

一种基于双重流传输特性的 P2P 流量检测方案

吴 敏,王汝传

(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要: P2P 流量逐渐占据了互联网主要流量,在对 Internet 起巨大推动作用的同时,也带来了因资源过度占用而引起的网络拥塞以及安全隐患等问题,妨碍了正常的网络业务的开展。介绍了各种 P2P 流量识别方法及特点;然后提出一种基于双重流传输特性的局域网内 P2P 流量检测方法,即基于连续动态端口变化和 P2P 应用独特的上下行流量比率特征。该方法的创新之处在于仅使用一部分包基本统计信息,无需检测数据净荷就可以识别 P2P 流量。实验结果显示该方法克服了传统的基于 payload 特征的方法不能检测加密和未知 P2P 应用的缺点,具有较高检测效率和合适的检测精度。

关键词: 对等网络;流量检测;净荷

中图分类号: TP393.08

文献标识码: A

文章编号: 1673-629X(2010)10-0066-04

A Mechanism for P2P-Traffic Detection Based on Two-Fold Traffic Transmission Features

WU Min, WANG Ru-chuan

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

Abstract: P2P traffic has taken great portions in the network traffic. While having a significant impact on the Internet, it brings serious problems such as network congestion and traffic hindrance caused by the excessive occupation in the bandwidth. Introduces methods in identifying P2P traffic and their characters, then proposes a method of P2P traffic identifying in local area network based on two-fold traffic transmission features, namely, successive port change and its unique character in the ratio of upload and download traffic volume. The novelty of the proposed method is that it only utilizes some basic statistical information of packets instead of the inspection of data payload. Experimental results show that the method has achieved some improvements in identifying payload-encrypted and unknown P2P traffic which is hard for traditional payload-method to fulfill and has low cost and proper accuracy.

Key words: P2P network; traffic identification; payload

0 引 言

随着 P2P 网络^[1]技术的兴起, P2P 流量逐渐成为互联网流量的重要组成部分。精确地识别 P2P 流量对于有效地管理网络和合理地利用网络资源都具有重要意义。

目前 P2P 流量检测技术大致有以下三类:基于端口的检测技术,深层数据包检测技术和基于流量特征

的检测技术。

但是由于越来越多的 P2P 应用采用了端口跳变、负载加密等流量隐藏技术,使得原来的基于端口和深层数据包检测技术等 P2P 流量识别方法已经逐步被淘汰^[2~4]。而且,人们需要实时地识别出 P2P 流才能够实现对流量的控制,以提高网络的性能。

基于流量特征的检测技术主要利用 P2P 应用在传输层表现出来的流量特征去识别 P2P 应用。这类方法通常是借用统计学领域和机器学习理论^[5,6]通用的一些概念去分析 P2P 应用在传输层的特征信息。该方法是几乎不需要其他额外的硬件或者软件,不需要任何关于应用层协议的信息,并能够识别加密的和未知的 P2P 流量,因而近年来国内外广泛关注这种利用流统计方式去测量 P2P 流量的方式^[7~11]。但关于流量特征进行的检测,大都属于离线分析,不能进行实时 P2P 流测量。文献[12]提出了一种基于上下行流量比率(uds)的识别办法,但该方法计算一定时间间

收稿日期:2009-10-01;修回日期:2010-01-11

基金项目: 国家自然科学基金(60973139, 60773041);江苏省自然科学基金(BK2008451);省级现代服务业发展专项资金;江苏省高校科技创新计划项目(CX09B-153Z, CX08B-086Z);南京邮电大学青蓝工程项目(NY206034, NY208011);江苏省六大高峰人才科技项目(2008118)

作者简介: 吴 敏(1976-),女,博士生,讲师,研究方向为移动代理技术、流量检测、分布式计算、计算机密码学、人工智能;王汝传,教授,博士生导师,研究方向为计算机软件理论、移动代理技术、分布式计算、计算机安全和网络计算。

007-12-18 22:34:06.533601 58.217.134.166	221.227.139.233	TCP	3387 > 28188 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.560895 58.217.134.166	70.48.222.142	TCP	3388 > 2448 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.604741 58.217.134.166	122.126.69.17	TCP	3357 > 20696 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.616084 58.217.134.166	91.65.4.248	TCP	3390 > 6346 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.668655 58.217.134.166	82.245.155.155	TCP	3391 > 6346 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.705051 58.217.134.166	59.32.142.97	TCP	3358 > 26075 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.768295 58.217.134.166	82.229.119.158	TCP	3383 > 30264 [ACK] Seq=1 Ack=1 win=65535
007-12-18 22:34:06.770909 58.217.134.166	82.229.119.158	TCP	3383 > 30264 [PSH, ACK] Seq=1 Ack=1 win=65535
007-12-18 22:34:06.815307 58.217.134.166	219.68.226.112	TCP	3392 > 17953 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.905708 58.217.134.166	220.140.193.5	TCP	3381 > 23076 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.905807 58.217.134.166	83.24.170.216	TCP	3359 > 6346 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:06.912517 58.217.134.166	58.221.6.162	TCP	3393 > 27630 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:07.046226 58.217.134.166	219.84.215.204	TCP	3394 > 16881 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:07.106358 58.217.134.166	59.126.172.74	TCP	3360 > 41711 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:07.106464 58.217.134.166	62.241.53.15	TCP	3361 > 4242 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:07.206680 58.217.134.166	62.241.53.16	TCP	3362 > 4242 [SYN] Seq=0 Len=0 MSS=1414
007-12-18 22:34:07.206786 58.217.134.166	81.56.116.29	TCP	3364 > 6346 [SYN] Seq=0 Len=0 MSS=1414

图 1 pplive 抓包截图

隔内同一 IP 地址的总体流量比率,具有一定的局限性,即只能判别出运行单一 P2P 应用的节点,而对于混合流量识别,由于其比率不再满足文中提出区分范围,因而无法有效检测出 P2P 流量。

文中在文献[12]基础上提出一种基于连续动态端口和流量特征的实时 P2P 流量检测方法。该方法的创新之处在于首先观察连接的动态端口特征进行在线实时 P2P 流量的判别,对无法识别的连接计算一定时间内的上下行流量比率,从而实现 P2P 流量的准确识别。该方法只使用报文大小和包基本信息,具有较高检测效率和合适的检测精度。

1 双重 P2P 流量特征

1.1 连续端口特征

文献[6]提出了使用 {IP, port} 识别 P2P 网络流的方法:通过检测所有源 {srcIP, srcport} 和目的 {dstIP, dstport}, 判断特定连接 IP 数与特定连接端口数是否相同,相同则认定为 P2P 流;反之,认定为非 P2P 流。为了找出 P2P 流量的端口特征,研究了几种流行的网络服务在 1 分钟内源端口的动态变化,这几种服务是 web, bitTorrent, PPlive, edonkey, Limewire, Bearshare, FTP。研究结果表明:许多 P2P 应用,包括文件共享和流媒体应用,其源端口呈现出连续动态的变化特征,图 1 是利用 Analyzer^[8] 的报文捕获截图,运行软件为 PPlive,而非 P2P 的应用如 FTP,则表现出很稳定的特征。当然,web 流量也有类似的特征,但可以通过下文描述的上下行比率法进行区分。其特征描述满足下列条件:

$$\begin{cases} \text{Diff}(\text{srcport}) = D \text{ iff}(\text{dstIP}) \\ \text{Dist}(\text{srcport}) \leq \text{threshold1} \end{cases}$$

Threshold1, 根据文献[6],一般设为 10。

1.2 一定时间内 uds 特征

对于 P2P 节点,它既是客户机,又是服务器。据此,文献[7]提出通过统计一定时间内(1 分钟)五种不同 P2P 应用(Maze, PPlive, BitTorrent, eDonkey 及 thunder)的上下行比率,分析其比率的分布范围,从而实现未知流量的判别。但该方法存在两个缺陷:其一,没有分析非 P2P 应用如 web、FTP 流量的上下行比率特点,从而不能很好地判别非 P2P 流量与 P2P 流量;其二,该方法对混合流量检测失败,主要原因在于在混合流量中,其上行流量和混合下行流量的比率已经不再满足该文提出特征。文中在文献[6]基础上,通过实际网络抓包实验,研究分析了几种 P2P 应用和 web 应用、FTP 服务在稳定数据传输过程中的流量特点。这几种 P2P 应用分别为:比特彗星、PPlive、Maze 和迅雷。试验方法是分别在不同主机上一次运行一种软件,抓取一定时间内(根据文献[7],时间间隔取 1 分钟)每条连接流量。对每组应用共获得 140 组的上下行流量,计算并分析其下行与上行比率,如图 2 所示。观察发现一般网络服务如 FTP 等流量的 uds 大多超过 5,而 P2P 应用则在 [0~2] 内。

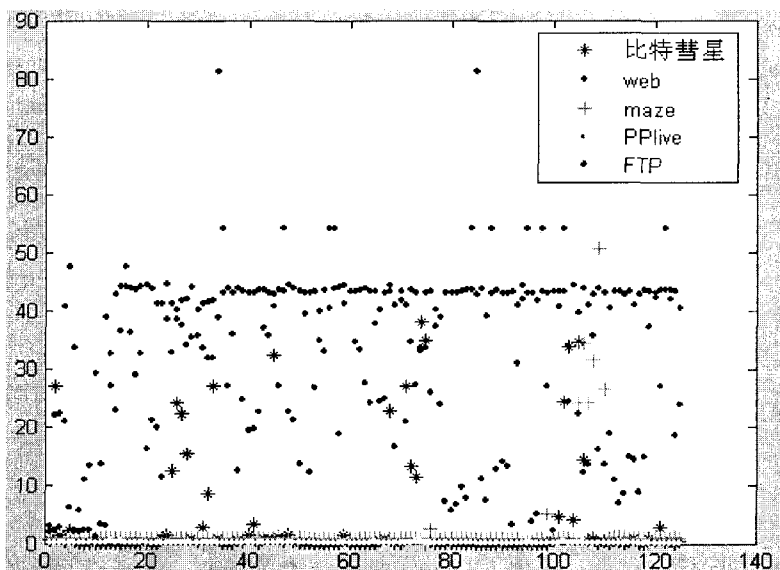


图 2 各网络服务 1 分钟内下行上行比率分布

2 基于流传输特性的 P2P 流量检测方法

2.1 检测流程

根据 P2P 流的这两种特征,提出一种新型的 P2P 流量识别方法,如图 3 所示。

(1)捕获报文,取出该报文的(SrcIP、SrcPort、DestIP、DestPort、Size),观察其端口是否与 P2P 节点集 P2PTable 中同一 IP 记录对应的端口满足连续动态变化的特征,若满足说明该包为 P2P 数据包,直接用该端口更新 P2PTable 中相应记录对应的端口。如果不满足端口匹配特征,则判断(SrcIP、SrcPort)或(DestIP、DestPort)中哪个属于本地局域网节点,对于前者记 status 为 0,后者记为 1,同时将其 IP、Port、status 及 size 提供给 FlowAnalyzer 模块。

(2)一定时间内(根据文献[6],确定最优间隔为 6 秒钟)将未通过 P2PTable 表匹配的网络数据包更新记入 FlowTable。到达相应时间间隔后,首先计算局域网各主机各端口流量总和,如果达到某最低门限(该值通过观察数据采集阶段各 P2P 应用的总体流量,设定其最小值确定,如 300b/s),则计算该连接的上下行比率 Rate,并判别是否是 P2P 流量。将成功识别为 P2P 应用的连接,添加或更新至 P2PTable 后清空 FlowTable 中记录。

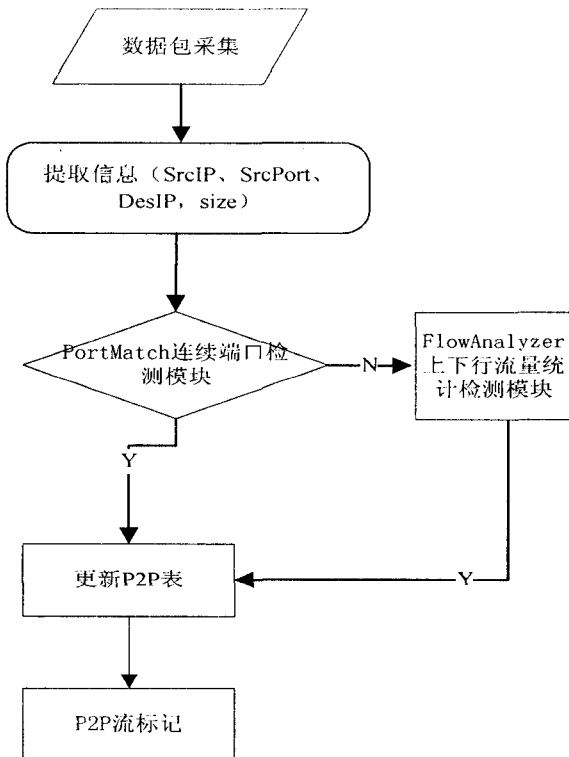


图 3 基于流传输特性的 P2P 流量检测方法

2.2 连续端口检测模块 PortMatch

建立以节点 IP 地址为索引关键字的 hash 表,称为 P2PTable。对于只运行单个 P2P 应用程序的局域

网节点,表中只有唯一一条记录,对于同时运行多个 P2P 应用程序的局域网节点,同一个 IP 可能会对于不同范围的端口。算法如图 4 所示。

```

if SrcIp in local net range
{
    Status=0
    IP=SrcIp
    Port=SrcPort
}
else
{
    Status=1
    IP=DestIp
    Port=DestPort
}
for each record in P2PTable
if find(IP)
{
    if(dis(Port) <= threshold)
        Discard it
    else
        Goto FlowAnalyzer Module with( IP, Port, status, size)
}
  
```

图 4 连续端口检测模块 PortMatch 算法

2.3 上下行流量检测模块 FlowAnalyzer

建立以节点 IP 地址为索引关键字的 hash 表,称为 FlowTable。其中每个 IP 在表中至多有一行记录。表定义结构为:IP(整型)、Port(整型)、IncomingTraffic(整型)、OutgoingTraffic(整型)、Rate(整型),算法如图 5 所示。

3 实验结果分析

3.1 效率验证

为了验证文中提出的双重特征在 P2P 流量识别的有效性和准确性,设计了一个原型系统,并将之部署在校园网的一台入侵检测的服务器上,这样校园网的所有对外流量都通过镜像流经该服务器。在实验中应用的 P2P 协议包括: eMule、BitTorrent、迅雷、PPlive、QQlive、Poco、FTP 服务以及 web 应用,为了进行检测效果的对比,在该服务器上同时部署了 Analyzer 软件^[13],后者是完全通过特征关键字达到 P2P 流识别目的的。

检测结果如下: PPlive(FN: 2.12, FP: 0)、Maze(FN: 100, FP: 0)、Emule(FN: 2.68, FP: 0)、Kazaa(FN: 5.45, FP: 0)、BitTorrent(FN: 25, FP: 0)

在数据采集和分析阶段,研究发现确实存在部分

web 连接,其下行与上行流量比率也符合 P2P 特征,但这些连接的上下行总流量远小于 P2P 应用中各连接的流量,因此在上下行流量检测模块 FlowAnalyze 中通过设置总体流量的门限值,忽略了这部分连接的相应计算,因而在一定程度上降低了误检率,提高了检测精度。

```

While(! interval)
{
    if ! find(IP,Port)
    {
        Add(IP,Port)in FlowTable
        if(status)
            IncomingTraffic = size
        else
            OutgoingTraffic = size
    }
    else
        add IncomingTraffic or OutgoingTraffic according to the
status
    While (FlowTable! = NULL)
    {
        for each record in FlowTable
        if (IncomingTraffic + OutgoingTraffic) > threshold2
        compute Rate with IncomingTraffic/OutgoingTraffic
        if Rate in[0,1]
            UpateP2PTable with (IP,Port)
            Clear all record in FlowTable
    }
}

```

图 5 上下行流量检测模块 FlowAnalyze 算法

3.2 效率验证

服务器配置如下:2.8G CPU、512M 内存, NIC: 100Mbps。其 CPU 占有率为 20%, 使用总体内存 180M。

4 结束语

在分析了当前 P2P 流识别方案的实现原理和特点的基础上,提出了一种基于动态端口和流量特征的实时局域网 P2P 流量检测方法。该方案既可以部署在局域网内能实时获取网络报文头的网络管理主机上,也可部署在路由器的网络处理器上,同样适用于使用任意端口的 P2P 应用和加密的 P2P 协议,有效地提高了 P2P 流的识别率和识别的速度。但如同大多数基于流识别的方案一样,该方案并不能精确识别具体的 P2P 应用协议,因此一个可选的方案如下:捕获具体的报文内容然后根据 P2PTable 离线进行特征字匹配,可以方便今后进行基于服务类型的 P2P 流量管

理。

下一步的研究目标是通过仿真实验和原型系统运行,确定流统计的精确时间间隔等,从而进一步完善方案。然后实现该方案和特征字匹配的融合,从而更准确判别 P2P 流量并进行应用级分类。

参考文献:

- [1] 吴国庆. 对等网络技术研究[J]. 计算机技术与发展, 2008, 18(7): 100-104.
- [2] 蒋海明, 张剑英, 王青青, 等. P2P 流量检测与分析[J]. 计算机技术与发展, 2008, 18(7): 74-76.
- [3] 吴 敏, 王汝传. 基于主机的 P2P 流量检测与控制方案[J]. 计算机技术与发展, 2009, 19(10): 26-30.
- [4] 王 锐. P2P 流量检测技术研究[D]. 长沙: 国防科学技术大学, 2006.
- [5] McGraw-Hill. 机器学习[M]. 曾华军, 张银奎, 等译. 北京: 机械工业出版社, 2003.
- [6] 威 滕. 数据挖掘: 实用机器学习技术[J]. 北京: 机械工业出版社, 2006.
- [7] Karagiannis T, Broido A, Faloutsos M, et al. Transport Layer Identification of P2P Traffic[C]//In: Proc. of ACM SIGCOMM IMC. Taormina, Sicily, Italy: [s. n.], 2004: 121-134.
- [8] Moore A, Zuev D. Internet traffic classification using bayesi analysis[C]//Proceedings of International Conference on Measurement and Modeling of Computer Systems. [s. l.]: [s. n.], 2005: 50-60.
- [9] Zuev D, Moore A. Traffic classification using a statistical approach[J]. Lecture Notes in Computer Science, 2005, 3431: 321-324.
- [10] Constantinou F, Mavrommatis P. Identifying Known and Unknown Peer-to-Peer Traffic[C]//Proceedings of Fifth IEEE International Symposium on Network Computing and Applications. [s. l.]: [s. n.], 2006: 93-102.
- [11] Li Wei, Canini M, Moore A W, et al. Efficient application identification and the temporal and spatial stability of classification schema[J]. Computer Networks, 2009, 53: 790-809.
- [12] Liu Hui, Feng Wenfeng, Huang Yongfeng, et al. A Peer-to-Peer Traffic Identification Method Using Machine Learning[C]//International Conference on Networking, Architecture, and Storage(NAS2007). Guilin, China: IEEE Computer Society, 2007: 155-160.
- [13] Risso F, Baldini A, Bonomi F. Extending the NetPDL Language to Support Traffic Classification[C]//Global Telecommunications Conference, 2007. GLOBECOM apos; 07. [s. l.]: [s. n.], 2007: 22-27.