ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 419- 422 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

A Novel classification approach for Firewall Log data Analysis

[1]P.Sowjanya, [2]V.Srikanth
[1]Final M.Tech Student, [2]Assistant Professor
[1,2]Dept of Computer Science and Eng
[1,2]Pydah College of Engineering, Visakhapatnam, AP, India

Abstract: Firewall log data analysis always an interesting research issue in the field of internet traffic analysis , Even though various approaches available for minimizing internet traffic over heads during network traffic , may not be optimal for all possible samples. In this paper we are proposing an efficient classification approach for internet traffic by analyzing the behavior of the nodes for allowing or disconnection of the incoming node by computing the posterior probabilities of the factors with respect to the node.

INTRODUCTION

Various researchers proposed different approaches for classifying the network traffic or identify the anonymous node either by clustering, signature mechanisms and classification mechanisms. In clustering mechanisms we group the similar type of data objects based on the similarity between the data objects, by selecting the initial data points or centroids.

Various network intrusion detection systems proposed by the various researchers by using different approaches like state of are Classification, clustering and statistics based. Some of the tools for analysis of network intrusion detection are Snort, tcpdump, Network flight Recorder and other.

Snort: Snort one of the simple and efficient tool for intrusion detection and prevention and it also an open

Source tool, it detects the attacks like Denial of service attacks and its preprocessor detects the anomalies in the packet header and packet payload feature available in the snort.

Tcpdump: It is some extent similar to the Snort tool in packet sniffing bust user output is not that much better than the snort and packet payload feature not available in this tool.

Network Flight Recorder

It is a much more flexible network analysis tool than the snort in most of the cases. It contains complete feature set of rules than snort but is not quick and as easy as snort Features like Packet filtering,IP fragmentation and TCP stream decoding available and is the first probuch which defeated anti NIDS attacks.

Address Resolution Protocol (ARL) converts Ip-address to MAC address through network layer to data link layer during the transmission of data packet from one host to another host. In this protocol an ARP request the network and receiver responds with the MAC address and sender asks request the receiver for the MAC address then receiver responds with MAC and IP address. It is becoming vulnerable due to stateless and lack of authentication mechanism.

	Snort	Tcp dump
Packet sniffing	Available	Available
Packet payload feature	Available	Not available
GUI	More user friendly than Tcp dump	User friendly
lookup of host names and ports	Not Available	Available

Identifying the unauthorized access, malicious activities or unauthorized user over network is known as intrusion detection, analyzing the malicious Behavior is still an important research issue because we cannot blame anyone/node as intruder without accurate results.

Various researchers proposed Network Intrusion detection and prevention system to identify and prevent the

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 419- 422 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

malicious behavior of the connected node. Most of the traditional approaches of intrusion detection system works based on the signature based mechanisms. In signature based intrusion detection mechanism, signature of the node can be compared and monitor with known signatures or identities in the database for the intruder, it works like an Anti Virus software. The main drawback with signature based approach is, it cannot identify emerged and unknown identity or signature while matching with signatures in the database and one more drawback with maintenance of signatures of known nodes, it obviously gives the opportunity to the intruder to change is way of breaking attacks .

To overcome the drawbacks of the previous approach, researchers concentrated on the traffic classification approaches over the statistical parameters. In this approach we forward the testing sample feature set to the known training dataset to analyze the testing sample behavior. So many approaches proposed for these classifications, every mechanism has their own pros and cons.

Payload-Based Classification

Another approach to classify packets is to analyze the packetpayload or use deep packet inspection (DPI) technology. They classify the packets based on the signature in the packet payload, and it has been routed as the most accurate classification method, with 100% of packets correctly classified if the payload is not encrypted [3]. The signature is unique strings in the payload that distinguish the target packets from other traffic packets. Every protocol has its distinct way of communication that differs from other protocols. There are communication patterns in the payload of the packets. We can set up rules to analyze the packet payload to match those communication patterns in order to classify the application.

For example, according to [3], "MAILFROM","RCPT TO" and "DATA", as in Figure 1, are the commands that appear in the payload of SMTP packets. Therefore, we can create rules to match the plaintext in the packet payload to classify SMTP packets. The problems include: users may encrypt the payload to avoid detection, and some countries forbid doing payload inspection to protect user information privacy. Furthermore, the classifier will experience heavy operational load because it needs to constantly update the application signature to make sure it contains the signature of all the latest applications.

RELATED WORK

Various clustering and classification mechanisms available for classify or analyze the behavior of the nodes in the network traffic. The major drawback with clustering process is the random selection of the centroid it may leads to the local optimal, it means results or clusters depends upon the selection of the centroid.

While not strictly classification, Floyd & Paxson [9] observe that simple (Poisson) models are unable to effectively capture some network characteristics. However, they did and a Poisson process could describe a number of events caused directly by the user; such as telnet packets within rows and connection arrivals for ftp-data. This paper is not the forum for a survey of the entire Machine Learning field. However, our approach may be contrasted with previous work which attempts the classification of network traffic. We present a sample of such papers here. Rough net al. [10] performs classification of traffic rows into a small number of classes suitable for Quality of Service applications. The authors identify the set of useful features that will allow discrimination between classes. Limited by the amount of training data, the authors demonstrate the performance of nearest neighbor, LDA and QDA algorithms using several suitable features. McGregoretal. [11] Seek to identify traffic with similar observable properties and apply an untrained classifier to this problem.

Eavesdropping or sniffing is a process of monitors network traffic which is transmitted over network, attackers uses various sniffing tools to monitor the in and out flow of network traffic. Attacker can analyze the data packets size , source and destination of the packets. Monitor in house staff who are doing unauthorized or unassigned tasks. Provide Strong encryption, decryption algorithms (for confidentiality) and Key Exchange protocols (For secure key generation).It gives unformatted data to the attacker even though it is hacked. Continuous monitoring of network traffic required for both in house and outside staff (Accessing remotely)

PROPOSED SYSTEM

We are proposing an efficient internet traffic classification over log data or training dataset which consists of source ip-address or name, Destination ip-address and port number, type of protocol and number of packets transmitted from source to destination. When a node connects if retrieves the meta data i.e. testing dataset and forwards to the training dataset .both training and testing datasets CAN Be forwarded to Bayesian classifier for analyzing the behavior of the connected node.

We proposed a novel and efficient trust computation mechanism with naive Bayesian classifier by analyzing the

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 419- 422 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

new agent information with existing agent information, by classifying the feature sets or characteristics of the agent. This approach shows optimal results than the traditional trust computation approaches.

In our approach we proposes an efficient classification based approach for analyzing the anonymous users over network traffic and calculates the trust measures based on the training data with the anonymous testing data. Our architecture contributes with the following modules like Analysis agent, Neighborhood node, Classifier and data collection and preprocess as follows

1) Analysis agent –Analysis agent or Home Agent is present in the system and it monitors its own system continuously. If an attacker sends any packet to gather information or broadcast through this system, it calls the classifier construction to find out the attacks. If an attack has been made, it will filter the respective system from the global networks.

2) Neighbouring node - Any system in the network transfer any information to some other system, it broadcast through intermediate system. Before it transfer the message, it send mobile agent to the neighbouring node and gather all the information and it return back to the system and it calls classifier rule to find out the attacks. If there is no suspicious activity, then it will forward the message to neighbouring node.

3) Data collection - Data collection module is included for each anomaly detection subsystem to collect the values of features for corresponding layer in an system. Normal profile is created using the data collected during the normal scenario. Attack data is collected during the attack scenario.

4) Data pre-process - The audit data is collected in a file and it is smoothed so that it can be used for anomaly detection. Data pre-process is a technique to process the information with the test train data. In the entire layer anomaly detection systems, the above mentioned pre-processing technique is used

For the classification process we are using Bayesian classifier for analyzing the neighbor node testing data with the training information. Bayesian classifier is defined by a set C of classes and a set A of attributes. A generic class belonging to C is denoted by c_j and a generic attribute belonging to A as A_i. Consider a database D with a set of attribute values and the class label of the case. The training of the Naïve Bayesian Classifier consists of the estimation of the conditional probability distribution of each attribute, given the class.

In our example we will consider a synthetic dataset which consists of various anonymous and non anonymous users node names, type of protocols and number of packets transmitted and class labels, that is considered as our feature set C ($c_1,c_c,\ldots\ldots c_n$) for training of system and calculates overall probability for positive class and negative class and then calculate the posterior probability with respect to all features ,finally calculate the trust probability.

Algorithm to classify malicious agent

Sample space: set of agent

H= Hypothesis that X is an agent

P(H/X) is our confidence that X is an agent

P(H) is Prior Probability of H, i.e. the probability that any given data sample is an agent regardless of its behavior.

P(H/X) is based on more information, P(H) is independent of X

Estimating probabilities

P(X), P(H), and P(X/H) may be estimated from given data

Bayes Theorem

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)}$$

Steps Involved:

1. Each data sample is of the type

X=(xi) i =1(1)n, where xi is the values of X for attribute Ai

2. Suppose there are m classes Ci, i=1(1)m.

X Î Ciiff

P(Ci|X) > P(Cj|X) for 1£ j £ m, j¹i

i.e BC assigns X to class Ci having highest posterior probability conditioned on X

The class for which P(Ci|X) is maximized is called the maximum posterior hypothesis.

From Bayes Theorem

3. P(X) is constant. Only need be maximized.

 If class prior probabilities not known, then assume all classes to be equally likely

 Otherwise maximize

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 419- 422 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

P(Ci) = Si/S

Problem: computing P(X|Ci) is unfeasible!

4. Naïve assumption: attribute independence

P(X|Ci) = P(x1,…,xn|C) = PP(xk|C)

5. In order to classify an unknown sample X, evaluate for each class Ci. Sample X is assigned to the class Ci iff P(X|Ci)P(Ci) > P(X|Cj) P(Cj) for 1£ j £ m, j¹i

In the above classification algorithm, computes the posterior probabilities of the input samples with respect to the data records in the training dataset over all positive and negative probabilities, analyzes the network traffic with positive and negative probabilities.

FUTURE WORK

Preprocessing is the basic step before analyzing the behaviors of the nodes because most of the intrusion detection detection systems directly or indirectly deals with mining or neural network or other approachesbefore analyzing the testing sample behavior best training sample ,both should be preprocessed. Usually preprocessing includes

- Removal of redundant records from the training and testing datasets
- Feature extraction is one more important factor before applying any classification approach various feature selection approaches available Principle component analysis and DDC Provision for conversion of categorical data to numerical data.

CONCLUSION

We are concluding our research work with efficient classification approach by analyzing the anonymous behaviors of the log data packet analysis with their respective posterior probabilities of the individual attribute and final class labels to compute final probabilities of the connected node.

REFERENCES

1) Internet assigned numbers authority (IANA),http://www.iana.org/assignments/port-number (last accessed October, 2009)

2) A. Madhukar, C. Williamson, A longitudinal study of p2ptraffic classification, in: MASCOTS '06: Proceedings of the14th IEEE International Symposium on Modeling, Analysis,and Simulation, IEEE Computer Society, Washington, DC,USA, 2006, pp. 179–188.doi:http://dx.doi.org/10.1109/MASCOTS.2006.6.

3) J. Klensin, SIMPLE MAIL TRANSFER PROTOCOL, IETF RFC 821, April 2001; http://www.ietf.org/rfc/rfc2821.txt

[4] Bro intrusion detection system - Bro overview, http://broids. org, as of August 14, 2007.

[5] V. Paxson, "Bro*: A system for detecting network intruders in real-time*," Computer Networks, no. 31(23-24), pp. 2435–2463, 1999.

[6] Azzouna, Nadia Ben and Guillemin, Fabrice, *Analysis of ADSL Trafficon an IP Backbone Link*, IEEE IEEE Global Telecommunications Conference 2003, San Francisco, USA,December 2003.

[7] Cho, Kenjiro, Fukuda, Kenshue, Esaki, Hiroshi and Kato, Akira, *The Impact and Implications of the Growth inResidential User-to-User Traffic*, ACM SIGCOMM 2006, Pisa, Italy, September 2006.

[8] Balachandran, Anand; Voelker, Geoffrey M.; Bahl, Paramvir and Ragan,P. Venkat, *Characterizing user behavior andnetwork performance in a public wireless LAN*, Proceedings of the 2002 ACM SIGMETRICS International Conference
on Measurement and Modeling of Computer Systems, pp. 195-205, 2002.

[9] Internet assigned numbers authority (IANA), http://www.iana.org/assignments/port-number (last accessed October, 2009)

BIOGRAPHIES

V.Srikanth is currently working as Associate Professor in Pydah College of Engineering, Visakhapatnam, AP, India. He has 17 years of experience and his interested areas are cloud computing, network security.