

# 基于主被动连接的 P2P 节点识别算法

程春玲, 周 芸, 徐小龙

(南京邮电大学 计算机学院, 江苏 南京 210003)

**摘 要:** P2P 应用近年来取得了飞速的发展, 在推动 Internet 发展的同时也带来了大量带宽占用和网络安全问题。P2P 节点的识别是 P2P 流量检测中一类重要的方法。通过理论和实验两方面分析了基于连接成功率的节点识别方法对 P2P 流媒体识别存在误判率高的缺点, 并在此基础上提出基于主被动连接的识别算法。该算法在基于连接成功率算法基础上, 通过计算节点在单位时间内, 被动连接和主动连接数的比值进一步判断。实验结果显示, 改进后方法比原方法有更小的误判率和漏判率, 提高了检测的准确度。

**关键词:** P2P; 节点识别; 连接成功率; 主动连接; 被动连接

**中图分类号:** TP393

**文献标识码:** A

**文章编号:** 1673-629X(2010)12-0050-04

## Identification Algorithm of P2P Peers Based on Initiative and Passive Connections

CHENG Chun-ling, ZHOU Yun, XU Xiao-long

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:** With the wide application of the P2P, it has brought bandwidth occupation and safety problems. The identification of P2P peers is one of the important methods in P2P measurement. Firstly, analyzes the "connection responded success rate" in theory, then find the high false positive in identification of P2P streaming media by experiment. So the identification algorithm of P2P peers based on initiative and passive connections is presented. The algorithm computes the ratio of passive connections and initiative connections in unit time to identify P2P peer after connection responded success rate method. The experiment results show that the algorithm proposed in this paper can identify P2P peers more accurately, with less false positive and false negative.

**Key words:** peer-to-peer; node identification; connection responded success rate; initiative connection; passive connection

## 0 引 言

随着 P2P 应用的不断扩展, P2P 业务不仅充分利用了网络带宽也过度消耗了网络资源。据统计, P2P 业务的带宽占用比率在高峰期达到 80%~90%<sup>[1]</sup>, 因此精确地识别 P2P 流量对于有效管理网络、合理利用网络资源具有重要意义。P2P 节点的识别作为 P2P 流量识别技术中的重要部分, 其准确识别可以有效地定位 P2P 流量, 从而在网络繁忙时段能够对其加以限制, 合理分配带宽资源。

现有的 P2P 节点识别方法有基于端口、基于应用签名和基于连接特征三类<sup>[2~4]</sup>。其中, 基于连接特征

的 P2P 节点识别方法主要依据 P2P 节点有别于传统节点的一些基本连接特性, 在使用较少传输层信息的情况下, 简单、快速地识别 P2P 节点。根据利用的连接特征不同, 主要分为利用节点同时作为服务器和客户机的双重角色特征、利用节点所连接的主机数量和利用上下行流量比三种。除此之外, 还有一些其它方法和综合方法, 例如: Karagiannis 等人<sup>[5]</sup>通过对 Gnutella、eDonkey、FastTrack 等著名 P2P 应用研究发现同时使用 TCP 和 UDP 传输协议的多为 P2P 节点; 文献[6]则认为在短时间发出大量连接的可判断为 P2P 应用; 文献[7]依据 P2P 节点在单位时间内连接的目的子网数和单位时间内连接的目的 IP 数与有效连接数的比值都大于普通节点的特性, 提出的 P2P 节点识别法大大减少了误判和漏判取得较好的识别效果; 文献[8]则采用了一个基于应用层有效载荷特征和传输层流量特征的双层特征的高效混合检测方法; 文献[9]提出一种被称为 "connection responded success rate"

收稿日期: 2010-03-23; 修回日期: 2010-06-11

基金项目: 教育部博士点基金(20093223120001); 江苏省科技支撑计划(BE2009158); 江苏省高校自然科学基金项目(09KJB520010); 教育部专项研究课题(2009117)

作者简介: 程春玲(1972-), 女, 副教授, CCF 会员, 研究方向为对等计算、信息安全、数据挖掘等。

(以下称“连接成功率”)的方法,基于P2P网络的高度分散、自我组织的本质特性来检测P2P节点。

连接成功率法使用的传输层信息较少、计算简便,但其以P2P文件共享应用为例,而未考虑P2P另一类重要应用——P2P流媒体。文中通过理论和实验两方面分析,发现该方法对传统的P2P文件共享系统检测准确率较高,而在以PPStream为代表的流媒体中则存在较高的误判。因此文中对文献[9]中基于连接成功率的P2P节点识别方法进行改进,提出基于主被动连接的P2P节点识别方法,以提高检测的准确度。

## 1 基于连接成功率方法的不足

文献[9]中基于连接成功率的P2P节点识别算法的依据是:P2P网络具有高度动态性,系统为了减少主机随机退出给P2P网络连通性带来的影响以及保证自身稳定的传输速度,需要频繁地连接其它节点。又由于被选连接节点的随机性,因此,P2P节点的连接成功率低下,而普通非P2P节点由于主要连接对象是Internet上各种服务器,C/S模式下的服务器为了保证服务可获得而基本处于稳定状态,因此其连接成功率相对较高。因此,文献[9]通过计算一定周期内节点的TCP连接成功率来判断该节点是否P2P节点。连接成功率 $R_C$ 为:

$$R_C = b/a \quad (1)$$

其中, $b$ 为监测主机对不同目的IP发送的SYN包的数目, $a$ 为该主机从不同源IP收到的SYN/ACK包的数目。

基于连接成功率的P2P节点识别方法对BT和Emule两种P2P文件共享应用检测准确率较高,但却忽视了对P2P流媒体的考虑,如校园网中常用的PPstream。不同于在BT对等点之间只使用TCP通信,PPstream、PPlive等网络电视的控制信息是使用UDP协议,而视频数据传输PPstream使用TCP协议<sup>[10]</sup>,PPlive则既利用TCP也利用UDP。由于使用了UDP传输控制信息确认节点是否拥有所需信息,使得TCP连接成功率偏高。另外,P2P流媒体区别于P2P文件下载应用的一个特征是P2P流媒体节点连接上足够的播放源后,在播放缓冲区中数据不足之前,很少主动向外发起连接。这种情况下,根据主动发起TCP连接的回应率来判断P2P节点的连接成功率法,其检测精度必然将大大降低。

通过对校园网内流量数据进行一个月的监控观察发现:BT下载一直占总应用带宽的70%左右,以PPStream为主的流媒体应用占10%以上,HTTP占不到10%。文中以BT和PPstream为例,对文件共享和流

媒体这两种在校园网中消耗主要带宽的P2P应用进行实验,使用wireshark1.1.2在校园网环境下对多台机器进行数据抓包,统计P2P节点和HTTP节点在2h内的连接成功率( $R_C$ 值),结果如图1所示。

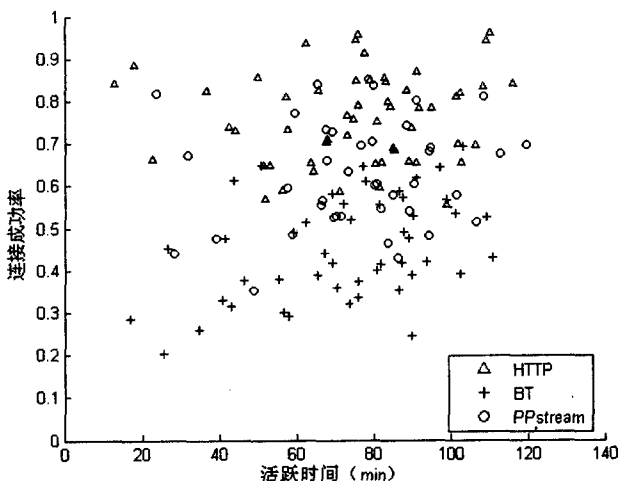


图1 P2P和HTTP节点 $R_C$ 值统计分布

从图1中可以看出,相比P2P节点,HTTP节点的 $R_C$ 值确实如前面的分析,取值较大,普遍集中在[0.6, 0.9]之间,BT节点基本在[0.25, 0.65]之间,实验结果范围与文献[7]中一致。而对于PPstream节点,从图中可以观察到:其也有较大的连接成功率, $R_C$ 值在[0.55, 0.9]之间。

由此可见,基于连接成功率的方法对于BT、eMule等文件传输型P2P应用,其检测准确率相当高。因为系统为获取更快、更稳定的下载速度会不断地尝试连接更多的节点,从而导致了P2P节点连接成功率的持续低下。而以PPstream为代表的一类P2P流媒体在网络条件好的情况下,则不需不断发起新连接从而使其连接成功率普遍偏高。这时如果仅靠连接成功率来判断势必出现很多误判,所以文中提出一种基于主被动连接的P2P节点监测方法,首先采用连接成功率进行初判,然后对可疑节点采用主被动连接数方法加以精确检测。

## 2 基于主被动连接的P2P节点识别算法

### 2.1 算法依据

在P2P网络中,节点不仅作为客户端主动(Initiative)向别的节点发起连接(连接其它节点的监听端口),同时还会承担服务器的功能,有大量被动(Passive)连接(被其它节点连接其监听端口)。针对P2P应用的平均连接时间比较长的特征,Raffaele Bolla, Marco Canini在大量实验的基础上提出P2P流量应分为信号流(signaling)和数据下载流(downloading)两个阶段<sup>[11]</sup>。在这两个阶段所展现的P2P特征有显著的不

同,数据流比信号流更稳定,且更能反映当前的连接模式。文献[10]通过实验发现在校园网内,以 PPStream 为代表的一类流媒体的上传流量明显大于下载流量,客户端向对等点上传数据比从节点下载要分散的多,即:客户端向大多数的对等点上传数据,这种情况下必然存在大量被动连接。

从总体上看, P2P 节点的主被动连接数相差较小。而没有运行 P2P 应用程序的主机要么只是单纯作为客户端,要么只是单纯作为服务器,它们的主被动连接数是不平衡的,表现出比较大的差值。因此,统计在单位时间内主机的被动连接数与主动连接数的比值,并与使用传统网络应用的主机的实验观测值作比较,就可以判断该节点是否是 P2P 节点。

## 2.2 主被动连接率的计算

根据以上分析并考虑实际中多用 TCP 进行数据传输,所以文中主被动连接的计算也是针对 TCP 而言。定义主被动连接率为:节点的被动连接数和主动连接数的比值,表示为  $R_{PI}$ ,则:

$$R_{PI} = C_P / C_I \quad (2)$$

其中  $C_P$ 、 $C_I$  分别为单位时间内监测主机的被动连接数和主动连接数。考虑到 P2P 应用一般都会提供一个监听端口供其他主机连接,因此在实际计算中,被动连接数  $C_P$  取值为一个周期内与被检测主机同一端口相连的不同 IP 数目。主动连接数  $C_I$  则取值为一个周期内检测主机主动发起连接的目的 IP 数目。

对于 P2P 节点,  $R_{PI}$  应在一个稳定的范围内,即:存在阈值  $\alpha$  和  $\beta$ ,使得  $\alpha < R_{PI} < \beta$ 。对于阈值  $\alpha$  和  $\beta$  的确定,文中通过对校园网中 P2P 主机和非 P2P 主机(HTTP 客户端主机)使用 wireshark1.1.2 在不同时段多次抓包数据进行分析,结果如图 2 所示。

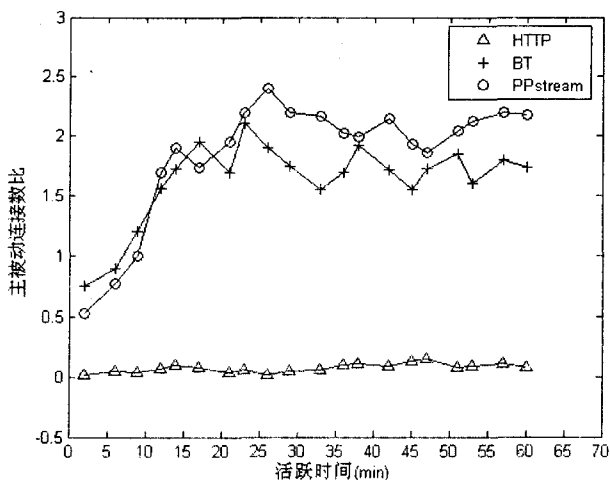


图 2 P2P 节点和非 P2P 节点平均  $R_{PI}$  值统计分布

从图 2 可以看出, P2P 节点的  $R_{PI}$  值明显大于 HTTP 节点,传统非 P2P 客户端节点因为几乎没有被

动连接而使  $R_{PI}$  值几乎为 0。文中根据多次实验结果,将阈值  $\alpha$  取为 0.5,  $\beta$  取为 2.5 时,检测准确度较高,使得在此范围内被判断为 P2P 节点的误判和漏判综合最小。

## 2.3 算法描述

考虑到基于连接成功率识别算法的计算量比基于主被动连接数的计算量小,且对 BT 等传统 P2P 文件共享应用的识别比较准确,所以,为减少识别时间,文中将二者融合,即:先用基于连接成功率算法加以判断,然后对用连接成功率算法无法准确判断的可疑节点再采用主被动连接数算法进一步识别。这里,可对连接成功率  $R_C$  的取值设置一阈值  $\mu$ ,使得对于实际  $R_C$  值低于  $\mu$  的节点可正确判断为 P2P 节点,而高于  $\mu$  的为可疑节点,有待进一步判断。阈值  $\mu$  的取值较低时,可提高判断的准确率;而取值较高,可能存在一定误判,但识别速度较快。文中的  $\mu$  值是通过多次实验获得的,取值为 0.6。

具体识别算法如下:

1) 对于待检测节点首先按公式(1)周期性计算平均连接成功率  $R_C$  值。对于  $R_C$  值低于阈值  $\mu$  的判断为 P2P 节点。

2) 对于  $R_C$  值大于阈值  $\mu$  的列为可疑节点。

3) 对可疑节点采用公式(2)作进一步判断,计算其在本周期内的主被动连接率  $R_{PI}$ 。如果其值在  $[\alpha, \beta]$  区间内,则认为该节点为 P2P 节点,否则为非 P2P 节点。

在实际应用中,主被动连接数法判断 P2P 节点的上下限阈值  $\alpha$ 、 $\beta$  和连接成功率法中的阈值  $\mu$  的取值在不同网络中可能不同,需要根据实际数据获得。

## 3 实验及性能分析

P2P 节点识别方法的优劣主要可以从误报率和漏报率两方面来衡量。误报 FP(False Positive)为将非 P2P 节点识别为 P2P 节点所引起的错误判断。漏报 FN(False Negative)为将 P2P 节点误识别为非 P2P 节点,从而漏报了 P2P 节点。误报的原因主要是一些非 P2P 应用在一些特殊情况下表现出和 P2P 应用相似的行为特征。

为验证文中方法的实际效果,实验使用 wire-Shark1.1.2 抓包软件从校园网环境抓取了不同时段、不同机器上的流量数据进行分析。本实验将节点分为三组:BT 节点, PPstream 节点, HTTP 客户端节点(不运行 P2P 应用),采用上述阈值进行分析判断。由于基于应用签名的识别技术<sup>[12]</sup>(亦称为 Payload 特征法)识别准确率较高,因此文中分别从误报率 FP 和漏报

率 FN 两方面,对文中所提算法和 Payload 特征法、仅使用连接成功率的方法进行比较,实验结果如表 1、表 2 所示。

表 1 三种识别方法误报率 FP 的比较

节点类型	连接成功率法	文中方法	Payload 特征法
HTTP 节点	2.33%	0.83%	—
BT 节点	2.13%	1.21%	0
PPstream 节点	3.67%	2.56%	0

表 2 三种识别方法漏报率 FN 的比较

节点类型	连接成功率法	文中方法	Payload 特征法
HTTP 节点	1.45%	1.17%	—
BT 节点	7.12%	3.82%	1.78%
PPstream 节点	54.7%	5.32%	4.56%

从表 1 中可以看出文中方法在 PPstream 节点和 BT、HTTP 节点的识别中误报率较连接成功率法均有所减少,减少幅度都在 1% 左右。此外,文中识别法在这 3 类节点识别中与准确性得到广泛承认的 Payload 特征法相比 FP 值小于 3%,明显小于公认的 5% 的合理范围,可以认为该算法的误报率是可接受的。

漏报率可以从表 2 看出,连接成功率法在 PPstream 节点的识别中 FN 高达 50% 以上,这主要是由于数据稳定期主动连接较少所致。改进后的文中方法由于采用主被动连接数比值法作为辅助判断而大大减少了这种漏报,使漏报保持在 5% 左右,接近于 Payload 特征法。在 BT 节点和 HTTP 节点的识别中,文中方法 FN 值虽略高于 Payload 特征法但较连接成功率法仍有一定减少,减少范围分别在 1% 和 3% 左右,这主要是由于主被动连接数的辅助判断进一步提高了识别精度。

综上所述,文中提出的基于主被动连接的识别方法无论在误报率还是漏报率方面较文献[9]中仅使用连接成功率的方法都有一定提高,尤其弥补了原方法在以 PPstream 为代表的一类流媒体节点识别中漏报严重的缺陷。虽然与 Payload 特征法相比误报和漏报都略高,但文中算法计算简单、占用资源较少、识别速度快,且在节点识别的范围内该算法的准确度是可以接受的。另外,基于 Payload 特征的方法不可用于检测加密的 P2P 流量以及 Payload 特征未知的 P2P 流量,且会涉及隐私问题,因此不适用于大规模流量环境或安全网络中,而文中所提方法不受此限制,因此基于主被动连接的识别法具有更广泛的应用范围。

## 4 结束语

P2P 节点的识别作为 P2P 网络测量领域的一类重要的方法正在逐步发展。文中通过对节点连接特征的

分析和实验,针对连接成功率法在流媒体节点检测中误报较大的不足,提出主被动连接算法来识别 P2P 节点,实验表明,该算法能有效地提高 P2P 节点识别的精度。文中连接特征的计算是基于 TCP 的,而对于 UDP 的情况有待进一步研究解决。另外,由于受网络带宽、资源稀缺度等因素的影响,P2P 应用所表现出的连接特征具有一定的不稳定性,下一步的研究将结合参数方法、半参数方法和非参数估计等统计学方法解决异常点对整体判断的影响。

## 参考文献:

- [1] Azzouna N B,Guillemin F. Impact of peer-to-peer applications on wide area network traffic: an experimental approach [C]//Proc of IEEE Global Telecommunications Conference. Texas, USA: IEEE Computer Society, 2004: 1544-1548.
- [2] 黄烟波,周磊戈. 基于流特征 P2P 流量识别方法研究[J]. 计算机技术与发展, 2009, 19(9): 46-48.
- [3] 袁雪美,王 晖,张 鑫,等. P2P 流量识别技术综述[J]. 计算机应用, 2009, 29(S2): 11-15.
- [4] 温 超,郑雪峰,戚 翔,等. 基于流量分析的 P2P 协议识别方法的研究[J]. 微计算机应用, 2007, 28(7): 714-717.
- [5] Karagiannis T, Broido A, Michalis, et al. Transport layer identification of P2P traffic [C]//Proc of the 4th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2004: 121-134.
- [6] 裴 江,卢选民,周亚建. 一种基于节点状态的 P2P 流量检测模型[J]. 计算机应用, 2009, 29(3): 662-664.
- [7] 鲁文斌,杨家海,刘洪波. 基于节点连接模式的 P2P 节点识别算法[J]. 清华大学学报: 自然科学版, 2009, 49(7): 1029-1033.
- [8] 王春枝,李 涛. 基于双层特征的 P2P 流量检测[J]. 计算机技术与发展, 2009, 19(7): 238-241.
- [9] Zhou Li - Juan, Li Zhi - Tang, Hao Tu. Proposition and provement of a TCP feature of P2P traffic - an example of BitTorrent and Emule [C]//International Conference on Communications and Networking in China. Shanghai: IEEE, 2007: 61-65.
- [10] 李代玲. 基于网络测量的 PPStream 网络电视系统研究 [D]. 北京: 北京交通大学电子信息工程学院, 2008.
- [11] Bolla R, Canini M. On the Double - Faced Nature of P2P Traffic Department of Communication [C]//In Proceedings of the Sixteenth Euromicro Conference on Parallel, Distributed and Network Based Processing. Toulouse, France: IEEE Computer Society, 2008: 524-530.
- [12] Sen S, Spatscheck O, Wang Dong - mei. Accurate, scalable in network identification of P2P traffic using application signatures [C]//Proc of the 13th International Conference on WorldWideWeb. New York: ACM Press, 2004: 512-521.