Traffic classification based on Port Connection Pattern

Guang Cheng , Song Wang
School of Computer Science & Engineering, Southeast University,
Key Laboratory of Computer Network and Information Integration, Ministry of Education
Key Laboratory of Computer Network Technology of Jiangsu, Nanjing 210096, China
gcheng@njnet.edu.cn

Abstract— **While the traditional Web, Email and other Internet applications are still used large-scale, Internet applications, such as typical P2P resource sharing, streaming, games, instant messaging and other new applications emerging, is generating the rapid growth traffic. Connection pattern approach is to observe and identify connection patterns of host behavior at the transport layer. This method has no access to packet payload, no knowledge of port numbers and no additional information other than what current flow collectors provide. But traffic classification method based on host connection pattern can't distinguish some similar applications provided by one server. This paper analyzes the connection pattern between servers for SMTP, POP3, Web, BT traffic, and the distribution of simultaneous connections for each server port, and then propose a new traffic classified approach based on port connection pattern and the simultaneous connection number. The classified method can distinguish WEB, SMTP, POP3, and BT traffic each port in a server without any knowledge of port. The experimental result shows that this method is effective.**

Keywords- traffic classification; port connection pattern; simultaneous connection number

I. INTRODUCTION

While the rapid development of Internet, P2P, streaming video and other new applications are appeared, traffic in the Internet have translated from simple data to multiple types of information, and network traffic and behavior changed greatly. Traffic classification method is to infer the application type from raw traffic. Analyzing the distribution of various applications can strengthen network management, to improve network security and network utilization by optimizing network configuration. Traditional traffic classification methods are mainly base on TCP/UDP port, deep packet inspection [1] and machine learning methods [2]. Because there are some deficiencies in traditional identification methods, traffic classification algorithm based on host connection patterns of interaction has become a main research field in recent years [3]. This method established the visual representation of transport-layer interactions for various applications, classifying traffic by matching these graphs. It's difficulty to distinguish several applications on the same host using this method.

In this paper, we analyze the interaction of POP3, SMTP, Web, and BT, to find that these applications have the following differences (The "server" in BT refers to the "hot host", what means the host with more connected peers): (1) **Interaction Behavior between servers:** Mail Server needs connecting to other servers when using SMTP protocol sending E-mail, the connection is a one-way transmission, and destination port generally fixed; BT is two-way transmissions between the hot hosts.2) **the number of Simultaneous Connections:** Client connects to a server port with multiple ports, so the number of client ports is the number of simultaneous connections. Because webpage usually contains plenty of elements, so the browser generally establishes multiple connections, which can fetch different elements simultaneously in order to decrease page loading time. However, the simultaneous connections number of Mail server and BT is close to 1. To combine these differences, we propose a new traffic classification approach based on port connection pattern, which can distinguish variety applications on one server.

This paper is organized as follows. Section II gives the related work about traffic classification. Section III analyzes the features both connection pattern between severs and simultaneous connection. We propose the traffic classification method based on the connection pattern and simultaneous connection in Section IV. Experimental evaluation based on the proposed method is presented in Section V. The conclusion is given in Section VI.

II. RELATED WORKEDS

Traditional Internet applications are used by the IANA under a unified service fixed port [4], the majority of standard applications are known to use a fixed port to provide connectivity services, such as Web use port 80, DNS uses port 53, and Telnet uses port 23 and so on. The first generation P2P also uses non-uniform fixed ports for data transfer. Deep Packet Inspection (DPI) uses protocol analysis and data reduction technique to extract the application layer and then match some special string to determine the specific application of network traffic. Subhabrata Sen [5] analyzed the proposal features of Gnutella, eDonkey, Direbteonnect, BitTorrent, and KazaA. Ohzahata [6] introduced the PZP system winny, and Kang [7] detected the chat packets. At present, many commercial and open source application identification solutions are based on the DPI methods, including the L7-filter [8], Microsoft common application signatures [9], Cisco's products DPI technology [10]. The DPI and port based

978-1-4244-9763-8/11/$26.00 ©2011 IEEE

approaches are easy to identify well-known traffic, but cannot find unknown one.

Machine learning methods, supervised learning methods and unsupervised clustering methods, have been applied to traffic classification. Supervised learning methods included neural network [11], Bayes theory [12], and Support Vector Machine [13]. Unsupervised clustering methods had K-Means [14], AutoClass, DBSCAN [15], and entropy-based profiling [16]. However, machine learning methods depend on training dataset correctly, which is difficult to construct for real Internet traffic. BLINC [3] focuses on communication flow structure among hosts to propose a multilevel traffic classification.

III. TRAFFIC FEATRUE ANALYSIS

This paper analyzes the network traffic trace of WEB, POP3, SMTP and BT, to summarize the characteristics of the interaction behavior between servers and the number of simultaneous connections. On this basis, a classification algorithm is proposed to identify traffic in each port.

A. Experimental Data

We collected the experimental data from the two different aspects. The first kind data are collected from a 10Gbps backbone link in China Education Research Network (CERNET) Southeast China (North) regional network center. These data with 112GB, collected from 2008-08-20 15:00:00 to 16:00:00, are pre-identified by L7-filter. L7-filter is a classifier for Linux's Netfilter that identifies packets based on application layer data. It can classify packets as Kazaa, HTTP, Jabber, Citrix, Bittorrent, FTP, Gnucleus, eDonkey2000, etc.

The second method is to collect data from the local host. We designed a packet capture system, based on winpcap, which can obtain full message content and packet arrive time. Before calling capture module, first the system can obtain current process information, then get all connections and listening TCP/UDP ports, finally combine these information by PID (process ID) to find the process name of ports. The captured packets are saved into different files according to its process name. The packet capture system has collected over 100,000 flow records with 10GB traffic totally.

B. Connection Pattern Between Servers

Four kinds of applications, Web, SMTP, POP3, BT, will be analyzed according to the two team experimental data. From the traces, we know that the used ports in a mail server are greater than 2, because the mail server not only provides SMTP and POP3 service, but also needs to connect to other servers to send mail. The mechanism to delivery E-mail is that firstly user sends an email to server, and then server sends it to destination mail server. A source mail server connects a destination server with a random port in SMTP. Figure 1 is an example about the connection pattern of mail application.

The used ports in a web server are almost not more than 2 ports, that is 80 port and 443 port for encrypt. A web server rarely communicates others web servers. Since the downloaded elements are not necessarily in the same server, so in the web connection graph, there might be a user access to multiple Web servers at the same time. If lots of local ports of a host access the same port in the different destination hosts simultaneously, then we can judge that the destination port is likely providing web service. Figure 2(a) is an example about the connection pattern of web application example.

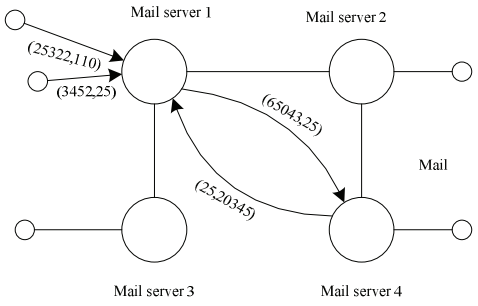

Figure 1. connection pattern of mail application example

The used ports in the BT hot host are more than 1. After detailed analysis of utorrent-2.0 packets, we found the port in utorrent preference using most traffic, and using both TCP and UDP protocols at the same time. We name this port as main-port. In addition, some other ports would also be used to connect peers. We can get the following rules: (1) All the destination IP addresses connect to a particular port of source IP (such as 10663), which is the port set in program preference; (2) The connection between destination IP and source IP may use two other ports, but destination port has to connect source main-port firstly. Figure 2(b) is an example about the connection pattern of BT application example.

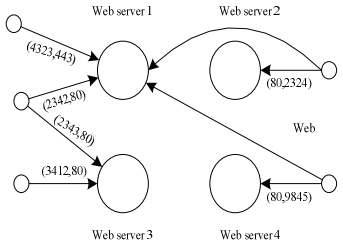

Figure 2(a)

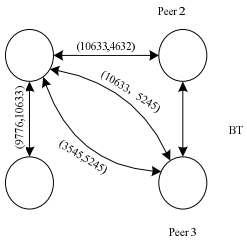

Figure 2(b)

Figure 2. connections pattern of Web and BT application example

C. The Number of Simultaneous Connections

Clients connect to a server port with multiple ports, and the number of client ports is defined to the number of simultaneous connections. We calculated simultaneous connections between each destination IP address and (source IP address, source Port) pair. the largest number of simultaneous connections in the (srcIP, Port) pair is the maximum simultaneous connection. In the program realization, we recorded duration of connections between the (srcIP, Port) pair and (dstIP, dstPort) pairs for each (srcIP, Port) pair. The largest number of connections in a moment is the maximum simultaneous connection. The distribution of simultaneous connections in the four kinds of applications is showed from Figure 3 to Figure 6.

The main port of BT will accept a large number of connections. But when local host connects to other hosts, every

port in the BT server only establishes one connection. In Figure 3, the port with the maximum number 2 is main-port. The average number of simultaneous connections is about 1.

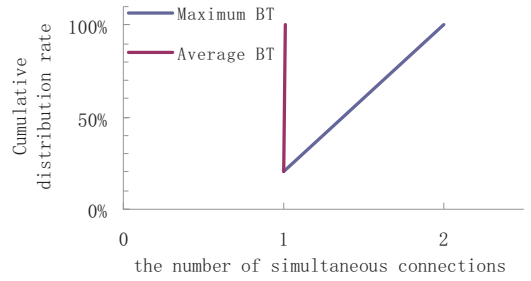

Figure 3. BT applications within 5 minutes maximum concurrent connection port and the average number of concurrent connections

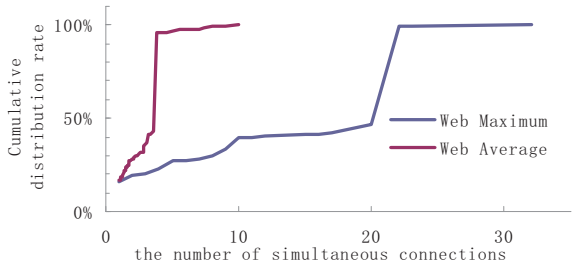

Figure 4. WEB applications within 5 minutes maximum concurrent connection port and the average number of concurrent connections

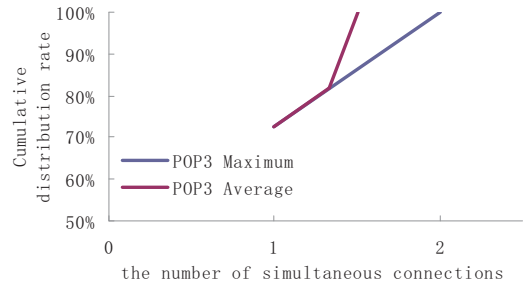

Figure 5. POP3 applications within 5 minutes maximum concurrent connection port and the average number of concurrent connections

Webpage usually contains many elements. In order to decrease page loading time, the browser generally establishes multiple connections, which can fetch different elements simultaneously. So the number of simultaneous connections of web is larger. Figure 4 shows that 75% packet of server port's maximum number of simultaneous connections is larger than 5, and about 70% packet of server port's average number of simultaneous connections is larger than 2.

Mail Server provides POP3 and SMTP. Since it's usually speed insensitive, it usually only create one connection at one time. The maximum number and the average number of simultaneous connections for most POP3 port is less than 2, 70% is only 1 in Figure 5. Simultaneous connections of SMTP are larger than POP3 in Figure 6. The maximum number is 5, and its average number is less than 2.

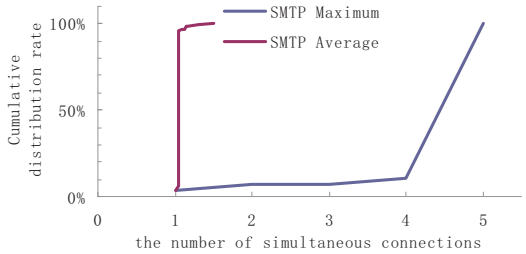

Figure 6. SMTP applications within 5 minutes maximum concurrent connection port and the average number of concurrent connections

IV. CLASSIFICATION ALGORITHM

We still use some conclusions given by Thomas [3]: Firstly, most IP addresses act as clients having the minimum number of destinations IP addresses. Thus, the identification of the small number of servers can retrospectively pinpoint the clients, and lead to the classification of a large portion of the traffic, while limiting the amount of associated overhead. Secondly, if a host uses a small number of source ports, typically less or equal to two for every flow, then this host is likely providing a service. Thirdly, we also use the average packet size of per-flow.

Base on the analysis of the experimental data, suppose the application type of (IP, Port) pair in a short time is fixed. Since user or server usually run multiple applications simultaneously, it's wrong to assume a host using only one application. But different applications using different ports in a short time mostly, so it's reasonable to assume (IP, Port) pair using only one application.

For a hot host in a network, given that the used port number is P, the number of flows connecting to it is F. The number of IPs connecting to the host is M. The average number of simultaneous connections is C. The maximum number of simultaneous connections is C_m. The number of IPs connecting to a port P_i is M_{pi}, and the number of ports connecting to this port is P_{pi}. The average length of outgoing flow is V_{pi+}, and the average length of incoming flow is V_{pi-}. If a port $M_{pi} > M_{ps}$, then the port is defined as a service port. We found M_{pi} value of client equals 1 generally, and M_{pi} of server is larger than 1, so in this algorithm M_{ps} is defined as 1. A hot host, referred to the hosts with large M, may be Mail server, Web server, or BT hot Peer. The hot host in this algorithm is defined a host with $M > M_d$. BLINC algorithm chose $M_d = 4$, so we also uses the same M_d for accuracy.

As following, we will give the traffic classification algorithm according to port connection pattern.

(1) First, find the following two cases: If a port P_i , its $M_{pi} > M_{ps}$, and exists multiple ports connecting to a same port of other host. For each port P_i in compliance with situation 1, we can do the following judgment: **First case)** if $C >= C_{web}$ and $C_m >= C_{webm}$, the hot host provides Web service, P_i is Web service port. **Second case**) If P_i complies both two cases, this port is BT port, P_i, ..., P_j are all BT ports.

(2) Secondly, if C <= C_{web} or C_m <= C_{web}, then to find out whether the hosts which connecting with P_i also use the same P_i as service port, if it is so, then P_i is mail port, and to decide the specific protocol type according to V_{pi+} and V_{pi-}. If it is not, to judge the application type according to V_{pi+} and V_{pi-} directly.

(3) Thirdly, in second case, the destination IP is named as a hot host, and the destination port as P_i is identified as the first case.

(4) After identify the application type of P_i, if pinpoint the ports connecting with P_i retrospectively, then the remaining connection can not be identified.

According to the analysis of the number of simultaneous connections, for BT and Mail, the average number is less than 2, and their the maximum number is not more than 5. So here C_{web} is 2, C_{webm} is 6. The algorithm can only distinguish between mail, web and bt.

V. EXPERIMENTAL ANALYSIS

We collected packets relating with a mail server as test set from CERNET data. This mail server is a webmail system providing smtp, pop3 and web services. We tried to use the algorithm to identify the test set. The result is shown in Table 2. The ports listed here is only to test the correctness of results.

TABLE I. RESULT OF MAIL AND WEB IDENTIFICATION

IP	PORT	APPLICATION
202.119.112.48	25	SMTP
220.181.12.75	25	SMTP
202.119.112.48	80	HTTP
202.119.112.48	110	POP3
209.85.217.28	50345,63103	SMTP(sender)
202.119.112.48	22	UNKNOWN

The algorithm successfully distinguishes SMTP, HTTP, and POP3 applications in this server (202.119.112.48). We also list the mail servers (220.181.12.75, 209.85.217.28) which connected to this server (202.119.112.48). It proves this algorithm works functionally. Besides it found this server also using 22 port as server port, the algorithms have not considered this application, so it didn't works.

Because web server or BT server rarely uses BT at the same time, so we used the utorrent data collected by winpcap to test the algorithm. Table 3 shows that our algorithm can identify utorrent port correctly with 97% ratio.

TABLE II. RESULT OF BT IDENTIFICATION

application	Identified port number	used port number	corrected ratio
utorrent	749	772	97%

VI. CONCLUSIONS

This paper considers the interaction between servers, and we propose a new traffic identification algorithm based on the port connection patterns and the number of simultaneous connections. It can identify variety applications on a same server. We analyzed the differences of servers' interaction of SMTP, POP3, Web, and BT, and summarized the differences of the number of simultaneous connections based on the actual data, then gave the identification algorithm. Finally the experimental result showed that this method is feasible in the actual data. In the future, we will continue to analyze the connection pattern and simultaneous of other applications, such as FTP, TELNET, and so on to classify more applications.

ACKNOWLEDGMENT

This work was supported by the National Grand Fundamental Research 973 program of China under Grant No. 2009CB320505, the National Nature Science Foundation of China under Grant No. 60973123, the national science & technology pillar program during the eleventh five-year plan period under Grant No. 2008BAH37B04, and Qing Lan Project.

REFERENCES

[1] Andrew M, Konstantina P. Toward the accurate identification of network applications [C] / / PAM, March 2005.

[2] Zhang Haining. The research on the application of flow feature selection in P2P traffic identification[D].Beijing: Beijing University of Posts and Telecommunications, college of Software Engineering, 2008:14 ~ 17.

[3] Thomas K, Konstantina P, Michalis F. BLINC: multilevel traffic classification in the dark [C].ACM SIGCOMM Computer Communication Review, v.35 n.4, October 2005.

[4] PORT NUMBERS, http://www.iana.org/assignments/port-numbers.

[5] Subhabrata Sen, Oliver Spatscheck, Dongmei Wang, Accurate, scalable in-network identification of p2p traffic using application signatures, Proceedings of the 13th international conference on World Wide Web, 2004: 512-521.

[6] S Ohzahata, Y Hagiwara, M Terada, K Kawashima, A Traffic Identification Method and Evaluations for a Pure P2P Application, Proceeding of PAM 2005.

[7] HJ Kang, MS Kim, JW Hong, Streaming Media and Multimedia Conferencing Traffic Analysis Using Payload Examination, ETRI Journal, 2004, 26(3): 203-217.

[8] Application Layer Packet Classifier for Linux, http://l7-filter.sourceforge.net/.

[9] Common Application Signatures , on Line 2009.4, http://technet.microsoft.com/en-us/library/cc302520.aspx

[10] White Paper, MANAGING PEER-TO-PEER TRAFFIC WITH CISCO SERVICE CONTROL TECHNOLOGY, on line 2009.4, http://www.cisco.com/en/US/prod/collateral/ps7045/ps6129/ps6133/ps6150/prod_white_paper0900aecd8023500d.pdf

[11] J. McHugh, R. McLeod, and V. Nagaonkar, Passive Network Forensics: Behavioural Classification of Network Hosts Based on Connecti on Patterns, ACM SIGOPS Volume 42 Issue 3, pp. 99–111, April 2008.

[12] A. W. Moore and D. Zuev, Internet Traffic Classification Using Bayesian Analysis Techniques, ACM SIGMETRICS'05, pp. 50–60, June 2005.

[13] H. Kim, kc claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Y. Lee, Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices, ACM CoNEXT 2008, December 2008.

[14] L. Bernaille and R. Teixeira, Early Recognition of Encrypted Applications, PAM 2007, pp. 165–175, April 2007.

[15] J. Erman, M. Arlitt, and A. Mahanti, Traffic Classification Using Clustering Algorithms, ACM SIGCOMM'06 Minenet workshop, pp. 281–286, September 2006.

[16] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, Profiling internet backbone traffic: behavior models and applications, ACM SIGCOMM'05, pp. 169–180, 2005.