第41卷 第6A期 2014年6月
计 算 机 科 学 Computer Science
Vol. 41 No. 6A June 2014

一种基于IP地址随机测度的P2P主机识别算法

柳 斌 周丽娟
(华中科技大学网络与计算中心 武汉 430074)

摘 要 P2P流量的迅猛增长加剧了网络拥塞状况,P2P流量识别为网络管理提供了基本的技术支持。首先分析了P2P节点远端地址分布特性,在此基础上,提出了IP地址随机测度的特征用于衡量远端IP地址的分散性,并给出了一种基于IP地址随机测度的P2P主机识别算法。实验表明该算法能有效识别P2P主机,误报率低。

关键词 P2P,IP地址随机测度,启发式算法
中图法分类号 TP393 **文献标识码** A

Heuristic Method for Identifying P2P Application Based on IP Address Entropy

LIU Bin ZHOU Li-juan
(Huazhong University of Science & Technology, Wuhan 430074, China)

Abstract The rapid increase of P2P traffic worsens the congestion of network while P2P traffic identification becomes the basic technical support for network management. The heuristic method for identifying P2P application was studied. Firstly, the behaviors that are inherent to P2P host were explored. The behaviors were translated to metrics: remote hosts' IP address entropy. An algorithm based on remote hosts' IP address entropy characteristics was proposed. The algorithm showed low false positive in experiment.

Keywords P2P, IP address entropy, Heuristic method

1 引言

随着因特网的迅速发展,对等网络(P2P)已成为Internet上重要的应用之一。在P2P方式下,每个对等实体(peer)既是服务的提供者,又是服务的享用者。通过将服务器的负载分散到众多peer中,有效地减轻了服务器的负载。P2P应用在不断发展的同时,也给网络管理带来了许多新的问题,如P2P应用消耗了大量带宽,版权纠纷,内容监控困难等。如何准确识别网络中P2P主机是P2P管理中重要的问题。

作为一种充分利用客户端资源的新型应用,P2P应用在网络层和传输层表现出许多与传统应用不同的特征,可以通过检测这些特征间接地发现P2P应用。这类启发式算法的优点在于:①具有发现新的P2P应用和加密P2P应用的能力;②只分析数据包头部,不需要检测数据包的载荷部分。目前的启发式算法也存在一些局限性,表现为:①不适合实时分析。很多启发式算法使用的特征复杂,如连接成功率,主机之间交换的数据量,连接的方向,存在双向连接,高端口以及TCP的传输标志位等,这些特征计算复杂,存储开销大,当数据量大时,无法进行实时分析。②误报率高。很多启发式算法提取的特征并不是P2P内在的特征,而仅关注于某些特定P2P协议中的细微特征,或者偶然性特征,一旦P2P协议这些细微特征有了改变,算法将产生很高的误报率,如同时使用UDP和TCP协议特征,实际上DNS流、游戏流也同时使用UDP和TCP协议。

本文首先分析了P2P主机的远端地址的行为特点,提取了IP地址随机测度的特征。在此基础上,提出了一种基于IP地址随机测度P2P主机识别算法。实验表明,该算法能有效识别出P2P主机,且误报率较低。

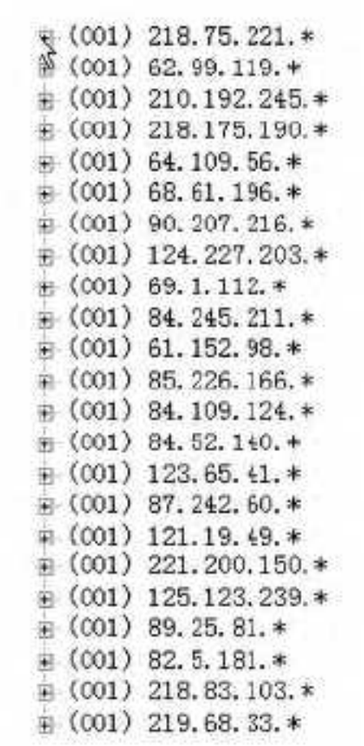
2 IP地址随机测度

2.1 IP地址随机测度定义

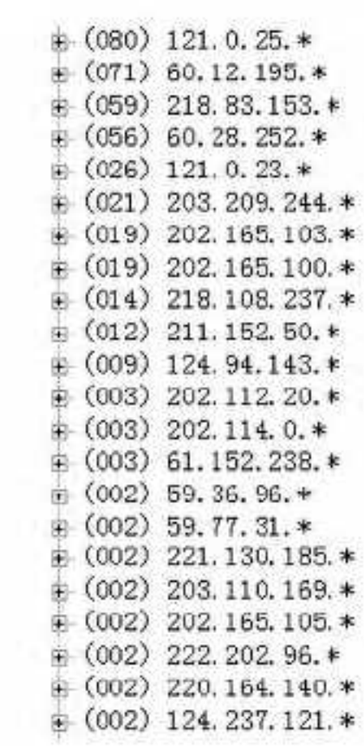
IP地址随机测度定义观察P2P连接中对端地址的分布特点。图1(a)给出某主机进行BT应用时的连接情况。其中,右边表示为与主机互联的对端IP地址的子网号,左边括号内为该子网中的连接数。图1(b)给出该主机进行WWW访问时的连接情况。在图1(a)的例子中,主机进行BT下载时共产生了23条连接,这23个连接的对端IP地址分布在23个不同的C类子网中,地址分布比较均匀,平均每个C类子网内一条连接。WWW应用一共产生了410条连接,这410条连接的对端IP地址范围集中在22个子网中,平均每个C类子网内有18.6条连接。其中,在121.0.25.0这个C类子网中有80条连接,在60.12.195.*这个子网中有71条连接。

观察多种P2P系统的远端IP地址分布,发现也有类似的特点。如何来描述这种地址分布的特征呢?信息论中用熵作为不确定性的度量,可以借助信息论中熵的思想作为地址分布集中程度的度量。

柳 斌(1971—),男,博士,副教授,主要研究方向为网络管理,E-mail:bliu@hust.edu.cn;周丽娟(1975—),女,博士,讲师,主要研究方向为网络安全。



(a)BT 应用时的连接情况



(b)WWW 访问时的连接情况

图 1 某主机的连接情况

定义 1(IP 地址熵) 假设 T 为 n 个对端 IP 地址集合，这些 IP 地址分别属于 k 个子网。子网定义为 m 位前缀相同的 IP 地址的集合，$m=16\sim24$。其中第 i 个子网在 T 中出现的概率为 $P_k, k=1,2,3,\cdots,K, P_k=\dfrac{\text{子网}_k\text{中的 IP 地址数}}{n}$，$P_k>0, \sum_{k=1}^{K}P_k=1$，$n$ 为集合 T 中总的 IP 地址数。第 k 个子网的信息量为 $\log\dfrac{1}{P_k}$。

K 个子网信息量的数学期望定义为 IP 地址熵，如下式所示：

$$H(p_1,p_2,\cdots,p_k)=\sum_{k=1}^{K}P_k\log\frac{1}{P_k} \tag{1}$$

如果 $k=1$，也就是 T 中的所有样本在一个子网里，那么 $H(p_1,p_2,\cdots,p_k)=0$，达到最小值。

定理 1(最大熵定理) T 中子网以等概率出现，即 $p_1=p_2\cdots=p_k$，也就是每个 IP 地址就是一个子网，那么 $H_{\max}(p_1,p_2,\cdots,p_k)=\log n$，达到最大值。

定义 2 IP 地址随机测度(E)是 IP 地址熵和最大 IP 地址熵的比值，表示 IP 地址的随机程度。

$$E=\frac{H(p_1,p_2,\cdots,p_k)}{H_{\max}(p_1,p_2,\cdots,p_k)}=\frac{H(p_1,p_2,\cdots p_k)}{\log n} \tag{2}$$

由定义可知，$0\leqslant E\leqslant1$，E 表示随机程度，E 越接近 1，表示信息随机程度越大；E 越接近 0，表示确定性信息越大。

图 2 给出了 BT，Emule 每隔 10 秒计算的 IP 地址随机测度。图 3—图 5 分别给出了 PPLive，PPstream，Dns，Email 和 Web 每隔 10 秒计算的 IP 地址随机测度。表 1 给出了 IP 地址随机测度的统计结果。

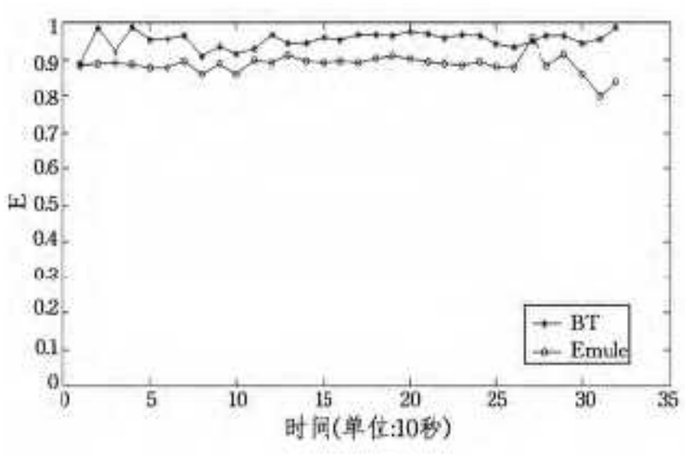


图 2 BT，Emule 的 IP 地址随机测度

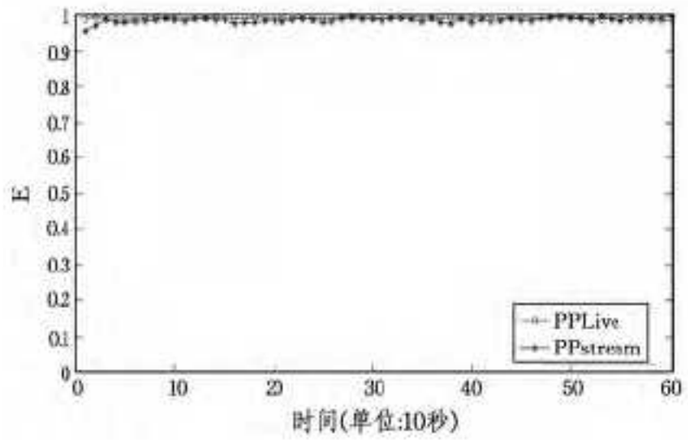


图 3 PPLive，PPstream 的 IP 地址随机测度

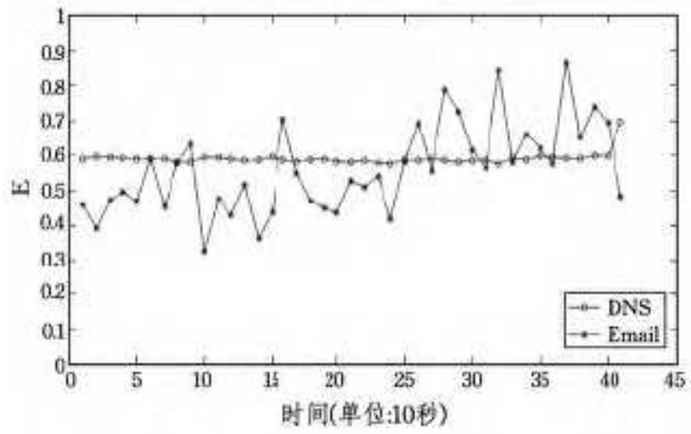


图 4 Dns，Email 的 IP 地址随机测度

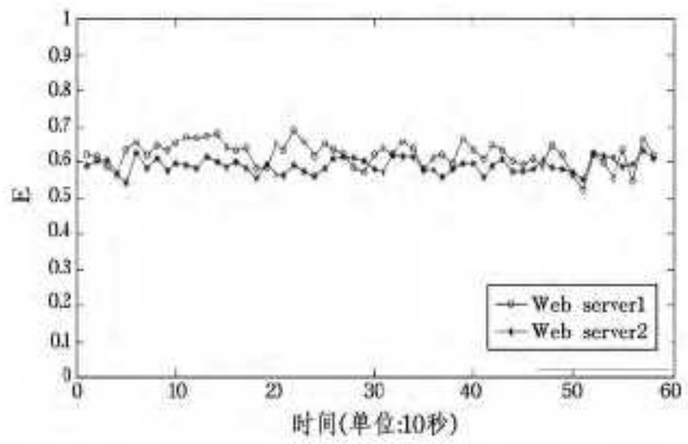


图 5 Webserver 的 IP 地址随机测度

表 1 各种应用的 IP 地址随机测度统计值

数据集	最大值	最小值	平均值	中位数	标准差
BT	0.98	0.88	0.95	0.95	0.02
Emule	0.96	0.8	0.89	0.89	0.03
PPLive	0.99	0.98	0.99	0.99	0.002
PPstream	0.997	0.96	0.99	0.985	0.007
Dns	0.7	0.58	0.59	0.59	0.01
Email	0.87	0.33	0.56	0.55	0.02
Web server1	0.69	0.52	0.62	0.62	0.03
Web server2	0.63	0.54	0.59	0.59	0.02

从上述的实验结果可以看出：

(1)P2P 应用与非 P2P 应用的 IP 地址随机测度有很明显的区别。所有的 P2P 应用中 IP 地址随机测度最小值为 0.8，而绝大多数的非 P2P 应用 IP 地址随机测度不大于 0.7。容易混淆的 Dns 的 IP 地址随机测度的平均值为 0.59。只有 Email 的 IP 地址随机测度值出现了一次 0.87，但是其平均值只有 0.56。

(2)IP 地址随机测度值变化比较稳定(最大值和最小值的差异较小)，特别是 PPLive 和 PPstream 应用。IP 地址随机测度比较稳定有利于设置阈值。

2.2 P2P 节点远端地址分布特性分析

IP 地址随机测度越接近 1，表示信息随机程度越大，IP 地址随机测度接近 0，表示确定性信息越大。实验表明 P2P 连接中地址分布比较均匀，分析原因：

(1)资源的分散性

P2P 优秀下载性能的一个重要因素是采用了分片机制。如 BT 中将原始文件分成若干个 Piece(其他的 P2P 系统也有

类似的机制)。分片机制使得资源很快从一台机器分散到了多台机器,串行下载变成了并行的下载。从连接的角度看,从少量连接变成了多条连接。在混合式的 P2P 系统中,Tracker 收集所有参与节点的状态,并给新加入节点一个随机的节点列表。随机选择节点时只会考虑到该节点是否在同一个群(下载相同的文件)中,不会考虑到这些 Peer 是否相邻、距离以及带宽等因素。因此,正是分片机制导致集中的资源迅速地分散到了各个节点,下载节点将从多个节点获得资源,而不会集中在某点。这种资源分散性在连接上表现为 IP 地址分布比较分散。在 C/S 模式中,某些应用也会产生多条连接,如 WWW 服务下载页面时也会产生多条连接,但是由于资源比较集中(集中在 1,2 台主机上),导致目的 IP 地址比较集中。因此可以说,IP 地址随机测度实际上反映了 P2P 系统与 C/S 系统在资源分布上的区别,即反映的是 P2P 网络的本质特征。

(2)动态性

P2P 网络是一个由很多临时节点组成的网络,这些主机可能会有不同的处理能力或连接速度,特别是随机离开这个网络的行为是由用户自身决定的。这使得 P2P 的网络拓扑很不稳定。P2P 节点为了保证它的下载速度不会因为部分用户的随机离开而下降,会不断地向其它 P2P 节点发起新的连接,同时会继续不断尝试连接离开 P2P 网络的节点,这也导致连接不断增多,IP 地址随机测度增加。

3 基于 IP 地址随机测度的 P2P 主机识别算法

在 IP 地址随机测度特征的基础上,给出一种基于 IP 地址随机测度的 P2P 主机识别算法,算法流程如下:

步骤 1 流量镜像

通过分光器或者交换机的镜像功能将网络流量镜像到系统上。

步骤 2 基于哈希算法的流记录表构建

流定义为在某段时间内具有相同源 IP 地址、目的 IP 地址、源端口、目的端口及协议的一系列数据包的集合。采用哈希的方法将数据包转换成流,记录每条流的源 IP 地址、目的 IP 地址、源端口、目的端口、协议、数据包个数以及字节数信息。流记录表的构建过程是:首先建立一个空的流记录表,记录表的大小为 65536,再通过以下步骤的循环实现所述流记录表的更新,构建出所述的流记录表。

1)从网卡中读取一个数据包,提取数据包的源 IP 地址、目的 IP 地址、源端口,目的端口及协议进行哈希运算;

2)根据计算的哈希值在流记录表中查找此数据包对应的流记录是否已经存在,如果存在转到 3),否则转到 4);

3) 更新此条流记录信息,包括数据包个数和字节数,转 1);

4) 新建一条流记录并插入到流记录表中,转 1)。

步骤 3 计算节点的 IP 地址随机测度

对于每个节点计算其 IP 地址随机测度。计算 IP 地址随机测度需要查找该节点所对应的所有目的 IP 地址所在的目的 IP 地址子网,统计每个目的 IP 地址子网中的流数。由于查找目的子网操作非常频繁,某些 P2P 节点 5 分钟的目的子网个数甚至上万,这样对于某个节点每条流都需要进行多次匹配才能找到目标子网。采用了一种树状结构存储目的 IP 地址的子网号,每一个节点都有一棵目的 IP 地址子网树,如图 6 所示。

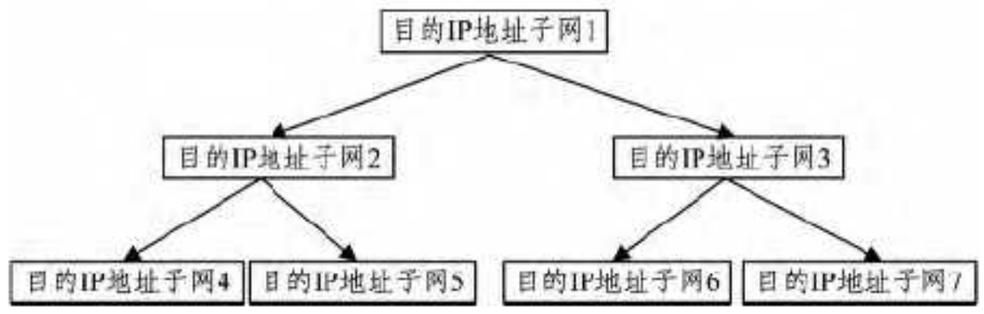


图 6 目的 IP 地址的子网树结构

IP 地址随机测度计算过程是首先遍历流记录表,建立各节点的目的 IP 地址子网树,然后一次性处理节点数据计算 IP 随机测度。

具体步骤如下:

1)从节点记录表中取一节点,计算其目的 IP 地址的子网号,方法是将目的 IP 地址与掩码 255.255.255.0 相与。

2)将目的 IP 地址子网号在该节点的目的 IP 地址子网树上进行搜索。若该目的 IP 地址子网号与该节点的目的 IP 地址子网树的根节点子网号相等,转 3),否则,转 4)。

3)更新该节点的目的 IP 地址子网树中根节点结构中流记录数字段,将流记录数加 1。转 5)。

4)新建一个目的 IP 地址子网节点结构,插入到目的 IP 地址子网树中,该结构中包括子网号和流记录数字段。插入目的 IP 地址子网树的方法是:将目的 IP 地址子网号与节点子网号进行比较,若该子网号等于节点子网号,则更新该节点结构中的流记录数字段,将流记录数加 1。转 5)。若该子网号大于节点子网号,则插入到该节点的左子树中,若该子网号小于节点子网号,则插入到该节点的右子树中,逐级递归。

5)流记录表是否遍历完,若完成,转 6);否则转 1)。

6)统计每个节点的目的 IP 地址子网树中节点的流记录数,按照 IP 地址随机测度公式,计算每节点的 IP 地址随机测度。

步骤 4 P2P 节点识别

IP 地址随机测度取值区间设为[0.9,1]。若节点的 IP 地址随机测度在此取值区间内,即判定为 P2P 节点,否则为非 P2P 节点。

4 实验与结果分析

在一个局域网环境中进行了测试。对交换机的出口进行端口镜像,测量点在镜像端口上可以捕获到局域网出口的所有流量。局域网内有 50 台机器,出口带宽 100Mbps。

采用检测率和误报率评价算法性能。检测率定义为正确划分为 P2P 的主机数除以 P2P 主机的总数。误报率定义为误分为 P2P 的主机数除以非 P2P 主机的总数。

首先进行有效性测试,验证算法能否实时识别出运行 P2P 的主机。在局域网内不同的主机上安装了不同 P2P 应用的客户端,包括 Bitcomet,Btspirit,PPLive,QQLive,Maze,Emule,PPstream,Sopcast,Uusee 和 Thunder 一共 10 种流行的 P2P 应用。所有客户端均采用默认配置。考虑 P2P 系统会随着时间的变化处于不同状态,如 BT 节点从下载状态到做种状态。随机抽取了 10 个时间点进行识别测试。启动测试系统运行 2 分钟后记录下识别结果。采用双向连接法、高端口法两种算法与本算法进行对比测试,其中,M_E,M_B,M_H

(下转第 338 页)

cation, 17th Int. Conf., CAV 2005. volume 3576 of Lecture Notes in Computer Science. Springer, 2005: 291-295

[28] Zaks A, Pnueli A. Covac: Compiler validation by program analysis of the cross-product[C] // FM 2008: Formal Methods, 15th International Symposium on Formal Methods. volume 5014 of Lecture Notes in Computer Science. Springer, 2008: 35-51

[29] Leroy X. Formal verification of a realistic compiler[M] // Communications of the ACM, 2009

[30] Leroy X. The Compcert verified compiler, software and commented proof[OL]. http://compcert.inria.fr/, Jan. 2010

[31] Fang Yi. Translation Validation of Optimizing Compilers[D]. New York University, 2005

[32] Goldberg B, Zuck L, Barret C. Into the loops: Practical issues in translation validation for optimizing compilers[C] // Proc. Workshop Compiler Optimization Meets Compiler Verification (COCV 2004). volume 132 of Electronic Notes in Theoretical Computer Science. Elsevier, 2005: 53-71

[33] Pnueli A, Zaks A. Translation validation of interprocedural optimizations[C] // Proceedings of the 4th International Workshop on Software Verification and Validation(SVV). 2006

[34] Stepp M, Tate R, Lerner S. Equality-Based translation validator for LLVM[C] // CAV'11: Proceedings of the 23rd International Conference on Computer Aided Verification. 2011

[35] Barthe G, Demange D, Pichardie D. A formally verified SSA-based middle-end-Static Single Assignment meets CompCert[C] // ESOP'12. 2012

[36] 何炎祥, 刘陶, 吴伟. 可信编译器关键技术研究[J]. 计算机工程与科学, 2010, 32(8): 1-6

(上接第 302 页)

表示使用 IP 地址随机测度、双向连接率、高端口法 3 个特征的方法。结果如表 2、表 3 所列。

表 2 3 种算法 10 次测试的检测率

次数	M_B	M_E	M_H
1	0.9	0.9	0.9
2	0.8	0.8	0.9
3	0.7	0.8	0.9
4	0.8	0.8	0.9
5	0.6	0.8	0.9
6	0.9	0.7	0.9
7	0.7	0.7	0.9
8	0.7	0.8	0.9
9	0.8	0.8	0.9
10	0.7	0.8	0.9

表 3 3 种算法 10 次测试的误报率

次数	M_B	M_E	M_H
1	0.15	0.025	0
2	0.075	0	0.025
3	0.075	0	0.025
4	0.15	0.025	0.025
5	0.1	0	0.025
6	0.075	0.075	0
7	0.15	0	0.025
8	0.15	0	0.025
9	0.1	0	0.025
10	0.075	0.025	0

从实验结果看出, IP 地址随机测度是比较理想的特征, 具有很高的检测率和较低的误报率。高端口特征与具体的实验环境和人为因素有关。实验中所有 P2P 客户端采用的是缺省配置, 监听端口均为高端口, 局域网中其他的应用也比较单纯, 主要运行一些常规的应用, 如浏览网页, 收发邮件等。而在实际更复杂的网络环境中, 可能存在 P2P 用户指定低端口为监听端口, 一些特殊的非 P2P 应用如 SQL 等也采用高端口_高端口等情况。因此, 高端口_高端口特征在这样一个局域网环境内有效并不能表明其在大规模的网络环境中就一定有效。

对算法误报率做进一步的测试, 容易造成误判的应用主要是一些服务器如 Dns 服务器、Mail 服务、视频服务器、Web 服务器等。对校园网的服务器网段进行了镜像, 服务器网段包括校园网 Dns 服务器、Email 服务器、Web 服务器和若干托管的服务器, 一共 15 台服务器。网段里没有使用 P2P 的主机, 构成了一个没有 P2P 的环境。同样抽样测试 10 次, 每次系统运行 2 分钟。误报率如表 4 所列。

表 4 10 次测试的误报率

次数	M_B	M_E	M_H
1	0.13	0	0.06
2	0.33	0.06	0.06
3	0.13	0	0.06
4	0.13	0	0.06
5	0.26	0.06	0.06
6	0.13	0	0.06
7	0.26	0.06	0.06
8	0.13	0	0.06
9	0.26	0.06	0.06
10	0.13	0	0.06

M_H 方法每次都有误报, 因为网段中某台服务器的监听端口为 8000, 所以 M_H 每次测试都产生了误报, 这也说明实际环境中使用 M_H 是很不准确的。M_B 方法 10 次检测结果不稳定, 波动比较大。进一步分析 M_B 方法每次都误报 Dns 服务器、Email 服务器, 这与单独实验是一致的, 对 Web 服务也产生了几次误报。Dns, Email 应用都表现出既有输入连接, 又有输出连接, 且连接数多, 与 P2P 的连接特征十分类似, 造成 M_B 方法误报率很高。M_E 方法较好, 误报率低, 且每次误报都是 Email 服务器, 可进一步调整门限。

结束语 P2P 网络中, 每个节点同时承担服务器和客户机的双重功能, 这是与 C/S 结构的根本区别, 对 P2P 主机远端地址分布进行了研究。由于资源随机分布在不同的 P2P 节点上, 导致了 P2P 主机连接的远端地址的均匀分布。据此, 提取了 P2P 主机的 IP 地址随机测度特征。实验表明该特征能有效地区分 P2P 主机和非 P2P 主机, 该特征计算简单, 适合实时处理。

参 考 文 献

[1] 鲁刚, 张宏莉, 叶麟. P2P 流量识别[J]. 软件学报, 2011, 12(6): 1281-1285

[2] 张琛, 王亮, 熊文柱. P2P 僵尸网络的检测技术[J]. 计算机应用, 2010, 6(2): 117-119

[3] 李鑫, 刘东林. 基于统计特征的 P2P 流量检测方法[J]. 计算机工程, 2010, 36(5): 114-116