A Hybrid Technique to Identify Peer-to-Peer Internet Traffic

Max Bhatia
University Institute of Engineering and Technology.
Panjab University, Chandigarh (India)

Sakshi Kaushal
University Institute of Engineering and Technology.
Panjab University, Chandigarh (India)

ABSTRACT
Over the past few years, Peer-to-Peer traffic has been consuming a lot of Internet traffic bandwidth and is still rising which brings great difficulties to network management. Traditional classification techniques such as port based and payload based have significant limitations. Hence, newer statistical approaches are adopted for P2P identification. P2P traffic uses both TCP and UDP protocols for communication. This paper provides a technique to identify P2P traffic, which focuses on significant TCP and UDP features and utilizes C4.5 decision tree algorithm.

Keywords
P2P traffic identification, TCP traffic, UDP traffic, C4.5 decision tree

1. INTRODUCTION
Network traffic identification is one of the major challenging task in past few years. The tasks of network engineers include meeting application performance, meeting bandwidth requirement of customers, managing bandwidth consumption, apply security rules, fault diagnosis, performing accurate accounting for billing, etc. In order to accomplish all these tasks, it is necessary to understand network traffic properties, which would help to improve network performance by developing better architecture. Therefore, identification of network traffic is of great importance. With the help of identification results, an enterprise or a service provider can protect and manage network resources [18].

There has been rapid growth in the Internet traffic over the past few years [25]. This is due to the fact that various types of applications are evolving rapidly and are being used. Traditional applications such as Http, SMTP, etc. no longer dominate the Internet traffic. Peer-to-Peer (P2P) traffic is dominating the Internet traffic over the past few years [3]. There has been significant trend in the recent years where P2P file sharing has been under use. The leading content which is shared by P2P applications are audio and video files which tend to be large in size [1], in addition to illegal file sharing. So, the wide use of P2P application nowadays, account for more than 60% of total network traffic [2][3][4][5]; which becomes the main user of network bandwidth.

Accurate P2P identification is crucial of other network activities. The first generation P2P traffic was easy to identify as it utilized well-known port numbers [6] and ISPs can easily identify and classify network traffic [7]. But, it can no longer be used now as current P2P applications tend to disguise their traffic by using arbitrary port numbers in order to thwart firewalls and network management applications.

Some applications even use TCP port number 80 for communication in order to hide P2P traffic [18][25]. To solve these problems, payload-based technique was utilized. This approach directly compared the stored signatures to packets from applications in order to classify them accurately. However, payload-based technique also confronts many limitations as new P2P protocols keep upgrading, direct inspection of payload breach privacy policies of organizations, many applications encrypting traffic or using proprietary protocols, *etc* [14]. Therefore, statistical based techniques are utilized to identify P2P traffic, which makes use of transport or network layer statistics such as packet length, inter-arrival time, *etc*. which is independent of inspection of packet payload and port numbers [16][17]. This paper proposes P2P traffic identification methodology which identifies both TCP and UDP protocols which are used for communication.

The rest of the paper is organized as follows. In the next section, some important related literatures are presented. Section 3 proposed the hybrid identification process. Section 4 presents performance evaluation and finally, conclusion is given in section 5.

2. RELATED WORK
Identifying the network flows by using statistical properties of network traffic is not new. Such methods assume that statistical properties of network traffic are unique for different applications and can be utilized to identify applications from each other [15]. The commonly used statistical features are flow duration, packet inter-arrival time, packet size, bytes transferred, number of packets, etc. Yan H *et al.* [8] utilized flow duration, the number of packets in the request (response) direction and the size of first (second) data packet in the request (response) direction to identify P2P and other applications. Chen *et al.* [9] focused on identifying P2P file-sharing traffic proposing 2 characteristics: discreteness of remote host (RHD) and discreteness of remote ports [RPD] to identify BT-like traffic. Mei-feng *et al.* [10] investigated and obtained traffic characteristics: Ack-Len ab and Ack-Len ba; which are the data volumes sent by the communication parties continuously and classified P2P and other network traffic using C4.5 decision tree in the early time when flow arrived. But, this technique is only suitable for TCP flow and cannot be applied for UDP flow. Wei-ming [11] identified P2P traffic which mainly focused on UDP traffic produced by P2P applications. They revealed three significant features of UDP traffic with respect to transport layer behaviors and packet size distribution, i.e. unique local port number, unique UDP protocol pair and two-point distribution of packet size.

The technique utilized by Mei-feng *et al.* [10] works only for TCP protocol to identify P2P traffic and cannot be used for UDP protocol which is also currently being used by newer P2P applications for communication. Hence, a hybrid technique is proposed in this paper which is suitable for identifying P2P application associated with both TCP and UDP flow. The process which utilizes TCP flow takes total data length which is sent by the peer before it receives the first acknowledgement (Ack) packet, as the characteristic. Here it is assumed that the two sides that communicate are A and B and uses Ack-Len ab to represent total data length which is sent from A to B before first Ack packet arrived. Conversely, Ack-Len ba is represented in same way. The process which utilizes UDP flow checks for three significant UDP characteristics: two unique UDP connection characteristics and one packet size characteristic for P2P traffic. These characteristics appear together only in UDP traffic produced by P2P software.

3. HYBRID TECHNIQUE TO IDENTIFY P2P TRAFFIC

The hybrid technique is the combination of 2 techniques which takes TCP and UDP separately to identify P2P traffic. Therefore, it has been divided into 2 processes: TCP process and UDP process as discussed below.

3.1 TCP Process:

For P2P applications utilizing TCP for communication, there is no distinction between client and server because both sides support download and upload simultaneously i.e. both sides transfer data to each other [10]. The data volume first sent by both sides to each other is large, to thousands of bytes. The data volumes first sent by peers are different for different P2P applications, but in general they are non-zero. After the connection is established between both the peers, one peer sends its data to the other. Then, after receiver has received the data, it will send its own data to the sender and piggybacking the acknowledgement. After the sender has received this data as well as acknowledgement, it goes on to send its new data and also piggybacking the acknowledgement. Therefore, in this process, one peer can download data from the other and provide the data it is having for the other peer.

Pseudo-code for TCP process:

```
//        Seq. no. → Sequence number
//        Ack. no. → Acknowledgement number
//           EoF → End of File
//           Src. IP → Source IP
//           Dst. IP → Destination IP

Begin
  For each stream/flow:
    loop(until EoF)
      Extract socket pairs: Src IP and port, Dst IP and port;
      Capture first Seq and Ack no. between socket pairs;
      Then, Compute:
            Data transferred by A = Seq(A) – Ack(B);
            Data transferred by B = Seq(B) – Ack(A);
       if(data transferred by both A and B)
             then, P2P identified
    end loop
End
```

Firstly, captured traffic is filtered to get TCP traffic and then its flow construction takes place. A flow is a collection of packets which meet specific flow specifications and researchers often use 5-tuples (source IP, source port, destination IP, destination port, protocol) to define a flow. Now, 2 TCP features of packets needs to be considered: Sequence number (Seq_no) and Acknowledgement number (Ack_no). After the connection is established, peer A sends the data it has for peer B. Here, data transferred by peer A is calculated as: $(Data)_A = (Ack\_no)_B - (Seq\_no)_A$; where $(Ack\_no)_B$ is the acknowledgement number sent by B to A and $(Seq\_no)_A$ is the sequence number of A. Similarly, data transferred by peer B is calculated as: $(Data)_B = (Ack\_no)_A - (Seq\_no)_B$. If it is observed that the first data volume transferred by both peers A and B is non-zero (thus reflecting download and upload behavior), then P2P is identified.

3.2 UDP Process:

For P2P applications utilizing UDP for communication, there exists three unique characteristics which do not appear together either in TCP or UDP traffic produced by non-P2P software [11]. These are:

1) Almost all UDP traffic of local host transfers by fixed port number.
2) Nearly all remote hosts utilize single port number to communicate with local host.
3) A couple of packet sizes are monopoly in UDP packet size distribution.

Wei-ming found that more than 99.5% UDP packets use fixed port number [11]. UDP packets which utilize different port numbers are the DNS packets which are utilized by P2P applications to obtain index servers' IP address. The justification that, nearly all remote hosts use single port number to communicate with local host, is that P2P protocol is designed to distribute traffic evenly in order to avoid overloading any peer. Hence, number of ports is limited to one in implementation of popular P2P software. Also, it is found that size of packets produced by P2P applications is relatively fixed. Most UDP packets in P2P traffic are used to request and answer; where request packets are small (72 bytes) while answer ones are large (1392 bytes).

Pseudo-code for UDP process:

```
//        Seq. no. → Sequence number
//        Ack. no. → Acknowledgement number
//           EoF → End of File
//           Src. IP → Source IP
//           Dst. IP → Destination IP

Begin
  For each stream/flow:
    loop(until EoF)
      Extract socket pairs: Src IP and port, Dst IP and port;
      Capture and store:  1) local host port no.
                          2) no. of ports used by remote peer
                          3) packet size
       if(local host use fixed port)
            if(remote peer use single port)
               if(packet size= 1392 or 72)
                     then, P2P identified
    end loop
End
```

To identify P2P traffic which utilize UDP for communication, firstly captured traffic is filtered to get UDP traffic and then its flow construction takes place. Now, 3 UDP characteristics, as mentioned, need to be examined for each flow. Only if these 3 characteristics are met, P2P traffic is reported to be found.

Fig. 1 depicts the whole procedure for identifying P2P traffic. Firstly, capturing of network packets take place by a software [26]. Then, packets are classified based on TCP and UDP protocol. Before examining the characteristics of both TCP and UDP individually, streams of packets have to be constructed. In case of TCP, if data volume first transferred by both the hosts (sender and receiver), then P2P is identified. In case of UDP, if all 3 characteristics (as mentioned) are met, then P2P is identified.

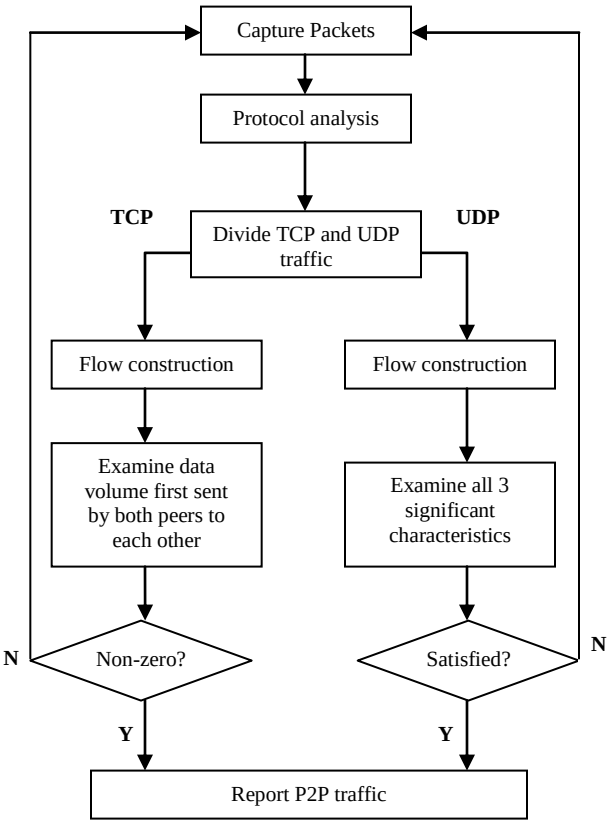

Fig. 1: Procedure for P2P traffic identification

4. SIMULATION and EVALUATION

In this section, the datasets that are used for simulation purpose and the results obtained by using them are discussed.

4.1 Data Preparation and Evaluation Criteria:

Two datasets are used in simulation study. The first one is downloaded from the Internet, namely Dataset 1. The second one is captured from our own campus area network, namely Dataset 2. Both the datasets have been classified into two categories: P2P and non-P2P; where non-P2P refers to all other kinds of traffic such as HTTP, FTP, SMTP, *etc*. Accuracy, Recall and Precision are used to evaluate the classification performance. Some parameters are defined as follows:

- True Positive (TP): Percentage of members of class X correctly classified as belonging to class X.
- True Negative (TN): Percentage of members of other classes correctly classified as not belonging to class X.
- False Positive (FP): Percentage of members of other classes incorrectly classified as belonging to class X.
- False Negative (FN): Percentage of members of class X incorrectly classified as not belonging to class X.

The Recall and Precision for a class is defined as:

$$D_{Recall} = \frac{TP}{TP+FN}$$

$$D_{Precision} = \frac{TP}{TP+FP}$$

where Recall is the proportion of samples of class X correctly classified as belonging to class X and Precision determines the proportion of actual samples of class X among those classified as class X. If classification has N types of applications, then accuracy is defined as:

$$D_{Accuracy} = \frac{\sum_{i=1}^{N}(TP_i)}{\sum_{i=1}^{N}(TP_i+FP_i)}$$

where Accuracy is overall effectiveness of classification, and reflects the predictive power of classification model.

4.2 Simulation Results:

This simulation study is based on Weka 3.6 platform [27], and C4.5 decision tree algorithm [12] is used for classification due to its good classification speed [16]. Platform for experiment is a PC with Windows 7 system, Intel Core i3 2.4 GHz CPU, DDR3 4GB memory.

Fig. 2, Fig. 3 and Fig. 4 shows classification accuracy, recall and precision when hybrid technique is applied on both datasets.

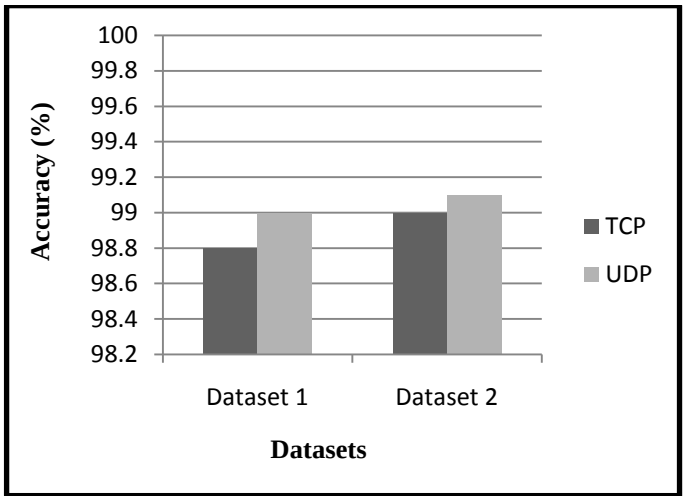

Fig. 2: Classification Accuracy of Dataset 1and Dataset 2

Fig. 2 shows that classification accuracy of TCP traffic is somewhat less than 99% for Dataset 1 and equals 99% for Dataset 2; whereas for UDP traffic, it is 99% for Dataset1 and greater than 99% for Dataset 2. The y-axis of Fig.2 represents the percentage value of accuracy and x-axis represents datasets (i.e, Dataset 1 and Dataset 2) used in simulation.

Fig.3 represents recall values of Dataset 1 and Dataset 2, where TCP traffic gives the value of 0.98 and 1, respectively; and UDP traffic gives value of 1 and 0.99, respectively.

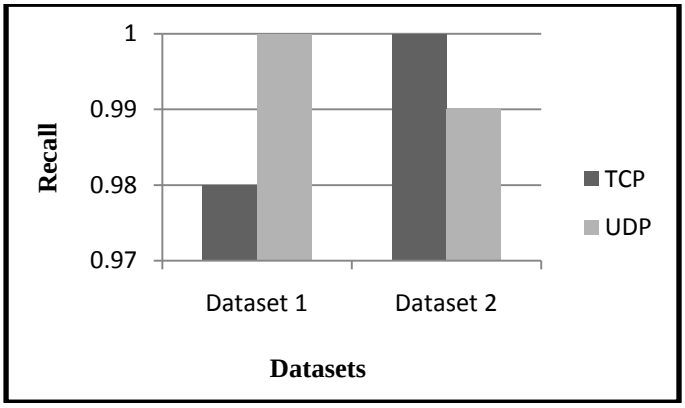

Fig. 3: Recall of Dataset 1and Dataset 2

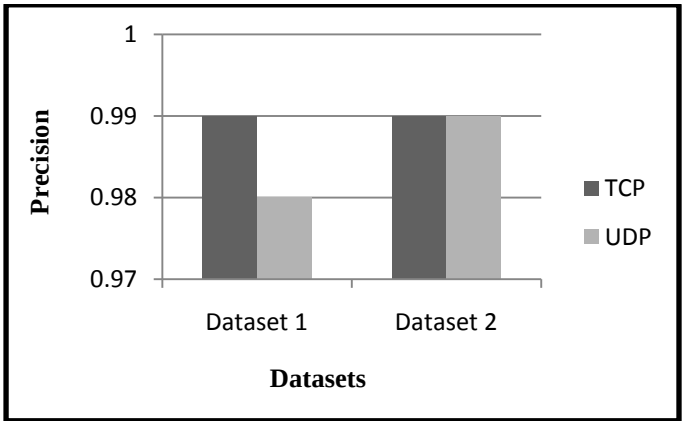

Fig. 4: Precision of Dataset 1and Dataset 2

Also, from Fig. 4 it can be seen that precision value of TCP traffic for both Datasets 1 and 2 is equal to 0.99 and precision value of UDP traffic for Datasets 1 and 2 are 0.98 and 0.99, respectively.

Fig. 5 shows accuracy comparison of proposed hybrid technique (represented by Exp-result) with the techniques used by Mei-feng and Jing-tao for TCP traffic [10] and Wei-ming for UDP traffic [11] (represented by Ref-result), to identify P2P traffic.

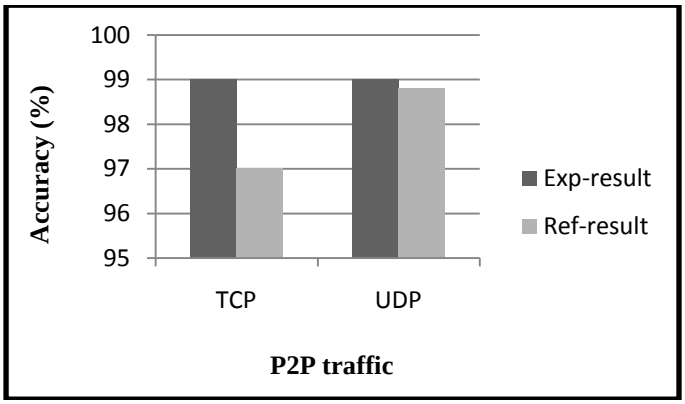

Fig. 5: Accuracy comparison of Hybrid Technique

The datasets used in this experiment are smaller and do not contain much of the P2P traffic and also they do not have mixed dataset which includes traffic from various P2P applications. But the technique used in this paper overcomes the shortcoming of technique used by Mei-feng and Jing-tao [10], which identifies P2P traffic only for TCP protocol; whereas the technique proposed in this paper identifies P2P traffic for both TCP and UDP protocols and hence rectifies its problem.

5. CONCLUSION

In this paper, a hybrid technique to identify P2P traffic which is classified by C4.5 decision tree algorithm into P2P and non-P2P traffic is presented. This technique is divided into two parts, one of which is called TCP process which works on TCP traffic and other is called UDP process which utilizes UDP traffic. TCP process classifies traffic based on characteristics which includes the data sent continuously by the communicating parties and if it is found that first data volume transferred by both communicating peers in a flow is non-zero, then P2P traffic is reported. On the other hand, if all three significant characteristics (unique local port number, unique UDP protocol pair and two-point distribution of packet size) of UDP which are unique for P2P traffic are satisfied, then P2P traffic is reported.

The experimental results produced by this proposed hybrid technique gives more than 99% accuracy and is able to work on both TCP and UDP traffic, which is utilized by different P2P applications for communication. It is evaluated on offline datasets. Also, datasets utilized are smaller and contained small amount of P2P traffic. Hence, future work includes evaluating this technique for online P2P identification and using larger as well as mixed datasets which includes traffic from various P2P applications.

6. REFERENCES

[1] Sen.S, Jia Wang, "Analyzing peer-to-peer traffic across large networks", Networking, IEEE/ACM Transactions on, Volume 12, Issue 2, April 2004 pp. 219–232.

[2] Li Dai, Jie Yang, Li Lin, "A comprehensive system for P2P classification**",** in 2nd IEEE International Conference, pp. 561-563, September 2010.

[3] Mehdi Mohammadi, Bijan Raahemi, Ahmad Akbari, Hossein Moeinzadeh, Babak Nasershari, "Genetic-based minimum classification error mapping for accurate identifying Peer-to-Peer applications in the internet traffic", Expert Systems with Applications, Vol. 38, pp. 6417-6423, June 2011.

[4] HuiLin Chu, HongBo Yi, XingMing Zhang, "A New P2P Traffic Identification Methodology Based on Flow Statistics", in 3rd IEEE International Conference, pp. 277-281, May 2011.

[5] Ram Keralapura, Antonio Nucci, Chen-Nee Chuah , "A novel self-learning architecture for p2p traffic classification in high speed networks", Computer Networks, Vol. 54, pp. 1055-1068, May 2010.

[6] A. Gerber, J. Houle, H. Nguyen, M. Roughan, S. Sen, "P2P The Gorilla in the Cable", in National Cable and Telecommunications Association (NCTA) 2003 National Show, Chicago, IL, June 8-11, 2003.

[7] S. Sen, J Wang, "Analyzing peer-to-peer traffic across large networks", in IEEE/ACM Transactions on Networking, vol.12, no.2, pp.219-232, 2004.

[8] Yan H, Chiu D M, Lui J C S, "Profiling and identification of P2P traffic", Computer Networks, 2009, 53(6): 849-863.

[9] Chen W Q, Gong J, Ding W, "Identifying file-sharing P2P traffic based on traffic characteristics", The Journal of China Universities of Posts and Telecommunications, 2008, 15(4): 112-120.

[10] Mei-feng SUN, Jing-tao CHEN, "Research of Traffic characteristics for real time online traffic classification", The Journal of China Universities of Posts and

Telecommunications, Computer Networks, vol. 18, no. 3, pp. 92-98, June 2011.

[11] Wei-ming Hong, "A Novel method for P2P traffic identification", Procedia Engineering, International Conference on Power Electronics and Engineering Application, vol. 23, pp. 204-209, 2011.

[12] Quinlan J R. C4.5: Program for machine learning. San Mateo, CA, USA: Morgan Kaufman, 1993.

[13] Min-huo HONG, Ren-tao GU, Hong-xiang WANG, Yong-mei SUN, Yue-feng JI, "Identifying online traffic based on property of TCP flow", The Journal of China Universities of Posts and Telecommunications, vol. 16, no. 3, pp. 84-88, June 2009.

[14] Mei-feng SUN, Jing-tao CHEN ,"Research of Traffic characteristics for real time online traffic classification", The Journal of China Universities of Posts and Telecommunications, Comuter Networks,vol. 18, no. 3, pp. 92-98, June 2011.

[15] Chun-Nan Lu, Chun-Ying Huang, Ying-Dar Lin, Yuan-Cheng Lai, "Session level flow classification by packet size distribution and session grouping", Computer Networks, vol. 56,no.1, pp. 260-272, Jan 2012.

[16] Nguyen, T.T.T., Armitage, G., "A survey of techniques for internet traffic classification using machine learning", in: IEEE Communications Surveys and Tutorials, vol. 10, no.4, Jan 2009.

[17] Callado, A.; Kamienski, C.; Szabo, G.; Gero, B.; Kelner, J.; Fernandes, S.; Sadok, D.,"A Survey on Internet Traffic Identification", in: IEEE Communications Surveys and Tutorials, vol. 11, no. 3, pp. 37-52, Aug 2009.

[18] Murat Soysal, Ece Guran Schmidt,"Machine Learning Algorithms for accurate flow based network traffic classification: Evaluation and Comparison", Performance Evaluation, vol. 67, no. 6, pp. 451-467, June,2010.

[19] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in Proc. Passive and Active Measurement Workshop (PAM2005), Boston, MA, USA, March/April 2005.

[20] 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.

[21] T. Karagiannis, A. Broido, N. Brownlee, K.C. Claffy, M. Faloutsos, "Is P2P dying or just hiding", in: IEEE Global Telecommunications Conference, GLOBECOM 04, 2004.

[22] S. Sen , C. Spatscheck, D. Wang, "Accurate, scalable innetwork identification of P2P traffic using application signature", in: 13th International Conference on World Wide Web, 2004.

[23] T. Karagiannis, A. Broido, M. Faloutsos, K.C. Claffy, "Transport layer identification of P2P traffic", in: 4th ACM SIGCOMM Conference on Internet Measurement, 2004.

[24] K. Wang, S.J. Stolfo, "Anomalous payload-based network intrusion detection", in: Lecture Notes in Computer Science, Springer, Berlin, 2004.

[25] John Hurley, Emi Garcia-Palacios and Sakir Sezer, "Classification of P2P and HTTP Using Specific Protocol Characteristics", Lecture notes in The Internet of the Future, 15th Open European Summer School and IFIP TC6.6 Workshop, EUNICE 2009, Barcelona, Spain, Proceedings, September 7-9, 2009.

[26] Wireshark, Available: http://www.wireshark.org/

[27] Weka, Available: http://www.cs.waikato.ac.nz/ml/Weka

/