P2P Traffic Identification Method Based on Traffic Statistical Characteristics

Bing Hu and Zhixin Sun(✉)

Key Laboratory of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Ministry of Education, Nanjing, China
sunzx@njupt.edu.cn

Abstract. Different from traditional P2P traffic identification methods based on keywords or well-known ports, this paper presents a new identification method of P2P traffic based on the most basic characteristics of P2P protocol. We compare and analyze P2P traffic and traditional C/S traffic using statistical analysis techniques, and obtain two statistical characteristics of P2P traffic: continuity and multi-connectivity. In addition, the mechanism of sliding window is introduced in the quantification of the statistical characteristics to establish a P2P traffic identification model. Finally, we develop a P2P traffic identification emulation system based on the proposed model, and the experiment results reveal that this new model can identify known and unknown P2P traffic effectively.

Keywords: Traffic identification · Statistical characteristic · Sliding window · Peer-to-Peer (P2P)

1 Introduction

The concept of Peer-to-Peer (P2P) network service is first proposed by Steve Crocker in 1969 [1]. Each peer in the P2P network can communicate or share files with any other peers independently and directly. In recent years, P2P technology has been paid widespread attention. The main application areas of P2P technology include information sharing, real-time communication, online games, financial service, information retrieval and network storage [2].

P2P technology is not only able to provide fast and efficient file sharing and low-cost and high-availability computing and storage resources sharing, but also holds a strong network connectivity as well as more direct and flexible information communication capability. However, the benefits of P2P technology have also brought some new problems. (1) Due to the high transmission speed of P2P, its users occupy 60%–80% of the network bandwidth, while leaving limited bandwidth for non-P2P users, which can easily cause the obstruction of enterprise and ISP bottleneck link. (2) The vast volume of P2P downloads consumes large amounts of network bandwidth, which prevents other users from accessing the network resources normally. (3) Continuous high-speed downloads of P2P users not only increase the load of network equipment, but also easily cause link congestion at peak time.

D.-S. Huang et al. (Eds.): ICIC 2017, Part I, LNCS 10361, pp. 260–272, 2017.
DOI: 10.1007/978-3-319-63309-1_24

Hence, it is urgent to identify P2P traffic and control it properly [3]. In this paper, based on the analysis of the characteristics of P2P traffic through experiments, we take the statistical characteristics of P2P traffic, continuity and multi-connectivity, as the basis of P2P traffic identification model to identify P2P traffic with known and unknown protocols. Finally, the validity of the proposed method is verified by developing a P2P traffic identification system.

The main contributions of this paper are highlighted as follows:

(1) From the perspective of the monitoring host, we analyze the characteristics of P2P traffic through experiments and the continuity and multi-connectivity of P2P traffic are selected as the basis of traffic identification model.
(2) A multi-connectivity quantization method is proposed based on an improved address aggregation method, which not only counts the number of IP address, but also saves a lot of space.
(3) According to the proposed traffic identification model, we develop a P2P traffic identification system to identify P2P traffic.

The rest of this paper is organized as below. Section 2 illustrates the related work. In Sect. 3, we describe the analysis of P2P traffic statistical characteristics. The methods for identifying P2P traffic and the strategies of quantifying its traffic characteristics are illustrated in detail in Sect. 4. In order to verify the effectiveness and practicality of the identification method, Sect. 5 presents a simulation experiment. Conclusions are provided in the last section.

2 Related Work

During the past decade, many approaches employ statistical methods to identify the P2P traffic are proposed. Reference [4] proposes a method based on the characteristics of the transmission layer communication mode of P2P application, such as the characteristics of communication protocol in transport layer and the relationship between ports and destination IP, to identify P2P traffic. More depth analysis of various kinds of traffic (such as P2P, WEB and mail) are carried out in [5]. According to the basic characteristics of various kinds of traffic, a new traffic classification and identification method, called BLINC, is proposed [5]. BLINC takes the network host as the research object, and the location analysis of network host is carried out from three aspects: (1) Social level, revealing the popularity of the monitoring host via the number of hosts that communicate with it. (2) Functional level, the role a host plays, server or client. (3) Application level, the communication characteristics of the host in transport layer. Reference [6] proposes a method that uses fundamental characteristics of P2P protocols, such as a large network diameter and the presence of many hosts acting both as servers and clients, without any application-specific information, to identify both known and unknown P2P protocols in a simple and efficient way. Reference [7] uses the statistical characteristics of P2P traffic, such as connection success rate and connection response success rate, to identify and classify P2P traffic. Based on the method proposed in [4], Ref. [8] combines all the characteristics of P2P traffic in [4] with a new characteristic, DNS query log feature, to create P2P hosts feature set. And then use

flexible neural tree (FNT) model to identify P2P hosts and traffic. Reference [9] presents a novel approach which relies on counting some special traffic that are appearing frequently and steadily in the traffic generated by specific P2P applications to identify P2P traffic at a fine-grained level. Besides, it is also able to identify encrypted traffic in real-time. Reference [10] develops a framework named CUFTI (Core Users Finding and Traffic Identification) to identify and manage P2P traffic of core users, i.e. long-lived peers. The model employed the direction and payload length of the control packets at the beginning of the traffic as the characteristics to perform traffic identification. Furthermore, the model can be employed for real-time traffic identification.

According to the related works, we can find that identification method based on traffic characteristics is an effective way for identifying P2P traffic with known P2P protocols and unknown P2P protocols. In this paper, we propose a new P2P traffic identification approach based on the statistical characteristics of P2P traffic, which can effectively identify known and unknown P2P traffic.

3 Characteristics of P2P Traffic

Due to the difference between P2P network and C/S network, their traffic also shows many differences. Compared with the traditional C/S traffic, P2P traffic presents many characteristics, such as higher transfer speed, heavier traffic and occupying more bandwidth. We analyze the current popular multiple P2P application software with using Ethereal to capture P2P traffic, and with the help of the statistical function of Ethereal to study the P2P traffic deeply. Then, compared with normal applications traffic, two important statistical characteristics of P2P traffic are found, strong continuity and multi-connectivity, which will be elaborated in the next.

3.1 Continuity

Network traffic of the host computer which is running P2P applications is not necessarily all P2P traffic. In order to distinguish the normal application traffic and the P2P application traffic, the average packet length of P2P traffic is analyzed. We use the "IO Graphs" analysis method of Ethereal analysis tool to analyze the byte rate of the traffic involved in the monitoring host. The average packet length of the eMule traffic (P2P traffic) with the time interval 0.1 s and 1 s is illustrated in Fig. 1, respectively. And Fig. 2 shows the results of the Thunder analysis. Thunder is a popular P2P download software in China. As can be seen from the figures, with the increase of analysis granularity, the average packet length of P2P traffic tends to be a fixed large size.

The average packet length of web traffic (non-P2P traffic) is showed in Fig. 3. It can be seen by comparing Figs. 2 and 3. When the analysis granularity is small (such as 0.1 s), the average packet length of both Thunder and web traffics fluctuates greatly with time. However, when the analysis granularity is increased to 1 s, the average packet length of web traffic still fluctuates greatly, and sometimes tend to be 0, but the average packet length of P2P traffic is stable.

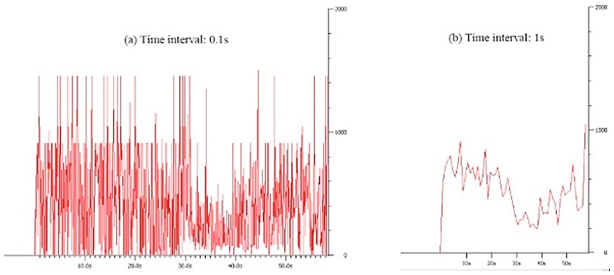

Fig. 1. Average packet length of eMule traffic.

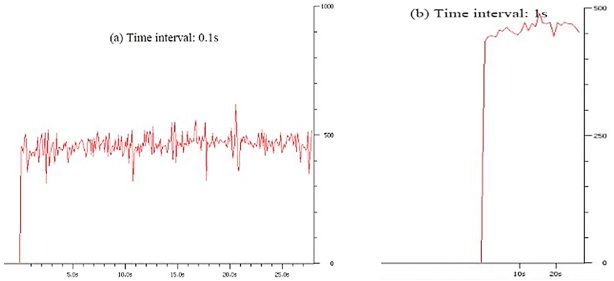

Fig. 2. Average packet length of Thunder traffic.

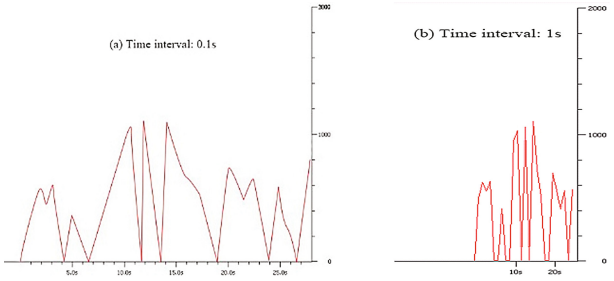

Fig. 3. Average packet length of web traffic.

The analysis results of the average packet length for FTP traffic (non-P2P traffic) are shown in Fig. 4. Compared with Fig. 1, average packet length of FTP traffic exhibits the same characteristics as that of eMule traffic. Therefore, FTP traffic (non-P2P traffic) and eMule (P2P traffic) cannot be distinguished only based on the characteristic of the average packet length.

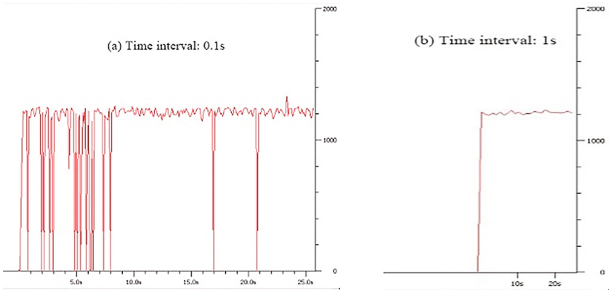

Fig. 4. Average packet length of FTP traffic.

3.2 Multi-connectivity

One of the most important characteristics of a P2P network is its strong connectivity. Each peer in the P2P network can communicate directly with any other peer, that is, any peer act as a client to obtain services from other peers while also serving as a server to provide services to other multiple peers. Additionally, almost all of the current P2P protocols support multi-thread, multi-point and broken point download. Hence, a peer can obtain the different parts of the same service from several other peers at the same time to improve the information transmission rate.

We still utilize Ethereal to capture the P2P traffic of the monitoring host and further use its "conversations" method to establish a "time-connection" table for each kind of traffic. The table records the average number of connections of various traffic, which indicates the average number of hosts that communicated with the monitoring host at different times. As Figs. 5(a) and (b) show, when the monitoring host is downloading files using Thunder or eMule, the average number of connections remains high. Moreover, when P2P peer is no longer downloading files, the number of connections is still high with some tens generally. The connections are mainly the query messages sent by other P2P peers to ensure the connectivity of P2P network. However, under normal circumstances, except a few servers, the number of connections for the traditional C/S network host is generally not very high. In Figs. 5(c) and (d), whether it is FTP downloads or web browsing with 15 web pages open, the number of hosts that connected with the monitoring host is very small with the range being from 5 to 15. Unlike P2P network, in the C/S network, client hosts only communicate with the server rather than with other client hosts.

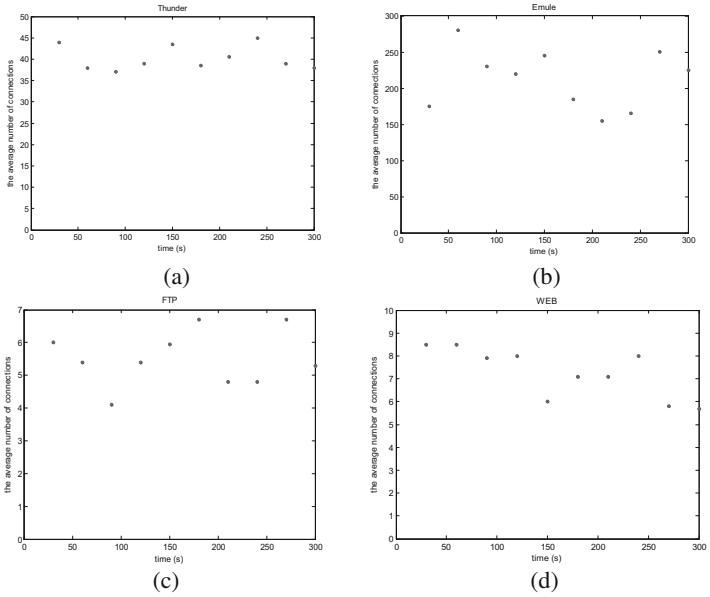

Fig. 5. Average number of connections.

4 Proposed P2P Traffic Identification Method

In this section, a P2P traffic identification method is proposed, which is based on the statistical characteristics of P2P traffic and the sliding window mechanism. The symbols used in this paper are illustrated as follows.

ti: Basic unit time for collecting the raw data of the traffic.

$f_t = (f_{1,t}, f_{2,t})$: The statistical state of the network traffic during the unit time t; $f_{1,t}$: quantized value of the traffic continuity; $f_{2,t}$: quantized value of the traffic multi-connectivity.

W_f: The size of sliding window.

TB_{ti}: The total number of bytes obtained by monitoring the host traffic during the unit time *ti*.

Statistical characteristics of P2P traffic are descriptive statistical characteristics, which need to be quantified to be used in the identification model. In addition, the quantized value can effectively reflect the strength degree of the traffic characteristics. In this paper, the quantized value of traffic characteristics is $f_{i,t} \in [0, 1],\ i = 1, 2$. The closer the value is to 1, the more obvious the corresponding characteristics. For example, the continuity of the traffic of $f_{1,t} = 0.8$ is stronger than that of $f_{1,t} = 0.6$.

4.1 Application of Sliding Window in Quantification of Traffic Characteristics

Quantifying characteristics of P2P traffic are quantifying the statistical information of specific characteristic within a unit time. However, use only the original statistics

information of the traffic during the current unit time can hardly describe the statistical characteristics of the current traffic accurately, and it is most likely to lead to a great leap in the statistical characteristics of the traffic. In order to describe the change of the network traffic accurately, and to ensure that the quantized value of the traffic characteristics are still continuity, sliding window mechanism is introduced into the quantification of P2P traffic characteristics. Therefore, the quantized values for each characteristic of P2P traffic are not determined by the original statistics information about the current window, but the integrated information of "W_f sending windows" in sliding window mechanism.

Suppose that a unit time is a window. All of the original statistics information from the current time backward through the W_f units of the statistical time is processed through a specific quantification strategy to obtain the quantized value for the characteristics of the current traffic. The aforementioned is the basic idea of characteristic quantification based on sliding window mechanism. Here, we take the quantification of traffic continuity for example to illustrate the application of sliding window mechanism in the quantification of traffic characteristics. Assume that the size of the statistics window is 4, i.e. $W_f = 4$, and the specific steps for the quantification strategy of the traffic continuity based on sliding window mechanism are described in Fig. 6.

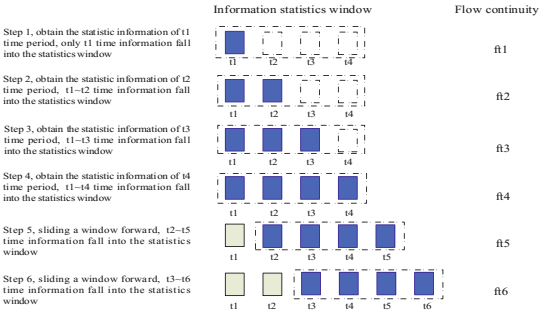

Fig. 6. Procedure of quantifying flow continuity.

(1) Initially, as step 1 shows, though the window size is 4, $t1$ is the only effective window. We can only get the statistical value TB_{t1} during the unit time $t1$, so the quantification value of the traffic continuity $f_{1,t1}$ is only decided by TB_{t1}, and the $\mathrm{TB}_{ti},\ i = 2, 3, \cdots$ is 0 by default. Then defining $f_{1,t1} = \begin{cases} 1, & TB_{t1} > 0 \\ 0, & TB_{t1} = 0 \end{cases}$. Sliding window is not moved in this step.

(2) As shown in step 2 and 3, TB_{t2} and TB_{t3} are collected during the unit time $t2$ and $t3$ respectively. The quantification value of the traffic continuity is determined by the windows that have already collected its information and the $\mathrm{TB}_{ti},\ i = 4, 5, \cdots$ is 0 by default. So the quantifying formula is

$$f_{1,ti} = \frac{\sum_{n=1}^{i} TB_n}{i},\ i = 2,3;\ TB_n = \begin{cases} 1, & TB_{tn} > 0 \\ 0, & TB_{tn} = 0 \end{cases},\ n = 1,2,3,\cdots$$

After the step 3, the window is still not fully occupied, so sliding window keep the same.

(3) As step 4 shows, the statistics of the fourth window is obtained and the four windows are all occupied, so the quantification value $f_{1,t4}$ is determined by all the statistical results of 4 time spans in the sliding window. Then defining

$$f_{1,t4} = \frac{\sum_{n=1}^{4} TB_n}{4};\ TB_n = \begin{cases} 1, & TB_{tn} > 0 \\ 0, & TB_{tn} = 0 \end{cases},\ n = 1,2,3,\cdots$$

(4) Information statistics window should move forward with a unit of time window to obtain the statistics of the fifth window, so the quantifying formula is

$$f_{1,t5} = \frac{\sum_{n=5-W_f+1}^{5} TB_n}{W_f};\ TB_n = \begin{cases} 1, & TB_{tn} > 0 \\ 0, & TB_{tn} = 0 \end{cases},\ n = 1,2,3,\cdots$$

(5) Then, continue to move forward a window to vacate a new window for obtaining the statistics of the following time span.

4.2 Quantification of Traffic Continuity

P2P traffic continuity represents the continuity of the network traffic that the P2P host participated in. If the network traffic of the monitoring host is not empty in a certain period of time, the host has a strong traffic continuity. Otherwise, if the paroxysm or return to 0 phenomenon occurs in the network traffic of the monitoring host, it means the traffic continuity of the host is weak. According to the research in Sect. 2, P2P peers have no difference from common network hosts when they do not participate in resource downloading or information sharing, in this case, the continuity of their network traffic should be very weak. In contrast, strong continuity of network traffic will appear when the hosts participate in resource downloading or information sharing.

If the total network traffic of the monitoring host is 0 in a unit time, it is considered that the host's traffic appears return to zero phenomenon in the time unit. Otherwise the host's traffic does not return to zero. Calculate the proportion of the time windows that does not return to zero from the current time back to the W_f time units. This ratio is the quantized value of the traffic continuity. As time forwards, statistical windows move forward and the statistics of those windows is always the traffic statistics of the latest W_f time units. Then, the algorithm of calculating $f_{1,t}$ is shown in Algorithm 1.

Algorithm 1

temp=0;
for the size of statistics window is W_f
  if $TB_{ti} \neq 0$
    temp++;
  end if
end for
$f_{1,t} = \frac{temp}{W_f}$

4.3 Quantification of Traffic Multi-connectivity

Compared with the traditional web services clients, P2P peers have the characteristics of multi-connectivity. In a unit time, the amount of the hosts that communication with the monitoring host can effectively reflect the multi-connectivity property of P2P traffic. In this paper, we use the number of IP addresses that communicate with the monitoring host to represent the multi-connectivity of the traffic.

Since the quantification of multi-connectivity only needs to record the number of IP addresses that communicate with the monitoring host, and do not need to know the specific IP address. We propose an improved address aggregation method which can not only count the number of IP addresses, but also save a lot of space.

We divide the IP address into K segments and each segment take up a space of L bits. Set up a hash mapping table with a size of $K \times 2^L$ bits. As Fig. 7 shows, divide IPv4 address into six domains: a, b, c, d, e, f, where five domains occupy 6 bits respectively, the remaining domain only 2 bits. So we can set up a 6×64 hash mapping table, using the decimal value mapping function to map the 6 domains respectively. Since the IPv4 address is 32 bits, there are only 4 valid spaces in the last column of the hash mapping table and the remaining 60 spaces will never be mapped to. Therefore, the actual valid space of the hash mapping table is

$$hashlength * hashwidth + ((oxooo1) \ll (32\%hashwidth)) - hashlength$$

where hashwidth indicates K and hashlength is 2^L. The result that the IPv4 address 10.10.10.1 expressed in dotted decimal notation form through hash mapping is represented in Table 1.

For a packet received within a unit time, if its source IP address or destination IP address is the address of the monitoring host, the six domains of the corresponding destination IP address or source IP address are mapped into the address aggregation mapping table, at the same time, add 1 to the corresponding domain value in the address aggregation mapping table respectively. The aggregation degree of the corresponding domain of the hosts that communicate with the monitoring host is recorded by the domain value. Then, the algorithm of quantifying the multi-connectivity of network traffic by using address aggregation technology is showed in Algorithm 2.

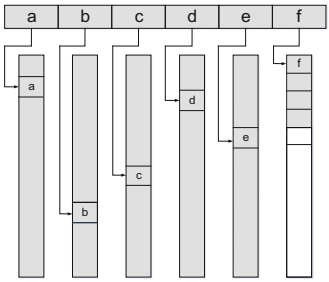

Fig. 7. hashTable: address aggregation mapping table.

Table 1. An example for address aggregation hash mapping.

IP address in dotted decimal notation form	10.10.10.1					
Binary form	00	001010	000010	100000	101000	000001
Hash value	0	10	2	32	40	1
Corresponding domain	f	e	d	c	b	a

Algorithm 2

temp=0;
k=0;
for the size of statistics window is W_f
k=the number of the nonzero filed in the table hashTable during the current unit time ti ;
temp=temp+k;
end for

$$f_{2,t} = \frac{temp}{hashlength * hashwidth + ((oxooo1) << (32\%hashwidth)) - hashlength}$$

P2P traffic identification and control system only record the aggregation information of the hosts that communicate with the monitoring host in the latest W_f unit time, and the statistics of the traffic multi-connectivity are carried out in accordance with Algorithm 2 during every time unit. Since the traffic multi-connectivity of the monitoring host is from the traffic statistics in the latest W_f time units, it basically reflects the real-time changes of the current traffic. Furthermore, the system only needs to keep W_f address aggregation mapping table with a fixed size for every host in the local network to achieve the statistics of traffic multi-connectivity with saving a lot of space. On the other hand, the realization of the method is very simple and effective.

5 Emulation

In order to verify the effectiveness of the proposed identification method, we develop a P2P traffic identification emulation system using VC 6.0. We install a number of different P2P software on a monitoring host, and make the host become a node in the P2P network. The P2P traffic identification emulation system is also installed on the monitoring host. The system will identify and analyze the network traffic when the host runs different P2P software. In this section, we select three popular P2P applications as the representative of P2P traffic, namely: eMule, Thunder and Vagaa. Then continue to run 24 h on the monitoring host. Similarly, all kinds of non-P2P applications are run randomly on the monitoring host, such as, FTP download and web browsing, the running time is also 24 h. The P2P traffic identification emulation system captures and stores all the network traffic generated during the application runs at the experimental network export, and sets the capture data into experimental data set. The system combines the two features mentioned above to form a feature vector, and uses the support vector machine algorithm to identify the P2P traffic.

In the system, users can set up the IP address of the monitoring host, analysis interval and the size of the analysis window. First, we reveal the influence of sliding window size on the identification accuracy by setting different size of the window. As Fig. 8 shows, when the sliding window size is set to 4, the identification accuracy is high, and with the increase of the window size, the identification accuracy is not improved obviously. Therefore, we set the sliding window size to 4 in the next experiment.

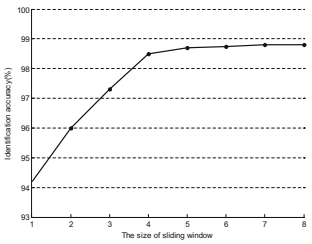

Fig. 8. The influence of sliding window size on the identification accuracy

Next, we exhibit the traffic analysis graph of P2P traffic identification emulation system for difference applications, where the red line represents the curve of traffic continuity quantization and the blue line represents the curve of traffic multi-connectivity quantization, and the identification result of the current traffic is displayed in the right box of the system. Figure 9 shows the traffic analysis graph of a P2P software with unknown protocol, i.e. Vagaa. In the test, Vagaa downloads two video files at the rate of 37 KB/s and 64 KB/s, respectively. It can be seen from the figure, the continuity quantization value of the Vagaa traffic is mainly around 1.0 and its multi-connectivity quantization value is changed between 0.4 and 0.8. The system determined the current traffic is P2P traffic.

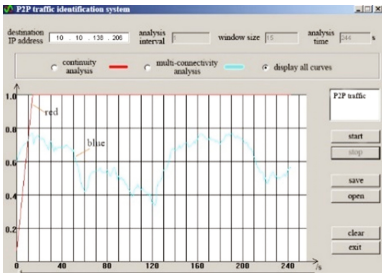

Fig. 9. Analysis graph of Vagaa traffic

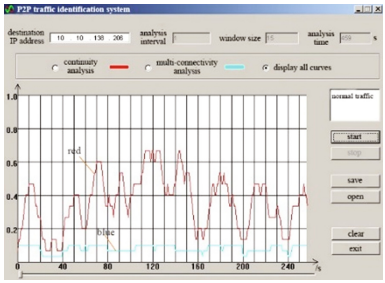

Fig. 10. Analysis graph of web traffic.

Figure 10 is the analysis graph of web traffic. As the figure shows, its continuity quantification values fluctuate greatly between 0.0 and 0.7. It indicates that the continuity of web traffic is weak. Although in the process of web traffic analysis, we open up to 20 different web pages continuously, the multi-connectivity quantification values generally stay below 0.2. The identification result shown in the right box is normal traffic, i.e. non-P2P traffic.

6 Conclusions

In this paper, we study the characteristics of P2P traffic start from the basic characteristics of a P2P network, and take the statistical characteristics of P2P traffic, continuity and multi-connectivity, as the basis of P2P traffic identification model. Compared with the traditional P2P traffic identification methods, the P2P identification method based on statistical characteristics presented in this paper has many advantages. First, the traffic identification method is easy to carry out without knowing the content of traffic itself. Secondly, the method can identify P2P traffic with known and unknown protocols. Admittedly, the P2P traffic identification method proposed in this paper cannot identify the specific traffic type that comes from which P2P software, which will serve as our future research direction.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61373135, No. 61672299).

References

1. Horng, M.F., Chen, C.W., Chuang, C.S.: Identification and analysis of P2P traffic-an example of bittorrent. In: Proceedings of the First International Conference on Innovative Computing, Information and Control, pp. 266–269, 30 August–1 September 2006
2. Chen, Z.Q., Delis, A., Wei, P.: Identification and management of sessions generated by instant messaging and Peer-to-Peer systems. Int. J. Coop. Inf. Syst. **17**, 1–51 (2008)

3. Bhatia, M., Rai, M.K.: Identifying P2P traffic: a survey. Peer-to-Peer Netw. Appl. **9**, 1–22 (2016)
4. Karagiannis, T., Broido, A., Faloutsos, M.: Transport layer identification of P2P traffic. In: Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement, pp. 121–134 (2004)
5. Karagiannis, T., Papagiannaki, K., Faloutsos, M.: BLINC: multilevel traffic classification in the dark. In: Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 229–240, October 2005
6. Constantinou, F., Mavrommatis, P.: Identifying known and unknown Peer-to-Peer traffic. In: Proceedings of the 5th IEEE International Symposium on Network Computing and Applications, pp. 93–102, 24–26 July 2006
7. Matsuda, T., Nakamura, F., Wakahara, Y.: Traffic features fit for P2P discrimination. In: Proceedings of the 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, pp. 230–235, 9–10 November 2005
8. Chen, Z., Wang, H., Peng, L.: A novel method of P2P hosts detection based on flexible neural tree. In: Proceedings of the of 6th International Conference on Intelligent Systems Design and Applications, pp. 556–561, 16–18 October 2006
9. He, J., Yang, Y., Qiao, Y.: Fine-grained P2P traffic classification by simply counting flows. Front. Inf. Technol. Electron. Eng. **16**, 391–403 (2005)
10. Qin, T., Wang, L., Zhao, D.: CUFTI: methods for core users finding and traffic identification in P2P systems. Peer-to-Peer Netw. Appl. **9**, 424–435 (2016)