对等网络应用中的网络统计特征分析*

李一鹏[1)†] 任 勇[1)] 袁 坚[1)] 王 钺[1)] 黄小红[2)] 山秀明[1)]

1)(清华大学电子工程系,北京 100084)

2)(北京邮电大学网络技术研究院,北京 100876)

(2010 年 5 月 7 日收到;2010 年 8 月 19 日收到修改稿)

本文基于实测数据抽象出用户网络与资源网络,探讨了对等网络中用户、资源及其内部的相互作用关系,发掘并分析了其内在的网络统计特征. 分析结果表明,用户节点度值及权值呈分段分布,体现了其各异的活跃性;网络资源的流行度差异明显,度值和权值近似呈幂律分布. 用户网络与资源网络存在分簇结构,少数簇中含大量节点,多数簇所含节点数量较少. 用户网络中,同簇内的用户有着相似的兴趣趋向,不同簇用户间兴趣趋向存在着差异,资源网络各簇中不同类别的资源间呈现出明显的关联性.

关键词: 对等网络, 簇结构, 网络统计特征

PACS: 89. 20. Hh

1. 引 言

对等网络(peer-to-peer network, P2P 网络)已经成为当今互联网络中最重要的应用之一,其占用的网络带宽目前已超过 50%. 近些年来,在涉及复杂网络及人类动力学特征的研究方面,出现许多颇具启发性的研究成果[1—6]. 结合复杂网络的研究进展,P2P 网络丰富的应用特征也引起了广泛的关注.

人们直观上理解的对等网络,是用户间形成的能力均等的通信网络. 对这样一种通信网络的特性分析逐渐成为一个研究的热点. 首先,针对 P2P 网络这种通信网络的结构特征,文献[7]探讨了 Gnutella 网络的用户度值、网络半径和度相关性,给出了直观的认识. 其次,从 P2P 网络建模角度,研究人员引入各种连边机理[8,9]和偏向性随机游走策略[10],探索了网络拓扑的构建方法. 更进一步,人们开始分别从用户和资源两个角度,对 BearShare 和 LimeWire 两种 P2P 网络进行统计分析[11]. 遗憾的是,这些研究忽略了对等网络中用户与资源间的联系. 另外,Iamnitchi 在研究中提及数据共享图(data-sharing graph)的概念[12],用户根据对相同资源的兴趣形成连边,并形成具有小世界特性的网络拓扑,这为进一步的研究提供了启示.

用户和资源是对等网络中的两个核心元素,而网络中丰富的统计特征,正是源于二者之间复杂的相互作用关系. 一方面,具有自然人属性的用户对资源有着各异的兴趣,对相同资源感兴趣的用户间频繁通信,兴趣差异较大的用户间连接稀疏. 这种差异使用户间形成的网络可能呈现出人类社交网络的特征:如拓扑的层次化结构、局部及全局特性等. 另一方面,由于同一用户会涉及多种感兴趣的资源,这就在各类资源之间构成一种关联. 因此,从用户与资源之间相互作用关系的角度入手,就可不再局限于单纯的用户行为特征或资源流行度分析,从而引发对网络应用丰富内涵的深入思考,并对相关应用的理论建模及系统设计起到指导作用.

本文基于对等网络实测数据,结合复杂网络与人类行为动力学的研究思路,从一个新的视角探索对等网络中用户、资源及其内部的作用关系,发掘其中的网络统计特征并加以分析. 在网络建模方面,我们将用户对资源下载关系的基本网络剥离,构建了两种网络形式:用户网络与资源网络. 在网络统计特征方面,我们发现用户节点度值分布的分段特性,体现了其各异的活跃性,资源节点度值分布近似符合幂律特性,反映了资源流行度的差异. 进一步分析发现,两种网络呈现明显分簇结构:少数簇中含大量节点,多数簇所含节点数量较少. 用

* 国家自然科学基金(批准号: 60932005), 国家重点基础研究发展计划(批准号:2007CB307100, 2007CB307105)资助的课题.

† E-mail:yp-li05@mails.tsinghua.edu.cn

户网络中,同簇内的用户有着相似的兴趣趋向,不同簇用户间兴趣趋向存在着差异,资源网络各簇中不同类别的资源间呈现出明显的关联性.

2. 对等网络应用数据与建模

我们分析的数据来源于一个对等网络的服务提供者,数据收集的时间为2009年10月20日至31日. 数据中包含81531条用户对各种资源下载信息的详细记录,标明了用户编号、资源编号及资源内容信息等,共涉及10368个独立用户和7376个不同资源.

本文采用二部图表达用户与资源的关系,进而剥离这种基本关系形成用户网络与资源网络. 如图1(a)的二部图所示,连接用户A与资源a的边表示A下载或上传a,或者说用户A与其他用户共享资源a. 为了凸显用户与资源间的相互作用关系,我们分别从用户与资源的角度来重新审视图1(a)中的基本下载关系.

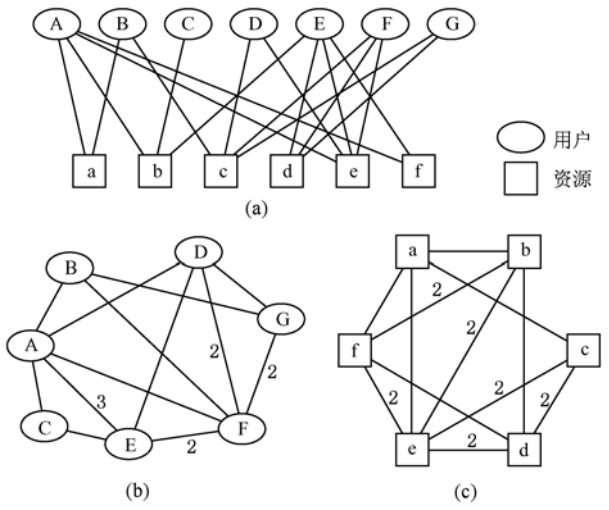

图1 (a)用户对资源的下载关系;(b)用户网络;(c)资源网络

首先,我们构造出一种反映用户关系的用户网络. 以同一资源为端点并连接多个用户的边,表示这些用户间存在以共享该资源为目的的通信连接. 如图1(b)的用户网络所示,A与C因共享资源b而连接,A与E因共享资源b,e和f形成权值为3的边. 这种以用户兴趣为依据建立的边不同于单纯的数据传输连接,从统计意义上反映了兴趣驱动下用户的行为特征及变化.

接着,我们构造出一种反映资源关系的资源网络. 以同一用户为端点并连接多个资源的边,不仅表明用户多样的下载兴趣,更重要的是使看似独立的各类资源间产生关联. 图1(c)的资源网络中,a与b因A的下载而连接,b与f因A与E的共享而建立权值为2的边.

本文基于上述用户网络与资源网络,首先统计节点度值、权值及聚类系数等基本信息,接着分析两种网络拓扑的簇结构特征,并在此基础上深入的探讨多用户间和各类资源间存在的差异及联系.

3. 网络基本统计信息

节点的度值和权值是单个节点的属性,直观上看,节点度值与权值越大就意味着该节点在网络中越重要;进一步,聚类系数刻画了节点之间连接的紧密程度. 这几个参数,反映了节点本身及节点间的基本统计信息.

3.1. 度值及权值

统计节点度值及权值时,用邻接矩阵 $\boldsymbol{A}=\{a_{ij}\}$ 和加权邻接矩阵 $\boldsymbol{A}_w=\{w_{ij}\}$ 表示节点间的连接关系,其中

$$a_{ij}=\begin{cases}1, & \text{节点 } i \text{ 与 } j \text{ 有边连接},\\ 0, & \text{节点 } i \text{ 与 } j \text{ 无边连接},\end{cases} \tag{1}$$

$$w_{ij}=(\text{节点 } i \text{ 与 } j \text{ 之间的边权值}), \tag{2}$$

则节点 i 度值为 $d_i=\sum_j a_{ij}$, 节点 i 权值为 $w_i=\sum_j a_{ij}w_{ij}$.

从图2中可以看到,用户节点的度值及权值呈现分段特性,而资源节点的度值和权值近似幂律分布. 图2(a)中,$d>10^3$ 的用户度值波动平缓,说明这些用户共享活动频繁,较为活跃;$d<10^3$ 的用户间差别显著,其活跃性波动较大. 图2(b)中节点度值和权值的近似幂律分布,体现出各资源间较大的流行度差异,存在着被多数用户下载的热门资源和少数用户感兴趣的冷门资源. 另外,两种网络中节点度值和权值在双对数坐标下满足一定关系,用户网络中 $w\sim d^{1.0404}$,资源网络中 $w\sim d^{1.0757}$.

3.2. 聚类系数

宏观层次上,网络的聚类系数(clustering coefficient)C 反映了节点间连接的紧密程度[13],其中 $C\in[0,1]$,且越大表明连接越紧密. 本文统计的用户网络中 $C=0.5275$,资源网络中 $C=0.5866$.

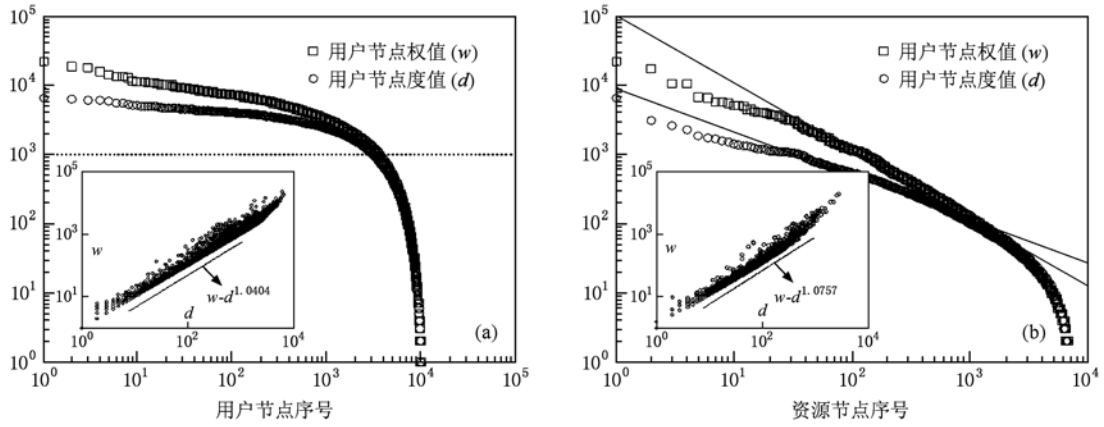

图2 (a)用户网络度值与权值分布;(b)资源网络度值与权值分布

在微观层次上,各节点及其邻居间的连接关系存在着差异,考察不同度值的节点其聚类系数的变化情况,有助于发掘网络内在的拓扑结构特征.

节点 i 的聚类系数 c_i 是指以 i 为顶点的三角形数量与和它相连的三元组数量的比值,表示为

$$c_i = \frac{2}{d_i(d_i-1)}\sum_{j,k} a_{ij}a_{ik}a_{jk}. \qquad (3)$$

类似的,节点 i 的加权聚类系数 $c_i^{(w)}$[14] 为

$$c_i^{(w)} = \frac{2}{w_i(d_i-1)}\sum_{j,k}\frac{w_{ij}+w_{ik}}{2}a_{ij}a_{ik}a_{jk}. \qquad (4)$$

图3 显示了随着节点度值或权值的增长,节点聚类系数的变化情况. 当节点的度值逐渐增大时,部分节点的聚类系数较高,其邻居节点间连接紧密;部分节点的聚类系数小,说明其邻居节点间连接稀疏. 如图3(a)中用户网络节点聚类系数随度值的增长大幅波动,而资源网络中节点的聚类系数迅速降低.

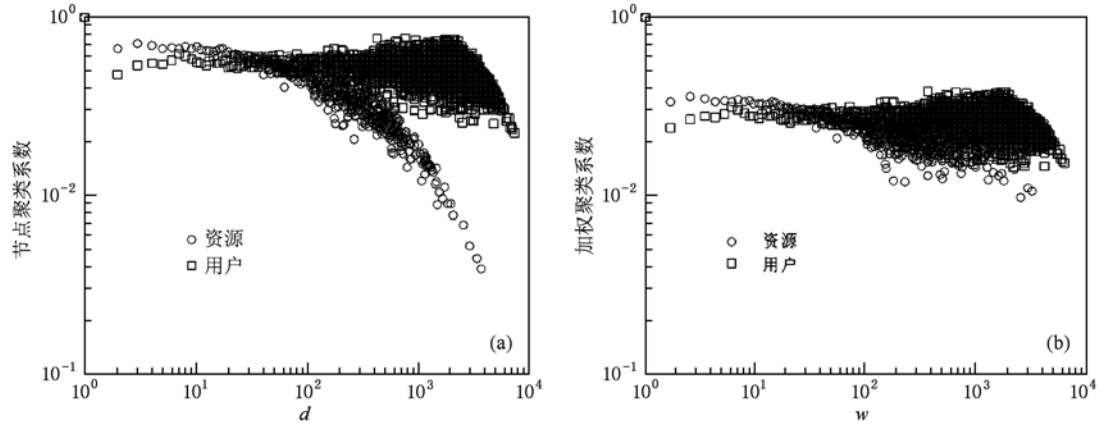

图3 用户网络与资源网络节点聚类系数(a)与加权聚类系数(b)

4. 对等网络的簇结构特征

用户网络与资源网络节点聚类系数随度值变化而出现的波动,反映了网络内在的结构特点,对这种内在拓扑结构特征的发掘是进一步分析的基础.

Newman 和 Girvan[15] 提出网络层次化分簇算法,并不断改进[16]. 该算法定义了衡量网络社团化程度的参数

$$Q = \sum_i (e_{ii} - a_i^2), \qquad (5)$$

其中 e_{ii} 表示两个端点都属于簇 i 的边占网络所有边的比例,a_i 表示一个端点或两端点均属于簇 i 的边占所有边的比例. 计算过程中 Q 值达到最大时,分簇结束. 其中 $Q_{\max}\in(0,1)$,$Q_{\max}>0.3$ 时,网络呈现明显的簇结构,$Q_{\max}>0.5$ 时,分簇结果较好[16].

分析结果显示,用户网络 $Q_{max}=0.75$,资源网络 $Q_{max}=0.61$. 如图4(a)所示,横坐标表示独立节点被合并成簇的个数,纵坐标表示合并过程中 Q 值的增长. 计算时先假定各节点独立成簇,遍历节点对并读取邻接矩阵信息,若所取两点合并成一簇使 Q 值增长最大,则合并. 如此迭代,直到 Q 值不再增长.

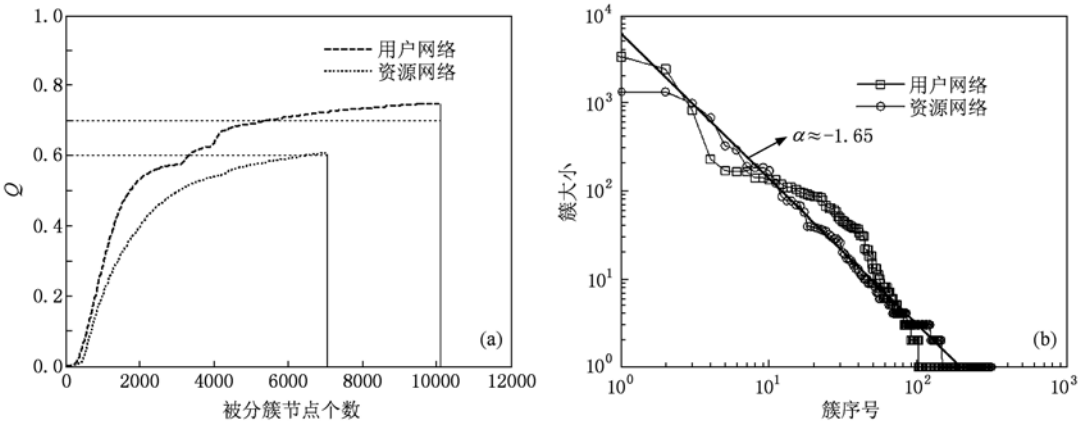

图4 节点分簇过程中的Q值变化(a)与分簇结果(b)

进一步分析发现,用户网络形成287个簇,孤立节点占总数的2.77%;资源网络形成308个簇,孤立节点占总数的4.39%. 图4(b)显示了两种网络各簇所包含的节点数量,均为少数簇包含大量节点,多数簇含节点较少,且资源网络簇大小符合 $\alpha\approx-1.65$ 的幂律分布.

5. 分簇特性的分析结果

在分析了网络的拓扑结构后,让我们重新回到对等网络应用的背景下,探寻此拓扑结构特征所反映出的多用户与各资源间的差异和联系,加深对对等网络应用丰富内涵的理解.

为方便分析,首先对涉及到的资源进行分类和预处理. 本文分析的数据包含7376个不同资源,分为电影、电视剧等9大类,每一大类包含若干子类,共355子类,如表1所示. 由于存在命名不规范等现象,某些资源无法被正确分类,可被正确识别并分类的为7085个,占96.05%.

表1 资源的分类

资源类别	电影	电视剧	音乐	游戏	动漫	综艺	软件	资料	体育
子类数量	114	17	29	12	24	16	11	113	19

5.1. 用户网络中的共性与差异

用户网络是用户间共享资源而形成的网络,资源是用户间建立连接的媒介,用户间的连接关系反映了网络的拓扑结构特征. 这提示我们从用户下载的资源角度入手,揭示存在于全网连接的用户之间的共性,和分属不同簇的用户间的差异.

首先,根据下载关系形成的用户网络是一个连通的网络,不存在孤岛现象,这表明所有用户在共享资源时体现出了一定的共性. 如图5中对簇2,7,18和30簇的分析所示,各簇用户下载的资源涵盖了所有种类,且都是电影占据主要部分. 所有用户兴趣趋向的共性,使其建立了复杂的连接关系,保证了整个网络的连通性. 这种共性也存在于人类交谈网络中,如参与不同讨论小组的人都对某个共同话题有着浓厚的兴趣[17].

其次,在去掉了某些用户之间的连边后,我们发现部分用户间的连接依然紧密并且形成簇,而簇与簇之间的连接较为稀疏,这反映出同簇内用户共享兴趣相似,而各簇间用户兴趣趋向存在差异. 图5中,第2簇用户下载较多的是电影(a)和体育(b),而簇3中用户感兴趣的主要是电影(c)、游戏(d)和动漫(e),簇17和30中依旧是电影占主要部分,并且各种类资源所占比例不同. 兴趣趋向的差异,使用户间的连接或紧密或稀疏,形成了层次化的网络结构.

多个用户兴趣趋向的共性和差异,使用户网络拓扑存在着一定的稳定结构. 突出对等网络用户作

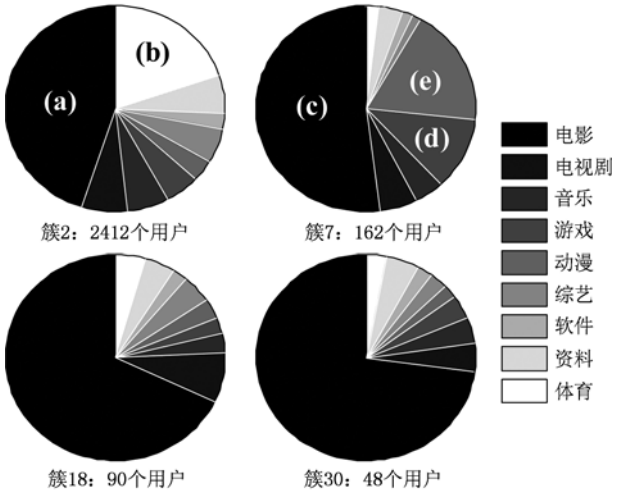

图5 用户兴趣趋向的共性与差异

为人的特性,有助于相关应用的理论建模,并为基于用户兴趣的协议设计提供了新的出发点.

5.2. 资源网络中的微观特性

各个不同的资源因用户的下载而建立连接,形成了资源网络. 网络中的每个节点代表一个资源,每一个簇都是多个具体资源的聚合,分属于不同簇的资源和处于同一簇内的资源间存在着一些微观层面上的性质.

我们从资源网络中选取第 2,6,13 和 17 簇,并辅以各资源被下载次数的比例灰度图(颜色越深表明被下载次数越多,资源越流行)加以对照分析,如图 6 所示. 横坐标表示资源的分类,纵坐标表示每个资源子类在相应簇中所占比例.

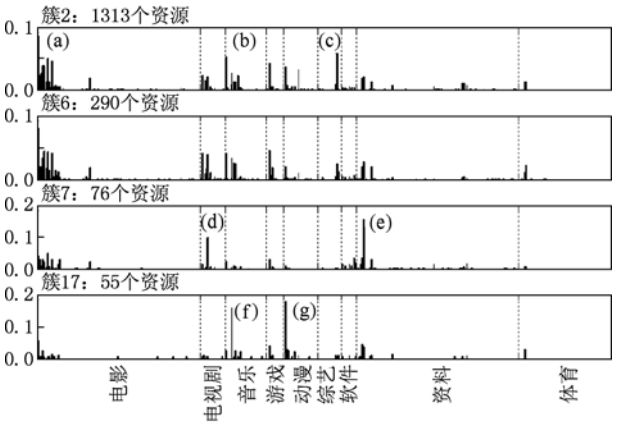

图6 各类资源之间的关联关系

从总体上看,各簇均包含了所有类别的资源,但比例不同. 具体到簇的层面上,每一簇均有不同类别的资源间存在很强的关联性,但相互关联的具体类别与其流行度有密切的关系,这里的关联性强是指某簇中某几类资源所占比例较大且近似相当. 在包含资源个数较多的大簇中,产生关联的资源多属于流行度较高的类别,如第 2 簇中的电影(a)、音乐(b)和综艺(c)等. 而包含资源较少的小簇中,产生关联的具体资源的流行度虽然不高,但某些用户稍显特殊的兴趣偏好仍然促成了他们之间的紧密连接,如第 13 簇中的电视剧(d)和各类资料(e),第 17 簇中的音乐(f)和动漫(g).

1995 年,Blischok 在对某一商店特定时间段内的售出物品分析后,发现了啤酒与尿布之间强烈的关联性,进而推动了零售商销售模式的改进[18]. 本文对于具体资源类别关联性的分析,希望能够有助于实际网络应用中资源推送机制的设计,提高资源推送准确率的同时降低服务器的维护开销.

6. 结 论

对于对等网络应用中的网络统计特征的发掘和分析,有助于人们理解这种典型网络应用的丰富内涵. 本文基于对等网络服务提供者的日志数据,抽象出用户网络与资源网络,结合复杂网络及人类行为动力学的研究方法进行统计分析. 从实验结果可以看到:1)用户节点度值及权值呈分段分布,体现了用户各异的活跃性,资源节点度值的近似幂律特性分布表明各资源的流行度差异明显. 2)两种网络具有明显的分簇结构,少数簇含有大量节点,多数簇所含节点数量较少,资源网络簇大小呈现 $\alpha \approx -1.65$ 的幂律分布. 3)用户网络中,用户的兴趣趋向存在着共性,而各簇间用户的兴趣存在差异;资源网络中,不同类别的资源间呈现出明显的关联性.

根据实测数据形成的用户网络与资源网络中,节点由带有权值的边连接在一起. 加权边反映了节点之间相互作用的强度(strength),在实际应用背景下有着特定的物理意义:如连接两个用户的加权边反映了二者兴趣广度上的关联性. 这种加权网络近些年来得到了复杂网络研究的广泛关注[19—22],出现了诸多研究模型. 对这种实际应用背景下加权网络的分析,是进一步工作的重点.

感谢北京邮电大学北邮人 BT 技术团队的支持,尤其要感谢网络技术研究院丛群老师和田旭博士在数据提取方面给予的协助.

[1] Liu F, Shan X M, Ren Y, Zhang J, Ma Z X 2004 *Acta Phys. Sin.* **53** 273 (in Chinese) [刘 锋、山秀明、任 勇、张军、马正新 2004 物理学报 **53** 273]

[2] Zhang P P, He Y, Zhou T, Su P P, Chang H, Zhou Y P, Wang B H, He D R 2006 *Acta Phys. Sin.* **55** 1 (in Chinese) [张培培、何 阅、周 涛、苏蓓蓓、常 慧、周月平、汪秉宏、何大韧 2006 物理学报 **55** 1]

[3] Wang L, Zhou S H, Yuan J, Ren Y, Shan X M 2007 *Acta Phys. Sin.* **56** 36 (in Chinese) [王 磊、周淑华、袁 坚、任 勇、山秀明 2007 物理学报 **56** 36]

[4] Zhang H F, Michael S, Fu X C, Wang B H 2009 *Chin. Phys.* B **18** 9

[5] Wei W F 2009 *Acta Phys. Sin.* **58** 4 (in Chinese) [尉伟峰 2009 物理学报 **58** 4]

[6] Guo J L 2010 *Acta Phys. Sin.* **59** 6 (in Chinese) [郭进利 2010 物理学报 **59** 6]

[7] Wang F, Moreno Y, Sun Y R 2006 *Phys. Rev.* E **73** 036123

[8] Sarshar M, Roychowdhury V 2004 *Phys. Rev.* E **69** 026101

[9] Yoon S, Lee S, Yook S H, Kim Y 2007 *Phys. Rev.* E **75** 046114

[10] Lee S, Yook S H, Kim Y 2009 *Phys. Rev.* E **80** 017102

[11] Daniel S, Reza R, Subhabrata S 2008 *IEEE Trans. on Networking* **16** 2

[12] Iamnitchi A, Ripeanu M, Foster I 2003 *Info Com*

[13] Wang X F, Li X, Chen G R 2006 *Complex Network Theory and Application* (Beijing: Tsinghua Press) p10 (in Chinese) [汪小凡、李翔、陈关荣 2006 复杂网络理论及其应用 (北京: 清华大学出版社) 第 10 页]

[14] Barrat A, Barthélemy M, Pastor S R, Vespignani A 2004 *Proc. Natl. Acad. Sci. U. S. A.* **101** 3747

[15] Newman M, Girvan M 2004 *Phys. Rev.* E **69** 026113

[16] Clauset A, Newman M, Moore C 2004 *Phys. Rev.* E **70** 066111

[17] Rosvall M, Sneppen K 2009 *Phys. Rev.* E **79** 026111

[18] Blischok T 1995 *Chain Store Age Executive with Shopping Center Age* **71** 3

[19] Pan Z F, Wang X F 2006 *Acta Phys. Sin.* **55** 8 (in Chinese) [潘灶烽、汪小帆 2006 物理学报 **55** 8]

[20] Qin S, Dai G Z, Wang L, Fan M 2007 *Acta Phys. Sin.* **56** 11 (in Chinese) [覃 森、戴冠中、王 林、范 明 2007 物理学报 **56** 11]

[21] Xu Q X, Xu X J 2009 *Chin. Phys.* B **18** 3

[22] Pu C L, Pei W J 2010 *Acta Phys. Sin.* **59** 6 (in Chinese) [濮存来、裴文江 2010 物理学报 **59** 6]

Network statistical analysis in peer-to-peer application*

Li Yi-Peng[1)†] Ren Yong[1)] Yuan Jian[1)] Wang Yue[1)] Huang Xiao-Hong[2)] Shan Xiu-Ming[1)]

1) (*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

2) (*Institute of Networking Technology, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

(Received 7 May 2010; revised manuscript received 19 August 2010)

Abstract

The rich statistical characteristics in peer-to-peer (p2p) network have recently attracted much research interest. This paper reveals the internal network statistical characteristics in the user network and resource network, both of which are abstracted from the real application downloading logs. The two-segment degree and weight distribution of user nodes indicate the dynamic of p2p users, and the similar power-law distribution of resource nodes shows the popularity diversity. Furthermore, we found that these two networks have the inherent cluster structure, only minority of clusters contain a large number of nodes, and the majority have fewer nodes in it. In user network, users in the same cluster have similar file-sharing interest, in contrast to the different user interest between clusters; meanwhile, there are obvious correlations between different resource categories in resource clusters.

Keywords: peer-to-peer network, cluster structure, network statistical characteristic

PACS: 89.20.Hh

* Project supported by the National Natural Science Foundation of China (Grants No. 60932005), and the State Key Development Program for Basic Research of China (Grant Nos. 2007CB307100, 2007CB307105).

† E-mail: yp-li05@mails.tsinghua.edu.cn