在线社交媒体数据抽样方法的比较研究

崔颖安[1),3),4)]， 李雪[2)]，王志晓[1),3)] 张德运[1)]

[1)](西安交通大学电信学院， 西安 710049)

[2)](陕西师范大学国际商学院 西安 710062)

[3)] (西安理工大学计算机科学与工程学院， 西安 710048)

[4)] (网络计算与安全技术陕西省重点实验室(西安理工大学)， 西安 710048)

摘 要 社交媒体数据是行动者自组织关系的集合，其内部蕴含了多层次的社会实体关系，传统的抽样方法难以处理其内生的复杂性、不确定性以及涌现性，因此社交媒体抽样方法的研究对于社会计算这一新兴研究领域具有重要的研究价值和实践意义。本文首先按照社交媒体抽样技术发展的演进轨迹，对广度优先抽样法、点—边抽样法、用户均匀抽样法、同伴推动抽样法以及随机行走抽样法的基本思想、概率化控制能力、应用效果进行了全面的分析和比较，介绍了各类方法的特点与不足。其次根据领域问题研究的需要，使用社交媒体数据对上述方法进行了实际测试。测试结果表明现有抽样方法在微观层次（节点）和中观层次（子群）通过有效的节点规模扩张和概率控制，能够满足节点异质性与子群内聚性抽样的要求，但在宏观层次上却无法准确刻画由于局部凝聚子群再组织所表现出的涌现性。最后以此为依据指出社交媒体数据抽样未来需要进一步深入研究的问题。

关键词 在线社交媒体;社会计算;社交网络;抽样技术;马尔科夫随机行走;抽样评价

中图法分类号 TP301

A Comparison on Methodologies of Sampling Online Social Media

Cui Ying-An[1), 3), 4)], Li Xue[2)], Wang Zhi-Xiao[1), 3)], Zhang De-Yun[1)]

[1)](School of Electronic and Information Engineering, Xi'an JiaoTong University, Xi'an 710049)

[2)](School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048)

[3)](Key laboratory of Network Computing and Security Technology Ministry of Shaanxi Province(Xi'an University of Technology), Xi'an 710048)

[4)](School of International Business, Shannxi Normal of University, Xi'an 710062)

Abstract This paper provides an overview and information useful for the sampling of online social media. The collections of actors' self-organized relations constitute the data of online social media that contain multi-level social relations. The traditional sampling methods are difficult to deal with their endogenous complexity, uncertainty and the whole emergence. So, the sampling method of online social media is significance and importance for the social computing, which is an emerging field in recent years. In accordance with the evolution of technology trajectory on the sampling of online social media, this paper introduces the characteristics and shortcomings of these methods, such as the Breadth-First-Search sampling, the "Node-edge" sampling, the Uniform sampling of UserIDs, the Respondent Driven Sampling and the Random Walk sampling based on its thinking, probabilistic control, and the results of application with a comprehensive analysis, which uses the

本课题得到教育部中央高校基金“在线社会化营销口碑群体发现(No.13SZYB01)；陕西省社科联重大理论与现实问题研究项目“陕西休闲农业的在线社会化媒体营销机制研究”(No. 2013C124)；中国电信“社会化媒体大数据云服务商业模式的研究”的资助（NO.SN2012-YS-13709）。崔颖安，男，1975 年生，博士，E-mail: cuiyan@xaut.edu.cn，讲师，主要研究领域为社会化媒体抽样、大数据分析与社交商务。李雪，女，1974 年生，博士后，E-mail: lixue@snnu.edu.cn，讲师，主要研究领域为社会计算、口碑营销、在线社会化营销。 王志晓，男，1977 年生，博士，E-mail: wzxiao@xjtu.edu.cn，讲师，主要研究领域为在线社会网络动力学分析、社会计算，张德运，男，1949 年生，博士，教授，博士生导师，E-mail: zhangdeyun@xjtu.edu.cn，主要研究领域为网络体系结构、网络工程、复杂系统分析。

microblog data for the actual test based on the needs of the field. The research results show that the existing sampling methods at the micro-level (nodes) and meso-level (subgroup) simply by node expansion and the probability of effective control to meet the node heterogeneity and organizational subgroups sampling requirements, but at the macro level, the local cohesion of subgroups and their organization of emergent properties cannot portrayed accurately. At last, this paper pointed out the future research directions of the online social media sampling.

Key words online social media; social computing; social network; sampling techniques; markov random walk; sampling evaluation

1 引言

社交媒体是帮助用户自主创造内容、群体化意见分享、自组织建立社会网络的新型互联网应用，典型的社交媒体包括博客、微博、维基、视频分享网站、社交网站、点评社区等。根据中国互联网信息中心发布的研究报告:截止 2013 年 6 月底，博客、微博、社交网站、视频分享网站国内的注册用户数已经达到 19.06 亿，日均访问用户数约 23.34 亿人次，是最活跃的互联网应用。广大网民使用社交媒体传递信息、建立联系、表达情感，以自我认同的价值为中心，构造新型的网络化组织，给当代中国社会带来前所未有的影响，因而有关社交媒体的研究已经成为社交商务、市场营销、数据挖掘、舆情评测、知识管理等多个领域学者共同关注的热点研究问题[1-5]。

尽管不同学科对社交媒体研究的侧重点各不相同，但是支撑这些研究的基础均依赖于社交媒体数据。社交媒体数据与传统行为科学数据相比有两个突出的特点[6-8]：一是数据规模大，其获取、存储、检索、展示的成本很高，是典型的大数据，通常情况下，对总体进行分析不具有可行性，只能选择抽样分析；二是数据结构复杂，社交媒体数据是行动者自组织关系的集合，其内部蕴含了多层次的社会实体关系，传统的抽样方法难以处理如此复杂的内生相干性。

研究表明：经过长期运营的社交媒体系统内部通过择优连接确实形成了一些高入度的节点（例如微博大 V），但是这些节点之间缺少直接联系，散乱的分布在不同的凝聚子群内部，没有形成全局意义上的核心群体。在宏观尺度上不具备“核心-边缘”的结构特征，“去中心性”是社交媒体内部结构的重要特征。

在不同凝聚子群内部，用户的交互特性非常复杂，有 4 种典型的互动模式。包括直连模式（边缘节点直接连接到核心节点，此时边缘节点、核心节点同属一个凝聚子群）、中转模式（边缘节点通过第三者连接到核心节点，此时边缘节点、第三者、核心节点同属一个凝聚子群），结构洞模式（某一凝聚子群内的边缘节点通过第三者连接到另一凝聚子群的边缘节点），多边模式（某一凝聚子群内的大量边缘节点直接连接到另一凝聚子群的边缘节点或核心节点，且联系具有互惠性）。

对于大型社交媒体，传统的抽样方法难以处理宏观层次以及各凝聚子群内部的复杂相干关系，暴露出不少问题。首先是总体抽样框不知，缺少先验信息的指导，只能选用非概率抽样进行探索，非概率抽样方法自身的不足导致抽样信度和效度缺乏有效的保证。其次样本选取方法存在不足，已有抽样方法只能选择相继关系作为入样单元。除了直连模式以外，其它模式下样本点的判定非常困难，因此高度节点过度入样的问题时有发生且难以克服。纵然加入纠偏机制，也难以处理结构洞模式与多边模式，样本结构与总体结构的相似性难以得到保证。最后是抽样效率低下，收敛判断困难、燃烧预热时间不确定、抽样并行化困难。

已有实证研究表明不恰当的抽样会扭曲在线社会网络的结构关系及动力学特性[9-11]，因此社交媒体抽样方法的研究就成为社会计算这一新兴研究领域的基础科学问题。唯有构建理论完备、易于实施的抽样方法，才能正确认识社交媒体主体间性关系的本质特征，克服其内生的复杂性、不确定性以及涌现性，为后续问题的研究打好基础，提高相关研究的准确性和可信性。

本文第 2 节简要介绍了统计学中有关抽样的基础知识；第 3 节对社交媒体抽样研究工作现状进行了介绍；第 4 节目前主流的社交媒体数据抽样方法进行了实际测试与比较分析；第 5 节对全文进行了总结，并展望了下一步的研究工作。

2 预备知识

抽样是一种非全面调查方法，随着现代统计科学的不断发展，已经形成较完善的理论体系，为了便于阅读，本文对抽样相关基础知识作简要介绍，如需深入了解，可参考文献[12]。

2.1 抽样理论基础

大数定理:设 $X_1, X_2, ...X_n$ 是相互独立，服从同一分布的随机变量序列，具有数学期望 $E(X_k)=\mu$，取任意 n 个单元的算术平均 $\frac{1}{n}\sum_{k=1}^{n}X_k$，则对任意 $\varepsilon>0$，有

$$\lim P\{|\frac{1}{n}\sum_{j=1}^{n}X_i-\mu|<\varepsilon\}=1 \quad (1)$$ [12]

大数定律说明了随机事件的稳定性，尤其当 n 越大时，算术平均值 $\frac{1}{n}\sum_{k=1}^{n}X_k$ 越趋近于 μ，这说明样本数量越多，其统计特性就会越接近总体的分布特性。

中心极限定理：设 $X_1, X_2, ...X_n$ 是相互独立，服从同一分布的随机变量序列，具有数学期望 $E(X_k)=\mu$，方差 $D(X_k)=\sigma^2$，则随机变量之和 $\sum_{k=1}^{n}X_k$ 的标准化变量如式（2）：

$$\lim_{n\to\infty}Y_n=\lim_{n\to\infty}(\sum_{k=1}^{n}X_k-E(\sum_{k=1}^{n}X_k)/\sqrt{D(\sum_{k=1}^{n}X_k)})$$

$$=\sum_{k=1}^{n}(X_k-n\mu)/\sqrt{n}\sigma \sim N(0,1) \quad (2)$$ [12]

式（2）说明无论总体是什么分布，从中抽取容量为 n 的样本时，只要 n 足够大，其样本平均数的分布就趋于数学期望为 μ，方差为 σ^2/n 的正态分布。显然，当 n 越大时，抽样误差越小，因而通过合理的抽样完全可以获取总体的特征。

为了使得抽样序列的选择既具有随机性还具有概率特征，MCMC（Markov Chain Monte Carlo）方法广泛的用于样本抽取中。MCMC 是一种在各学科广泛使用的随机化方法，其基本思想是构造一个具有非周期、不可约特性且平稳分布为 $p(x)$ 的 Markov 链，如果马尔科夫链足够长，其模拟的数值可视为目标分布 $p(x)$ 的独立样本以用于各种统计推断，其过程如下：

（1）概率空间 (Ω,F,P) 的随机变量 $X=\{X^t,t\in T\}$，使用随机变量 X 构造一个转移概率为 $p(x)$ 的 Markov 链，其分布函数具有平稳性。

（2）随机选择 X 中的点 X_0 做为初始种子，而后使用 Markov 链生成相应的点序列 $X_1.....X_n$。

（3）对某个常数 m 和充分大的常数 n，在概率区间 $[m,n]$ 上的函数 $f(x)$ 的数学期望可表示为：

$$E(f)=\frac{1}{n-m}\sum_{t=m+1}^{n}f(X^t) \quad (3)$$ [12]

其转移概率函数为：$p(x\to B)=\int_B p(x,x')dx$。

2.2 样本抽选方法

根据样本抽选方法，抽样方法可以分为非概率抽样和概率抽样。概率抽样是以概率理论和随机原则为依据来控制样本的抽中概率，非概率抽样是调查者根据自己的方便或主观判断抽取样本，没有确定的抽中概率，详细分类如图 1 所示[13-14]。

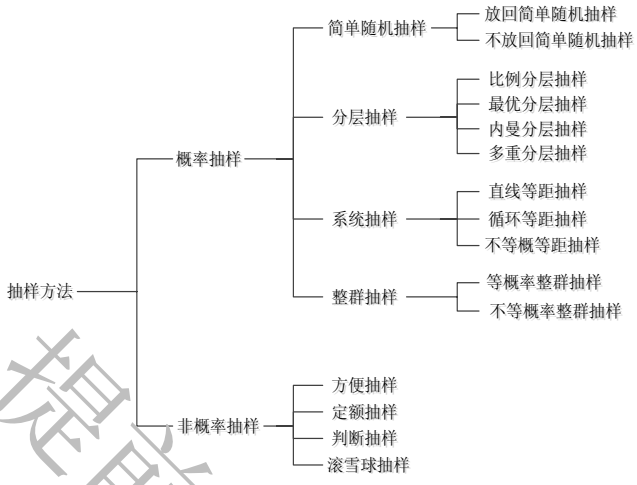

图 1 样本抽选方法

（1）概率抽样方法：

① 简单随机抽样(Simple Random Sampling): 若总体由 N 个单位组成，从中随机抽取 n 个样本，根据样本在使用完以后是否放回总体，可进一步分为放回抽样和不放回抽样。简单随机抽样方法的优点是实施方便，估算简单，缺点是抽样指导性差，抽取样本规模大，后续数据处理工作量大。

② 分层抽样(Stratification Sampling): 将总体 N 个单位按其属性特征分成 L 个互不重复的子总体，每个子总体称为层，而后在每个层中独立的抽取样本单位。分层抽样的优点是适用领域广、抽样精度高、易于组织与管理，缺点是对层的临界特性要求高，只有层间差异非常明显才能获得较好的分层效果。

③ 系统抽样(Systematic Sampling): 将 N 个总体单元按一定顺序排列，先随机抽取一个单元作为

样本的第一个单元，即起始单元，然后按照某种确定的规则抽取其他样本单元。系统抽样的优点是方法简单、易于实施，缺点是若总体中的单元具有周期性，抽样准确性就会受到影响，另外方差计算困难。

④ **整群抽样(Cluster Sampling):**将总体中N个单位分解成不同的子群（群内由若干个有联系的基本单元组成），然后以群为抽样单位，从总体中随机抽取一部分群，对被选中群内的所有单元进行研究。整群抽样的优点是不需编制庞大的抽样框，适合样本间存在联系的大型调查对象，缺点是群的划分方法与调查对象具有相关性，群的划分对抽样质量影响明显。

（2）非概率抽样方法：

① **方便抽样(Convenience Sampling):**调查者以自己方便的方式抽取偶然得到的样本，最典型的方便抽样是“街头拦人法”。方便抽样的优点是易于实施，代价较小，缺点是样本代表性差，有很大的偶然性。

② **定额抽样(Quota Sampling):**调查者先将总体按某种特征划分成不同的组，然后在配额内以主观判断选定样本作为研究对象。定额抽样和分层抽样的相同之处是对总体进行分组，不同之处是分层抽样按概率原则在层内抽选样本，而定额抽样选取样本是主观的。定额抽样的优点是能够缩小抽样范围，减少抽样成本，缺点是确定额度困难，需多次探索。

③ **判断抽样(Judgment Sampling)**研究人员根据调查目的和主观经验，从总体中选择最具代表性的样本。判断抽样的优点是可以用于总体难以确定的研究对象，缺点是受研究人员的主观倾向性影响大，一旦主观判断失误，则易引起较大的抽样偏差。

④**滚雪球抽样(Snowball Sampling):**先选取若干符合特征的样本构成最初的调查对象，然后依靠他们提供新的调查对象，随着调查的推进，样本如同滚雪球般由小变大。滚雪球抽样方法的优点是能够很方便的找到被调查者，适用于探索性研究，缺点是样本之间必须存在联系且愿意保持和提供这种联系。

3 研究工作现状

综合国内外相关研究文献，目前常用的社交媒体抽样方法包括广度优先抽样法、点—边构造法、用户均匀抽样法、同伴推动抽样法以及随机行走系列抽样方法，下面分别予以介绍：

3.1 典型抽样方法

（1）广度优先法(Breadth-First-Search Sampling, BFS)

BFS 的基本思想是从网络中选择初始节点入放入先进先出队列，而后搜索与之相邻的所有节点，如果这些节点在队列中尚未出现，则节点入队并记录该节点的父节点，反之则从队首取另一节点对其相邻节点做相同处理直至队列为空。由于 BFS 方法可以彻底搜索整个网络，加之易于编程实现，因而在多个领域得到广泛应用。

文献[15]使用 58 台服务器组成的集群对 Flickr、LiveJournal、Orkut、YouTube 的关系数据分别进行抓取，为了验证 BFS 的抽样质量，使用 WCC（weak connected component）指标对抽样数据与总体进行了比较分析，结果显示抽样准确率在 85%-95%。该文提出的局部全连通子网抓取思想、反向链接数据补充方法以及抽样效果评价方法对后续的研究者产生了较大的影响。文献[16]对 Facebook 用户照片点评行为与在线社会关系的相关性进行了研究，使用文献[15]中的方法对排名前 22 位地区的用户数据进行抽样分析并对旧金山地区的抽样数据进行了验证，结果显示抽样准确率高达 99%。文献[17]使用文献[16]中的抽样方法对 Facebook 的社区特征进行了研究，为了验证抽样数据的准确性，作者将抽样数据与文献[18-19]的结果进行了比较分析，结果显示本次抽样数据与以往抽样数据具有一致性，都表现出较小的网络直径(diameter)和较高的簇类系数(clustering coefficient)。

尽管 BFS 具有较强的样本获取能力，但是该方法在使用过程中也暴露出不少问题：文献[20-21]指出 BFS 方法抽样准确性高度依赖样本规模（图 2 所示）。如果样本规模低于 40%，抽样与总体就会有较大的偏差，不能为后续研究提供准确的数据。文献[22]指出随着样本规模的不断增大，BFS 方法的抽样效率越来越低，新样本发现能力呈现出快速衰减趋势。文献[21-22]指出 BFS 方法存在高度节点过度入样的问题，文献[23]对这一现象进行了理论分析，如式（4）所示。由（4）可知 BFS 高度节点过度入样并非偶然误差，而是 BFS 方法的内生缺陷。

$$\left\langle q_k^{BFS}\right\rangle(f)=\sum_k k\cdot\left(\sum_l \frac{p_k(1-(1-t(f))^k)}{p_l(1-(1-t(f))^l)}\right) \quad (4)$$ [23]

为了解决上述问题，文献[24]认为只要扩大 BFS 抽样规模即可对冲高度节点的过度入样。尽管该方法看似简单，但是其可行性却很差。因为大型社交媒体即使只提高 1%的入样率，就会增加上百万个节点、数千万条关系，为后续分析带来很多不便。文献[25]试图给出 BFS 的一般性抽样纠偏方法，如式（5）所示：

$$\hat{p}_k^{BFS}=\frac{\hat{q}_k}{1-(1-t(f))^k}\Big/\left(\sum_l \frac{q_l}{1-(1-t(f))^l}\right) \quad (5)$$ [25]

然而，该方法的缺陷正如作者在文中所指出的那样：p^k 与 $t(f)$ 两者互为因果，就如同“蛋与鸡”的关系一样，因此该方法实用性有限，只有理论参考价值。

文献[25]利用 BFS 非成回抽样的特性，将其改造成“主—从”式抽样架构。用“主”将 BFS 队列中的节点分配到不同的机器中，用“从”对不同节点的社会网络进行并行遍历以提高抽样效率。对于该方法，要有清醒的认识：其实质是通过并行化在单位时间内通过增加样本规模来提高抽样效率，而不是通过调整样本的入样概率来提高抽样质量。

BFS 是一种图遍历方法，在抽样理论中，BFS 属于非概率抽样中的滚雪球方法。将其用于社交媒体抽样时，可以充分利用其“竭尽所能，向前探索”的特性得到一个局部全连通的子网。这个子网以及子网内部元素的选取并不遵循随机原则，因而无法按照概率特性对候选样本进行控制，这是将 BFS 方法用于抽样时存在的根本问题，因此该方法很难进行调整和控制。

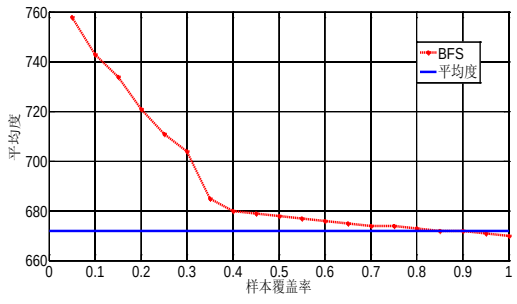

图 2 BFS 抽样质量与样本规模的关系

（2）“点—边”构造法（Node-Edge Sampling , NES）

“点—边”构造法是随机点抽样法与随机边抽样法及其系列改进方法的总称[26-27]，随机点抽样法与随机边抽样的基本思想都是以等概率不放回的方式随机抽取一定数量的节点或者边，而后对这些节点或边的关系进行分析从而导出抽样子网。

典型的随机点抽样改进方法包括 RPN 方法(Random PageRank Node)和 RDN 方法(Random Degree Node)。这两种方法的共同之处都是将传统的等概率抽样改进为不等概抽样，抽样结果有意偏向高度节点，希望以最小的代价尽可能多的获取核心节点（信息）。不同之处是前者使用 PageRank 作为度量标准，后者使用节点度作为度量标准。显然以上两种改进方法都是针对搜索引擎抽样而设计的特异性方法，不适合社交媒体的抽样。

典型的随机边抽样改进方法包括 RNE 方法(Random Node Edge) ,HYVE（Hybrid Vertex/Edge）方法和 TIES（Totally-Induced Edge Sampling）方法。这些改进方法的共同之处是采用“点—边”相结合的方法共同进行抽样而不是仅使用“边”作为抽样对象，其中最具实用性、抽样效果相对最好的是 TIES 方法。文献[27]给出了 TIES 方法，其抽样过程分为两步，首先进行传统的边抽样，而后根据总体补充样本节点之间的关系即可。例如使用边抽样法获得关系 $e_1=(v_1,v_2),e_2=(v_3,v_4)$，而后检查原图，如果节点 v_1,v_3 存在关系则补充相应的关系，直至达到预设采样规模。文中以 HepPH, Twitter, ConMat, PU-Email，Facebook, Enron-Email 为研究对象进行了抽样测试，结果显示 TIES 要优于其他“点—边”抽样法，能够有效提高样本覆盖率，RE,RN,TIES 抽样效果对比如图 3 所示。

尽管“点—边”构造法可以用于社交媒体抽样，但是该方法可用的前提条件是总体已知，因为只有总体已知，才能等概率的获取节点或者关系。在抽样理论中，“点—边”构造法属于非概率抽样中的定额抽样方法。很明显，“定额”这一特定的前提条件在多数情况下难以具备，因此该方法实用性较差，在实际抽样中较少使用。

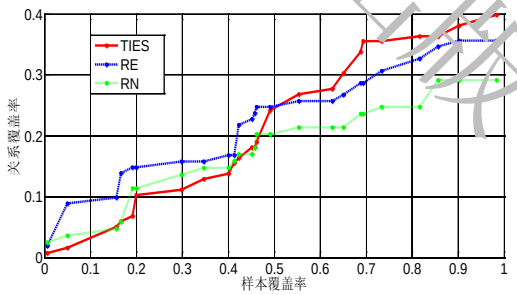

图 3 TIES, RN, RE 抽样覆盖率

（3）用户均匀抽样法(Uniform Sampling of UserIDs，UNIS)

UNI Sample 是一种针对社交媒体的特异性抽样方法，它的主要作用是为其他抽样研究提供一个可供参考的真实总体，以验证其他抽样方法的有效性。UNI Sample 方法分为两步：首先根据抽样对象的编码规则生成规模为[0,MaxUserID]的样本空间，用程序自动验证每一个 UserID，删除不存在的 UserID，保留有效 UserID，形成与客观实际相一致的总体。而后根据抽样质量的要求，等概率的使用 UserID 作为种子进行关系发现，直至达到抽样要求。文献[17, 18, 24]中采用 UNI Sample 方法进行了总体探测，并依此作为标准对其他抽样方法进行了比较分析。

尽管 UNI Sample 方法比 BFS 方法探索总体的准确率和效率要高得多，但是该方法还是存在两个难以治愈的“顽疾”。首当其冲的问题就是总体的估计能力有限。如果研究对象的 ID 编码规则不具有规律性或者研究对象的 ID 编码空间过大（64 位或更大），准确估测总体就非常困难。事实上，大量社交媒体的 UserID 编码规则都不具有规律性且编码长度不固定，该方法的适用性就受到了很大的挑战。另外该方法还有一个颇受质疑的问题——等概率使用 UserID 作为入样控制依据能否准确刻画社交媒体用户的内在关系特性？图 4 给出了真实社交媒体 UserID 的分布特征，图中表明 UserID 并不具有均匀分布的特性。如果针对这样的数据采用 UNI Sample 方法进行抽样，就会产生很大偏差。为了减少误差，不得不抽取更多的节点，这既加大了抽样的成本又增加了后续研究的难度。

UNI Sample 方法虽然具有一定的局限性，但是该方法对于社交媒体的抽样研究具有重要的意义。因为这是目前唯一能够揭示社交媒体总体特性，将其从“黑盒子”变成“白盒子”的抽样方法。在抽样理论中，UNI Sample 方法属于概率抽样中的系统抽样方法。这意味着统计学中有关样本容量确定、总体指标估计、抽样误差分析、信度分析等经典理论就可能用于社交媒体这一新生事物，从而有效的指导抽样。

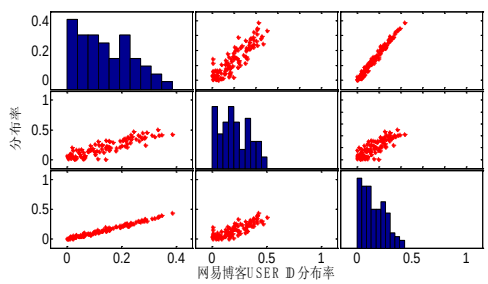

图 4 User ID 的实际分布

（4）同伴推动采样法(Respondent Driven Sampling，RDS)

RDS 是由滚雪球方法演变而来，但又与滚雪球方法存在较大差异的抽样方法，它主要用于总体与样本都难以确定社会现象的研究。该方法从调查对象的社会网络中招募同伴参加研究，并通过样本特征对总体做出渐进无偏估计和推断，目前已在艾滋病、吸毒、同性恋群体的研究中得到广泛应用。

传统的滚雪球方法主要存在三方面的问题：一是抽样结果与初始种子选取具有相关性；二是存在大群体、高度节点、活跃子群过度抽样的问题；三是抽样受后继关系影响明显，样本过度同质化导致代表性不足。Heckathorn 在 1997 年采用事后分层抽样的思想建立了 RDS 方法，经过 3 次大的修改与完善[28-30]，形成了目前较完善的 RDS 方法。该方法主要包括三部分内容：一是利用马尔科夫链的平稳性解决了初始种子敏感性的问题；二是使用汉森—赫维茨估计法（式（6））进行分层入样控制以减少抽样偏差；三是控制后继样本的入样个数，该方法规定每一样本至多只能选取 3-4 个后继样本以减少样本同质性偏差。

$$\tilde{y}_{HH}=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i} \qquad (6)\ [13]$$

文献[31]为了研究微博用户之间互粉以及发帖行为特征，使用 RDS 方法对 Twitter 进行了 15X24 小时的数据采集，采集信息包括用户基本信息、朋友列表、微博内容等。为了确保抽样数据的准确性，以 UNI 抽样为标准，与 MHRW、TimeLine 抽样方法进行了比较分析，结果显示 RDS 抽样与 UNI 方法最接近(图 6)，抽样质量好于 MHRW、TimeLine。文献[32]采用 RDS 与 MHRW 抽样方法对静态网络（随机网络、BA 无标度网络、小世界网络）及动态 P2P 网络（Gnutella 协议）分别进行了抽样比较测试，结果显示在静态网络中随着抽样规模的不断扩大，RDS 的抽样质量要好于 MRW 方法，对于具有复杂结构特征的分层无标度网络，尽管 RDS 方法的抽样质量仍优于 MRHW,但是抽样偏差也比较大。

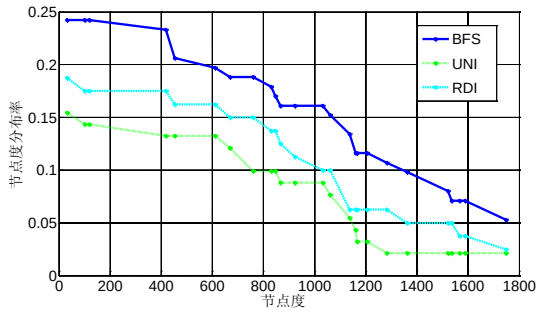

图 5　RDS 抽样与 UNI Sample 拟合效果

文献[33]注意到有些社交媒体除了提供常用的朋友列表功能以外还提供圈子、点评等功能，这些社会化功能可以有效的组织用户，形成活跃的社区群体，因此有必要全面的抽取社交媒体内部的各类关系。作者在综合以上各类社会关系的基础上，采用 RDS 方法对 Last.fm 进行了抽样测试，结果显示 RDS 方法能够很好的进行复合关系的抽样，同时还可以比单独关系抽样更快的获得平稳分布。

文献[34]对标准 RDS 方法进行了简单调整，采用随机行走获取后继样本（将每次取 3-4 个后继样本改为只取 1 个），采用调和平均数进行样本估计。为了验证改进的有效性，以平均度作为指标对 Twitter(2009 年 7 月)的总体数据进行了抽样测试，结果显示抽样效果明显好于均匀抽样。该文以网络平均度（式（7））为指标对网络的结构特征进行了分析，式（7）表明内部关系的不均匀分布是 RDS 不等概估计方法成立的客观基础，因而该方法才会获得比均匀抽样更好的抽样效果。

$$\bar{\upsilon}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i^W}\quad，其中\ d_i^W=\frac{B}{(a+i)^b}\qquad（7）^{[34]}$$

尽管 RDS 方法用于社交媒体抽样的案例并不多，但是其意义重大。从抽样方法的技术演进来看，RDS 方法是一道分水岭，该方法对应的是概率抽样中的分层不等概抽样方法，它是真正通过概率调整来控制入样节点，从而使其较好的刻画出总体的特性。文献[32]对 RDS 方法适用性的分析非常富有启发性，它再一次验证了物质决定意识这一基本哲学原理。就社交媒体抽样这一特定研究领域而言，最佳方案应当是建立一种洞察总体内部结构特性的机制，而不是单纯依靠规模扩张的蛮干式抽样。

（5）随机行走抽样法（Random Walk Sampling, RWS）

随机行走是一种经典的随机化方法，在图论领域得到了广泛应用。文献[35-36]给出了随机行走在图中的收敛特性，因而将随机行走用于社交媒体抽样时必须对其进行完善，以确保每一个节点都能具有相同的入样概率。

文献[37]介绍了两种简单的 RW 改进方法：Random Jump（RJ）方法与 Forest Fire(FF)方法。RJ 方法将标准 RW 方法相邻节点间的游走调整为以 15%的概率在整个图内进行跳转，以解决抽样陷入局部子网的问题。FF 方法将网络中的关系看成一颗分层树，以此为依据调整随机游走的规则，使得样本间的关系逐步具有超线性（度分布的无标度性）和高密度（网络直径变短）的特性。显然这两种改进方法都有明显的局限性，因为社交媒体内部的关系通常都是由多种分布特性混合而成，用一种随机行走的抽样控制机制或者控制参数难以准确的抽取复杂样本。因而后来的研究者采用 MCMC 方法替代了标准 RW 方法及其改进方法，主要包括以下三种典型方法：

①Metropolis-Hastings Random Walk

Metropolis-Hastings 算法是一种典型的 Markov Chain Monte Carlo(MCMC)，它使用提议分布函数进行抽样控制以构造一个具有非周期、不可约、遍历特性且与总体分布 $p(x)$ 一致的 Markov 链[38-39]，该方法能够确保生成的马尔科夫链具有细致平衡的特性，因而可以很好的满足社交媒体抽样的需要，MHRW 标准算法的概率转移核如式（8）：

$$a(x,x')=\min\left\{\frac{f(x')q(x',x)}{f(x)q(x,x')},1\right\}\qquad（8）^{[38]}$$

文献[40]对 Twitter 中不同类型用户的信息传播特征进行了研究，为了确保研究结果的正确，采用多种方法进行数据抽样并与标准数据集进行了对比，结果表明 MHRW 抽样不会受到用户自身特征的影响，能够确保初始种子选取的无后效性，可以无偏等概获取样本。标准 MHRW 方法与标准 RW 方法抽样效果对比如图 6 所示。

文献[18][19][24]以 Facebook 为研究对象，对 BFS 方法，标准 RW 方法、RWRW（RDS）方法,MHRW 方法进行了比较测试，结果显示 MHRW 和 RWRW（RDS）方法在抽样质量和效率上都明显优于 BFS 和标准 RW 方法，适用于大型社交媒体抽样。文章同时还对 RWRW（RDS）方法和 MHRW 方法进行了比较研究，结果显示两者在抽样质量及收敛速度上具有相似性，属于旗鼓相当的方法。文献[41]以 Wiki,Epinion,SlashDot 为研究对象综合比

较了 MHRW 和 BFS 方法,得出与上文相似的结论。

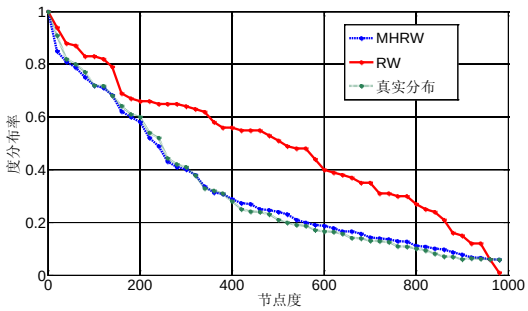

图 6 标准 MHRW 方法与标准 RW 方法抽样效果对比

文献[42]改进了标准的 MHRW 方法，增加了随机化跳跃参数 w，使其能够在所有的样本中选取下一抽样节点而不仅仅是选择其相邻节点。这种改进给大型社交媒体抽样带来了三方面的益处：一是能够避免在高度聚集的子网内过度采样；二是提高了采样方法对稀疏网络的适应能力，三是在确保平稳性的前提下提高了收敛速度。文中以 Buzznet 和 Berk-Stan 为研究对象进行了抽样测试，结果显示改进方法抽样效果明显好于 BFS 及标准的 MHRW 方法。其概率转移核如式（9）：

$$p_{u,\upsilon}=\begin{cases}\min(\frac{1-p}{k_u},\frac{1-p}{k_\upsilon})+\frac{p}{|V|} & \text{if } \upsilon \text{ is u's neighbor}\\ 1-p-\sum_{w\neq u}\min(\frac{1-p}{k_u},\frac{1-p}{kw})+\frac{p}{|V|} & \text{if } \upsilon=\text{u}\\ \frac{p}{|V|} & \text{otherwise}\end{cases} \quad (9)\ [42]$$

文献[43]使用模拟退火算法对标准 MHRW 进行了改进，期望抽样算法具有全局最优或者局部最优社区发现能力，使其能够按照社区特性进行抽样。文章以 PPI network, EPinion, HEP-PH 为研究对象进行了抽样测试，结果显示改进后的方法对复杂结构的适应能力明显提高，既能有效处理稀疏网络也能有效抽取高聚集的网络，但是美中不足的是增大了计算复杂性，降低了抽样效率，模拟退火的概率控制方法如式（10）：

$$a(S,S')=\min(1,(\frac{\Delta_{G,\sigma}(S)}{\Delta_{G,\sigma}(S')})^{\frac{p}{T}}) \quad (10)\ [43]$$

文献[44]注意到标准 MHRW 可能会出现低度节点的过度入样，提出了一种改进的 MHRW 算法称为 USRS 方法，该方法降低了低度节点的选中概率，防止其过度入样。文中以 RenRen 作为研究对象，使用归一化误差（NMSE）方法进行了抽样评价，结果显示该方法可以提高收敛速度，防止低度节点的过度入样，其概率转移核如式（11）：

$$p_{u,\upsilon}=\begin{cases}\frac{1}{k_u}\cdot\min(1,\frac{k_u}{k_\upsilon})+\Delta r_u/N_u & \text{if } \upsilon\in \text{R}_u,\\ 1-\sum_{y\neq u}P_{u,y}-\Delta r_u & \text{if } \upsilon=u,\\ 0 & otherwise.\end{cases} \quad (11)\ [44]$$

需要特别注意的是：USRS 方法成立的先决条如式（12）所示：

$$\forall i\in V,\sum_{j=1}^{|V|}P_{j,i}=\sum_{j=1}^{|V|}P_{i,j}=1 \quad (12)\ [44]$$

由上式可见，该方法只适用于对称关系抽样，因此适用面较窄。另外如何判断低度节点的过度采样也是一个难以度量的问题，给算法实用带来一定困难。

②Directed Unbiased Random Walk（DURW）是针对有向图的抽样方法。文献[45]指出非对称关系抽样主要有两方面的困难：一是零出度节点（sink node）对随机行走的影响，使得样本序列难以获得平稳分布；二是难以准确估计入度分布。

DURW 对标准的 RW 方法进行了改进，其基本思想是根据有向原图的出度关系进行节点抽样，若遇到 sink node，则进行节点自举，寻找新的样本点做种子以继续抽样。该方法的优点是能够较好的刻画出总体的出度关系，同时还可以灵活的进行跳转控制（包括 W 权值的控制、样本容量控制、随机化跃迁程度控制）。出度分布控制方法如式（13）：

$$\hat{\phi}_j=\frac{1}{n}\sum_{i=1}^{n}\frac{h_j(s_i)}{\hat{\pi}(s_i)},j=0,1,...$$
$$h_j(\upsilon)=\begin{cases}1 & \text{if the outdegree of } \upsilon \text{ in } G_d \text{ is } j,\\ 0 & \text{otherwise}\end{cases} \quad (13)\ [45]$$

S 的计算方法如式（14）：

$$S=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w+\deg(s_i)} \quad (14)\ [45]$$

文中以 MySpace，Flickr，Youtube，Wiki，LiveJournal 作为研究对象进行了抽样测试，结果显示该方法的出度分布与总体接近，具有较高的一致性。但是该方法还存在不少问题，首先确定权值非常困难，虽然作者给出了权值的经验估计方法（50%无向平均度—100%无向平均度），但是对于一个总体不知的“黑盒子”，如何才能获得平均度，如果不能获得平均度,该方法就不能实用。其次如何在抽样前确定样本容量，确定依据是什么？另外作者在文中指出，该方法入度分布的估计与实际情况偏离太大，因而该方法还需要进一步的完善。尽管该方

法还存在不足，但是其抽样思想仍然具有较高的借鉴价值，因为它采用参数去控制样本选取，如果参数选择恰当，抽样效率和质量就会很高。

③Random Walks with uniform Restarts（RWuR）

Random Walks with uniform Restarts（RWuR）是一种将 UNI Sample 抽样与 RW 抽样相结合的混合抽样方法。文献[46]指出标准 RW 在抽样过程中存在高度节点过度入样和收敛效率较低的不足。作者给出了图的谱隙（spectral gap）与抽样误差的关系，如式（15）所示：

$$\mathrm{var}_{\pi}(\hat{f}_k) \propto \frac{2\theta_k^2}{\prod_k \delta B}(1+\frac{\delta}{2B}) \quad (15)^{[46]}$$

以及与图的谱隙收敛效率的关系，如式（16）所示：

$$\sup_f \lim_{B\to\infty} \frac{\mathrm{var}(\hat{\hat{f}}) + bias(f)}{\mathrm{var}_{\pi}(\hat{f})} = \frac{1+\lambda_2}{1-\lambda_2} \cdot \frac{2-\delta}{\delta} \quad (16)^{[46]}$$

以此为依据，作者提出了 RWuR 方法，其概率转移核如式（17）

$$\hat{p}_{ij} = \begin{cases} \dfrac{\alpha/n+1}{d_i+\alpha}, & \text{if } i \text{ has a link to } j, \\ \dfrac{\alpha/n}{d_i+\alpha}, & \text{if } i \text{ does not have a link to } j. \end{cases} \quad (17)^{[46]}$$

应该说,RWuR 方法是一种具有深刻洞察力的抽样方法，它揭示了抽样现象与图的谱隙之间的内在联系，从而建立了“改进概率转移核→增大谱隙→控制抽样路径”这一非常新颖的抽样方法。从抽样思想上看，RWuR 方法与 DURW 相似，都属于参数控制法，但是 RWuR 方法将 DURW 向前推动了一大步，因为该方法采用谱隙作为辅助量进行调控更具有客观性。在实现上，RWuR 也非常巧妙，在一个概率转移核里既能控制随机行走也能通过 UNI Sample 进行节点跳转，该方法比以往的固定概率跳转方法更具有灵活性，同时也与总体特征更具有一致性。文中以 LiveJournal 与标准 BA 无标度网络作为研究对象进行了抽样测试，结果显示 RWuR 明显优于标准的随机行走方法，该方法既能减少高度节点的过度入样，还能缩短随机行走的混合时间，加速收敛到平稳分布。

当然 RWuR 方法也有不足，首先该方法成立的基础是已知总体信息，但是现实研究中很难获取总体信息，因此该方法的实用性受到了影响。另外参数 α 的选择缺乏有效的指导，对 α 不恰当的选择就会影响抽样的准确性，因而该方法还需进一步的完善，使其具有更好的实用性和操作性。

MHRW，DURW，RWuR 是随机行走法中最具代表性的方法，三个方法共同之处是具有扎实的理论基础，能够以概率理论和随机化原则进行样本控制，在抽样理论中，以上三种方法均属于不等概抽样。不同之处是 MHRW 将抽样过程转化为随机序列的拟合过程，后两者则通过参数进行抽样控制。两类方法各具特色，拟合法的适用性好，但是燃烧预热（Markov burn in）时间可能很长，抽样效率低，对于复杂分布拟合效果差；参数法的针对性强，一旦参数选择恰当，抽样效率和质量就会很好，但是参数值的调优还需要进一步的研究，以提高此类方法的适用性。

本文介绍了 5 类抽样方法，这 5 类方法代表了社交媒体抽样研究的演进与发展。BFS 方法属于遍历方法，在搜索引擎的页面抓取中大量使用，由于社交媒体抽样隐含了数据抓取这一问题，因而研究者最先将 BFS 方法用于抽样研究。应该说这是历史形成的原因，并非最佳选择。从应用效果来看，BFS 方法不能进行概率控制，其样本控制能力非常有限，在这样的背景下，“点—边”构造法应运而生。无论随机点抽样还是随机边抽样，其基本思想都是把社交媒体内部的节点或者关系看成独立样本。很明显，“点—边”构造法对样本相干性的忽视严重违背了社交媒体内生的结构特性，其实用效果比 BFS 方法还要差，因而这类方法也只是昙花一现。

BFS 方法与“点—边”构造法的不足激发了研究者探索总体的欲望，UNI Sample 方法就随之而现。与 BFS 方法相比，UNI Sample 方法既能发现连通子网也能发现离散的孤立节点，其总体探索能力和效率确实要好得多。更为重要的是一旦总体可知，社交媒体就有可能从无概率抽样转化为有概率抽样，抽样质量和抽样效率就有可能大幅度的提高，这一改变为未来的研究留下充分的探索空间。当然也必须看到 UNI Sample 方法毕竟是一种特异性方法，只能用于社交媒体的抽样研究，对于其他具有相似复杂相干关系的化学分子网络、生物网络则无能为力。

从社会科学领域借鉴 RDS 方法用于社交媒体抽样研究具有重要的里程碑意义。从抽样效果来看，RDS 方法明显好于 BFS 方法和“点—边”构造法。更为重要的是 RDS 给后续研究提供了一些重要的启发：概率化控制、随机化探索、平稳性保

证、收敛判断是提高社交媒体抽样质量的基本方法，为随机行走系列方法的研究起到了承上启下的作用。

随机行走及其改进方法的大量出现是在借鉴 RDS 方法的优点基础之上，将社交媒体抽样研究推进到了一个新的阶段。除了保持对抽样效率的关注以外，研究者更加重视抽样方法对社交媒体内部复杂结构的探索能力，更加重视细节特性的抽样控制，使抽样的与总体更具一致性。可以预见，未来围绕着总体相干性及随机行走可塑性的研究成果将会大量出现。

3.2 抽样评价方法

（1）样本评价指标

在完成抽样以后需要对抽样结果进行合理的评价以验证抽样的有效性，文献[47-50]给出了若干统计评价指标用来描述网络的特征。总的来看现有文献给出的指标基本一致，主要包括度分布、簇类系数、网络密度、网络距离等常用统计指标。截止目前，文献[48]使用 47 个评价指标对抽样网络进行了统计与分析，该文是目前选择评价指标最多的文献。

作为总体的估计量，一般情况下选取的评价指标越多越能全面说明样本与总体是否具有一致性，当然指标过多，就会影响评价的效率。经过分析筛选，本文给出了可以用于社交媒体抽样评价的主要网络评价指标（共计 28 个），如表 1 所示。在进行抽样时，读者可以根据自己的需要合理选择，也可以进一步阅读文献[47-50]更全面了解复杂网络统计指标。

表 1 抽样评价指标

测试指标	符号	参考文献
网络连通性评价关键指标		
Mean geodesic distance	ℓ	文献[51]
Harmonic mean distance	h	文献[53]
Vulnerability	V	文献[54]
lth moment	M_l	文献[55]
Matching index of edge i, j	u_{ij}	文献[56]
网络聚集性评价关键指标		
Network clustering coefficient	C and $\tilde{C}$	文献[57]
Weighted clustering coefficient	C^w	文献[58]
Cyclic coefficient	Θ	文献[50]
Betweenness centrality	B_i	文献[58]
Central point dominance	CPD	文献[59]
网络交互性评价关键指标		
Maximum degree	$k_{\max}$	文献[60]
Mean degree of the neighbors	$k_{nn}(k)$	文献[61]
Degree-degree correlation coefficient	r	文献[62]
Assortativity coefficient	$\tilde{Q}, Q$	文献[63]
Degree Distribution entropy	$H(i)$	文献[64]
网络层次性评价关键指标		
Modularity	Q	文献[65]
Significance profile	SP_i	文献[66]
Subgraph centrality	SC	文献[67]
Hierarchical clustering coefficient	Crs	文献[68]
Convergence ratio	$cvd(i)$	文献[69]
网络异质性评价关键指标		
Average search information	S	文献[50]
Access information	A_i	文献[71]
Hide information	H_i	文献[72]
Target entropy	T	文献[70]
Road entropy	R	文献[71]

（2）K-S 检验方法

K-S 检验是一种常用的检验方法,主要用于检验经验分布函数与总体分布函数间的差异显著性。该方法由 Kolmogorov 检验和 Smirnv 检验组成，前者用于检验一个给定样本 X 是否服从某个已知的概率分布 $F_x(x)$,后者检验两个样本 X_1 、X_2 是否服从同一概率分布。其工作原理如下：

设 $x_{(1)},...,x_{(N_0)}$ 是样本 $x_1,...,x_{N_0}$ 的有序统计量，因 $F_n(x)$ 由样本得到，只在 N_0 个样本处存在函数值，且为单调非降的阶梯函数。

$$F_n(x_{(i)}) = \frac{i}{N_0}, i = 1,...,N_0$$

$$D_{N_0} = \max\left\{\left|F(x_{(i)}) - \frac{i-1}{N_0}\right|, \left|F(x_{(i)}) - \frac{i}{N_0}\right|, i = 1,...,N_0\right\} \quad (18)$$ [14]

式(18)说明可以用 D_{N_0} 作为检验统计量，当 H_0 为真时， D_{N_0} 趋向于较小值，当 D_{N_0} 过大时则拒绝 H_0 。

实施 K-S 检验时，不需要对数据分组，减少了划分区间的麻烦,这样就不必因区间划分而舍弃部分样本数据，可以提高检验的质量。另外 K-S 检验能够处理任意长度的样本群体，对大样本群体的适应能力要好于 S-W 检验，这个特点非常适合社交媒体抽样检验的需要，因此 K-S 检验已经成为社交媒体抽样研究中最主流的检验方法。

4 抽样比较分析

4.1 测试环境搭建

测试系统由两部分构成：一部分是数据抓取子

系统，另一部分是数据分析子系统。数据抓取子系统由 40 台 2-CPU 机架服务器组成。通过 CPU 绑定技术，每个机架服务器运行四个独立的数据抓取线程，抓取子系统总体并发规模控制在 280-300，采集的数据集中存放在存储系统的数据库中。

数据分析子系统由 3 台 64 位 Linux 高性能服务器组成，每台服务器分别负责不同社交媒体的抽样指标的分析，彼此独立，互不干扰。整个系统部署在高速 SAN 网络中，公网接入是 50M 光纤专线，测试环境拓扑如图 7 所示。

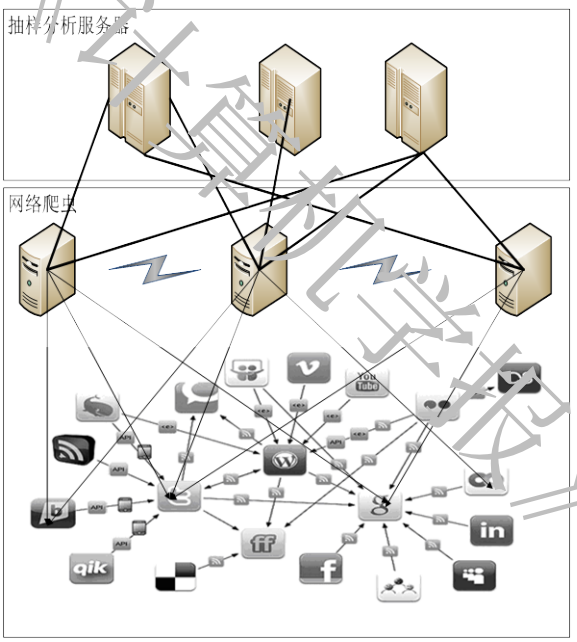

图 7 抽样测试环境

4.2 测试对象说明

我们选择网易微博（t.163.com）、蘑菇街（www.mogujie.com）、优酷（www.youku.com）作为测试对象，选择原因：首先是其数据规模庞大，使用频繁，抽样与总体对比效果明显；二是系统运营时间长，用户行为趋于稳定，具有抽样研究的稳定性基础；三是内部结构复杂，具有测试典型性。

4.3 测试数据获取

测试数据获取分为三个步骤：首先人工植入一些初始种子，要求种子必须具有出度和入度关系，不能是孤立节点。而后从初始种子作为出发点，使用标准 BFS 方法（深度只需达到 3-5 度即可）获得其相继节点。通过初始种子的采集，可以有效扩大种子的范围，形成初始数据集。

其次是将初始数据集分布到不同的服务器中，再以等概率随机扫描初始数据集，按照个体网络（ego-network）对其相继关系进行分析。分析的内容包括相继关系的基本属性（网名、地区等），出度关系、入度关系等信息。完成分析以后，对该记录做出标志，避免重复分析，同时还将新发现的节点并入待分析的数据集。

最后在每日 12 点与 0 点时，守护进程调用数据清理程序，对不必要的重复数据进行清理，如果在采集过程中出现数据不一致的情况（社会网络中的节点已注销），需要清除该节点形成的各种关系。另外对于微博数据还需根据出度关系完善部分节点的入度信息，同时还要清理明显的僵尸粉。

使用上述方法持续采集了近 180×24 小时，已很难发现新的节点，可以认为达到了对测试对象连通子网总体数据的收集。此处需要特别指明的是 BFS 并非最佳大型在线社会网络总体数据采集方法。该方法随着采集样本的不断增加，并行化冲突非常明显，采集效率较低，因而我们采用等概率随机扫描法去捕获总体。

在完成数据采集以后，我们对所有测试对象进行了较全面的社会网络分析，从其构成上看：有对称关系构成的小世界网络；也有包含大量孤立节点的随机网络；还有由意见领袖组成的单向无标度网络。这些单纯结构组织在一起时，就表现出非常复杂的链接关系。由于测试数据集可以达到总体规模，因而选择这样的研究对象可以充分测试出不同抽样算法的质量特性，能够确保测试的无偏性和有效性，测试数据基本统计信息如表 2 所示。

表 2 测试数据集

抽样测试数据集			
指标	网易微博	优酷	蘑菇街
节点数	47,058,432	4,391,911	1,124,695
关系数	1,473,393,5[illegible]	74,602,487	20,249,370
入度指数	-2.432	-1.3[illegible]1	-1.571
出度指数	-0.087	-0.193	-0.169
度相关系数	0.147	0.274	0.029
平均路径长度	97.82	41.23	153.57
簇类系数	0.063	0.127	0.011

4.4 测试结果分析

根据前文所述，我们选择了 6 种有代表性的抽样方法，按照 5%的步距对蘑菇街、优酷与网易微博进行了抽样实测。由于蘑菇街、优酷与网易微博表现出相同的抽样特性，为了方便读者阅读，仅给

出网易微博的测试过程数据。6 组实验数据如表 3～表 8 所示。

测试数据的拟合优度采用 K-S 检验进行判断，检验命题包括 H0：测试数据所来自的分布符合总体的分布，H1：测试数据所来自的分布不符合总体的分布。如果取较宽松的显著性水平 α =0.1，根据表 8 的结果，RDS 与 MHRW 方法各项指标累积分布函数的 K-S 检验结果最低，H0 假设成立。可以认为在统计学意义上，样本与总体已具有一致性，且显著性水平较其他方法更有优势，这与前文 3.1 小节中提到的研究结果完全一致。

表 8 中，RDS 方法与 MHRW 方法对不同的指标似乎表现出不同的优势，例如 RDS 方法的入度拟合度较好，MHRW 方法度相关系数的拟合度较好。此处需要说明的是：这种比较优势并不具有规律性，如果保持一切测试条件不变，重新抽样，结果就可能改变，这既说明了抽样本身具有随机性，也说明 RDS 方法和 MHRW 方法本身的稳定性不足，还需进一步的优化与完善。

表 8 中统计结果差异的主要原因是不同抽样方法纠偏能力所致。BFS、UNI Sample 以及“点—边”抽样法不具有样本纠偏能力，属于有偏抽样，因而样本选取质量不高，抽样结果较差。RDS 和随机行走系列抽样方法属于不等概有放回抽样，其中 RDS 方法使用汉森—赫维茨（Hansen-Hurwitz ）估计法进行样本纠偏，随机行走系列方法中采取霍维茨—汤普森（Horvitz-Thompson）或汉森—赫维茨（Hansen-hurwitz ）方法进行样本纠偏，以上两种纠偏方法可以确保抽样按照渐进无偏的方式达到收敛。此处需要强调的是：使用 RDS、RW 系列方法获得无偏抽样的关键在于马尔科夫链燃烧预热以及收敛稳态的判断，尤其收敛稳态的判断尚无完备的理论支撑，多数情况下需靠研究者主观判定。

表 3 网易微博 5% Sample K-S 检验

抽样方法	抽样指标						
	入度	出度	相关系数	网络直径	平均路径长度	簇类系数	弱连通组件
BFS	0.2150	0.2473	0.1921	0.2168	0.2160	0.2431	0.3436
TIES	0.1938	0.2460	0.1673	0.1929	0.1759	0.1981	0.3512
UNI	0.2104	0.2816	0.2072	0.2293	0.1768	0.2402	0.3617
RDS	0.2161	0.2581	0.2118	0.1787	0.2161	0.2315	0.2936
MHRW	0.1744	0.2648	0.2481	0.1938	0.2127	0.2205	0.3077
RWuR	0.1967	0.2774	0.2417	0.2170	0.2759	0.1997	0.3479

表 4 网易微博 10% Sample K-S 检验

抽样方法	抽样指标						
	入度	出度	相关系数	网络直径	平均路径长度	簇类系数	弱连通组件
BFS	0.2032	0.2204	0.2377	0.2769	0.2041	0.1956	0.3205
TIES	0.1942	0.2121	0.1804	0.2364	0.1681	0.1910	0.3168
UNI	0.1991	0.1732	0.1962	0.1810	0.1689	0.2099	0.2310
RDS	0.1422	0.1513	0.1803	0.1961	0.1342	0.2148	0.2812
MHRW	0.1967	0.1781	0.1790	0.1427	0.1512	0.1964	0.2701
RWuR	0.1868	0.1695	0.1974	0.1601	0.1390	0.1925	0.2938

表 5 网易微博 15% Sample K-S 检验

抽样方法	抽样指标						
	入度	出度	相关系数	网络直径	平均路径长度	簇类系数	弱连通组件
BFS	0.1845	0.2076	0.2313	0.2575	0.2019	0.1832	0.2705
TIES	0.1857	0.2073	0.1946	0.1969	0.1769	0.1857	0.2717
UNI	0.1760	0.2168	0.1676	0.1820	0.1730	0.2688	0.3578

RDS	0.1396	0.1538	0.1946	0.1618	0.1707	0.1987	0.2274
MHRW	0.1711	0.1849	0.1893	0.1630	0.1682	0.1770	0.3148
RWuR	0.2082	0.1850	0.1078	0.1407	0.1159	0.1877	0.2945

表 6　网易微博　20% Sample　K-S 检验

抽样方法	抽样指标						
	入度	出度	相关系数	网络直径	平均路径长度	簇类系数	弱连通组件
BFS	0.1514	0.1597	0.1726	0.1640	0.2004	0.1107	0.2333
TIES	0.1014	0.2077	0.1828	0.1499	0.1586	0.1506	0.1828
UNI	0.1878	0.1947	0.1721	0.1809	0.1903	0.1691	0.2906
RDS	0.1594	0.1734	0.1828	0.1979	0.1620	0.1602	0.1739
MHRW	0.1648	0.1832	0.0870	0.1552	0.1597	0.1875	0.2830
RWuR	0.1797	0.1482	0.0944	0.1357	0.1015	0.1768	0.2827

表 7　网易微博　25% Sample K-S 检验

抽样方法	抽样指标						
	入度	出度	相关系数	网络直径	平均路径长度	簇类系数	弱连通组件
BFS	0.1408	0.1660	0.1711	0.1378	0.1966	0.1074	0.1898
TIES	0.1055	0.2115	0.1569	0.1140	0.1804	0.1087	0.1736
UNI	0.1706	0.1804	0.1[illegible]7	0.1794	0.1970	0.1636	0.1905
RDS	0.1306	0.1472	0.1[illegible]2	0.1533	0.1865	0.1466	0.1722
MHRW	0.1754	0.1793	0.1[illegible]6	0.1539	0.1283	0.1595	0.2081
RWuR	0.1533	0.1674	0.0927	0.1431	0.1114	0.1648	0.2164

表 8　网易微博　30% Sample K-S 检验

抽样方法	抽样指标						
	入度	出度	相关系数	网络直径	平均路径长度	簇类系数	弱连通组件
BFS	0.1311	0.1430	0.1701	0.1293	0.1875	0.1274	0.2098
TIES	0.1138	0.2071	0.1654	0.1249	0.1709	0.1063	0.1754
UNI	0.1870	0.1798	0.1503	0.1703	0.2108	0.1753	0.2205
RDS	0.0583	0.0429	0.1156	0.1039	0.1058	0.0967	0.0973
MHRW	0.0832	0.0401	0.0874	0.0875	0.0702	0.1168	0.0702
RWuR	0.1467	0.1374	0.1299	0.1401	0.1507	0.14[illegible]	0.18997

根据本课题组《在线社会化口碑营销群体发现》研究的需要，我们采用相同方法对样本与总体的口碑群体进行了比较分析，谱系分析如图 8 所示，该图反映的是口碑群体内部分层聚集的结构与聚集程度。从图中可以看出：随着采样率不断提高，抽样群体内部结构的分层关系不断发生变化，这说明抽样群体内部的结构不具有稳定性，即使在样本覆盖率达到 30%时，抽样数据与总体在分层结构上仍然差异显著。

在线口碑群体营销是将营销活动嵌入到在线社会网络的一类特殊营销行为，其关键是考察行动者社会关系网络（结构）与营销行为（功能）之间相互作用的关系。从领域问题（市场营销）的角度来看，图 8 中总体与抽样的差异说明不同品牌群体的同嗜性结构发生了改变，这种变化可以理解为是由群体内部互动关系变异所致。也就是说口碑群体内部信息的传播关系以及由此建立起的信任关系发生了根本的变化，从而使得口碑群体的集体意识发生了改变，呈现出不同的结构特性。因此我们可

以认为样本数据形成的口碑群体在角色、位置、从属关系与总体相比存在明显差异，不能满足领域问题研究的需要。

针对以上问题，我们认为社交媒体抽样可以理解为以下六种能力的综合：

（1）关系发现能力

抽样开始以后，样本数量不断增加，这意味着抽样子网中的节点不断增多。节点有两类，一类是无关系的孤立点，孤立点的增加不会对其他节点产生影响，抽样网络的连通性也不会发生变化；另一类是有关系的联系点，联系点的增加会对抽样网络的连通性产生影响。本文 3.1 节中提到的抽样方法均能通过抽样获得连通关系，逐步恢复社交媒体内部的社会网络，因而 K-S 检验效果随着样本规模的扩大变得越来越好。

（2 样本概率控制能力

随着节点规模的继续增加，节点度的统计规律性开始显现，也就是说样本序列可以按照节点度的数量特性进行分类。基于事后分层思想的 RDS 方法就开始表现出优势，在较小的样本规模下，样本与总体具有较高的一致性，尤其是入度与出度这两个指标，明显好于其他抽样方法，如表 4 所示。

（3）凝聚子群自组织能力

当节点再增加到一定规模时，样本节点间的关系开始发挥作用，这些节点组织在一起，形成或大或小的凝聚子群，可以明显反映这种变化的指标是变大的簇类系数。此时 BFS 抽样方法获取全连通子网的特性开始发挥作用，表现出较好的抽样特性，如表 6 所示。结合表 4 与表 5 的数据，我们注意到网络的簇类系数很小，这说明此时网络内部的组织关系尚未形成，因而我们可以判断 RDS 抽样方法的优势在于发现独立样本的统计规律性，而不是相干样本的主体间性关系。

（4）样本纠偏能力

随着样本的不断增加，局部凝聚子群内部的关系会逐步完善。这就要求抽样方法合理的选择子群内部的节点，其实质就是要求抽样过程中能够有意纠偏，防止高度节点的过度入样，准确的表达子群内部的横向与纵向层次的相干特性。MHRW 表现出较明显的优势，如表 7 所示。

（5）样本全局扩散能力

凝聚子群的规模扩张到一定程度以后，对节点规模的增长将不再敏感，这时面临的问题就是如何从一个庞大的凝聚子群跳出，获取其他抽样节点。很明显，那些只能获取相继样本的抽样方法跳出能力差，在原来的凝聚子群中继续抽样就是做无用功，反之能够全局获取样本的抽样方法则表现出较好的抽样特性，表 7 中 RwuR 表现出了优势。

（6）总体组织涌现能力

当抽样过程发展到这一阶段时，已经完成了两步工作：一是通过规模扩张克服了节点的构成复杂性；二是通过概率控制能力克服了局部凝聚子群内部的组织复杂性。当前所面临的关键任务就是通过选择不同凝聚子群之间的重叠关系来形成更大的子群乃至整体网络。从谱系图反映的信息来看，子群间重叠关系的选择有很多不足，从表 8 的数据也可以看出，抽样质量并没有因为节点规模的扩张而提高，由此我们可以判断在形成更高一级子网乃至整体网络时，抽样数据与总体表现出了较大的偏差。

现实存在的社交媒体内部总是由多种关系组织在一起，因而它在微观（节点）——中观（群体）——宏观（整体网）三个层次都存在不同的规律。微观是单纯数量特性所表现出的统计规律性、中观是节点凝聚子群所表现出的内聚性，而宏观则是不同凝聚子群再组织所表现出的涌现性。

相较而言，微观层次的规律容易捕获，只要样本数量达到一定规模即可；中观层次的规律尽管比较复杂，可以通过完善抽样方法的概率控制能力形成具有内聚性的凝聚子群，但是现有抽样方法对于宏观层次整体性（涌现）的刻画还有很多不足，因而当我们将抽样数据用于领域问题的研究时就会表现出诸多问题。本文在此首先提出以上问题，希望能够得到研究者的关注。

综合以上分析，我们可以看出已有的社交媒体抽样方法既各具特色，又有共同的缺陷，没有一种方法具备所有的能力，因而在未来还需要更深入的研究以改进现有方法的不足。本文选择 7 种不同的抽样方法进行了比较分析，结果如表 9 所示，供读者参考。

根据总体是否已知，可以将社交媒体抽样分为线上抽样与线下抽样。考虑到多数研究者缺少获得总体以及使用总体的大规模计算条件，本文重点综述了线上抽样方法。对于有条件获得总体的研究者，本文提到的线上抽样方法也可以用于线下抽样。

总体已知可以为抽样提供先验知识，但是这并不意味着现有方法在总体已知的条件下就能完全

有效的进行社交媒体抽样。首先，现有抽样方法的机制并不合理，已有方法都属于单阶段抽样。由于单阶段抽样方法在抽样过程中不能“变轨”，因此很难用适量的样本准确刻画社交媒体内部存在的多种复杂分布特征。其次，样本选取方法存在不足，现有方法只能选择相继关系作为入样单元，这样就导致大型马尔科夫链难以并行化、取样局部性陷入、马尔科夫链燃烧预热等诸多问题。因而不论是总体已知或是总体未知，现有抽样方法仍然存在较大改进与完善的空间，需要科研工作者进一步的深入研究。

表 9 抽样方法质量特性分析表

	评价指标	抽样方法						
		BFS	TIES	UNI	RDS	MHRW	DURW	RWuR
抽样能力评价	抽样类别	无概率 滚雪球	无概率 定额抽样	有概率 系统抽样	有概率 分层抽样	有概率 不等概抽样	有概率 不等概抽样	有概率 不等概抽样
	总体发现能力	不具备	不具备	具备	不具备	不具备	不具备	不具备
	抽样纠偏能力	差	差	差	较好	较好	较好	较好
	样本概率控制能力	不具备	不具备	具备	具备	具备	具备	具备
	样本结构	全连通	离散结构	离散结构	全连通	全连通	全连通	离散结构
	凝聚子群适应能力	较好	一般	一般	一般	较好	一般	一般
	有向图抽样能力	不具备	不具备	不具备	具备	具备	具备	具备
	样本推荐方式	相继 样本	相继 样本	相继 样本	相继 样本	相继 样本	全局 发现	全局 发现
	不回答样本影响	影响突出	影响较小	不受影响	影响突出	影响突出	不受影响	不受影响
	接近总体分布速度	慢	慢	慢	快	较快	与参数有关	与参数有关
抽样效率评价	种子选取影响	影响	不影响	不影响	不影响	不影响	不影响	不影响
	可并行化能力	可并行	可并行	可并行	可并行	可并行	可并行	可并行
	收敛性判断	不具备	不具备	不具备	比较困难	比较困难	很难	很难
	计算复杂性	低	低	低	中	中	中	高
	参数可调性	不需调节	不需调节	不需调节	不需调节	不需调节	调节困难	调节困难

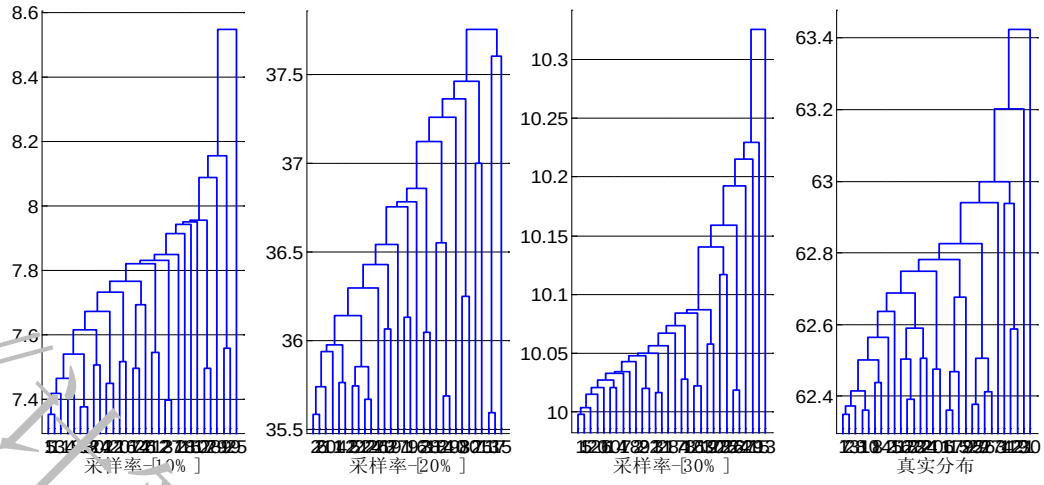

图 8 口碑群体谱系图

5 总结与展望

从社会学的角度来看，关系是一种客观实在，其表现形式为结构，因而社交媒体抽样研究的实质就是建立一种可以洞察关系结构的机制，使之能够反映各种抽样对象的社会存在。如果以此为标准，目前的研究忽视了社交化媒体内部的相干性，将丰富的主体间性关系解构为简单的统计指标，这样的抽样结果必然丢失了总体内部蕴含的大量信息，因此还需要从以下三个方面进行完善：

（1）抽样方法

现有的社交媒体抽样方法是从单一的宏观结构指标来理解其关系特性，忽视了社交媒体不同凝聚子群之间及其内部的相干性，属于简单还原论的方法。因此未来研究的首要任务就是要建立社交媒体内部结构关系的认知方法，将其由“黑盒子”变成“白盒子”，而后将统计学中有关样本容量估计、信度分析等理论合理的用于社交媒体这一新生事物，使抽样结果与总体更具一致性。

（2）抽样技术

抽样技术的本质是随机化探索方法，就目前的研究成果而言，随机行走是一种理论完备，适合于社交媒体使用的随机化方法，该方法在未来的研究中仍具有应用价值。从社交媒体抽样这一特定应用领域来看，还需对传统的随机行走进行完善，以模块化方法进行节点抽样，这样既能够提高算法对各种复杂结构的适应能力，还能提高收敛速度，加快达到平稳分布。

（3）抽样评价

结构与功能是一对相互依存又相互作用不可分离的基本哲学范畴，现有的研究仅采用拓扑特性指标（结构）进行抽样评价是必要而不充分的。还需要根据社交媒体自身的特点，研究动力学评价指标，以便更加全面的评价抽样质量，确保抽样结果具有完整的社会学含义。

致 谢：中国电信陕西分公司 IDC 提供了大规模计算环境，在此表示感谢。

参 考 文 献

[1] Kim W, Jeong O R, Lee S W. On social web sites. Information Systems, 2010, 35(2): 215-236

[2] Heidemann J, Klier M, Probst F. Online social networks: a survey of a global phenomenon. Computer Networks, 2012, 56(10): 3866-3878

[3] Siew Sin, Khalil MdNor, Ameen M Agaga. Factors affecting Malaysian young consumer’s online purchase intention in social media websites. Social and Behavioral Sciences, 2012, 40(8):323-333

[4] Kaplan A M, Haenlein M. Users of the world, unite the challenges and opportunities of social media. Business Horizons, 2010, 53(1): 59-68

[5] William G. Building stronger brands through online communities. Sloan Management Review, 2012, 41(3):189-203

[6] Wang Yuan-Zhuo, Jin Xiao-Long, Cheng Xue-Qi. Network big data: present and future. Chinese Journal of Computers. 2013 6(36):1125-1138 (in Chinese)
(王元卓, 靳小龙, 程学旗. 网络大数据:现状与展望. 计算机学报, 2013 6(36):1125-1138)

[7] Lynch C. Big data: how do your data grow?. Nature, 2008, 455(7209): 28-29

[8] McAfee A, Brynjolfsson E. Big data: the management revolution. Harvard Business Review, 2012, 90(10): 60-66

[9] Kossinets G. Effects of missing data in social networks. Social Networks, 2006, 28(3): 247-268

[10] Yangbo, Duan Wen-Qi, Chen Zhong. The Effect of sampling on multipe structural properties of complex networks. Journal of Shang Hai Jiao Tong Unviersity, 2007, 12(41): 1979-1984
(杨波, 段文奇, 陈忠. 抽样对复杂网络多重结构特征的影响. 上海交通大学学报, 2007, 12(41): 1979-1984)

[11] Yangbo, Chen Ying. Effect of sampling on growth mechanisms of complex networks. Journal of Shang Hai Jiao Tong University, 2013, 47(3) 479-484
(杨波, 陈颖. 抽样对复杂网络生长机制的影响[J]. 上海交通大学学报, 2013,47(3), 479-484

[12] Lohr Sharon L. Sampling: design and analysis. Second Edition. Boston: Brooks/Cole Cengage Learning, 2010

[13] Kish L. Survey Sampling. New York: Wiley & Sons, 1965

[14] Levy P S, Lemeshow S. Sampling of populations: methods and applications. New York: Wiley & Sons, 2013

[15] Mislove A, Marcon M, Gummadi K P, et al. Measurement and analysis of online social networks//Proceedings of the 7th ACM sigcomm conference on Internet measurement. San Diego, USA, 2007: 29-42

[16] Traud A L, Kelsic E D, Mucha P J, et al. Comparing community structure to characteristics in online collegiate social networks. SIAM Review, 2011, 53(3): 526-543

[17] Ferrara E. A large-scale community structure analysis in Facebook. EPJ Data Science, 2012, 1(1) 1-30

[18] Gjoka M, Kurant M, Butts C. Markopoulou A walking in Facebook: a case study of unbiased sampling of OSNs//Proceedings of the 29th conference on information communications. New York, USA 2010:2498-2506

[19] Viswanath B, Mislove A, Cha M, et al. On the evolution of user interaction in Facebook//Proceedings of the ACM workshop on online social networks. Florida, USA, 2009: 37-42

[20] Kwak, H, Park, H, Moon, S. What is Twitter, a social network or a news media?//Proceedings of the 19th international conference on World wide web. Carolina, USA, 2010:591-600

[21] Haewoon Kwak, Seungyeop Han, Yong Yeol Ahn, Sue Moon, Hawoong Jeong. Impact of snowball sampling ratios on network characteristics estimation: A case study of Cyworld. Daejeon: KAIST, Technical Report:CS-TR-2006-262, 2006

[22] Ye S, Lang J, Wu F. Crawling online social graphs// Proceedings of the 12th International Asia-Pacific Web. Busan, South Korea, 2010: 236-242

[23] Kurant M, Markopoulou A, Thiran P. Towards unbiased BFS sampling. Selected Areas in Communications, 2011, 29(9): 1799-1809

[24] Gjoka M, Kurant M, Butts C T. Walking in facebook: a case study of unbiased sampling of OSNs//Proceedings of the 29th International Conference on Computer Communications. San Diego, California, USA.2010: 1-9

[25] Chau D H, Pandit S, Wang S. Parallel crawling for online social networks//Proceedings of the 16th international conference on World Wide Web. Pennsylvania, USA, 2007: 1283-1284

[26] Krishnamurthy V, Faloutsos M, Chrobak M. Sampling large internet topologies for simulation purposes. Computer Networks, 2007, 51(15): 4284-4302

[27] Ahmed N, Neville J, Kompella RR. Network sampling via edge-based node selection with graph induction. West Lafayette: Purdue University, Computer Science Technical:11-016, 2011

[28] Heckathorn D. Respondent-driven sampling: a new approach to the study of hidden populations. Social Problems, 1997, 56(12):174-199

[29] Heckathorn D. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Social Problems, 2002, 49(1): 11-34

[30] Salganik M J, Heckathorn D. Sampling and estimation in hidden populations using respondent driven sampling. Sociological Methodology, 2004, 34(1): 193-240

[31] Salehi M, Rabiee H R, Nabavi N. Characterizing Twitter with Respondent-Driven Sampling//Proceedings of the 9th conference on autonomic and secure computing. Texas, USA, 2011: 1211-1217

[32] Rasti A H, Torkjazi M, Rejaie R. Respondent-driven sampling for characterizing unstructured overlays//Proceedings of the International Conference on Computer Communications Mini-Conference. Janeiro, Brazil, 2009: 2701-2705

[33] Gjoka M, Butts C T, Kurant M, et al. Multi-graph sampling of online social networks. Selected Areas in Communications, 2011, 29(9): 1893-1905

[34] Lu J, Li D. Sampling online social networks by random walk// Proceedings of the ACM international workshop on hot topics on interdisciplinary social networks research. Beijing, China, 2012: 33-40

[35] Noh J D, Rieger H. Random walks on complex networks. Physical Review Letters, 2004, 92(11): 701-712.

[36] Weiss G H, Rubin R J. Random walks: theory and selected applications. Advance Chemistry Physics, 1983, 52: 363-505

[37] Leskovec J, Faloutsos C. Sampling from large graphs//Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. Philadelphia, USA, 2006: 631-636

[38] Gilks W R, Richardson S, Spiegelhalter D J. Markov chain Monte Carlo in practice. London: Chapman & Hall, 1996

[39] Gamerman D, Lopes H F. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Second Edition.Boca Raton,FL: Chapman & Hall/CRC, 2006

[40] Krishnamurthy B, Gill P, Arlitt M. A few chirps about twitter//Proceedings of the first workshop on online social networks. Glasgow,Scotland UK, 2008: 19-24

[41] Wang T, Chen Y, Zhang Z, et al. Unbiased sampling in directed social graph. Sigcomm Computer Communication Review. 2010, 40(4): 401-402

[42] Jin L, Chen Y, Hui P, et al. Albatross sampling: robust and effective hybrid vertex sampling for social graphs//Proceedings of the 3th international workshop on mobi-arch. Beijing, China, 2011: 11-16

[43] Hubler C, Kriegel H P, Borgwardt K, et al. Metropolis algorithms for representative sub-graph sampling//Proceedings of the international conference on data mining. Pisa, Italia, 2008: 283-292

[44] Wang D, Li Z, Xie G.Towards unbiased sampling of online social networks//Proceedings of the international conference on communications. Kyoto, Japan, 2011: 1-5

[45] Ribeiro B, Wang P, Murai F, et al. Sampling directed graphs with random walks//Proceedings of the International Conference on Computer Communications. Orlando, FL USA, 2012: 1692-1700

[46] Avrachenkov K, Ribeiro B, Towsley D. Improving random walk estimation accuracy with uniform restarts//Proceedings of the algorithms and models for the web-graph. Berlin:Germany, 2010: 98-109

[47] Costa L F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: a survey of measurements. Advances in Physics, 2007, 56(1): 167-242

[48] Airoldi E M, Bai X, Carley K M. Network sampling and classification: an investigation of network model representations. Decision Support Systems, 2011, 51(3): 506-518

[49] Newman M E J. The structure and function of complex networks. SIAM Review, 2003, 45(2): 167-256

[50] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: structure and dynamics. Physics Reports, 2006, 424(4): 175-308

[51] Scott J. Social network analysis. Third Edition.Los Angeles: Sage publication, 2012

[52] Newman M E J, Park J. Why social networks are different from other types of networks. Physical Review E, 2003, 68(3): 036122-036131

[53] Barabási A L,Oltvai Z N. Network biology: understanding the cells functional organization. Nature, 2004, 5(2):101–113

[54] Amaral L A N, Ottino J M. Complex networks. European Physical Journal B, 2004, 74(1): 147–162

[55] Barrat A, Barthélemy M, Pastor-Satorras R. The architecture of complex weighted networks//Proceedings of the National Academy of Science. Las Vegas, USA, 2004:3747–3752

[56] Kaiser M, Hilgetag C C. Edge vulnerability in neural and metabolic networks. Biological Cybernetics, 2004, 90(5) :311–317

[57] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology. Computer Communication Review, 1999, 29(4):251–262

[58] Gastner M T, Newman M E J. The spatial structure of networks. The European Physical Journal B, 2006, 49(2):247-255

[59] Albert R, Albert I, Nakarado G L. Structural vulnerability of the north american power grid. Physical Review E, 2004, 69(2):103-118

[60] Boccara N. Modeling Complex Systems. Second Edition.New York:Springer,2010

[61] Strogatz S. Sync: The Emerging Science of Spontaneous Order. London:Penguin Books Limited, 2004

[62] Duda R O, Hart P E, Stork D G. Pattern Classification. New York: John Wiley & Sons, 2001

[63] Guimerà R, Díaz-Guilera A, Vega-Redondo F, Cabrales A, Arenas A , et al. Optimal network topologies for local search with congestion. Physical Review Letters, 2002: 89(24): 701-719

[64] Monasson R. Diffusion, localization and dispersion relations on “small-world” lattices. European Physical Journal B, 1999, 12(555), 34-42

[65] Zhou S, Mondragon R J. The rich-club phenomenon in the internet topology. Communications Letters, 2004, 8(3):180–182

[66] Latora V, Marchiori M. Is the Boston subway a small-world network?. Physical Review A, 2002, 314(1): 109-113

[67] Catanzaro M, Caldarelli G, Pietronero L. Assortative model for social networks. Physical Review E, 2004, 70(3): 037101-037109

[68] Madar N, Kalisky T, Cohen R. Immunization and epidemic dynamics in complex networks. The European Physical Journal B-condensed Matter and Complex Systems, 2004, 38(2): 269-276

[69] Tao Z, Zhong-qian F, Bing-hong W. Epidemic dynamics on complex networks. Nature, 2006, 16(5): 452-457.

[70] Newman M E J, Strogatz S H, Watts D J. Random graphs with arbitrary degree distributions and their applications. Physical Review E, 2001, 64(2): 026118

[71] Chung F, Lu L. The average distances in random graphs with given expected degrees. National Academy of Sciences, 2002, 99(25): 15879-15882

[72] Cohen R, Havlin S. Scale-free networks are ultra small. Physical Review Letters, 2003, 90(5): 058701

Cui Ying-an, born in 1975，Ph.D. His research interests include online social media sampling, big data analytics and social business.

Li Xue born in 1974，Postdoctor, Her research interests include social business, internet word of mouth and social media marketing.

Wang Zhi-Xiao born in 1977，Ph.D, His research interests include social computing, online social network dynamic.

Zhang De-Yun born in 1949，Ph.D, His research interests include network engineering, network architecture, complex system.

Background

The data of online social media is the relationship between the actors self-organization set, whose contains a multi-level social entity relationship. The traditional sampling methods are difficult to deal with its endogenous complexity, uncertainty and the emergence of resistance, so it has important research value and practical significance of online social media sampling method for social computing, which is an emerging field in recent years. In order to improve the performance and availability of online social media sampling technique, there have been a lot of methods proposed by researchers in recent years.

In this survey, we introduce the characteristics and shortcomings of the bfs method, the “point-edge” sampling method, the uniform sample of userids method, the respondent driven sampling method and the random walk method based on its thinking, probabilistic control, and the results of application with a comprehensive analysis and comparison. using the microblog data on various methods for the actual test based on the needs of the field, the research results show that the existing sampling methods at the micro-level (nodes) and meso-level (group) simply by node expansion and the probability of effective control to meet the node heterogeneity and organizational subgroups sampling requirements, but at the macro level, It is can not accurately portray the local cohesion to organize subgroups demonstrated emergent properties. At last, This paper pointed out that the future research directions of the online social media sampling.