ISSN:0975-9646

A Survey: Peer-to-Peer Traffic Identification technologies

Prof S. R. Patil[#1], Suraj S. Dangat[*2]

[#] *Dept. of Computer Engineering, SIT, Savitribai Phule Pune University*
309/310, Off Mumbai-Pune Expressway, Kusgaon BK, Lonavala, Maharashtra 410401, India

[*] *Dept. of Computer Engineering, SIT, Savitribai Phule Pune University*
At Post - Kiwale, Tal-Haveli, Dist - Pune, Maharashtra 412101, India

***Abstract*— The Peer-to-Peer (P2P) traffic plays a vital role in promoting the Internet applications. But, P2P has also caused network congestion and safety problems because of resource occupation (especially bandwidth). To ensure the network services which is provided by the network, it is necessary to have a control over the P2P traffic. So, it is necessary to identify this P2P traffic. There are various approaches available for it. In this paper we first briefly introduce P2P technology. Then, we have made a survey on the overall progress in P2P traffic classification technologies. Finally we outline the present research challenges and future developments.**

***Keywords*— P2P; Peer-to-Peer, Traffic identification, Traffic Classification technologies**

I. INTRODUCTION

A computer network or data network is a telecommunications network that allows computers to exchange data. In computer networks, networked computing devices pass data to each other along data connections. The connections (network links) between nodes are established using either cable media or wireless media. The best 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. Files can be shared directly between the different systems on the network without the need of a central server. Means, each computer on a P2P network becomes a file server as well as a client. With the extensive application of P2P technology, P2P applications take up a large amount of network bandwidth, which increase the burden of the network. According to the statistics, P2P applications account for 60% to 80% of total ISP business and become the largest consumer of network bandwidth. Therefore, the key to solve the problem of bandwidth congestion is that we limit the users who use large amount of bandwidth to protect those who use a small amount of bandwidth when the network resource constraints. On the contrary, when the network resources are available, we remove these restrictions so that each user can use the lines efficiently [1]. How to effectively control the network resources and how to effectively control the P2P traffic have become quite important. Therefore, P2P traffic identification is a key technology for effective control. The rest of this paper is structured as follow. In section 2, we briefly introduce Details of P2P. Section 3, we analyse the advantage and disadvantage of the mainstream identification technology for P2P traffic. Section 4, we discuss the Future research & developments in P2P traffic identification. Section 5, provides conclusion.

II. DETAILS OF P2P

A. *Overview of P2P*

First, P2P is a distributed computing model for sharing and managing of mass information resources in network. The main idea is that the status of all the nodes is full equal, and each node has dual role (Client and Server). In the traditional C / S network architecture, the Central Server collects all resources on the Internet. In C/S network architecture; it is very difficult to achieve transparent communication and abilities integration among Central Servers according to user needs. So, they (Servers) become a bottleneck for open network and capacity expansion. On the contrary, P2P network architecture does not exist central nodes (Servers) during media communication. Each node's status is equal and can be peer-to-peer communication. The advantage of this network structure is resources shared by P2P node each other. Resources is no longer focused on the Central Servers, but distributed in the edge of the P2P network nodes. P2P technology enables business system to evolve from centralization to decentralization. The architecture of P2P network has overcome the bottleneck caused by the concentration of nodes, reduced the cost of network's construction and usage, increased utilization of the system and network equipment.

B. *Problems caused by P2P application*

In recent years, P2Pnetwork technology has developed rapidly. There are variety of P2P applications are available for file sharing, sharing CPU resources, distributed storage, distributed collaboration environment, and so on [3]. These applications are very useful for the world. But, these P2P applications having certain properties like no centre, a loose, distributed properties and it arises some problems such as Bandwidth issues, Copyright issues and Security issues.

1) Bandwidth occupation: Statistics show that about 60 percent of the bandwidth is occupied by P2P applications and from these 60%, 80% of which were occupied by P2P file-sharing. But these P2P file-sharing users are very low in number. They are only 5% of the total number of Internet users. Because P2P has many characterizes, such as large flow, connection for a long time, automatic operation, regardless of time running, and so on. Therefore, P2P applications can take up more bandwidth. With the increase in the number of P2P users, network traffic will 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 increase network traffic. 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.

2) Copyright issues: P2P applications are useful for sharing information and software's. It may cause the copyright of data and software piracy issues. Despite the current Gnutella, Kazaa, and other P2P sharing software to promote its core server does not store any of the content of the protection of property rights. However, it is undeniable the prosperity of P2P software has accelerated the spread of the piracy and increased the difficulty of protection of intellectual property rights.

3) *Security issues:* Since P2P applications having certain characteristics like it is decentralized, it is uncontrolled, it pusses spontaneous behaviour and it has anonymous release character. The transmission route selected by the P2P application is vulnerable to different kinds of viruses. So, In P2P application, how to control the security of information is a big problem. These applications will unavoidable threat user privacy and network security.

Therefore, considering intensions of ISP and network security issues, it would like to be able to effectively identify P2P traffic. There are various techniques available for it. Following section gives details about it.

III. TECHNIQUES FOR PEER-TO-PEER TRAFFIC IDENTIFICATION

A. Port-based Classification

P2P Traffic classification by using port number is the simplest and traditional method. It identifies the application traffic by identifying the application type from the port number in the transport layer [4, 8]. For example, TCP port 80 is HTTP traffic, TCP port 1214 is Kazaa P2P traffic and so on. This approach is extremely easy to implement and it gives very little overhead on the traffic classifier. It was successful method because many traditional applications use port numbers assigned by or registered with the IANA. Now a day, such traditional port-based technique has become less accurate because of several reasons. These are, At First, Many applications no longer use fixed, predictable port numbers.

Means they use random ports. Second, some P2P applications use dynamic ports which are not known in advance [9]. Table.1 gives port numbers of commonly used P2P protocols.

TABLE I
PORT NUMBERS OF COMMONLY USED P2P PROTOCOLS

Protocol	Transport Layer Protocol	Default Port No.
BitTorrent	TCP	6861-6889
Edonkey	TCP / UDP	4661/4665
eMule	TCP	4661-4662
Fasttrack	TCP	1214
Gnutella	TCP	6346-6347
MP2P	TCP	41170
Thunder	TCP	3076-3077
WinMax	TCP / UDP	5690
KazaA	TCP	1214/80
Freenet	TCP	19114/8081
Napstar	TCP	5566/6666/6677/6699-6701
Skype	TCP	80/443/1024

B. Payload-based Classification

To remove the drawbacks of port-based classification method, several payload-based analysis techniques have been proposed [5-10, 12]. Most protocols contain a protocol specific string in the payload (namely signatures in some literatures) that can be used for identification. These strings are public information and can also be determined by examining a number of network traffic traces. Subhabrata et al. [10] presented an analysis of a number of P2P protocols and their signatures. For example "http/1" corresponds to the application HTTP services. By comparing every packet payload with a pool of previously determined signatures, this method can identify application traffic more accurately than the traditional method.

DPI was first pointed out by Karagiannis and other scholars [12], and then Sen and other scholars analyses the characteristics of 5 kinds of P2P protocols (Gnutella, donkey, DirebtConnect, BitTorrent and Kazaa) [10], proposed feature based on P2P traffic detection methods, and verify that the false positive of the method is less than 10%. Thomas Karagiannis collected payload keywords from eight kinds of popular P2P protocols [11]. Based on application layer signatures, Holger Bleul et al [10] proposed a simple, effective, flexible P2P traffic measurement method, and the method is easily extended to the new P2P applications. [5], introduced a kind of bait nodes, and analysed the traffic from Japanese popular P2P system Winny through application layer signature. Studies show that these approaches work very well for today's Internet traffic, including P2P flows. In fact, commercial bandwidth management tools use application signature matching to enhance robustness of classification. The main benefits include: high accuracy and robustness, and has a good classification functions.

However, there are some disadvantages too. Firstly, these methods identify only P2P traffic for those signature is known and it is unable to classify any other traffic, but maintaining the updates of signatures. Secondly, some newer-generation P2P applications are incorporating various strategies to avoid detection. Third, these techniques typically require increased processing and storage capacity. In order to solve these problems, artificial intelligence and data mining methods are introduced to DPI. Table.2 shows the signatures of some P2P applications.

TABLE III
SIGNATURES OF P2P APPLICATIONS.

Protocols	Signatures
BitTorrent	"0x13Bit"
eDonkey	0xe319010000
Gnutella	"GNUT", "GIV"
Kazaa	"X-Kazaa"
MP2P	Go!!,MD5,SIZ0x20

C. Feature-based Classification

Given the shortcomings of port- and payload-based approaches for detecting P2P traffic, the research community started developing the new techniques which are less dependent on particular individual applications, but focused on capturing and extracting commonalities in the behaviour of P2P applications which is based on layer-3/layer-4 information. We refer feature-based technique. This kind of approach is to classify traffic based on the analysis of non-stationary i.e. ''hidden" transition patterns of traffic flows. Such nonlinear properties cannot be affected by payload encryption or dynamic port change and hence cannot be easily masqueraded. These methods provide a promising alternative for traffic classification.

D. Hybrid Classification Method

There are also some hybrid P2P traffic classifiers available. This classifier includes most of the proposed methods for improving classification accuracy. For example, [2] proposes a novel two-stage P2P traffic classifier. In the first stage, it uses an algorithm which is fast, light-weight. It exploits the temporal correlation of flows to clearly separate P2P traffic from the rest of the traffic. In the second stage, it uses the signature extraction algorithm. This algorithm is used to accurately identify signatures of several known and unknown P2P protocols.

IV. FUTURE CHALLENGES & DEVELOPMENTS IN PEER-TO-PEER TRAFFIC IDENTIFICATION TECHNIQUES

For analyzing the working principal of P2P applications the enough amount of study on P2P traffic must be done. The detailed understanding of P2P networks will helpful for improving design of these applications and for evaluating the impact of these applications on network. However, P2P applications are even harder to identify than the traditional network applications because of complexity of these applications. Furthermore, some new P2P applications will come in future, so for identification and classification of traffic of these application a new novel approach need to be develop.

Deeper understanding of different flow characteristic and then combining various flow characteristics organically is also having lot of research.

Uses of ideas of artificial neural networks to identify P2P traffic are the latest research direction. The successful growth of artificial intelligence has put a great challenge of incorporating this new field in P2P traffic identification. Use of neural networks can also be effective in this field.

V. CONCLUSION

This survey paper explains various techniques available for classifying Peer-to-Peer traffic. Each of these techniques has their own advantages and disadvantages. In addition, all of the techniques have one main drawback that they are not feasible for real-time classification in high speed ISP because of some reasons. These are, At First, the algorithms used by most of the techniques are time consuming and it is applied to every flow which is seen by classifier and thus, traffic rate becomes extremely difficult. Second, Most of the above technique could not identify individual applications; they are just designed to identify high level application classes. Third, if some new P2P application comes in future then none of above technique is able to identify the new application traffic. So, a new generic technique needs to be developed that can help us to identify P2P traffic of any applications. This requires a detailed knowledge of already existing techniques and their loopholes. So that researchers can propose ideas to overcome the weakness and develop a much stronger approach to deal with it.

ACKNOWLEDGMENT

We are thankful to Mr. T. J. Parvat and Mr. P. J. Pandit, for the encouragement and the support they have extended to us for completing this review paper. We are also thankful to the Mrs. S. R. Patil for their support to make this paper analysis good as it is.

REFERENCES

[1] Jingyu Wang, Jiyuan Zhang, Yuesheng Tan," Research of P2P Traffic Identification Based on Traffic Characteristics" Inner Mongolia University of Science &Technology, Baotou, China, IEEE 2011.

[2] R. Keralapura, A. Nucci, C.N. Chuah, "A novel self-learning architecture for p2p traffic classification in high speed networks,"Computer Networks, in press, doi:10.1016/j.comnet.2009.10.009.

[3] YANG Fa- yi. Development of P2P Networks [J]. Network Computer Knowledge and Technology，2007 (01), pp. 80、118.

[4] W. Sears, Z. Yu, and Y. Guan, "An adaptive reputation based trust framework for Peer-to-Peer application," Proc. the International Symposium on Network Computing and Applications (NCA 2005),IEEE Press, July 2005, pp. 13-20, doi:10.1109/NCA.2005.6.

[5] H. Bleul, E.P. Rathgeb,"A Simple, Efficient and Flexible Approach to Measure Multi-Protocol Peer-To-Peer Traffic," Proc. IEEE International Conference on Networking (ICN' 05), IEEE Press, 2005. pp. 606-616, doi: 10.1007/b107118.

[6] S. Ohzahata, Y. Hagiwara, M. Terada, et al, "A traffic identification method and evaluations for a pure P2P application," Proc. 2005 Passive and Active Measurement (PAM'05), Springer-Verlag, Mar. 2005, pp. 55-68, doi: 10.1007/b135479.

[7] P. Haffner, S. Sen, O. Spatscheck, "Acas: Automated construction of application signatures," Proc. 2005 ACM SIGCOMM workshop on Mining network data, ACM Press, Aug. 2005, pp. 197-202, doi:10.1145/1080173.1080183.

[8] S. Sen and J. Wang, "Analyzing peer-to-peer traffic across large networks," IEEE/ACM Transactions on Networking, vol. 12, pp. 219-232, April 2004.

[9] T. Karagiannis, A. Broido, N. Brownlee, K.C. Claffy and M. Faloutsos, "Is P2P dying or just hiding," Proc. IEEE Globecom 2004, IEEE Press, Nov. 2004, pp. 1532-1538, doi: 10.1109/GLOCOM.2004.1378239.

[10] 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.

[11] T. Karagiannis, A. Broido, M. Faloutsos, "Transport Layer Identification of P2P Traffic," Proc. the 4th ACM SIGCOMM conference on Internet measurement, ACM Press, Oct. 2004, pp. 121-134, doi: 10.1145/1028788.1028804.

[12] T. Karagiannis, A. Broido, N. Brownlee, et al, "File-sharing in the internet: A characterization of p2p traffic in the backbone," 2003, UCRiverside, Technical report.