P2P Traffic Classification for Residential Network

Channary Thay
thay.channary@gmail.com

Vasaka Visoottiviseth
vasaka.vis@mahidol.ac.th

Sophon Mongkolluksamee
theohncom@gmail.com

Faculty of Information and Communication Technology
Mahidol University, Nakhon Pathom, Thailand

***Abstract*—Excessive bandwidth consuming by peer-to-peer (P2P) applications is one of serious problems in residential networks such as in dorms, apartments and even Small and Medium-sized Enterprises (SMEs) networks which have a limited bandwidth. P2P file sharing and P2P streaming applications usually are the cause of this problem. To share the bandwidth fairly among users, the traffic of these applications needs to be classified and filtered out. However, traditional port-based and payload-based classification will fail when the applications use dynamic ports, port disguise and payload encryption. In this paper, we present the classification technique that based on characteristics of number of peer connection and number of traffic in both incoming and outgoing direction for 5-minute duration to classify the P2P traffic. We make use of decision tree J48 to model and classify the traffic. Experimental results over three well-known P2P applications (BitTorrent, Skype and SopCast) confirm that this technique can detect the existence of P2P traffic from the background traffic with 100% accuracy and can classify three types of P2P applications with 90% accuracy.**

Keywords*—decision-tree j48; peer-to-peer application; traffic classification; application identification

I. Introduction

There are huge amount of traffic from various applications exchanged over the Internet. The Internet traffic has increased linearly since 1993. Besides, traffic from P2P applications and video contents dominate the Internet presently. Nadia Ben Azzouna, et.al [5] confirms that around 49% of Internet traffic was P2P application traffic. P2P file sharing such as BitTorrent held about 35% to 70% of all Internet traffic in 2013 [10]. It is also predicted that in 2025 the video communication traffic will hold most of the Internet traffic [4]. File sharing, media stream and instant messaging are three main application types that normally apply a P2P technology. These applications usually consume high network bandwidth and affect to other applications' performance running in the same network. The P2P traffic needs to be classified and managed in order to be fair with other applications. However, classifying P2P traffic is challenging because of its complexities. Thomas Karagiannis, et.al [6] found that about 30% to 70% of total P2P traffic used dynamic ports. Naimul Basher, et.al [7] found 90% of P2P applications which used random ports and about 80% of P2P file-sharing in the network in 2011 [8]. Moreover, some applications also use an encryption technique in their communication.

There are two problems that focused in this research. The first problem is how to identify the existing of P2P traffic in the network under background traffic. The second is how to classify specific types of P2P application from background traffic. We also want to detect P2P traffic in semi-real time manner (within five minutes time interval). The six features base on characteristics of connections and traffic volume for TCP and UDP traffic in both incoming and outgoing from investigated host are extracted. The six features consist of: (1) number of incoming peers, (2) number of outgoing peers, (3) ratio of incoming TCP, (4) ratio of outgoing TCP, (5) ratio of incoming UDP and (6) ratio of outgoing UDP. These traffic features are passed to the decision tree J48 module in the classification process.

To evaluate our technique, we select three popular P2P applications, which are Skype, BitTorrent and SopCast, in our experiment. We use Web traffic as background traffic for this study. Skype was first released in 2003 by KaZaa, and it has three types of node: Skype server, Super node, and Skype client. They run on both TCP and UDP protocol [7]. BitTorrent is a P2P file-sharing that allows peers to download or upload files to the other peers at the same time even though a peer in the network connection is disconnected [8]. Zhe Yang, et.al [9] found that BitTorrent uses both TCP and UDP protocol to upload and download files. SopCast is a kind of P2PTV which allows users to create their own channel to broadcast video streaming, and it mostly relies on UDP protocol [10].

The contributions of this paper are: (1) it can help the network administrator in small networks to detect the existence of P2P traffic in their network. They may later filter out or block those P2P applications, (2) the proposed model and technique to classify multiple P2P applications could help further research in P2P traffic classification.

This paper is organized as follow. Section 2 discusses related works. Section 3 explains our datasets used in the experiments. Section 4 describes our selected features. Section 5 pinpoints proposed work. Section 6 mentions about experiments and results. Lastly, we express discussion and conclude our work in section 7.

978-1-4673-7825-3/15/$31.00 ©2015 IEEE

II. Related Works

Port-based classification techniques are the simplest methods for classifying P2P traffic. These techniques use pre-defined protocols and port numbers to classify target P2P applications [11]. However, the results of these methods become invalid when facing with a dynamic ports scheme of modern P2P applications. Subhabrata Sen, et.al [12] proposed a payload-based method. It works well on dynamic ports and provides high accurate results by using signatures that are created from packet payload. However, it cannot classify encrypted traffic and requires high computational resources in classification process. Sasan Adibi [13] proposed a flow-based method. This technique exploits flow statistics instead of using packet payloads and makes use a machine learning technique in classification. It can classify on encrypted packet payloads but it focused only on a single flow. Communication behaviors of host/application can also be used for classifying P2P traffic. BLINC [14] uses communication behaviors between hosts in a network to classify P2P traffic. It can classify P2P traffic but cannot identify specific P2P application. Chen-Chi Wu, et.al [15] use signaling activities on both host level and message level in 5-minute monitoring time interval to recognize P2P applications. It works well to identify P2P application when there is only single running P2P application on a host. However, it cannot show high performance when there are many running P2P applications on a host.

III. DATASET

The network topology that we use to capture traffic for this study is presented in Fig.1. The traffic of selected P2P applications (BitTorrent, Skype and SopCast) and Web are generated by a laptop. The Web traffic is used to represent the background traffic. These applications communicate with other peers over ADSL network. The laptop's specification is Windows 7 64-bits with 2 GB of RAM, Intel CPU Core i3-2330M @ 2.20 GHz, and 500 GB of Hard Drive. The ADSL network has the download speed of 5.44 Mbps and the upload speed of 2.83 Mbps. We used Wireshark, which is a well-known network analyzer tool, for capturing network traffic and recording them into packet trace file (PCAP) format.

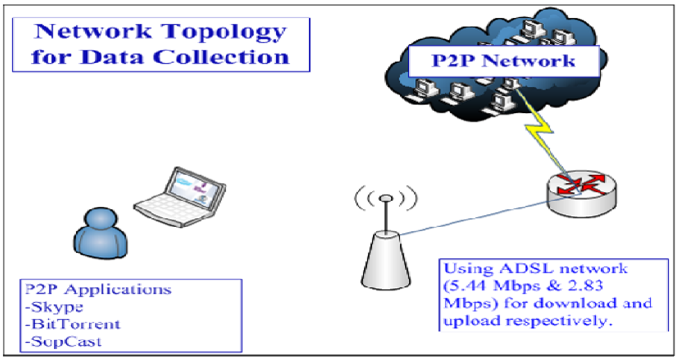

Fig. 1. Network topology for data collection

There are totally 13 trace files that are created for evaluating our proposed technique. Each traffic trace file is two hours long trace. Four of the 13 traces are pure traffic of Skype, BitTorrent, SopCast and Web. Each application is separately run on the laptop while capturing traffic. The next three of 13 traces are mixture of two P2P traffic. Two P2P applicatons are run at the same time while capturing traffic, which are [Skype, BitTorrent], [Skype, SopCast] and [BitTorrent, SopCast]. The remain six traces are the mixture of six P2P traffic (three of pure P2P and three of mixture of two P2P) with Web traffic. The P2P traffic and Web traffic are merged by using tcprewrite and tcpreplay programs.

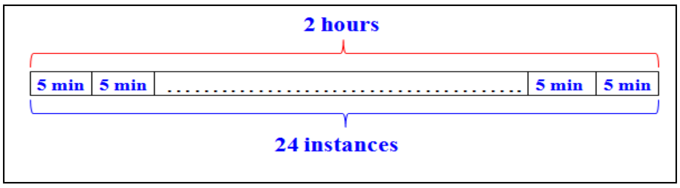

Fig. 2. Splitted trace files into five-minute time interval

We want to classify P2P traffic in a short peroid. Therefore, the long raw traffic files are need to be shorten before extracting traffic features for passing to the modeling and classifying processes. The two-hour trace file is splited by five minutes as shown in Fig.2. Thus, there are 24 instances per each trace file. The 5-minute interval are selected based on the study of [17]. Moreover, based on our observation, 5-minute time interval gives us a good performance in classification performance.

IV. Traffic Features

A. Selected Features

There are six selected features used to model and classify the selected P2P applications. These features come from our literature review [17] and our study on the real P2P traffic. The features can be classified into two parts: characteristics of peer connections and traffic volume. Moreover, from our study, we found that both directions of traffic on investigated host need to be considered in order to gain high classification performance. These six features are: the number of incoming peers, number of outgoing peers, ratio of incoming TCP, ratio of outgoing TCP, ratio of incoming UDP, and ratio of outgoing UDP.

The numbers of incoming and outgoing peer features are very simple features. Number of peers is counted without concerning about a difference of protocols. The characteristics in numbers of incoming and outgoing peer for Skype, BitTorrent and SopCast are presented in Fig. 3, Fig. 4 and Fig.5 respectively. For Skype, in the first 30 minutes, number of incoming peers and number of outgoing peers are high because it tries to connect to the other peers in the Internet. For BitTorrent, number of incoming peers and number of outgoing peers seem equal for every five minutes of timestamp. However, at 55th, 65th, 95th and 105th minutes of timestamp, we found that number of outgoing peers are quite higher than number of incoming peers. For SopCast, number of incoming peers and number of outgoing peers were almost equal. It is less than 100 peers. Moreover, we found that number of outgoing peers are higher than number of incoming peers for every five minutes of timestamp.

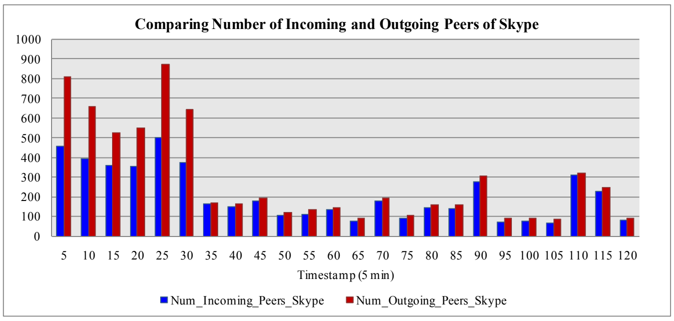

Fig. 3. Comparison on number of incoming and outgoing peers of Skype

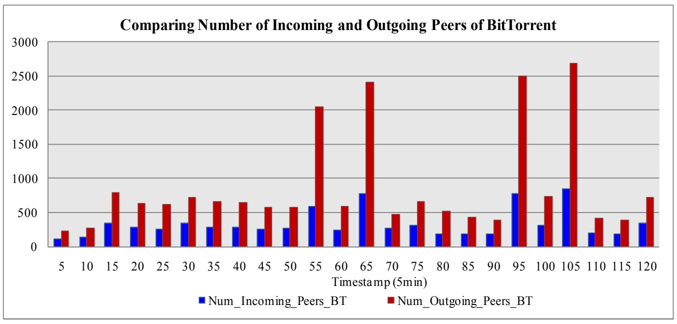

Fig. 4. Comparison on number of incoming and outgoing peers of BitTorrent

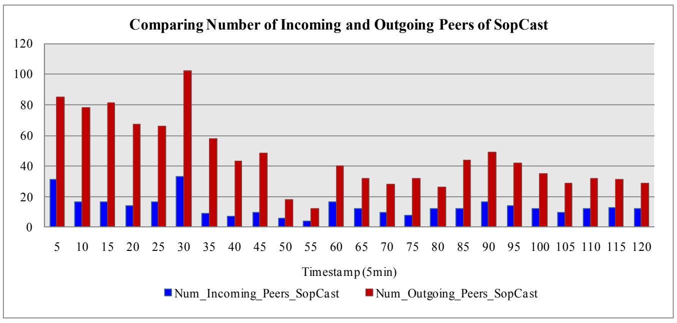

Fig. 5. Comparison on number of incoming and outgoing peers of SopCast

The ratios of incoming and outgoing traffic are number of packets in each direction divided by total number of packets. The characteristics of ratios of incoming and outgoing for Skype and BitTorrent are presented in Fig.6 and Fig.7. For SopCast, it did not use TCP protocol to transfer or receive the video content, so it means that the ratio of incoming TCP and the ratio of outgoing TCP of SopCast were zero. For Skype, the ratios of incoming and outgoing TCP are mostly equal for every 5-minute of time interval. However, at 75th minute of timestamp, the ratio of outgoing TCP is more than 90% while the ratio of incoming TCP is about 30%. For BitTorrent, the ratios of incoming and outgoing TCP are mostly equal for every 5-minute time interval.

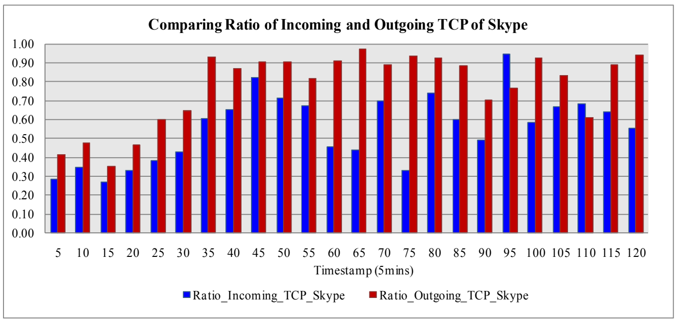

Fig. 6. Comparison on ratios of incoming and outgoing TCP for Skype

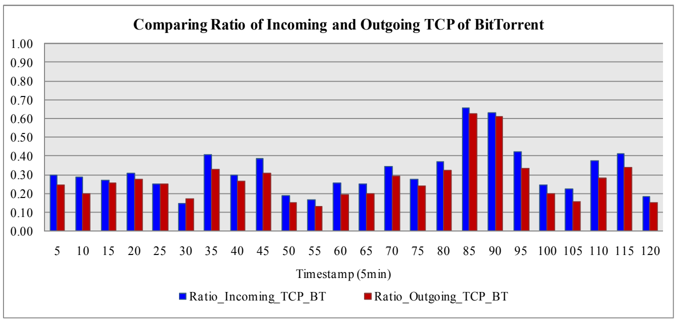

Fig. 7. Comparison on ratios of incoming and outgoing TCP for BitTorrent

The ratios of incoming and outgoing UDP of Skype and BitTorrent are shown in Fig.8 and Fig.9, respectively. For SopCast, it uses only an UDP protocol. It means that the ratios of incoming and outgoing UDP of SopCast are 100%.

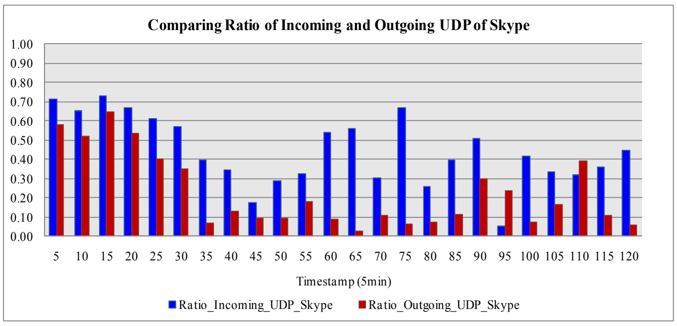

Fig. 8. Comparison on ratio of incoming and outgoing UDP for Skype

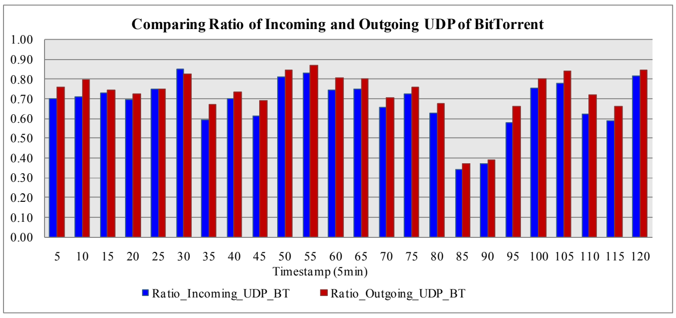

Fig. 9. Comparison on ratio of incoming and outgoing UDP for BitTorrent

For Skype, the ratios of incoming and outgoing UDP of the first 30 minutes are equal. However, after 30 minutes, the ratio of incoming UDP was higher than the ratio of outgoing UDP. For BitTorrent, the ratio of incoming UDP and the ratio of outgoing UDP are mostly equal.

B. *Summarization of Feature Selection on each Traffic Types*

There are four traffic types in this study; and each traffic type contains three traffic trace files as shown in TABLE I to TABLE IV. Each table is a summary of six selected features within 2 hours of our dataset.

TABLE I. SUMMARIZATION ON PURE P2P TRAFFIC

Selected Features	Skype	BitTorrent	SopCast
No. of incoming peers	5,010	7,944	326
No. of outgoing peers	6,918	20,629	1,107
Ratio of incoming TCP	0.56	0.32	0.00
Ratio of outgoing TCP	0.77	0.27	0.00
Ratio of incoming UDP	0.44	0.68	1.00
Ratio of outgoing UDP	0.23	0.73	1.00

TABLE II. SUMMARIZATION ON PURE P2P TRAFFIC WITH BACKGROUND TRAFFIC

Selected Features	Skype and Web	BitTorrent and Web	SopCast and Web
No. of incoming peers	5,733	8,356	1,042
No. of outgoing peers	7,667	21,098	1,863
Ratio of incoming TCP	0.61	0.34	0.10
Ratio of outgoing TCP	0.79	0.29	0.08
Ratio of incoming UDP	0.39	0.66	0.90
Ratio of outgoing UDP	0.21	0.71	0.92

TABLE III. SUMMARIZATION ON MERGED TWO P2P TRAFFICS

Selected Features	Skype and BitTorrent	Skype and SopCast	BitTorrent And SopCast
No. of incoming peers	7,208	3,266	2,510
No. of outgoing peers	15,588	3,854	5,364
Ratio incoming TCP	0.28	0.05	0.22
Ratio outgoing TCP	0.25	0.13	0.17
Ratio incoming UDP	0.72	0.95	0.78
Ratio outgoing UDP	0.75	0.87	0.83

TABLE IV. SUMMARIZATION ON MERGED TWO P2P TRAFFICSWITH BACKGROUND TRAFFIC

Selected Features	Skype, BitTorrent and Web	Skype, SopCast and Web	BitTorrent, SopCast and Web
No. of incoming peers	7,921	3,982	3,226
No. of outgoing peers	16,366	4,631	6,141
Ratio incoming TCP	0.31	0.10	0.26
Ratio outgoing TCP	0.27	0.17	0.21
Ratio incoming UDP	0.69	0.90	0.74
Ratio outgoing UDP	0.73	0.83	0.79

TABLE V below shows the date rate of the merged traffics of P2P traffics and background traffic within 2 hours.

TABLE V. DATA RATES OF MERGED TRAFFICS (MBPS)

Merged Traffics	Ratio P2P Traffic	Ratio Web Traffic
Skype and Web	1.51	0.18
BitTorrent and Web	3.93	0.18
SopCast and Web	1.91	0.18
Skype and BitTorrent and Web	6.71	0.18
Skype and SopCast and Web	3.16	0.18
BitTorrent and SopCast and Web	5.89	0.18

V. PROPOSED WORK

This research aims to answer two different problem statements. Thus, there are two steps in the experiments. In the first step, we try to answer the first problem whether there is P2P traffic in the network or not by creating two classes: P2P and Web. In the second step, we try to answer the second problem if there is P2P traffic in the network, what types of specific P2P applications in the network by creating three classes: Skype, BitTorrent, and SopCast. Besides, there are three processes: (1) capturing traffic and filtering only TCP and UDP protocol on any hosts that run IPv4, (2) manual feature selection from selected P2P applications, and (3) classing traffic with a decision tree J48 technique. A decision tree classifier is a popular technique. It provides the faster learning speed when comparing to other techniques. It is simple and easy to understand by end users. It supports for using with numeric-type feature, which our features are, and also supports well with the small amount of datasets. Therefore, the decision tree J48 is selected for this study.

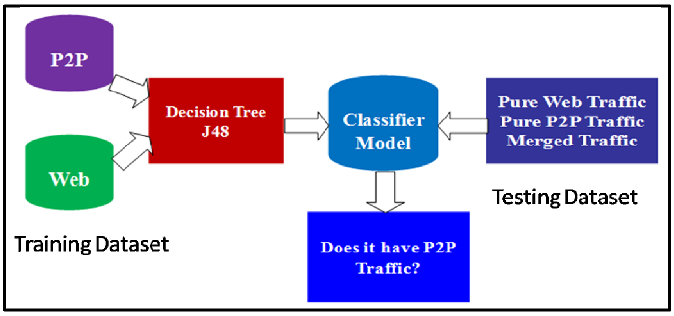

Fig. 10. Overview of the first-step experiment

Furthermore, to answer the first problem, we created models to predict the existence of P2P traffic in the network as in Fig. 10 above. The overview of first proposed work contained two kinds of datasets. The training dataset has two classes of P2P and Web containing 143 instances with 6, 5, 4, 3, and 2 features respectively by removing feature one by one based on the splitting attribute of previous model. The testing datasets are pure Web traffic, pure P2P traffics, and two kinds of merged traffics (single P2P traffic with background traffic and merged two P2P traffics with background traffic). Moreover, one trace file of testing set contained 24 instances; except the Web traffic contained 71 instances in order to balance with the amount of P2P traffics, and then applied to classifier model to detect P2P traffic in the network.

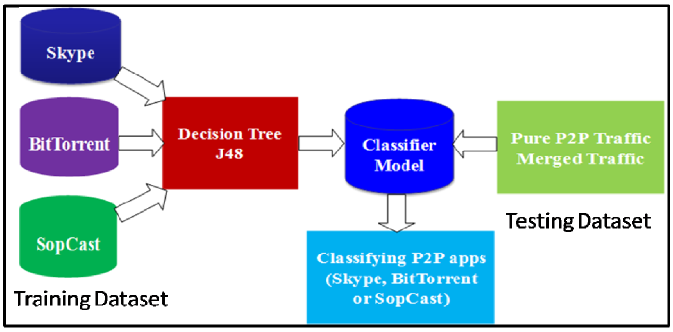

Fig. 11. Overview of the second-step experiment

To answer the second problem, we created other models as shown in Fig. 11 above. The models contain two datasets. The training dataset has three classes: Skype, BitTorrent, and SopCast, containing 72 instances, and 6, 5, 4, 3, and 2 atrributes respectively based on the splitting attribute from the previous model. The testing datasets include pure P2P traffic, single P2P traffic with background traffic, merged two P2P traffics, and merged two P2P traffics with background traffic. They were used to test the models to classify specific types of P2P applications from the background traffic of the Web traffic.

VI. Experiments and Results

To answer the first problem, we created the training set that contained two classes: P2P and Web, 143 instances with 6, 5, 4, 3, and 2 attributes respectively by dividing into five tests.The first test (six features) that tested on four types of unknown traffic, we found the accuracy is 100% ofclassifying P2P traffic in the network by usingthe ratio of incoming TCP with number of outgoing peers as Fig. 8.The processing time for the classificationwas about 55 milliseconds.

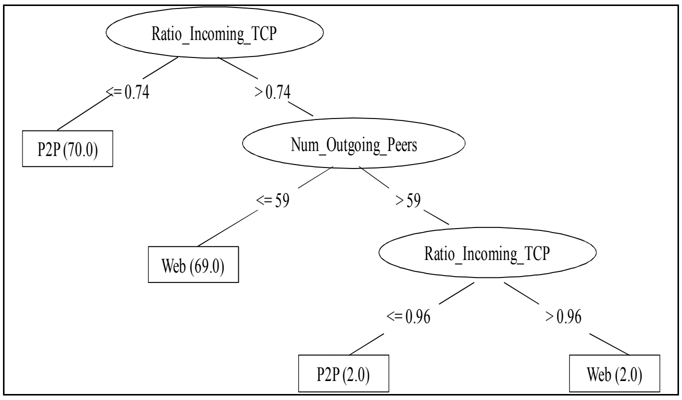

Fig. 12. Decision tree ofthe first test on six features for the first step

Next, we remove the feature that is in the top of the decision tree, so that the J48 can create another decision tree. We keep removingthe feature one by one until the training dataset contained two features to build models and to test the unknown traffic. Forthe first step of the experiment, we got the results as shown in TABLE VI. The six features were No. of incoming and outgoing peers, the ratios of incoming and outgoing TCP, and the ratios of incoming and outgoing UDP. The five features were the number of incoming and outgoing peers, the ratios of outgoing TCP, and the ratios of incoming and outgoing UDP. The four features were the number of incoming and outgoing peers, the ratios of outgoing TCP and UDP. The three features were the number of incoming peers, the ratios of outgoing TCP and UDP, and the two features were the number of incoming peers, and the ratio of outgoing UDP.

TABLE VI. SUMMARIZATION OF THE RESULTS ON THE FIRST STEP

Combination of Feature Selections	Selected Features from Models	Accuracy	Misclassi fication
6-Features	Ratio of incoming TCP and No. of outgoing peers	100%	0%
5-Features	Ratio of incoming UDP and No. of outgoing peers	100%	0%
4-Features	No. of outgoing peers and ratio of outgoing TCP	99.87%	0.13%
3-Features	Ratio of outgoing TCP and No. of incoming peers	99.75%	0.25%
2-Features	Ratio of outgoing UDP and No. of incoming peers	99.75%	0.25%

A. The Second Step of The Experiment

To answer the second problem, we created the training dataset that contained three classes: Skype, BitTorrent and SopCast, 72 instances with 6, 5, 4, 3, and 2 attributes respectively. Among these five tests, the 4th test (three features) produced highest accuracy of 90% by selecting the ratio of outgoing UDP as shown in Fig.9. The processing time for the classification was about 53 milliseconds.

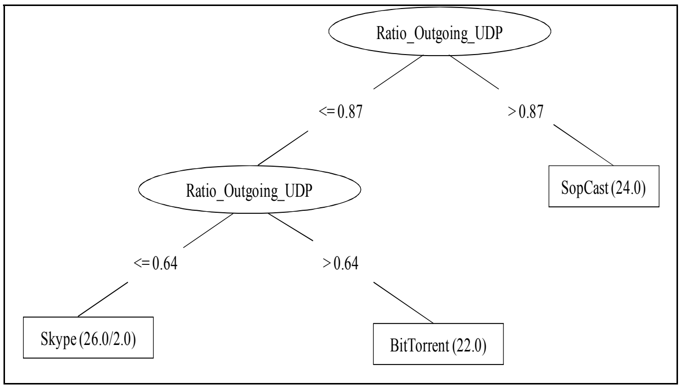

Fig. 13. Decision tree ofthe fourth test on three features for the second step

Similar to the first step, we keep removing the feature one by one until the training set contained only two featuresin order to find the better accuracy. For the second step of the experiment, we got the results as shown in TABLE VII.

TABLE VII. SUMMARIZATION OF THE RESULTS ON THE SECOND STEP

Combination of Feature Selections	Selected Features from Model	Accuracy	Misclassifi cation
6-Features	Ratio of incoming and outgoing TCP	77%	23%
5-Features	Ratio of incoming UDP and ratio of outgoing TCP	85%	15%
4-Features	Ratio of outgoing TCP	77%	23%
3-Features	Ratio of outgoing UDP	90%	10%
2-Features	No. of incoming and outgoing peers	78%	22%

VII. DISCUSSION AND CONCLUSION

Based on results in the experiments, we found that some applications are misclassified because the rate of each application was different and mostly depended on the network speed we captured the packets. If the rate is changed, our accuracy may be changed. Besides, we did not use shorter time interval because the collected data were captured from the small residential network with limited bandwidth.

This research tried to find out a technique that can classify P2P applications from the background traffic. Through experiments, we confirm that this technique could detect P2P traffic in the network with 100%accuracy, and could classify specific P2P applications from the background traffic of the Web traffic with 90% accuracy. However, the results here are limited to only these datasets that we captured from an ADSL network. For the future work, we may try other machine learning techniques in order to gain the higher accuracy.

REFERENCES

[1] N. Ben Azzouna, F. Guillemin, “Analysis of ADSL traffic on an IP backbone link,” Global Telecommunications Conference, 2003, p. 3742 – 3746. 2003.

[2] C. Wilson, A. Mislove, “P2P and BitTorrent,” CS 4700 / CS 5700 Network Fundamentals 2013,website: http://www.ccs.neu.edu/home/cbw/networks.html.

[3] S. Helal, C. Lee, D. Purandare,M. Zhang, J. David Mol, “The State of the Art of P2P Video Streaming,” 2014.

[4] T. Karagiannis, A. Broido, N. Brownlee, k claffy, M. Faloutsos, File-sharing in the Internet: “A characterization of P2P traffic in the backbone,” 2003, UC Riverside. p. 13.

[5] N. Basher, A. Mahanti, A. Mahanti, C. Williamson, M. Arlitt, “A comparative analysis of Web and Peer-to-Peer traffic,” in Proceedings of the 17th international conference on World Wide Web 2008. p. 287-296.

[6] Y.H Moon, J. Nah, J. Yoo, J. Jang, H. Kwon, S. Koh, J. Gu, “Apparatus-And-Method-For-Managing-P2P-Traffic,” 2009.

[7] S. Guha , N. Daswani, R. Jain, “An Experimental Study of the Skype Peer-to-Peer VoIP System,” IPTPS'06 - Proceedings of The 5th International Workshop on Peer-to-Peer Systems 2006: Santa Barbara, CA, USA.

[8] Government of the HKSAR, “PEER-TO-PEER NETWORK 2008,” The Government of the Hong Kong Special Administrative Region, 2008.

[9] Z. Yang, L. Li, Q. Ji, Y. Zhu, “Cocktail method for BitTorrent traffic identification in real time,”Journalof Computers, January 2012.

[10] N. Cascarano, F. Risso, A. Este, F. Gringoli, A. Finamore, M. Mellia, “Comparing P2PTV Traffic Classifiers,” Communications (ICC), 2010 IEEE International Conference on 2010: p. 1 – 6.

[11] T. Karagiannis, A. Broido, N. Brownlee, kc claffy, M. Faloutsos, “Is P2P dying or just hiding?,”[P2P traffic measurement] Global Telecommunications Conference, 2004. 3: p. 1532 - 1538.

[12] S. Sen, O. Spatscheck, D. Wang, “Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures,” ACM SIGCOMM Computer Communication Review 2004 Conference2004.

[13] S. Adibi, “Traffic Classification – Packet-, Flow-, and Application-based Approaches,” (IJACSA) International Journal of Advanced ComputerScience and Applications, 2010. 1: p. 6-15.

[14] T. Karagiannis, K. Papagiannaki, M. Faloutsos, “BLINC: multilevel traffic classification in the dark,” ACM SIGCOMM Computer Communication Review - Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications 2005. p. 229-240.

[15] C.C Wu, K.T Chen, Y.C Chang, C.L Lei, “Peer-to-Peer Application Recognition Based on Signaling Activity,” Communications, 2009. ICC '09. IEEE International Conference, 2009: p. 1 - 5.