# Identifying BT-like P2P Traffic by the Discreteness of Remote Hosts

W.Q. Cheng[1,2], J. Gong[1], W. Ding[1]
[1]Computer College, Southeast University
[2]Computer College, Nanjing University of Posts and Communications
Nanjing, Jiangsu Province, PRC
Email: wqcheng@njnet.edu.cn, jgong@njnet.edu.cn, wding@njnet.edu.cn

*Abstract*—**By analyzing application protocols and traffic, we find that the most striking distinguish between BitTorrent (BT)-like peer-to-peer (P2P) applications' traffic and traditional as well as other P2P (such as Skype) applications' traffic of a single user may be the dissimilarity in the distribution of remote hosts involved. Therefore, we propose a method based on Discreteness of Remote Hosts (RHD) to identify BT-like traffic. In this method, traffic for each user host in a stub network need be monitored at the border of the stub network and classified into flows. At intervals concurrent TCP and UDP flows for a single host should be grouped respectively by what stub network the remote host of each flow belongs to, and then calculate instant RHDs for TCP and UDP flows respectively. For any user host, if the sum of two average RHDs for a period of time exceeds specific threshold, then we can deduce that the host has used BT-like P2P application. The method proposed here is a simple traffic characteristic-based traffic classification method. It is more suitable for identifying protean BT-like P2P application than usual content-based methods such as those based on port numbers or application signatures. Experiments results reveal that our method can effectively recognize BT-like traffic and may be particularly appropriate for use to restrict BT-like traffic during working hours if needed.**

***Keywords-traffic identification; traffic characteristics; P2P; concurrent flows; discreteness of remote hosts***

## I. INTRODUCTION

Today exist many popular peer-to-peer (P2P) systems tailored for sharing large files, network TV, or music on the Internet, such as BT, PPLive˒ eMule, FastTrack, eDonkey, PPStream, KuGoo[1-3]. These P2P software can overcome the limits of the traditional (Client/Server) download mode, and the more users downloading the same file or enjoying the same network TV or music performance, the faster download bit rate or the more fluent play users will get. The P2P systems can provide fast sharing of information resource by sufficiently exploiting the peer communication capability to occupy more bandwidth than traditional applications. To guarantee the use of critical applications, some ISPs, enterprise networks or campus networks may hope to limit the use of BT-like applications during working hours or rush hours on Internet. To describe conveniently, we classify BT-like P2P applications that usually establish as many concurrent P2P connections as possible during sharing contents (large files or rich media) as class I, while those that often establish few or a few concurrent P2P connections as class II, such as QQ (a popular P2P application in China, providing instant messaging and voice services). It is unnecessary to forbid class II P2P applications even during rush hours since they only occupy a little bandwidth in general. Due to the high cost, it is not very possible and economic to restrict P2P efficiently at the core network. It may be a sensible choice to control BT-like traffic at the borders of stub networks, where it is convenient to enforce policy-based control of traffic on host granularity with acceptable control costs. This paper is to manage to identify which hosts have been generating class I P2P traffic, so as to render the restriction of class I P2P traffic on host granularity possible.

Recently, the commonly used methods for application recognition or traffic identification are content-based, such as based on port numbers or application signatures [3-6]. But, due to the randomness of design and implementation of P2P protocol and software, as well as the lack of adaptation and scalability of these methods, identification rules or even identification software must be updated correspondingly to recognize new versions of known P2P traffic. Moreover, these methods are usually incapable of identifying encrypted or unfamiliar P2P traffic.

In this paper we propose a new method that takes comparatively steady non-content characteristic of application, traffic characteristic as the basis to identify class I P2P traffic. By the observe and analysis of single host's traffic, as well as the analysis of the protocols of applications, we found that the maximum distinguish between BT-like traffic of single user and the traffic of traditional application as well as class II P2P application may lie in the distribution characteristic of remote hosts involved. Therefore, we define a simple metric "Remote Hosts' Discreteness" (RHD) to be used to discern whether user traffic contained class I traffic therein. Practical tests reveal that BT-like traffic can be detected quite soon using this method with fairly high accuracy.

## II. RHD-BASED IDENTIFICATION OF BT-LIKE TRAFFIC

### A. Definitions of TCP/UDP flows

Both traffic characteristic analysis and identification of class I P2P traffic on the borders of stub networks, need classify packets into flows. In this paper, a ***flow*** is defined by 5-tuple {***local IP, local port, remote IP, remote port, protocol***}, and a flow is considered to have expired if no packets belonging to the flow have been observed for a certain period of time.

IP packets that shuttle between specific local endpoint (local IP, local port) and specific remote endpoint (remote IP, remote port), carry transport-level PDU (protocol data unit), and arrive under specified timeout constraints belong to a TCP or UDP flow. An IP packet not belonging to any active flow (see below) will belong to a new TCP or UDP flow. A flow has two states:

(1) **S_ACTIVE**, i.e. active state. The state for a new flow is S_ACTIVE, and it remains unchanged until no new packet of the flow arrives for longer than specified timeout interval.

(2) **S_TIMEOUT**, i.e. timeout state. If no new packet of a flow arrives for longer than specified timeout interval, the state of the flow turns to S_TIMEOUT. The flow timeout interval for both TCP and UDP flows can be set as 4 seconds.

Flows with the state of S_ACTIVE are termed ***active flows***, and active flows that coexist at a time are called ***concurrent flows*** at that time.

### B. *Definition of remote hosts' discreteness (RHD)*

With regard to concurrent flows of a single host, the more proportion of flows of which remote hosts belong to the same stub network, the less discreteness of these flows' remote hosts. Referring to the entropy principle in information theory and communication theory, we define Remote Hosts' Discreteness (RHD) of concurrent TCP or UDP flows as follows:

$$D_t = \frac{1}{n}\sum_{i=1}^{m}\log_2(n/x_i) \qquad (1)$$

where $n$ denotes number of concurrent flows at time $t$, $m$ denotes number of stub networks that remote hosts of the flows belong to (obviously, $m \leqslant n$), and $x_i$ denotes number of flows with their remote hosts residing in network $i$. The network prefix length of stub networks can be set as 23. The RHDs should be calculated for concurrent TCP or UDP flows respectively.

### C. *RHD comparison between various application traffic*

#### 1) *RHD for BT traffic of a single host*

The number of concurrent flows for BT traffic of a single host fluctuates with the amount and states of peers, and there are great probabilities that RHD has fairly high numerical value anyway.

#### 2) *RHD for traditional and class II P2P application traffic of a single host*

During a user visiting a Web site, the number of concurrent flows may be great at some times, but the RHDs are always low because different flows often have the same remote host and remote port. During a user accessing a FTP server, the number of concurrent flows remains few, and the RHD remains 0 due to duplicate remote host for both control and data connections. The RHD will increase when a user accessing multiple FTP servers. However, this case seldom occurs. The reasons may be that seldom users have such habit, and the obtainable data rates usually dissatisfy users. The RHD for Email traffic of a single host is also low. As to QQ or Skype traffic of a single host, the RHD usually remains low too and moderate stable, in that only a small amount of concurrent flows exist therein.

### D. ***Algorithm***

Our RHD-based method is built on the analysis of RHD and other characteristics of various application traffic on a single host, as well as user behavior characteristics. The method needs to monitor traffic of every internal host at the border of a stub network. The algorithm is described as follows:

**Criteria 1 (C1):** ***Average RHD-based identification***: If the sum of average RHD for TCP flows and that for UDP flows of a host's traffic during an measurement interval (e.g. $\boldsymbol{T}$ =10s) is greater than threshold $D_{sumOfAvg}$ (less than $D_{sum}$, e.g. 2.3), we assert that BT-like traffic is contained in the host's traffic.

*1) Monitor every in/outbound IP packet and classify each of them into a specific flow based on the 5-tuples and flow timeout.*

Flows are grouped based on local host (local IP). A flow record includes flow keys, start time of the flow, the arrival time of the last packet of the flow. No packet payloads need be recorded.

*2) For* ***each active host H*** *during* ***every*** *measurement interval (****T*** *seconds) do:*

*a) At every* **G** *seconds (***G<T***/20):*

i) Check whether each active TCP or UDP flow of host ***H*** has timed out, and update the state of timed out flows; calculate instant RHD of concurrent TCP and UDP flows of the host **respectively**, and RHD is taken as zero if number of concurrent flows is zero;

ii) Update average RHD of TCP and UDP flows of the host **respectively**.

*b) Adopt Criteria 1 to distinguish whether BT-like P2P traffic is contained in current traffic of host H.*

*c) Delete information of timed out flows.*

The method proposed in this paper is used to identify whether traffic of user hosts (excluding traditional servers) in a stub network contains BT-like traffic. The traffic of traditional servers (e.g. Web, FTP, Email, TELNET servers) should be omitted.

## REFERENCES

[1] http://www.pplive.com/zh-cn/index.html.

[2] Dejan S. Milojicic, Vana Kalogeraki, Rajan Lukose, etc. Peer-to-Peer Computing, HP Laboratories Palo Alto, HPL-2002-57.

[3] Cisco MAR. Network-Based Application Recognition and Distributed Network-Based Application Recognition [EB/OL]. http://www.cisco.com/en/US/products/ps6616/products_ios_protocol_group_home.html. 2005-8-26.

[4] TS Choi, CH Kim, SH Yoon, et al. Content-aware Internet application traffic measurement and analysis[C]. In: 2004 IEEE/IFIP Network Operations and Management Symposium (NOMS 2004), 2004. 511~524.

[5] Andrew W. Moore and Konstantina Papagiannaki. Toward the accurate identification of network applications[C]. In: Passive & Active Measurement Workshop 2005 (PAM2005), Boston, MA (2005).

[6] S. Sen, O. Spatscheck, and D. Wang. Accurate, scalable in-network identification of p2p traffic using application signatures[C]. In: Proceedings of the 13th international conference on World Wide Web (WWW 2004), 2004. 512~ 521.