VoIP Traffic Identification Based on Host and Flow Behavior Analysis

Bing Li, Zhigang Jin
Tianjin University
Tianjin, China
libingice@tju.edu.cn

Maode Ma
Nanyang Technological University
Singapore, Singapore

***Abstract*—More and more VoIP applications have emerged with the development of network and multimedia coding techniques. Their traffic identification is meaningful to for network management and application optimization. In this paper, a new traffic identification scheme, which combines traffic flow statistic analysis with host behavior estimation, is proposed to identify the VoIP traffic at transport layer. The host IP addresses and the port numbers in the network are examined as the host behavior to distinguish the VoIP traffic from traditional traffic. The inter-packet time has been modeled with the self-adaptive estimation. The experiment results show that our scheme could obtain a stable performance. At the same time, the proposed scheme could maintain its validity when existing VoIP applications are updated or new ones admitted. Both the accuracy and flexibility have been improved.**

Keywords-traffic identification; VOIP; host behavior

I. INTRODUCTION

More and more VoIP (Voice over Internet Protocol) applications have emerged with the development of networks and multimedia encoding techniques. Their traffic identification could help the network managers to analyze the performance of the Internet with VoIP applications. At the same time it could also help the software engineers to improve the quality of the VoIP applications further. So identification of VoIP traffic is necessary and important to both ISP managers and researchers.

The traditional method of traffic identification is to monitor the communication port at the transport layer. The specific port number is occupied by the specific application [1]. However, for the VoIP application, ephemeral random ports are usually used for the transmission of voice and video data. This implies that the traditional approach has not been suitable for the identification of VoIP applications.

Due to the limitation of traditional method, many new solutions have been proposed. In order to protect the user privacy, most studies have focused on the transport layer approach which only utilizes the information at the transport layer. The statistics clustering technique and some heuristic approaches have been applied [2-8].

In the statistics clustering methods, many statistical variables on the features of the traffic flows have been exploited as parameters of clustering. The Naive Bayes estimator and Bayesian neural network were applied to categorize the traffic in [2, 3]. An unsupervised approach based on Expectation Maximization (EM) algorithm was proposed for probability clustering in [4]. However, these solutions can only identify entire flow since the duration of flow usually as the parameter in algorithms. [5] proposed a method for the VoIP identification based on the flow level behavior (FLB). Only the packet size and inter-arrival time in a short period have been analyzed. Although it can be implemented with part of the traffic flow, it lacks the stability and flexibility for various VoIP applications. At the other hand, some heuristic methods have been explored based on the characteristics of the host behaviors [6-8]. In [6], an approach based on the behavior of P2P peers has been proposed. And then multilevel traffic classification in the dark, named BLINC, was developed in [7]. Concurrent to [7], a methodology based on data mining and information theoretic technique was proposed to discover the behavioral patterns of the hosts and the services provided by the hosts in [8]. However, these algorithms assume that only one network application has been in execution at one host. In reality, multiple applications may coexist. In this case, these algorithms are difficult to achieve a satisfied high accuracy of the traffic identification.

To overcome the limitations of existing solutions, we propose a new traffic identification scheme based on the host and the flow behavior analysis to identify the VoIP traffic at the transport layer. The major contribution of the proposal is to combine the statistic analysis on the traffic with the host behavior analysis. By our proposed scheme, the analysis of host behavior can distinguish the VoIP traffic from the traditional traffic. The feature of the VoIP flow is modeled to distinguish similar traffic. The self-adaptive estimated value and the relevance of the sequence have been considered.

The remainder of the paper is structured as follows. In Section II, we describe the features of VoIP applications and the corresponding models developed. In Section III, we present our solution to identify the VoIP traffic in detail. In Section IV,

978-1-4244-3709-2/10/$25.00 ©2010 IEEE

we show the experiments and the results analysis. Finally, in Section V, we conclude the paper.

II. Modeling of VoIP Applications

In this session we exploit the features of the VoIP traffic with respects of host behavior and traffic flow behavior. Based on the observation of the features the VoIP traffic, we have developed the models for the behaviors of the host and traffic flow, correspondingly.

A. Host Behavior

Hosts, with the execution of different types of network applications, usually behave differently due to structures and models of the applications. The port is focused in our study.

Traditional network applications usually adopt the Client/Server model by which the client connecting to the server would often initiate more than one concurrent connection in order to download objects in parallel. Take the web application for example. The web server with IP address W listens to port 80. The client with IP address C occupies a number of ports to connect to port 80 of W to obtain the objects maintained at the web page with pipeline mechanism to reduce the time for downloading. However, in the VoIP applications, the end-to-end communication is usually established between hosts after authentication. The pipeline mechanism is not necessary for any hosts in the session, where the numbers of connecting ports of the communication hosts are equal, or the difference will be extremely small.

Based on the difference of end-to-end structure and Client/Server structure, the number of ports used by the applications is the basis for the host behavior analysis. The source-destination IP pair can be represented as (srcIP, dstIP). The number of the source ports with the srcIP can be indicated by pair.srcPort.num and the number of the destination ports with the dstIP can be indicated by pair.dstPort.num. The difference, D between pair.srcPort.num and pair.srcPort.num should be small enough for the VoIP applications. D = |pair.srcPort.num – pair.dstPort.num| < threshold, where the threshold is a system parameter to differentiate the two distinct host behavior models.

B. Feature of Traffic Flow

Inter-packet arrival time is typical parameters to describe a traffic flow. Different applications may usually have different values to satisfy their specific requirements.

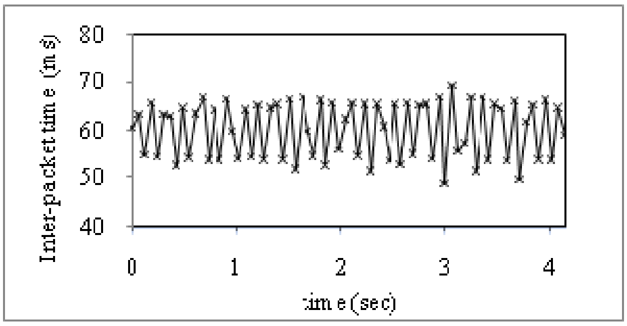

Figure 1 Inter-packet time of QQ voice

In the VoIP applications, the source host samples the voice or/and video signal and encodes them into frames with certain sampling period. The host assembles voice or video frames into packets. The inter-packet time (t) at the source host is either too short when the packets are from the same frame, or similar to the sample period *T* when the packets are from different frames. In this analysis, we will ignore the short inter-packet time and only consider the one similar to *T*. In the operation of a network, random jitter makes the inter-packet time to be a random variable with mean value as *T*.

As many traces of the VoIP traffic show, the waveform of the continuous inter-packet time usually shows as regular serrated and relevancy of adjacent inter- packet time. Figure 1 shows a fragment of the inter-packet time in a trace file of QQ voice. If the previous t is large, the next one will be small. Similarly if previous is small, next will be large. Based on the above observation, we propose to use two variables. One is the evaluation for the large inter-packet time (EL) and the other is evaluation for the small inter-packet time (ES) to estimate the large and small thresholds of inter-packet time. They can be derived with iteration following (1) and (2), respectively, where *a* is the smooth coefficient and t_i is the *i*th inter-packet time.

$$\begin{cases} EL_i = a * EL_{i-1} + (1-a)t_i & t_i - ES_{i-1} > EL_{i-1} - t_i \\ EL_i = EL_{i-1} & otherwise \end{cases} \tag{1}$$

$$\begin{cases} ES_i = a * ES_{i-1} + (1-a)t_i & t_i - ES_{i-1} < EL_{i-1} - t_i \\ ES_i = ES_{i-1} & otherwise \end{cases} \tag{2}$$

According to analysis of relevance of adjacent inter-packet time, some restrains should be satisfied. When the previous inter-packet time, t_{i-1} is larger than EL_{i-2}, the current inter-packet time t_i should be smaller than EL_{i-1}. Similarly, t_i should be larger than ES_{i-1} when t_{i-1} is smaller than ES_{i-2}.

At the same time, a ratio, R has been introduced which can be expressed by (3). *R* should guarantee that the difference between the large and small values is not too big since VoIP applications are sensitive to delay and cannot bear too much jitter.

$$R_i = ES_i/EL_i \tag{3}$$

III. Proposed Traffic Identification Scheme

Our proposed traffic identification algorithm works at the transport layer. Before our identification, the traffic should be pre-processed into flows, that is the packets with the same value of a five-tuple (transport protocol, source IP address, source port, destination IP address, destination port). At the same time timestamp is given to each packet.

The whole identification algorithm can be described as follows with three steps.

1. Examine the source-destination IP pair (srcIP, dstIP). If the difference D is smaller than the threshold *d*, the traffic between srcIP and dstIP could be considered as similar VoIP and should be identified further. Otherwise, it is filtered out as non-VoIP traffic.

2. Utilize (3) and (4) to calculate the values of EL_i *and* ES_i by the self-adaptive manner. And then calculate the ratio, R_i and set the threshold, γ for R_i. The VoIP traffic should satisfy two conditions below: (a) No continuous t_i values are larger or smaller than the corresponding EL_{i-1} or ES_{i-1}; (b) R_i should be larger than threshold γ

Assume that n is the number of continuous inter-packet times that satisfy condition (a) and (b), and $n= n(L) + n(S)$. $n(L)$ is the number of t_i that is near to EL_i while $n(S)$ is the number of t_i that is near to ES_i. The difference between $n(L)$ and $n(S)$ is $n(D)$. If n can reach the threshold μ and $n(D)$ could be smaller than the threshold δ , the traffic flow will be identified as VoIP traffic.

The values of γ and δ should be carefully set.

3. The VoIP traffic is usually bidirectional. If a flow is labeled as a VoIP traffic flow, the reverse flow will also be the VoIP traffic flow.

The whole scheme has been described in Fig. 2 with a flowchart. In Step 1 the host behavior feature is examined. The traffic based on traditional C/S structure is filtered out directly since VoIP data traffic is based on end-to-end. In the Step 2, the features of the traffic are examined and the inter-packet time is extracted to model according to the codec at the source host of VoIP application. Since the VoIP application is symmetrical, the reserve is also considered as VoIP flow in Step 3.

Compared to typical statistical clustering algorithms, the number of statistical variables has been reduced and no training is required before the identification.

IV. EXPERIMENTAL EVALUATION

In order to measure the performance of our proposal, the experiments have been implemented on the actual Internet traces. The FLB algorithm [5] has also been implemented to compare the efficiency of our proposed scheme.

Two performance metrics, named fault negatives (FN) and fault negatives (FN), are used to evaluate the experiments precisely. FN is the ratio of the number of the VoIP flows that are wrongly identified to the total number of VoIP flows FP is the ratio of the number of the non-VoIP flows that are mistakenly classified to the total number of the non-VoIP flows. They can be computed by (4) and (5), respectively. N is the total number of traffic flows while N_v is the number of VoIP flows. The number of identified VoIP flows is denoted with N_E and the number of correctly identified traffic flows as the VoIP flows is denoted with N_{EC}. Better performance can be obtained when both values of FP% and FN% are small.

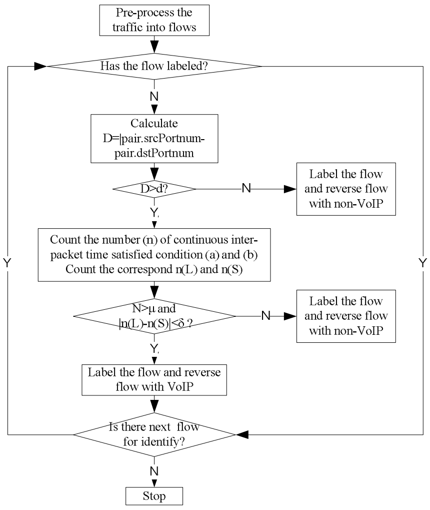
Pre-process the traffic into flows
Has the flow labeled?
N
Calculate D=|pair.srcPortnum-pair.dstPortnum
D>d?
N
Label the flow and reverse flow with non-VoIP
Y,
Count the number (n) of continuous inter-packet time satisfied condition (a) and (b) Count the correspond n(L) and n(S)
N>μ and |n(L)-n(S)|<δ ?
N
Label the flow and reverse flow with non-VoIP
Y,
Label the flow and reverse flow with VoIP
Is there next flow for identify?
Y
Y
N
Stop

Figure 2 Flowchart of our scheme

$$FP\% = \frac{N_E - N_{EC}}{N - N_V} * 100\% \tag{4}$$

$$FN\% = \frac{N_V - N_{EC}}{N_V} * 100\% \tag{5}$$

Our experiments have been divided into two parts: (A) experiment on the VoIP traffic to validate FN, (B) experiment on the non-VoIP traffic to validate FP.

A. Experiment on Non-VoIP Traffic

The trace of OC 48, provided by CAIDA [9], has been selected as the non-VoIP traffic set. Its description has been presented in TABLE I. The trace traffic has been captured on a west coast peering link for a large ISP.

The experiment results have been presented in TABLE II. The values of FP% in both our scheme and FLB are very small, especially in our scheme. It shows that both our scheme and FLB could efficiently filter out non-VoIP traffic. In the experiment our scheme keeps accurate identification for all flows while the FLB takes mistakes for 20 flows.

B. Experiment on VoIP Traffic

In this experiment, Skype, as the most popular and successful VoIP application over the Internet, has been selected to perform the traffic identification.

The trace file, named PT Skype, contains the Skype voice and video traffic which was captured on the main link of Politecnico di Torino (PT). Its detailed description has been listed in the TABLE III. Before the experiment, signaling traffic has been filtered out from the original file since we will perform traffic identification based on the voice and video traffic flows.

TABLE I TRACE BENCHMARK OF NON-VOIP

Trace Name	**Start Time**	**End Time**	**Packet Size**	**Trace Size**
OC 48	2003-04-14 07:00 UTC	2003-04-14 07:05 UTC	48 Bytes	593MB

TABLE II EXPERIMENT RESULT ON NON-VOIP TRACE

Trace Name	**Flow Number**	**Identified Flow Number**		**FP%**	
		Our scheme	**FLB**	**Our scheme**	**FLB**
OC48	6529	0	20	0	0.31

The experiment results have been presented in TABLE IV. The value of FN% in our scheme has remained small (9.72%) while that in FLB has been very large (44.4%). It shows that our scheme can identify VoIP traffic with high accuracy and outperform FLB. The terrible value of FN% in FLB implies that the characteristics selected by FLB are not shared by all VoIP traffic flows. And at the same time the impact of other network applications to VoIP traffic is not considered in FLB. This makes the result more inflexible. So many of flows are taken mistake to identify as Non-VoIP traffic. In contrast, our scheme has good advantage. It exhibits higher flexibility and could effectively filter out most of the disturbance brought out by the complicated Internet. In the transmission over the Internet, the relative values of inter-packet time nearly still maintains stable without much change. And at the same time, the self-adaptive estimated value on *EL* and *ES* can adapt the complicated Internet. The technique to take relative value and self-adaptive estimated value as identification criterion makes our scheme more advantage than the FLB and other existing schemes.

In the PT Skype trace, there exist 14 fault negative flows detected by our scheme. The analysis on the 14 flows has revealed that at least, 6 flows contain many larger packets, which are different from normal audio packets. They are supposed be the video packets. The identification accuracy will be much affected by the additional video traffic with audio and video data packets integrated into the same flow. On the other hand, the increase of the traffic load in the Internet could cause the congestions seriously to reduce the accuracy of the traffic identification. The 14 flows concentrate at two time periods. In fact 8 flows distribute in the 10:00 to 11:00 CET and 6 flows distribute in 14:00 to 15:00 CET. During the two periods the traffic load in Internet is high.

We have further studied the impact of the different codecs to the identification of the VoIP traffic by our scheme. We have four popular codecs selected for the examination. They are iSAC, EG711 A, EG711 U and iPCMwB, proprietary solutions of Global IP Sound [29]. The experiment result has been presented in TABLE V. It is clear that our scheme can identify VoIP flows that are encoded by popular codecs with very low FN% values. For the codecs of iSAC over TCP and EG711-U, the FN% values can even reach to zero. This reveals that our algorithm hassuccessfully extracted the common features of different codecs and is able to identify flows with various popular codecs.

TABLE III TRACE BENCHMARK OF VOIP

Trace Name	Traffic Type	Start Time	End Time	Packet Size	File Size	Flow Number
PT skype	Skype voice and video flows	2006-05-29 10:01 CET	2006-06-02 10:00 CET	38 Bytes	1.3 GB	144

TABLE IV EXPERIMENT RESULT ON VOIP TRACE

Trace Name	Flow Number	Identified Flow Number		FN%	
		Our scheme	FLB	Our scheme	FLB
PT skype	144	130	80	9.72	44.4

TABLE V EXPERIMENT RESULT ON VARIOUS VOIP CODEC

Codec	File Size	Flow Number	Identified Flow Number	FN%
iSAC over UDP	8.3MB	74	72	2.73
iSAC over TCP	7.3MB	60	60	0
EG711 U	22MB	56	56	0
EG711A	16MB	60	58	3.33
iPCMwB	24MB	58	54	6.98

V. CONCLUSION

In this paper, we have designed new models to describe the host behavior and traffic flow behavior for the purpose to identify the VoIP traffic. Then, based on it, as our major contribution, we have proposed a new traffic identification algorithm to identify VoIP traffic at transport layer. Extensive experiments have shown that our proposal could obtain a good performance with high accuracy of the traffic identification. At the same time, the proposed scheme could maintain its validity, when the existing VoIP applications are updated or new ones admitted. Both of the accuracy and flexibility of the traffic identification have been improved by our algorithm.

ACKNOWLEDGMENT

This research was supported in part by China High Technology Plan (863 grant No. 2008AA01A320) and the Tianjin NSF Project (grant No. 08JCYBJC14200).

REFERENCES

[1] Internet Assigned Numbers Authority, http://www.iana.org

[2] A. W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques", *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, 2005, pp. 50-60.

[3] T. Auld, A. W. Moore and S. F. Gull, "Bayesian Neural Networks for Internet Traffic Classification," *IEEE Transactions on Neural Networks*, Vol. 18, January 2007, pp. 223-239.

[4] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow Clustering Using Machine Learning Techniques," *Proceedings of Passive and Active Network Measurement*, April 2004, pp. 205-214.

[5] T. Okabe, T. Kitamura, and T. Shizuno, "Statistical Traffic Identification Method Based on Flow-Level Behavior for Fair VoIP service," *Proceedings of IEEE Workshop on VoIP Management and Security*, April 2006, pp. 35-40.

[6] T. Karagiannis, A. Broido, M. Faloutsos, and K. claffy, "Transport Layer Identification of P2P Traffic," *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement,* 2004, pp. 121-134.

[7] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, 2005, pp. 229-240.

[8] K. Xu, Z. Zhang, S. Bhattacharyya, "Profiling Internet Backbone Traffic: Behavior Models and Applications," *Proceedings of 2005 conference on Applications, technologies, architectures, and protocols for computer communications,* 2005, pp. 169-180.

[9] CAIDA web site, http://www.caida.org/home