_i\right]$, 1<=j<=5 and 1<=i<=5

The probability distribution of observation symbol in state j, $B = \{b_j(k)\}$

Where

$b_j(k) = P[V_k\ at\ t | q_t = S_i]$, 1<=j<=5, 1<=k<=6 if j=1,2 or 3 else 1<=k<=2

The (figure 2) shows the transition of different states, where the links connected represents transition probability of the states.

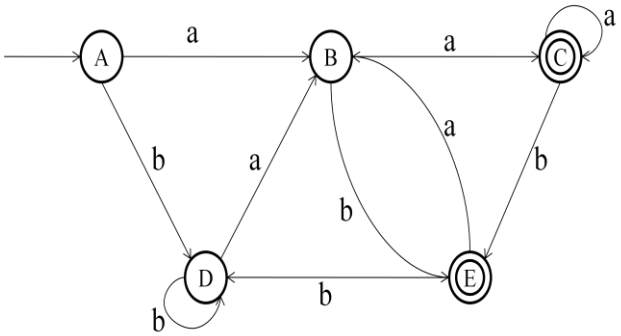

Figure 2: Transitions Probability of States.

Welch proposed a hidden markov model which contains and starts with the estimate initial state and likelihood value is used to find the local maxima value.

Table 1. State Distribution Value

States	Initial State Distribution value (π_i)
1	0.000482
2	0.249375
3	0.079432
4	0.36835
5	0.246293

Table 2. Average Probability Distribution

States	1	2	3	4	5
1	0.1658	0.1356	0.2659	0.1853	0.157
2	0.0465	0.2850	0.1759	0.2845	0.1745
3	0.0275	0.1385	0.2759	0.1773	0.2942
4	0.5843	0.1853	0.0649	0.0023	0.0046
5	0.1395	0.1548	0.1844	0.2753	0.2352

After parameter estimation step, Forward Procedure is applied for training HMM.

The forward variable:=P (OO1,OO2,OO3,OO4,OO5, qt = Si $|\lambda$) (1)

The forward variable 'P' indicates the probability of the partial observation sequence OO1, OO2, OO3, OO4, and OO5, and the state Si at time t, given the model. λ

Observation sequences OO1, OO2, OO3, OO4, and OO5 represent the discrete observation symbol number of the state's SS1, SS2, SS3, SS4, and SS5 respectively. Thus, in this case values of OO1, OO2, OO3 ranges from 1 to 6 and for OO4 and OO5 it is either 1 or 2. Steps involved in the Forward Procedure are described using equations (2), (3), and (4):

Initialization of the forward variable value

$$\alpha \text{ t(i)} = \pi \text{ i} * \text{bi (OO1)} \quad (2)$$

where 1 <= i <= 5

Induction step of the Forward Procedure

$$\alpha_{(t+1)}(j) = \left[\sum_{i=1}^{5} \alpha_t(i) * a_{ij} \right] * b_j(O_{t+1}) \quad (3)$$

Where 1<=t<=T-1 and 1<=j<=5.

Termination step of the Forward Procedure

$$P(O|\lambda) = \sum_{i=1}^{5} \alpha_t(i) \quad (4)$$

Thus, P (O $|\lambda$) is the sum of all the π t (i) values.

Now the probability distribution of each of the state is computed and then fused the probability distribution from each of the state to get an average probability distribution, which is then compared with each of the individual state and then according to the probability distribution the anomalies are classified as normal, medium or high type of anomaly.

5. CONCLUSION

Here in this paper a survey various techniques implemented for the detection of anomalies and intrusions in the network are discussed. Also the Hidden Markov model is used for the detection of anomalies and intrusion is discussed. The proposed methodology implemented here is an efficient technique for the detection of network anomalies and intrusions.

Although the methodology proposed here provides efficient detection of anomalies but further enhancements can be done for the classification of anomalies and improving the detection ratio will also be done in the future.

6. ACKNOWLEDGMENT

I would like to express my deep gratitude to Asst. Prof. Mr. Vikas Sharma and Asst. Prof. Mr. Sanjay Sharma, Computer

Science Department, Oriental Institute of Science and Technology, Bhopal, my research supervisors, for their guidance, enthusiastic encouragement and useful critiques of this research work

7. REFERENCES

[1] Gao, Debin, Michael K. Reiter, and Dawn Song. "Beyond output voting: Detecting compromised replicas using HMM-based behavioral distance", IEEE Transactions on Dependable and Secure Computing, vol. 6, no. 2, pp. 96-110, 2009.

[2] Gao, Debin, Michael K. Reiter, and Dawn Song "Behavioral distance measurement using hidden markov models", In Recent Advances in Intrusion Detection, pp. 19-40, Springer Berlin Heidelberg, 2006.

[3] Tan, Kymie, John McHugh, and Kevin Killourhy "Hiding intrusions: From the abnormal to the normal and beyond", In Information Hiding, pp. 1-17, Springer Berlin Heidelberg, 2003.

[4] Wagner, David, and Paolo Soto "Mimicry attacks on host-based intrusion detection systems", In Proceedings of the 9th ACM Conference on Computer and Communications Security, pp. 255-264, 2002.

[5] Thottan M., Ji C. "Anomaly Detection in IP Networks", IEEE Transaction on Signal Processing, Special Issue of Signal Processing in Networking, Vol. 51, No. 8, pp. 2191-2204, 2003.

[6] Wespi, Andreas, Marc Dacier, and Hervé Debar "Intrusion detection using variable-length audit trail patterns", In Recent advances in intrusion detection, pp. 110-129, Springer Berlin Heidelberg, 2000.

[7] Stephanie Forrest, Steven A. Hofmeyr, Anil Somayaji, and Thomas A. Longstaff "A sense of self for Unix processes", In Proceedings of the 1996 IEEE Symposium on Research in Security and Privacy, pp. 120–128, IEEE Computer Society Press, May 1996.

[8] Sekar, R., Mugdha Bendre, Dinakar Dhurjati, and Pradeep Bollineni "A fast automaton-based method for detecting anomalous program behaviors", In Proceedings 2001 IEEE Symposium on Security and Privacy, pp. 144-155, IEEE, 2001.

[9] Gao, Debin, Michael K. Reiter, and Dawn Song "Gray-box extraction of execution graphs for anomaly detection", In Proceedings of the 11th ACM conference on Computer and communications security, pp. 318-329, 2004.

[10] Feng, Henry Hanping, Oleg M. Kolesnikov, Prahlad Fogla, Wenke Lee, and Weibo Gong "Anomaly detection using call stack information", In Proceedings of IEEE Symposium on Security and Privacy, pp. 62-75, 2003.

[11] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, J. Srivastava "A comparative Study of Anomaly Detection Schemes in Network Intrusion Detection", A technical report, 2003.

[12] Wassim El-Hajj, Fadi Aloul, Zouheir Trabelsi and Nazar Zaki "On Detecting Port Scanning using Fuzzy Based Intrusion Detection System", International Wireless Communications and Mobile Computing Conference (IWCMC '08), pp. 105 – 110, 2008.

[13] Leonardo Aniello, Giorgia Lodi and Roberto Baldoni "Inter-Domain Stealthy Port Scan Detection through Complex Event Processing", In Proceedings of the 13th European Workshop on Dependable Computing, pp. 67 - 72, 2011.

[14] Tan, Xiaobin, and Hongsheng Xi "Hidden semi-Markov model for anomaly detection", Applied Mathematics and Computation, vol. 205, no. 2, pp. 562 – 567, Elsevier 2008.

[15] Zhu Lin and Zhu- Can- Shi "Research into the Network Security Model Blended of Data Stream Mining and Intrusion Detection System", in Proceedings of the IEEE 7th International Conference on Computer Science & Education (ICCSE 2012), pp. 498 – 499, 2012.

[16] Kyung Choi, Xinyi Chen, Shi Li, Mihui Kim, Kijoon Chae, and JungChan Na "Intrusion Detection of NSM Based DoS Attacks Using Data Mining in Smart Grid", OPEN ACCESS Energies, vol. 5, pp. 4091-4109, 2012.

[17] Divya Pal Singh, Pankaj Sharma, Ashish Kumar "Detection of Spoofing attacks in Wireless network and their Remedies", International Journal of Research Review in Engineering Science and Technology (IJRREST), Volume 1, Issue1, June 2012.

[18] A. Rahul, S.K.Prashanth, B.Suresh kumar, G.Arun "Detection of Intruders and Flooding In Voip Using IDS, Jacobson Fast and Hellinger Distance Algorithms", IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 2, Issue 2, PP 30-36, July-Aug. 2012.

[19] Emil Kuriakose John and Sumaiya Thaseen "Efficient Defense System For IP Spoofing In Networks", computer Science & Information Technology (CS & IT), pp. 185–193, 2012.