International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 3, March - 2014

A Survey on Network Traffic Classification Techniques

Aafa J S
Computer Science and Engineering
Sree Chitra Thirunal College of Engineering, Trivandrum

Soja Salim
Computer Science and Engineering
Sree Chitra Thirunal College of Engineering Trivandrum

***Abstract*— Efficient identification and classification of network traffic plays an important role in network management system. It has many advantages such as diagnostic monitoring, flow prioritization and traffic policing/shaping. Moreover, during network congestion it is vital to identify the type of traffic that flows over the network. In that situation network operators usually underutilize the link capacity. Some applications require different QoS requirements. Because of these reasons traffic classification has to be performed in the middle of the network.**

Classification technique based on the port numbers of the packets were used during early days of the Internet. Later to address the issues specified by the port based approach is overcome by payload based technique which inspect the payload content. However, it may leak the privacy of the packet. Due to security issues, statistical features of the packet were examined to identify the application that generated them. This paper makes a survey on different types of classification algorithms for IP traffic. Moreover, the performance of each technique is discussed. Finally a summary of these techniques is given.

Keywords: QoS, traffic classification, security.

I. INTRODUCTION

As the internet grows as the most critical communication infrastructure, internet service providers attempt to provide privacy, security, reliability and multiple service qualities thereby giving best-effort architecture. Since the last few years we were experienced with a number and variety of applications over internet such as real time, interactive, corporate and bulk data transfer application. These may cause some network security risks. Looking on one side, some applications require lot of bandwidth thereby congest the network and thus reduces the network performance. On the other side, some may result in the distribution of malicious codes such as Virus and Trojan horse. These may leaks the privacy. So proper classification of network traffic according to their application that generated them should be done to such as to prioritize, protect or prevent some traffic. Network traffic identification is crucial due to various reasons such as security monitoring, accounting, forecasting long term provisioning, QoS measurements etc. It is also useful to address the security problems including lawful interception and intrusion detection [1].

Real time application classification has the ability to solve most of the network management problems for ISPs and equipment vendors. Classification is performed using different techniques. During early days of Internet, transport-layer or UDP port numbers are used to identify a particular traffic. This method was proposed under the assumption that most applications consistently use well-defined TCP or UDP port numbers which are registered in IANA (Internet Assigned Number Authority). However, nowadays some popular applications usually use port numbers dynamically or use port numbers of other applications to hide their identity. For example, those that uses peer-to-peer (P2P) file sharing. Thus more recently port based approach becomes less liable. Later techniques rely on deep packet inspection to identify the type of application that generated them. Packets having the same 5-tuple of source address and port, destination address and port, protocol type are considered to be coming from the same application and also belong to the same flow. But this approach is losing its effectiveness bases on the assumption that third parties those are affiliated with source or destination can view the packet's payload. Likewise, payload can be interpreted because the classifier knows the syntax of the packet. Packet's confidentiality can be maintained by simply encrypting it including the TCP/UDP port numbers. Payload inspection approach is impossible for encrypted data. While integrity preservation may impose heavy operational overhead on the network- to stay ahead from gradual changes on packet payloads, commercial devices need repeated updates.

Later researches on this area investigated newer approaches capable of inferring application level usage patterns. Some statistical features related to traffic such as packet length and packet inter-arrival time helps to cluster IP traffic flows into classes of same traffic characteristics or patterns. Researchers are looking forward to machine learning techniques for IP traffic classification. After following a number of steps, the future unknown traffic is identified and classified. The ML classifier is trained to associate set of features (flow duration or inter-arrival time, maximum or minimum packet length in each direction) of with known traffic classes. This paper reviews the different traffic classification techniques in the IP network.

II. IMPORTANCE OF IP TRAFFIC CLASSIFICATION

The importance of network traffic classification can be defined in terms of two factors: Quality of Service and lawful interception [6]. In enterprise networks, it is desirable to provide different Quality of Service to traffic from different applications. This is done because different applications may have their own QoS requirements. During the time of any

network congestion, network operators may under-utilize the link capacity. But there should be a mechanism to differentiate users with different requirements and charge for the Quality of Service they receive. Several techniques have been discovered to classify the traffic flows based on the QoS that they require. Emerging QoS enables products as well as automated QoS architectures rely to a great extend on traffic classification techniques.

Several lawful interception services are implemented by Government at various level of abstraction. For example, in the telephony world, law enforcement body may intercept the data such as who called whom or simply tap the call while progressing. With the help of traffic classification techniques internet service providers can differentiate which endpoints are sending packets and when and can identify which application a particular person of interest is using at any given time.

III. REVIEW OF NETWORK TRAFFIC CLASSIFICATION TECHNIQUES

IP traffic classification techniques can be broadly classified into five categories: port-based approach, deep packet inspection classification, payload based classification, statistical features classification and machine learning classification. We first discuss about the port-based approach.

Port-based approach: During the early days of internet, the common practices for traffic classification rely on the usage of port numbers in the transport layer found in the TCP or UDP header [4]. These are registered with IANA to define a well known application. Classification using port numbers are faster and simpler, however, several researcher have proved that it performs poorly. There are three ranges of TCP/UDP ports: Registered ports, the Dynamic/Private ports and Well known ports. Usually a TCP connection starts with a three way handshaking (SYN, SYN-ACK, and ACK). The dynamically chosen port number by the client is specified on the packet. During the entire period of that session both sender and receiver uses the same pair of port numbers to identify themselves on the network. Application is identified by searching the port numbers in the list of registered ports in IANA. Since the port to application mapping is not well defined, network operators move to implement other classification techniques. Researchers addresses the issues associated with port based approach. P2P applications usually use ports [2] which are not registered in IANA. In the case of FTP, for data transfer dynamic negotiation of server port is used. Some attacks such as Denial of Service use port numbers of some application to which they didn't belong. In that case the traffic is mistakenly associated to that application to which the port number represents. Another disadvantage associated with this approach is that applications with different QoS requirements use the same port number which would not liable for traffic classification based on QoS. Moore and Papagiannaki founded that the port based approach using IANA list acquire only 70% byte accuracy. In some cases encryption of packets restrict the extraction port number from the TCP/UDP headers. These problems led the researchers to find an alternative way to classify the network traffic which rely on deep packet inspection.

Deep packet inspection: To address the issues related with port based classification technique, most of the current network devices use session and application layer information to identify the type traffic that the packet represents. These techniques are commonly employed to identify P2P applications and for intrusion detection [1]. Packets that employ same source and destination port/address and protocol are considered to be belong to the same flow. However, it may leaks the privacy of the data in some way. Encrypted payloads are not subjected to deep packet inspection.

Statistical signature based classification: In this case some statistical features about the packets such as packet length and interarrival time between consecutive packets are used to classify them [5]. These statistical features are sometimes called protocol fingerprint. The idea is to classify traffic flows or it may provide behaviour of the traffic. Existing studies have shown the relationship between traffic flows and application that generated them. Certain features such as packet length, flow duration and interarrival time shows the behaviour of protocols.

A. Review of techniques

i. Traffic classification based on clustering algorithms:

Different types of clustering algorithms [7] have been proposed such as K-Means, DBSCAN, AutoClass etc. K-Means is the simplest and quickest algorithm. K-Means is an unsupervised machine learning technique which classifies TCP based applications by using the first few packets in a traffic flow. This type of examination is done on the assumption that first few packets in a TCP connection captures the application's negotiation phase and is unique for each application. In the training phase, from a given dataset, the algorithm classifies the objects into k disjoint clusters. Here objects refer to each flow. Then the square error is minimized within a cluster thereby maximizing the homogeneity. Square error is the square of the distance between each object and the centre of cluster. The new flow is assigned to cluster to which the distance is minimum. The output of the learning algorithm consists of description of each cluster and composition of its application. In the classification phase, packets are transformed to a bidirectional flow. The sizes of first few packets are used to map the flow to a spatial representation. The flow is assigned to the most prevalent application in the cluster. AutoClass is based on the unsupervised Bayesian classifier. During training phase, a subset of flow data is selected using sampling. New flows are classified once the classifier learnt the clusters.

ii. Classification based on Bayesian techniques:

Internet traffic classification using Naive Bayes technique [3] is a supervised machine learning technique. Flow contents such as port numbers, flow length and time between consecutive flows are used to train the classifier. Moreover, to train the classifier 248 full-flow based features were used. The chosen traffic for application was categorized into different groups such as database, mail services, games and multimedia, www, p2p, bulk data transfer and attack. The flows are subjected to probabilistic class assignment. When a flow comes its posterior probability of class membership is

calculated against each class. Then the flow is assigned to that class to which maximum probability is attained. The authors used two metrics for performance evaluation: Accuracy and Trust. Later the work is extended with use of neural network approach [8]. The authors proposed a Bayesian framework which classifies traffic without the use of any port or host address. A multilayer perceptron classification network is used for assigning probabilities to flows. 246 flow features are used as input to the first layer of network. The output represents ten classes of membership to which a particular flow belongs. Class membership is determined by calculating the probability density function.

iii. Statistical fingerprint based classification

Statistical protocol fingerprint based classification algorithm proposed in 2007 classifies the network traffic based on three features related to a packet [9]: packet size, inter arrival time and arrival order. For classification, an algorithm based on anomaly score threshold is used. In the training phase, the training dataset i.e., pre-labelled flow from the application to be classified are analyzed and construct the protocol fingerprint. The protocol fingerprint is then indicated as a PDF(Probabilistic Density Function) vector. For all the i^{th} pairs of $P_i = \{s_i, \nabla t_i\}$ PDF_i is build. Here s_i denotes the packet size and ∇t_i denotes the inter arrival time between packet i and i-1. For the unknown flow, the algorithm checks if there is at least one PDF whose description is compatible with the behaviour of that flow. Then associate the flow to that PDF which describes it better. The statistical distance between the unknown flow and PDF is given by a measure called anomaly score which have values ranges from 0 to 1. This distance describes the correlation between i^{th} packet of the flow and the application layer protocol indicated by the given PDF. Higher the value, higher the chance that the flow is generated by that application.

The technique has a disadvantage that the classifier assumes that it will always capture the first few packets in a flow. The classifier is unaware of the packet loss and packet reordering. If the classifier misses first few packets in a flow, then it can't construct the correct protocol fingerprint.

iv. Correlation information based classification:

Network traffic classification using correlation information [10] between the traffics improves the classification performance. The paper describes a novel parametric approach and the performance benefits were detailed through theoretically and experimentally. The input IP packets go through a series of operations such as pre-processing, feature extraction, flow correlation analysis and robust classification. During the pre-processing step traffic flows are constructed from IP packets which are crossing through the network by examining the IP header. Each flow is represented by some statistical features. These features are relevant for constructing efficient classification model. The features extracted by the authors are number of packets transferred in unidirection, volume of bytes transferred in unidirection, minimum, maximum, mean and standard deviation of packet sized in unidirection and minimum, maximum, mean and standard deviation of inter packet time in unidirection. Flow correlation analysis is done to extract the correlated information in the traffic flows. Based on the correlation information and features extracted, the robust classification module classifies traffic into application based classes. To model the correlation information, Bag of Flow is used. BoF consists of traffic flows which are correlates with each other and are generated by the same application. For example the correlation information consists of three tuple {destination port, destination ip, protocol}. The authors used a Maximum-Likelihood classifier which would be based on the Bayesian decision theory. The non parametric approach for the classifier is given by the equation:

$$\omega^* = \arg\min_{\omega} \frac{1}{\|Q\|} \sum_{x \epsilon Q} \min_{x'} \left\| x - x' \right\|^2$$

where Q if the flow to be classified and ω^* denotes the class to which the flow is to be associated. The performance is evaluated by the parameters accuracy and F-measure. The experiment is done on two types of data sets: wide and isp. the wide data set is comprised of 182 k traffic flows chosen from the wide trace. The data set is recognized by the DPI tool. While the isp consists of 200 k flows which are randomly selected from 11 major classes.

Later the work is extended by using an aggregated predictor which is based on the Naive Bayes theory. For each testing flow, NB algorithm produce the set of posterior probabilities as predictions. The flow is assigned to the application to which it attains a maximum probability. Testing is done on the same data set as used by the previous method. Performance evaluation shows that this method improves the performance.

v. Multistage classifier

Multistage classifier [11] uses three major approaches such as port based approach, deep packet inspection and statistical based approach to classify the network traffic. Two databases are used: port database which contains the port numbers and the corresponding application class and a signature database which consists of the signature associated with a packet and corresponding application. The algorithm first checks if the newly arrived packet belongs to an existing session. If not, new session is created to which it belongs otherwise it is added to the corresponding existing session. Then first checks the port database if the port number of the packet exists in it. If it is there, the packet is classified to that application that the port corresponds to. If both of these methods fails to classify the traffic, then it is classified based on the statistical approach. The session of packets is classified by using SVM (Support Vector Machine) algorithm. The authors used eight popular applications to train the classifier: Skype, POP3, PPLive, SMTP, BitTorrent, QQ, eDonkey and MSN. The performance is evaluated by using False Positive and False Negative rates.

B. *Performance evaluation of techniques*

We have seen a number of techniques for the traffic classification. Each method test on different training set of data. The K-Means algorithm proposed in 2006 classified one hour packet trace of TCP flows. The results show that the method achieved an accuracy of 49% for the Auckland dataset and for Calgary data set it achieved an accuracy of 67%. As the number of clusters increases the overall accuracy also get increases. For the AutoClass algorithm, the cluster parameters and number of clusters are determined automatically. This

method outperforms the K-Means algorithm by achieving an accuracy of 92.4% for Auckland and 88.7% for Calgary dataset.

Bayesian analysis uses accuracy and trust metrics to evaluate the performance of the classifier. The simple Bayesian analysis technique achieves an accuracy of 65% for the whole population of flow features. The extended work of Bayesian technique i.e. with the use of neural network achieves and accuracy of 99% for data trained in one day and an accuracy of 95% for data trained and tested eight months apart.

Statistical protocol fingerprint based approach gain an accuracy of more than 91% for the classification of three applications: POP3, HTTP and SMTP.

The performance of classification method using correlation information is evaluated using the parameter F-measure.

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

Precision denotes the ratio of correctly classified flows over all predicted flows in a class while recall is the ration of correctly classified flows over all ground flows in a class. The approach is implemented on three methods: AVG-NN, MIN-NN and MVT-NN. Experimental results showed that by using correlated information, it outperforms the ordinary NN classifier when a small set of training data are available. The increase in overall accuracy ranges from 10 to 20 percent. Moreover, the classification time may ranges from 2 to 5 seconds for the wide and isp data sets. The extended work using aggregated correlation information further improves the performance. The experimental results showed that BoF-NB achieves a better classification performance than other methods due to its ability to utilize flow correlation information.

Multistage classifier tested the algorithm with eight applications. The performance was evaluated by using False Positive and False Negative rates. For all the eight applications, the method achieved a False Positive rate ranges from 0 to 1.5 and a False Negative rate ranges from 0 to 2.2. The identification rate achieved by this method is 98.15%.

IV. SUMMARY OF TECHNIQUES

Techniques	Features	Traffic considered
K-Means	• Total number of packets. • Packet length. • Flow duration. Mean inter-arrival time.	Web, P2P, FTP
AutoClass	• Flow size. • Flow duration. • Packet length statistics. • Inter-arrival time statistics.	• HTTP, DNS, Telnet, Half-Life, SMTP, FTP.
Bayesian techniques	Total 128 features: • Flow duraction. • Packet inter-arrival time. • TCP port. Payload size statistics.	• P2P, Buck, Services, Mail, a large range of database.
Bayesian neural network	• Number of packets transferred in unidirection. • Volume of bytes transferred in unidirection. • Packet size statistics.	• P2P, Buck, Services, Mail, a large range of database.
Aggregating correlated Naive Bayes predictions	• Number of packets transferred in unidirection. • Volume of bytes transferred in unidirection. • Packet size statistics.	• P2P, WWW, DNS, CHAT, FTP and MAIL.
Multistage classifier	• Number of packets. • Number of packets received and sent in a session. Maximum and variance of packet length.	• SMTP, POP3, QQ, MSN, BitTorrent, eDonkey, PPLive, Skype.

V. CONCLUSION

This paper surveys different techniques that are used for network traffic classification during the peak period of 2004 to 2013. When compared to port based and payload based techniques, the statistical feature approach achieved a better performance by giving a good identification rate. The use of AutoClass, correlation information and neural networks, the methods achieves an accuracy of more than 90% for a various range of application traffic. Early techniques were based on the static and offline analysis of traffic. But now researchers are addressing the issues for implementing a better classifier over the network.

However, there is still a lot of space for research in this field. While most of the approaches are implemented for certain applications, the work has to be extended to apply for a wide variety of applications. There still a question arises that how can the classifier maintain its performance when a packet loss occurs.

REFERENCES

[1] Snort - The de facto standard for intrusion detection/prevention, http://www.snort.org, as of August 14, 2007.
[2] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in network identification of P2P traffic using application signatures," in *WWW2004*, New York, NY, USA, May 2004.
[3] A. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in *ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) 2005*, Banff, Alberta, Canada, June 2005.
[4] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: Multilevel traffic classification in the dark," in *Proc. of the Special Interest Group on Data Communication conference (SIGCOMM) 2005*, Philadelphia, PA, USA, August 2005.
[5] 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 *Proc. ACM/SIGCOMM Internet Measurement Conference (IMC) 2004*, Taormina, Sicily, Italy, October 2004.
[6] K. Claffy, "Internet traffic characterisation," PhD Thesis, University of California, San Diego, 1994.
[7] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms," in *MineNet '06: Proc. 2006 SIGCOMM workshop on Mining network data*. New York, NY, USA: ACM Press, 2006, pp. 281–286.
[8] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for Internet traffic classification," *IEEE Trans. Neural Networks*, no. 1, pp. 223–239, January 2007.
[9] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," *SIGCOMM Comput. Commun. Rev.*, vol. 37, no. 1, pp. 5–16, 2007.
[10] Jun Zhang, Chao Chen, Yang Xiang, Waneli Zhou Yong Xiang, "Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions," *IEEE Trans. On Inf. Forensics and Security*, Vol. 8, No. 1, Jan 2013.
[11] DU Min, CHEN Xingshu, "Online Internet Traffic Identification Algorithm Based on Multistage Classifier," *IEEE Trans. On Inf. Forensics and Security, Vol. 8, No. 1, Jun 2013*.