软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(4):781−796 [doi: 10.13328/j.cnki.jos.004566]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

社会化媒体大数据多阶段整群抽样方法*

崔颖安[1,2], 李 雪[3], 王志晓[1], 张德运[1]

[1](西安交通大学 电信学院,陕西 西安 710049)

[2](西安理工大学 计算机科学与工程学院,陕西 西安 710048)

[3](陕西师范大学 国际商学院,陕西 西安 710062)

通讯作者: 崔颖安, E-mail: cuiyan@xaut.edu.cn

摘 要: 在线社会化媒体大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,因此,在线社会化媒体大数据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理论和应用价值.现有抽样方法存在大型马尔可夫链难以并行化、样本局部性陷入、马尔可夫链燃烧预热等问题.针对这些问题,提出了在线社会化媒体大数据整群多阶段抽样方法 OSM-MSCS.该方法首先进行整群分解,将总体分解成若干小型凝聚子群;而后,使用动态延迟拒绝方法对凝聚子群内部的关系抽样;最后,使用 Gibbs 方法完成不同凝聚子群之间相干关系的筛选,从而获得整个样本序列.实验结果表明,OSM-MSCS 方法能够有效地对各种结构特征的在线社会化媒体大数据进行抽样,从"个体地位-群体凝聚性-整体结构性"这 3 个层次进行综合评价,其抽样效果要明显好于 MHRW 和 BFS 这两种最主流的抽样方法.

关键词: 在线社会化媒体;大数据;马尔可夫蒙特卡洛方法;多阶段整群抽样

中图法分类号: TP311 **文献标识码**: A

中文引用格式: 崔颖安,李雪,王志晓,张德运.社会化媒体大数据多阶段整群抽样方法.软件学报,2014,25(4):781−796. http://www.jos.org.cn/1000-9825/4566.htm

英文引用格式: Cui YA, Li X, Wang ZX, Zhang DY. Sampling online social media big data based multi stage cluster method. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):781−796 (in Chinese). http://www.jos.org.cn/1000-9825/4566.htm

Sampling Online Social Media Big Data Based Multi Stage Cluster Method

CUI Ying-An[1,2], LI Xue[3], WANG Zhi-Xiao[1], ZHANG De-Yun[1]

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

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

[3](Internatiional Business School, Shaanxi Normal of University, Xi'an 710062, China)

Corresponding author: CUI Ying-An, E-mail: cuiyan@xaut.edu.cn

Abstract: The big data from online social media represents the relationship between the actors' self-organization. It contains multi-level social entity relationship. As an emerging field in recent years, online social media sampling method has important research value and practical significance in social computing. However, there are some problems in existing methods. For example, large Markov chain is difficult to parallelize, sampling is easy to be trapped in local, and there is concerns with Markov chain burn-in process. To address those issues, the paper presents a multi stage cluster sampling for online social media big data (OSM-MSCS). The proposed method first decomposes integral cluster into small cohesive subgroups, then uses delay rejection (DR) to sample typical online social relationship with parallel processing, and finally uses Gibbs sampling methods to choose interaction relationship in different cohesive subgroups to

* 基金项目: 教育部中央高校基金(13SZYB01); 陕西省社科联重大理论与现实问题研究项目(2013C124); 中国电信"社会化媒体大数据云服务商业模式的研究"项目(SN2012-YS-13709)

收稿时间: 2013-09-11; 修改时间: 2013-12-18; 定稿时间: 2014-01-27

obtain the random sequence. Experimental results show that OSM-MSCS is an effective method for online social media big data, and its sampling technique is better than BFS and MHRW.

Key words: online social media; big data; Markov chain Monte Carlo; multi stage cluster sampling

在线社会化媒体是用户自主创造内容、群体化意见分享、自组织建立社会网络的新型互联网应用,典型的在线社会化媒体包括博客、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.广大网民使用在线社会化媒体传递信息、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,有关在线社会化媒体的研究已成为社交商务、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[1,2].

尽管不同学科对在线社会化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在线社会化媒体数据.在线社会化媒体数据与传统行为科学数据相比有两个突出的特点:

- 一是数据规模大,其获取、存储、使用的成本很高,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.在线社会化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.

实证研究表明,不恰当的抽样会扭曲在线社会化媒体的结构关系及动力学特性.因此,在线社会化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,才能正确认识在线社会化媒体主体间性关系的本质特征,克服其内生的复杂性,为后续问题的研究打好基础,提高研究的准确性和可信性[3,4].

本文第 1 节对在线社会化媒体大数据抽样研究工作现状进行介绍.第 2 节给出多阶段整群抽样方法的总体框架以及各部分的详细说明.第 3 节对测试环境、测试数据集、测试评价方法进行简要介绍.第 4 节对多阶段整群抽样方法进行实际测试与分析.第 5 节对全文进行总结,并展望下一步的研究工作.

1 相关工作

综合国内外相关研究文献,目前常用的在线社会化媒体大数据抽样方法包括广度优先法、“点-边”构造法、用户均匀抽样法、同伴推动法以及随机行走系列方法,下面分别予以介绍和评述.

(1) 广度优先法(breadth-first-search,简称 BFS)

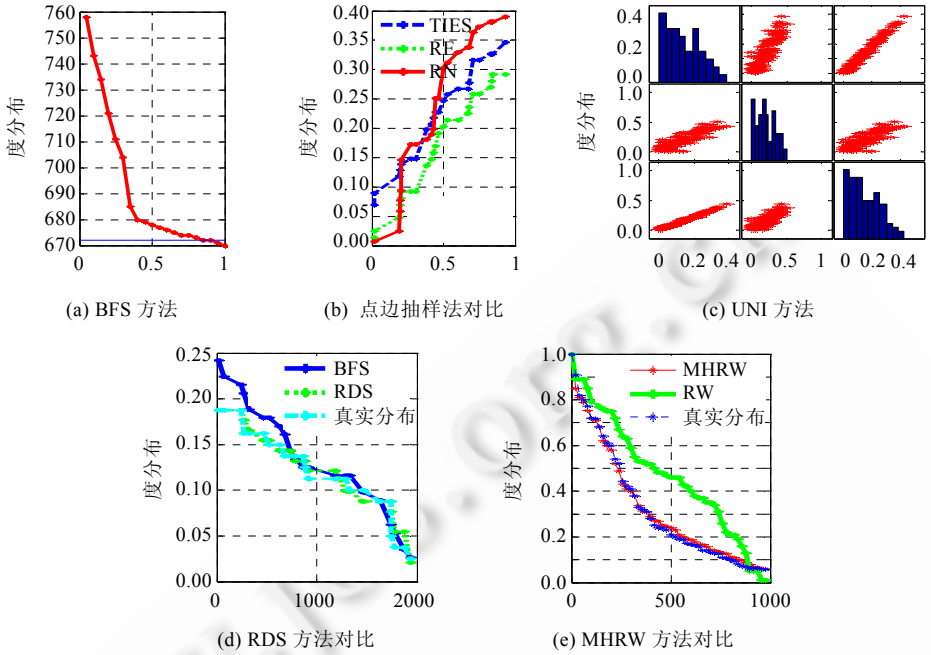
基本思想:从网络中选择初始节点放入先进先出队列,而后搜索与之相邻的所有节点,如果这些节点在队列中尚未出现,则节点入队并记录该节点的父节点;反之,则从队首取另一节点对其相邻节点作相同处理,直至队列为空.由于 BFS 方法可以彻底搜索整个网络,加之易于编程实现,因而在多个领域得到广泛应用.

典型案例:文献[5]使用 58 台服务器组成的集群对 Flickr,LiveJournal,Orkut,YouTube 的在线关系数据分别进行抓取,结果显示,抽样准确率在 85%~95%.文献[6,7]对 FaceBook 的在线关系数据进行了研究,结果显示,本次抽样数据与前人的研究结果具有一致性,都表现出较高的簇类系数和较小的网络直径.图 1(a)给出了 BFS 抽样质量与样本规模的关系,从该图中可以看出,BFS 方法的抽样质量高度依赖样本规模.

(2) “点-边”构造法(node-edge sampling,简称 NES)

基本思想:“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,随机点抽样法与随机边抽样法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(random pagerank node)方法和 RDN(random degree node)方法.典型的随机边抽样改进方法包括 RNE(random node edge)方法、HYVE(hybrid vertex/edge)方法和 TIES(totally-induced edge sampling)方法.

典型案例:“点-边”构造法中最具实用性、抽样效果相对最好的是 TIES 方法.文献[8]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 抽象效果对比如图 1(b)所示.



(a) BFS 方法 (b) 点边抽样法对比 (c) UNI 方法

(d) RDS 方法对比 (e) MHRW 方法对比

Fig.1 Comparsion of existing sampling method for social media

图 1 现有在线社会化媒体大数据抽样比较分析

(3) 用户均匀抽样法(uniform sample of UserIDs,简称 UNI Sample)

基本思想:UNI Sample 是一种针对在线社会化媒体的特异性抽样方法,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,形成与客观实际相一致的总体;而后,根据抽样质量的要求,等概率地使用 UserID 作为种子进行关系发现,直至达到抽样要求.

典型案例:文献[9]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,结果显示,UNI Sample 对于内部关系稀疏的在线社会化媒体具有一定的优势.图 1(c)给出了某真实博客系统 UserID 的分布特征,对于此类不规则分布的在线社会化媒体,UNI Sample 抽样效果并不理想.

(4) 同伴推动采样法(respondent driven sampling,简称 RDS)

基本思想:RDS 类似于广度优先法,但在 3 个方面作了大幅度的调整:一是利用马尔可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.

典型案例:文献[10]使用 RDS 方法对 Twitter 进行了 15×24 小时的数据采集,结果显示,RDS 抽样与总体最接近(如图 1(d)所示).文献[11]采用 RDS 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:RDS 在静态网络具有较好的抽样效果,但是对于动态网络,抽样偏差较大.

(5) 随机行走法(metropolis-hastings random walk)

基本思想:随机行走是一种经典的随机化方法,文献[4]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 方法将网络内的关系看成一棵分层的树,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(度分布具有无标度性)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,因为在社会化媒体内部的关系通常都是由多种分布特性混合而成,用一种抽样机制或者控制参数难以准确抽取复杂分布.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.Metropolis-Hastings 算法是一种典型的马尔可夫蒙特卡洛方法(Markov chain Monte Carlo),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[12,13].该方法能够确保生成的马尔可夫链具有细致平衡的特性,因而可以满足在线社会化媒体抽样的需要.

典型案例:文献[14]对 Twitter 中不同类型用户的信息传播特征进行了研究,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,可以无偏等概地获取样本.文献[15]以 FaceBook 为研究对象,对 BFS 方法、标准 RW 方法进行了比较研究,结果显示:MHRW 在抽样质量上明显优于 BFS 和标准 RW 方法(如图 1(e)所示),可以用于大型在线社会化媒体抽样.文献[16]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.

文献[17]使用模拟退火算法对标准 MHRW 进行了改进,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,但该方法的计算复杂性过大,抽样效率较低.文献[18]利用图的谱隙来修正抽样误差,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,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,抽样效果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.

以上 5 类抽样方法代表了在线社会化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,由于在线社会化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,其基本思想都是把在线社会化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,严重违背了在线社会化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愿望,UNI Sampl 方法随之出现.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,其总体探索能力确实要好得多.更为重要的是,一旦总体可知,在线社会化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地提高,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,只能用于在线社会化媒体的抽样研究,对于其他具有相似复杂相干关系的分子网络、生物网络则无能为力.

RDS 方法的出现对于在线社会化媒体大数据抽样研究具有重要的意义.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、收敛判断是提高在线社会化媒体大数据抽样质量的必要方法,为后续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,将在线社会化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,国外的研究者更加重视抽样方法对在线社会化网络内部复杂结构的探索能力,更加重视细节特性的抽样控制,使抽样样本与总体更具一致性.国内已有学者敏感地意识到抽样会对其相关研究质量产生非常大的影响,因此,有必要进一步深入研究在线社会化媒体大数据的抽样方法,为领域问题的研究提供高质量的数据.

小结:尽管以上抽样方法各不相同,但是它们都存在诸多共性问题.首先是抽样机制不合理,现有方法都属于单阶段抽样.由于单阶段抽样方法难以处理集成多种概率分布特征于一体的复杂网络,因此很难用适量的样本准确刻画在线社会化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.RDS 方法与随机行走

法只能选择相继关系作为入样单元,这样就导致大型马尔可夫链难以并行化、样本局部性陷入、马尔可夫链燃烧预热等诸多问题.BFS 方法、“点-边”构造法、用户均匀抽样法缺少样本的概率化控制能力,因而抽样过程中难以调控,对复杂概率分布的总体适应能力较差,抽样效果不佳;另外,抽样评价标准过于宽松,仅从宏观尺度将样本与总体进行比较远远不能达到领域问题研究的需要,例如抽样数据中凝聚子群的结构是否与总体相似?关键行动者的地位是否一致等.如果不能确保这些指标的相似,抽样数据就会成为最大的误差源.

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,因而在线社会化媒体大数据抽样研究的实质就是建立一种可以洞察结构的机制,使之能够反映抽样对象内在的社会存在.如果以此为标准,单阶段抽样法忽视了在线社会化媒体内部的相干性,将丰富的主体间性关系还原为点的集合,破坏了在线社会化媒体不同凝聚子群之间及其内部的相干性.因此,我们研究的关键就是要建立能够洞察在线社会化媒体内部结构特征的机制,使其能够有效地指导抽样选择.

2 多阶段整群抽样方法

2.1 在线社会化媒体结构特征分析

已有研究表明:经过长期运营的在线社会化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,散乱地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,其内部是由多个具有局部中心化特征且包含大量嵌套关系的凝聚子群组合而成,因而“去中心性”是在线社会化媒体内部结构的重要特征.

在不同凝聚子群内部,用户的交互特性也比较复杂,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,联系具有互惠性).

结构的去中心性与联系的多样性使得在线社会化媒体表现出许多看似矛盾的规律,例如社区结构的整体无序与局部有序、信息传播路径的宏观稳定与微观混沌等,因而,如何准确地把握在线社会化媒体内部这些既统一又对立的规律,就成为理解在线社会化媒体特征的关键所在.

2.2 总体框架

抽样理论中,将若干有联系的基本单元所组成的相干样本集合称为群.抽样时,先将总体按照某种标准划分为不同的子群,而后以每个子群为抽样单元,以随机的原则从中抽取若干子群进行研究,这种方法就是整群抽样.整群抽样中,如果群规模较大,则没有必要研究初级抽样单元(PSU)中的所有二级抽样单元(SSU),而是从每个被抽中的 PSU 中对 SSU 再抽取子样本,这就是两阶段整群抽样.以此类推,若抽样单元不断细分,则称此方法为多阶段整群抽样(multi stage cluster sampling)[19,20].

以在线社会化媒体的结构特征作为切入点,可以采用多阶段整群抽样方法(online social media-multi stage cluster sampling,简称 OSM-MSCS)对在线社会化媒体大数据进行抽样分析.该方法的总体框架依次由数据感知、整群分解、子群抽样以及抽样评价这 4 部分构成,如图 2 所示.

数据感知子系统的主要作用是完成各类社会化媒体数据的采集与封装,采集规则通过 Meta-Data 配置在规则引擎中,而后依据页面元素的点击流程自动进行数据抓取与关系矩阵的封装.

整群分解子系统的主要作用是将社会化媒体大数据按照“群内关系紧密,群间关系稀疏”的原则划分成不同的凝聚子群,而后分析每一凝聚子群内部的分布特征,将在线社会化媒体由“黑盒子”变成“白盒子”,为后续样本的选取提供更加合理的指导.

子群抽样子系统由群内并行抽样和群间关系整合两部分构成.群内并行抽样采用 Delayed rejection(DR)方法,以并行化方式对每一个首代父图(包括该父图内的嵌套子群)进行样本抽选.这样既能提高样本入样率,还能

避免马尔可夫链的燃烧预热(Markov burn in).在完成群内并行抽样后,采用 Gibbs 抽样方法对群间关系进行专门处理,选取适当的抽取策略处理子群间的相干关系,从而连通整个网络.

抽样评价子系统的主要作用是采用简单估计法判定抽样的准确性,确保抽样数据能够满足在领域问题研究中的需要.主要指标包括常用的整体网统计指标、凝聚子群统计指标、关键行动者地位统计指标.

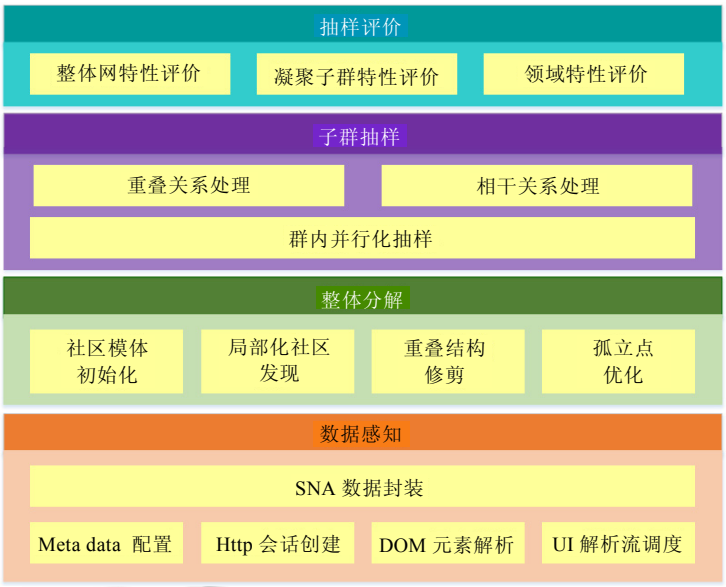


Fig.2 General framework of multi stage cluster sampling

图 2 整群多阶段抽样方法总体框架

2.3 详细介绍

(1) 整群分解

本文使用 SLPA(speaker-listener label propagation algorithm)社区发现方法进行整群分解[21,22],该方法能够处理各种网络的各种复杂特性,包括有向关系、重叠社区、嵌套社区以及非社区节点等,具有较好的社区发现能力和应用性能,可以满足行动者规模在百万级的大型在线社会化媒体整群分解的需要.考虑到不同社区发现方法对整群分解的影响,本文还采用 CORPA 方法(community overlap propagation algorithm)[23]与 SLPA 方法进行比较研究,以验证多阶段整群抽样方法的有效性,整群分解数据结构见表 1.

Table 1 Data structure of the integral cluster decomposition

表 1 整群分解数据结构

社区编码	父系成员	子系成员	社区成员
$Commu\text{-}ID_l$	$\{C\text{-}ID_b\}$	$\{C\text{-}ID_m;\ldots;C\text{-}ID_t\}$	$\{user\text{-}ID_{1a};user\text{-}ID_{1b};user\text{-}ID_{1c};\ldots;user\text{-}ID_{1m}\}$
$Commu\text{-}ID_m$	$\{C\text{-}ID_c\}$	$\{C\text{-}ID_b;\ldots;C\text{-}ID_f\}$	$\{user\text{-}ID_{2xa};user\text{-}ID_{2b};user\text{-}ID_{2c};\ldots;user\text{-}ID_{2p}\}$
…	{…;…}	{…;…}	
$Commu\text{-}ID_t$	$\{C\text{-}ID_g\}$	$\{C\text{-}ID_b;\ldots;C\text{-}ID_y\}$	$\{user\text{-}ID_{pa};user\text{-}ID_{pb};user\text{-}ID_{pc};\ldots;user\text{-}ID_{px}\}$

由于在线社会化媒体的数据规模非常庞大,我们可以依据在线社会化媒体群内联系紧密、群间联系稀疏的客观特性,将整体网分解成不同的子网,即,根据在线社会化媒体客观的结构特性,将大型马尔可夫链分解成小规模的子链.由于不同凝聚子群之间的关系非常稀疏(见表 2),因而可以将每一个首代父图(包括其嵌套子图)作为 PSU(群间关系不列入抽样对象),这样就可以将在线社会化媒体大数据抽样转化为不相干凝聚子群的并行抽样.

Table 2 Proportional distribution of relations in different coffesive groups

表 2 群间关系分布比例

抽样方法	网易微博			蘑菇街			优酷		
	群内关系	群间关系	孤立点	群内关系	群间关系	孤立点	群内关系	群间关系	孤立点
SLPA (%)	87.17	1.62	11.21	91.19	1.32	7.49	76.83	0.8	22.37
CORPA (%)	85.65	3.12	11.21	90.87	1.64	7.49	74.28	3.35	22.37

(2) 子群抽样

本文使用 Delayed rejection(DR)方法进行子群抽样,该算法的主要思想是:在进行 MHRW 抽样时,如果候选点被拒绝,算法不是立即放弃该样本,而是提出一个二级推荐分布,计算由此二级推荐分布取得候选点的接受概率,以此为依据来确定候选样本的取舍.DR 方法可以设定多级推荐分布对候选点进行取舍,高级别的推荐分布根据以前候选点被接受或拒绝的情况进行设定,这样既能提高候选样本的入样率,还能对推荐分布进行局部调整,以适应凝聚子群内部关系的复杂分布[24,25].

DR 算法的具体工作过程为:令 Markov 链当前所在状态为 $X^{(t)}=x$,在标准 M-H 算法中,由推荐分布 $q_1(X^{(t)},\cdot)$ 抽取候选点 Y_1,其接受概率为公式(1),此时,若候选点 Y_1 被拒绝,则 Markov 链保留原状态 $X_{n+1}=x$.

$$A(X^{(t)},Y_1)=1\wedge\left\{\frac{\pi(Y_1)q_1(Y_1,X^{(t)})}{\pi(X^{(t)})q_1(X^{(t)},Y_1)}\right\}=1\wedge\frac{N_1}{D_1} \tag{1}$$

其中,"$\wedge$"表示取二者中较小的值.

在 DR 算法中,算法设定一个二级推荐分布,该推荐分布不但取决于 Markov 链的当前状态,也与第 1 级推荐分布的接受和拒绝情况有关,令该二级推荐分布为 $q_2(X,Y_1,\cdot)$,在其中抽取第 2 级候选点 Y_2,此时,相应的样本提议概率为公式(2):

$$A_2(X^{(t)},Y_1,Y_2)=1\wedge\left\{\frac{\pi(Y_2)q_1(Y_2,Y_1)q_2(Y_2,Y_1,X^{(t)})[1-A_1(Y_2,Y_1)]}{\pi(X^{(t)})q_1(X^{(t)},Y_1)q_2(X^{(t)},Y_1,Y_2)[1-A_1(X^{(t)},Y_1)]}\right\}=1\wedge\frac{N_2}{D_2} \tag{2}$$

以此类推,若用 $q_i(\cdot|\cdot)$表示第 i 级移动的建议分布,则第 i 级候选点接受概率为公式(3):

$$A_i(X^{(t)},Y_1,...,Y_i)=1\wedge\left\{\begin{array}{l}\dfrac{\pi(Y_i)q_1(Y_i,Y_{i-1})q_2(Y_i,Y_{i-1},Y_{i-2})...q_i(Y_i,Y_{i-1},...,X^{(t)})}{\pi(X^{(t)})q_1(X^{(t)},Y_1)q_2(X^{(t)},Y_1,Y_2)...q_i(X^{(t)},Y_1,...,Y_i)}\times\\ \dfrac{[1-A_1(Y_i,Y_{i-1})][1-A_2(Y_i,Y_{i-1},Y_{i-2})]...[1-A_{i-1}(Y_i,Y_{i-1},...,Y_1)]}{[1-A_1(X^{(t)},Y_1)][1-A_2(X^{(t)},Y_1,Y_2)]...[1-A_{i-1}(X^{(t)},Y_1,...,Y_{i-1})]}\end{array}\right\} \tag{3}$$

在将 DR 方法用于在线社会化媒体大数据抽样时,首先需要解决的问题就是如何获得总体的概率密度 $\pi(X^{(t)})$和精确的提议分布函数 $q_1(Y_1,X^{(t)})$.通常情况下,准确拟合 $\pi(X^{(t)})$和 $q_1(Y_1,X^{(t)})$不具有可行性,即使本文通过整群分解大大降低了网络的结构复杂性,也很难以较低的成本拟合不同子群的复杂分布,因而需要改进 DR 方法以适应各种复杂分布.

就在线社会化媒体这一特定的抽样对象而言,所有的候选点都是客观存在的真实节点,不是随机生成的虚拟数据.以此特征为切入点,我们对标准 DR 方法作了两方面的改进:首先,在 DR 方法的第 1 层采用 Hansen-Hurwitz 进行不等概入样控制,调整 MHRW 转移核的计算方法,减少高度节点的过度入样,使得更多低度候选节点能成为样本点;其次,在 DR 方法的第 2 层利用已有的入样单元估计抽样序列的协方差,而后用协方差不断动态调整 DR 的转移核,从而影响后续样本点的选择,使得局部抽样更加细腻、准确.因而,我们将此改进方法称为 DDR(dynamic delay rejected).

DDR 方法的关键是在获得一定规模的样本点(例如 100 的整数倍)以后,不断更新协方差矩阵,协方差矩阵计算方法如公式(4)所示:

$$C_n=s_dCov(X_0,X_1,...,X_k)+s_d\varepsilon I_d \tag{4}$$

其中,ε是确保 C_n 为奇异矩阵的较小正数;s_d 是控制延迟拒绝的比例因子,取值范围为[0,1];I_d 为 d 维单位矩阵,本文 d 为 2;参与协方差运算的变量分别是度分布系数 X_0 和簇类分布系数 X_1.对于 $Cov(X_0,X_1,\ldots,X_k)$,可以表示为公

式(5):

$$Cov(X_0,X_1,...,X_k)=\frac{1}{k}\left[\sum_{i=0}^{k}X_iX_i^T-(k+1)\overline{X_k}\,\overline{X_k^T}\right] \tag{5}$$

其中, $X_k=\frac{\sum_{i=1}^{k}X_i}{(k+1)}$.将其代入公式(5)即可得到下一时刻的协方差递归公式.

OSM-MSCS 使用 DDR 方法进行子群抽样有两个明显的优点:一是快,二是准.

所谓快是相对 MHRW 方法而言的.将 MHRW 方法用于在线社会化媒体抽样时,由于总体分布未知,因而不得不将 MHRW 方法简化为公式(6),很显然,这样将舍弃大量有效节点,导致抽样效率低下;DDR 方法使用 Hansen-Hurwitz 方法进行第 1 层抽样控制,这样,$\pi(X^{(t)})$和 $q(X^{(t)},Y)$就可以分别按照节点在网络中的分布特性与连通特性进行概率估计,减少了高度节点的过度入样,明显地提高了有效样本的入样率.

$$P_{v,w}^{MH}=\begin{cases}\dfrac{1}{k_v}\min\left(1,\dfrac{k_v}{k_w}\right), & \text{if } w \text{ is a neighbor of } v\\ 1-\sum\limits_{y\neq v}P_{v,y}^{MH}, & \text{if } w=v\end{cases} \tag{6}$$

所谓准是相对 DR 方法而言的.DDR 采用协方差矩阵进行二级提议分布的控制.由于协方差矩阵是根据入样单元实时动态生成,因而样本方差与总体(局部区域)的差异较小,这样就能较明显地提高局部抽样质量.另外,DDR 方法还能有效获取子群内部低度行动者之间的相干关系,这就使得子群内部的连通性与真实总体更接近.总的来看,DDR 方法在一级提议分布实现了“精挑”,在二级提议分布实现了“细选”,这样就能更好地满足复杂嵌套结构的子群抽样需求,使得样本与总体具有更高的一致性.

(3) 群间重组

本文使用 Gibbs 方法对子群间的关系进行抽样,该方法属于马尔可夫更新机制的范畴,它采用分而治之的思想(divide and conquer),即,推断一组参数时,假定其他参数固定且已知.令 X_j 代表某种随机变量或同组的几个随机变量,第 j 组变量的边际分布为 $P(X_j)$.根据 Gibbs 抽样方法,从 $P(X_j)$中抽样的步骤如下[26]:

第 1 步:给定任意初始向量 $X_j^{(0)}=(X_j^{(0)},X_j^{(1)},...,X_j^{(k)})$;

第 2 步:从条件分布 $P(\theta_1\mid X_2^{(0)},...,X_k^{(0)})$ 中抽取样本 $X_1^{(1)}$,从 $P(X_2\mid X_1^{(1)},X_3^{(0)},...,X_k^{(0)})$ 中抽取样本 $X_2^{(1)}$;

第 3 步:从 $P(X_k\mid X_1^{(1)},X_2^{(1)},...,X_{j-1}^{(1)},X_{j+1}^{(0)},...,X_k^{(0)})$ 中抽取样本 $X_j^{(1)}$;

第 4 步:从 $P(X_j\mid X_1^{(1)},...,X_{j-1}^{(1)},...,X_{k-1}^{(1)})$ 中抽取样本 $X_k^{(1)}$.

由以上过程即完成 $X^{(0)}$到 $X^{(1)}=(X_1^{(1)},...,X_{j-1}^{(1)},...,X_k^{(1)})$ 的 1 次转移,$X^{(1)}$为马尔可夫链的 1 次实现值;

第 5 步:返回第 2 步,经过 t 次迭代,得到 $X^{(t)}=(X_1^{(t)},...,X_k^{(t)})$, 并最终得到马尔可夫链的实现值.从而,马尔可夫链转移概率函数为公式(7):

$$P(X,X^*)=P(X_1\mid X_2,...,X_k)P(X_2\mid X_1^*,...,X_k)...P(X_k\mid X_1^*,...,X_{k-1}^*) \tag{7}$$

由不同的 $X_{(0)}$出发,当 $t\to\infty$,在遍历条件下,可以认为各时刻 $X^{(t)}$的边际分布为平稳分布,此时它收敛并可以被看作是样本的仿真观测点.根据 Markov 链的遍历性,所有基于 MCMC 的统计推断都是在假定 Markov 链已经收敛的条件下进行的.所以,MCMC 的收敛性诊断对于用模拟的样本进行模型推断和估计极为重要.本文采用 Geweke-Diagnostic 方法对 Gibbs 抽样序列进行收敛性诊断[27].

OSM-MSCS 选用 Gibbs 方法用于群间关系重组,选择 Gibbs 方法是因为在完成首代父图内部行动者的抽样以后,剩余的抽样对象只是首代父图之间的相干关系.由于首代父图之间的相干关系非常少且全部可知(通过整群分解可以获得),这样就可以直接选用某种筛选策略(例如等概入样、结构洞入样、hop 行动者入样等)选取若干首代父图之间的相干关系,而后通过 Gibbs 抽样调整其概率转移矩阵获得平稳分布,这样就使得每一个后继节点都能成为入样节点,从而提高抽样的整体效率.

3 实验测试

3.1 测试环境

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 4 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在线社会化媒体大数据的抽样分析,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 3 所示.

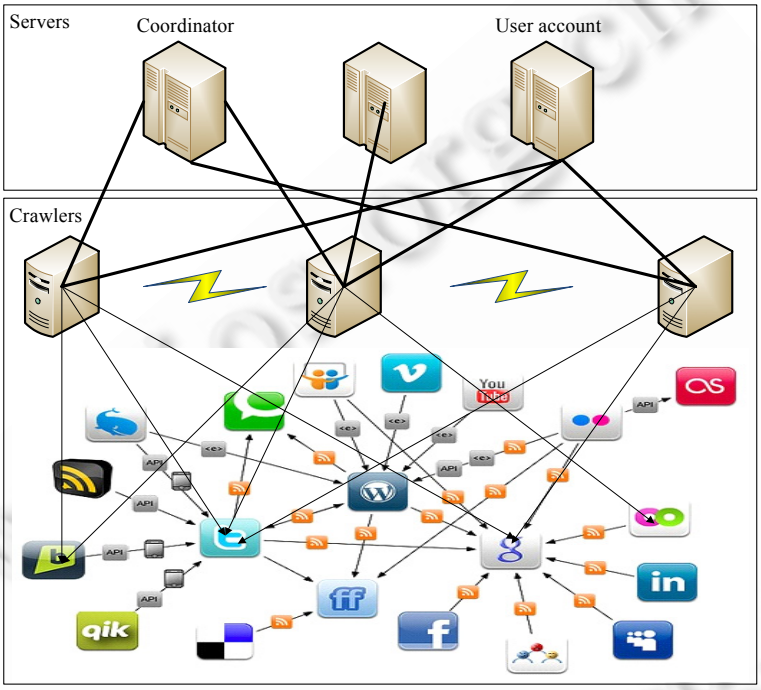


Fig.3 Test environmenting

图 3 测试环境

3.2 测试数据集

我们选择网易微博(http://t.163.com/)、蘑菇街(www.mogujie.com)、优酷(www.youku.com)作为测试对象(见表 3).选择以上社会化媒体的主要原因是: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.

Table 3 Test data-set of online social meida

表 3 在线社会化媒体测试数据集

统计指标	测试对象		
	网易微博	蘑菇街	优酷
Nodes	47 058 432	1 124 695	4 391 911
Edges	1 473 399 506	20 249 370	74 662 487
In-Degree law exponent	−2.432	−1.491	−1.371
Out-Degree law exponent	−0.087	−0.169	−0.193
Degree correlation	0.147	0.274	0.089
Avg path length	97.82	41.23	153.67
Cluster cofficient	0.063	0.127	0.011

在测试之前,我们对以上数据进行了较为全面的社会网络分析,从其构成上看:有由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,表现出非常复杂的链接关系.因而,选择这样的研究对象可以充分测试出多阶段整群抽样方法的质量特性.

3.3 测试检验方法

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 $F_x(x)$,后者检验两个样本是否服从同一概率分布.其工作原理如下:

设 $X=(X_{(1)},X_{(2)},\ldots,X_{(n)})$是样本 $X=(X_1,X_2,\ldots,X_n)$的有序统计量,因 $F_x(x)$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,假设检验结果为公式(8):

$$F_n(x_{(t)})=\frac{i}{N_0},i=1,...,N_0$$
$$D_{N_0}=\max\left\{\left|F(x_{(i)})-\frac{i-1}{N_0}\right|,\left|F(x_{(i)})-\frac{i}{N_0}\right|\right\} \tag{8}$$

式(8)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,这个特点非常适合在线社会化媒体大数据抽样检验的需要,因此,K-S 检验已经成为在线社会化媒体抽样研究中最主流的检验方法.

4 实验结果与分析

4.1 整群分解分析

(1) 整群分解效果比较分析

本文使用 SLAP 与 CORPA 方法对实验数据集进行了整群分解,谱系图分析结果如图 4 所示.从图中的数据可以明显看出:不同的社区分解方法对同一对象分解效果差异很大,其内部子群分解特征有很大的不同.

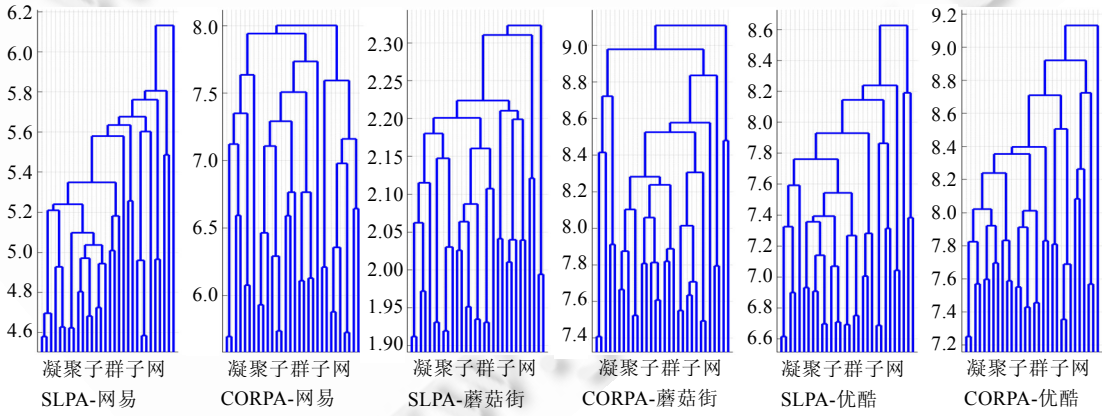


Fig.4 Comparative analysis of cluster decomposition dendrogram

图 4 整群分解谱系图比较分析

为了更加直观地说明不同社区分解方法对整群分解的影响,我们采用相对熵(K-L divergence)对比不同分解方法的差异性,统计结果见表 4.

Table 4 Comparative analysis of integral cluster decomposition

表 4 整群分解比较分析

相对熵	网易微博	蘑菇街	优酷
$D_{KL}(P_{\mathrm{SLPA}}\|\|P_{\mathrm{CORPA}})$	$6.205.10^{-3}$	$3.769.10^{-3}$	$8.107.10^{-3}$
$D_{KL}(P_{\mathrm{CORPA}}\|\|P_{\mathrm{SLPA}})$	$8.122.10^{-3}$	$5.207.10^{-3}$	$6.315.10^{-3}$

结合谱系图的直接观察与相对熵的定量分析我们可以看出:不同社区分解方法不论是在子群分解规模还是子群间的嵌套关系以及由此产生的子群聚类特性,都存在显著的差异.

(2) 子群结构特征比较分析

群体动力学认为:在每一个凝聚子群内部,一定存在着组织之轴,轴心可能是单个的点或者点集,轴距是一系列级联的延伸关系,这些复杂关系形成“场”.群体的行动受到场关系的影响和约束,即,群体结构决定群体功能.为了观察不同子群内部的结构特征,我们选择度中心势、接近中心势、中介中心势这 3 个指标对子群的结构特性进行比较研究.

① 度中心势

度中心势是衡量凝聚子群内不同行动者之间向中心聚集的整体性指标,该指标越大,表明群体与中心的关系越紧密,网络整合能力越强;反之,则表示群体凝聚力不高.星型网络取到该指标的极大值 1,环型网络取到该指标的极小值 0.该指标的计算公式为公式(9):

$$C_D=\frac{\sum_{i=1}^{g}[C_D(n^*)-C_D(n_i)]}{[(g-1)(g-2)]} \tag{9}$$

其中,$C_D(n^*)$为最大度中心势.

② 接近中心势

接近中心势是衡量凝聚子群内的不同行动者之间关系亲疏的整体性指标,该指标越大,表明群体内存在大量有直接关系的行动者;反之,则表示群体内不同行动者的关系比较疏远,缺少直接关系.星型网络取到该指标的极大值 1,完全网络取到该指标的极小值 0.该指标的计算公式为公式(10):

$$C_c=\frac{\sum_{i=1}^{g}[C'_C(n^*)-C'_C(n_i)]}{[(g-2)(g-1)]/(2g-3)} \tag{10}$$

其中,$C_C(n^*)$为最大接近中心势.

③ 中介中心势

中介中心势是衡量凝聚子群内部关系中介程度的整体性指标,该指标越大,表明凝聚子群内部结构洞越多,即,不同行动者只有通过其他行动者才能发生联系;反之,则表示群体内部结构洞较少,不同群体的相干关系少,网络结构相对单纯.星型网络取到该指标的极大值 1,当所有行动者有的中介度相等时,则取得该值的最小值 0.该指标的计算公式为公式(11):

$$C_B=\frac{2\sum_{i=1}^{g}[C_B(n^*)-C_B(n_i)]}{[(g-1)^2(g-2)]} \tag{11}$$

其中,$C_B(n^*)$为最大中介中心势.

图 5 反映的是不同测试对象凝聚子群关系特性的比较分析,从结果来看:尽管采取的整群分解方法不同,但是仍然可以表现出相似的规律.3 个测试对象中,网易微博的度中心势最高.这是因为网易微博中存在众多不同领域的意见领袖,大量用户都关注了自己关心的意见领袖(粉丝过千万的意见领袖被关注率在 73.27%,粉丝过百万的意见领袖被关注率在 50.33%),形成了类似星型网络的高度聚集结构.相比较而言,蘑菇街的度中心势最低,蘑菇街是一个专门面向青年女性的社会化营销平台,该系统根据青年女性的时尚需求设立了不同的频道(包包、配饰、鞋子等),由于频道分类很细致,因而难以形成具有极高人气的意见领袖,其度中心势最低.但是蘑菇街提供了诸多类似“晒货”之类促进社交的服务,极大地激发了不同行动者之间的联系,因而其接近中心势达到了所有研究对象中的最大值.也就是说,蘑菇街系统中,行动者之间的直接关系丰富,间接关系很少,形成了大量类

似于小世界网络的凝聚子群.优酷在所有研究对象中结构最单纯,是典型的随机网络.该网络内包括了大量既无入度也无出度的孤立节点,占比高达 67.24%,凝聚子群的规模也比较小,仅有 13.83%的用户建立了明确的好友关系,是所有测试对象中的最低值,而且凝聚子群之间的相干关系也很少,占比仅有 1.83/%.而网易微博却达到了 8.79%(低度行动者之间的相干关系占比 59.31%,高度行动者之间的相干关系占比 5.07%,低度行动者与高度行动者相干关系占比 35.62%),取得所有测试对象的最高值.由此我们可以看出:3 个测试对象中,内部结构最为复杂的是网易微博,其内部网络不仅规模庞大,凝聚子群数量多,而且凝聚子群之间存在大量不同类型的相干关系,这给抽样带来不少困难.

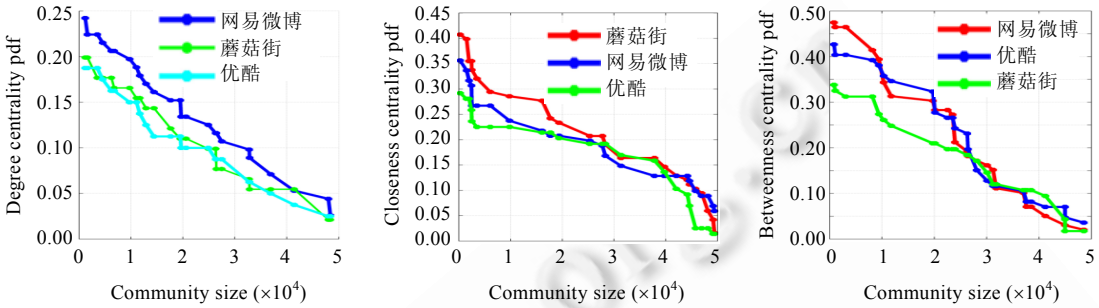


Fig.5 Structural features analysis of cohesive group

图 5 凝聚子群结构特性分析

尽管不同的社区分解算法的子群分解结果有很大差异,但是子群内部的结构特征仍然具有较高的相似性(K-S 检验置信度>=92%).由此我们可以认为:即使选取不同的整群分解方法,凝聚子群内部的结构特性在宏观尺度上仍然能够保持其客观性,并不会因为分解方法不同而改变其内部的结构特征,这一特征确保了 OSM-MSSC 方法具有较高的可行性和稳定性.

4.2 子群抽样分析

通过整群分解,OSM-MSCS 可以根据首代父图内行动者的概率分布特征,按照不等概、有放回的方式进行群内样本选取.不等概控制方法采用汉森-赫维茨估计法,依据不同凝聚子群内关系的古典概率分布特征进行入样控制,确保样本与总体具有相似的子群分布特征.OSM-MSCS 也具有较高的抽样效率,与其他方法抽样覆盖率与抽样质量的对比如图 6 所示,由图中曲线的信息数据可以看出,OSM-MSCS 方法的点-边覆盖能力优于其他抽样方法.

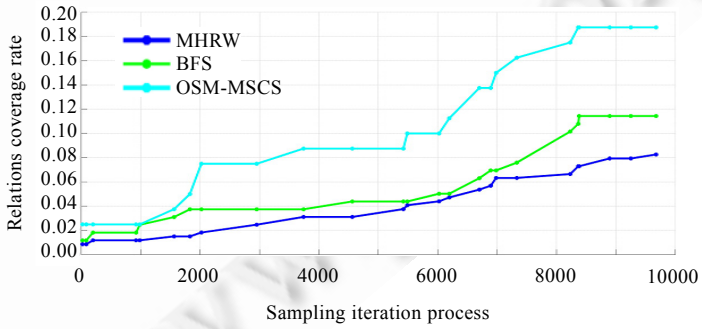


Fig.6 Comparative analysis of relations coverage rate

图 6 抽样覆盖率比较分析

OSM-MSCS 与其他方法相比具有较高的入样率,其根本原因在于其抽样机制的改变.MHRW 和 BFS 方法很难并行化,因为行动者之间的相干关系全部纠缠在一起,缺少有效解耦,这就导致并行化时难以克服样本选取

冲突.OSM-MSCS 利用整群分解实现了相干关系的有效解耦,因而可以对每一个首代父图并行抽样,从而提高了整体的抽样效率.另外,OSM-MSCS 在子群抽样时不需要马尔可夫燃烧预热,这也提高了抽样的效率[28].

4.3 群间重组分析

通过并行化子群抽样完成首代父图内行动者的筛选以后,使用 Gibbs 方法进行首代父图间相干关系的抽样.由于 Gibbs 抽样需要构造初始样本序列,为了确保样本与总体连通能力的相似,我们选择群间最活跃的 hop 节点构造 Gibbs 抽样初始序列.经过实际抽样测试(30%的取样率),网络半径(radius)与直径(diameter)与总体相差不超过 6.8%,表明由抽样关系形成的网络中心向外延伸以及其整体规模尺度与总体具有相似性.抽样平均路径长度(avg path length)与总体相差不超过 3%,说明抽样数据能够按照总体的特性将不同子群有效地连接起来,确保整体网络的连通性.网间连通性的抽样效果见表 5.

Table 5 Comparative analysis of connectivity

表 5 连通性分析

评价指标	抽样数据连通性误差率(%)		
	网易微博	优酷	蘑菇街
Radius	5.87	6.25	4.77
Diameter	6.73	3.27	5.87
Avg len	1.59	2.07	3.02
Scc	3.83	5.45	5.63

OSM-MSCS 之所以能够取得较好的连通性,关键在于在中观和宏观尺度上分别建立了有效的连接机制:在中观尺度:每一个首代父图内部使用 DR 方法将其嵌套子群内低度行动者的相干关系有效地保留下来,准确地还原了嵌套子群的“小连通”关系;在宏观尺度:使用 Gibbs 方法对不同首代父图之间的相干关系建立了平稳分布,确保了首代父图之间“大连通”关系.综合样本数据在宏观尺度与中观尺度的连通特性,从而有效地刻画了整体在线社会化媒体的连通特性.

4.4 整体抽样效果分析

在完成以上所有工作以后,我们从整体特性、凝聚子群特性、关键节点这 3 个维度对不同抽样方法进行了综合比较分析,结果如下:

① 宏观特性(见表 6~表 8)

Table 6 K-S statistics of YouKu.com

表 6 优酷 K-S 检验

抽样方法	优酷【30%】样本覆盖率抽样比较分析						
	In-Degree	Out-Degree	Correlation	Diameter	Avg.Path.Len	Cluster coefficien	wcc
OSM-MCS	0.054 9	0.041 3	0.061 2	0.042 8	0.089 0	0.042 8	0.069 8
MHRW	0.096 7	0.077 4	0.100 7	0.087 0	0.126 9	0.109 7	0.127 9
FF	0.112 1	0.094 8	0.142 9	0.098 5	0.135 2	0.147 8	0.122 9
BFS	0.103 2	0.120 4	0.137 7	0.126 9	0.144 1	0.145 6	0.130 5

Table 7 K-S statistics of Netease microblog

表 7 网易微博 K-S 检验

抽样方法	网易微博【30%】样本覆盖率抽样比较分析						
	In-Degree	Out-Degree	Correlation	Diameter	Avg.Path.Len	Cluster coefficien	wcc
OSM-MCS	0.073 2	0.089 0	0.094 9	0.057 8	0.048 7	0.059 0	0.031 7
MHRW	0.189 1	0.173 2	0.176 2	0.171 0	0.158 9	0.179 9	0.231 0
FF	0.207 8	0.247 3	0.192 9	0.264 6	0.176 1	0.298 7	0.365 4
BFS	0.203 2	0.220 4	0.237 7	0.276 9	0.204 1	0.295 6	0.320 5

Table 8 K-S statistics of Mogujie.com

表 8 蘑菇街 K-S 检验

抽样方法	蘑菇街【30%】样本覆盖率抽样比较分析						
	In-Degree	Out-Degree	Correlation	Diameter	Avg.Path.Len	Cluster coefficien	wcc
OSM-MCS	0.046 7	0.037 4	0.029 9	0.040 1	0.070 7	0.044 9	0.019 7
MHRW	0.216 1	0.158 1	0.121 8	0.178 7	0.184 3	0.200 5	0.143 6
FF	0.228 1	0.231 6	0.196 7	0.222 8	0.196 5	0.237 6	0.287 3
BFS	0.210 4	0.281 6	0.207 2	0.229 3	0.176 8	0.240 2	0.261 7

② 凝聚子群特性(见表 9)

Table 9 K-S statistics of cohesive group

表 9 凝聚子群 K-S 检验

抽样方法	统计指标								
	Degree centrality			Closeness centrality			Between centrality		
	网易微博	蘑菇街	优酷	网易微博	蘑菇街	优酷	网易微博	蘑菇街	优酷
OSM-MCS	0.023 7	0.017 1	0.086 5	0.017 5	0.025 9	0.032 8	0.018 7	0.009 8	0.024 6
MHRW	0.092 8	0.213 9	0.176 7	0.157 8	0.097 8	0.148 8	0.128 9	0.083 9	0.115 8
FF	0.246 7	0.235 6	0.170 8	0.242 1	0.144 3	0.197 8	0.153 5	0.104 7	0.154 9
BFS	0.236 5	0.289 9	0.187 6	0.276 5	0.154 4	0.201 9	0.146 8	0.137 7	0.156 4

③ 关键节点地位特性(见表 10)

Table 10 Similarity of opinion leader between sampling and population

表 10 关键节点地位相似性

抽样方法	统计指标			
	高度相似比例(%)	中度相似比例(%)	低度相似(%)	不相似(%)
OSM-MCS	64.57	19.21	11.67	4.55
MHRW	28.38	17.29	25.56	28.97
FF	26.67	18.27	23.86	48.27
BFS	10.79	13.46	16.14	59.43

表 6~表 8 反映了不同抽样方法的综合比较结果.从以上数据可以看出,不同抽样方法对在线社会化媒体宏观层次上所表现出的抽样能力存在明显的差异.对于内部关系稀疏的视频分享网站优酷,每一种抽样方法都能达到宏观尺度上的一致性,但是对于内部结构复杂的网易微博和蘑菇街,OSM-MSCS 的 K-S 检验效果要明显好于 MHRW,MHRW 要明显好于 FF,FF 与 BFS 各有优势.表 9 反映的是不同抽样方法对在线社会化媒体内部凝聚子群特性的刻画能力,OSM-MSCS 抽样方法表现出比较明显的优势,凝聚子群统计估计量 K-S 检验结果表明与总体最接近,即使对于蘑菇街和网易微博这样内部社区规模庞大、嵌套和重叠关系复杂的在线社会化媒体仍然能够较好地表现出其中的社区特性.从表中反映的数据可以看出:BFS 与 FF 的凝聚子群抽样能力很有限,甚至可以认为存在严重的问题,该方法有关凝聚子群 K-S 检验的所有结果都存在较大的偏差,这表明 BFS 与 FF 方法并不是一种能够适合大型在线社会化媒体的抽样方法.表 10 反映的是关键节点的地位关系,表中数据采用皮尔森相关系数对排名前 0.1%的高入度与高出度行动者进行了规则相关性分析,而后依据相关程度将其分为 4 档.很明显,OSM-MSCS 是所有抽样方法中效果最好的方法.高度相似比例与中度相似比例总计达到 83.78%,MHRW 的相似性达到 45.67%,而 BFS 的相似性仅仅达到 24.25%.综合以上数据我们发现:如果按照“整体特性-子群特性-关键节点特性”的方式对抽样进行评价,不同抽样方法间的差距表现得非常明显.显然,差距并非来自抽样规模,根本原因是因为不同抽样方法对复杂结构适应能力的不同.为了确保测试结果的稳定,我们采用以上抽样方法对测试对象反复进行了测试,测试结果与上述数据相符,因而可以确定分析结果的稳定.

从在线社会化媒体大数据抽样发展过程来看:在抽样初期,随着入样单元的不断增加,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.对于简单的数量规律,以上抽样方法都具有足够的适应性,但当节点再扩大到一定规模时,样本间的关系开始发挥作用,这些样本组织在一起,形成或大或小的凝聚子群,由此而形成的内聚性对不同抽样方法提出了新的挑战.BFS 方法由于只能获得相继关系,如果

抽样规模不大幅提升,则其凝聚子群发现能力非常有限.FF 方法的缺陷在于其确定性的抽样方法难以适应复杂的网络结构,越是复杂分布,FF 表现出的偏差越大.MHRW 方法虽然可以筛选样本,但是该方法一旦进入一个规模较大且关系紧密的子群时就会陷入其中,很难跳出.由于大型在线社会化媒体内部包含众多具有嵌套关系的大型凝聚子群,因而 MHRW 方法对凝聚子群的抽样效果并不好.从实验数据的结果可以看出:唯有 OSM-MSCS 方法能够利用 DR 样本舍选的特性,在不同嵌套子群之间相互切换,适应在线社会化媒体内部的复杂结构.

当抽样规模继续膨胀时,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;二是克服了局部凝聚子群内部的组织复杂性.此时所面临的关键任务就是通过选择不同凝聚子群之间的相干关系来形成更大的子群乃至整体网络,即,形成整体涌现特征.BFS 和 MHRW 都没有能力理解首代父图相干关系的特征,因而其群间关系的选择带有很大的盲目性,最终导致的结果就是样本不能忠实地反映总体的连通性.相比较而言,OSM-MSCS 方法可以有效地识别首代父图的群间关系,并能通过 Gibbs 抽样获得首代父图群间关系的平稳分布,因此可以确保整体网络的连通与整体网络结构的完备.

5 总结与展望

OSM-MSCS 抽样方法与以往的研究工作相比,主要贡献包括如下 3 个方面:一是改变了在线社会化媒体大数据的抽样机制,将其由单阶段抽样转化为多阶段整群抽样,使在线社会化媒体由“黑盒子”变成了“白盒子”,为抽样提供了与客观事实更相近的样本概率控制方法,同时还为未来抽样信度与效度的分析打下了良好基础;二是利用整群分解,将整体网络分解成不同的凝聚子群,使得在线社会化媒体并行抽样成为可能,从而较好地避免了样本局部性陷入、马尔可夫链燃烧预热等问题,提高了抽样效率和抽样质量;三是完善了在线社会化媒体大数据抽样估计量评价标准,从“个体地位-群体凝聚性-整体结构性”这 3 个层次进行综合评价,使得抽样数据能够更好地满足领域问题研究的需要.

未来仍需要对 OSM-MSCS 方法作进一步的优化:一是要优化子群抽样的方法,使其能够更好地适应子群内的嵌套结构,最理想的情况应当是针对一种特定的结构就有一种与之对应的抽样方法;二是优化首代父图相干关系的选择策略,使其能够适应各种复杂分布,既能确保群间连通性,还能与总体更加一致.

致谢 感谢中国电信提供的大规模计算环境以及审稿人的建议.

References:

[1] Yunus M. Building Social Business: The New Kind of Capitalism That Serves Humanity’s Most Pressing Needs. Philadelphia: Public Affairs, 2011. 2−17.

[2] Leung L. Generational differences in content generation in social media: The roles of the gratifications sought and of narcissism. Computers in Human Behavior, 2013,29(3):997−1006. [doi: 10.1016/j.chb.2012.12.028]

[3] Becchetti L, Castillo C, Donato D, Fazzone A. A comparison of sampling techniques for Web graph characterization. In: Proc. of the Workshop on Link Analysis (LinkKDD 2006). New York: ACM Press, 2006. http://ailab.ijs.si/dunja/linkkdd2006/Papers/becchetti.pdf [doi: 10.1.1.69.1736]

[4] Leskovec J, Faloutsos C. Sampling from large graphs. In: Proc. of the 12th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2006. 631−636. [doi: 10.1145/1150402.1150479]

[5] Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B. Measurement and analysis of online social networks. In: Proc. of the 7th ACM SIGCOMM Conf. on Internet Measurement. New York: ACM Press, 2007. 29−42. [doi: 10.1145/1298306.1298311]

[6] Amanda LT, Peter JM, Mason AP. Social structure of Facebook networks. Physica A, 2012,391:4165−4180. [doi: 10.1016/j.physa.2011.12.021

[7] Ferrara E. A large-scale community structure analysis in Facebook. EPJ Data Science, 2012,1(1):1−30. [doi: 10.1140/epjds1]

[8] Ahmed N, Neville J, Kompella R. Network sampling via edge-based node selection with graph induction. Computer Science Technical Reports, 11-016, 2011. 1−10.

[9] Gjoka M, Kurant M, Butts CT, Markopoulou A. Practical recommendations on crawling online social networks. IEEE Journal on Selected Areas in Communications, 2011,29(9):1872−1892. [doi: 10.1109/JSAC.2011.111011]

[10] Goel S, Salganik MJ. Assessing respondent-driven sampling. Proc. of the National Academy of Sciences, 2010,107(15):6743−6747. [doi: 10.1073/pnas.1000261107]

[11] Rasti AH, Torkjazi M, Rejaie R, Duffield N, Willinger W, Stutzbach D. Evaluating sampling techniques for large dynamic graphs. Technical Report, CIS-TR-08-01, Oregon: Department of Computer and Information Science, University of Oregon, 2008. 1−14.

[12] Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. London: Chapman & Hall/CRC, 1996. 67−72.

[13] Lovász L. Random walks on graphs: A survey. Combinatorics, Paul Erdos is Eighty, 1993,2(1):1−46.

[14] Gjoka M, Kurant M, Butts C, Markopoulou A. Walking in Facebook: A case study of unbiased sampling of OSNs. In: Proc. of the 29th Conf. on Information Communications. New York: IEEE Press, 2010. 2498−2506. http://dl.acm.org/citation.cfm?id=1833840 [doi: 10.1109/INFCOM.2010.5462078]

[15] Wang TY, Chen Y, Zhang ZB, Xu TY, Jin L, Hui P, Deng BX, Li X. Understanding graph sampling algorithms for social network analysis. In: Proc. of the 31st Int'l Conf. on Distributed Computing Systems Workshops (ICDCSW). New York: IEEE Press, 2011. 123−128. [doi: 10.1109/ICDCSW.2011.34]

[16] Jin L, Chen Y, Hui P, Ding C, Wang TY, Vasilakos AV, Deng BX, Li X. Albatross sampling: Robust and effective hybrid vertex sampling for social graphs. In: Proc. of the 3rd ACM Int'l Workshop on MobiArch. New York: ACM Press, 2011. 11−16. [doi: 10.1145/2000172.2000178]

[17] Hubler C, Kriegel HP, Borgwardt K, Ghahramani Z. Metropolis algorithms for representative subgraph sampling. In: Proc. of the 8th IEEE Int'l Conf. on Data Mining (ICDM 2008). New York: IEEE Press, 2008. 283−292. [doi: 10.1109/ICDM.2008.124]

[18] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 1997,57(1):97−109.

[19] Leskovec J, Lang KJ, Dasgupta A, Mahoney MW. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Mathematics, 2009,6(1):29−123. [doi: 10.1080/15427951.2009.10129177]

[20] Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature, 2010,466(7307):761−764. [doi: 10.1038/nature09182]

[21] Lancichinetti A, Kivelä M, Saramäki J, Fortunato S. Characterizing the community structure of complex networks. PloS One, 2010,5(8):e11796. [doi: 10.1371/journal.pone.0011796]

[22] Xie J, Szymanski BK. Towards linear time overlapping community detection in social networks. In: Advances in Knowledge Discovery and Data Mining, Vol.7302. Berlin, Heidelberg: Springer-Verlag, 2012. 25−36. [doi: 10.1007/978-3-642-30220-6_3]

[23] Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010,12(10):103018.

[24] Mira A. On Metropolis-Hastings algorithms with delayed rejection. Metron, 2001,59(3-4):231−234. [doi: 10.1088/1367-2630/12/10/103018]

[25] Trias M, Vecchio A, Veitch J. Delayed rejection schemes for efficient Markov-chain Monte-Carlo sampling of multimoda, distributions. arXiv preprint arXiv, 2207,0904:2009.

[26] Martinez WL. Computational statistics in MATLAB. Wiley Reviews: Computational Statistics, 2011,3(1):69−74.

[27] Mengersen K, Knight S, Robert CP. MCMC: How do we know when to stop? In: Proc. of the Bulletin of the Int'l Statistical. 1999. 58.

[28] Roberts G, Rosenthal J. Examples of adaptive MCMC. Journal of Computational and Graphical Statistics, 2009,18(2):349−367.

崔颖安(1975－),男,陕西西安人,博士,讲师,主要研究领域为社会化媒体大数据抽样,大数据分析.
E-mail: cuiyan@xaut.edu.cn

王志晓(1977－),男,博士,主要研究领域为在线社会网络动力学分析.
E-mail: wangzhx@xjtu.edu.cn

李雪(1974－),女,博士,讲师,主要研究领域为社会化营销,口碑营销.
E-mail: lixue@snnu.edu.cn

张德运(1949－),男,教授,博士生导师,主要研究领域为复杂网络,传感器网络,智能家庭网络.
E-mail: zhangdeyun@xjtu.edu.cn