doi: 10.7690/bgzdh.2013.04.012

复杂网络局部社区发现算法

李星[1,2]，钟志农[1]，景宁[1]，伍勇[1]
（1. 国防科学技术大学电子科学与工程学院，长沙 410073；2. 中国人民解放军 63880 部队，河南 洛阳 410073）

摘要：针对传统的社区发现算法在处理网络局部特性时具有局限性的问题，提出一种基于聚簇优先遍历以及二次切割方法的局部社区发现算法。该算法基于改进流模型的思想，从网络的局部拓扑结构出发，利用节点的排序结果，不依赖于先验知识的新的局部社区划分准则进行节点截断，降低了算法的复杂度。在聚簇优先遍历的基础上通过二次切割的思想实现局部社区发现，应用于网络整体数据无法获取的情况下进行社区发现，最后结合基准数据进行算例分析。分析结果表明，该算法能够较好地发现网络的局部社区结构。

关键词：社区发现；网络分析；节点重要等级
中图分类号：TJ02　　**文献标志码**：A

A Local Community Detection Algorithm on Complex Network

Li Xing[1,2], Zhong Zhinong[1], Jing Ning[1], Wu Yong[1]
(1. *School of Electronic Science & Engineering, National University of Defense Technology, Changsha* 410073, *China*;
2. *No*. 63880 *Unit of PLA, Luoyang* 410073, *China*)

Abstract: For traditional communities detection algorithm with limitations in dealing with a network of local characteristics, a new algorithm of local communities detection based on clustered first search and the second cut method was proposed in this paper. The algorithm is based on the idea of improved flow model, start from the local topology structure of the network, use sequence results of node and carry out node cut by new local community cutting principle which is not depending on priori knowledge, and reduce the complexity of the algorithm. Realize local community search by second cut method based on clustered first search. This algorithm can be applied to community searching that overall network data can not be obtained. At last, combine with based data to carry out simulation analysis. The analysis results show that the algorithm can discover local community structure of network effectively.

Key words: local community detection; network analysis; node importance level

0　引言

自然网络都有一定的聚簇结构，其聚簇内节点的边比较稠密，聚簇之间的边比较稀疏[1-3]。准确识别网络的自然聚簇结构称之为社区发现，可广泛应用于搜索引擎优化。用户行为分析，蛋白质功能分析以及科学数据的处理和可视化，已经成为近年来研究的热点。社区发现问题最早可以追溯到 Gibson 分析网页链接特性的研究工作[1]。2002 年 Newman 提出一种层次聚类的方法，第一次从理论上解决了这个问题[2]。

自从社区发现问题提出以来，几乎所有的社区发现算法针对的都是网络的整体拓扑结构，计算结果也是网络的所有社区划分结果[2-4]，这些算法的计算复杂度都比较高，对一个 m 条边，n 个节点的网络，GN 算法的计算复杂度是 $O(m^2n)$[2]，文献[5]的计算复杂度是 $O(n(m+n))$，文献[6]的计算复杂度是 $O(n^3)$。这些基于网络整体结构的社区发现算法在处理一些只涉及网络局部特性的问题，如信息检索、人际关系时，具有以下局限性：1) 在分析网络的局部关系时，不需要分析整个网络的关系，否则当网络数据集较大时，分析效率较低。如分析某个人的社会关系时，没有必要考虑全人类的交互信息；2) 在一些情况下，网络不能获取完整的数据，如 Internet 网络、社会关系网络，这也限制了这些算法的应用。为此需要一种只考虑网络局部拓扑结构的社区发现方法，即局部社区发现方法。

局部社区发现是指定一个种子节点，然后挖掘出该节点所属的自然社区，而不管网络中其他节点的拓扑关系，网络中剩余的节点拓扑结构将视为一个“黑盒子”。由于局部社区发现具有很强的实际应用价值，目前国际上已有学者对局部社区发现问题展开了研究。笔者基于改进流模型的思想，提出了基于聚簇优先遍历及二次切割方法的局部社区发现算法，用较低的复杂度解决了局部社区发现问题。

1　局部社区发现算法研究现状

2009 年 Fortunato 提出一种基于社区适应度(fitness)的社区评价方法，并利用该评价方法提出一种局部社区发现算法——ASJ 算法[7]，一个子图

收稿日期：2012-10-11；**修回日期**：2012-11-10
基金项目：国家高技术研究发展计划(863 计划)；主题项目(2011AA120300)；湖南省自然科学基金资助项目(11JJ4028)
作者简介：李星(1985—)，男，山西人，在读硕士，从事数据挖掘和信息处理研究。

G 的社区的适应度可定义为

$$f(G)=\frac{K_{\text{in}}^{G}}{(K_{\text{in}}^{G}+K_{\text{out}}^{G})^{\alpha}} \tag{1}$$

其中：K_{in}^{G} 为其子图的内度之和；K_{out}^{G} 为子图的外度之和；α 为一个控制因子。$f(G)$越大，此社区就越健壮。该算法的思想是将种子节点 v_{seed} 作为初始的社区 G，将使 $f(G)$增值最大的邻居节点加入到社区内，更新局部社区 G，并重新计算该局部社区的所有邻居节点的适应度。直到 G 所有的邻居节点都遍历过，且都不能找到使 $f(G)$提高的点，就认为 G 是节点 v_{seed} 的自然社区。该算法参数 α 的选择非常重要，参数较小的话会导致整个网络是一个局部社区。

2009 年 Papadopoulos 分析网络的拓扑结构，在介数中心性的基础上提出了边桥度的概念，一条边的桥度是衡量该边成为横跨社区之间边的概率，边桥度越大，这条边就更有可能成为社区之间的桥边。结合边桥度提出一种局部社区发现——SAA 算法[8]，边桥度的定义为

$$b_{\text{L}}(e_{\text{st}})=1-\frac{|N(s)\cap N(t)|}{\min[(d(s)-1),(d(t)-1)]} \tag{2}$$

其中，$N(s)$表示节点邻居节点集合，该算法的思路是指定一个门限，首先指定种子为初始社区，比较社区和其邻居节点之间边的桥度，当桥度小于门限时，就将该节点加入到社区，否则就不加入，重复这个过程，直到社区所有的邻居节点都不能加入，这时可以认为社区指向社区外的边都是桥边。

ASJ算法和SAA算法都不需考虑整个网络的拓扑结构，通过搜索邻居节点的方式即可获取种子节点的局部社区，但这 2 种算法都是通过局部网络拓扑结构来进行扩散计算，邻居节点的选择是随机的，这样其一跳节点的扩散会对结果造成一定的影响。如果先能对与种子节点相关联的邻居节点进行聚簇计算，然后再对聚簇的节点进行分析来提取局部社区，将可能避免上述的问题。笔者正是基于这种“聚簇优先”的思想，提出了一种局部社区发现算法。

2　局部社区发现算法

2.1　流模型的引入

Newman 在解决网络桥节点发现的问题时，引入了流模型的概念，即将网络视为一个电网，每条边视为一个电阻，利用节点的电压或电流来计算网络拓扑结构。该模型能较好地反映出网络的自然聚簇结构[9]，WU 引入这种思想提出了一种社区发现算法——WF 算法[4]。该算法通过随机指定 2 个点 $v_{\text{one}},v_{\text{zero}}$，分别作为 1 电极和 0 电极，然后计算网络中所有节点的电压值，电压靠接近 1 的那些点更有可能和 v_{one} 处在同一个社区，这一点同样适用于 0 电极。迭代计算网络中节点的电压值，将节点按照电压值进行降序排序得出节点的排序结果 $\boldsymbol{\Phi}$，向量 $\boldsymbol{\Phi}$ 就是与节点 v_{one} 处在一个社区并且与节点不在一个社区的概率排序。这样将 $\boldsymbol{\Phi}$ 中前面的节点截取出来，就可以得到节点的自然社区。由于社区内部边密度较大，社区之间边密度较小[1-2,7]，那么同一个社区内的节点，电压值变化较小，社区之间节点的电压值变化较大。搜索 $\boldsymbol{\Phi}$ 中节点电压变化较大的点，进行截断，就可以进行社区划分。

WF 算法存在一定缺陷：1) 精确求解节点的电压值收敛较慢，对于只有 34 个节点的 karat club 网络，电压精确到 0.01 时，需要迭代 200 次以上；2) 划分社区需要依赖于难以获取的先验知识(社区规模)；3) 需要随机选取，如果处在一个社区，这个社区就会被强行拆开，因此需要多次选取统计才能进行社区划分；4) WF 算法可以应用在局部社区发现，但仍然需要整个网络的拓扑结构。

2.2　基于改进流模型的聚簇优先节点遍历

流模型能够较好地反映出节点的自然聚簇结构，但是受选取的影响，一个结构紧密的自然聚簇，可能被强行分开；因此，在算法的开始，笔者默认除种子节点外的其他节点电压值都为 0。当考虑加权网络时，$R_{ij}=1/\omega_{ij}$，ω_{ij} 为边的权重；当考虑非加权网络时，所有的电阻为 1，霍夫曼指出电路中每一个节点流入的电流和流出的电流是相等的；因此，网络中节点的电压值满足

$$\text{Vot}_i=\frac{1}{K_i}\sum_{e_i\in\text{nab}(e_i)}\text{Vot}_j \tag{3}$$

其中 K_i 为节点的度，可用迭代的方式计算网络中节点的电压值，用矩阵的方式表达上述迭代过程为

$$\text{Vot}^{l}=B\text{Vot}^{l-1}\quad \text{Vot}^{l}(v_{\text{one}})=1\quad B_{ij}=A_{ij}/\sum\nolimits_{j}A_{ij} \tag{4}$$

其中 Vot^{l} 为第 l 次迭代各个节点的电压向量 $\text{Vot}^{0}=(0,0...1...0)$，$\boldsymbol{A}$ 为网络的邻接矩阵。利用式 (4) 进行迭代，那么迭代的结果按照电压进行降序排序，得到的节点 $\boldsymbol{\Phi}$ 序列就是网络中任意一个节点与 v_{one} 的亲密程度排名，这样与 v_{one} 处在一个社区的节点，其电压值就会大于不在一个社区的节点。式 (3) 可以理解为：将当前网络电流离散化，每一次迭代结点根据其邻居节点的电压更新当前的电压。但这就导致了只要迭代次数足够长，所有节点的电压都为 1。经过试验论证，在迭代的过程中，节点电压排序结果 $\boldsymbol{\Phi}$ 的收敛速度是很快的，即使是 3 000 个节点

的网络也只需要迭代 15 次，其节点电压的顺序 $\boldsymbol{\Phi}$ 已经稳定。因此笔者可以设计一种新的局部社区划分准则，只利用节点的排序结果，而不依赖于先验知识来进行节点截断，即可提取局部社区，这样既利用了该模型的优点，又降低了算法的复杂度，就不必精确求解节点的电压值了。

2.3 局部社区评价

评价社区优劣普遍采用 Q 函数[10]，但 Q 函数是针对网络整体的社区划分目标函数，如果用在局部社区分割，最佳的切割往往是网络的二分[11]。在无法获取整个网络数据的前提下，很难评价局部社区。社区是图内与系统其他部分联系很少的局部子图。在某种程度上，社区可以看作是具有自主性的相互分离的实体。因此将它们独立于整个图之外进行评估具有重要意义。由于社区的定义是，社区内部边的密度大于社区边的密度；因此，笔者可以使用下列准则来判断一个局部社区的优劣：

$$\psi(S)=\sum_{\mathrm{E}}\omega_{ij}/\Omega(S) \tag{5}$$

其中 E 为节点集合 S 的内部边集合，$\Omega(s)=\sum_j\sum_i\omega_{ij},v_i,v_j\in S$ 表示节点集合的边权值之和。这样 $\psi(s)$ 值越大，社区边的密度或者边权重的密度就越大，这个社区就越紧凑。结合聚簇优先遍历的优点以及局部社区评价准则，指定种子节点 v_{seed}，计算 v_{seed} 聚簇优先遍历向量 $\boldsymbol{\Phi}$，依次将 $\boldsymbol{\Phi}$ 中的前 i 个节点取出，作为一个社区，计算其社区积分 $\psi(i)$，当取的节点和 v_{seed} 在同一个社区内，当前的社区会越来紧凑，其 $\psi(i)$ 值将会不断增大，当取到其他社区的节点时，社区的质量开始变差，$\psi(i)$ 值将会减小，这时 $\psi(i)$ 值将会出现一个明显的峰值（随机网络中没有聚簇结果，不会出现这种情况），如果在峰值处进行截断，前面的节点就是 v_{seed} 节点的社区覆盖范围。

2.4 局部社区发现算法

指定种子节点 v_{seed} 后，将该点的电压初始化为 1，根据式 (4) 计算种子节点的按聚簇优先遍历的结果 $\boldsymbol{\Phi}$（式 (4) 只是矩阵的表达形式，实际计算的时候只考虑电压不为 0 的节点的邻居节点）。每一次迭代都会更新到 v_{seed} 的下一跳节点，电压的更新过程可以看作是一个种子节点电压的局部扩散过程。由于 v_{seed} 的局部社区结构只占网络中的很小一部分，因而不必计算所有节点电压，只需计算种子节点局部拓扑结构的电压值，就可以进行社区发现。因此，可以指定一个计算范围 N，在迭代过程中，被覆盖的节点，其电压值就会大于 0，这样当电压值大于 0 的节点数大于 N 时，将这部分节点按照电压进行降序排序，将前 N 个节点截取出来，形成一个子网络。由于自然网络的直径很小，因此只需要迭代较少的次数就可以进行截取，但由于这时节点的序列还没有收敛，因此在子网络上根据式 (4) 完成剩余的迭代。上述思想为本文算法对网络的第 1 次切割，该过程可以视为对网络局部拓扑结构的一次采样，在该子图上进行计算社区积分，寻找最优的局部社区结构，进行第 2 次切割，就可以挖掘出种子节点的局部社区结构，这样只需要利用一部分数据，就可以实现局部社区发现。

结合社区评价准则，笔者提出一种基于聚簇优先的局部社区挖掘算法，算法流程如下。

输入：网络 E={V,E}，搜索数据大小 N，迭代次数 l，种子节点 v_{seed}。

输出：局部社区 G''。

1) 根据种子节点初始化网络电压。

2) 迭代计算节点的电压。

3) 判断当前电压大于 0 的节点是否超过 N，如果没有，转 2)。

4) 根据节点的电压值进行降序排序，截取前 N 个节点，组成网络 G'。

5) 在网络 G'中继续迭代计算每个节点电压值。

6) 根据节点的电压降序排序，得出排序结果。

7) 遍历 $\boldsymbol{\Phi}$，不断将前 i 个节点加入到集合 G''，计算当前节点集合的社区积分 $\psi(i)$，依据下列准则，判断 $\psi(i)$ 是不是一个峰值

$$\psi(i)=\max(\psi(i-4:i+4))$$

$$C\leftarrow i$$

8) 在峰值集合 C 中按照下列准则寻找最优值，即最尖锐的一个峰

$$i=\arg\max(\psi(i-4+\psi(i+4)-2\psi(i))$$

9) 将 $\boldsymbol{\Phi}$ 中前 i 个节点就是局部社区 G''。

如果遍历 N 个节点，还没找到峰值时，就将当前的 N 个节点作为局部社区输出，这时 2 次切割的位置相同。其物理意义是指定的 N 小于局部社区的自然聚簇大小，因为聚簇优先遍历思想就是将更有可能和种子节点处在一个社区的节点排在前面，因此只将结果中与最亲密的 N 个节点作为结果输出。当局部社区本身规模较大，或用户不关注局部社区的全部的节点信息情况下，这一点是可行的。

3　算法分析

为了证明上述算法的可靠性，笔者使用网络分析基准数据 football 网络[12]进行试验。该网络反应了美国高校橄榄球联赛 2000 赛季的对阵情况。网络节点表示参赛的橄榄球队，边表示 2 支球队在常规赛季进行过比赛。根据地理位置，联盟中的全部 115 支球队被分成 12 个联盟。根据赛程安排，同一联盟内部球队的比赛比和其他联盟球队之间的比赛更为频繁。该网络具有明显的网络社区结构，是验证社区挖掘算法的常用基准网络。

参数设置为 $v_{\text{seed}}=32$，$l=10$，$N=40$。执行 3 次迭代后进行第 1 次截断，在截断后的网络上完成剩余的 7 次迭代，最后得到节点按聚簇优先遍历结果 $\boldsymbol{\Phi}$。由于篇幅有限，只给出向量 $\boldsymbol{\Phi}$ 一部分结果。$\boldsymbol{\Phi}$={32,39,100,47,13,64,106,15,6,60,2}。根据向量 $\boldsymbol{\Phi}$ 以及社区积分函数，计算 ψ 值寻找最优值进行截断，最后节点 32 的自然社区节点为 32,39,100,47,13,64,106,15,6,20,2，这个划分结果与文献[2−4,9]中的社区划分结果相同。截断后的子网络上最终的局部社区划分结果如图 1 所示。

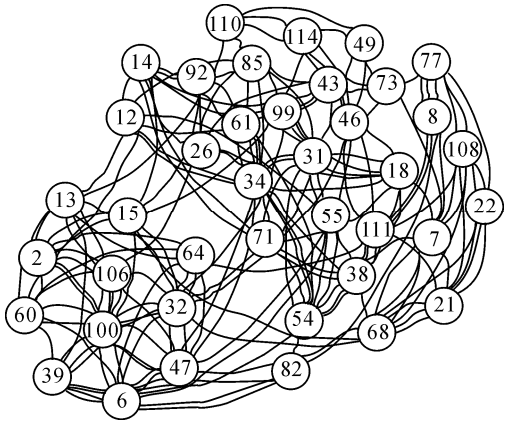


图 1　局部社区发现结果

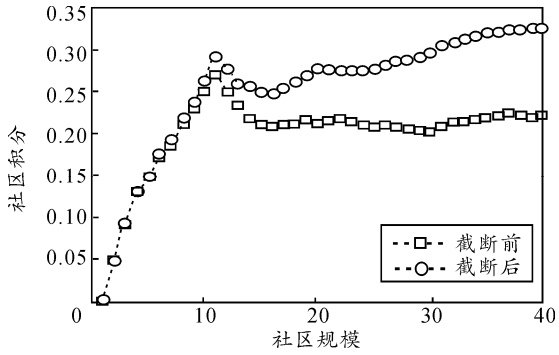


图 2　截断前后 ψ 值分布

为了分析第 1 次截断对 ψ 值的影响，笔者比较截断和不截断 ψ 值分布，为了公平起见，只和截断前 ψ 值的 40 个节点进行比较，试验结果如图 2 所示。从图中可以看出，尽管进行了截断，但并没有影响峰值的位置，因此对原始网络进行截断后仍然能进行正确的局部社区划分。

3.1　复杂度分析

利用 2 次切割的思想迭代计算局部拓扑结构的电压值，其复杂度是 Nl，节点排序的复杂度为 N[13]。同时由于第 2 次的排序结果和第 1 次结果相差不大，也就是说第 2 次排序是在一个顺序已经基本确定的序列上进行的，第 2 次切割的复杂度为 c（c 为社区自然规模大小），这样整体的复杂度为 $O(Nl+N+c)$。

3.2　算法性能分析

首先分析参数 l 对算法结果的影响。图 3 中，x 轴表示迭代次数，y 轴表示两相邻 2 次迭代所对应的结点序列的差异。若 l 次迭代结点序列为 $\boldsymbol{\Phi}^{(l)}$，$l-1$ 次迭代结点序列为 $\boldsymbol{\Phi}^{(l-1)}$，那么 y 轴的函数值为 $y(l)=N(\{i\,|\,\boldsymbol{\Phi}^{(l)}(i)\neq\boldsymbol{\Phi}^{(l-1)}(i),i\in(1,n)\})$，其中，$N()$ 表示集合中元素的个数。为了验证 l 的取值对节点排序的影响，笔者随机生成 6 个不同规模的网络，网络规模分别为(500 1 000 1 500 2 000 2 500 3 000)，网络的边数为节点数的 10 倍，同时网络均为连通网络。测试 6 个不同网络在不同的迭代次数的排序结果收敛情况，试验结果如图 3 所示。从图 3 可以看出，对 3 000 个节点，30 000 条边的随机网络，也能在 15 步内收敛。由于自然网络具有小世界特性，其网络直径的随节点个数的增加是非常缓慢的[14]；因此，同等规模的自然网络，其收敛更快。在试验过程中发现迭代次数的经验值可取为 $5\log(n)$。

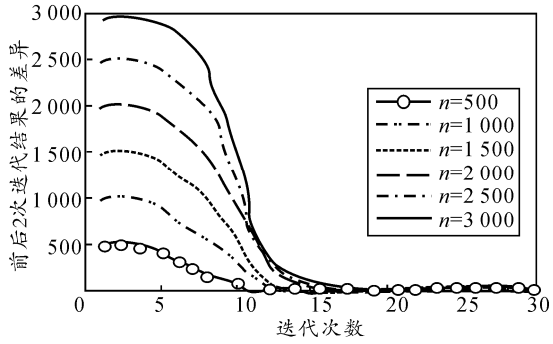
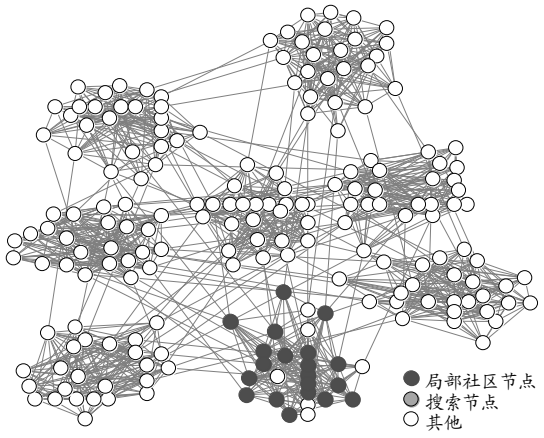
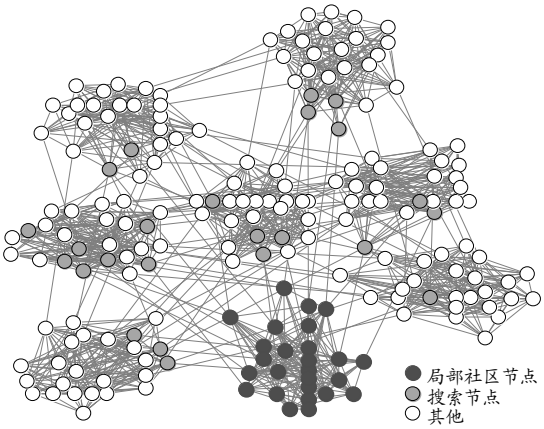


图 3　l 收敛分布

为了分析 N 对社区发现结果的影响，笔者生成一组网络，该网络 n=200，m=1 500；存在 8 个社区，社区内度的比例为 c=0.9。v_{seed} 所属社区的规模为 29，整个网络社区的规模设置为 20～40。分别取 N=20,50，计算结果如图 4 所示，分别用不同色阶的点分别表示搜索节点和局部社区节点。



(a) N=20



(b) N=50

图 4　局部社区搜索示意图

从图 4(a)可以看出，当 N 取太小时，不能很好地发现局部社区，只返回了局部社区中 v_{seed} 较近的点，图 4(b)中当 N=50 的时候，能够很好地发现该局部社区结构；因此，N 值没有必要取太大。局部社区发现的原则是将整体的网络数据视为一个“黑盒子”，只关心局部的聚簇结构。如果要寻找到 ψ 值分布的峰值，N 值不能取太小，只有 N 大于种子节点的自然社区归属，才能搜索到种子节点自然社区之外的节点，有这些节点才能“衬托”出种子节点的自然社区，如图 4 所示。N 的取值虽然为用户指定，但是从结果可以看出，只要 N大于局部聚簇规模的 2 倍，结果就对 N不敏感，因此 N不能算做一种先验知识。

4　结束语

笔者利用二次切割的思想，结合新的局部社区评价准则提出一种局部社区发现算法。通过网络基准数据进行测试，证明了算法的可行性。和同类算法比较，该算法在性能上有了一定的改进，并有以下特点：1) 利用网络部分数据实现局部社区发现；2) 不依赖于先验知识。该算法虽然利用了部分网络数据进行局部社区发现，但在 2 次截断时，还是利用了一些不必要的节点信息，同时，在绘制社区积分分布图时，其峰值的识别只采用了一种比较简单的形式，即连续 2 个节点值降低并且连续 2 个节点值升高，并没有考虑更为严格的判决，这些问题将在下一步的工作中进行修正。

参考文献：

[1] Gibson D, Kleinberg J, Raghavan P. Inferring web communities from link topology[C]. In HYPERTEXT’98: Proceedings of the ninth ACM conference on Hypertext and hypermedia, New York, NY, USA, ACM, 1998: 225–234.

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

[3] Fortunato S. Community detection in graphs[J]. Phys. Rep, 2010, 486(3–5): 75–174.

[4] Wu F, Huberman BA. Finding communities in linear time.A Physics approaeh[J]. European Physieal Journal B, 2004, 38(2): 331−338.

[5] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 69:066133, 2004.

[6] Haijun Zhou, Reinhard Lipowsky. Network Brownian Motion: A New Method to Measure Vertex-Vertex Proximity and to Identify Communities and Subcommunities[M]. ICCS 2004, LNCS 3038, 2004: 1062–1069.

[7] Lancichinetti A, Fortunato S, Kertész J. Detecting the verlapping and hierarchical community structure in complex etworks[J]. New Journal of Physics, 2009, 11(3): 033015.

[8] Papadopoulos S, Skusa A, Vakali A, et al. Bridge bounding: A local approach for efcient community discovery in complex networks[R]. Technical Report arXiv: 0902.0871, Feb 2009.

[9] Newman MEJ. A measure of betweenness centrality based on random w[J]. arXiv:cond-mat/0309045v1 [cond-mat.stat-mech] 1 Sep 2003.

[10] James P. Bagrow, Erik, Bollt M. A Local Method for Detecting Communities[J]. arXiv: cond-mat/0412482v2 [cond-mat.dis-nn] 23 Mar 2005.

[11] Jorg Reichardt and Stefan Bornholdt.Statistical Mechanics of Community Detection[J]. arXiv: cond-mat/ 0603718v1 [cond-mat.dis-nn] 27 Mar 2006.

[12] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc. Natl. Acad. Sci. USA 99, 8271-8276 (2002).

[13] Cormen T H, Leiserson C E, Rivest R L, et al. Introduction to algorithms[J]. 2nd Ed., p. 189, p. 168 [11] M. E. J. Newman, and M. Girvan. Finding and evaluating community structure in networks. PHYSICAL REVIEW E 69, 026113 ~2004.

[14] Matthieu Latapy, Clemence Magnien. Measuring Fundamental Properties of Real-World Complex Networks[J]. arXiv:cs / 0609115v2 [cs.NI] 1 Mar 2007.