网络社区发现优化：基于随机游走的边权预处理方法

刘 阳* 季新生 刘彩霞
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要：在网络日趋复杂化、巨大化的背景下，仅依靠网络拓扑特征难以提高现有社区发现算法的精确度和性能。该文提出一种优化网络社区发现的边权预处理方法，基于马尔可夫随机游走理论建模社区结构对复杂网络行为的影响，根据多重随机游走对网络连接的遍历情况，重新衡量网络边权。预处理后的边权作为网络拓扑的有效补充信息，能够将网络社区结构去模糊化，从而改善现有算法的社区发现性能。对于一些典型的计算机生成网络和真实网络，经实验验证：该预处理方法能够有效提升现有部分社区发现算法的准确性和效率。

关键词：社会网络；社区发现；预处理；随机游走；边权
中图分类号：TP393 **文献标识码**： A **文章编号**：1009-5896(2013)10-2335-06
DOI: 10.3724/SP.J.1146.2012.01676

Optimizing Community Detection Using the Pre-processing of Edge Weighted Based on Random Walk in Networks

Liu Yang Ji Xin-sheng Liu Cai-xia
(*National Digital Switching System Engineering & Technological R&D Center*, *Zhengzhou* 450002, *China*)

Abstract: In the context of social network becomes more and more complicated and huge, it is difficult to improve the accuracy and performance of existing community detection algorithms only relying on the network topological features. Based on Markov random walk theory, this paper proposes a method of edge weighted pre-processing for optimizing community detection, models community structures how to influence on the complex network behaviors. According to the situation of multiple random walk traverses on the network links, the network edges weight is reset, and makes it as the network topology effective supplementary information to promote the network community structure defuzzification, thus the performance of the existing algorithms is improved for community detection. For a set of typical benchmark computer-generated networks and real-world network data sets, the experimental results show that the pre-processing method can effectively improve the accuracy and efficiency of some existing community detection algorithms.

Key words: Social network; Community detection; Pre-processing; Random walk; Edge weighted

1 引言

近年来，随着移动互联网和云计算等技术的推动，以社交网络、维基百科、博客、播客等为代表的社会网络应用迅速普及，使得这些复杂网络系统不断地深入我们的生活。社区结构作为复杂网络中广泛存在的重要属性，对揭示网络的模块化、异质性等特征，研究网络内部相互作用、网络功能演化都有重要意义。许多实际问题都可以建模为复杂网络社区发现问题，如社会关系群落识别、社交网络的兴趣小组、根据消费行为圈定用户类型、自组织网络成簇、蛋白质相互作用等等。

现有社区发现方法大多依据网络拓扑信息衡量节点、边的相似度进行聚类，进而发现网络的社区结构；按照社区发现的基本策略[1]可以归纳为两大类：启发式方法，包括 GN(Girvan Newman)算法[2]、FEC(Finding and Extracting Communities)算法[3]等；基于优化的方法，包括 Fast Newman 算法[4]、CNM 算法[5]、BGLL 算法[6]等，此外还有基于层次聚类、图分割等思想的方法。近年来，基于随机游走的启发式社区发现方法开始引起人们的关注。Newman[7]提出了通过随机游走模拟网络消息传播来衡量节点介数以辅助社区发现，与最短路径、模拟电流两种衡量方法相比，其复杂度低，效果更好。Cai 等人[8]提出了基于随机游走的重叠社区发现算法，每个节点都作为源节点发生随机游走并生成其遍历节点集，根据集合间的重叠系数迭代合并得到重叠社区。Rosvall 等人[9]提出基于节点连接关系的随机游走社区发现方法，根据随机游走对节点的访

2012-12-24 收到，2013-05-17 改回
国家 863 计划项目(2011AA010605)资助课题
*通信作者：刘阳 liuyang198610@163.com

问频率将节点和社区编码，力求最小化平均编码长度的过程也即社区发现的过程。Alahakoon 等人[10]提出了一种基于 k-path 随机游走的节点介数替代计算方法，并验证了其在改善社区发现复杂度和精确度方面的优势。上述社区发现方法引入随机游走理论衡量节点、边的相似度，实质上还是沿用利用网络拓扑信息进行节点、边聚类的思想。而基于邻接矩阵的网络拓扑信息对网络社区结构的表征能力有限，当网络社区结构不明显、社区规模尺度差异较大时，就会出现分辨精度低、分辨极限(resolution limit)等问题。

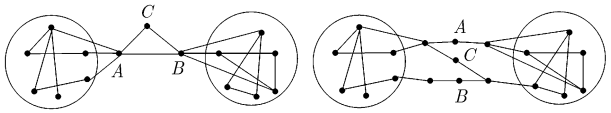
基于上述分析，本文在文献[10]的基础上提出了一种基于随机游走的边权预处理方法，通过重新衡量网络边权重为社区发现算法提供网络拓扑补充信息以优化网络社区发现。在该算法中，随机游走的agent(代理)根据所在节点的边权生成相应大小的路径选择概率，按照该概率分布随机选择下一跳路径；根据多重随机游走对网络连接的遍历情况，重新衡量网络边权。最终获得的边权刻画了网络连接对网络行为传输、扩散的重要性和贡献度，使网络内部的结构特征体现在网络连接的权重上。

本文第 2 节分析了基于最短路径衡量网络边权的局限；第 3 节介绍了基于随机游走的边权预处理模型及其算法；第 4 节给出了实验结果和分析；第 5 节对全文进行总结。

2 问题描述

近年来，有研究和分析[11]表明刻画网络边权对社区发现的重要影响：(1)节点由于边连接状态不同，产生、传播消息的能力是不均衡的。(2)节点对于消息传播的边选择具有不均衡性。这表明：网络行为动力学过程对边的选择造就了社区结构背景下不均衡的边权重分布，同时这种具有差异性的边权也是形成网络社区的内在因素，对社区发现也具有指导作用。

传统基于最短路径的边权重度量方法依赖于网络的拓扑信息，除了复杂度较高，往往会忽视最短路径周围节点的重要性。这种不足使得网络边权信息无法充分应用于网络社区结构的深入挖掘，限制了社区发现准确性和精确度。如图 1 中网络 1，网络 2 所示，节点 A, B 由于位于社区间的最短路径上，边介数较高；而节点 C 没有最短路径直接经过，边介数较低。而在实际网络中，即使不了解边和节点是否位于最短路径，人们也会通过可转达的第三方来传递信息。因此，节点 C 很可能在社区间交互扮演重要角色，但这种重要性在基于最短路径的衡量方法中却难以体现。



(a)网络1 (b)网络2

图 1 易引起边介数衡量错误的典型网络拓扑图

随机游走的过程反映了网络中心性、平均距离等结构特征，通过随机游走来衡量影响网络行为的边权，能够避免单纯依靠邻接矩阵拓扑信息带来的误差。尽管每次随机游走过程只反映 agent 的局部观点和选择性，但通过一定数量级的多重随机游走融合已有的“路标”信息(以前随机游走设置的边权)，最终能够获取全局意义上的网络边权重图。

3 基于随机游走的边权预处理

3.1 算法模型

复杂网络中的随机游走是指以网络节点为载体，按照一定概率从网络上任意一点转移到与之相连的邻居节点的状态转移过程。令 G 表示某给定复杂网络，其邻接矩阵为 $\boldsymbol{A}=|a_{ij}|_{n\times n}$，$a_{ij}$ 表示节点 i 和 j 的连接状况。边 e_i 权重初始化为 $w_{in}(e_i)=a_{ij}/|\boldsymbol{V}|$，$|\boldsymbol{V}|$ 表示网络 G 中的节点个数。根据复杂网络的统计规律：网络中两节点距离越远，存在网络行为交互的可能性越小，相互影响力越弱。本算法模型设定 agent 在复杂网络中随机游走的生命周期有限，将其限制在一定的跳数范围内。由此推导出 k-path[10]的定义。

定义 1 k-path: k-path 表示 agent 在 G 中随机游走，每步以一定的概率选择一条边作为到达邻居节点的游走路径，直至游走完 k 步或满足其它停止条件。$k>0$ 且 $k\in N$。

定义 2 k-path 边介数：令 $e\in E$ 表示图 G 的一条边，其 k-path 边介数 $L_k(e)$ 表示在所有源节点在 k 步随机游走过程中经过边 e 的概率。

由于复杂网络社区结构的存在使社区内的网络行为要比社区间更为频繁，因而 agent 在随机游走过程中停留在社区内的概率大于游走到其它社区的概率。由此，本文以随机游走的转移概率作为启发式规则，根据边的遍历频率不断更新其边权，使社区间的边权被逐渐弱化，社区内部的边权逐渐加强，使网络的社区结构逐步趋于明显，从而实现对复杂网络的社区发现的优化。

3.2 算法描述

根据上文所述，可将算法模型分解为 k-path 随机游走和边权评价两个部分：

(1)每轮随机游走前选定 1 个源节点，共完成 λ

轮随机游走(组成包含 λ 个节点的随机游走源节点集 $\boldsymbol{V}_s$)。每轮有 k 步(k-path)的随机游走过程中，agent 根据所在节点的边权 $w(e_{ij})$ 计算其路径选择概率，边权越大，选择的概率越大；反之，则越小。

(2)每步随机游走完成后，对被遍历边加权其初始权重 $w_{in}(e_i)$，使重新衡量的边权能够影响后续的随机游走路径选择；所有随机游走完成后，即得到该网络预处理后的最终边权。

为保证随机游走对网络的有效覆盖，在每次随机游走之前选择当前网络中没有遍历的，归一化节点度 $D_n(v_n)=\left|E(v_n)\right|/|\boldsymbol{V}|$ 最大的节点作为随机游走的源节点。$D_n(v_n)$ 越大，表示节点 v_n 在图 G 的连通性越好。在第 j 轮的游走过程中，当 agent 到达节点 v_n 时，v_n 任意一条边 e_i 的权重表示为 $w(e_{ij})$，其被选中的概率为

$$P\left(e_{ij}\right)=\frac{w\left(e_{ij}\right)}{\sum\limits_{e_i\in E(v_n)} w\left(e_{ij}\right)} \tag{1}$$

且 $\sum_{e_i\in E(v_n)} P(e_{ij})=1$。Agent 根据式(1)计算各边的选择概率，按照该概率分布随机选择其下一步游走的边连接。定义 $\delta(e_{ij})$ 标注边 e_i 在第 j 轮随机游走过程中的遍历状态为

$$\delta\left(e_{ij}\right)=\begin{cases}1, & e_i\text{被遍历}\\ 0, & e_i\text{未被遍历}\end{cases} \tag{2}$$

当从源节点集 $\boldsymbol{V}_s$ 出发的所有随机游走完成后，边 e_i 的最终权重为

$$L_k(e_i)=\sum_{v\in\boldsymbol{V}_s} P\left(e_i\right)=\sum_{v\in\boldsymbol{V}_s} P\left(e_i|v\right)P(v) \tag{3}$$

$L_k(e_i)$ 即为其 k-path 边介数，$P(v)$ 表示节点 v 被选作随机游走源节点的概率。通过 λ 轮的 k-path 随机游走来重新设置 e_i 的边权重，$L_k(e_i)$ 最终刻画了网络 G 在信息传播、行为交互过程中选择 e_i 连接的倾向性和趋势。

当网络存在局部特殊拓扑结构(孤立连接、孤立环路)时，或其它极端条件下，agent 在某条边上重复遍历，或形成局部小型环路，会导致反复经过的边的权重被放大化。为避免随机游走 agent 陷入这种“地形陷阱”，在随机游走过程中采取如下的控制策略来避免这种情况：只有 agent 当前所在节点还有存在边未被经过，且步数小于 k，则随机游走才能够继续进行；否则，此次随机游走结束。

3.3 算法分析

由于直接计算 $L_k(e_i)$ 的复杂度较大，为减小计算量，本算法实际获取的是其近似值：

$$\hat{L}_k\left(e_i\right)=w\left(e_i\right)\big/\lambda \tag{4}$$

定义变量 $T(e_{ij})=E\left[\sum_{p=1}^{k}\delta\left(e_{ij}\right)\right]$，由于 $\delta(e_{ij})$ 是0,1变量，则上式可转化为 $E[\delta(e_{ij})]=P(\delta(e_{ij})=1)=P(e,j,v)$，$v$ 是边 e_i 在第 j 轮随机游走的源节点。根据贝叶斯公式，使 $P(e_i,j,v)=P(e_i,j|v)P(v)$。由于在每轮随机游走过程中，边 e_i 最多只能被遍历一次，$P(e_i,j|v)$ 为 0,1 变量。那么从节点 v 出发的边 e_i 的 k-path 遍历权重为

$$\sum_{p=1}^{k} P\left(e_i,j|v\right)=P\left(e_i|v\right) \tag{5}$$

从而可以得出

$$\begin{aligned}E\left[T\left(e_{ij}\right)\right]&=E\left[\sum_{v\in\boldsymbol{V}_s}\sum_{i=1}^{k}P\left(e_i,j|v\right)P(v)\right]\\&=E\left[\sum_{v\in\boldsymbol{V}_s}P\left(e_i|v\right)P(v)\right]\end{aligned} \tag{6}$$

根据式(3)和式(4)，可以推导得到

$$E\left[\hat{L}_k\left(e_i\right)\right]=\frac{1}{\lambda}\sum_{\boldsymbol{V}_s}E\left[\sum_{v\in\boldsymbol{V}_s}P\left(e_i|v\right)P(v)\right]=L_k\left(e_i\right) \tag{7}$$

本文预处理方法需要设置的参数包括：随机游走步数 k，源节点个数 λ。为保证随机游走的覆盖效果，必须满足在 k 步之内，agent 能够从源节点游走到网络中任意其它节点。对于一般无标度网络和小世界网络来说，网络平均直径长度较小，根据本文实验所采用网络的规模，设置 k 的经验值区间为 [6,20]。对于 λ 的设置，则要求 $\hat{L}_k(e)$ 与期望值的偏差足够小

$$P\left(\left|\hat{L}_k(e)-L_k(e_i)\right|\geq\xi\right)\leq 2\exp\left(-2\lambda\xi^2\right) \tag{8}$$

根据 Hoeffding 不等式[16]可知：当 $\lambda=O(|\boldsymbol{V}|\cdot\log|\boldsymbol{V}|)$ 时，近似值 $\hat{L}_k(e_i)$ 与 $L_k(e_i)$ 的误差 $\sigma\leq 2/|\boldsymbol{V}|$。对于 $\boldsymbol{V}=\{v_1,\cdots,v_n\}$，$|E|=m$ 的网络，随机游走的边权预处理方法的复杂度取决于随机游走次数 λ 和游走步数 k。其中运算量较大包括边权值的更新和随机游走的循环过程，最好情况是 k 步之内就经过重复节点使游走结束，时间复杂度为 $O(\lambda\log|km|)$；最坏情况是每轮都完成 k 步随机游走，时间复杂度为 $O(\lambda|k|)$。

4　实验

为了验证 k-path 随机游走的边权预处理方法的性能，本文利用人工网络和真实网络分别对其进行测试，并给出了预处理前后的参数对比分析。算法实验环境为：处理器 Interl(R)Core(TM)2 2.83 GHz，内存 4.00 GB，硬盘 500 G, Microsoft Windows 7 Professional 的 64 位操作系统，编程环境 Matlab R2010b。

4.1 人工测试网络

人工测试网络采用 LFR(Lancichinetti Fortunato Radicchi)程序[12]生成的基准模拟网络数据。LFR 能够灵活生成较高质量的测试网络数据，在网络社区发现验证方面应用广泛。实验按照表 1 的参数设置分别生成 4 组网络测试数据集，每组包含 10 个网络，所有网络的度分布和社区规模分布的幂指数分别为 2, 1，编号为组 1 到组 4。其中，参数 mix 表示网络中社区间的连接数占所有边总和的比例，mix 越小，测试网络的社区结构越模糊。

随机游走步数 k 是影响边权预处理方法的重要参数，在 3.3 节的算法分析中给出了 k 的经验值区间。从 4 组网络测试数据集各随机抽取 1 个网络，编号为网络 1 到网络 4，通过实验分析取不同的 k 值对预处理后边权分布的影响。从图 2 可以发现：低边权分布比较密集，高边权分布比较稀疏。这是因为大部分连接由于被遍历次数较少，边权重较低，形成了“重尾”；而高权重连接分布的数量级比较小，在双对数坐标下分布比较均匀。当 $\lambda = O(|\boldsymbol{V}|\log|\boldsymbol{V}|)$，$k \subseteq [6,20]$ 时，在不同网络条件下，边权都可以迅速完成收敛。这说明源节点选择和游走策略对凸显网络社区结构起到了作用。在同一网络不同 k 值条件下，边权分布的收敛趋势大致相同；随着 k 值增大，小规模网络的边权增幅更加明显。对于相同规模的网络，社区结构明显的网络的边权差距更大，区别更为明显。由于 k =20 时，网络边权分布相对稳定，且区分度更好，因此接下来的实验若没有特别说明，都设置 k 值为 20。

实验选择 Fast-Newman, FEC, BGLL 3 种代表性的算法，基于评估社区发现精度的两个常用指

表 1 LFR 生成的测试网络配置

测试网络数据	节点数	平均节点度	最小社区节点数	最大社区节点数	mix	最大节点度
组 1	1000	15	20	120	0.2	50
组 2	1000	15	20	120	0.4	50
组 3	10000	15	20	120	0.2	50
组 4	10000	15	20	120	0.4	50

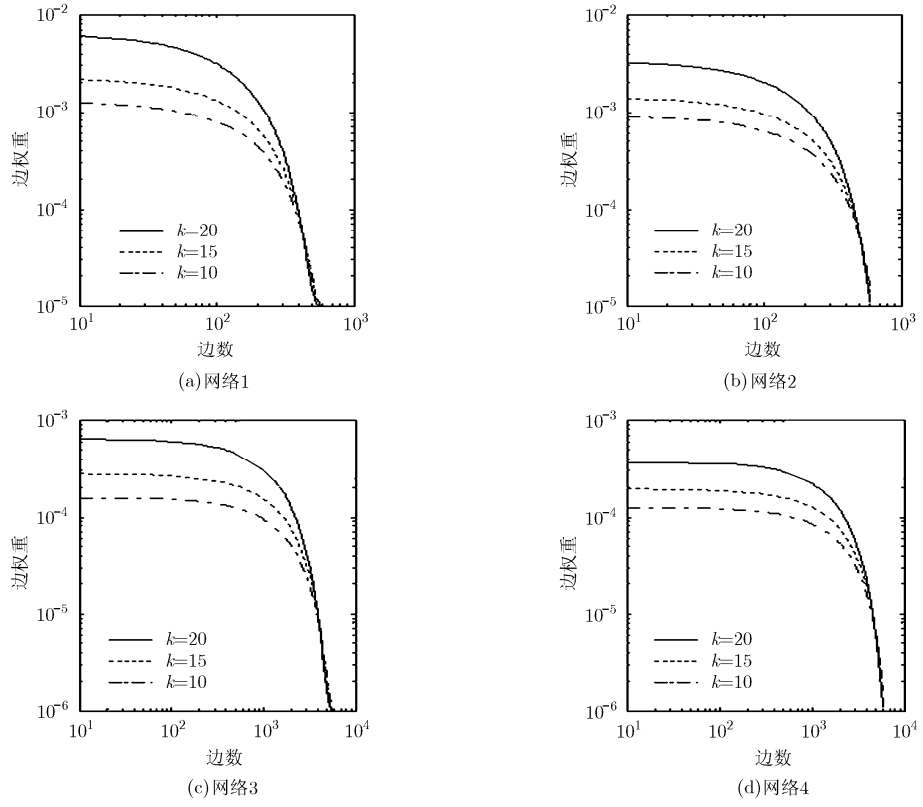


图 2 网络 1~网络 4 的 k-path 随机游走的边权分布的拟合曲线图

标：模块度(Modularity)和归一化互信息 NMI[18] (Normalized Mutual Information)来对比其未经预处理和经预处理(Preprocess, Pre)的社区发现结果，从而验证该预处理方法的有效性。由于人工测试网络社区结构已知，将 4 组测试数据集每组随机抽取 5 个网络，分别完成 3 种算法经 k-path 随机游走边权预处理和未经该预处理条件下的社区发现，将社区发现结果与 LFR 程序提供的实际基准社区结构进行对比，计算其各组测试数据集的 NMI 均值作为最终结果。分析表 2 中 NMI 值的变化趋势可知：在网络规模较小，社区结构明显时，3 种算法都具有较高的社区发现精度。但是当网络规模变大，社区结构模糊度增加时，社区发现精度明显下降。

如表 2 所示，经预处理后，组 1 到组 4 的 NMI 值平均增幅分别为 0.0206, 0.0251, 0.0368, 0.0448。随着测试网络 mix 系数增加，网络规模增大，反映社区发现效果的 NMI 值的增幅也呈现扩大趋势。在网络规模较小，社区结构明显条件下，基于随机游走的边权预处理对 3 种算法的社区发现精度改善并不明显。当网络规模增大，社区结构趋于模糊，边权预处理的效果才得以显现，使得 NMI 增幅相对较大。这说明基于随机游走的边权预处理方法能够有效地改善社区结构模糊化的问题，从而提升现有网络社区发现算法的社区发现性能。

表 2 未经预处理和经预处理的社区结构 NMI 值对比

NMI	组 1	组 2	组 3	组 4
Fast-Newman	0.8439	0.6681	0.7533	0.7145
Pre Fast-Newman	0.8915/0.0476↑	0.7122/0.0441↑	0.7981/0.0448↑	0.7752/0.0607↑
FEC	0.9803	0.9245	0.8462	0.7786
Pre FEC	0.9925/0.0122↑	0.9480/0.0235↑	0.8795/0.0333↑	0.8160/0.0374↑
BGLL	0.9827	0.9773	0.8638	0.7649
Pre BGLL	0.9947/0.012↑	0.9849/0.0076↑	0.8962/0.0324↑	0.8012/0.0363↑

4.2 真实网络

真实网络相比计算机生成的测试网络更不规则，因而其社区结构往往也更加复杂。这里采用 3 个具有不同规模的典型真实网络数据集：Cond-dt 2003[4], Hep-th[19], Polblooks[20]来进一步测试预处理算法性能。Cond-dt 2003 是 1995 年 1 月 1 日到 2003 年 6 月 30 日收集的凝固态物质物理合著网络，包含 31163 个节点，120029 条边；Hep-th 是 1995 年至 1999 年收集的高能物理合著网络，包含 8361 个节点，1571 条边；Polbooks 是 Krebs 建立的美国政治图书网络，包含 105 个节点，441 条边。按照书中的观点倾向可分为保守派 49 个，自由派 43 个，中间派 13 个，分别用 C, L, N 表示。

由于 Polblooks 社区结构已知，可以用来测试社区发现算法的精确度。表 3 给出了 3 种算法在预处理前后的效果对比，other 表示算法在 3 个社区之外，多划分出了 1 个社区。由于网络规模较小，随机游走步数 k=10。与 Polblooks 实际社区结构进行比较，发现：经边权预处理后的各算法的社区发现精确度得到改善。即使 Fast-Newman 经预处理后划分出 4 个社区，但社区大小和分布已经大大接近于真实网络。

表 4 给出了预处理前后 3 种算法社区发现的 Q 值对比。由于网络结构不同，且社区发现算法的原理和方法不同，预处理对社区发现的改善也有差异。Polblooks, Cond-dt 2003, Hep-th 这 3 个真实网络经预处理后社区发现的 Q 值平均增幅分别为 0.0296, 0.0483, 0.0420。分析表 4 可知：经随机游走的边权预处理后，3 种算法社区发现的精确度都有了不同程度的改善。当网络社区规模变化较大且模糊时，预处理后的边权信息能够帮助算法减少社区发现过程中的错分、误分，使其更接近真实的社区结构。虽然预处理过程增加了部分计算量，但获得的边权信息强化了网络的社区结构，能够有效地改善网络社区发现算法的性能。

表 3 Polbooks:未经预处理和经预处理的 3 种算法社区发现结果对比

社区节点个数	C	L	N	other
Polbooks	49	43	13	–
FN	79	11	15	–
Pre FN	61	19	18	7
FEC	40	50	7	8
Pre FEC	43	47	15	–
BGLL	57	32	9	7
Pre BGLL	53	41	11	–

5 结论

本文提出了一种服务于优化网络社区发现的基于随机游走的边权预处理方法，根据边在随机游走过程中被遍历的频率重新衡量其权重。最终得到的边权是多重随机游走提供的全局网络信息，不仅反映了其连通能力，更体现了其在社区中的重要性和贡献度，从而使网络社区结构更明显。实验证明：随机游走的预处理方法获取的边权，可以作为网络拓扑的有效补充信息，改善网络社区发现算法的性能。但该预处理方法不适用于某些特殊的有向网络，它限制了 agent 的自由跳转，造成部分边遍历受限，使随机游走不能实现完整有效的遍历覆盖。同时，实际网络中节点属性对社区结构也有重要影响。今后的研究将围绕在随机游走的预处理过程中同时衡量节点权重和边权重，以期进一步提升社区发现算法的性能。

表 4 未经预处理和经预处理的社区结构 Q 值对比

Q	Polblooks	Cond-dt 2003	Hep-th
Fast-Newman	0.4959	0.7214	0.7520
Pre Fast-Newman	0.5206/0.0247↑	0.7558/0.0344↑	0.7865/0.0345↑
FEC	0.4583	0.5844	0.6227
Pre FEC	0.5033/0.0450↑	0.6162/0.0318↑	0.6554/0.0327↑
BGLL	0.4986	0.6415	0.6312
Pre BGLL	0.5176/0.0190↑	0.7201/0.0786↑	0.6901/0.0589↑

参 考 文 献

[1] 杨博, 刘大有, Liu Ji-ming, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54–66.
Yang Bo, Liu Da-you, Liu Ji-ming, *et al*.. Complex network clustering algorithms[J]. *Journal of Software,* 2009, 20(1): 54–66.

[2] Girvan M and Newman M E J. Community structure in social and biological networks[J]. *Proceedings of the National Academy of Science*, 2002, 9(12): 7821–7826.

[3] Yang B, Cheung W K, and Liu J M. Community mining from signed social networks[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(10): 1333–1348.

[4] Newman M E J. Fast algorithm for detecting community structure in networks[J]. *Physical Review E*, 2004, 69(6): 066133.

[5] Blondel V D, Guillaume J L, Lambiotte R, *et al*.. Fast unfolding of communities in large networks[J]. *Journal of Statistical Mechanics*: *Theory and Experiment*, 2008, COI: 10.1088/1742-5468/2008/10/910008.

[6] Newman M E J. A measure of betweenness centrality based on random walks[J]. *Social Networks*, 2005, 27(1): 39–54.

[7] Cai Bing-jing, Wang Hai-ying, Zheng Hui-ru, *et al*.. An improved random walk based clustering algorithm for community detection in complex networks[C]. 2011 IEEE International Conference, on Systems, Man,and Cybernetics, Alaska, USA, Oct. 2011: 2162–2167.

[8] Rosvall M and Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. *PNAS*, 2008, 105(4): 1118–1123.

[9] Alahakoon T, Tripathi R, Kourtellis N, *et al*.. K-path centrality: a new centrality measure in social networks[C]. Proceedings of 4th Workshop on Social Network Systems, Salzburg, Austria, 2011: 1–6.

[10] Ferrara E. Community structure discovery in Facebook[J]. *International Journal of Social Network Mining*, 2012, 1(1): 67–90.

[11] Hoeffding W. Probability inequalities for sums of bounded random variables[J]. *Journal of the American Statistical Association*, 1963, 58(301): 13–30.

[12] Lancichinetti A, Fortunato S, and Radicchi F. Benchmark graphs for testing community detection algorithms[J]. *Physical Review E*, 2008, 78(4): 046110.

[13] Newman M E J. The structure of scientific collaboration networks[J]. *Proceedings of the National Academy of Sciences*,2001, 98(2): 404–409.

[14] Adamic L A and Glance N. The political blogosphere and the 2004 US election: divided they blog[C]. Proceedings of the 3rd International Workshop on the Weblogging Ecosystem, New York, USA: ACM, 2005: 36–43.

刘 阳： 男，1986 年生，博士生，研究方向为社会网络分析、数据挖掘.

季新生： 男，1969 年生，教授，博士生导师，主要研究领域为电信网安全、移动通信安全.

刘彩霞： 女，1974 年生，副教授，硕士生导师，主要研究领域为移动通信安全、社会网络分析.