对等网络流量检测技术

陆 庆，周世杰，秦志光，吴春江

(电子科技大学计算机科学与工程学院 成都 610054)

【摘要】对等网络(P2P)应用的飞速发展，丰富了互联网的内容，但其流量的爆发式增长和不加限制的带宽占用，不仅给互联网基础设施带来巨大冲击，也给Internet服务提供商(ISP)和应用服务提供商(ASP)的高级服务部署带来了很多问题。开展高效、准确的P2P流量实时识别与过滤相关技术研究，不仅有利于合理利用互联网基础设施、P2P技术和合理部署P2P应用，还有利于制止非法内容在P2P网络中的传播。该文通过对现有P2P流量识别算法进行深入研究和对比分析，为P2P流量有效管理和合理规范提供了技术参考。

关 键 词 内容监管； 对等网络； 流量过滤； 流量识别
中图分类号 TP393.08 文献标识码 A

Research on the Technology of Peer-to-Peer Network Traffic Identification

LU Qing，ZHOU Shi-jie，QIN Zhi-guang，WU Chun-jiang

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 610054)

Abstract The rapid development of Peer-to-Peer (P2P) network has enriched the performance of Internet. However, with the violent increment of data flow and bandwidth using with no restrict, the P2P applications have brought huge impact to Internet base establishment and the advance service of Internet Service Provider (ISP) and Application Service Provider (ASP). Therefore, to utilize the Internet base establishment and P2P technology correctly and to process P2P applications effectively, the technology of P2P network traffic identification and traffic filtering should be researched while restraining the illegal content being transfered in P2P network. This paper focuses on the deep research and comparison of available algorithms of P2P network traffic identification, and gives a technology reference and standard for the P2P network traffic management.

Key words content monitoring; peer-to-peer network; traffic filtering; traffic identification

近年来，对等网络(Peer-to-Peer，P2P)的用户规模、应用类型和流量均呈爆发式增长。P2P应用类型也已从文件共享扩展到语音、视频等应用领域。中国互联网实际流量模式分析报告表明，P2P流量已约占整个互联网流量的60%。

为此，国外网络设备生产商和网络服务提供商相继推出了针对P2P流量识别与监管的产品或技术。P2P流量检测设备包括网络缓存设备、应用层流量管理设备、流统计状态路由器和智能防火墙等。主要厂商及产品包括Cisco公司的NetFlow 技术[1]、Allot的故障恢复流量管理方案[2]、CacheLogic公司的CacheLogic P2P管理方案[3]、Verso Technologies 的NetSpective系列产品[4]等。

但是，国内对于P2P流量识别技术的研究工作较少，不仅缺乏高质量学术论文，也缺乏高效的P2P多媒体内容识别与过滤产品。从产品角度来看，国内部分网络设备生产商虽然推出了P2P流量监控的相关产品，如华为的SecPath 1800F防火墙[5]、Eudemon500、1000防火墙[6]和CAPTECH的网络管理软件——网络慧眼CAP[7]，但由于这些产品采用的都是深层数据包检测技术，因此在性能、开销等方面存在很多问题。

因此，开展高效、准确的P2P流量(尤其是多媒体内容)实时识别与过滤的相关技术研究，有利于合理利用互联网基础设施、P2P技术和合理部署P2P应用，有利于制止非法内容在P2P网络中的传播，也有助于维护中国互联网的健康环境和营造一个和谐的网络社会。本文针对现有的P2P流量识别算法进行对比分析，为P2P流量有效管理和合理规范提供技术参考。

收稿日期：2007 – 05 – 10
基金项目：国家自然科学基金(60473090)；国家242信息安全专题计划项目(2006B19)
作者简介：陆 庆(1964 –)，女，高级工程师，主要从事对等计算及其安全技术方面的研究.

1 对等网络简介及对等网络流量检测的困难性

对等网络是一种分布式网络，其中的参与者共享他们所拥有的一部分硬件资源(处理能力、存储能力等)，这些共享资源要由网络提供服务和内容，且能被其他节点(Peer)直接访问而无需经过中间实体。网络中的参与者既是资源提供者(即服务器)，又是资源获取者(即客户)。对等网络的代表性应用是文件共享(如Napster)。

但是，P2P不仅仅是用于文件共享，它还包括建立基于P2P形式的通信网络、P2P计算或其他资源的共享等很多方面。P2P最根本的思想，同时也是它与客户/服务器模型(C/S)最显著的区别在于网络中的节点既可以获取其他节点的资源或服务，同时又是资源或服务的提供者，即兼具Client和Server的双重身份。一般P2P网络中每一个节点所拥有的权利和义务都是对等的，包括通信、服务和资源消费。

从分类来看，可以将P2P分为纯P2P和混合(Hybrid)P2P两种模式。纯P2P网络中不存在中心实体或服务器，从网络中移去任何一个单独的、任意的终端实体，都不会对网络中的服务带来大的损失。而混合P2P网络中则需要有中心实体来提供部分必要的网络服务，如保存元信息、提供索引或路由、提供安全检验等。

对等网络的快速识别与分类，不仅为运营商系统服务质量(QoS)提供技术支持，也可以为对等网络上的内容监管(如恶意代码识别、病毒防御)提供保障。但是，由于对等网络的内在特性，其流量识别存在以下特殊性：

(1) 不确定性。由于对等网络应用的多样性(如文件共享、语言通信、视频通信等)，因此对等网络流量不仅在流量特征上，而且在行为特征上也表现出不确定性。此外，对等网络中节点的动态性，也增加了对等网络流量的不确定性。这种流量的不确定性，为实现对等网络的流量识别带来了诸多困难。

(2) 海量性。对等网络不仅应用多种多样，而且规模极大，如文件共享式P2P系统BitTortent同时在线节点一般可高达100万以上。对等网络流量的海量性，不仅给流量的实时检测带来严重的性能问题，也为流量的存在带来了挑战。

(3) 加密性。由于对等网络属于应用层，因此为了躲避内容监管，现有P2P系统均对其载荷进行了加密处理。加密特性使得常规的模式识别算法很难直接应用于对等网络中。因此，必须寻求新的流量检测技术与方法，才能解决P2P流量识别的准确性和可靠性问题。

上述的特殊性给对等网络的流量进行正确、高效和实时识别带来了很多困难。从技术层面来看，现有P2P流量检测技术大致可分为基于流量特征的识别方法(Transport Layer Identification，TLI)和基于深层数据包的识别方法(Deep Packet Inspection，DPI)。此外，网络设备提供商和安全产品提供商也开展了P2P流量识别与监管的研发工作。以下分别对两种方法进行深入分析。

2 基于流量特征的P2P流量识别技术

在P2P系统中，每个节点既是客户也是服务器，这种节点充当双重角色的特点，也使得P2P应用在传输层表现出与其他网络应用(如HTTP、FTP、DNS、EMAIL等)不同的流量特征。因此，基于流量特征的P2P流量检测方法的基本思想是，通过对传输层数据包(包括TCP和UDP数据包)的分析，并结合P2P系统所表现出来的流量特征来识别某个网络流是否属于P2P。这类方法包括：TCP/UDP端口识别技术、网络直径分析技术、节点角色分析技术、协议对分析技术和地址端口对分析技术等。

TCP/UDP端口识别技术采用固定的服务端口的特点来识别第一代P2P系统。如文献[7]首次提出了P2P流量识别问题，利用端口识别技术对Fast-Track、Gnutella和Direct-Connect三种具有代表性的P2P系统的流量特征进行了分析。现有P2P系统所采用的常用服务端口如表1所示。但由于许多P2P应用为躲避流量审计与过滤，往往采用了随机端口技术，因此TCP/UDP端口识别技术存在严重的漏报问题。

网络直径分析技术利用了P2P系统所组成的逻辑网络具有网络直径大这一特点。在P2P系统中，节点之间需要建立连接。与物理连接不同，P2P系统中的连接是逻辑连接，因此所形成的P2P网络属于逻辑网络。

文献[8]通过记录网络中每个节点与其他节点建立连接的情况而得到P2P系统的逻辑连接拓扑图，并计算其网络直径。文献[8]的研究结果表明，与其他网络应用所形成的逻辑网络相比，P2P系统所形成的逻辑网络具有更大的直径。因此，如果网络直径超过某个门限值，则该网络中的节点就是P2P节点，相应的流量也即是P2P流量。

由于网络直径的计算需要记录整个网络的连接状态，因此不仅存储和计算开销大，而且也不支持

P2P流量的实时识别与过滤。

表1　现有**P2P**系统的常用服务端口

P2P系统	常用的服务端口
Edonkey (eMule, xMule)	TCP 2323, 3306, 4242, 4500, 4501, TCP 4661-4674, 4677, 4678, 7778
FastTrack(older KaZaA)	TCP 1214, 1215, 1331,1337, 1683, 4329
BitTorrent	TCP 6881-6889
Gnutella	TCP 6346, 6347
MP2P	TCP 41170, 10240-20480, 22321
DirectConnect(DC++, BCDC++)	TCP 411, 412, 1364-1383, 4702,4703, 4662
ShareShare	TCP 6399, UDP 6388, 6733, 6777
Freenet	TCP 19114, 8081
Napster (FileNavigator,WinMX)	TCP 5555, 6666, 6677, 6688,6699-6701, 6257
SoulSeek	TCP 2234, 5534
Blubster	TCP 41170
Morpheus	6346/6347 TCP/UDP

节点角色分析技术利用了P2P系统中每个节点具有双重角色的特点。P2P系统中的每个节点，既是客户端，也是服务器。因此，如果可以通过判断某个逻辑网络中具有这种双重角色的节点数，就可以确定该网络是否为P2P网络。如文献[8]通过记录并计算网络中同时充当客户和服务器两个角色的节点数，发现如果该数超过某个门限值，则这些节点所形成的网络就是P2P网络，而该网络中的节点就是P2P节点，相应的流量也即是P2P流量。

与网络直径分析一样，节点角色分析技术也需要记录整个网络的连接状态，因此同样面临存储与计算开销大、无法供P2P流量的实时识别与过滤功能等问题。

协议对分析技术利用了P2P系统可能同时使用TCP和UDP协议的特点。实际分析结果表明，P2P系统一般采用UDP来发送命令等控制信息，而采用TCP协议来传输数据。在一般的应用中，极少出现同时使用UDP协议和TCP协议的情况。因此，可以利用P2P系统的这个特征来识别P2P流量。如文献[9]所采用的协议对分析技术中，通过判断在时间t内，某个“源-目的IP地址对”之间是否同时使用了TCP和UDP协议，如果是，则这两个节点之间的流量就有可能是P2P流量；反之，则可能不是P2P流量。由于DNS等应用也会同时使用TCP和UDP协议，与上述分析技术产生混淆，因此协议对分析技术存在严重的误报问题。

地址端口对分析技术[9]也是利用了P2P系统中节点角色多重性的特点。地址端口对分析技术的依据是，在P2P系统中，每个节点既是客户端，也是服务器。为了能够接受其他节点建立连接的请求，每个节点都需要广播自己的IP地址和提供服务的端口(记为{目的IP，目的Port}，简称目的地址端口对)。而为了与其他节点建立连接，每个节点随机选择一个源端口，使用自己的IP地址(记为{源IP，源Port}，简称源地址端口对)，并利用其他节点所广播的IP地址和端口对信息来建立连接。由于每个节点与另外一个节点建立连接时，不论是源节点还是目的节点，都使用随机源端口技术，因此对于广播了目的地址端口对的节点A来说，与自己建立了连接的源IP地址数和源端口数应大致相同。相反，其他应用(如HTTP)往往需要建立多个连接来传送数据，因此来自于同一个源IP的节点可能采用不同的源端口，与Web服务器建立多条连接，其源IP数与源端口数往往不同。为此，在单位时间t内，如果网络流的源IP数与源端口数相同，则该流量可能就是P2P流量。地址端口对分析技术具有性能高的优点，但是缺乏实时识别与过滤的能力。

除了上述有关P2P流量检测技术外，还有基于流量特征的P2P流量检测技术。文献[10]通过两种方法来识别BitTorrent流量：

(1) 许多节点向同一个节点发送了大量数据且在目的节点出现握手数据包；

(2)某个节点广播了大量UDP数据包，并随之发送了大量握手数据包。

文献[11]利用P2P系统的连接错误率等TCP流的特征来识别P2P流量。

文献[12]结合Skype具有“中继”的特性，通过考察P2P流量的如下特征参数来识别网络流是否为P2P：开始时间差、结束时间差、流的速率、两个流的时间相关系数。文献[12]通过实验证明，具有中继特性的Skype流量有如下特征：开始时间差一般小于5 s；结束时间差一般小于5 s；进入流的比特率与出来流的比特率大小基本相等；两个P2P流的时间相关系数不小于0.37。因此，可以利用这四个特征参数来识别具有中继特征的Skype及其他P2P流量。

3　基于应用层数据检测的P2P流量识别技术

该类方法是通过协议分析与还原技术，提取P2P

应用层数据(即P2P载荷Payload)，通过分析P2P载荷所包含的协议特征值(Signature)，来判断是否属于P2P应用。因此，这类方法也被叫做深层数据包检测技术(Deep Packet Inspection，DPI)。在深层数据包检测技术中，通过对具体的P2P协议及其对应的P2P系统的载荷进行特征提取，建立特征库。对于流经的实时网络流，采用模式匹配算法，判断其中是否包含特征库中的特征串。如果特征匹配成功，该网络流就是P2P数据。

文献[13-17]采用DPI技术识别P2P流量。文献[13]对Gnutella、Edonkey、DirebtConnect、BitTorrent和Kazaa的协议特征进行分析，并据此对应用层数据进行分析来识别是否是P2P流量。文献[17]利用应用层数据分析技术，对多媒体流量进行识别分析。

此外，也有少量文献探讨了结合基于流量特征方法和基于应用层数据的检测技术。文献[15]对比了三种P2P流量识别技术：端口分析技术、应用层签名、传输层分析。文献[16]引入了诱饵节点，并结合应用层签名分析技术对日本流行的P2P系统Winny的流量进行了分析。

4 两种P2P流量识别技术的优缺点及比较分析

基于流量特征的检测技术，其优点包括可扩展性好、性能高和可识别加密数据流。

可扩展性好是指该方法仅利用了P2P应用所具有普适性流量特征，不仅可以发现已有的P2P流量，也可以识别新的、符合普适性流量特征的P2P流量。

性能高是由于不需要对协议进行解析和还原，且也不需要对P2P应用载荷进行分析因此计算开销和存储开销小，识别算法性能高。

可识别加密P2P流量是由于基于流量特征的检测技术，不依赖具体的P2P应用载荷，因此，数据是否加密对检测算法没有影响。

但是，基于流量特征的P2P流量识别技术也具有很多不足，其主要缺点包括：准确性差、健壮性差、缺乏流量分类功能等。

有两个因素决定了基于流量特征的P2P流量识别技术存在准确性差的缺点。第一个因素是P2P流量特征不一定唯一：很多流量特征都不是P2P流量唯一的，其他应用也有可能表现出这种流量特征来。因此该方法存在误判问题，即将不是P2P流量的网络流，误认为是P2P流量。第二个因素是网络环境复杂：如由于不对称路由和丢包、重传现象的存在，导致无法精确确定流量特征，从而有可能对P2P 流量检测的精确度造成影响。

健壮性差是指由于不能处理数据包丢失、重组等，因此不能适应复杂的P2P应用。

缺乏流量分类功能是指由于传输层流量特征一般不能明确指示应用层协议类型，所以这种方法对P2P 应用分类的能力较弱。而对P2P应用进行细分类，对于执行P2P流量监管措施(如禁封、限速、提供服务质量QoS等)非常重要。

深层数据包检测技术易于理解、升级方便、维护简单，是目前运用最普遍的P2P流量识别方法。其主要优点包括：准确性高、健壮性好及具有分类功能等。

准确性高是由于该方法执行精确特征匹配，因此极少存在误判问题。

健壮性好是由于可以处理数据包丢失、重组等，因此能适应复杂的P2P应用。

具有分类功能是由于深层数据包检测技术可以依据不同P2P应用的载荷特征来准确分类P2P应用，因此可以为实施P2P流量监管策略提供准确的信息。

而深层数据包检测技术的缺点包括：可扩展性差、缺乏加密数据分析功能、性能低等。

可扩展性差是由于该方法对新P2P应用的流量识别具有滞后性，即在未升级特征库前无法检测新的P2P应用，必须找到新应用的载荷特征后，才能对该应用实施有效检测。

缺乏加密数据分析功能是指由于P2P载荷加密，隐藏了P2P应用的协议和数据特征，因此深层数据包检测技术对加密P2P 应用的检测能力非常有限。

性能低是指由于需要完成协议解析还原和特征匹配等操作，因此计算和存储开销大，流量检测算法性能低。载荷特征越复杂，检测代价越高，算法性能也越差。

表2是各种P2P流量识别算法的比较。

表2 现有的**P2P**流量识别方法

识别算法类型	准确性	可扩展	健壮性	性能	分类功能	加密数据分析
TLI	较差	好	差	好	无	支持
DPI	好	较差	较好	差	有	不支持

5 结 束 语

基于流量特征和基于应用数据分析技术是目前主要的P2P流量识别方法。从P2P流量识别的技术现状来看，基于应用数据分析技术的深层数据分析方

法DPI由于具有准确性高、健壮性好和分类功能，且过去的P2P大都未加密，因此是P2P流量识别的主要方法。但是，基于DPI技术也面临诸如如何提供检测算法的性能、如何支持对加密数据的分析、如何更新P2P应用特征库等问题。同样，基于流量特征的P2P流量识别方法虽然具有性能高、可扩展性好的优点，但由于准确性差，在实际应用中也面临诸多困难。此外，现有方法都以离线数据分析为主，缺乏P2P流量的实时识别能力。

从本质来看，基于流量特征的检测属于启发式方法，而深层数据分析属于精确匹配方法。如果能够结合这两种方法的优点，就有可能设计出一个准确、高效的P2P流量实时识别算法。为此，研究启发式深层数据分析实时识别算法将是进一步研究的主要内容。

参 考 文 献

[1] CISCO. NetFlow技术白皮书[EB/OL]. http:∥www.cisco.com/, 2007-08-12.

[2] ALLOT. 故障恢复流量管理方案[EB/OL]. http://www.allot. com, 2007-08-12.

[3] CACHELOGIC. CacheLogic P2P 管理方案[EB/OL]. http ://www.cachelogic.com, 2007-08-12.

[4] VERSO TECHNOLOGIES. NetSpective 系列产品白皮书[EB/OL]. http ://www.verso.com/, 2007-08-12.

[5] 华为公司. P2P 流量(BT流量)监管技术白皮书[EB/OL]. http: //www. huawei.com, 2007-08-12.

[6] 华为公司. Eudemon系列防火墙解决方案[EB/OL]. http: //www.huawei.com, 2007-08-12.

[7] SAN Subhabrata, WANG Jia, Analyzing peer-to-peer traffic across large networks[C]// IEEE/ACM Trans on Networking. [S.l.]: [s.n.], 2004: 219-232.

[8] FIVOS C, PANAYIOTIS M, Identifying known and unknown peer-to-peer traffic[C]// In Proc. of Fifth IEEE International Symposium on Network Computing and Applications. [S.l.]: [s.n.], 2006: 93-102.

[9] THOMAS K, ANDRE B, MICHALIS F. Transport layer identification of P2P traffic[C]// In: Proc of the 4th ACM SIGCOMM Conf on Internet Measurement. [S.l.]: [s.n.], 2004: 121-134.

[10] HONG Mong-fong, CHEN Chun-wei, CHUANG Chin-shun. Identification and analysis of P2P traffic-an example of bit torrent[C]//Proc of Int'l Conf on Innovative Computing, Information and Control (ICICIC2006). [S.l.]: [s.n.], 2006, 2: 266-269.

[11] ZHOU Li-juan, LI Zhi-tong, LIU Bin. P2P traffic identification by TCP flow analysis [C]//Proceedings of International Workshop on Networking, Architecture, and Storages, Shenyang. Los Alamitos: IEEE Computer Society, 2006: 47-50.

[12] SUH K, FIGUEIREDO D R, KUROSE J, et al. Characterizing and detecting skype-relayed traffic[C]// Proceedings of IEEE Conference on Computer Communications (INFOCOM 2006). Barcelona: [s.n.], 2006: 1-12.

[13] SEN S, SPATSCHECK O, WANG Dong-mei. Accurate, scalable in-network identification of P2P traffic using application signatures[C]//Proceedings of the 13th International Conference on World Wide Web. New York: ACM, 2004:512-521.

[14] WANG Rui, LIU Yang, YANG Yue-xiang, et al. Solving the app-level classification problem of P2P traffic via optimized support vector machines[C]//Proceedings of Sixth International Conference on Intelligent Systems Design and Applications (ISDA '06). Los Alamitos: IEEE Computer Society, 2006, 2: 534-539.

[15] MADHUKAR A, WILLIAMSON C. A longitudinal study of P2P traffic classification[C]//Proceedings of 14th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems. Los Alamitos: IEEE Computer Society, 2006: 179-188.

[16] OHZAHATA S, HAGIWARA Y, TERADA M, et al. A traffic identification method and evaluations for a pure P2P application[C]//Proceedings of 2005 Passive and Active Measurement (PAM'05). Boston; Berlin: Springer-Verlag, 2005: 55-68.

[17] KANG H J, KIM M S, HONG J W. Streaming media and multimedia conferencing traffic analysis using payload examination[J]. ETRI Journal, 2004, 26(3): 203-217.

[18] CAPTECH. Computer and Telecommunication Systems (MASCOTS 2006)[EB/OL]. http://www.captech.net.cn/cn/, 2007-08-12.

[19] MARCELL P, DANG T D, GEFFERTH A[J]. Identification and analysis of peer-to-peer traffic[J]. Journal of Communication, 2006, 1(7): 36-46.

编　辑　张　俊