社团结构迭代快速探测算法

李慧嘉[1)] 李爱华[1)] 李慧颖[2)]

[1)](中央财经大学管理科学与工程学院 北京 100081)

[2)](清华大学自动化系 北京 100084)

摘 要 作为复杂网络研究的重要组成部分,社团结构分析对于理解和分析现实世界中各种社会、工程和生物等系统具有非常重要的意义.该文利用动态迭代技术,提出了一种新型的社团探测技术,能够准确而快速地识别网络中的社团结构.首先引入一种动态系统,可以使社团归属从随机状态逐步收敛到最优划分,进一步利用严格的数学分析给出了社团归属在离散时间内收敛到最优的条件.该文创新性地提出了划分指标函数的一般化形式,通过选择不同的参数,可以引申到几乎所有著名的指标函数.为了使动态系统不需要任何参数选择即可完成向最优社团的收敛,文中设计了一种新颖的图生成模型,使得算法能在无参数的情况下方便高效的运行.该算法具有较高的效率,计算复杂性分析显示算法需要的时间与稀疏网络节点的数量呈线性关系.最后,文中将算法应用到人工网络和实际网络中,结果显示算法不仅具有极高的准确性,还能够高效地应用于大规模现实网络的分析和计算中.

关键词 社团结构;动态系统;指标函数;线性复杂度;层次结构;社交网络

中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2017.00970

Fast Community Detection Algorithm Via Dynamical Iteration

LI Hui-Jia[1)] LI Ai-Hua[1)] LI Hui-Ying[2)]

[1)](*School of Management Science and Engineering, Central University of Finance and Economics, Beijing* 100081)

[2)](*Department of Automation, Tsinghua University, Beijing* 100084)

Abstract A feature, observable in many networks, is the presence of community structures, i.e. clusters of vertices which are densely connected to each other while less connected to the vertices outside. The community structure identification is an important problem in a wide range of applications such as marketing in social networks and study of protein interaction networks. Usually, the community members have more properties in common among themselves than with nonmembers and detecting community structure helps analyzing and searching the network with less effort. However, most existing approaches fall into the categories of either optimization based or heuristic methods which do not meet both speed and accuracy requirements simultaneously. In this paper, a new fast and accurate community detection algorithm is proposed based on dynamic system in complex networks. First, a discrete-time dynamic system is introduced to describe the assignment of community memberships, and the conditions driving the convergence of dynamics trajectory to the optimal situation are formulated. A new algorithm is proposed which can be generalized to unify the conventional algorithms widely applied. Furthermore, a new type graph generative model is designed which performs the algorithm free of the parameters. Our algorithm is highly efficient: the computational complexity analysis shows that the required time is linearly dependent on the number of all nodes in a sparse network. We perform extensive

收稿日期:2016-05-12;在线出版日期:2016-10-19.本课题得到国家自然科学基金项目(71401194,71401188,91324203,11131009)、中央财经大学"青年英才"培育支持项目(QYP1603)资助.**李慧嘉**,男,1985年生,博士,副教授,主要研究方向为数据挖掘、复杂网络、信息检索.E-mail:hjli@amss.ac.cn.**李爱华**,女,1978年生,博士,副教授,主要研究方向为数据挖掘、管理决策.**李慧颖**,女,1983年生,博士,助理研究员,主要研究方向为生物信息学、复杂网络、数据挖掘.

simulations with synthetic and also real-world benchmark networks to verify the algorithmic performance. The results showed that the proposed method does not face the resolution limit problem and performs very well.

Keywords　community structure; dynamic system; quality function; linear time; hierarchical structure; social networks

1　引　言

随着信息技术的发展，很多现实世界系统可以用复杂网络（complex network）来模拟[1-4]，例如Internet、WWW 网络、交通运输网、电力传输网、科学家合作网、微博关联网、蛋白质交互网和基因关联网络等[5-11]. 伴随着新型数据处理技术特别是大数据技术的产生和发展，复杂网络特性分析获得了越来越多的关注. 社团结构作为复杂网络的一个重要特征，它将网络划分成具有紧密内在联系的子群[12-14]，而同一子群中的节点通常有共同的属性和关联. 准确而快速地发现社团结构能够有效地优化网络系统的预测、控制和演化，例如设计网络推荐系统进行精确地营销[4,15-16]，药物靶点和基因关联点的定位[12,17-18]，网络社交媒体发展趋势的预测等[13,15,19-20].

目前虽然已有不少经典的探测方法，但是许多基本的社团相关问题仍然没有非常清晰的解释. 例如，社会或生物系统中社团结构的深层次意义是什么？为什么社团结构在网络中会自然地呈现出层次结构[21-24]？虽然经典优化方法和启发式算法被广泛应用到社团探测中，但它们的基本思想是比较社团内外的直观属性和拓扑特性[25-28]. 为了达到一定的准确性，这些方法需要很高的计算复杂度，而且结果会出现一些不可避免的缺陷，如分辨率限制[11]和错误识别[13]等. 因此设计一个准确而高效的社团探测算法并进一步刻画拓扑结构隐藏特征（比如动态迭代[29-34]）非常具有挑战性.

为了能快速而准确地发现最优的社团结构，本文设计出一种新颖的动态系统（如图 1 所示），可将社团归属从随机状态逐步迭代收敛到使指标函数最大的全局最优划分. 利用严格的数学推导，我们进一步给出社团归属收敛到最优的条件. 总体来说，本文创新点主要有 3 个方面：

（1）根据严格的数学推导，我们证明了通过迭代动态系统，可以最优化特定指标函数以获取理想的社团分区，即动态迭代过程的固定的稳定值. 特定条件下，最优社团划分可以通过迭代动态系统收敛得出.

（2）本文提出一种新颖的一般化目标函数，能够将著名的指标函数整合为统一的形式，这样我们可以通过选择不同的参数，来找到合适的指标函数.

（3）我们进一步设计了一种新颖的图生成模型（graph generative model），使得动态系统不需要任何参数选择即可完成向最优社团的收敛，方便高效.

利用动态系统，我们提出了一个新型的动态社团探测算法，在稀疏网络上其时间复杂度和节点数目线性相关. 进一步，为了确定社团的最优数目，本文利用 Markov 状态转移矩阵及其特征系统给出了具体而严格的证明. 最后，实验结果表明，本算法不仅具有较高的准确性，还能够方便地应用到大规模网络的计算和分析中.

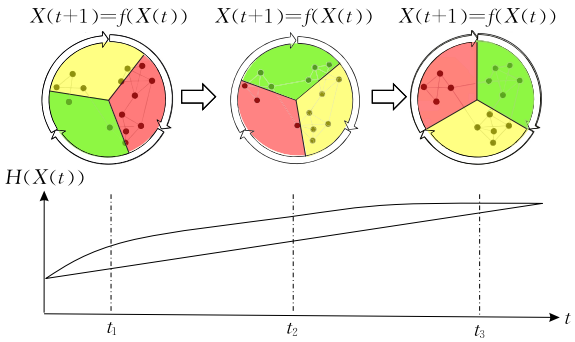


图 1　根据迭代机制，为了得到最优社团划分 $X(t)$，我们提出一种高效的动态迭代系统 $X(t+1)=f(X(t))$，最大化方程 H 并得到 $X(t)$收敛的最优结果（其中，指标函数 H 用来衡量 t 时刻的社团归属 $X(t)$的优劣，网络中不同的社团用不同的扇区表示，可以很容易看出通过迭代收敛，最右端的划分是最优的）

2　相关研究

近年来社团探测技术成为网络分析的热点并涌现出多种探测算法. 一般地，每种方法由两部分组成：评估社团划分的质量函数和相应的划分算法. 在众多质量函数中，模块度函数 Q 是最广为人知的. 假设 $\boldsymbol{A}$ 是 G 的邻接矩阵，σ_i是节点 i 的社团标志，当

$\sigma_i=\sigma_j$时有$\delta(\sigma_i,\sigma_j)=1$，否则$\delta(\sigma_i,\sigma_j)=0$. 如式(1)所示，模块度函数$Q$的增加代表了社团划分的优化. 因此，社团探测问题可以转化成为模块度函数Q的最大化问题.

$$Q(\{\sigma\})=-\frac{1}{2m}\sum_{i\neq j}(A_{ij}-p_{ij})\delta(\sigma_i,\sigma_j) \quad (1)$$

Potts模型是另一种应用广泛的社团探测技术，它通过将社团的标号用自旋状态(spin states)来表示，进而利用系统的自旋能量函数[35-36]来评估社团划分的质量. Reichardt & Bornholdt(RB)[15]提出了一个泛汉密尔顿函数来刻画能量函数，

$$H_{\mathrm{RB}}(\{\sigma\})=-\frac{1}{2}\sum_{i\neq j}(A_{ij}-\gamma_{\mathrm{RB}}p_{ij})\delta(\sigma_i,\sigma_j) \quad (2)$$

其中：γ_{RB}为汉密尔顿函数的分辨率参数；γ_{RB}，$p_{ij}\in\mathbb{R}$. RB模型可以拓展到两个典型的子模型：

(1) RBER模型. 所有的边都具有相同的连接概率，其能量函数如下：

$$H_{\mathrm{RBER}}(\{\sigma\})=-\frac{1}{2}\sum_{i\neq j}(A_{ij}-\gamma_{\mathrm{RB}}p)\delta(\sigma_i,\sigma_j) \quad (3)$$

(2) RBCM模型. 边的连接概率与网络的度分布相关，其能量函数如下：

$$H_{\mathrm{RBCM}}(\{\sigma\})=-\frac{1}{2}\sum_{i\neq j}\left(A_{ij}-\gamma_{\mathrm{RB}}\frac{k_ik_j}{2m}\right)\delta(\sigma_i,\sigma_j) \quad (4)$$

其中k_i是节点i的度. 值得注意的是，当$\gamma_{\mathrm{RB}}=1$时，RBCM模型的能量函数等同于模块度函数Q.

另外，Hofman和Wiggins[16]提出了一个一般化的概率模型并定义了相关的能量函数：

$$H_{\mathrm{HW}}(\{\sigma\})=-\frac{1}{2}\sum_{i\neq j}(W_LA_{ij}-W_G)\delta(\sigma_i,\sigma_j)+\frac{1}{2}\sum_{\mu=1}^{K}h_\mu\sum_{i=1}^{n}\delta(\sigma_i,\mu) \quad (5)$$

其中：$W_G=\log\frac{1-p_{\mathrm{out}}}{1-p_{\mathrm{in}}}$，$W_L=\log\frac{p_{\mathrm{out}}}{p_{\mathrm{in}}}+W_G$，$p_{\mathrm{in}}$代表两个节点属于相同社团的概率，$p_{\mathrm{out}}$代表两个节点属于不同社团的概率，该能量函数通过贝叶斯方法进行最优化处理.

进一步，Ronhovde和Nussinov[17]提出的一种无分辨率限制的模型，其质量函数如下：

$$H_{\mathrm{RN}}(\{\sigma\})=-\frac{1}{2}\sum_{i\neq j}(\boldsymbol{W}_{ij}A_{ij}-\gamma\boldsymbol{W}_{ij}\overline{A}_{ij})\delta(\sigma_i,\sigma_j) \quad (6)$$

其中：当$i\neq j$时，$\overline{A_{ij}}=1-A_{ij}$，否则$\overline{A_{ij}}=0$. 此外，$\boldsymbol{W}=[\boldsymbol{W}_{ij}]$是一个一般化的加权矩阵，为存在和缺失的边进行赋权.

算法方面，Newman-Fast(NF)算法[18]是一种基于模块度优化的迭代算法，它采用贪婪方法逐步更新社团归属，使得模块度的增量最大. Blondel等人[10]提出了Louvain算法，它初始时分配给每个节点不同的社团标号，使得每个节点都是一个独立社团，然后逐步迭代社团归属直至模块度函数无法再提高. OCR方法[37]是一种基于同步动态系统的社团探测方法，它得益于动态网络中紧密连接的顶点具有较大的局部同步性能. 标签传播方法是一种接近线性时间的社团探测方法[19]，算法以分配给每个节点不同的社团标号为起点，然后根据最近邻节点的标号来更新自己的社团标号[36-42]直至稳定状态. 有趣的是，标签传播等同于局部能量最小值的零温度Potts模型[22]：

$$H_{\mathrm{KPM}}(\{\sigma\})=-\sum_{i\neq j}A_{ij}\delta(\sigma_i,\sigma_j) \quad (7)$$

另外常见的社团探测方法还包括模拟退火算法(SA)[20]、外部优化算法[21,25-26]、最大化期望值算法[24,28]、贝叶斯推理方法[22,27]、变分贝叶斯方法等[16]和演化迭代方法[36,43-44]等.

3 问题标准化

3.1 概 念

我们首先给出一些基本定义，一些主要的参数在表1中进行了说明.

(1) c，n，m分别表示网络中社团、点和边的数目.

(2) l_μ^{in}和l_μ^{out}分别表示社团μ的团内边和团间边数量.

(3) p_μ^{in}表示社团μ团内边的比例，p_μ^{out}表示社团μ团间边的比例.

易得

$$p_\mu^{\mathrm{in}}=\frac{2l_\mu^{\mathrm{in}}}{n_\mu(n_\mu-1)},\quad p_\mu^{\mathrm{out}}=\frac{l_\mu^{\mathrm{out}}}{n_\mu(n-n_\mu)} \quad (8)$$

表1 主要参数说明

符号	说明	符号	说明
c	社团数量	p_μ^{in}	社团μ团内边的比例
n	节点数量	p_μ^{out}	社团μ团间边的比例
m	边数量	$x_{i\mu}$	节点i属于社团μ的概率
l_μ^{in}	社团μ团内边数量	$k_{i\mu}$	节点i与社团μ的节点-社团程度
l_μ^{out}	社团μ团间边数量	$Q_{i\mu}$	节点i基于社团μ的质量函数
n_μ	社团μ内节点数量		

一个节点可以以不同的概率分属不同的社团. 考虑一个由n个节点，m条边，c个社团构成的网络G，本文引入向量$\boldsymbol{X}_i(t)=[x_{i\mu}(t)]$，$\mu=\langle1,2,\cdots,c\rangle$，其中$x_{i\mu}(t)$表示在时刻$t$时，节点$i$属于社团$\mu$的概

率. 易得

$$0 \le x_{i\mu}(t) \le 1$$
$$\sum_{\mu=1}^{c} x_{i\mu}(t) = 1, \quad \forall i \in V \tag{9}$$

集合 $X(t)=[x_{i\mu}(t)]$, $i=\{1,2,\cdots,n\}$ 为在时刻 t 时网络社团归属的集合.

定义 1(硬划分). 硬划分中的每个点 i 在 t 时刻仅属于特定社团 c,它的充要条件是向量 $\boldsymbol{X}_i(t)$ 的第 c 个元素为 1,其余值都等于 0. 在硬划分中,不存在重叠节点(overlapping nodes).

定义 2(节点-社团度). 我们将节点 i 在社团 μ 中的节点-社团度表示为 $k_{i\mu}$,其定义为节点 i 连接到社团 μ 的边数. 其数学化表示为

$$k_{i\mu} = \sum_{\alpha} x_{\alpha\mu} A_{\alpha i} \tag{10}$$

利用社团归属 $X(t)$,在硬划分的条件下,一些常用的概念 $n_\mu, l_\mu^{\text{in}}, l_\mu^{\text{out}}, p_\mu^{\text{in}}, p_\mu^{\text{out}}$ 可以重新表示为

$$n_\mu(t) = \sum_i x_{i\mu}(t) \tag{11}$$

$$l_\mu^{\text{in}} = \frac{1}{2}\sum_i \sum_{j\ne i} x_{i\mu}(t) x_{j\mu}(t) A_{ij} \tag{12}$$

$$l_\mu^{\text{out}} = \sum_i \sum_{j\ne i} x_{i\mu}(t)(1-x_{j\mu}(t)) A_{ij} \tag{13}$$

$$\begin{aligned} p_\mu^{\text{in}} &= \frac{\sum_i \sum_{j\ne i} x_{i\mu}(t) x_{j\mu}(t) A_{ij}}{\sum_i \sum_{j\ne i} x_{i\mu}(t) x_{j\mu}(t)} \\ &= \frac{\sum_i \sum_{j\ne i} x_{i\mu}(t) x_{j\mu}(t) A_{ij}}{\sum_i x_{i\mu}(t)(n_\mu(t)-x_{i\mu}(t))} \\ &= \frac{\sum_i \sum_{j\ne i} x_{i\mu}(t) x_{j\mu}(t) A_{ij}}{n_\mu^2(t) - \sum_i x_{i\mu}^2(t)} \end{aligned} \tag{14}$$

$$\begin{aligned} p_\mu^{\text{out}} &= \frac{\sum_i \sum_{j\ne i} x_{i\mu}(t)(1-x_{j\mu}(t)) A_{ij}}{\sum_i \sum_{j\ne i} x_{i\mu}(t)(1-x_{j\mu}(t))} \\ &= \frac{\sum_i \sum_{j\ne i} x_{i\mu}(t)(1-x_{j\mu}(t)) A_{ij}}{\sum_i \sum_{j\ne i} x_{i\mu}(t) - \sum_i \sum_{j\ne i} x_{i\mu}(t) x_{j\mu}(t)} \\ &= \frac{\sum_i \sum_{j\ne i} x_{i\mu}(t)(1-x_{j\mu}(t)) A_{ij}}{(n-1)n_\mu - \left(n_\mu^2(t) - \sum_i x_{i\mu}^2(t)\right)} \end{aligned} \tag{15}$$

3.2 基于迭代的新型动态系统

相比与大多数经典算法,如层次分析法和启发式算法,利用动态系统更新归属向量(即 $\boldsymbol{X}(t)=[x_{i\mu}(t)]$)到一个理想的固定值是非常简单的,其计算复杂度非常小,便于在大规模网络上进行计算和分析. 虽然有一些其他的动态系统,如随机游走和协同系统,也能够将归属矩阵收敛到一个稳定的状态,但是它们仅仅是对归属矩阵进行迭代而无法保证收敛到最优值(最优化特定的指标函数 Q). 为了快速而准确地探测社团结构,我们提出一种新型的动态迭代算法. 这里,首先给出离散时间下的动态系统:

$$\boldsymbol{x}_{i\mu}(t+1) = \frac{\boldsymbol{x}_{i\mu}(t)\mathrm{e}^{Q_{i\mu}(\boldsymbol{x}_{*\mu}(t))}}{\sum_{\tilde{\mu}} \boldsymbol{x}_{i\tilde{\mu}}(t)\mathrm{e}^{Q_{i\tilde{\mu}}(\boldsymbol{x}_{*\tilde{\mu}}(t))}} \tag{16}$$

其中:$Q_{i\mu}(\boldsymbol{x}_{*\mu}(t+1)):R^n\to R$ 是节点 i 归属于社团 μ 的质量函数,$\boldsymbol{x}_{*\mu}(t)$ 是一个 n 维向量,第 i 维是 $\boldsymbol{x}_{i\mu}(t)$,函数 Q 满足下列约束条件:

$$\frac{\partial Q_{i\mu}}{\partial \boldsymbol{x}_{i\mu}} = 0 \tag{17}$$

由此,我们推断出了几条关于动态系统的重要性质.

性质 1. 离散的动态系统有两类平凡(trivial)稳定点,分别为:

(1) 硬社团结构的社团归属,即 $x_{i\mu}(t)=1$ 和 $x_{i\tilde{\mu}}(t)=0$, $\forall\tilde{\mu}\ne\mu$;

(2) 均匀的社团归属,即 $x_{i\mu}(t)=\dfrac{1}{c}$, $i=\{1,2,\cdots,n\}$, $\mu=\{1,2,\cdots,c\}$.

证明. 当 $x_{i\mu}(t)=1$ 和 $x_{i\tilde{\mu}}(t)=0$ 时, $\forall\tilde{\mu}\ne\mu$,得到

$$x_{i\mu}(t+1) = \frac{x_{i\mu}(t)\mathrm{e}^{Q_{i\mu}(x_{*\mu}(t))}}{\sum_{\tilde{\mu}} x_{i\tilde{\mu}}(t)\mathrm{e}^{Q_{i\tilde{\mu}}(x_{*\tilde{\mu}}(t))}} = \frac{x_{i\mu}(t)\mathrm{e}^{Q_{i\mu}(x_{*\mu}(t))}}{x_{i\mu}(t)\mathrm{e}^{Q_{i\mu}(x_{*\mu}(t))} + \sum_{\mu\ne\tilde{\mu}}(0\times \mathrm{e}^{Q_{i\tilde{\mu}}(x_{*\tilde{\mu}}(t))})} = 1 \tag{18}$$

同理,对于所有的 i 有 $x_{i\mu}(t)=\dfrac{1}{c}$,这意味着 $x_{i\mu}(t+1)=\dfrac{1}{c}$ 对于所有的 i 成立. 证毕.

为了表述方便,此后用 $Q_{i\mu}(t)$ 代替 $Q_{i\mu}(x_{*\mu}(t))$.

定理 1. 考虑一个一般化离散(连续)时间的动态系统,x^* 为相应的动态系统曲线的稳定值[24].

(1) 当且仅当 Jacobian 矩阵 $\dfrac{\partial F}{\partial x}(x^*)$ 所有的特征值 λ_i 满足 $|\lambda_i|<1(\mathrm{Re}(\lambda_i<0))$ 时,x^* 是渐进稳定的.

(2) 当 Jacobian 矩阵 $\dfrac{\partial F}{\partial x}(x^*)$ 只要有一个特征值 λ_i 满足 $|\lambda_i|>1(\mathrm{Re}(\lambda_i>0))$ 时,x^* 是不稳定的.

定理 2. 对于任意的硬社团归属 $X(t)$,记 μ_i^* 为归属值 $x_{i\mu_i^*}(t)=1$ 时的社团. 那么当

$$\forall i,\mu\neq\mu_i^*,\ Q_{i\mu}(t)<Q_{i\mu_i^*}(t) \tag{19}$$

动态系统(16)在 $X(t)$有一渐进的稳定点.如果对于某节点对 $i\mu$,$\forall i,\mu\neq\mu_i^*$,$Q_{i\mu}(t)>Q_{i\mu_i^*}(t)$,则此稳定点不固定.

证明. 为了确定动态系统在 $X(t)$处的 Jacobian 矩阵,系统(16)的 $n\times K$ 个状态变量和相应的导数计算如下

$$\frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu}(t)}=\frac{\mathrm{e}^{Q_{i\mu}(t)}\sum\limits_{\mu\neq\mu}x_{i\tilde{\mu}}(t)\mathrm{e}^{Q_{i\tilde{\mu}}(t)}}{\left(\sum\limits_{\tilde{\mu}}x_{i\tilde{\mu}}(t)\mathrm{e}^{Q_{i\tilde{\mu}}(t)}\right)^2}$$

$$\left.\frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu}(t)}\right|_{x_{i\mu}(t)=1}=0,$$

$$\left.\frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu}(t)}\right|_{\substack{x_{i\mu}(t)=0\\x_{i\tilde{\mu}}(t)=1}}=\frac{\mathrm{e}^{Q_{i\mu}(t)}}{\mathrm{e}^{Q_{i\tilde{\mu}}(t)}} \tag{20}$$

$$\frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu'}(t)}=-\frac{x_{i\mu}(t)\mathrm{e}^{Q_{i\mu}(t)}\mathrm{e}^{Q_{i\mu'}(t)}}{\left(\sum_{\tilde{\mu}}x_{i\tilde{\mu}}(t)\mathrm{e}^{Q_{i\tilde{\mu}}(t)}\right)^2}$$

$$\left.\frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu'}(t)}\right|_{x_{i\mu}(t)=1}=-\frac{\mathrm{e}^{Q_{i\mu'}(t)}}{\mathrm{e}^{Q_{i\mu}(t)}},\qquad \left.\frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu'}(t)}\right|_{x_{i\mu}(t)=0}=0, \tag{21}$$

$$\frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu}(t)}=\frac{x_{i\mu}(t)\mathrm{e}^{Q_{i\mu}(t)}\left(\frac{\partial Q_{i\mu}(t)}{\partial x_{j\mu}(t)}\right)\sum\limits_{\mu\neq\mu}x_{i\tilde{\mu}}(t)\mathrm{e}^{Q_{i\tilde{\mu}}(t)}}{\left(\sum\limits_{\tilde{\mu}}x_{i\tilde{\mu}}(t)\mathrm{e}^{Q_{i\tilde{\mu}}(t)}\right)^2},i\neq j$$

$$\left.\frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu}(t)}\right|_{x_{i\mu}(t)=1}=0,\qquad \left.\frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu}(t)}\right|_{x_{i\mu}(t)=0}=0 \tag{22}$$

$$\frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu'}(t)}=\frac{x_{i\mu}(t)\mathrm{e}^{Q_{i\mu}(t)}\left(\frac{\partial Q_{i\mu}(t)}{\partial x_{j\mu'}(t)}\right)x_{i\mu'}(t)\mathrm{e}^{Q_{i\mu'}(t)}}{\left(\sum\limits_{\tilde{\mu}}x_{i\tilde{\mu}}(t)\mathrm{e}^{Q_{i\tilde{\mu}}(t)}\right)^2},\ i\neq j$$

$$\left.\frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu'}(t)}\right|_{x_{i\mu}(t)=1}=0,\qquad \left.\frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu'}(t)}\right|_{x_{i\mu}(t)=0}=0 \tag{23}$$

对所有的 μ 和 $\tilde{\mu}$,假设 $\boldsymbol{J}_i$ 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,在方程(20)~(23)中运用合适的排列和替换,$\boldsymbol{J}_{ii}$ 在 $X(t)$的值为

$$\boldsymbol{J}_{ii}=\left.\begin{bmatrix}0&-\frac{\mathrm{e}^{Q_{i1}(t)}}{\mathrm{e}^{Q_{i\mu_i^*}(t)}}&\cdots&-\frac{\mathrm{e}^{Q_{ik}(t)}}{\mathrm{e}^{Q_{i\mu_i^*}(t)}}\\0&\frac{\mathrm{e}^{Q_{i2}(t)}}{\mathrm{e}^{Q_{i\mu_i^*}(t)}}&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\frac{\mathrm{e}^{Q_{ik}(t)}}{\mathrm{e}^{Q_{i\mu_i^*}(t)}}\end{bmatrix}\right|_{X(t)} \tag{24}$$

基于硬划分下的稳定点,整个 Jacobian 矩阵可以由以下矩阵表示:

$$\boldsymbol{J}=\left.\begin{bmatrix}J_{11}&0&0\\0&J_{ii}&0\\0&0&J_{nn}\end{bmatrix}\right|_{X(t)} \tag{25}$$

很显然,基于硬划分下稳定点的 Jacobian 矩阵的特征值为[45]

$$\lambda_{i\mu_i^*}=0,\ \lambda_{i\mu}=\frac{\mathrm{e}^{Q_{i\mu}(t)}}{\mathrm{e}^{Q_{i\mu_i^*}(t)}},\ i=1,2,\cdots,n,\ \mu\neq\mu_i^* \tag{26}$$

基于定理 1,动态系统在 $X(t)$有一个渐进式的稳定点,即如果

$$\forall i,\mu,\ |\lambda_{i\mu}|<1 \tag{27}$$

已经对所有 $\lambda_{i\mu_i^*}$ 都满足,而且剩余也满足的话需要满足下面条件:

$$\forall i,\mu\neq\mu_i^*,\ |\lambda_{i\mu}|=\left|\frac{\mathrm{e}^{Q_{i\mu}(t)}}{\mathrm{e}^{Q_{i\mu_i^*}(t)}}\right|<1 \tag{28}$$

即

$$\forall i,\mu\neq\mu_i^*,\ \mathrm{e}^{Q_{i\mu}(t)}<\mathrm{e}^{Q_{i\mu_i^*}(t)}\Rightarrow Q_{i\mu}(t)<Q_{i\mu_i^*}(t) \tag{29}$$

另一方面,如果某个点 i 和社团 μ,$\mu\neq\mu_i^*$,我们可以得到 $Q_{i\mu}(t)>Q_{i\mu_i^*}(t)$,那么这个稳定点不固定.

证毕.

定理 2 直观地说明了函数值 $Q_{i\mu}(t)$ 是“节点 i 属于社团 μ_i^*”的质量指标.因此,如果硬社团归属的稳定值是固定的,那么所有节点的相关质量函数都应该是最大的,这可以通过迭代动态系统(16)来实现.相应地,社团探测问题可以简化为设计合适的 Q 函数,用来衡量单个点在所有社团的归属程度.

到目前为止,我们介绍一种基于迭代的动态系统以及分析了其收敛的具体条件.接下来,我们为算法设计一种具有代表性的一般化社团探测指标.考虑下列方程 H:

$$H(X(t))=\frac{1}{2}\sum_{\mu}\sum_{i=1}^{n}x_{i\mu}(t)Q_{i\mu}(t) \tag{30}$$

我们下面证明方程 H 的全局最优值(global maximum)是动态系统(16)的固定的稳定值.

定理 3. 如果记 $\boldsymbol{X}(t)$为社团归属向量,那么最

大化方程 H 得到的社团归属向量 $\boldsymbol{X}(t)$ 的全局最优值是动态系统(16)的固定的稳定值.

证明.　如果 $\boldsymbol{X}(t)$ 为最大化方程 H 得到的社团归属向量的全局最优值,利用反证法,假定 $\boldsymbol{X}(t)$ 是动态系统的不固定的稳定值,那么会有

$$\exists i,\ \exists\tilde{\mu}\neq\mu,\ Q_{i\tilde{\mu}(t)}>Q_{i\mu_i^*}(t) \tag{31}$$

其中:μ_i^* 是社团标号,会使得 $x_{i\mu_i^*}(t)=1$.如果 $x_{i\mu_i^*}(t)=1$ 和 $x_{i\tilde{\mu}}(t)=0$ 的值互换,那么会有 $x_{i\mu_i^*}(t)=0$ 和 $x_{i\tilde{\mu}}(t)=1$,则方程 H 的值的相应变化如下:

$$H(X(t))|_{x_{i\mu_i^*}(t)=0,x_{i\tilde{\mu}}=1}-H(X(t))|_{x_{i\mu_i^*}(t)=1,x_{i\tilde{\mu}}=0}=Q_{i\tilde{\mu}}(t)-Q_{i\mu_i^*}(t)>0 \tag{32}$$

这就意味着函数 H 在节点 i 属于社团 $\tilde{\mu}$ 而非社团 μ_i^* 时有更大的值,这是矛盾的.因此我们开始的假设是无效的,进而可以验证最大化方程 H 得到的社团归属 $X(t)$ 的全局最优值是动态系统的固定的稳定点.　　证毕.

通过定理 3,我们容易看出如果最大化 H 函数时得到的硬划分社团结构是真实的社团结构,那么经过迭代动态系统(16)就会收敛到真实的社团结构上.

3.3　方程 Q 的一般化形式和拓展形式

我们发现,对于许多社团探测算法,如模块度算法和 Potts 模型,优化目标函数有如下的一般形式:

$$E(t)=-\frac{1}{2}\sum_{\mu}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}f_{\mu}^{+}A_{ij}x_{i\mu}(t)x_{j\mu}(t)-\sum_{j=1}^{n}f_{\mu}^{-}(1-A_{ij})x_{i\mu}(t)x_{j\mu}(t)\Big)+\sum_{\mu}R_{\mu} \tag{33}$$

其中:f_{μ}^{+} 为奖励项系数;f_{μ}^{-} 为惩罚项系数;R_{μ} 为一般冗余项,式(33)可以表述为

$$\begin{aligned}E(t)=&-\frac{1}{2}\sum_{\mu}\left[\frac{2\sum_{j=1}^{n}x_{j\mu}(t)}{l_{\mu}(t)}R_{\mu}+\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}f_{\mu}^{+}A_{ij}x_{i\mu}(t)x_{j\mu}(t)-\sum_{j=1}^{n}f_{\mu}^{-}(1-A_{ij})x_{i\mu}(t)x_{j\mu}(t)\Big)\right]\\=&-\frac{1}{2}\sum_{\mu}\sum_{i=1}^{n}x_{i\mu}\Big(\sum_{j=1}^{n}f_{\mu}^{+}A_{ij}x_{j\mu}(t)-\sum_{j=1}^{n}f_{\mu}^{-}(1-A_{ij})x_{j\mu}(t)+\frac{2}{l_{\mu}(t)}R_{\mu}\Big)\end{aligned} \tag{34}$$

其中:$l_{\mu}(t)=\sum_{j=1}^{n}x_{j\mu}(t)$ 是社团 μ 在时刻 t 时的规模(节点个数).

事实上,我们可以将不同类型的指标函数解释为其具有不同的奖励参数和惩罚参数.每一种不同的参数都有自己的内在原因、优势和缺点.我们定义下列方程:

$$Q_{i\mu}(t)=\sum_{j=1}^{n}f_{\mu}^{+}A_{ij}x_{j\mu}(t)-\sum_{j=1}^{n}f_{\mu}^{-}(1-A_{ij})x_{j\mu}(t)+R_{i\mu} \tag{35}$$

选择 $R_{i\mu}$,使 $\frac{\partial R_{i\mu}}{\partial x_{i\mu}(t)}=0$,$R_u=\sum_{i=1}^{n}R_{i\mu}$,例如 $R_{i\mu}=\frac{2}{l_{\mu}}R_u$,可以发现,此时方程(33)明显符合方程(30)定义的能量函数通用形式.

有趣的是,对于特定的社团归属 $x_{i\mu}(t)$,方程(35)所定义的 $Q_{i\mu}$ 可以引申到几乎所有的著名指标函数,因此它是具有代表性的一般化指标:

(1) Hofman&wiggins 模型[16]

$$f_{\mu}^{+}=\log\frac{p^{\text{in}}}{p^{\text{out}}},\ f_{\mu}^{-}=\log\frac{1-p^{\text{out}}}{1-p^{\text{in}}},\ R_{\mu}=l_{\mu}\log\pi_{\mu} \tag{36}$$

(2) Ronhovde&Nussinov 模型[17]

$$f_{\mu}^{+}=1,\ f_{\mu}^{-}=\min_{\mu}p_{\text{in},\mu},\ R_{\mu}=0 \tag{37}$$

(3) RB Potts 模型(Erdos-Renyi 空模型[15]

$$f_{\mu}^{+}=1-\gamma_{\text{RB}}p,\ f_{\mu}^{-}=\gamma_{\text{RB}}p,\ R_{\mu}=0 \tag{38}$$

(4) RB Potts 模型(configuration 空模型)[15]

$$f_{\mu}^{+}=1-\frac{\gamma_{\text{RB}}}{2m},\ f_{\mu}^{-}=\frac{\gamma_{\text{RB}}}{2m},\ R_{\mu}=\sum_{i>j}\frac{\gamma_{\text{RB}}}{2m}(k_ik_j-1)x_{i\mu}(t)x_{j\mu}(t) \tag{39}$$

其中:k_i 是节点 i 的度;m 是网络中边的数量.

(5) 模块度指标[14]

$$f_{\mu}^{+}=1,f_{\mu}^{-}=\frac{k_ik_j}{2m},R_{\mu}=\sum_{i>j}\frac{1}{2m}(k_ik_j-1)x_{i\mu}(t)x_{j\mu}(t) \tag{40}$$

其中:k_i 是节点 i 的度;m 是网络中边的数量.

(6) 标号传播模型[19]

$$f_{\mu}^{+}=1,\ f_{\mu}^{-}=0,\ R_{\mu}=0 \tag{41}$$

4　自动选择参数的社团检测算法

在式(35)中,目标函数 $Q_{i\mu}(t)$ 的系数必须提前给定,然而当运用图生成模型(graph generative model)得到时,$Q_{i\mu}(t)$ 中的系数便能够自动获得.事实上,我们可以通过获知网络特征和相关社团结构的先验知识来确定这些系数.文献[16]展示了由 Hofman 和 Wiggins 提出的一种基于贝叶斯分析的

图生成模型,在此模型中系数设定为

$$f_{\mu}^{+}=\log\frac{p^{\text{in}}}{p^{\text{out}}},\ f_{\mu}^{-}=\log\frac{1-p^{\text{out}}}{1-p^{\text{in}}},\ R_{\mu}=l_{\mu}\log\pi_{\mu} \tag{42}$$

然而在文中作者假定所有网络中的社团都有同样的边的比率,即 $p_1^{\text{in}}=p_2^{\text{in}}=\cdots=p_c^{\text{in}}$ 和 $p_1^{\text{out}}=p_2^{\text{out}}=\cdots=p_c^{\text{out}}$. 显然这个假设并不是对多数网络都有效,例如在 LFR 网络[23,42]中,各个社团的边的比率是显著不同的.

这里,我们提出一个新型的图生成模型,它考虑了边比率在概率上的差异,提供了具有更好兼容性的版本.本模型实际上是一个广义的 Hofman-Wiggins 模型,模型中我们假定在网络中的每一个社团都有自己独有的概率,即对于社团 μ,有特定的 p_{μ}^{in} 和 p_{μ}^{out} 与之对应.所以,对于包含社团结构 $\{q\}$ 的网络 G,我们提出下列似然函数

$$\begin{aligned}P&=p(G\mid\{q\})\\&=\prod_{\kappa}\Bigg[\overbrace{(p_{\kappa}^{\text{in}})^{\left(\sum_{i>j,i,j\in\kappa}A_{ij}\right)}}^{\text{I}}\cdot\overbrace{(1-p_{\kappa}^{\text{in}})^{\left(\sum_{i>j,i,j\in\kappa}\overline{A_{ij}}\right)}}^{\text{II}}\cdot\\&\quad\overbrace{(p_{\kappa}^{\text{out}})^{\left(\sum_{i>j,i\in\kappa}A_{ij}-\sum_{i>j,i,j\in\kappa}A_{ij}\right)}}^{\text{III}}\cdot\\&\quad\overbrace{(1-p_{\kappa}^{\text{out}})^{\left(\sum_{i>j,i\in\kappa}\overline{A_{ij}}-\sum_{i>j,i,j\in\kappa}\overline{A_{ij}}\right)}}^{\text{IV}}\cdot\overbrace{\pi_{\kappa}^{n_{\kappa}}}^{\text{V}}\Bigg]\end{aligned} \tag{43}$$

其中:$\overline{A_{ij}}=1-A_{ij}$,$\pi_{\kappa}$ 为将任一节点分配到社团 μ 中的概率,即

$$\pi_{\kappa}=\frac{n_{\kappa}}{n} \tag{44}$$

在方程(43)中,多项式中的项 I(III)对应于社团内(间)现存的边,而项 II(IV)对应于社团内(间)缺失的边.依据每个社团中节点的数量,项 V 定义了网络的划分.我们可以将等式(43)化简为

$$\begin{aligned}P=\prod_{\kappa}\Bigg[&\left(\frac{p_{\kappa}^{\text{in}}}{p_{\kappa}^{\text{out}}}\right)^{\sum_{i>j,i,j\in\kappa}A_{ij}}\left(\frac{1-p_{\kappa}^{\text{in}}}{1-p_{\kappa}^{\text{out}}}\right)^{\sum_{i>j,i,j\in\kappa}\overline{A_{ij}}}\\&(p_{\kappa}^{\text{out}})^{\sum_{i>j,i,j\in\kappa}A_{ij}}(1-p_{\kappa}^{\text{out}})^{\sum_{i>j,i,j\in\kappa}J_{ij}}\pi_{\kappa}^{n_{\kappa}}\Bigg]\end{aligned} \tag{45}$$

因此,t 时刻的对数似然函数,可以被整理表示为一个硬划分 $X(t)$ 的形式:

$$\begin{aligned}LP(t)&=\log P(t)\\&=\frac{1}{2}\sum_{\kappa}\Bigg[\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)x_{j\kappa}(t)A_{ij}\log\left(\frac{p_{\kappa}^{\text{in}}}{p_{\kappa}^{\text{out}}}\right)+\\&\quad\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)x_{j\kappa}(t)J_{ij}\log\left(\frac{1-p_{\kappa}^{\text{in}}}{1-p_{\kappa}^{\text{out}}}\right)+\\&\quad\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)A_{ij}\log(p_{\kappa}^{\text{out}})+\\&\quad\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)\overline{A_{ij}}\log(1-p_{\kappa}^{\text{out}})+\\&\quad2n_{\kappa}(t)\log\left(\frac{n_{\kappa}}{n}\right)\Bigg]\end{aligned} \tag{46}$$

通过比较式(46)和式(33),可以看出式(46)中的头两项分别表示奖励项和惩罚项,而其余的为冗余项.此外,我们还可以引入参数 γ_1 和 γ_2,用来分别控制各个项的作用强度

$$\begin{aligned}H(t)=\frac{1}{2}\sum_{\kappa}\Bigg[&\sum_{i}\sum_{j\neq i}\gamma_1x_{i\kappa}(t)x_{j\kappa}(t)A_{ij}\log\left(\frac{p_{\kappa}^{\text{in}}}{p_{\kappa}^{\text{out}}}\right)+\\&\sum_{i}\sum_{j\neq i}\gamma_2x_{i\kappa}(t)x_{j\kappa}(t)\overline{A_{ij}}\log\left(\frac{1-p_{\kappa}^{\text{in}}}{1-p_{\kappa}^{\text{out}}}\right)-\\&\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)A_{ij}\log(p_{\kappa}^{\text{out}})+\\&\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)\overline{A_{ij}}\log(1-p_{\kappa}^{\text{out}})+\\&\sum_{i}2x_{i\kappa}(t)\log\left(\frac{n_{\kappa}}{n}\right)\Bigg]\end{aligned} \tag{47}$$

值得注意的是,等式(47)的算法复杂度是 $O(n^2)$,进一步它可以被整理为

$$\begin{aligned}H(t)=\frac{1}{2}\sum_{\kappa}\Bigg[&\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)x_{j\kappa}(t)A_{ij}\left(\gamma_1\log\left(\frac{p_{\kappa}^{\text{in}}}{p_{\kappa}^{\text{out}}}\right)-\right.\\&\left.\gamma_2\log\left(\frac{1-p_{\kappa}^{\text{in}}}{1-p_{\kappa}^{\text{out}}}\right)\right)+\gamma_2\log\left(\frac{1-p_{\kappa}^{\text{in}}}{1-p_{\kappa}^{\text{out}}}\right)\cdot\\&\sum_{i}x_{i\kappa}(t)\sum_{j\neq i}x_{j\kappa}(t)+\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)A_{ij}\log\cdot\\&\left(\frac{p_{\kappa}^{\text{out}}}{1-p_{\kappa}^{\text{out}}}\right)+\sum_{i}\sum_{j\neq i}x_{i\kappa}(t)\log(1-p_{\kappa}^{\text{out}})+\\&\sum_{i}2x_{i\kappa}(t)\log\left(\frac{n_{\kappa}}{n}\right)\Bigg]\end{aligned} \tag{48}$$

式(48)则拥有 $O(mc)$ 的计算复杂度,算法复杂度的详细讨论将在下一节进行.基于方程(35),可以得到模型(47)中的系数为

$$f_{\mu}^{+}=\gamma_1\log\left(\frac{p_{\kappa}^{\text{in}}}{p_{\kappa}^{\text{out}}}\right) \tag{49}$$

$$f_{\mu}^{-}=-\gamma_2\log\left(\frac{1-p_{\kappa}^{\text{in}}}{1-p_{\kappa}^{\text{out}}}\right) \tag{50}$$

$$\begin{aligned}R_{\mu}=\sum_{i}\sum_{j\neq i}\Bigg\{&A_{ij}\log\left(\frac{p_{\kappa}^{\text{out}}}{1-p_{\kappa}^{\text{out}}}\right)+\\&\log(1-p_{\kappa}^{\text{out}})+2\log\left(\frac{n_{\kappa}}{n}\right)\Bigg\}\end{aligned} \tag{51}$$

可以证明最小化质量函数(48)得到的硬划分社团结构 $X(t)$ 与真实的社团结构是一致的.我们容易看出在简单的情况下,当节点被放置在正确的社团中时能量函数(48)的值会减小,例如假定有两个完全相似的社团 C_1 和 C_2,其中 $p_{C_1}^{\text{in}}=p_{C_2}^{\text{in}}$,$p_{C_1}^{\text{out}}=p_{C_2}^{\text{out}}$,$n_{C_1}=n_{C_2}$.另外,假定有一个节点处于 C_1 和 C_2 之外,分别与 C_1 和 C_2 有 k_1 和 k_2 条连边.接下来,我们比较两种情况,即这个节点要么属于社团 C_1(情况 1),要

么属于社团 C_2(情况 2). 在情况 1 和 2 中,等式(48)的差异如下:

$$\Delta H_{1,2}=-\frac{1}{2}(k_1-k_2)\left(\gamma_1\log\left(\frac{p_\kappa^{\text{in}}}{p_\kappa^{\text{out}}}\right)-\gamma_2\log\left(\frac{1-p_\kappa^{\text{in}}}{1-p_\kappa^{\text{out}}}\right)\right) \tag{52}$$

在这个例子中,假设 $\gamma_1,\gamma_2>0$, $p_\kappa^{\text{in}}>p_\kappa^{\text{out}}$,那么如果 $k_2>k_1$ 则 $\Delta H_{1,2}>0$,反之亦然. 所以,为了使得 $H(t)$ 最大,这个新节点应该划归到它能获得更多连接的社团中.

根据定理 2 和定理 3,我们可以通过使用动态系统(16)来最大化一个质量函数. 这里,我们展示一种实现最大化的更加便捷的做法.

定理 4. 假设质量函数形如等式(30),并且对于所有的 i,μ,k,在任意一个硬划分情况下有下列条件:

$$\frac{\partial Q_{ik}(t)}{\partial x_{i\mu}(t)}=0 \tag{53}$$

那么最大化质量函数(30)得到硬划分归属 X,便是动态系统(54)的一个渐进稳定点.

$$x_{i\mu}(t+1)=\frac{x_{i\mu}(t)\mathrm{e}^{-\frac{\partial H(t)}{\partial x_{i\mu}(t)}}}{\sum_{\kappa=1}^{c}x_{i\kappa}(t)\mathrm{e}^{-\frac{\partial H(t)}{\partial x_{i\mu}(t)}}} \tag{54}$$

证明. 在等式(30)两边求导,我们可以得到

$$\frac{\partial H(t)}{\partial x_{i\mu}(t)}=\frac{1}{2}\left(Q_{i\mu}(t)+\sum_\kappa x_{i\kappa}(t)\frac{\partial Q_{i\kappa}(t)}{\partial x_{i\mu}(t)}\right) \tag{55}$$

假设 $X(t)$ 是一个硬划分,通过假设 $\frac{\partial Q_{ik}(t)}{\partial x_{i\mu}(t)}=0$,我们可以得到

$$\frac{\partial H(t)}{\partial x_{i\mu}(t)}=\frac{1}{2}Q_{i\mu}(t) \tag{56}$$

从而根据等式(56)和 $\frac{\partial Q_{ik}(t)}{\partial x_{i\mu}(t)}=0$,我们可以推断出动态系统(16)与动态系统(54)的 Jacobian 行列式在 $X(t)$ 上是一致的. 因此,如果 $X(t)$ 是最大化 H 得到的结果,那么根据定理 3 它就是等式(16)的一个渐进稳定点,从而也就是系统(54)的一个渐进稳定点. 证毕.

推论 1. 如果网络参数,即 p_μ^{in}, p_μ^{out} 和 n_μ, $\mu=\{1,2,\cdots,c\}$ 是已知的,并且最大化函数 H 能够得到硬划分 $X(t)$,那么 $X(t)$ 就能通过迭代动态系统(54)获得.

本文提出的算法是以定理 4 为基础的. 为此我们需要计算 $\frac{\partial H(t)}{\partial x_{i\mu}(t)}$ 使得 $\frac{\partial Q_{ik}(t)}{\partial x_{i\mu}(t)}=0$. 根据式(46),通过重整和变形,我们可以得到 $\frac{\partial H(t)}{\partial x_{i\mu}(t)}$ 为

$$\begin{aligned}\frac{\partial H(t)}{\partial x_{i\mu}(t)}=&\frac{1}{2}\gamma_1\left[(2l_\mu^{\text{in}})\frac{\partial\log\left(\frac{p_\mu^{\text{in}}}{p_\mu^{\text{out}}}\right)}{\partial x_{i\mu}(t)}+2k_{i\mu}\log\left(\frac{p_\mu^{\text{in}}}{p_\mu^{\text{out}}}\right)\right]+\\&\frac{1}{2}\gamma_2\left[\left(\frac{2l_\mu^{\text{in}}(1-p_\mu^{\text{in}})}{p_\mu^{\text{out}}}\right)\frac{\partial\log\left(\frac{1-p_\mu^{\text{in}}}{1-p_\mu^{\text{out}}}\right)}{\partial x_{i\mu}(t)}+\right.\\&\left.2(n_\mu-k_{i\mu}-x_{i\mu})\log\left(\frac{1-p_\mu^{\text{in}}}{1-p_\mu^{\text{out}}}\right)\right]+\\&\frac{1}{2}\left[\left(\sum_a k_{a\mu}\right)\frac{\partial\log(p_\mu^{\text{out}})}{\partial x_{i\mu}(t)}+k_i\log(p_\mu^{\text{out}})\right]+\\&\frac{1}{2}\left[(n_\mu(n-1)-\sum_a k_{a\mu})\frac{\partial\log(1-p_\mu^{\text{in}})}{\partial x_{i\mu}(t)}+\right.\\&\left.(n-k_i-1)\log(1-p_\mu^{\text{in}})\right]+\log\left(\frac{n_\mu}{n}\right)+1\end{aligned} \tag{57}$$

为了满足准则(53), γ_1 和 γ_2 应该被适当地选取. 从数值上来说,我们发现 $\gamma_1=\frac{1}{p_\mu^{\text{in}}}$ 和 $\gamma_2=1$ 能近似完美地满足这个准则. 通过将等式(57)中的这些系数替换,我们得到

$$\begin{aligned}\frac{\partial H(t)}{\partial x_{i\mu}(t)}=&\frac{1}{2p_\mu^{\text{in}}}\left[(2l_\mu^{\text{in}})\frac{\partial\log\left(\frac{p_\mu^{\text{in}}}{p_\mu^{\text{out}}}\right)}{\partial x_{i\mu}(t)}+2k_{i\mu}\log\left(\frac{p_\mu^{\text{in}}}{p_\mu^{\text{out}}}\right)\right]+\\&\frac{1}{2}\left[\left(\frac{2l_\mu^{\text{in}}(1-p_\mu^{\text{in}})}{p_\mu^{\text{out}}}\right)\frac{\partial\log\left(\frac{1-p_\mu^{\text{in}}}{1-p_\mu^{\text{out}}}\right)}{\partial x_{i\mu}(t)}+\right.\\&\left.2(n_\mu-k_{i\mu}-x_{i\mu})\log\left(\frac{1-p_\mu^{\text{in}}}{1-p_\mu^{\text{out}}}\right)\right]+\\&\frac{1}{2}\left[\left(\sum_a k_{a\mu}\right)\frac{\partial\log(p_\mu^{\text{out}})}{\partial x_{i\mu}(t)}+k_i\log(p_\mu^{\text{out}})\right]+\\&\frac{1}{2}\left[(n_\mu(n-1)-\sum_a k_{a\mu})\frac{\partial\log(1-p_\mu^{\text{in}})}{\partial x_{i\mu}(t)}+\right.\\&\left.(n-k_i-1)\log(1-p_\mu^{\text{in}})\right]+\log\left(\frac{n_\mu}{n}\right)+1\end{aligned} \tag{58}$$

在式(58)中, $n_\mu(t)$, $p_\mu^{\text{in}}(t)$ 和 $p_\mu^{\text{out}}(t)$ 都是以 $x_{i\mu}(t)$ 为自变量的方程,具体形式请见式(11)、(14)和式(15).

为了实现社团划分,接下来我们提出基于迭代动态系统(54)的快速算法(算法 1). 此外,在每一个步骤中,动态系统(54)以一个预先确定好的步数进行迭代,这也就是动态迭代. 具体的算法流程如算法 1 所示.

算法 1. 社团划分算法流程.

输入:网络 G,其节点数为 n,边数为 m. 最大迭代步数 $R_{\max}$

输出:社团归属矩阵 $\boldsymbol{X}$

1. 初始化 $X(0)$ 并且设置 X_{best}(在 4.2 节中阐述);

2. 重复

3. 利用等式(11)～(15)分别更新 n_μ, l_μ^{in}, l_μ^{out}, p_μ^{in} 和 p_μ^{out};

4. 利用式(58)计算$\frac{\partial H(t)}{\partial x_{iu}(t)}$;

5. 利用式(54)迭代归属矩阵 $\boldsymbol{X}(t)$;

6. 利用式(46)计算函数 $LP(t)$;

7. 如果已经达到迭代步数上限 $R_{\max}$,则转向步 8;否则,回到步 2;

8. 将满足 $LP(t)$最大化的归属 $\boldsymbol{X}(t)$记为 $\boldsymbol{X}_t$;

9. 如果 $LP(X_t)>LP_{\text{best}}$,令 $LP_{\text{best}}=LP(X_t)$并且令 $X_{\text{best}}=X_t$;

10. 如果达到迭代最大步数 $R_{\max}$,返回 X_{best};

否则,回到步 1.

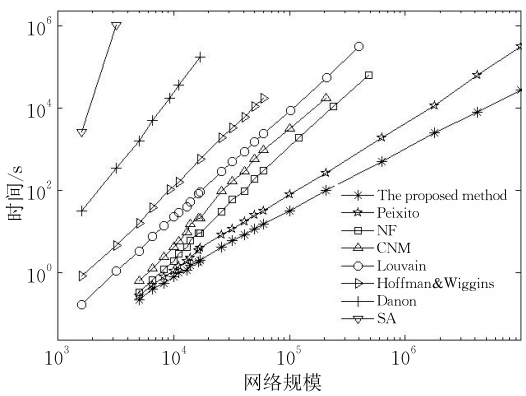
4.1 算法复杂度

在每步动态迭代中,计算对数似然函数(46)的复杂度是 $O(m\cdot c_{\max})$,其中 $c_{\max}$是社团数量的最大值.同时,对于拥有强社团结构的网络,该动态过程的收敛速度是相当快的.动态迭代的次数 $R_{\max}$并不直接取决于节点的数量,实际上,它与社团结构的模糊程度是相关的.因此,算法总的计算复杂度为 $O(m\cdot c_{\max})$,而对于稀疏网络,更会下降为 $O(n\cdot c_{\max})$.另外,计算复杂度与网络的拓扑结构相关.在许多网络中,社团的数目比网络的节点数目要小得多,例如空手道俱乐部网络[46],在这些实例中,对于稀疏网络,算法的计算复杂度为 $O(n)$.表 2 列出了一些著名算法的计算复杂度,通过比较可以看出我们的算法是非常快速的.然而,在一些特殊的网络中,社团的数量和网络的节点数量是相关的,例如团环网络(ring of cliques),其中的每一个团只有很少数量的节点.在这些网络中,复杂度仍然为 $O(n\cdot c_{\max})$.

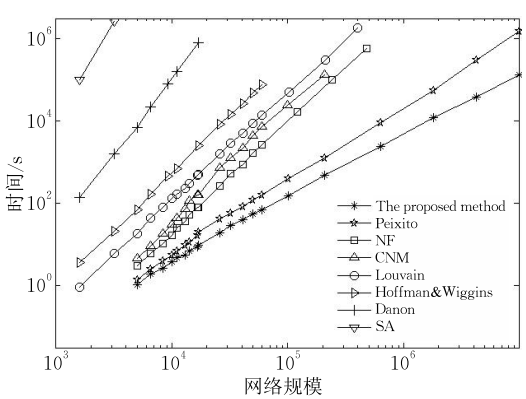
表 2 在稀疏网络中,著名算法计算复杂度分析比较

算法	计算复杂度
CNM[38]	$O(n\log^2 n)$
DA[39]	$O(n^2\log n)$
Louvain[10]	$O(n\log n)$
OCR-HK[37]	$O(n^2)$
Bayesian inference[22]	$O(n\log^a n)$
Variational Bayesian[16]	$O(n^{1.44})$
RN Potts model[17]	$O(n^{1.3})$

为了检验算法的性能,我们与 7 个著名的算法进行比较,其中包括 Louvain 方法[10]、Newman Fast(NF)方法[14]、Hoffman&Wiggins 方法[16]、SA 方法[20]、Peixito 方法[33]、Danon 方法[39]和 CNM 方法[40].同时我们也考虑网络的规模,检验其与算法性能的关系.这里,实验环境是一台拥有 2 GHz CPU 和 4 GB 内存的台式电脑,运行系统是 Windows 7,程序软件是 Matlab2010b.为了进行测试,我们利用 LFR 模型[23,42]生成稀疏网络,该模型中的社团结构是提前定义好的,但是网络中节点的连接情况是由参数控制的,包括节点数目 n、节点的平均度$\langle k\rangle$、最大节点度 $\max_k$、混合比例参数(Mixing Parameter)θ(每个节点与其他社团的节点共享 θ 比例的边)、最小社团规模 $\min_c$和最大社团规模 $\max_c$.在 LFR 网络中,混合比例参数(Mixing Parameter)θ用来调整网络的模糊性.θ的变化范围为$\lfloor 0,1\rfloor$,较大的 θ代表更加弱的社团结构.我们统一设定实验中网络的节点平均度$\langle k\rangle=8$(和 WWW 网络接近,为典型的稀疏网络)、最大节点度 $\max_k=20$、最小社团规模 $\min_c=10$ 和最大社团规模 $\max_c=100$(在网络规模固定的情况下,$\min_c$和 $\max_c$控制网络中社团的个数).为了证明结果的一般性,我们分别在 $\theta=0.2$ 和 $\theta=0.6$ 的两个 LFR 网络上进行试验,结果展示在图 2 中.首先看出我们的算法在所有的网络规模中都比其他的方法快;其次,我们方法的计算时间几乎是是随着网络的规模线性扩展的,因此可以在拥有数百万甚至数亿节点的大规模网络中进行实现.



(a) $\theta=0.2$



(b) $\theta=0.6$

图 2 不同规模的 LFR 网络下,各个算法运行时间的比较

4.2 初始化社团归属矩阵

在算法 1 的步 1 中,社团归属矩阵应该给定初始化形式.我们设计如下的方法,首先,初始归属矩

阵 $\boldsymbol{x}_{i\mu}(0)$被设定为

$$\boldsymbol{x}_{i\mu}(0)=\frac{1}{c}+y_{i\mu};\ i=\{1,2,\cdots,n\},\ \mu=\{1,2,\cdots,c\}$$
$$\text{s.t.}\ \sum_{\mu=1}^{c}y_{i\mu}=0 \tag{59}$$

其中：$y_{i\mu}$是一个小的高斯噪声. 这个策略可以避免在性质 1 中描述的第 2 类平凡解. 在接下来的步骤中，受益于之前迭代过程的最好结果，我们有以下改进策略：

$$x_{i\mu}(0)=x_{i\mu}^{\text{best}}+y_{i\mu};i=\{1,2,\cdots,n\},\mu=\{1,2,\cdots,c\}$$
$$\text{s.t.}\ \sum_{\mu=1}^{c}y_{i\mu}=0 \tag{60}$$

其中：$x_{i\mu}^{\text{best}}$是 X_{best}的(i,μ)元素. 换句话说，此过程通过迭代渐进地去改进结果. 值得注意的是，此步的高斯噪声的强度，即 $y_{i\mu}$，必须比上一次(式(59))的大很多；否则，动态过程将会收敛到之前的结果.

4.3　确定最优的社团个数

在运行算法之前，我们应当预先设定社团的数量，因为它并不是一个显著的先验信息. 在许多著名算法中，社团数量被隐式地提前指定，因为不同的社团个数会对划分质量产生极大地影响. 例如，CNM 算法[38]和 Louvain 算法[10]所考虑的初始值为 n，而对于二分迭代算法如 SA[20]和 DA[39]，初始值则设为 2.

在算法 1 中，我们只需要设定社团的初始个数为一个大于或者等于实际社团个数的值，因为额外的社团会随着动态系统在迭代过程中逐渐被合并掉，因此，这是一个粗糙且容易实现的步骤.

5　实　验

本文算法不仅在人工基准网络有良好的性能，另外在真实数据网络[47-53]特别是大规模网络上，如大型科学家合作网络，有着出色的表现.

5.1　人工网络

首先，我们将算法应用到著名的 GN 和 LFR 基准网络[27,41]. 实验结果表明，即使在含有模糊社团结构的网络中，算法也具有非常高的准确性.

图 3 展示了在 GN 基准网络[14,40]中不同算法性能的比较结果. GN 网络由 Newman 等人提出，共包含 $n=128$ 个节点，分成 4 个社团，每个社团包含 32 个节点. 假设属于相同社团的节点对以概率 Z_{in} 相连接，而属于不同社团的节点对以概率 Z_{out} 相连接. 网络中点的平均节点度固定为$\langle k\rangle=16$，可以得到 Z_{in}和 Z_{out}有关系 $31Z_{\text{in}}+96Z_{\text{out}}=16$. 我们发现，当 $Z_{\text{out}}\leqslant 7.4$ 时，基本上所有算法都具有较高的性能($NMI\geqslant 0.75$). 随着 Z_{out}的增加，基本上所有算法的划分效率都会急剧降低，但是本文算法直到 $Z_{\text{out}}\to 8$ 时依然具有最高的准确率，从而说明其划分能力的高效性.

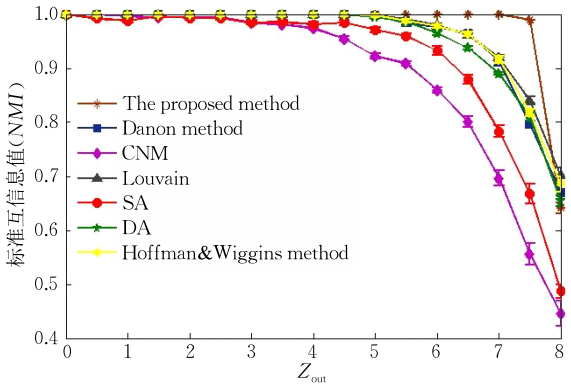


图 3　GN 网络中 7 种不同算法的标准互信息值(NMI)对比(图中的每个点表示 50 次不同实验结果的平均值)

为了进一步进行验证，我们将算法应用到 LFR 基准网络上. LFR 网络模型由 Lancichinetti 等人[23,42]提出，具有无标度特征的度分布和社团规模分布，因此更加接近现实中社会和生物系统. LFR 网络模型的生成由一些参数控制：节点数目 n、节点的平均度$\langle k\rangle$、最大节点度 $\max_k$、混合比例参数(Mixing Parameter)θ(每个节点与其他社团的节点共享 θ 比例的边)、最小社团规模 $\min_c$和最大社团规模 $\max_c$. θ 的变化范围为$\lfloor 0,1\rfloor$，用来调整网络的模糊性，较大的 θ 代表更加弱的社团结构. 在本次实验中，我们统一设定网络的节点平均度$\langle k\rangle=10$、最大节点度 $\max_k=20$，每个社团的规模有大小之分(由参数 $\min_c$和 $\max_c$控制，从而在网络规模固定的情况下可以调节社团的个数). 结果如图 4 所示，其中当网络规模为 $n=1000$ 时，其拥有的小(大)社团个数为 100(50)，当网络规模为 $n=5000$ 时，其拥有的小(大)社团个数为 300(200). 我们可以发现，在每一种情况下，本文算法都具有最高的标准互信息值(NMI)值.

另外，我们还将算法应用到包含 1000 个社团，每个社团包含 10 个节点的团环网络(ring of cliques)上. 这种团环网络[11]可以用来说明模块度指标(和一些其他的指标)有分辨率限制问题(resolution limit problem). 团环网络利用少量的边将一些团(完全图)连接起来，并使用团的数目和每个团的规模这两个参数进行控制. 结果显示，本文算法不仅能精确地发现每个团，而且不具有分辨率限制问题，进一步验证了算法的高效性.

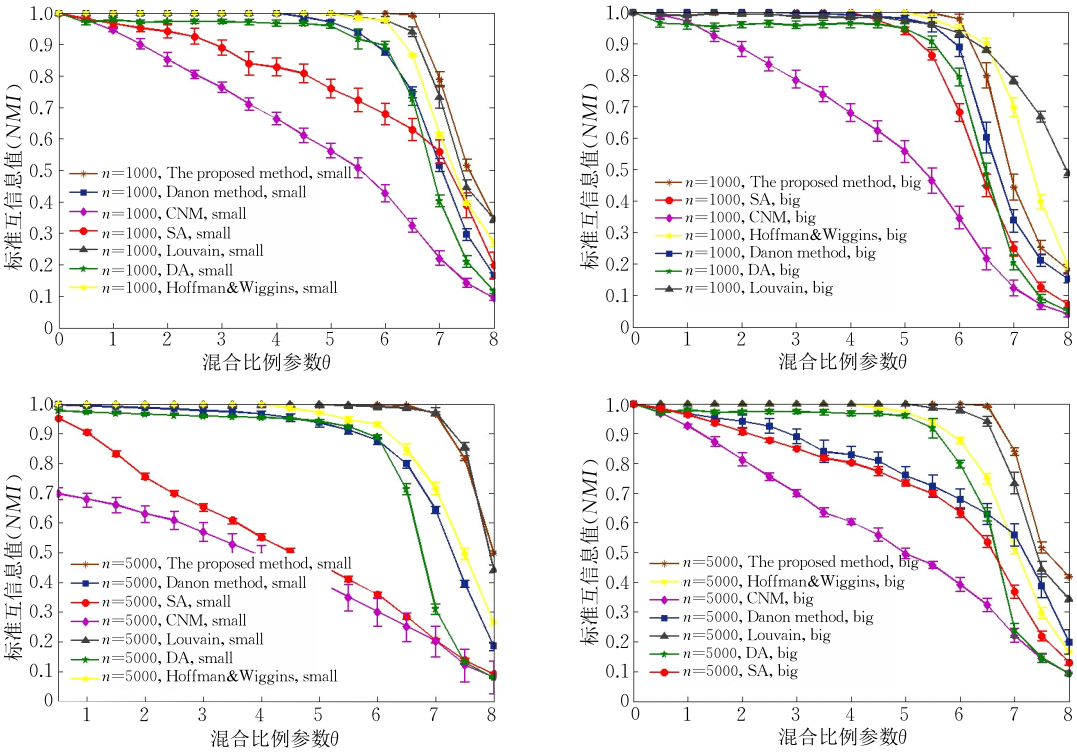


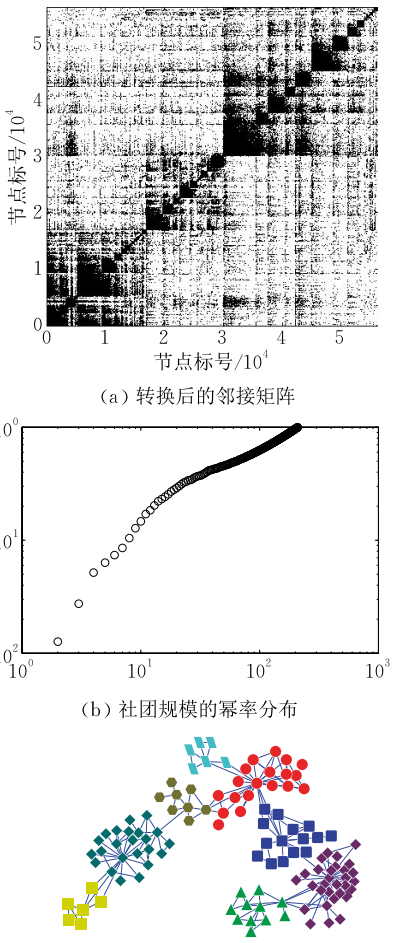
图 4 本文算法与一些著名算法的性能对比(每个点为 50 次计算结果的平均值)

5.2 科学家合作网络

进一步,为了验证算法的高效性,我们将算法应用到一个大规模网络——科学家合作网(Scientists Cooperation Network)[14,40] 上. 该网络包含 56 276 个点和 315 810 条加权边,代表了 56 276 位物理学家在预印本网站 arxiv.org 共同署名发表论文的合作关系. 我们根据算法的划分结果,输出转换后的邻接矩阵(相同社团内的节点被整合在一起,呈现出层次结构)进行可视化,如图 5(a)所示. 可以看出,和许多社会网络一样,科学家合作网呈现出较强的社团特征和层次特征,其中 3 个最大的社团具有现实意义,分别代表物理的主要研究领域:核物理、天体物理和凝聚态物理.

通过算法,我们从网络中共挖掘出 696 个社团. 图 5(b)以幂率坐标的形式展示了社团的规模分布情况,可以看出,这是一种典型的无标度分布,这和许多社会网络一致. 最大社团节点数为 202,最小社团节点数为 2,平均社团节点数为 81. 社团节点分布呈现不均匀特征,5.9% 的大规模社团包含大约 30.2%的节点,而剩余社团的规模均较小.

进一步,为了验证社团划分的准确性,我们截取并展示了一个包含 8 个社团的子图片段,不同的社团不用的形状表示,如图 5(c)所示. 可以容易看出,算法的划分结果非常好,而且该结果与文献[40-41]



(a) 转换后的邻接矩阵

(b) 社团规模的幂率分布

(c) 8个社团的子图片段,不同社团用不同的形状表示

图 5 算法在科学家合作网上的运算结果

完全相同，因此表明其非常适合在大规模网络上进行运算，对进一步研究分析大数据问题具有极其重大的价值.

5.3　更多现实网络

最后，为了更加充分地验证我们的算法，我们在7个著名的现实世界网络上进行试验，结果在表3中展示.为了方便地和现有方法进行比较，我们展示了现有方法在这些网络中得到的最高模块度(Modularity)Q值，这些最高值是通过比较大量划分算法后得到的，相关网络和最高Q值的文献出处也被展示出来.

表3　在现实网络中算法性能的分析比较

网络	网络参考文献	现有方法的最高Q值	最高Q值参考文献	本文算法的Q值
Email	[20]	0.579	[54]	0.568
US Football	[40]	0.606	[54]	0.602
Zachary Karate	[46]	0.420	[54]	0.416
Les Miserables	[50]	0.561	[54]	0.554
Dolphin	[51]	0.531	[54]	0.527
Jazz	[52]	0.446	[54]	0.439
PGP-Key signing	[53]	0.878	[54]	0.883

从表3中可以看出，本文算法的Q值非常接近现有方法的最高Q值，而且我们的算法具有一般方法所不具有的线性时间，其计算复杂性非常小.值得一提的是，在PGP-Key signing network上，本文算法比现有方法的最高值还高，充分展示了本文算法的高效性.

表4　在大规模网络数据中的算法性能分析

网络	n	m	R_g	R_m	NMI
Amazon	334863	925872	75149	73961	0.404
Dblp	317080	1049866	13477	14531	0.412
Youtube	1134890	2987624	8385	10156	0.337
Livejournal	3997962	34681189	287512	176351	0.304
Orkut	3072441	117185083	6288363	5963451	0.376
Friendster	65608366	1806067135	957154	937640	0.424

为了进一步验证算法应用的广泛性，我们在6个大规模网络数据上进行实验验证.这6个大规模网络数据是从Stanford Network Analysis Platform (SNAP)[①]下载的，并且提供了真实的网络划分以供比较验证.其中Amazon网络是一个产品网络，其中两个产品如果经常被同时购买，那么它们之间就有一条边；Dblp网络展示了科学家之间共同发表论文的关系；Youtube，Livejournal，Orkut和Friendster网络则是从不同社交网站中提取的社会关系网络.

我们将划分结果和一些网络关键属性在表4中展示出来，其中n和m分别代表网络的点数和边数，R_g和R_m分别代表真实社团数目和划分社团数目，NMI代表互信息值的值[23,42].通过比较很多现存的算法[54-56]，我们发现本文算法具有较高的准确性，非常适合于在百万节点以上的网络中进行实验分析.值得一提的是，我们发现划分社团数目R_m和真实社团数目R_g非常接近，从而验证了本算法具有非常好的模型选择功能.

6　结　论

本文提出了一种高效的动态社团划分算法，通过引入一种新型的基于离散时间的动态系统，来描述社团归属的从随机状态到最优划分的动态轨迹，并利用严格的数学分析找出了社团归属动态收敛到最优的条件.另外，本文还创新性地提出了一种划分指标函数的一般化形式，通过选择不同的参数，可以拓展到很多著名的划分指标函数.本算法非常高效，计算复杂度分析表明在稀疏网络中，算法需要的时间与网络节点数目呈线性关系.

此外，我们还可以对算法的应用进行更深入的探讨，例如：(1)如何进一步加快算法速度，使得其能够在超大规模的网络上准确划分社团，如包含几亿到几十亿节点的微博网络和生物网络上；(2)由于现实世界并不都是双向关系，网络中会出现单向或者混合方向的边，如何利用算法高效地处理这种混杂网络，是一个挑战.

致　谢　中国科学院管理学院石勇教授、中央财经大学刘志东教授对本文提出了很多建设性建议，特表示感谢.最后，感谢评审老师细致耐心的审查！

参　考　文　献

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

[2] Gong Mao-Guo, Cai Qing, Chen Xiao-Wei, Ma Li-Jia. Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition. IEEE Transactions on Evolutionary Computation, 2014, 18(1): 82-97

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

[4] Rubio-Largo A, Vega-Rodriguez M A, Gomez-Pulido J A, Sanchez-Perez Juan M. Multiobjective metaheuristics for

① http://snap.stanford.edu/data/index.html

traffic grooming in optical networks. IEEE Transactions on Evolutionary Computation, 2013, 17(4): 457-473

[5] Rolland T, Tasan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. Cell, 2014, 159(5): 1212-1226

[6] Liu Ying, Moser J, Aviyente S. Network community structure detection for directional neural networks inferred from multichannel multisubject EEG data. IEEE Transactions on Biomedical Engineering, 2014, 61(7): 1919-1930

[7] Gharavi H, Hu Bin. Multigate communication network for smart grid. Proceedings of the IEEE, 2011, 99(6): 1028-1045

[8] Tremblay N, Borgnat P. Graph wavelets for multiscale community mining. IEEE Transactions on Signal Processing, 2014, 62(20): 5227-5239

[9] Yang Bo, Liu Ji-Ming, Feng Jian-Feng. On the spectral characterization and scalable mining of network communities. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(2): 326-337

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

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

[12] Khadivi A, Rad A A, Hasler M. Network community detection enhancement by proper weighting. Physical Review E, 2011, 83(4): 046104

[13] Zhang Xiang-Sun, Wang Rui-Sheng, Wang Yong, et al. Modularity optimization in community detection of complex networks. Europhysics Letters, 2009, 87(3): 38002

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

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

[16] Hofman J M, Wiggins C H. Bayesian approach to network modularity. Physical Review Letters, 2008, 100(25): 258701

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

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

[19] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007, 76(3): 036106

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

[21] Duch J, Arenas A. Community detection in complex networks using extremal optimization. Physical Review E, 2005, 72(2): 027104

[22] Hastings M. B. Community detection as an inference problem. Physical Review E, 2006, 74(3): 035102

[23] Lancichinetti A, Fortunato S. Community detection algorithms: A comparative analysis. Physical Review E, 2009, 80(5): 056117

[24] Nadakuditi R, Newman M E J. Graph spectra and the detectability of community structure in networks. Physical Review Letters, 2012, 108(18): 188701

[25] Radicchi F. Driving interconnected networks to supercriticality. Physical Review X, 2014, 4(2): 021014

[26] Radicchi F. A paradox in community detection. Europhysics Letters, 2014, 106(3): 38001.

[27] Radicchi F. Detectability of communities in heterogeneous networks. Physical Review E, 2013, 88(1): 010801

[28] Zhang Pan, Moore C. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. Proceedings of the National Academy of Sciences, 2014, 111(51): 18144-18149

[29] Rosvall M, Bergstrom C. T. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 2008, 105(4): 1118-1123

[30] Esquivel A V, Rosvall M. Compression of flow can reveal overlapping modular organization in networks. Physical Review X, 2011, 1(2): 021025

[31] Peixoto T P. Parsimonious module inference in large networks. Physical Review Letters, 2013, 110(14): 148701

[32] Li Hui-Jia, Li Hui-Ying, Li Ai-Hua. Analysis of multi-scale stability in community structure. Chinese Journal of Computers, 2015, 38(2): 301-312(in Chinese)
(李慧嘉,李慧颖,李爱华. 多尺度的社团结构稳定性分析. 计算机学报, 2015, 38(2): 301-312)

[33] Peixoto T P. Model selection and hypothesis testing for large-scale network models with overlapping groups. Physical Review X, 2015, 5(1): 011033

[34] Aldecoa R, Marin I. Surprise maximization reveals the community structure of complex networks. Scientific reports, 2013, 3: 1060

[35] Li Hui-Jia and Zhang Xiang-Sun. Analysis of stability of community structure across multiple hierarchical levels. Europhysics Letters, 2013, 103(5): 58002

[36] Stanoev A, Smilkov D, Kocarev L. Identifying communities by influence dynamics in social networks. Physical Review E, 2011, 84(4): 046102

[37] Boccaletti S, Ivanchenko M, Latora V, Pluchino A. Detecting complex network modularity by dynamical clustering. Physical Review E, 2007, 75(4): 045102

[38] Clauset A, Newman M. E. J, Moore C. Finding community structure in very large networks. Physical Review E, 2004, 70(6): 066111

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

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

[41] Jiang Qi-Xia, Zhang Yan, Sun Mao-Song. Community detection on weighted networks: A variational bayesian method. Advances in Machine Learning. Springer Berlin, Heidelberg, 2009, 5828: 176-190

[42] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4): 046110-046115

[43] Shen Hua-Wei, Cheng Xue-Qi, Chen Hai-Qiang, et al. Information bottleneck based community detection in network. Chinese Journal of Computers, 2008, 31(4): 677-686 (in Chinese)
(沈华伟，程学旗，陈海强等. 基于信息瓶颈的社区发现. 计算机学报，2008, 31(4): 677-686)

[44] Tibely G, Kertesz J. On the equivalence of the label propagation method of community detection and a Potts model approach. Physica A, 2008, 387(19-20): 4982-4984

[45] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[46] Zachary W. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1977, 33(4): 452-473

[47] Brownlee A E I, McCall J A. W, Zhang Q. Fitness modelling with Markov networks. IEEE Transactions on Evolutionary Computation, 2013, 17(6): 862-879

[48] Echegoyen C, Mendiburu A, Santana R, Lozano J A. Toward understanding EDAs based on Bayesian networks through a quantitative analysis. IEEE Transactions on Evolutionary Computation, 2012, 16(2): 173-189

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

[50] Knuth D E. The Stanford GraphBase: A Platform for Combinatorial Computing, Reading. New York, USA: Addison-Wesley Professional, 1993, 37: 592

[51] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 2003, 54(4): 396-405

[52] Gleiser P, Danon L. List of edges of the network of jazz musicians. Advances in Complex Systems, 2003, 6: 565

[53] Boguna M, Pastor-Satorras R, Diaz-Guilera A, Arenas A. Models of social networks based on social distance attachment. Physical Review E, 2004, 70(5): 056122

[54] Agarwal G, Kempe D. Modularity-maximizing graph communities via mathematical programming. European Physical Journal B, 2008, 66(3): 409-418

[55] Bhatia N P, Szego G P. Stability Theory of Dynamical Systems. Berlin Heidelberg: Springer-Verlag, 2002

[56] Yang J, Leskovec J. Defining and evaluating network communities based on ground truth. Knowledge and Information Systems, 2015, 42(1): 181-213

LI Hui-Jia, born in 1985, Ph. D., associate professor. His research interests include data mining, complex networks, and information retrieval.

LI Ai-Hua, born in 1978, Ph. D., associate professor. Her research interests include data mining and management decision sciences.

LI Hui-Ying, born in 1983, Ph. D., research assistant. Her research interests include bioinformatics, complex networks, and data mining.

Background

Detecting community structure in networks has broad applications since nodes in the same community generally possess mutual properties or relationships relative to nodes interconnecting different communities. Typical examples include a social circle in which members have common interests, a group of proteins that work together for functional realization, and tweets under the same topic spreading the opinion. Distinguishing the communities facilitates the operation, control, and optimization of networked systems, such as developing tools for precision marketing, identifying target points for drug discovery, searching and mining online social networks for trend predictions and so on.

The traditional optimization or heuristic methods are usually used in present techniques, with the common logic of comparing the intuitive properties and external behaviors in and outside of the communities. However, to obtain an

acceptable accuracy, these methods usually have a high-level computational complexity. In this paper, we propose a new algorithm using dynamical system to realize the fast and exact detection of communities in complex networks. First, a discrete-time dynamical system is introduced to describe the assignment of community memberships, and the conditions driving the convergence of dynamics trajectory to the optimal situation is formulated. Further, we design a new type graph generative model, which performs the algorithm free of the parameters. By analyzing the eigenvalue gap of the Markovian transition matrix, we provide a mathematical theory to identify the optimum number of communities in a network, and the stability of community structure. Our algorithm can also be generalized to unify the conventional algorithms widely applied. Our algorithm is highly efficient: the computational complexity analysis shows that the required time is linearly dependent on the number of all nodes in a sparse network.

This research is supported by National Nature Science Foundation under Grant Nos. 91324203, 11131009, 71401194 and 71401188. The first two foundations aim to the study on complex networks from many scientific fields. The main focus is on three aspects: the statistical properties of complex networks, the model of the evolution of complex networks, the dynamics of complex networks. The study of community structure properties is a fundamental task of these works. The last two foundations aim at studying community structure characteristics in social network specifically. The authors have made some in-depth researches on the measure of community structure, the method to uncover the hierarchical and overlapping community structure in large networks, and its application on finding the intrinsic properties of community structure.