基于分组抽样的 P2P 超级节点推测

韦安明[①②] 王洪波[①] 程时端[①]
[①](北京邮电大学网络与交换技术国家重点实验室 北京 100876)
[②](广播电视规划院 北京 100866)

摘 要：基于小世界规律和幂率特征，该文提出了 P2P 超级节点检测方法 IPS。IPS 通过分组抽样检出 P2P 种子节点，然后根据 P2P 节点间的流量关联特征检出其余的 P2P 节点，并将连接度大的 P2P 节点作为超级节点输出。基于互联网数据的实验验证了该法的正确性。

关键词：P2P 网络；网络流量；网络测量；流量控制
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2009)06-1513-04

Inferring the P2P's Super Peers with Packet Sampling

Wei An-ming[①②] Wang Hong-bo[①] Cheng Shi-duan[①]
[①](*State Key Laboratory of Networking and Switching Technology*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)
[②](*Academy of Broadcasting Planning*, *Beijing* 100866, *China*)

Abstract: Based on the small-world and power-law characters of P2P, the IPS is proposed to identify the P2P super nodes. IPS adopts packets sampling to identify seed nodes. Then according to the traffic relationship between the P2P nodes, IPS identifies all the other P2P nodes originating from the seed nodes. Finally, IPS regards the nodes with large connection degree as super nodes and outputs them. The validity of proposed method is verified by the experiments using Internet traces.

Key words: Peer-to-Peer network; Network traffic; Network measurement; Traffic control

1 前言

文献[1–3]的研究表明 P2P 网络具有小世界规律和幂率特征，即小部分节点具有高连接度，它们对 P2P 网络的贡献要大于那些连接度小的节点。文献[4]发现 P2P 节点间的流量分布服从重尾分布，20.7%的节点传输了 90%的流量，2.3%的节点传输了 50%的流量。文献[1,3]的研究表明删除高连接度的节点会破坏 P2P 网络结构，降低节点间的连通性，达到减少 P2P 流量的目的。

但是，P2P 流量控制的前提是 P2P 节点发现。目前，基于固定端口、特征码匹配以及连接特征推测是最常用的 3 种 P2P 节点检测方法，其中连接特征推测法无需分析 IP 分组净荷，适用于检测各种 P2P 网络，但该法需多次遍历流量数据，需保存海量连接状态，难以实现[5]。本文提出的基于分组抽样的 P2P 超级节点推测方法(Inferring P2P's super nodes with packet Sampling，IPS)解决了这个难题。

2008-05-14 收到，2009-01-12 改回
国家自然科学基金(90604019，60502037)，高等学校博士学科点专项科研基金(200800131019)，国家 863 计划项目(2006AA01Z235，2007AA01Z206)和新世纪优秀人才支持计划(NECT-07-0109)资助课题

2 节点发现及超级节点推测

超级节点(super nodes)定义：用 DEG 表示节点连接度，$\mathrm{DEG}_{\mathrm{th}}$ 表示超级节点判定阈值，则满足 $\mathrm{DEG} > \mathrm{DEG}_{\mathrm{th}}$ 的 P2P 节点称为超级节点。

种子节点(seed node)定义：从分组样本中检出的 P2P 节点称为种子节点。

缩略语定义：SIP－源 IP 地址；SP－源端口；DIP－目的 IP 地址；DP－目的端口。

种子节点检出方法：在抽样周期 T 中到达的所有 5 元组流(指 SIP，DIP，SP，DP，协议类型 5 个要素定义的端到端的分组序列)，凡是满足以下 3 个规则之一的都被看成是 P2P 流，对应 SIP，DIP 的主机即称为 P2P 种子节点：(1)如果 {SIP,DIP} 地址对之间有 TCP 和 UDP 流量并发，则对应 SIP、DIP 的主机称为 P2P 节点；(2)如果将 {DIP,DP} 对定义为某个流(flow)的一个网络插口(socket)，则当连接到这个插口的的所有 $\{\mathrm{SIP}_i,\mathrm{SP}_i\}, i=0,\cdots,n$ 对中，当互斥的 SIP 数目与互斥的 SP 个数相等时，对应 $\mathrm{SIP}_i, i=0,\cdots,n$,DIP 的主机称为 P2P 节点；(3)源或目的端口为熟知 P2P 端口，对应 SIP，DIP 的主机称为 P2P 节点。

除了上文提到的 3 个规则外，还引入了第 4 个规则，称

为“节点关联”，定义如下：如果节点 A 已经被标明为 P2P 节点，那么所有和 A 有流量关系的节点都被当做候选的 P2P 节点。实际上，任何一种 P2P 网络的所有节点之间，通过多跳关联以后都可以构成一条路径。引入节点关联，解释了为什么通过种子节点可发现该 P2P 网络的其余节点。

IPS 结构如图 1 所示，由随机抽样，P2P 种子节点检测，节点关联 3 大模块构成，其基本思想：(1)对链路全流量分组用文献[6]中的方法进行随机抽样，然后用上述规则分析样本，得到 P2P 种子节点并插入节点关联列表中；(2)未被抽样的分组则与列表中的已经存在的 P2P 节点进行关联以检测新的节点；(3)对列表中的候选 P2P 节点进行替换以限制检测系统的资源消耗，同时输出超级节点。

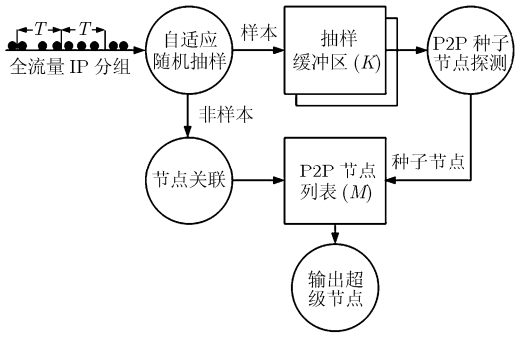

图 1 检测系统结构图

列表管理算法：新节点随机替换列表中寿命最长的非超级节点。以寿命因子 TTL(Time To Life)表示列表中节点的寿命，并设定最高寿命 $\mathrm{TTL}_{\max}$。在每个抽样周期 T 结束时，新节点 TTL 赋值 1，旧节点 TTL 加 1，扫描列表中 TTL $=\mathrm{TTL}_{\max}$ 的节点，将超级节点输出并清除非超级节点。

图 2，图 3 分别描述了种子节点插入表和节点关联操作过程，图 4 则图示了列表的内存布局，用双存储模式提高节点关联速度，CAM 器件用于内容匹配，SRAM 用于变量存储和管理。为提高速度，设计时使 SRAM 和 CAM 的表项地址相关联，即 CAM 中第 i 项的地址用作该项在 SRAM 中的地址偏移量，当第 1 个时钟周期过后如果 CAM 中的某项被命中，返回的地址在第 2 个时钟周期即作为 SRAM 对应项的访问地址。每个分组最长处理时间为 6 次 CAM 查找，4 次线性表删插及 1 次 SRAM 操作。按每次 CAM 操作耗时 1 时钟周期，最长线性表操作为从双向链表中删除节点并插入到头节点、耗时 7 访问周期计算，总时间最多消耗 35 个 SRAM 访问周期。按当前技术，CAM 操作耗时不足 4ns，SRAM 可达 2 至 5ns。若按每次操作耗时 4ns 计算，则 IPS 处理一个分组的总时间为 140ns。按照互联网分组平均长度 400 字节计算，10G 链路中平均分组到达间隔为 320ns，远大于 140ns，可以实现在线节点关联。

图 2–图 4 中，缩略语义解释列于表 1。

```
1  InsertSeedNode (SIP, SP, DIP, DP) {
2    If (sip, sp) not in SNL then {
3      SNL.ins (sip, sp);
4      HPL.ins (sip, sp);
5      TTL. (sip, sp) = 1;
6      DEG. (sip, sp) = 1;
7    }
8  }
```

图 2 种子节点插入列表的过程

```
1  SearchFunc (sip, sp, dip, dp) {
2    If sip in SNL then {
3      If sp in SNL then
4        If (dip, dp) not in HPL then {
5          HFL.ins (dip, dp);
6          SNL(sip, sp).DEG ++;
7        }
8      Else {
9        SNL.ins (sip, sp);
10       HFL.ins (dip, dp);
11       TTL.(sip, sp) = 1;
12       DEG.(sip, sp) = 1;
13     }
14   }
15   If dip in SNL then {
16     If dp in SNL then
17       If (sip, sp) not in HPL then {
18         HFL.ins (sip, sp);
19         SNL(dip, dp).DEG ++;
20       }
21     Else {
22       SNL.ins (dip, dp);
23       HFL.ins (sip, sp);
24       TTL.(dip, dp) = 1;
25       DEG.(dip, dp) = 1;
26     }
27   }
28 }
```

图 3 节点关联过程

3 参数选取与复杂度分析

抽样周期 T，抽样缓冲区大小 K 的选取影响抽样率，同时影响计算能力、存储的消耗。P2P 超级节点的形成是一个相对缓慢的过程，实验中取 T 为 1200s。抽样过程必须在 SRAM 中进行，因此 K 的值会影响系统设计复杂度和成本。

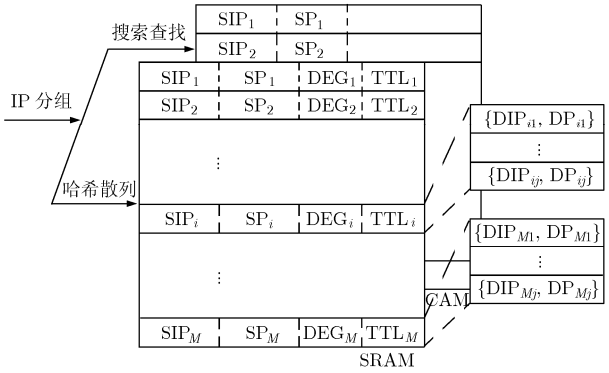

图 4 节点列表管理方法

表 1 缩略语义解释

名称	含义
SNL	超级节点关联列表，表中元素为{IP, Port}对
HFL	P2P 节点另一半列表，对应 SNL 中的 {IP, Port}对
TTL.(IP, Port)	设置 SNL 中(IP, Port)节点的寿命值
DEG.(IP, Port)	设置 SNL 中(xIP, xP)节点的连接度
SNL.ins(IP, Port)	将(IP, Port)插入 SNL 表中
HFL.ins(IP, Port)	将(IP, Port)插入 HFL 表中

节点关联列表大小 M 决定 SRAM 空间消耗，网络中 P2P 节点数影响 M 的选取，但从实验结果来看，尚未发现超级节点数量超过 1000 的情况，因此，这里取值 10000。

文献[7]研究发现，Gnutella（KaZaa）的超级节点倾向于连接 30~45（60~150）个叶节点，同时连接 30（40~60）个超级邻居节点。据此，设定 $\mathrm{DEG}_{\mathrm{th}}$ 为 60。

考虑到超级节点的形成相对于抽样周期是一个更加缓慢的过程，为避免由于抽样周期太短而漏检某些超级节点，设置了寿命因子 TTL。设置 TTL 可以使 P2P 节点在成为超级节点之前在列表中最多能够存活的时长为 $\mathrm{TTL}_{\max} \cdot T$。从实验结果来看，$\mathrm{TTL}_{\max}$ 取 1 和 3 的结果差别不是很明显，故无需严格界定该值，但推荐取值[1, 3]。

抽样模块中，假设 N 为测量周期 T 内到达的分组数，时间复杂度为 $O(N)$。种子节点算法复杂度为 $O(K^2)$。节点关联模块，时间复杂度为 $O(N)$。抽样模块无辅助存储空间需求，空间复杂度为 $O(1)$。种子节点检测模块消耗的内存较多，最大空间消耗为 $(S_{\mathrm{FT}} + S_{\mathrm{DDP}} + S_{\mathrm{SIP}} + S_{\mathrm{SP}})K$ 字节，但不考虑链表指针的空间消耗，复杂度为 $O(1)$，其中 $S_{\mathrm{FT}} = 13$ 为流表表项长度，$S_{\mathrm{DDP}} = 6$ 为{DIP, DP}对表项长度，$S_{\mathrm{SIP}} = 4$，$S_{\mathrm{SP}} = 2$ 分别为 SIP 和 SP 的表项长度。节点关联模块的存储空间由 CAM 和 SRAM 构成，为 $2M\,(S_{\mathrm{SIP}} + S_{\mathrm{SP}} + S_{\mathrm{DEG}} + S_{\mathrm{TTL}} + 6\,S_{\mathrm{Avg.HF}})$字节，空间复杂度为 $O(1)$，其中 $S_{\mathrm{DEG}} = 2$ 用于存储节点连接度，$S_{\mathrm{TTL}} = 1$ 用于存储寿命因子，M 为关联列表最多可保存的节点数，而 $S_{\mathrm{Avg.HF}}$ 为每个节点平均连接数。式中乘 2 是因为同时采用了 CAM 和 SRAM 用于保存节点信息。

4 实验

表 2 中描述的实验数据来自美国应用网络研究国家实验室。实验参数：(1) K 分别取 20000，200000，400000；(2) 考虑到 P2P 节点随意加入和退出网络这一行为，以及一旦节点连接成功，其下载过程相对稳定的特点，抽样周期 T 取定值 1200s，种子节点检测及超级节点输出均按该周期进行；(3) $\mathrm{DEG}_{\mathrm{th}} = 60$；(4)节点关联列表大小取 10000。

4.1 种子节点检测

结果如图 5 所示，每行 3 个子图分别对应每组数据的检测结果，每列 3 个子图分别对应不同抽样缓冲区的实验结果。图中标号含义如下：UDP/TCP 表示以 UDP/TCP 对规则检测出的种子节点数目；Port Based 表示以已知端口规则检测出的种子节点数目；Con. Mode 表示以{IP, Port}对连接模式规则检测出的种子节点数目；Sum 则表示去除前述 3 种方法所检测出的重复节点后得到互不重复的总的种子节点数目。

数据 1 中，随着抽样缓冲区变大，种子节点数目呈增长趋势，但趋势变缓，表明抽样缓冲区对种子节点检测影响不大，另外，端口法检出的种子节点数目虽远大连接模式法，但从端口法的结果非常接近总结果这个现象表明连接模式法检出的节点大部分已经由端口法检出，这印证了连接模式法可以用来检测 P2P 节点；数据 2 中，端口法只检出了小部分种子节点，而连接模式法检出了大部分，这可能是当地网管使用了基于端口的 P2P 流量控制设备，但基于非熟知 P2P

表 2 实验数据描述

编号	名称	大小	采集时间	链路	时长	分组数	分组速率
1	Leipzig-I	4.3GB	Nov.23.2002	OC3	6h	$72M$	$3.5K$
2	Auckland-VIII	2.3GB	Dec.2.2003	Ethernet	6h	$34M$	$1.7K$
3	San Diego-I	18.8GB	Jan.30.2004	Giga-Ethernet	6h	$265M$	$13.0K$

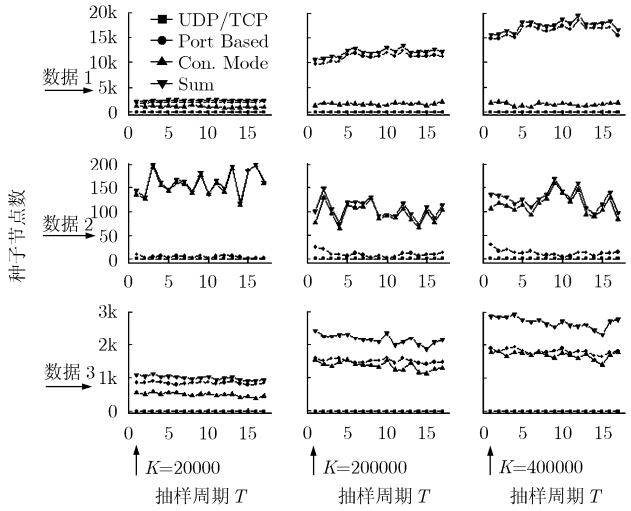

图 5 种子节点检测结果

端口的节点可以继续存活；数据 3 中，端口法检出的种子节点数目与连接模式法的非常接近，它们的加和接近总数。

各结果中，以规则 1 检出的种子节点数非常小。检出率过低可能是欠抽样造成的，以 50%的抽样率按规则 2 分析数据 1，得到了几十个种子节点。

4.2 超级节点检测

结果如图 6 所示，每行对应一组数据，每列则对应一个抽样缓冲区的值，图中实心方块表示检出超级节点数目(Inferred Super Nodes, ISN)，实心圆表示在检出的超级节点中采用已知 P2P 端口进行通信的节点数目(Known Port Super Nodes, KPSN)。

首先，随着抽样缓冲区的改变，种子节点数随着改变，但是从 3 组数据 9 组结果来看，超级节点数目无明显差别，同组数据在不同抽样缓冲区条件下的超级节点输出曲线几乎吻合，这表明抽样缓冲区(或抽样率)不直接决定超级节点检测结果，这支持了我们提出的用抽样的方法减少数据处理负担的观点。

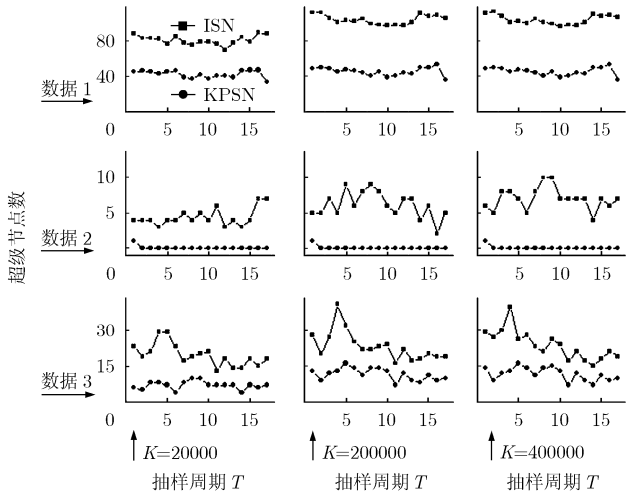

图 6 超级节点检测结果

其次，数据 1 中 KPSN 数目将近 ISN 数目的一半，这说明了两个问题：(1)2002 年，采用固定端口的 P2P 网络占优，这从输出的超级节点中端口号为 4662(4662 为 Edonkey 采用的端口)占很大部分可以得到证明；(2)IPS 检出的结果确实是 P2P 节点。数据 2 中情况相当不同，KPSN 几乎为 0。这可能是网管采用了 P2P 端口封锁技术的结果，也可能是网络中 P2P 流量本身就不大。数据 3 和数据 1 相类似，但分析超级节点端口号构成情况后发现，KPSN 中占主导的不再是 Edonkey，而是端口号为 6346 的 Gnutella。

第三，对每个检出的超级节点以熟知的 P2P 端口进行了验证后发现，除数据 2 外，检出的 50%以上的超级节点使用 P2P 熟知端口通信，因此，有理由相信剩下的那部分超级节点使用非熟知端口通信。

5 结束语

IPS 首先从分组样本获得种子节点，通过流量关联检出未知的 P2P 节点，并对节点列表进行管理，解决了连接特征推测法难以实现的问题，其有效性得到了实验印证。

参 考 文 献

[1] Saroiu S, Gummadi P K, and Gribble S D. A measurement study of peer-to-peer file sharing systems [C]. Proc. of the Multimedia Computing and Networking 2002, San Jose: SPIE, 2002: 156–170.

[2] Ripeanu M, Foster I, and Iamnitchi A. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design [J]. *IEEE Internet Computing Journal*, 2002, 6(1): 50–57.

[3] Stutzbach D, Rejaie R, and Sen S. Characterizing unstructured overlay topologies in modern P2P file-sharing systems [J]. *IEEE/ACM Trans. on Networking*, 2008, 16(2): 267–280.

[4] Zhang Y F, Lei L H, and Chen C J. Characterizing peer-to-peer traffic across Internet [C]. Proc. of the GCC 2003, Heidelberg: Springer-Verlag, 2004, LNCS 3032: 388–395.

[5] Karagiannis T, Broido A, Faloutsos M, and Claffy K C. Transport layer identification of P2P traffic [C]. ACM SIGCOMM/USENIX Internet Measurement Conference, Italy, 2004: 121–134.

[6] Wang Hong bo, Wei An ming, Lin Yu, and Cheng Shi duan. Time stratified packet sampling based on measurement buffer for flow measurement [J]. *Journal of Software*, 2006, 17(8): 1775–1784.

[7] Liang J, Kumar R, and Ross K W. The KaZaA overlay: A measurement study. http://cis.poly.edu/~ross/papers/Kazaa Overlay. pdf, 2004.

韦安明： 男，1975 年生，博士，研究方向为互联网、电台电视台制播网测量、安全等.

王洪波： 男，1975 年生，博士，副教授，研究方向为互联网测量、管理与安全、P2P 计算等.

程时端： 女，1940 年生，教授，博士生导师，研究方向为宽带网交换与路由、IP 网 QoS 控制、管理、测量等.