Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 3159-0040
Vol. 2 Issue 8, August - 2015

Identifying Peer-To-Peer Traffic Based On Traffic Characteristics

Suraj Sanjay Dangat
Dept. of Computer Engineering
SIT, Savitribai Phule Pune University
Lonavala, India
Surajdangat6691@gmail.com

Prof. S. R. Patil
Dept. of Computer Engineering
SIT, Savitribai Phule Pune University
Lonavala, India
Srp.sit@sinhgad.edu

***Abstract*—The P2P (Peer-to-Peer) network is dynamic, self-organized and has some other features. So, P2P traffic has become one of the most significant portions of the network traffic. But, it has also caused network congestion problems because of resource occupation (mainly bandwidth). Accurate identification of P2P traffic makes great sense for efficient network management and efficient utility of network resources. In this paper we address traffic identification technology, such as traffic identification by using network characteristics have been considered for solving the problem with different parameters, improved accuracy rate and efficiency. Experiments on various P2P applications demonstrate that the method is generic and it can be applied to most of P2P applications. Experimental result shows that the algorithm can identify P2P application accurately. In this paper we first briefly introduce P2P technology and then we made a short survey on the overall progress in P2P traffic identification technologies. Finally we discuss the proposed method and its result analysis.**

Keywords— P2P; Peer-to-Peer, Traffic characteristics, Peer-to-Peer Traffic identification

I. INTRODUCTION

A network is a group or system of interconnected people or some things. A computer network or data network is a telecommunications network in which a group of two or more computer systems linked together and it allows computers to exchange data. In computer networks, networked computing devices pass data to each other along network links and these links between nodes are established using either wired media or wireless media. The widely-known example of computer network is the Internet. In a P2P network, the "peers" are computer systems which are connected to each other via the Internet and peers are configured to allow certain files and folders to be shared with every peer or with selected peers. Files can be shared directly between the different peers on the network without the need of a central server. In short, each computer on a P2P network acts as a file server as well as a client.

With the extensive application of P2P technology, P2P applications consume a large amount of network bandwidth, which ultimately increase the burden of the network. According to the available statistics, P2P applications account for 60% to 80% of total ISP business and become the largest consumer of network bandwidth. According to the statistics, about 60 percent of the bandwidth is occupied by P2P applications and from these 60%, 80% of which were occupied by P2P file-sharing applications. But these P2P file-sharing users are very low in number and they are only 5% of the total number of Internet users. The P2P has many characterizes, such as large flow, automatic operation, connection for a long time, regardless of time running, and so on. Therefore, P2P applications can take up more bandwidth than other applications. As number of P2P user's increases, network traffic will also increase significantly. At the same time, increasing the size of the network will lead a large number of broadcast news to flood in the entire network and so the network traffic increases. In the end, it led to bottlenecks of the network and network congestion damages the services provided by the Internet Service Providers and common users [1].

The solution to this bandwidth congestion problem is that we limit the users who use large amount of bandwidth to protect those user who use a small amount of bandwidth when the network resource constraints. On the contrary, when there are no constraints on network resources, we remove these restrictions so that each user can use the lines efficiently. How to effectively control the available network resources and how to effectively control the P2P traffic have become quite important questions. Therefore, P2P traffic identification is a key technology for effective control over it.

P2P traffic which is produced by P2P application has different characteristics than the other applications. In P2P traffic identification based on traffic characteristics, these characteristics are used to identify the network traffic. This approach can detect P2P traffics with dynamic ports and encrypted transmission. This method easily detects the flow of encrypted payload and unknown payload characteristics of P2P traffic [2].

In recent years, researchers have been using traffic characteristics in the area of computer network for identifying P2P traffic. It is very relevant for the study of network traffic control, traffic congestion, resource utilization and it also has an economic importance. In this paper, the main focus is on three things i.e. on classifying all the incoming traffic in the network, identify the P2P traffic in the network and to minimize the rate of false detection and the rate of negative detection. The rest sections of this paper are

arranged as follow. In section II, we briefly introduce Details of P2P traffic classification technologies. In Section III, we briefly discuss the System Architecture. In Section IV, we represent proposed method by means of Mathematical model. In Section V, we analyze the Result of proposed system. In Section VI, we discuss the conclusion of proposed system.

II. EXISTING METHODS OF P2P TRAFFIC CLASSIFICATION TECHNOLOGIES

There are various techniques available for P2P traffic identification and classification. It includes Port-based classification, Payload-based classification, Feature-based classification and Hybrid classification method. It is discussed in following paragraphs.

Port-based classification method is the simplest and traditional method. It identifies the application traffic by identifying the application type and this application type is identified from the port number used in the transport layer. For example, TCP port 80/443/1024 is Skype traffic, TCP port 1214 is Kazaa P2P traffic and so on. This approach is extremely easy for implementation and it gives very little overhead on the traffic classifier. But, it also has some limitation and it becomes less accurate because of several reasons. These are, many applications uses random ports and some P2P applications use dynamic ports which are not known in advance [3].

To remove the drawbacks of port-based classification method, several payload-based classification techniques have been proposed. Most protocols contain a protocol specific string in the payload namely signatures that can be used for identification. The information about this strings are publicly available. Subhabrata [4] presented an analysis of a number of P2P application and their signatures. For example “0x13Bit” corresponds to the BitTorrent application. By comparing every packet payload with a set of previously determined signatures, this method can identify application traffic more accurately than the traditional Port-based method. The benefits of this method are high accuracy and robustness, and have a good classification functions. However, there are still some disadvantages of this method. These methods identify only P2P traffic for those signatures which is known in advance and it is unable to classify any other traffic.

Because of the disadvantages of these two methods, the research community started developing the new methods which are less dependent on particular individual applications, but focused on capturing and extracting common things in the behavior of P2P applications. We refer this as feature-based technique. This kind of approach is to classify traffic based on the analysis of hidden transition patterns of traffic flows. Such nonlinear properties cannot be affected by dynamic port change or payload encryption. These methods provide an alternative for effective traffic classification.

There are also some hybrid P2P traffic classification methods available. This approach includes most of the proposed methods for improving classification accuracy. Most of the traffic identification method of P2P based on traffic characteristics is to select particular feature from the P2P features [5, 6]. There are four main P2P features. First is that the number of other hosts that P2P hosts are connecting to be bigger than the traditional hosts. Second, P2P traffics of up and down are roughly equal and it is different from the traditional characteristics of the host. Third, P2P hosts are act as a both servers as well as clients, which differ from the traditional hosts. Fourth, the connection features of P2P hosts listening port are different from the traditional ones [7]. In the reference [8], the authors proposed a simple algorithm based on the first feature. In the second reference, the authors proposed a flow characteristics identification algorithm based on first and fourth feature and also uses factors that affect the efficiency of identification algorithm. Proposed system considers different features and uses K-means and Naive-Bayes algorithm to improve accuracy rate. It minimizes the rate of false detection and the rate of the negative detection. The next section describes the system architecture of the proposed system.

III. SYSTEM ARCHITECTURE

The system architecture of our proposed system is discussed briefly below. It is shown in following figure1. It has two phases, In Phase 1 we calculate the optimum thresholds and in Phase 2 we apply these optimum thresholds values for classifying P2P traffic. Input to the proposed system is all traffic. We get output as classified traffic. There are one preprocessing module and three main modules. In preprocessing module, means in Calculate Optimum Threshold We calculate the optimum threshold value in it. Main modules are namely Tracking Traffic Model, Traffic Identification Module and Traffic Classification.

A. Module I-Tracking Traffic Model

In this module, System tracks the traffic using Packet Sniffer and represents it in the form of array of packet objects. Input to this module is all traffic. The output of this module is Tracked Traffic. There are two steps in it and these are Packet Sniffer and Incoming Packets Dataset. In Packet Sniffer step all traffic is passed to the Packet Sniffer and it continually tracks the maximum traffic as much as possible. Then in Incoming Packets Dataset step, all traffic which is tracked by Packet Sniffer is representing in the form of Packet class objects. The array for this is maintained.

B. Module II-Traffic Identification Module

In this module three things take place and these are packet analysis, feature set, packet classification takes places. Input to this module is the array of packets. Output of this module is classified packets. There are three steps in it and these are Packet Analyzer, Feature Set and Packet Classification. In Packet Analyzer step, the array of packets which is tracked by Packet Sniffer is passed to Packet Analyzer to analyze each packet in details. Additional information which is getting by this step we represent and maintained using collection objects. In Feature Set

step, system maintains the particular features and its ideal value for packet classification. These features

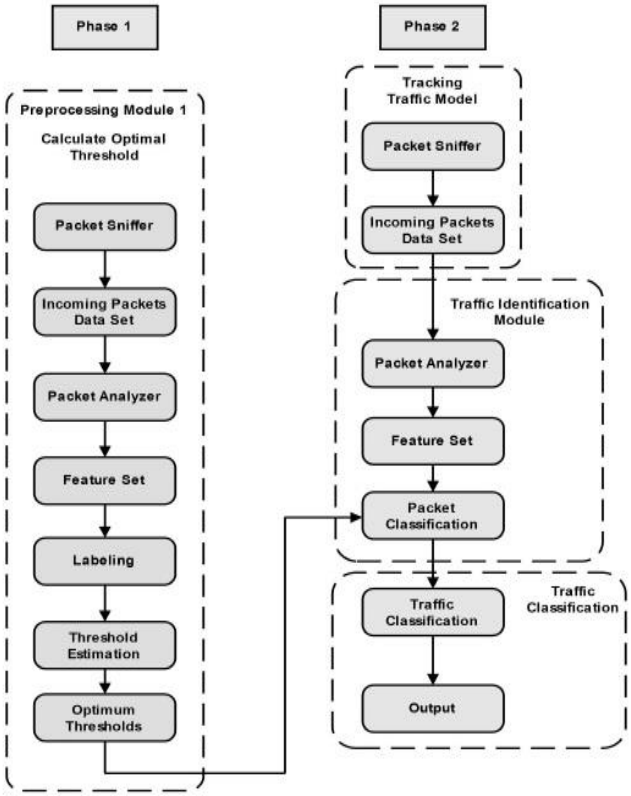

Fig. 1. State System architecture of proposed system

include packets destination IP, hop limit, packet length and port number. We determine the optimal threshold values for particular traffic in preprocessing module namely calculate Optimum Thresholds and it is part of it. We apply K-Means algorithm in it. First we consider trainee features of packet then we finds two centroids for packet length that is Centroid 1 for minimum value and Centroid 2 for its maximum value. K-means is one of the simplest unsupervised learning algorithms that solve the widely-known clustering problem. The procedure of K-means algorithm follows a simple and it is easy way to classify a given data set through a certain number of clusters assume as a k clusters, fixed a priori. The main idea is to define k centroids, one for each cluster and these centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other and then, the next step is to take each point belonging to a given data set and associate it to the nearest centroid.

In Packet Classification step all packets then classified by using Naive-Bayes classifier algorithm. The input to this step is calculated optimum threshold values, collection of feature set and tracked packet array. We consider the min and max values of packet length of packets in it which is calculated in preprocessing module. We also consider hop limit and port number features. The naive Bayes algorithm is a simple probabilistic classification algorithm. In simple terms, a naive Bayes classifier assumes that the presence or absence of a particular feature of a class is not related to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be a tomato if it is red, round, and about 3" in diameter. Even if these features depend on each other or upon the existence of the other features a naive Bayes classification algorithm considers all of these properties to independently contribute to the probability that this fruit is a tomato [9]. Same thing is applied for each packet. The naive Bayes consider all the properties of packet and its value to independently contribute to the probability that this packet is P2P packet or Non-P2P packet..

K-Means Algorithm:-

1. *Place k points into the space, each point represents the packet length of the distinct packet. These points represent initial group centroids.*
2. *Assign each point to the group that has the closest centroid.*
3. *When all points have been assigned, recalculate the positions of the K centroids.*
4. *Repeat step 2 and 3 until the centroids no longer move.*

Naïve - Bayes Algorithm:-

The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X=(x1, x2… xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2,.., An.

2. Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if

$$P(Ci|X) > P(Cj|X) \text{ for } 1 \le j \le m, j \ne i$$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(Ci \mid X) = P(X \mid Ci)\, P(Ci) / P(X)$$

3. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is P(C1)=P(C2)=…=P(Cm), and we could therefore maximize P(X | Ci). Otherwise we maximize P(X | Ci) P(Ci).

4. Given data sets with many attributes it would be extremely computationally expensive to compute P(X | Ci).

$$P(X \mid Ci) = P(X1 \mid Ci) * P(X2 \mid Ci) * \ldots\ldots P(Xm \mid Ci).$$

5.The predictable class label is the class Ci for which P(X | Ci) P(Ci) is the maximum.

C. Module III-Traffic Classification

TABLE I. MAPPING TABLE

Sr. No.	Function	Description
1	f1(P) → P	Tracking traffic by using Packet Sniffer.
2	f2(P) → PA	Analyze Packet using Packet Analyzer.
3	f3(PA) → D	For Updating Dataset of the system.
4	f4(D, FE) → PE	Packet of Particular Feature.
5	f5(PE, C, CL) → CL	To Find cluster.
6	F6(CL) → C	To get Classified Traffic.
7	f7(C) → O	To represent Output as Set.

In this, system classifies the traffic and represents the output to user in textual form. Input to this module is classified packets. Output of this module is classified traffic. There are two steps in it and these are Traffic Classification and Output. In Traffic Classification step we classify the particular traffic. In Output step system represents the output as two sets namely as P2P and Non-P2P traffic.

IV. MATHEMATICAL MODEL

A. System description by means of mathematical formulas

As Let, S is a system which is defined in following manner:

S = {P, PS, PA, FE, MD, KM, C, CL, NB, O, F|f1, f2, f3, f4, f5, f6, f7}

Where,

P = Set of Packets

PS = Packet Sniffer

PA = Set of Analyzed Packets

FE = Feature Extraction

MD = Manage Database

KM = K-Means Algorithm

C = Set of Centroids

CL = Set of Clusters

NB = Nave-Bayes Algorithm

O = Set of Output

SET THEORY:

P = {P0, P1, ..., Pn}

PA = {PA0, PA1, ..., PAn}

C = {C0, C1, ..., Cn}

CL = {CL0, CL1, ..., CLn}

O = {P2P, Non − P2P}

INPUT:

P = {P0, P1, ..., Pn}

OUTPUT:

O = {P2P, Non − P2P}

FUNCTIONS:

f1 = For Packet Sniffer to capture network traffic

f2 = For Packet Analyzing

f3 = For Updating Dataset

f4 = For Feature Extraction

f5 = For Finding Clusters

f6 = For getting Classified Traffic

f7 = Output

B. Function mapping table

Above functions can be mapped onto the elements of the set. It is shown in table I.

C. State transiction diagram

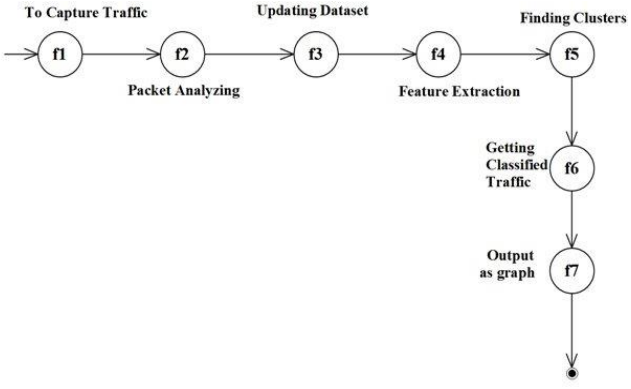

Fig. 2. State transition diagram of system

The basic idea of this diagram is to define a machine that has finite states. It represents the states of the system. Following figure represents the state transaction diagram of the proposed system.

V. ANALYSIS OF RESULT AND ALGORITHM PERFORMANCE

In campus network it is hard to determine the P2P host. Even we could determine the nodes which use P2P but it is hard to determine how much of traffic is generated by P2P and non-P2P applications. So we conduct the experiments in the laboratory. In that experiment we used some sets of experimental machines which is used for generating enough amount of P2P traffic. It is used for testing robustness and detection accuracy rate of the earlier and proposed methods. These experimental machines generate different types of P2P and Non-P2P traffic during different period of time. We set threshold value and we analyzed network for 1 week. Results of the experiments are shown in Table II.

As we are using K-means for determining min and max threshold values of properties of packet, it makes our analysis and proposed scheme more accurate. The time complexity of proposed method is less than the previous method as we are using Naïve-Bayes classification algorithm. Although the proposed method in this paper still having some weaknesses and false detection rate but they are acceptable.

TABLE I. EFFRCTIVENESS COMPARISON BETWEEN EARLIER TRAFFIC CHARACTERISTICS BASED METHOD AND METHOD IN THE PAPER

Application	Earlier traffic characteristics based method negative rate	Earlier traffic characteristics based method false rate	Expected result of negative rate in the paper	Expected result of false rate in the paper
BT	5.34%	2.86%	2.3%	1.86%
EMule	5.82%	3.23%	2.82%	1.23%
Non-P2P	4.26%	4.84%	2.18%	2.44%

VI. CONCLUSION

The traffic identification is very relevant for the study of network traffic control, traffic congestion, resource utilization and it also has an economic importance. With the P2P technology continues to progress, P2P client has experienced the phase of using fixed-port, random port, encrypted message and the tunnel mode. It is hard for people to identify packets of P2P protocols. Some of the traditional identification methods like the Port-based identification method, Payload-based identification method and Feature-based identification are not very effective. Traffic identification technologies such as traffic identification by using network characteristics have been considered for solving the problem with improved accuracy rate and improved algorithm efficiency. The previous work of P2P traffic identification based on traffic characteristics, it gives the high rate of false detection and high rate of negative detection and it is not efficient when we consider time as a factor. So we designed a system in such a way that it minimizes the rate of false detection and the rate of negative detection and classifies the traffic in efficient manner. The presented method in this paper can able to detect P2P traffics with dynamic ports and encrypted transmission, and the efficiency of the implementation of the algorithm is also improved to some extent. These techniques are useful for increasing the performance of the network. This system can easily extend for working with other regions of the network in order to solve the problems in it.

ACKNOWLEDGMENT

We are thankful to Mr. S. D. Baber and Mr. T. J. Parvat, for the encouragement and the support they have extended to us for completing this paper. We are also thankful to the Mrs. S. R. Patil and Mr. Sanjay Dangat for their support to make this paper as good as it is. We are also thankful to our family members and friends for patience and encouragement.

REFERENCES

[1] Yu-shui Geng, Tao Han and Xue-song Jiang ”The research of P2P traffic identification technology.” (978-1-4244-4589-9/09, 2009, IEEE).

[2] Jingyu Wang, Jiyuan Zhang, Yuesheng Tan ”Research of P2P Traffic Identification Based on Traffic Characteristics”(978-1-61284-774-0/11, 2011, IEEE).

[3] Jian Feng ”Research on the Technology of Peer-to-Peer Traffic Classification.”(978-1-4244-5567-6/10 ,2010 , International Symposium on Computer, Communication, Control and Automation, IEEE).

[4] S. Subhabrata, S. Oliver, D. Wang, “Accurate, scalable in-network identification of p2p traffic using application signatures,” Proc. The 13th international conference on World Wide Web, ACM Press, Oct. 2004, pp. 512-521, doi: 10.1145/988672.988742.

[5] Ke Xu , Ming Zhang , Mingjiang Ye , Dah Ming Chiu, Jianping Wu ”Identify P2P traffic by inspecting data transfer behavior.” (0140-3664, Elsevier B.V., 2010, Computer Communications).

[6] CHENG Wei-qing, GONG Jian, DING Wei ”Identifying file-sharing P2P traffic based on traffic characteristics.” (15(4): 112120, December 2008, paper number:10058885, Sciencedirect).

[7] Haiming Jiang, Jianying Zhang, Qingqing Wang, “P2P traffic detection and analysis”, J. Computer Technology and Development, 2008, 18 (7): 116-119.

[8] Chao Wen, Xuefeng ZHENG, “The study of P2P protocol identification method based on the traffic analysis” J. Micro Computer Applications, 2007 (7): 714-717

[9] Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao ”Decision Support in Heart Disease Prediction System using Naive Bayes” (ISSN : 0976-5166, Vol. 2 No. 2 Apr-May 2011, Indian Journal of Computer Science and Engineering (IJCSE))