An Empirical Model of Network Traffic Classification for Identifying Anonymous Behavior

Vasanthi Pampana[1], J Peter Praveen[2]

[1]Final MTech Student, [2]Assistant professor
[1,2] Computer Science and Engineering
[1,2] Visakha Institute Of Engineering And Technology

Abstract: Optimizing the internet traffic is always an important research issue in the field of network traffic classification, although various approaches available for minimizing the traffic over heads during the network traffic, they are not optimal. In this paper we are proposing an optimized classification approach for internet traffic by analyzing the behavior of the nodes for allowing or dis connection of the incoming node by computing the posterior probabilities of the factors with respect to the node.

I.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.

A parallel Signature based technique proposed by the some researchers, in this approach, they are analyzing the network traffic with parallel processing, "In this method, complete rule groups are spread across nodes. It is possible to use a packet duplicator to send every packet to every node for processing, or a traffic splitter to route each packet to the appropriate node. In this case, rules are clustered into rule groups based on source and destination ports. So, a traffic splitter could route packets based on port number

In port based classification firewall log can be classified or analyzed by the port, whether incoming node is accessing the open port or other than open port number of the server otherwise trust metrics are one of the factors to measure the authenticity of the user while communicating to the nodes in the network.

Port Based classification:

However, many researches claim the portnumber-based classification is not sufficient. Moore and Papagiannaki claimed the accuracy of port-based classification is around 70% during their experiment. Moreover, Madhukar and Williamson claimed in their research that the misclassification of port-based classification is between 30% and 70%. [1] The main reason for choosing static port numbers is to make the packet more able to go through the server firewalls. Many recent applications try to avoid the detection of firewall by hiding the port numbers. Some of the other applications use dynamic port numbers instead of static ones. And servers which share the same IP address will use un-standard port numbers.

Payload-Based Classification

Another approach to classify packets is to analyze the packet payload or use deep packet inspection (DPI) technology. They classify the packets based on the signature in the packet payload, and it has been touted 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], "MAIL FROM","RCPT TO" and "DATA", as in Figure 1, are the commands that appear in the payload of SMTP packets.

II. 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.

Major issues with the signature based mechanisms are time complexity and network overhead while transmission of the packets along with hash codes from the both ends. Various SVM based classification mechanisms available in terms of computing the probability in the occurrences of all positive and negative occurrences of the training sample datasets.

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 _eld. However, our approach may be contrasted with previous work which attempts the classification of network traffic. We present a sample of such papers here. Roughan *et al.* [10] perform 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. In contrast, McGregor *et al.* [11] seek to identify traffic with similar observable properties and apply an untrained classifier to this problem. The untrained classifier has the advantage of identifying groups/classes of traffic with similar properties but does not directly assist in understanding what or why applications have been grouped this way. However, such a technique may be suitable for applying the rest series of classification where the traffic is completely unknown and no previous classification has been previously applied

III. 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 new agent information with existing agent information, by classifying the feature sets or chatectertics 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, ie, 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

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

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 Ci iff

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

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

IV. 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 approaches before 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.

V.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 p2p traffic classification, in: MASCOTS '06: Proceedings of the 14th 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 TelecommunicationsConference 2003, San Francisco, USA, December 2003.

[7] Cho, Kenjiro, Fukuda, Kenshue, Esaki, Hiroshi and Kato, Akira, The Impact and Implications of the Growth in Residential 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 and network 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)

[10] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield. Class-of-Service Mapping for QoS: A statistical signature-based approach to IP traffic classification. In ACM SIGCOMM Internet Measurement Conference, Taormina, Sicily, Italy, 2004.

[11] A. McGregor, M. Hall, P. Lorier, and J. Brunskill. Flow Clustering Using Machine Learning Techniques. In Proceedings of the Fifth Passive and Active Measurement Workshop (PAM 2004), April 2004.

[12] A. Soule, K. Salamatian, N. Taft, R. Emilion, and K. Papagiannaki. Flow Classification by Histograms or How to Go on Safari in the Internet. In Proceedings of ACM Sigmetrics, New York, NY, June 2004.

[13] F. Hernandez-Campos, ´ A. B. Nobel, F. D. Smith, and K. Jeffay. Statistical clustering of internet communication patterns. In Proceedings of the 35th Symposium on the Interface of Computing Science and Statistics, Computing Science and Statistics, volume 35, July 2003.

[14] A. W. Moore, J. Hall, C. Kreibich, E. Harris, and I. Pratt. Architecture of a Network Monitor. In Passive & Active Measurement Workshop 2003 (PAM2003), La Jolla, CA, April 2003.

[15] A. W. Moore and D. Zuev. Discriminators for use in flow-based classification. Technical report, Intel Research, Cambridge, 2005