第 43 卷 第 6A 期
2016 年 6 月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

近似线性时间的社团结构动态演化挖掘算法

任泺锟 李慧嘉 贾传亮

(中央财经大学管理科学与工程学院 北京 100081)

摘 要 探测网络社团结构对于分析、设计复杂的自然或工程网络至关重要,然而现有的探测技术主要依托于最优化和启发式算法,不能兼顾计算效率和准确性。因此提出了一种基于演化迭代技术的动态社团探测算法,它能准确高效地发现网络中的社团结构。首先引入了一个离散时间的动态系统,通过描述社团划分收敛到特定指标最优的演化轨迹来确定社团划分。接着提出了一个一般化的指标函数,以确定网络中最优的社团数量及最稳定的社团结构。该指标函数极具概括性,改变相应的参数即可引申到各种已广泛应用的指标函数。针对参数选择的困难,利用图生成模型自动确定社团划分的指标函数。此算法效率很高,计算复杂度与稀疏网络中的节点数量呈近似线性关系。最后,在人工和真实网络中进行了大量的仿真实验来测试算法表现,结果显示所提算法能够揭示很多有价值的信息。

关键词 社团挖掘,演化计算,动态迭代系统,近似线性时间

中图法分类号 TP393 文献标识码 A

Near Linear Time Community Detection Algorithm Based on Dynamical Evolution

REN Luo-kun LI Hui-jia JIA Chuan-liang

(School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100081, China)

Abstract Detecting communities is is crucial for analyzing and designing complicated natural and engineering network. The existing community detection algorithms rely heavily on optimization and heuristic methods, which can not balance computational efficiency and accuracy simultaneously. Thus we proposed an evolutionary algorithm which uses a new dynamical system based on community membership vector to formulate the conditions driving the convergence of dynamics trajectory. Then, we proposed a quality function, which can unify the conventional algorithms by selecting appropriate parameters. Furthermore, considering the difficulty in choosing parameters, we established a graph generative model according to the network prior information, by which the optimum formalism of the quality function can be obtained automatically. Our algorithm is highly efficient and the computational complexity is nearly linear with the number of all nodes in a sparse network. Finally, extensive experiments were performed in both artificial and real networks, which reveal much useful information.

Keywords Community detection, Evolutionary computation, Dynamical iterative systems, Near linear time

1 引言

复杂网络可以模拟许多大规模的社会和生物系统,其中网络中的节点代表个体,边代表个体之间的联系[1,2],如因特网、万维网、智能电网、蛋白质交互网络、微博关联网和科学家合作网等[3-5]。社团结构作为网络的一个重要拓扑特征,将网络划分成若干个团内边密度较高的子图[6,7]。一般相同社团内的节点有共同的性质和相近的关系,精确划分社团结构对系统的控制、优化和协同有重要意义。

近年来涌现出了多种挖掘社团结构的算法,其中模块度函数是一种普遍的量化社团划分质量的方法,其函数值的增加代表了社团划分质量的提高。另外,Potts 模型在社团结构分析中的应用也非常广泛,它将自旋状态表示为社团标号,并且利用系统能量值来衡量社团划分的质量。假设 A 是 G 的邻接矩阵,σ_i 是节点 i 的社团标志,考虑函数(当 $\sigma_i=\sigma_j$ 时有 $(\delta(\sigma_i,\sigma_j)=1)$,否则 $(\delta(\sigma_i,\sigma_j)=0)$)。Reichardt 和 Bornholdt[14] 提出了一个一般化的汉密尔顿函数,

$$H_{RB}(\{\sigma\})=-\frac{1}{2}\sum_{i\neq j}(A_{ij}-\gamma p_{ij})\delta(\delta_i,\delta_j) \tag{1}$$

其中,σ 是社团成员标志的集合,$\gamma, p_{ij}\in R$。

近些年虽然已有不少经典的探测方法[8-10],但是鲜有能够快速而准确地发现社团结构的方法,更不用说在大规模网络上所需的线性时间。为了更好地探测现实世界中的社团结构,本文首先提出了一种新颖的离散时间的动态系统,它可以通过最优化特定指标函数得到相应的社团划分,该划分是动态系统轨迹的唯一的稳定值。在一定条件下,动态系统收敛得到的社团就是最优化分的结果。接下来,在整合现有指标函数的基础上,给出了指标函数的一般化形式。这样,只需选择不同的参数即可引申到不同的指标函数。进而,针对参数

本文受国家自然科学基金资助项目(71401194,91324203,11131009)资助。

任泺锟(1995—),女,硕士生,主要研究方向为复杂网络;李慧嘉(1985—),男,博士,讲师,主要研究方向为社会网络、数据挖掘、运筹学,E-mail:Hjli@amss.ac.cn;贾传亮(1979—),男,博士,副教授,主要研究方向为管理决策分析、运筹学。

不易选择的困难，根据图生成模型，利用网络先验信息，可以使动态系统自动找出最优划分。该算法十分高效，其时间复杂度在稀疏网络上为 $O(n)$，其中 n 是网络中包含节点的个数。最后，将算法应用到人工网络和现实世界网络中，结果显示此算法不仅拥有非常好的性能，还能够揭示很多有用的信息，如层次结构[6]和社团交互模式信息[19,24]等。

2 符号与概念

针对社团探测的具体问题，首先给出一些基本的概念：

(1) c,n,m 分别表示社团、节点和边的数量。

(2) l_{out}^{in} 和 l_{in}^{out} 分别表示社团 μ 团内边和团间边的数量。

(3) p_μ^{in} 表示社团 μ 团内边的数量与所有可能存在的边的数量的比值。p_μ^{out} 表示社团 μ 团间边的数量与所有可能存在的边的数量的比值。

容易得出

$$p_\mu^{in}=\frac{2l_\mu^{in}}{n_\mu(n_\mu-1)},p_\mu^{out}=\frac{2l_\mu^{out}}{n_\mu(n_\mu-1)} \tag{2}$$

为了开展下一步的分析，引入向量 $X_i(t)=[X_{i\mu}(t)],\mu=\langle 1,2,\cdots,c\rangle$，其中 $X_{i\mu}(t)$ 表示在时刻为 t 时，节点 i 属于社团 μ 的概率。运用 $X(t),n_\mu,l_\mu^{in},l_\mu^{out},p_\mu^{in},p_\mu^{out}$ 可以重写为：

$$n_\mu(t)=\sum_i x_{i\mu}(t) \tag{3}$$

$$l_\mu^{in}=\frac{1}{2}\sum_i\sum_{j\neq i}x_{i\mu}(t)x_{i\mu}(t)A_{ij} \tag{4}$$

$$l_\mu^{out}=\sum_i\sum_{j\neq i}x_{i\mu}(t)(1-x_{j\mu}(t))A_{ij} \tag{5}$$

$$p_\mu^{in}=\frac{\sum_i\sum_{j\neq i}x_{i\mu}(t)x_{j\mu}(t)A_{ij}}{\sum_i\sum_{j\neq i}x_{i\mu}(t)x_{j\mu}(t)} \tag{6}$$

$$p_\mu^{out}=\frac{\sum_i\sum_{j\neq i}x_{i\mu}(t)(1-x_{j\mu}(t))A_{ij}}{\sum_i\sum_{j\neq i}x_{i\mu}(t)(1-x_{j\mu}(t))} \tag{7}$$

另外，还可以针对特定社团结构定义社团交互强度 C_{ks} 和社团重要性 I_μ：

(1) C_{ks} 定义为 k,s 两个社团之间相互连边的比例。

(2) I_μ 定义为特定社团 μ 与其他每个社团交互强度的加和，$I_\mu=\sum_s C_{\mu s}$。

本文提出了一种新颖的动态系统来寻找最优的社团归属 $X(t)$（用不同的灰度代替），指标函数 H 用来衡量 t 时刻的社团归属 $X(t)$ 的优劣。这里利用动态系统 $X(t+1)=f(X(t))$，可以得到 $X(t)$ 的固定的最优结果，如图 1 所示。

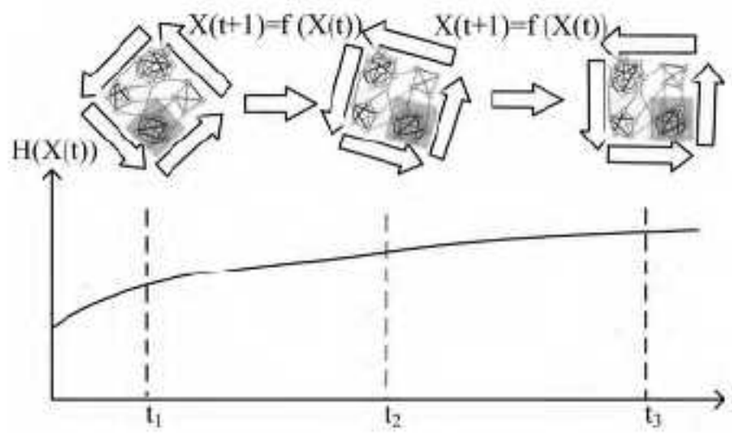


图 1

3 基于线性迭代的动态系统

为了在特定网络中进行社团划分，首先提出一种新型的动态演化系统：

$$x_{i\mu}(t+1)=\frac{f(x_{i\mu}(t))e^{Q_{i\mu}(x_{*\mu}(t))}}{\sum_{\tilde{\mu}}f(x_{i\tilde{\mu}}(t))e^{Q_{i\tilde{\mu}}(x_{*\tilde{\mu}}(t))}} \tag{8}$$

其中 $Q_{i\mu}(x_{*\mu}(t+1))$ 是节点 i 归属于社团 μ 的质量函数，$x_{*\mu}=(x_{i\mu})$，$f(x_{i\mu})=a\cdot x_{i\mu},a\neq 0\in R$，可以得到以下定理。

定理 1　如果 $X(t)$ 代表硬社团归属矩阵，假设 μ_i^* 为 $x_{i\mu_i^*}(t)=1$ 时的社团。那么当

$$\forall i,\mu\neq\mu_i^*,Q_{i\mu}(t)<Q_{i\mu_i^*}(t) \tag{9}$$

动态系统(9)在 $X(t)$ 有一渐进的稳定点。

证明：对所有的 μ 和 $\tilde{\mu}$，网络的 Jacobian 矩阵 J_i 为一个包含 $\frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu}(t)}$ 的 $K\times K$ 的矩阵块。接着，通过计算动态系统式(8)的 $n\times K$ 个状态变量和相应的导数，并运用合适的排列和替换，J_{ii} 在 $X(t)$ 的值为：

$$J_{ii}=\begin{bmatrix}0 & -\frac{e^{Q_{i1}(t)}}{e^{Q_{i\mu_i^*}(t)}} & \cdots & -\frac{e^{Q_{ik}(t)}}{e^{Q_{i\mu_i^*}(t)}}\\ 0 & \frac{e^{Q_{i2}(t)}}{e^{Q_{i\mu_i^*}(t)}} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{e^{Q_{ik}(t)}}{e^{Q_{i\mu_i^*}(t)}}\end{bmatrix}\Bigg|_{X(t)} \tag{10}$$

硬社团归属矩阵的稳定点的 Jacobian 矩阵可以由以下矩阵表示：

$$J=\begin{bmatrix}J_{11} & 0 & 0\\ 0 & J_{\tilde{\mu}} & 0\\ 0 & 0 & J_{nn}\end{bmatrix}\Bigg|_{X(t)} \tag{11}$$

很显然，其特征值为：

$$\lambda_{i\mu_i^*}=0,\lambda_{i\mu}=\frac{e^{Q_{i\mu}(t)}}{e^{Q_{i\mu_i^*}(t)}},i=1,2,\cdots,n,\mu\neq\mu_i^* \tag{12}$$

如果对 $\forall i,\mu$ 都有 $|\lambda_{i\mu}|<1$，即：

$$\forall i,\mu\neq\mu_i^*,|\lambda_{i\mu}|=|\frac{e^{Q_{i\mu}(t)}}{e^{Q_{i\mu_i^*}(t)}}|<1\Rightarrow Q_{i\mu}(t)<Q_{i\mu_i^*}(t) \tag{13}$$

那么对任一 $\lambda_{i\mu^*}$，动态系统在 $X(t)$ 上都有一个渐进式的稳定点。

证毕。

定理 1 直观地说明了我们可以利用迭代动态系统式(9)使得函数值 $Q_{i\mu}(t)$ 达到最大，并得到相应的最优的归属矩阵 X，如图 1 所示，动态系统式(9)会收敛到代表最优划分的稳定点。

接下来，算法的任务简化为设计一种高效的社团探测指标。通过研究发现，对于如模块度和 Potts 模型等众多方法，目标函数有下列一般形式：

$$E(t)=-\frac{1}{2}\sum_\mu\sum_{i=1}^n(\sum_{j=1}^n f_\mu^+A_{ij}x_{i\mu}(t)x_{j\mu}(t)-\sum_{j=1}^n f_\mu^-(1-A_{ij})x_{i\mu}(t)x_{j\mu}(t)+\sum R_\mu \tag{14}$$

事实上，不同的目标函数，如 Hofman and Wiggins 模型[26]和 Ronhovde and Nussinov 模型[28]，可以通过选择不同的奖励参数 f_μ^+ 和惩罚参数 f_μ^- 来体现。本文引入一个新颖的一般化方程 Q：

$$Q_{i\mu}(t)=\sum_{j=1}^n f_\mu^+A_{ij}x_{j\mu}(t)-\sum_{j=1}^n f_\mu^-(1-A_{ij})x_{j\mu}(t)+R_{i\mu} \tag{15}$$

可以选择参数 $R_{i\mu}$，使 $\frac{\partial R_{i\mu}}{\partial x_{i\mu}(t)}=0,R_{mu}=\sum_{i=1}^n R_{i\mu}$，例如 $R_{i\mu}=\frac{2}{l_\mu}R_{mu}$。通过选择不同的参数，式(15)可以引申到很多著名的指标函数，如下所示。

(1) Hofman and wiggins 模型[26]

$$f_\mu^+=\log\frac{p^{in}}{p^{out}},f_\mu^-=\log\frac{1-p^{out}}{1-p^{in}},R_\mu=l_\mu\log\pi_\mu \tag{16}$$

(2) Ronhovde and Nussinov 模型[28]

$$f_\mu^+=1, f_\mu^-=\min\ p_{in,\mu}, R_\mu=0 \tag{17}$$

(3) RB Potts 模型[14]

$$f_\mu^+=1-\gamma_{RB}p, f_\mu^-=\gamma_{RB}p, R_\mu=0 \tag{18}$$

4 自动确定函数 Q 的社团检测

基于前面的分析，指标函数式(15)中 $Q_{i_\mu}(t)$ 的系数必须提前给定。出于此目的，Hofman 和 Wiggins 给出的一个基于贝叶斯分析的理论模型[26]中将函数的这些系数设定为：

$$f_\mu^+=\log\frac{p^{in}}{p^{out}}, f_\mu^-=\log\frac{1-p^{out}}{1-p^{in}}, R_\mu=l_\mu\log\ \pi_\mu \tag{19}$$

然而该工作必须假定网络中的社团有同样的比率，例如，$p_1^{in}=p_{12}^{in}=\cdots=p_c^{in}$ 和 $p_1^{out}=p_{12}^{out}=\cdots=p_c^{out}$。这显然对大多数网络都不是有效的。

针对上述缺点，假定网络中的每一个社团都有自己的概率特性，并提出了一个具有更高兼容性的模型。进一步，本文提出了一个关联图 G 和社团结构 $\{q\}$ 的似然函数：

$$\begin{aligned}P&=p(G|\{q\})\\&=\prod_\kappa[\overbrace{(p_\kappa^{in})(\sum_{i>j,i,j\in\kappa}A_{ij})}^{\mathrm{I}}\cdot\overbrace{(1-p_\kappa^{in})(\sum_{i>j,i,j\in\kappa}J_{ij})}^{\mathrm{II}}\cdot\\&\overbrace{(p_\kappa^{out})(\sum_{i>j,i,j\in\kappa}A_{ij}-\sum_{i>j,i,j\in\kappa}A_{ij})}^{\mathrm{III}}\cdot\\&\overbrace{(1-p_\kappa^{out})(\sum_{i>j,i,j\in\kappa}J_{ij}-\sum_{i>j,i,j\in\kappa}J_{ij})}^{\mathrm{IV}}\cdot\overbrace{\pi_\kappa^{n_\kappa}}^{\mathrm{V}}]\end{aligned} \tag{20}$$

这里 $J_{ij}=1-A_{ij}$，π_κ 是节点 k 分配到社团 μ 中的概率，例如

$$\pi_\kappa=\frac{n_\kappa}{n} \tag{21}$$

在式(20)中，项 I(III) 对应社团内(间)现存的链接，项 II(IV) 与社团内(间)缺少的链接相关。项 V 依据每个社团中节点的数量定义了图的划分。可以将式(20)重新改写为：

$$\begin{aligned}P=\prod_\kappa[&(\frac{p_\kappa^{in}}{p_\kappa^{out}})^{\sum_{i>j,i,j\in\kappa}A_{ij}}(\frac{1-p_\kappa^{in}}{1-p_\kappa^{out}})^{\sum_{i>j,i,j\in\kappa}J_{ij}}(p_\kappa^{out})^{\sum_{i>j,i,j\in\kappa}A_{ij}}\\&(1-p_\kappa^{out})^{\sum_{i>j,i,j\in\kappa}J_{ij}}\pi_\kappa^{n_\kappa}]\end{aligned} \tag{22}$$

在 t 时刻运用 $X(t)$，模型中的对数似然函数化为如下形式：

$$\begin{aligned}LP(t)&=\log P(t)\\&=\frac{1}{2}\sum_\kappa\{\sum_i\sum_{j\neq i}x_{i\kappa}(t)x_{j\kappa}(t)A_{ij}\log(\frac{p_\kappa^{in}}{p_\kappa^{out}})+\sum_i\sum_{j\neq i}x_{i\kappa}(t)\\&x_{j\kappa}(t)J_{ij}\log(\frac{1-p_\kappa^{in}}{1-p_\kappa^{out}})+\sum_i\sum_{j\neq i}x_{i\kappa}(t)A_{ij}\log(p_\kappa^{out})+\\&\sum_i\sum_{j\neq i}x_{i\kappa}(t)J_{ij}\log(1-p_\kappa^{out})+2n_\kappa(t)\log(\frac{n_\kappa}{n})\}\end{aligned} \tag{23}$$

当把式(23)映射到式(14)中时，可以看到除了式(23)的第一项和第二项分别是奖励项和惩罚项(通过一个负号表示)外，剩余的项均是正则项。另外引入了两个变量 γ_1 和 γ_2 来控制奖励项、惩罚项和正则项的贡献度：

$$\begin{aligned}Q(t)=\frac{1}{2}\sum_\kappa\{&\sum_i\sum_{j\neq i}\gamma_1x_{i\kappa}(t)x_{j\kappa}(t)A_{ij}\log(\frac{p_\kappa^{in}}{p_\kappa^{out}})+\sum_i\sum_{j\neq i}\gamma_2x_{i\kappa}(t)\\&x_{j\kappa}(t)J_{ij}\log(\frac{1-p_\kappa^{in}}{1-p_\kappa^{out}})-\sum_i\sum_{j\neq i}x_{i\kappa}(t)A_{ij}\log(p_\kappa^{out})+\\&\sum_i\sum_{j\neq i}x_{i\kappa}(t)J_{ij}\log(1-p_\kappa^{out})+\sum_i2x_{i\kappa}(t)\log(\frac{n_\kappa}{n})\}\end{aligned} \tag{24}$$

值得注意的是，式(24)的算法复杂度是 $O(n^2)$。经过变换得到：

$$\begin{aligned}Q(t)=\frac{1}{2}\sum_\kappa\{&\sum_i\sum_{j\neq i}x_{i\kappa}(t)x_{j\kappa}(t)A_{ij}(\log(\gamma_1\frac{p_\kappa^{in}}{p_\kappa^{out}})-\gamma_2\\&\log(\frac{1-p_\kappa^{in}}{1-p_\kappa^{out}}))+\gamma_2\log(\frac{1-p_\kappa^{in}}{1-p_\kappa^{out}})\sum_ix_{i\kappa}(t)\sum_{j\neq i}x_{j\kappa}(t)+\\&\sum_i\sum_{j\neq i}x_{i\kappa}(t)A_{ij}\log(\frac{p_\kappa^{out}}{1-p_\kappa^{out}})+\sum_i\sum_{j\neq i}x_{i\kappa}(t)\log(1-\\&p_\kappa^{out})+\sum_i2x_{i\kappa}(t)\log(\frac{n_\kappa}{n})\}\end{aligned} \tag{25}$$

式(25)的复杂度约为 $O(mc)$。在式(25)的基础上，可以计算出模型(15)中的系数为：

$$f_\mu^+=\gamma_1\log(\frac{p_\kappa^{in}}{p_\kappa^{out}}) \tag{26}$$

$$f_\mu^-=-\gamma_2\log(\frac{1-p_\kappa^{in}}{1-p_\kappa^{out}}) \tag{27}$$

$$R_\mu=\sum_i\sum_{j\neq i}\{A_{ij}\log(\frac{p_\kappa^{out}}{1-p_\kappa^{out}})+\log(1-p_\kappa^{out})+2\log(\frac{n_\kappa}{n})\} \tag{28}$$

将式(26)－式(28)代入式(25)，计算 $\frac{\partial Q(t)}{\partial x_{i_\mu}(t)}$：

$$\begin{aligned}\frac{\partial Q(t)}{\partial x_{i_\mu}(t)}=&\frac{1}{2}\gamma_1\{(2l_\mu^{in})\frac{\partial\log(\frac{p_\mu^{in}}{p_\mu^{out}})}{\partial x_{i_\mu}(t)}+2k_{i_\mu}\log(\frac{p_\mu^{in}}{p_\mu^{out}})\}+\frac{1}{2}\gamma_2\\&\{(\frac{2l_\mu^{in}(1-p_\mu^{in})}{p_\mu^{out}})\frac{\partial\log(\frac{1-p_\mu^{in}}{1-p_\mu^{out}})}{\partial x_{i_\mu}(t)}+2(n_\mu-k_{i_\mu}-\\&x_{i_\mu})\log(\frac{1-p_\mu^{in}}{1-p_\mu^{out}})\}+\frac{1}{2}\{(\sum_ak_{a_\mu})\frac{\partial\log(p_\mu^{out})}{\partial x_{i_\mu}(t)}+\\&k_i\log(p_\mu^{out})\}+\frac{1}{2}\{(n_\mu(n-1)-\sum_ak_{a_\mu})\\&\frac{\partial\log(1-p_\mu^{in})}{\partial x_{i_\mu}(t)}+(n-k_i-1)\log(1-p_\mu^{in})\}+\log\\&(\frac{n_\mu}{n})+1\end{aligned} \tag{29}$$

发现，当 $\gamma_1=\frac{1}{p_\mu^{in}}$ 和 $\gamma_2=1$ 时，有 $\frac{\partial Q_{ik}(t)}{\partial x_{i_\mu}(t)}=0$。将式(29)中的相关系数替换，得到

$$\begin{aligned}\frac{\partial Q(t)}{\partial x_{i_\mu}(t)}=&\frac{1}{2p_\mu^{in}}\{(2l_\mu^{in})\frac{\partial\log(\frac{p_\mu^{in}}{p_\mu^{out}})}{\partial x_{i_\mu}(t)}+2k_{i_\mu}\log(\frac{p_\mu^{in}}{p_\mu^{out}})\}+\\&\frac{1}{2}\{(\frac{2l_\mu^{in}(1-p_\mu^{in})}{p_\mu^{out}})\frac{\partial\log(\frac{1-p_\mu^{in}}{1-p_\mu^{out}})}{\partial x_{i_\mu}(t)}+2(n_\mu-k_{i_\mu}-\\&x_{i_\mu})\log(\frac{1-p_\mu^{in}}{1-p_\mu^{out}})\}+\frac{1}{2}\{(\sum_ak_{a_\mu})\frac{\partial\log(p_\mu^{out})}{\partial x_{i_\mu}(t)}+\\&k_i\log(p_\mu^{out})\}+\frac{1}{2}\{(n_\mu(n-1)-\sum_ak_{a_\mu})\\&\frac{\partial\log(1-p_\mu^{in})}{\partial x_{i_\mu}(t)}+(n-k_i-1)\log(1-p_\mu^{in})\}+\\&\log(\frac{n_\mu}{n})+1\end{aligned} \tag{30}$$

为了找到这样的社团，提出基于重复使用动态系统的迭代算法，其基本描述见算法 1。

算法 1　通过自动确定函数参数的社团检测算法

输入：网络 G，其节点数为 n，边数为 m；动态迭代的最大次数 T_{max}

输出：社团归属矩阵 X

1. 初始化 X(0)

重复

2. 使用式(3)－式(6)和式(7)分别更新 $n_\mu, l_\mu^{in}, l_\mu^{out}, p_\mu^{in}$ 和 p_μ^{out}；

3. 使用式(30)计算$\frac{\partial Q(t)}{\partial x_{i\mu}(t)}$;
4. 更新社团归属矩阵 $X(t)$;
5. 使用式(23)计算对数似然函数 $LP(t)$;
6. 如果达到动态迭代的最大值(T_{max}),则前往步骤 7;否则,回到步骤 1;
7. 考虑满足 $LP(t)$ 最大的函数 $X(t)$ 记为 X_1;
8. 如果 $LP(X_t)>LP_{best}$,令 $LP_{best}=LP(X_t)$ 并且令 $X_{best}=X$;
9. 如果达到步骤迭代的最大值(R_{max}),返回 X_{best};否则,回到步骤 1。

4.1 算法复杂度

在该算法每一次的迭代过程中,计算对数似然函数式(23)的时间复杂度是 $O(mc_{max})$,其中 c_{max} 是社团数量的最大值。动态系统的收敛速度非常快,算法总的复杂度为 $O(mc_{max})$。而对于稀疏网络,由于点的个数与边的数量呈线性关系,算法的复杂度下降为 $O(nc_{max})$。而对于一些社团个数较少的网络,如空手道俱乐部网络[27],算法复杂性则会降到 $O(n)$。在图 2 中,为了展示了本文算法的高效性,将其与一些著名算法比较,包括 Newman 快速算法(NF)[17]、CNM 算法[25]、Louvain 算法[11]、Danon 算法[23]、SA 算法[18]、Peixito 算法[16]和 Hoffman 算法[26]。

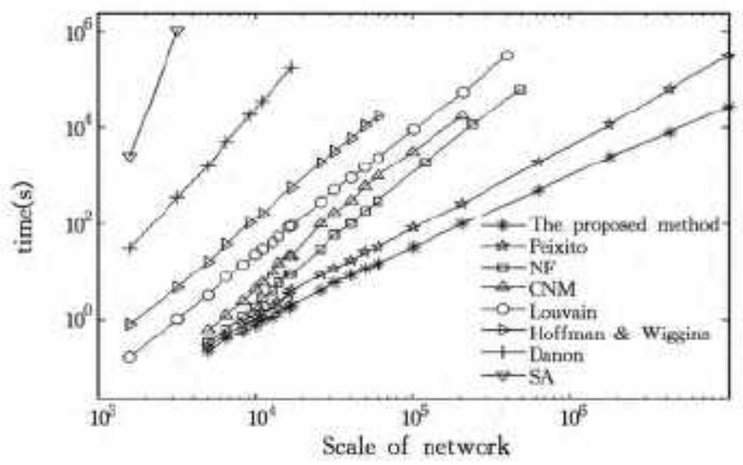


图 2 不同规模网络下各个算法运行时间的比较

在一台 2GHz CPU、4GB 内存、操作系统是 Windows 10 的台式电脑上运行此算法,程序计算软件是 Matlab 2012b。生成了一个具有 n 个节点和 $m=O(n)$ 条边的稀疏网络,并在拥有不同规模 n 的网络上运行算法。从图 2 可以观察到本文算法在任一规模为 n 的网络中的运行速度都优于其他 5 种算法,因此它可以方便地扩展到一些拥有数百万节点或者更大规模的实际网络中。

4.2 确定最优社团个数

尽管最优社团个数不是已知的先验信息,但是会极大地影响社团划分的结果。因此在运行算法之前,必须确定社团的数量。在许多算法中,社团数量被提前指定。例如,CNM[22]和 Louvain[11]算法考虑初始社团数量值为 n,而二分迭代式算法像 SA[16]和 DA[20]的初始值则为 2。这里利用经典的特征值理论,设 λ_μ 是网络拉普拉斯矩阵的第 μ 大特征值。从另一个角度来说,λ_μ 可以看成是社团强度的一种度量指标,特征值差 $\lambda_{\mu-1}-\lambda_\mu$ 可以认为是社团状态由 μ 转移($\mu-1$)的难度。可以通过计算最大的特征值差,得出最适合的社团个数 Λ 为:

$$\Lambda=\arg[\max_\mu(\lambda_{\mu-1}-\lambda_\mu)] \tag{31}$$

5 实验

本节分别在人工网络和现实网络上检验所提出的算法。结果表明,本算法可以有效且准确地发现复杂网络中的多尺度社团结构。

5.1 人工网络

本文用 GN 和 LFR 基准网络测试了此算法[12,21]。结果显示此算法即使在模糊簇状网络中也能有效探测社团。

在 GN 基准网络[13,17]中比较了不同算法的性能,如图 3 所示。可以看出,当团间边数目 $Z_{out}\leqslant 7.5$ 时,基本上所有算法都能找出正确的社团归属(互信信息值 $NMI\geqslant 0.9$)。但是随着团间边比例 Z_{out} 的增大,大多数算法无法找出正确的社团划分,而我们的算法直到 Z_{out} 接近 8 时仍然拥有最高的效率,验证了其划分的高效性。

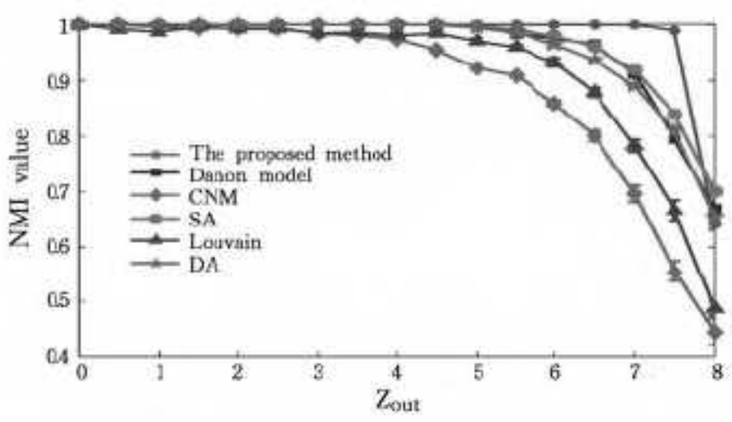
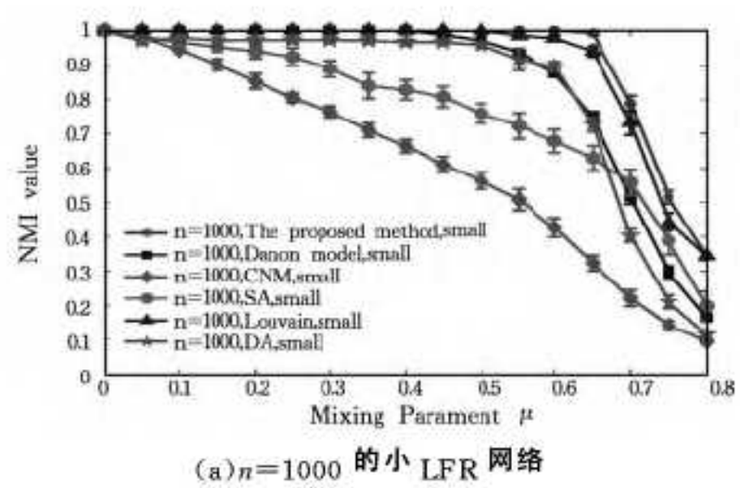
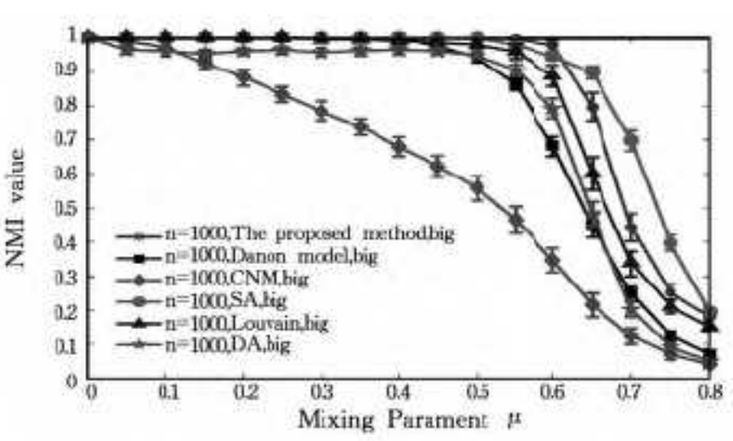


图 3 在 GN 网络中不同算法与本文算法的对比(平均每个点超过 50 次计算验证)

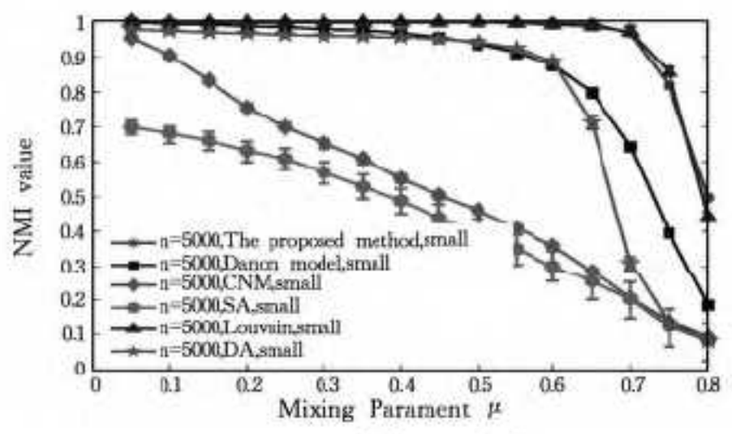
进一步在 LFR 基准网络上比较了不同算法的效率,结果如图 4 所示。



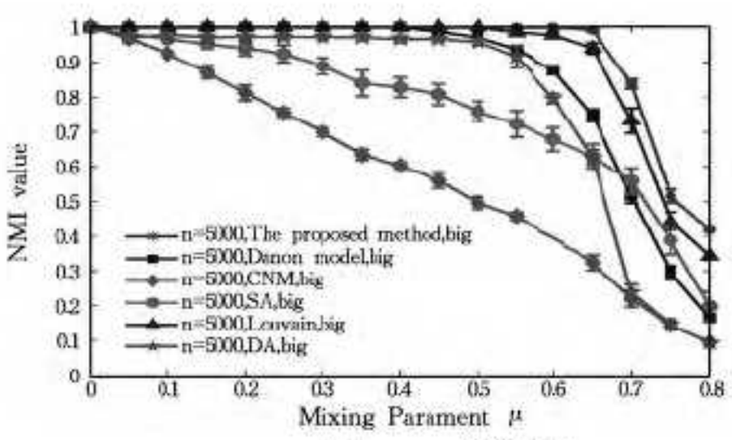
(a)$n=1000$ 的小 LFR 网络



(b)$n=1000$ 的大 LFR 网络



(c)$n=5000$ 的小 LFR 网络



(d)$n=5000$ 的大 LFR 基准网络

图 4 不同算法与本文算法的对比(每个点平均超过 50 次计算实现)

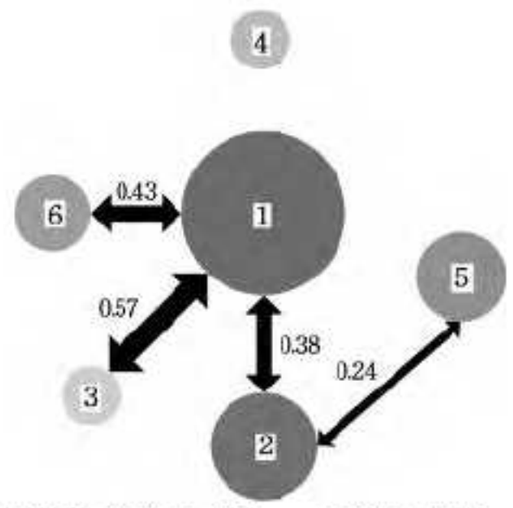
在本次实验中,每个社团的规模有大小之分。具体来说,

在 $n=1000$ 的网络中，大(小)社团的数目为 $100(50)$，$n=5000$ 的网络中，大(小)社团的数目为 $300(200)$。可以看出，我们的算法在各种网络配置中都几乎拥有最高的准确性，展示了其有效性和普适性。

除了 GN 和 LFR 网络，在拥有 1000 个团且每个团拥有 10 个节点的团环网络[12]上进行验证。团环网络将 K 个完全图用少量边连接起来，因此拥有两个参数即团的数目和每个团的规模。结果显示，此算法能精确地发现每个团，进而验证了其有效性。

5.2 揭示有效的社团交互模式

本文在真实社会网络上对算法进行了测试，利用“Les Miserables”网络来推测小说中各人物的角色定位。抽象社团网络描述了著名法国作家雨果的代表作“Les Miserables”(悲惨世界)中主要人物的关系。网络中节点代表主要人物，边代表所连接的两个人物在同一场景中出现过。在图 5 中，社团的颜色和大小分别由社团所包含的节点颜色和重要性决定。通过切断社团交互强度低于平均阈值(50%)的社团间联系，得到了社团间耦合模式图。通过观察，发现有两个中心节点位于中央，且有 4 个社团通过一些离群节点相互连接。



图中有中心节点(社团 1)，离群点(社团 4)，
普通社团(社团 2,3,5 和 6)，箭头的粗细代表耦合强。

图 5 推测“Les Miserables”网络中社团耦合模式图

为了推测相关人物的角色定位，主要研究了中心节点和离群节点。两个中心节点是小说的主人公 Valjean 和 Javert，他们的故事是贯穿全书的主线。这两个重要节点与网络内近 50% 的节点相互连接，进而他们与节点间的连线构成了整个网络。除了中心节点，离群点在构造网络中也起到了重要作用。例如，社团 4 几乎是与整个网络割裂的，然而它却包含 4 个非常有趣的人物节点，即 Blacheville、Listolier、Fameuil 和 Tholomyes，他们都是 Parision 的学生。此外，社团 5 中有 37 个离群点，他们看上去不起作用，然而正是这些离群点将社团 5 与两个中心节点和其他社团通过几条边连接了起来。

结束语 本文提出了一种新颖的动态社团探测算法，利用演化迭代技术高效而准确地揭示了网络中的社团结构。本算法十分高效，计算复杂度与稀疏网络规模呈线性关系。为了使得算法更加高效，一些问题需要更深入的探讨，例如如何在超大规模的网络(微博网络和生物网络)上进行准确的社团划分、如何处理非完全双向的混杂网络等。

参 考 文 献

[1] Wang X F, Chen G. Complex networks: Small-world, scale-free and beyond[J]. IEEE Circuits and System Magazine, 2003, 3(1): 6-20

[2] Newman M E J. Networks: an introduction[M]. Oxford University Press, 2010

[3] Lones M A, Caves A P, Stepney S, et al. Artificial biochemical networks: Evolving dynamical systems to control systems[J]. IEEE Transactions on Evolutionary Computation, 2012, 18(2): 145-166

[4] Palafox L, Noman N, Iba H. Reverse engineering of gene regulatory networks using dissipative particle swarm optimization[J]. IEEE Transactions on Evolutionary Computation, 2013, 17(4): 77-587

[5] Li H J, Daniels J. Social significance of community structure: Statistical view[J]. Physical Review E, 2015, 91(1): 012801

[6] Fortunato S. Community detection in graphs[J]. Physics Reports, 2010, 486(3-5): 75-174

[7] 李慧嘉. 基于信息扩散的多尺度重叠社团快速探测算法[J]. 计算机科学, 2014, 41(9): 125-131

[8] Pizzuti C. A multi-objective genetic algorithm to find communities in complex networks[J]. IEEE Transactions on Evolutionary Computation, 2012, 16(3): 418-430

[9] Liu Y, Moer J, Aviyente S. Network Community Structure Detection for Directional Neural Networks Inferred From Multichannel Multisubjective EEG Data[J]. IEEE Transactions on Biomedical Engineering, 2014, 61(7): 1919-1930

[10] 李慧嘉, 李慧颖, 李爱华. 多尺度的社团结构稳定性分析[J]. 计算机学报, 2015, 38(2): 301-312

[11] Blondel V D, Guillaume J, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): 10008

[14] Fortunato S, Barlemy M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(1): 36-41

[13] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113

[14] Reichardt J, Bornholdt S. Statistical mechanics of community detection[J]. Physical Review E, 2006, 74(1): 016110

[15] Hofman J M, Wiggins C H. Bayesian Approach to Network Modularity[J]. Physical Review Letters, 2008, 100(25): 258701

[16] Peixoto T P. Parsimonious module inference in large networks [J]. Physical review letters, 2013, 110(14): 148701

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

[18] Guimera R, Nunes L A, Amaral. Functional cartography of complex metabolic networks[J]. Nature, 2005, 433(7028): 895-900

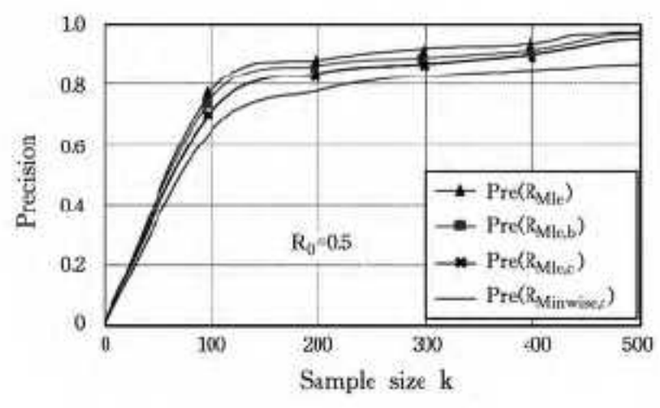
[19] 李慧嘉. 优化稳定性的多层次社团快速探测算法[J]. 小型微型计算机系统, 2015, 36(3): 566-571

[20] Duch J, Arenas A. Community detection in complex networks using external optimization[J]. Physical Review E, 2005, 72: 2

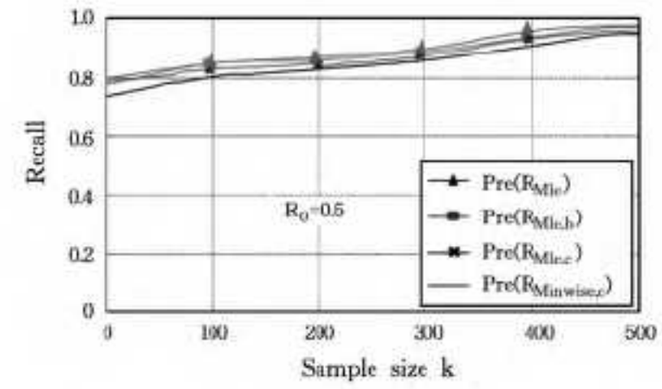
[21] Lancichinetti A, Fortunato S. Community detection algorithms: A comparative analysis[J]. Physical Review E, 2009, 80(5): 56117

[22] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical Review E, 2004, 70(6): 66111

(下转第 412 页)



(a)Precision of $R_0=0.5$



(b)Recall of $R_0=0.5$

图 2 R_{Mle},$R_{Mle,b}$, $R_{Mle,c}$, $R_{Minwise,c}$的准确率和召回率

(1)随着比对次数 k 的增加,估计精度会提升,表明 k 越大,极大似然估计子的估计方差越小,估计精度就越接近真实相似度。

(2)在相同的阈值 T 和比对次数 k 的条件下,准确率排序为:$Pre(R_{Mle})>Pre(R_{Mle,b})>Pre(R_{Mle,c})>Pre(R_{Minwise,c})$。例如,当 $T=0.5$、$k=300$ 时,准确率排序为:$Pre(R_{Mle})=93\%>Pre(R_{Mle,b})=89\%>Pre(R_{Mle,c})=87\%>Pre(R_{Minwise,c})=65\%$。

4.2 时间性能

图 3 显示了 3 种估计子(R_{Mle},$R_{Mle,c}$, $R_{Minwise,c}$)和 $R_{Mle,c}$ 增加过滤器 $R(T_{Mle,c})$ 的时间性能,当计算规模相同时,计算所需的时间排序为:$\mathrm{Time}(R(T_{Mle,c}))>\mathrm{Time}(R_{Minwise,c})>\mathrm{Time}(R_{Mle,c})>\mathrm{Time}(R_{Mle})$。

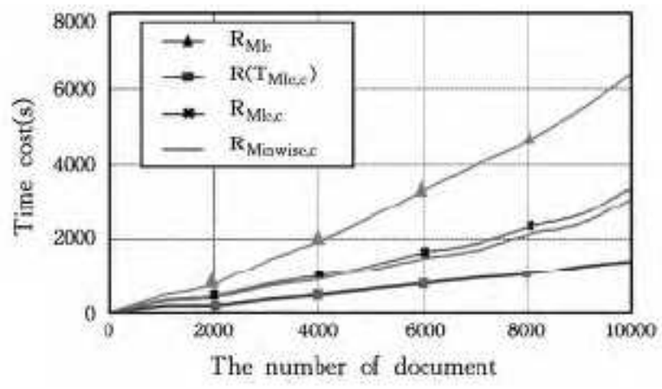


图 3 R_{Mle},$R_{Mle,c}$,$R_{Minwise,c}$和$R(T_{Mle,c})$的时间性能

综合图 2 和图 3 可知,根据不同的需求可以选择最适合自己的估计子:

(1) 如果将精度需求排在第一位,可以选择 R_{Mle},但是计算耗费的时间是最多的。

(2)如果想要兼顾精度和效率,可以选择 $R(T_{Mle,c})$,它的计算时间最少,并且在 $k>300$ 的条件下,$R_{Mle,c}$ 与 R_{Mle} 相比,准确度近乎相等。

结束语 本文提出了连接位极大似然动态过滤算法,通过实验对 $R(T_{Mle,c})$ 的可行性进行了分析。测试了 4 种估计子 R_{Mle},$R_{Mle,b}$,$R_{Mle,c}$, $R_{Minwise,c}$ 的准确率和召回率,其中 R_{Mle} 精度最高,$R_{Mle,c}$ 次之。测试了 3 种估计子(R_{Mle}, $R_{Mle,c}$, $R_{Minwise,c}$)和 $R_{Mle,c}$ 增加过滤器 $R(T_{Mle,c})$ 的时间性能,综合实验的结果表明 $R(T_{Mle,c})$ 是一种精度和效率兼顾的相似度估计算法,它的计算时间最少,并且在 $k>300$ 的条件下,$R_{Mle,c}$ 与 R_{Mle} 相比,准确度近乎相等。

参 考 文 献

[1] Broder A Z, Charikar M, Frieze A M, et al. Min-wise independent permutations [J]. Journal of Computer Systems and Sciences, 2000, 60(3): 630-659

[2] Kalpakis K, Tang S. Collaborative data gathering in wireless sensor networks using measurement co-occurrence [J]. Computer Communications, 2008, 31(10): 1979-1992

[3] Dourisboure Y, Geraci F, Pellegrini M. Extraction and classification of dense implicit communities in the web graph [J]. ACM Transactions on the Web (TWEB), 2009, 3(2): 1-36

[4] Bendersky M, Croft W B. Finding text reuse on the Web [C]// Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM'09). New York, USA: ACM, 2009: 262-271

[5] Buehrer G, Chellapilla K. A scalable pattern mining approach to web graph compression with communities [C]// Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08). New York, USA: ACM, 2008: 95-106

[6] Indyk P. A small approximately min-wise independent family of Hash functions [J]. Journal of Algorithm, 2001, 38(1): 84-90

[7] Charikar M S. Similarity estimation techniques from rounding algorithms [C]// Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing (STOC'02). New York, USA: ACM, 2002: 380-388

[8] Li P, König A C. b-bit minwise hashing [C]// Proceedings of the 19th International Conference on World Wide Web (WWW'10). New York, USA: ACM, 2010: 671-680

[9] Li P, König A C. Theory and applications of b-bit minwise hashing [J]. Communications of the ACM, 2011, 54(8): 101-109

[10] Yuan Xin-pan, Long Jun, Zhang Zu-ping, et al. Connected bit minwise hashing [J]. Journal of Computer Research and Development, 2013, 50(4): 883-890

(上接第 399 页)

[23] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification [J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(9): 09008

[24] 马英红,李慧嘉,张晓东. 赋权网络中的弱化免疫研究[J]. 管理科学学报, 2010, 13(10): 32-39

[25] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826

[26] Hoffman J M, Wiggins C H. Bayesian Approach to Network Modularity[J]. Physical Review Letters, 2008, 100(25): 258701

[27] Zachary W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 28: 452-47

[28] Ronhovde P, Nussinov Z. Local resolution-limit-free Potts model for community detection[J]. Physical Review E, 2010, 81(4): 046114