Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2015, Article ID 801934, 9 pages
http://dx.doi.org/10.1155/2015/801934

Research Article

Multilevel Bloom Filters for P2P Flows Identification Based on Cluster Analysis in Wireless Mesh Network

Xia-an Bi,[1] Xiaohui Wang,[1] Luyun Xu,[2] Sheng Chen,[3] and Hong Liu[1]

[1] *College of Mathematics and Computer Science, Hunan Normal University, Changsha, Hunan 410081, China*
[2] *Business School, Hunan University, Changsha, Hunan 410082, China*
[3] *College of Information Science and Engineering, Hunan University, Changsha, Hunan 410082, China*

Correspondence should be addressed to Xia-an Bi; bixiaan@hnu.edu.cn

Received 20 January 2015; Revised 15 May 2015; Accepted 27 May 2015

Academic Editor: Gabriella Bretti

With the development of wireless mesh networks and distributed computing, lots of new P2P services have been deployed and enrich the Internet contents and applications. The rapid growth of P2P flows brings great pressure to the regular network operation. So the effective flow identification and management of P2P applications become increasingly urgent. In this paper, we build a multilevel bloom filters data structure to identify the P2P flows through researches on the locality characteristics of P2P flows. Different level structure stores different numbers of P2P flow rules. According to the characteristics values of the P2P flows, we adjust the parameters of the data structure of bloom filters. The searching steps of the scheme traverse from the first level to the last level. Compared with the traditional algorithms, our method solves the drawbacks of previous schemes. The simulation results demonstrate that our algorithm effectively enhances the performance of P2P flows identification. Then we deploy our flow identification algorithm in the traffic monitoring sensors which belong to the network traffic monitoring system at the export link in the campus network. In the real environment, the experiment results demonstrate that our algorithm has a fast speed and high accuracy to identify the P2P flows; therefore, it is suitable for actual deployment.

1. Introduction

Wireless networks are getting more and more popular nowadays. As users have got used to wired infrastructure networks, it is becoming extremely indispensable for wireless networks to be committed to providing similar service features to them [1]. It is difficult to find the relevant contents and services because the users and the data associated with a variety of applications are distributed over various sites and devices [2]. Many resources in WMNs can be used efficiently, aiming to maximize the total throughput of the whole network. In these networks, the key to maximize aggregate throughput is the flow identification scheme plays [3].

As more and more people are interested in wireless mesh networks, making efforts to supply users with a similar quality of service is important to the ones who are adapted to networks with wired infrastructure [4]. P2P (peer-to-peer) has grown to be a network transmission technology of high efficiency because of the widespread adoption of broadband residential access. Furthermore, it takes advantage of the modern network technology as well as distributed computing technology. It is a kind of distributed network and its basic idea lies in changing the traditional Client/Server mode [5]. In recent years, P2P (peer-to-peer) network has gathered broad attentions because of the fact that the peer nodes have no need to be with the help of intermediate servers to achieve the purpose of communicating with each other. Besides, it has become a technology which has a bright future [6]. In the last decade, the number of applications based on P2P technology has been increasing, such as BT, PPLive and eDonkey, Thunder. Up to now, P2P systems have accounted for more than 60% of Internet traffic in China [7].

The purpose of using mesh routers as wireless mesh network's device (WMN) is to form a wireless backbone rather than such a wired network. It is wireless mess backbone network architecture in Figure 1. Acting as a server to each other, each mess node in the graph gets the services provided by the peer-to-peer node. Distributed network makes a great

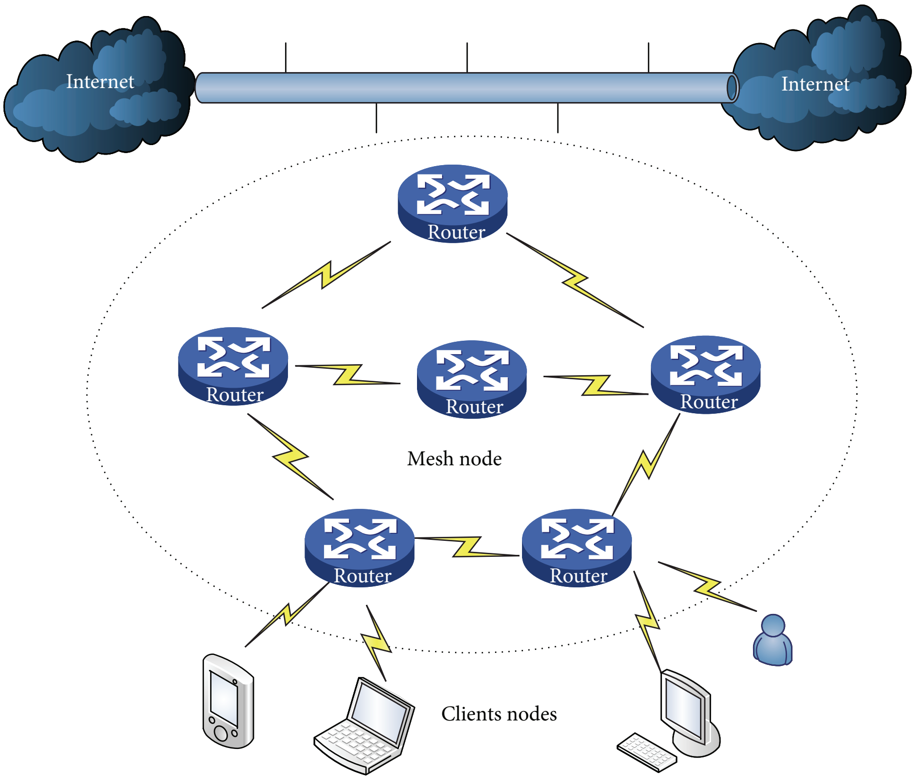

Figure 1: Wireless mess backbone network architecture.

difference in the flow distribution of network. Besides, it does reduce the stress on the storage server. What improved by the emergency of P2P network is the user experience along with enriching the Internet. However, the excessive growth in its flow as well as unlimited usage of bandwidth brings network congestion, increasing network packet loss and network delay. In other words, the network performance and quality of service are reduced in a great degree. Moreover, malicious code, reactionary, obscenity information, and piracy resources in the P2P network wantonly spread [8]. As a result, P2P applications can reduce the performance of the network greatly, sometimes making a rather adverse impact on the regular network services. It requires controlling and keeping an eye on the P2P flows continuously. Also, it needs to guarantee the regular operations of network services [9]. These are the motivations of P2P flows identification in wireless mesh network.

How can we manage the network bandwidth of P2P services as well as ensuring the quality of service? To maximize the users' satisfaction of P2P streaming in WMNs, The key technology is the flow identification scheme. It enables the network administrators to execute different control strategies according to different flow requirements, in order to achieve the effective management of P2P services [10, 11]. Therefore, the accurate identification and classification of P2P flows generally become the focus for network operators and service providers.

Over the past decades, the studies of flow identification of P2P services have been widely concerned [12–16], such as the port identification method [12, 13], the host-behavior characteristics analysis method [14], and the identification method based on flow statistical properties [15, 16]. But these methods cannot identify the P2P flows accurately and fast. So what we put forward in this paper is a high-performance P2P flow identification algorithm based on multilevel bloom filters.

The primary point of this paper can be divided into the following points:

(1) This paper finds that P2P flows have the locality characteristics of the time intervals of packets arrival and the length of packets in P2P flows.

(2) An efficient flow identification scheme based on the multilevel bloom filters is proposed for identifying P2P flows.

(3) A comprehensive set of experimental results demonstrate that our algorithm effectively enhances the performance of P2P flows identification.

The rest of the paper is organized as follows. Section 2 introduces the related work. In Section 3, we analyze the package lengths and time property features of P2P flows. Section 4 gives an efficient P2P flows identification scheme. Section 5 is the simulation evaluation, and Section 6 is the performance evaluation in real environment. Finally, Section 7 concludes the paper.

2. Related Works

In this section, we provide a brief discussion on the methods of P2P flow identification.

P2P flow identification method mainly has three categories: the port identification method [12, 13], the host-behavior characteristics analysis [14], and the identification method based on flow statistical properties [15, 16].

The port identification method [12] is the most primitive and simple network flow identification method. It is known that many traditional network applications use a fixed port. For example, HTTP flow uses port 80 and MSN uses ports 1863 and 80 and so forth [17]. Therefore it can quickly and efficiently identify the corresponding flow according to the port numbers, which has a low degree of complexity. However, with the development of new business, a lot of services use dynamic random port in order to prevent filtering. When facing such a network service, the port identification method is almost a failure and the classification accuracy is very low. Due to its simple and fast identification ability, the TCP/UDP port identification method is still used in high-speed network flow identification. A major concern in utilizing diverse strategies to change the port numbers of the new P2P applications aims at avoiding traffic identification. As a consequence, on account of incomplete and inaccurate identification results, there is no use of port-based method [13].

The host-behavior characteristics analysis [14] is mainly designed for P2P flow. The basic idea of this method is analyzing the data packet, summarizing P2P flow characteristics according to the analysis, and identifying the flow whether belongs to P2P applications [18]. In recent years, researchers have proposed many network measurement methods based on behavioral characteristics, and they have good scalability and high accuracy of identification [19]. Because of the part similarity of the network service model, those methods can only identify coarse-grained network services, and its memory consumption is very large. However, P2P applications and the regular applications cannot be discriminated by similar behaviors which cannot identify the traffic accurately.

The P2P identification method based on flow statistical properties is a solution overcoming the limits of port identification's and flow behavior characteristics analysis. It uses statistics on arrival time interval, duration and a series of characteristics of packets, and supervised or unsupervised machine learning methods to achieve services identification. Supervised machine learning [15] trains data to model and then classifies data directly on this model, while unsupervised machine learning [16] classifies data directly. The identification method based on machine learning has a better scalability and can identify the encrypted data flow, and its classifier also has a good scalability and flexibility. But they have low performance due to serious consumption of resources caused by signature searching in the payload of every packet.

3. Research on the Locality Characteristics of P2P Flows

3.1. The Locality Characteristics. In this section, we find the locality characteristics of the P2P flows through the research on the package lengths and time property features.

Over the past decade, researchers have revealed some statistical characteristics of the P2P flows through a large number of studies. It is also found that recently referenced file has a greater probability to be referenced again soon [20]. The researchers found that the P2P applications have some features such as synchronous upstream and downstream flow, fast transmission and high-capacity, wide distributed service points, and lack of security mechanisms [21]. These features determine that the P2P network has uncertainty, encryption, and large capacity.

For more comprehensive understanding and analysis of the characteristics of P2P flows, we select the average packet length and packet arrival time interval values to do experiments. Our purpose is to design an appropriate algorithm structure to identify P2P flows through the analysis of these values. We get the P2P packets from Internet and read the five-tuple information of the packets and then classify each packet to its own flow according to classification algorithm. Subsequently we get the corresponding time of the packet belonging to this flow and finally calculate the time attribute and the values of the packet length.

This paper used the P2P flows and did experiment on the time intervals of packet arrival and packet's average length. The results of the analysis are shown in Figures 2 and 3. As shown in Figure 2, it depicts average time interval of the arriving packet in the P2P flows. Obviously, as the number of flows increases, the attributes of packets gradually reduced and finally become stabilized when the number of flows grows up to 60. At this time the average time interval is about 1.7 seconds. As shown in Figure 3, it depicts the statistical value of the average length of the packet in the P2P flows. It shows that when the number of data flows reaches 130, the average packet length tends to be stable and around 100. We find the time intervals of packet arrival and the length of packet in P2P flows is less than other Internet flows, which is also in line with our analysis of the locality characteristics of P2P flows [22, 23].

3.2. The Mathematical Basis

Definition 1. The Minkowski distance of data x_t and mean u_t is

$$d(x_t, u_t) = \left[\sum_{k=1}^{m} \left| x_{tm} - u_{tm} \right|^p \right]^{1/p}, \qquad (1)$$

where $X_t = \{x_{t1}, x_{t2}, \ldots, x_{tp}\}$ is the time series sample and $u_t = \{u_{t1}, u_{t2}, \ldots, u_{tp}\}$ is the mean of the sample, and p denotes the number of the dimensions of the sample [24].

In our study, we assume that the historical time series sample $X = \{x_1, x_2, \ldots, x_t\}$ of Internet packets subjects to normal distribution. The package lengths and time property

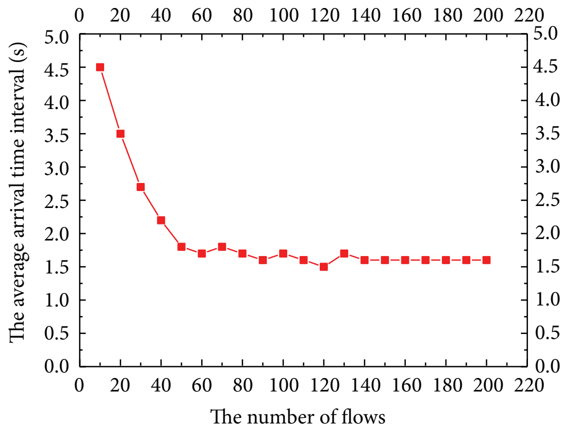

FIGURE 2: Statistics of the average arrival time interval.

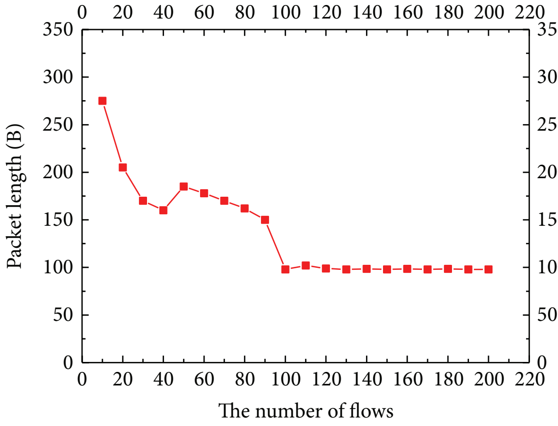

FIGURE 3: Statistics of the average packet length.

features make up the samples. Let us suppose that u denotes the mean and δ denotes the variance of sample X_t before time t. The distance from the newly generated sample data y to the mean u determines that the sample data y and the mean u share the same assigned cluster probability. The assigned cluster probability is as follows:

$$P(y,u) = \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{d(y,u)^2}{2\delta^2}\right), \qquad (2)$$

where $d(y,u)$ denotes the Minkowski distance of sample data y and the mean u.

From the equation we learn that the narrower the Minkowski distance between all of the newly produced sample data y and the mean u is, the bigger the value of $P(y,u)$ will become. In a certain period of time, the probability of the appearance of the data, closer to the mean u, is larger than the others in the similar time.

3.3. The Mathematical Model of Traffics Cluster Characteristics.

We give a quantitative analysis of the cluster characteristics of the real Internet traffics in this part. The study pays close attention to the package lengths along with the time property features.

The historical clustering sample coming into shape in the time period $[t-2r,t]$ along with cluster center u_A, cluster radius r, and sample variance δ_A is assumed to be similar prior information of the similar prior of time series sample $B\{x_t,\ldots,x_{t+k}\}$ in the continuous period of time $[t,t+k]$ [25]. It enables us to use a biased method to search the data of time series sample $B\{x_t,\ldots,x_{t+k}\}$ in the history cluster sample A. According to the influence that the time series sample $B\{x_t,\ldots,x_{t+k}\}$ in the period of time $[t,t+k]$ has on historical cluster sample center u_A, the new history cluster sample B can be generated. The similar prior information for subsequent time series sample can be provided by the newly produced cluster [26].

There exist two types of data: $x_{\text{new}} = \{x \mid x \in B,\ x \notin A\}$, $x_{\text{old}} = \{x \mid x \in B,\ x \in A\}$ in time series sample $B\{x_t,\ldots,x_{t+k}\}$ produced by machine after time t. We give the following definitions for the purpose of reflecting the impact which is produced by the sample $B\{x_{t+1},\ldots,x_{t+2r}\}$ on the historical cluster center u_A.

Definition 2. A deviate sample B_{new} can be made up from the data of $x_{\text{new}} = \{x \mid x \in B,\ x \notin A\}$ in the time series sample $B\{x_t,\ldots,x_{t+k}\}$. One can judge whether the data x_{t+i} in the $B\{x_{t+1},\ldots,x_{t+2r}\}$ belongs to the deviate sample B_{new} using the Pearson correlation function as follows:

$$S(x_{t+i},u_A) = \frac{\sum_{k=1}^{m}(x_{t+i,k}-\overline{x_{t+i}})(u_{A,k}-\overline{u_A})}{\sqrt{\sum_{k=1}^{m}(x_{t+i,k}-\overline{x_{t+i}})^2 * \sum_{k=1}^{m}(u_{A,k}-\overline{u_A})^2}}, \qquad (3)$$

where $x_{t+i} \in B$ and u_A is the clustering center of cluster A. Also, m is the number of the dimensions of the sample. We assume the value of the function is s. If the inequality $s \geq \alpha$ satisfies, we can know that $x_{t+i} \in B_{\text{new}}$. If the inequality $s < \alpha$ satisfies, we can know that $x_{t+i} \notin B_{\text{new}}$. According to the real situation, the value of α can be adjusted properly.

Definition 3. At first, the Minkowski distance between the data x_{t+i} in $B\{x_t,\ldots,x_{t+k}\}$ and the history cluster center u_A should be calculated, respectively. And then the probability of the same assigned cluster of x_{t+i} and u_A and Pearson correlation function $S(x_{t+i},u_A)$ should be calculated in the meantime. One can regard the product of the three as the deviate cost of x_{t+i} for historical cluster samples A. The deviate cost function is as follows:

$$G(x_{t+i},u_A) = P(x_{t+i},u_A) * S(x_{t+i},u_A) * d(x_{t+i},u_A). \qquad (4)$$

Definition 4. Historical cost function is as follows:

$$\text{Hc} = \sum_{i=1}^{k}\sum_{x_{t+i}\in B} G(x_{t+i},u_{t-1}) = \sum_{i=1}^{k}\sum_{x_{t+i}\in B} p(x_{t+i},u_{t-1}) \cdot \delta(B(x_{t+i}),A(u_{t-1}))\left\|x_{t+i}-u_{t-1}\right\|^2. \qquad (5)$$

The function is the total deviate cost of the data of the sample $B\{x_t,\ldots,x_{t+k}\}$ from the historical cluster samples

$A\{u_{t-1}\}_{u_{t-1}-r}^{u_{t-1}+r}$. When Hc $< \varepsilon$ is satisfied, take the u_B of the new time series sample B as a new cluster center to form a new cluster sample B.

Definition 5. One can evaluate the clustering quality of the new cluster sample B by calculating the Pearson correlation degree between new cluster center u_B and historical cluster center u_A. The objective function is shown as follows:

$$S\left(u_B,u_A\right)=\frac{\sum_{k=1}^{m}\left(u_{B,k}-\overline{u_B}\right)\left(u_{A,k}-\overline{u_A}\right)}{\sqrt{\sum_{k=1}^{m}\left(u_{B,k}-\overline{u_B}\right)^2 * \sum_{k=1}^{m}\left(u_{A,k}-\overline{u_A}\right)^2}}. \qquad (6)$$

For the purpose of better reflecting the changes of new time series samples as well as producing new clusters [27] in a faster and better way, we can adjust the parameter ε appropriately with the help of the cluster quality function $S(u_B, u_A)$.

A cluster algorithm for packet matching is given out in the next. In the time period of $[t-2r, t]$, a historical time series cluster sample A should be assumed at first.

(1) A packet x_{t+1} is produced by the Internet and added to the time series sample B at time $t+1$ and the rest can be done following this way. When it comes to time $t+2r$, the historical cost function of sample $B\{x_{t+1}, \ldots, x_{t+2r}\}$ should be calculated. If Hc $\le \varepsilon$, then a new cluster B is formed. Besides, we should calculate a new cluster u_B and a new variance δ_B.

(2) For the purpose of preferably reflecting new time series samples' changes and producing new clusters in a faster and better way, we should calculate the cluster quality function $S(u_B, u_A)$ to update the parameter ε appropriately.

4. Our Algorithm for Identifying P2P Flows

4.1. The Architecture of Our Algorithm. In this part, we design an efficient multilevel bloom filters algorithm to identify the P2P flows with high performance according to the locality characteristics of P2P flows.

Through the above experiments we get a detailed analysis of the locality characteristics of P2P flows; for example, the average time interval of packet arrival is stable at about 2 seconds as shown in Figure 2. Because of the quick update of the peer-to-peer flow nodes and the rapid transmission, the algorithm, respectively, stores the 0–30-second data flows in the first-level bloom filter, the 30–90-second data flows in the second-level bloom filter, and the remainder of the flows in the last-level bloom filter. If there are too many flows stored in the first-level bloom filter, in this case, the algorithm will consume more identification time. So we store more flows in the second-level bloom filter. The whole multilevel structure is designed as shown in Figure 4.

The algorithm adds a counter, respectively, in the three levels to calculate the amount of packets pertained to the same flow. When the packets of Internet flows enter, the algorithm firstly obtains the details of five-tuple information of the packets, and they are the source IP address (SA), the destination IP address (DA), the source port (SP), the destination port (DP), and protocol (Pro). Then our algorithm identifies which flows the packets belong to. The P2P flow identification step of our algorithm is the packet matching which is from the first-level to the last-level bloom filter. Firstly, when the incoming packet enters, the algorithm uses the hash functions to search out the flows, stored in the first-level bloom filter.

```
Loop
Search (in_packet, SA, DA, SP, DP, Pro)
{//get the five-tuple information
    for ($i = 1$; $i \le 3$; $i$++)
    {//traverse three level bloom filters;
        for ($i = 1$; $i \le k$; $i$++)
        {//calculate the hash values;
            Hashi = Hashi(SA, DA, SP, DP, Pro);
        }
        for ($i = 1$; $i \le k$; $i$++)
        {//compare the corresponding positions;
                if (Hashi == BF[$i$])
                //find out the matching rules;
                {return the rules;}
        }
    }
}
End loop for all packets
```

Algorithm 1: The searching procedure of the scheme.

The searching procedure can be described by a pseudocode as shown in Algorithm 1. The specific method can be described as follows. The five-tuple information (SA, DA, SP, DP, and Pro) is substituted into FL_Hash1, FL_Hash2, . . ., FL_Hashk and the result values are compared with the first-level bloom filter. If the algorithm can find the matching flow node, the searching procedure will stop. Otherwise the packet enters the second-level bloom filter. Similarly the five-tuple information is substituted into SL_Hash1, SL_Hash2, . . ., SL_Hashk and compares the values in the second-level bloom filter. If the matching flow is found, the searching step will stop. Otherwise the packet enters the third-level bloom filter. And the searching procedure continues to the last-level structure until the corresponding flow is found. Due to the locality characteristics of P2P flows, a newly arriving packet has a large probability of being found in the first level of the structure and the corresponding counter of the flow is directly updated.

Our algorithm is designed by bloom filters. Bloom filter has false positive, so we need to discuss the false positive probability of our approach. Assuming the length of each bloom filter is m bits, the number of rules of each virtual router is n. Based on existing research results in the paper [28], we calculate the number of hash functions through $K = \lceil \ln 2(m/n) \rceil$ to reduce the probability of false positive.

4.2. Dynamic Flow Aging and Update of the Multilevel Structure. With time elapsing, some flows have been out of use and

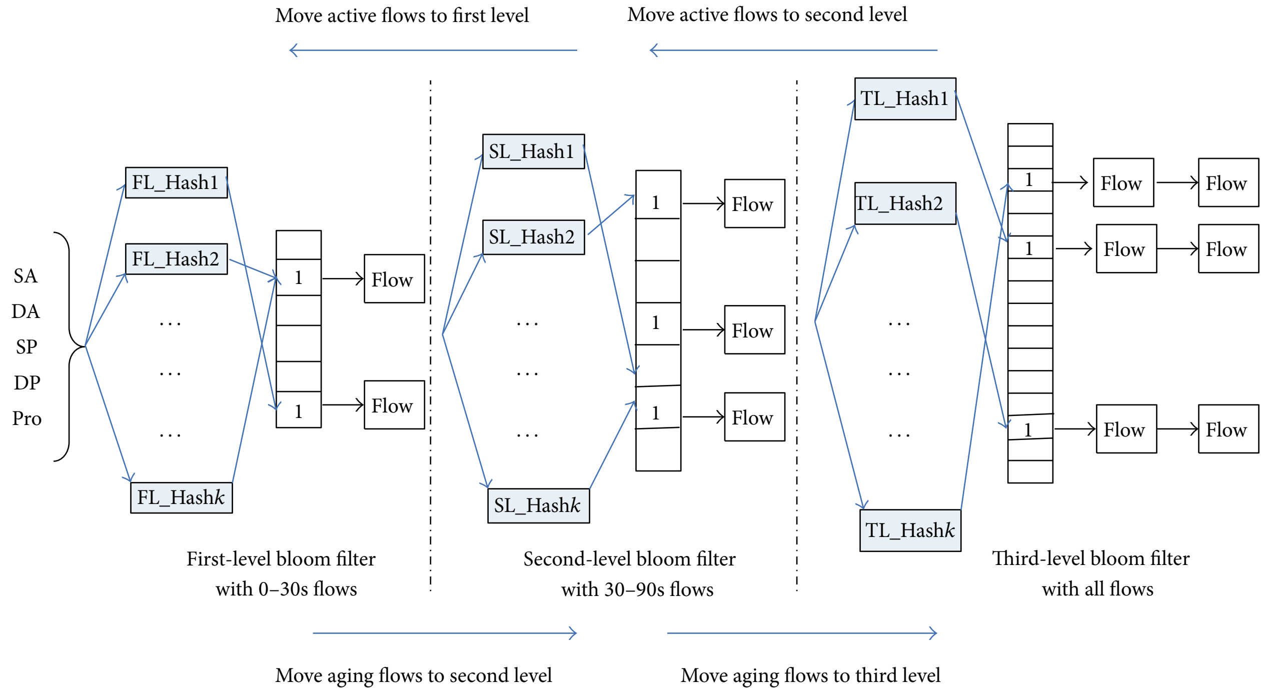
Move active flows to first level
Move active flows to second level
SA
DA
SP
DP
Pro
FL_Hash1
FL_Hash2
FL_Hashk
1
1
Flow
Flow
SL_Hash1
SL_Hash2
SL_Hashk
1
1
1
Flow
Flow
Flow
TL_Hash1
TL_Hash2
TL_Hashk
1
1
1
Flow
Flow
Flow
Flow
Flow
Flow
First-level bloom filter with 0–30s flows
Second-level bloom filter with 30–90s flows
Third-level bloom filter with all flows
Move aging flows to second level
Move aging flows to third level

Figure 4: The multilevel bloom filters' architecture of the flow identification scheme.

the corresponding records should be eliminated in our bloom filter data structure. The memory space which is released can be used for the following flows. According to the locality characteristics of the P2P flows, we use sample data packet to update the timestamp of the flows, instead of using timestamp of every packet to update the information. Therefore, the algorithm reduces lots of writing operations on the memory. With the calculation and analysis of the flow's timestamp we can get the inactive P2P flows. By experimental analyzing packets in the previous section, we define the flow whose reaching time exceeds 10 seconds as an inactive flow and will move these flow nodes from the first-level to the second- or the third-level bloom filter. And the algorithm alternately updates the flows whose reaching time is within 5 seconds from the last two levels to the first level. Through the dynamic update of the data structure, the algorithm greatly improves the flow matching speed and the utilization coefficient of storage resource.

5. Simulation Evaluation

In this section, we come up with the emulation experiments to compare the performances of our algorithm with the flow statistical properties (FSP) algorithm [14] in P2P flow identification. In the experiments, the metrics of performance include the memory access evaluating the searching performance.

5.1. Experimental Environment. This paper tests the performance of algorithms by employing PALAC (packet lookup and classification simulator) in the Linux operating system (Kernel Version 3.16). PALAC provides the performance evaluation with a discrete event simulation environment. It consists of the following modules: flow generation module, classifier description language module, event queue manager module, classification or lookup algorithms repository module, classifier update module, and statistics collection and query module.

5.2. The Evaluation with Two Types of Data Sets. Below this paper uses two group experiments to test and analyze the performance of the algorithms. We select data sets from National Laboratory for Application Network Research (NLANR) and the Chinese Academy of Sciences Institute (CASI). NLANR team has exploited data collection permitting identifying a wide range of issues, which ranges from network connectivity and commodity issues to high-performance network hardware and router problem. This data of NLANR is useful for longitudinal study of the Internet flows, and it can be available from the NLANR website. Through the study of the two kind data sets, we find that the distribution of packet in the data sets of Chinese Academy of Sciences Institute is relatively scattered, and NLANR is relatively concentrated. We use the flow generation module of PALAC to generate the P2P flows with the generating packet rate of 1 G bit/sec and the generating packet time of 30 minutes.

Figures 5 and 6 are the experimental results of FSP algorithm and our algorithm on packets scattered distribution (CASI data sets) and concentration (NLANR data sets) in P2P flows. Figure 5 shows the memory access performance of our algorithm has an average increase of 33.54% compared with the FSP algorithm when packets are in the relatively scattered case. Figure 6 shows our algorithm's memory access performance has an average increase of 35.17% compared with FSP

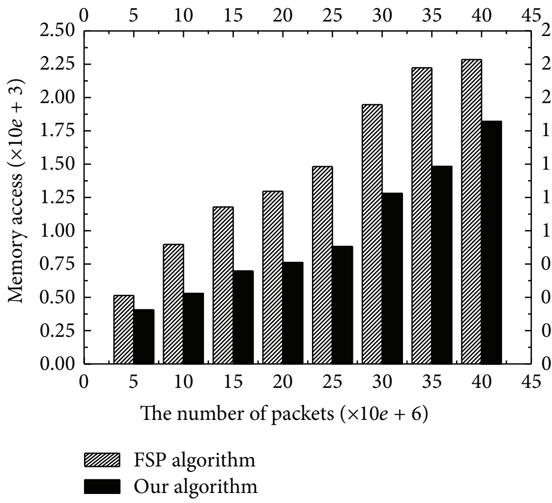

Figure 5: Performance comparison on CASI data sets.

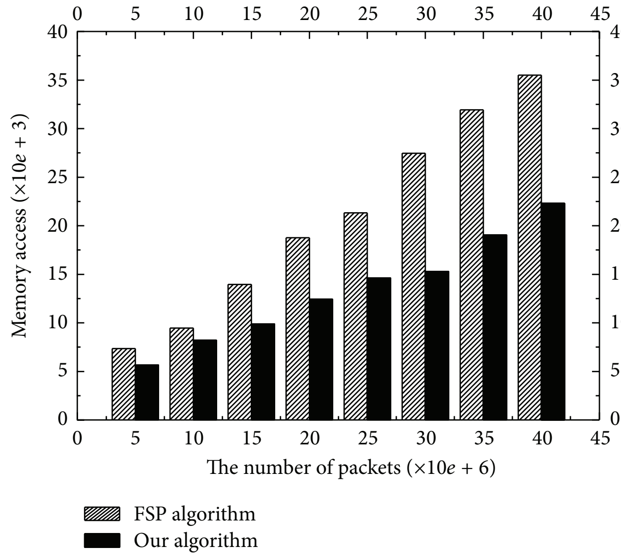

Figure 6: Performance comparison on NLANR data sets.

algorithm when packets are in the relative concentration case. This is because the packets of the P2P flows are relatively smaller than the packets of other Internet flows, but their transmission speed is greater than the other packets, which makes the P2P flows identification more difficult for the FSP algorithm. However, it becomes much easier to identity P2P flows for our algorithm.

6. Performance Evaluation in Real Environment

In this section, we present the experiments to compare the performances of our algorithm with the port-based and host-behavior-based algorithms in real environment. In the experiments, the metrics of performance include the memory access, evaluating the searching performance, and the identification precision, evaluating the accuracy of the algorithms.

6.1. Experimental Environment. In order to fully verify the practical performance of the packet classification algorithm, the algorithm and the rule sets should be written on the network traffic monitoring system to test the effect of the algorithms for the actual network traffic monitoring results and then improve our algorithm.

Figure 7 shows the deployment of the network traffic monitoring system at the export link in the campus network. The system is divided into the traffic monitoring sensors, the traffic data collector, the data storage center, the data analysis center, and the remote browser. The traffic monitor probe is deployed in the vicinity of the routers and the network servers and other kinds of network equipment, which is responsible for the data packets mirroring and identifying the data packets as the service traffic of the application layer, the experimental data as the real network traffic in campus network according to the packet classification algorithms. We use SmartBits 2000 network test platform to test the performance of the algorithms, to further improve our algorithm and the efficiency of the algorithm in practical application.

Below we use two group experiments to test and analyze the performance of the algorithms.

6.2. The Evaluation on Speed and Accuracy. Firstly, this group experiment is utilized to evaluate the speed of the three algorithms with the same experimental configuration. As shown in Figure 8, compared with the port-based algorithm and the host-behavior-based algorithm, the average memory access of our algorithm separately drops by 66% and 47%. This experiment demonstrates that our algorithm has a fast speed to identify the P2P flows.

Secondly, this group experiment is utilized to evaluate the accuracy of the three algorithms with the same experimental configuration. As shown in Figure 9, compared with the accuracy 26.92% of port-based algorithm and accuracy 53.25% of host-behavior-based algorithm, our algorithm has a high accuracy 87.25%. This experiment demonstrates that our algorithm is suitable for actual deployment.

7. Conclusions

As the Internet brings efficiency and convenience to people's life, study, and work, the Internet becomes more and more important as well as its influence; besides a large number of network applications came into being. Not only abundant traditional applications such as Web, FTP, Email, and Telnet but also a mass of new services exist in the network, for example, P2P, streaming media, virtual reality, and interactive online applications. A wide variety of network applications and a large number of Internet users have made the constitution of the Internet flows increasingly complex. Followed by this, the Internet flow identification technology has developed rapidly in the meanwhile.

In this paper, an efficient P2P flows identification scheme based on multilevel bloom filters is proposed. Through

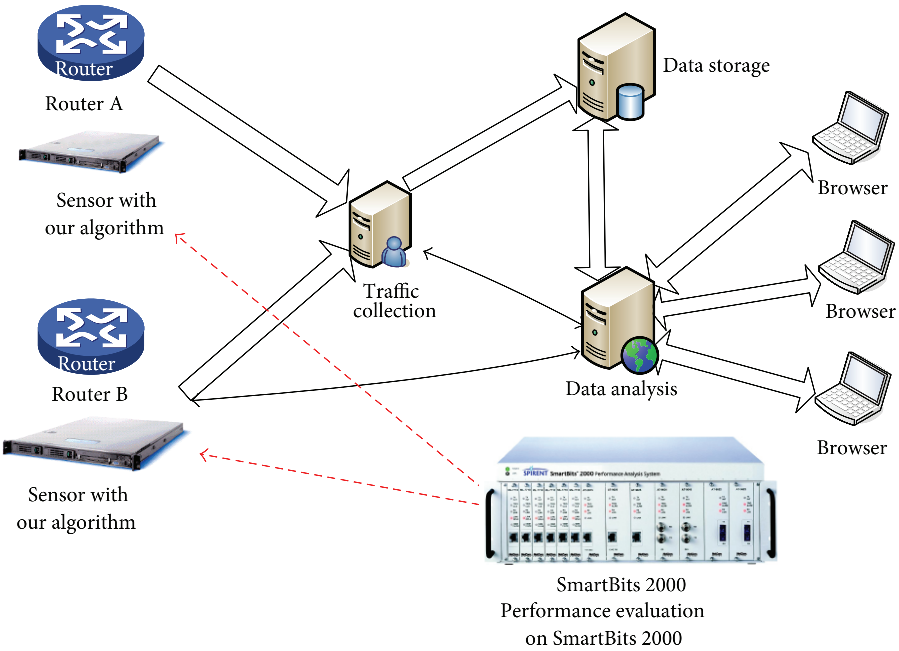

Figure 7: Performance evaluation on real environment.

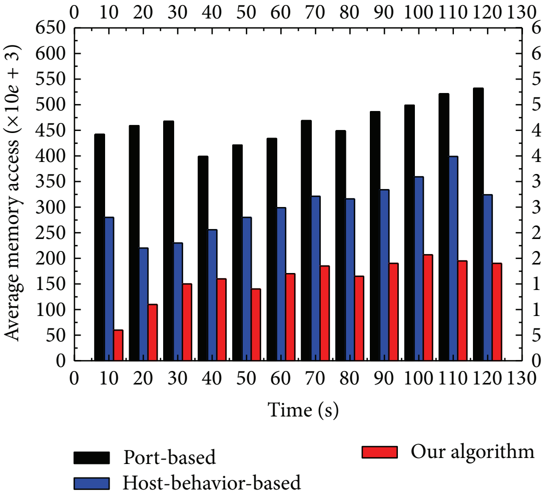

Figure 8: The memory access comparison.

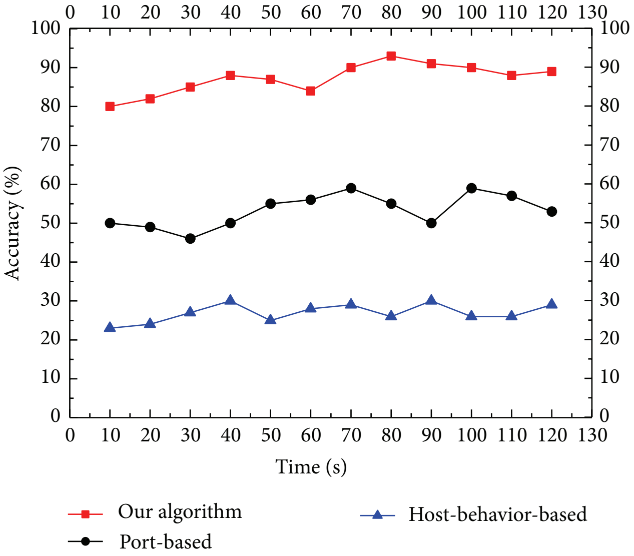

Figure 9: The accuracy comparison.

the study on the package lengths and time property features of P2P flows, the scheme is designed as a multilevel structure containing bloom filters. Different level structures store different numbers of flow rules, and the searching steps of the scheme traverse from first level to the last level. The simulation results demonstrate that our algorithm effectively enhances the performance of P2P flows identification.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Basic Research Program of China (973 Program, no. 2012CB315801); the National Science Foundation of China (nos. 61173167 and 61173168); the Scientific Research Foundation for Ph.D. Hunan Normal University (no. Math 120641); Youth Scientific Research Fund of Hunan Normal University (no. 11301); Scientific Research Fund of Hunan Provincial Education Department (no. 10C0941).

References

[1] Z. R. Zaidi, S. Hakami, B. Landfeldt, and T. Moors, "Real-time detection of traffic anomalies in wireless mesh networks," *Wireless Networks*, vol. 16, no. 6, pp. 1675–1689, 2010.

[2] N. Kumar, N. Chilamkurti, and J.-H. Lee, "Distributed context aware collaborative filtering approach for P2P service selection and recovery in wireless mesh networks," *Peer-to-Peer Networking and Applications*, vol. 5, no. 4, pp. 350–362, 2012.

[3] N. Kumar, N. Chilamkurti, and J.-H. Lee, "A novel minimum delay maximum flow multicast algorithm to construct a multicast tree in wireless mesh networks," *Computers & Mathematics with Applications*, vol. 63, no. 2, pp. 481–491, 2012.

[4] Z. R. Zaidi, S. Hakami, T. Moors, and B. Landfeldt, "Detection and identification of anomalies in wireless mesh networks using principal component analysis (PCA)," *Journal of Interconnection Networks*, vol. 10, no. 4, pp. 517–534, 2009.

[5] L. Li, G. Zhang, and A. Yao, "The model design of MP2P content distribution networks based on sphere clusters," *Journal of Computational Information Systems*, vol. 8, no. 4, pp. 1732–1743, 2012.

[6] J. Zhang, H. Duan, W. Liu, and J. Wu, "Anonymity analysis of P2P anonymous communication systems," *Computer Communications*, vol. 34, no. 3, pp. 358–366, 2011.

[7] J. Dong, X. Ren, D. Zuo, and H. Liu, "An adaptive failure detector based on quality of service in peer-to-peer networks," *Sensors*, vol. 14, no. 9, pp. 16617–16629, 2014.

[8] J. Ju, F. Fan, and J. Wu, "Analysis of model and key technology for P2P network route security evaluation with 2-tuple linguistic information," *Journal of Computational Information Systems*, vol. 9, no. 14, pp. 5529–5534, 2013.

[9] X.-A. Bi, D.-F. Zhang, X.-B. Yang, and S. Chen, "An efficient P2P traffic identification scheme," *International Journal of Digital Content Technology and Its Applications*, vol. 5, no. 12, pp. 459–467, 2011.

[10] L. Feng, X. Liao, Q. Han, and L. Song, "Modeling and analysis of peer-to-peer botnets," *Discrete Dynamics in Nature and Society*, vol. 2012, Article ID 865075, 18 pages, 2012.

[11] S. Zeng, L. Li, and D. Liao, "Path selection and bandwidth allocation for fixed and mobile peers in P2P streaming system," *Journal of Computational Information Systems*, vol. 8, no. 17, pp. 7163–7170, 2012.

[12] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: multilevel flow classification in the dark," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, New York, NY, USA, 2005.

[13] F. Constantinou and P. Mavrommantis, "Identifying known and unknown peer-to-peer flow," in *Proceedings of the IEENCA*, Washington, DC, USA, 2006.

[14] Y. Sawaya, A. Kubota, and Y. Miyake, "Detection of attackers in services using anomalous host behavior based on traffic flow statistics," in *Proceedings of the 11th IEEE/IPSJ International Symposium on Applications and the Internet (SAINT '11)*, pp. 353–359, Munich, Germany, July 2011.

[15] T. Nguyen and G. Armitage, "A survey of techniques for internet flow classification using machine learning," *IEEE Communications Surveys and Tutorials*, vol. 11, no. 3, pp. 37–52, 2008.

[16] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification," *ACM SIGCOMM Computer Communication Review*, vol. 30, no. 5, pp. 5–16, 2006.

[17] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy, "Transport layer identification of P2P traffic," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement (IMC '04)*, pp. 121–134, Taormina, Italy, October 2004.

[18] T. Liu and X. Chen, "A novel approach to detect P2P traffic based on program behavior analysis," in *Proceedings of the IEEE 2nd Annual Conference on Electrical and Control Engineering (ICECE '11)*, pp. 5677–5680, Yichang, China, September 2011.

[19] J. Zhang, R. Perdisci, W. Lee, U. Sarfraz, and X. Luo, "Detecting stealthy P2P botnets using statistical traffic fingerprints," in *Proceedings of the IEEE/IFIP 41st International Conference on Dependable Systems and Networks (DSN '11)*, pp. 121–132, Hong Kong, China, June 2011.

[20] H. Kang, M. Kim, and J. Hong, "A method on multimedia service flow monitoring and analysis," in *Proceedings of the 14th IEEE international workshop on DSOM*, Heidelberg, Germany, 2003.

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

[22] V. Carela-Espanol, P. Barlet-Ros, M. Solé-Simó, A. Dainotti, W. de Donato, and A. Pescapé, "K-dimensional trees for continuous traffic classification," in *Traffic Monitoring and Analysis*, vol. 6003 of *Lecture Notes in Computer Science*, pp. 141–154, Springer, Berlin, Germany, 2010.

[23] A. Dainotti, F. Gargiulo, L. I. Kuncheva, A. Pescapè, and C. Sansone, "Identification of traffic flows hiding behind TCP port 80," in *Proceedings of the IEEE International Conference on Communications (ICC '10)*, pp. 1–6, Cape Town, South Africa, May 2010.

[24] M. Polczynski and M. Polczynski, "Using the k-means clustering algorithm to classify features for choropleth maps," *Cartographica*, vol. 49, no. 1, pp. 69–75, 2014.

[25] R. Martino, P. Mazzotta, H. Bourdin et al., "LoCuSS: hydrostatic mass measurements of the high-LX cluster sample—cross-calibration of Chandra and XMM-Newton," *Monthly Notices of the Royal Astronomical Society*, vol. 443, no. 3, pp. 2342–2360, 2014.

[26] A. Hassan and R. Kouhy, "Time-series cross-sectional environmental performance and disclosure relationship: specific evidence from a less-developed country," *International Journal of Accounting and Economics Studies*, vol. 2, no. 2, pp. 60–73, 2014.

[27] J. B. MacQueen, "Some methods for classification and analysis of multivariate observation," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, Berkeley, Calif, USA, January 1967.

[28] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422–426, 1970.