基于随机网络集成模型的广义网络社区挖掘算法

杨 博 [1, 2] 刘 杰 [1, 2] 刘大有 [1, 2]

摘 要 根据结点的属性和链接关系, 现实世界中的复杂网络大多可分为同配网络和异配网络, 社区结构在这两类网络中均普遍存在. 准确地挖掘出两种不同类型网络的社区结构具有重要的理论意义和广泛的应用领域. 由于待处理的网络类型通常未知, 因而难以事先确定应当选择何种类型的网络社区挖掘算法才能获得有意义的社区结构. 针对该问题, 本文提出了广义网络社区概念, 力图将同配和异配网络社区结构统一起来. 本文提出了随机网络集成模型, 进而提出了广义网络社区挖掘算法 G - NCMA. 实验结果表明: 该算法能够在网络类型未知的前提下准确地挖掘出有意义的社区结构, 并能分析出所得社区的类型特征.

关键词 复杂网络, 社区挖掘, 随机网络, 极大似然估计

引用格式 杨博, 刘杰, 刘大有. 基于随机网络集成模型的广义网络社区挖掘算法. 自动化学报, 2012, **38**(5): 812−822

DOI 10.3724/SP.J.1004.2012.00812

A Random Network Ensemble Model Based Generalized Network Community Mining Algorithm

YANG Bo[1, 2] LIU Jie[1, 2] LIU Da-You[1, 2]

Abstract According to the attributes of nodes and the linkages between them, most real-world complex networks could be assortative and disassortative. Community structures are ubiquitous in both types of networks. The ability to discovery meaningful community structures from both types of networks is fundamental for theoretical research and practical applications. Since the types of exploratory networks to be processed are usually unknown beforehand, it is difficult to determine what specific algorithms should be applied to them to obtain meaningful community structures. To address this issue, a novel concept of generalized network community is proposed in order to unify two concepts of assortative and disassortative communities. Based on a random network ensemble model, a generalized community mining algorithm, called G - NCMA, is proposed. Experimental results demonstrate that the G - NCMA algorithm is able to properly mine potential communities from explorative networks, as well as to determine their respective types.

Key words Complex network, community mining, random network, maximum likelihood estimation

Citation Yang Bo, Liu Jie, Liu Da-You. A random network ensemble model based generalized network community mining algorithm. *Acta Automatica Sinica*, 2012, **38**(5): 812−822

社区结构是复杂网络最普遍的拓扑结构之一. 研究表明, 包括社会网络和生物网络在内的很多复杂网络都具有明显的社区结构 — 同区结点相互作用强而异区结点相互作用弱[1−2]. 网络社区结构挖掘算法能够帮助人们理解一个复杂的网络是如何基于一些基本网络构造模块组合而成的, 这对于深入理解网络拓扑结构、挖掘隐含模式和预测网络行为都具有十分重要的意义, 广泛应用于多个领域.

如果将网络社区挖掘问题看作是一个图分割问题, 则它是一个 NP 完全问题. 现有工作都试图给出一种对效率和精度进行很好折中的近似算法. 根据基本工作原理, 现有主要的网络社区挖掘算法可归结为两大类: 一类是基于优化的算法, 另一类是启发式算法. 前者通过优化预先定义的目标函数获取合理的社区结构, 如基于模块度 (Modularity) 的算法[2−5]; 后者没有显式的优化目标而采用预定义的启发式规则获取合理的社区结构[1, 6−8].

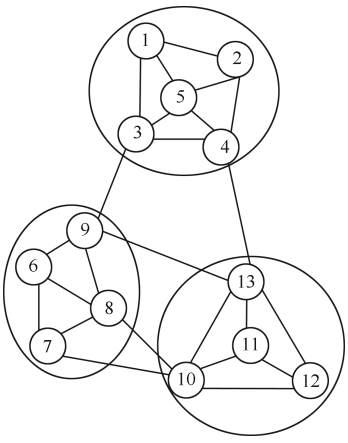
以上所述的社区挖掘算法都是针对同配网络, 而不能有效处理异配网络. 同配网络 (Assortative network) 是指与同等规模的随机网络相比, 网络中属性相似的结点具有较高的互连概率, 而属性相异

收稿日期 2010-12-21 录用日期 2011-12-19
Manuscript received December 21, 2010; accepted December 19, 2011
国家自然科学基金 (60873149, 60973088, 61133011, 61170092), 模式识别国家重点实验室开放课题, 中央高校基本科研业务费专项资金 (200903177), 教育部新世纪优秀人才支持计划 (NCET-11-0204) 资助
Supported by National Natural Science Foundation of China (60873149, 60973088, 61133011, 61170092), the Open Project Program of the National Laboratory of Pattern Recognition, the Fundamental Research Funds for the Central Universities (200903177), and Program for New Century Excellent Talents in University (NCET-11-0204)
本文责任编委 吕金虎
Recommended by Associate Editor LV Jin-Hu
1. 吉林大学计算机科学与技术学院 长春 130012 2. 吉林大学符号计算与知识工程教育部重点实验室 长春 130012
1. College of Computer Science and Technology, Jilin University, Changchun 130012 2. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012

的结点具有较低的互连概率. 具有相反性质的网络称为异配网络 (Disassortative network)[9]. 现实世界中, 同配网络和异配网络都广泛存在. 社会网络是常见的同配网络, 网络中社会地位、兴趣爱好等属性相似的个体倾向于相互交互. 某些生物网络, 如一些类型的蛋白质交互网络是异配网络, 网络中具有不同功能的蛋白质倾向于相互交互与作用. 还有些网络同时表现出同配和异配的性质. 如在 Web 网络中, 主题相关的网页倾向于相互连接, 表现出同配性质; 而中心结点 (Hub) 倾向于指向权威结点 (Authority), 则表现出异配性质.

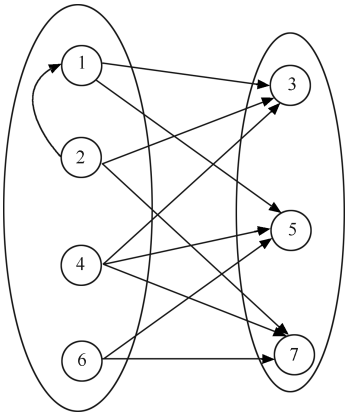
传统意义下的社区挖掘算法旨在根据网络拓扑结构将网络中的结点划分为 K 个不相交的社区, 使得同区结点相互作用强而异区结点相互作用弱. 作为传统社区挖掘问题的推广, 广义社区挖掘旨在根据网络拓扑结构将网络中的结点划分为不相交的 K 个社区, 使得被划分在相同社区中的结点具有相同或相似的性质. 对同配网络而言, 相同社区中的结点性质相似, 导致它们之间将会以较高的概率互连, 因而社区内的链接密度大, 而社区间的链接密度小. 同配网络的社区结构恰好对应传统意义下的社区结构. 对异配网络而言, 相同社区中的结点性质相似导致它们之间将会以较小概率互连, 而位于不同社区中的结点性质相异, 会以较大的概率互连, 因而社区内的链接密度小, 而社区间的链接密度大. 异配网络的社区结构恰好与传统意义下的社区结构相反. 图 1 给出了广义社区结构的一个简单示例. 图 1 (a) 是一个具有 3 个社区的同配网络, 社区内的连接密度远大于社区间的连接密度. 图 1 (b) 是一个具有 2 个社区的异配网络, 社区内的连接密度远小于社区间的连接密度.

现有社区挖掘方法大多仅能有效处理某一特定类型网络. 如采用前文所述的社区挖掘算法能够准确地挖掘出图 1 (a) 所示的同配网络社区结构, 但无法正确识别图 1 (b) 所示的异配网络社区结构.



(a) 具有 3 个社区的同配网络

(a) An assortative network with three communities



(b) 具有 2 个社区的异配网络

(b) A disassortative network with two communities

图 1　同配网络社区结构和异配网络社区结构示例

Fig. 1　The illustration of the community structures in assortative and disassortative networks

Newman 提出的 “模块度 (Modularity)” 概念被广泛用于刻画同配网络的社区结构, “模块” 值越大表明对应的同配网络社区结构越好. 很多社区挖掘算法都是最大化 “模块” 值的优化算法. Newman 进一步指出 “模块度” 也可用于刻画异配网络的社区结构[10], “模块” 值越小表明对应的异配网络社区结构越好. 据此, 可分别采用极大化或极小化 “模块” 值的优化方法挖掘出同配或异配网络中的社区结构. 然而, 采用以上思路的困难在于, 对于给定的网络, 通常难以事先判断它是同配网络还是异配网络, 因而无法确定应选择何种具体算法以获得正确挖掘结果.

Rosvall 和 Bergstrom 提出了基于信息论的网络社区挖掘方法 RB[11]. 假定网络的编码为 X, 根据网络社区结构对网络的压缩编码为 Y, 算法 RB 认为能够最小化条件熵 $H(X \mid Y)$ 与 Y 编码长度之和的社区结构为最优社区结构. 该工作以发现同配社区为主要目标, 但除了发现社区结构外, RB 还可发现中心 – 边缘结构. 当使用该算法从网络中挖掘社区结构或中心 – 边缘结构时, 需要明确告诉优化程序是否需要添加额外的 “链接约束”, 以惩罚不符合要求的网络划分. 因而, 尽管 RB 能够挖掘社区之外的结构, 但在其执行之前, 需要人工设置来决定挖掘何种结构.

Newman 等提出了基于混合模型 (Mixture model, MM) 的社区挖掘算法[12]. 在该算法中, 他们将社区看作是具有相似链接分布的结点构成的集合, 提出了基于链接分布表示网络似然的混合模型, 进而采用 EM (Expectation-maximization) 算法估计出模型参数, 获得网络的社区划分. 该工作探索了从未知类型的网络中挖掘有意义社区结构的方法, 对后续工作具有很好的借鉴意义.

首先给出网络的生成模型, 进而根据观察到的网络估计出生成模型的参数, 是 MM 算法的基本思想. 与该思想类似, 本文提出了广义社区挖掘方法 G - NCMA, 其基本思想是: 将网络社区挖掘问题看作随机网络集成问题, 即根据当前观察到的网络数据, 计算出对应网络出现似然最大的随机网络集成方案, 该方案中的每个随机子网络即为同配或异配网络中的一个社区. 该方法并没有根据社区内和社区间的链接密度差异去定义或者选择优化目标, 即没有显式地使用同配网络社区和异配网络社区的结构特征, 而是计算出一个能够最优拟合所观察网络数据的划分结构, 据此识别出不同类型网络中潜在有意义的社区结构.

本文提出的 G - NCMA 算法与 MM 算法都采用了极大似然思想设计优化目标, 但两种算法具有如下主要区别: 1) 优化目标函数的建模方法不同. G - NCMA 算法在计算网络似然时, 同时考虑了链接出现和不出现两种情况, 采用二重贝努利分布建立网络的似然函数; MM 算法在建立网络似然时仅考虑链接出现的情况, 是对真实似然的一种近似估计, 在非常稀疏或者社区结构定义良好的网络中, 由该近似产生的误差对计算结果的影响可被忽略. 2) 所采用的优化方法 (即参数估计方法) 不同. G - NCMA 采用局部搜索和模拟退火控制策略, 搜索优化目标的近似全局最优解, 而 MM 算法采用 EM 算法估计参数, 求出优化目标的局部最优解. 此外, EM 算法是一种点估计方法, 受初始输入和噪声影响较大. EM 算法的这些缺点导致 MM 算法的社区识别精度不高. 3) 相比于 MM 算法, G - NCMA 算法能够自动分析出各社区的类型特征, 以及输入网络的类型特征.

本文后面内容组织如下: 第 1 节提出随机网络集成模型; 第 2 节提出随机网络集成模型的参数估计算法, 即广义网络社区挖掘算法 G - NCMA; 第 3 节通过实验对 G - NCMA 算法进行了测试、比较、分析和评价; 第 4 节总结了本文的工作, 讨论了拟进一步开展的工作.

1 随机网络集成模型

随机网络集成模型的基本思想是: 从被观察网络出发, 基于机器学习的方法将其分割为指定数目的随机子网络, 这些随机子网络的集成能够最优或近似最优地拟合被观测网络的链接结构.

令 N 表示一个网络, $V(N)$ 表示网络中结点的集合, $E(N)$ 表示网络中边的集合. 令矩阵 A 表示网络 N 的邻接矩阵, 记为 $A=(a_{ij})_{n\times n}$, 其中 n 表示网络中的结点个数. 如果网络中存在从结点 i 到结点 j 的有向边 $\langle i,j\rangle$, 则 $a_{ij}=1$, 否则 $a_{ij}=0$.

令 $C=(C_1,\cdots,C_K)$ 表示网络 N 的一个 K–划分, 其中 $C_1,\cdots,C_K$ 表示 N 的 K 个子网络, 并且满足 $\bigcup_{i=1}^{K}V(C_i)=V(N)$ 和 $\bigcap_{i=1}^{K}V(C_i)=\varnothing$.

令 $P(x=j|y=k)$ 表示由结点 j 和子网络 C_k 定义的条件概率, 其含义为: 从 C_k 中任取一个结点 i, 网络中存在链接 $\langle i,j\rangle$ 的概率.

令 n 维向量 $\boldsymbol{t}_i=\langle a_{i,1},a_{i,2},\cdots,a_{i,n}\rangle$ 表示结点 i 与其他结点 (包括自身) 的拓扑链接结构. 令 $P(\boldsymbol{t}=\boldsymbol{t}_i)$ 表示拓扑结构 $\boldsymbol{t}_i$ 在网络 N 中出现的概率. 有如下的结论.

命题 1. 如果网络中的链接独立出现, C 为满足上述条件的一个 K–划分, 结点 i 属于子图 C_k, 则有:

$$P(\boldsymbol{t}=\boldsymbol{t}_i)=\frac{|C_k|}{n}\times \prod_{j=1}^{n}P(x=j|y=k)^{a_{ij}}(1-P(x=j|y=k))^{1-a_{ij}} \quad (1)$$

其中, $|C_k|$ 表示集合 C_k 中元素的个数.

证明. 由全概率式和条件概率式可得

$$P(\boldsymbol{t}=\boldsymbol{t}_i)=P(\boldsymbol{t}=\boldsymbol{t}_i\wedge(\vee_{l=1}^{K}(y=y_l)))=\sum_{l=1}^{K}P(\boldsymbol{t}=\boldsymbol{t}_i\wedge y=y_l)=\sum_{l=1}^{K}P(\boldsymbol{t}=\boldsymbol{t}_i|y=y_l)P(y=y_l)$$

因为结点 i 在子图 C_k 中, 且 $\bigcap_{l=1}^{K}V(C_l)=\varnothing$, 因此对 $l\neq k$, 有 $P(\boldsymbol{t}=\boldsymbol{t}_i|y=y_l)=0$. 进而,

$$P(\boldsymbol{t}=\boldsymbol{t}_i)=P(\boldsymbol{t}=\boldsymbol{t}_i|y=k)P(y=k)$$

先验概率可计算如下:

$$P(y=y_k)=\frac{|C_k|}{n} \quad (2)$$

因此有

$$P(\boldsymbol{t}=\boldsymbol{t}_i)=\frac{|C_k|}{n}P(\boldsymbol{t}=\boldsymbol{t}_i|y=y_k)$$

由链接独立出现的假设可得:

$$P(\boldsymbol{t}=\boldsymbol{t}_i|y=y_k)=\prod_{j=1}^{n}P(x=y|y=k)^{a_{ij}})(1-P(x=j|y=k))^{1-a_{ij}}$$

□

基于概率分布 $P(\boldsymbol{t}=\boldsymbol{t}_i)$, $1\leq i\leq n$, 网络 N 出现的

似然函数可定义为

$$\ln L(N) = \ln \prod_{i=1}^{n} P(\boldsymbol{t} = \boldsymbol{t}_i) \tag{3}$$

进而有

$$\begin{aligned}
&\ln L(N) = \sum_{i=1}^{n} \ln P(\boldsymbol{t} = \boldsymbol{t}_i) = \\
&\sum_{i=1}^{n} \ln \left(\frac{|C_{S(i)}|}{n} \prod_{j=1}^{n} \Big(P(x=j|y=S(i))^{a_{ij}} \times \right. \\
&\left. (1 - P(x=j|y=S(i)))^{1-a_{ij}} \Big) \right) = \\
&\sum_{i=1}^{n} \Bigg(\ln |C_{S(i)}| - \ln n + \\
&\sum_{j=1}^{n} \Big(a_{ij} \ln P(x=j|y=S(i)) + \\
&(1-a_{ij}) \ln(1 - P(x=j|y=S(i))) \Big) \Bigg) = \\
&\sum_{i=1}^{n} \Bigg(\ln |C_{S(i)}| - \ln n + \\
&\sum_{j=1}^{n} h(a_{ij}, P(x=j|y=S(i))) \Bigg)
\end{aligned} \tag{4}$$

其中, $S(i)$ 表示在划分 C 中结点 i 所在的子图序号, $h(x,y) = x \ln y + (1-x)\ln(1-y)$.

以上网络模型的 “随机性” 体现在链接独立假设. 设该模型的参数为 $\theta = \{\theta_1, \theta_2\}$, 其中

$$\theta_1 = \Big\{ |C_k| \mid 1 \le k \le K \Big\}$$
$$\theta_2 = \Big\{ P(x=j|y=k) \mid 1 \le j \le n, 1 \le k \le K \Big\}$$

令 $\ln L(N|\theta)$ 表示在给定参数 θ 下网络 N 出现的似然. 该模型参数估计的目标是: 估计出参数 θ^*, 使得被观察网络 N 出现的可能性最大, 即:

$$\theta^* = \arg\max_{\theta} \ln L(N|\theta)$$

第 2 节将讨论提出的基于模拟退火策略的局部搜索算法, 该方法能够有效地估计出 θ^* 的一个近似全局最优解.

令矩阵 $\Delta_c = (\delta_{pq})_{K\times K}$ 表示 K–划分 $C = (C_1, \cdots, C_K)$ 对应的子网络耦合矩阵, 其中 δ_{pq} 表示子网络 C_p 和 C_q 的耦合度, 其含义为: 分别从 C_p 和 C_q 中任取结点 i 和 j, 网络中存在链接 $\langle i,j\rangle$ 的概率. 根据估计出的参数 θ, δ_{pq} 可计算如下:

$$\delta_{pq} = \frac{1}{|C_q|} \sum_{j \in C_q} P(x=j|y=p) \tag{5}$$

同配社区中的结点倾向于相互连接, 因此同配社区会具有较大的自耦合度. 反之, 异配社区中的结点倾向于相互排斥, 因此异配社区会具有较小的自耦合度. 对于给定的网络 N 和 K–划分 C, 计算出的子网络耦合矩阵 Δ_c, 通过分析划分中每个子网络的自耦合度, 可分析出各个子网络的类型 (同配社区或异配社区), 进而分析出网络 N 的类型 (同配网络或异配网络). 具体步骤如下:

步骤 1. 设定一个阈值 δ';

步骤 2. 根据每个子网络的自耦合情况, 判断其类别. 具体为: 对于子网络 C_p, $1 \le p \le K$, 如果 $\delta_{pp} \ge \delta'$, 则 C_p 为同配社区, 否则 C_p 为异配社区;

步骤 3. 根据每个子网络的类型, 分析网络的类型. 具体为: 如果所有子网络都为同配社区, 则该网络为同配网络; 如果所有子网络都为异配社区, 则该网络为异配网络; 否则为同配/异配混合型网络.

在步骤 1 中, 阈值 δ' 的选取可采用多种策略, 本文采用简单的均值法, 即:

$$\delta' = \frac{1}{K^2} \sum_{p=1}^{K} \sum_{q=1}^{K} \delta_{pq} \tag{6}$$

2 广义社区挖掘算法

给定网络 N 的一个 K–划分 $C = (C_1, \cdots, C_K)$, 参数 $\theta = \{\theta_1, \theta_2\}$ 可由 C 唯一确定, 其中 θ_1 中的分量可由式 (2) 计算, θ_2 中分量计算如下:

$$P(x=j|y=k) = \frac{|\{\langle i,j\rangle \in E(N) | i \in C_k\}|}{|C_k|} \tag{7}$$

记 $\theta(C)$ 表示由划分 C 确定的参数, 从而所观察网络 N 出现的似然可以通过 C 求得, 记为 $\ln L(N|\theta(C))$. 至此, 参数估计问题转化为寻找最优网络划分, 使得:

$$C^* = \arg\max_{C} \ln L(N|\theta(C))$$

令

$$\begin{aligned}
&l(i) = \ln P(\boldsymbol{t} = \boldsymbol{t}_i) = \\
&\ln |C_{S(i)}| - \ln n + \sum_{j=1}^{n} h(a_{ij}, P(x=j|y=S(i)))
\end{aligned} \tag{8}$$

则 $l(i)$ 表示网络的局部似然函数, 即结点 i 与其他结点连接结构出现的似然. 不难证明, 网络的全局似

然函数是各局部似然函数之和, 即

$$\ln L(N) = \sum_{i=1}^{n} l(i) \tag{9}$$

式 (9) 是式 (4) 的一个局部化版本, 极大化网络全局似然函数可以转换为极大化各个结点的局部似然函数. 基于局部搜索方法, 本节给出极大化 $\ln L(N)$ 的优化算法, 描述如下:

算法 1. C = G-NCMA(N, K)

/* 输入参数分别表示网络和划分的子图个数, 输出参数 C 表示网络的一个近似最优 K–划分 */

步骤 1. 初始化划分 C

将结点 i ($1 \le i \le n$) 所在的子图序号 $S(i)$ 初始化为 1 和 K 之间的随机整数.

步骤 2. 初始化局部似然函数

步骤 2.1. 根据初始划分, 由式 (2) 和式 (7) 估计初始参数 θ;

步骤 2.2. 根据初始参数, 由式 (8) 计算初始局部似然函数 $l(i), 1 \le i \le n$;

步骤 2.3. 由式 (9) 计算初始全局似然函数 V.

步骤 3. 局部搜索过程

步骤 3.1. 以概率 r_1 选择具有最小局部似然函数值的结点 i, 以概率 $1 - r_1$ 随机选择一个结点 i;

步骤 3.2. 以概率 r_2 更新 $S(i)$ 使其最大化 $l(i)$, 以概率 $1 - r_2$ 更新 $S(i)$ 使得 $l(i)$ 增大;

步骤 3.3. 根据更新后的 $S(i)$, 由式 (2) 和式 (7) 更新参数 θ;

步骤 3.4. 根据更新后的 θ, 由式 (8) 计算各局部似然函数值;

步骤 3.5. 由式 (9) 更新全局似然函数值 V'.

步骤 4. 结束条件

如果 $V' \ge V$ 且迭代步数小于规定值, 则执行步骤 3, 否则终止算法.

以上算法采用贪心策略控制局部搜索过程: 若当前解不差于先前解, 保留当前解并继续探索, 直到找到局部极大值或达到规定迭代步数为止. 因此, 上述算法的停止条件为: 算法搜索到优化目标函数 $\ln L(N)$ 的一个局部极大值或搜索步数达到指定值.

为了跳出局部极值, 搜索到更接近全局最优的解, 可采用如下模拟退火控制策略代替贪心策略:

在步骤 3 之后, 根据 Metropolis 准则计算接受概率 p: 若 $V' \ge V$ 则 $p \leftarrow 1$; 否则 $p \leftarrow \exp(\frac{V'-V}{T})$. 以概率 p 决定是否接受步骤 3 中的更新. 其中, T 表示系统温度, 一次迭代之后, 更新为 $T \leftarrow \alpha \times T$, α 为 0 与 1 之间的常数. 当系统温度低于指定值时, 停止算法. 理论上, 当搜索过程充分长时, 模拟退火控制策略可保证搜索到优化目标的全局最优解.

步骤 3 的局部搜索过程是 G-NCMA 算法开销最大的操作, 其中, 步骤 3.1 的期望时间复杂性为 $O(r_1 n)$; 步骤 3.2 的期望时间复杂性为 $O((r_2 + 1)Kn/2)$; 式 (2) 的计算时间为 $O(|C_k|)$, C_k 表示第 k 个子网络中结点的个数, 因此计算 θ_1 的 K 个分量需要时间 $O(\sum_{k-1}^{k} |C_k|) = O(n)$. 式 (7) 的计算时间为 $O(d_k^i)$, d_k^i 表示第 k 个子网络中连接到结点 i 的结点个数, 因此计算 θ_2 的 Kn 个分量需要的时间是:

$$O\left(\sum_{i=1}^{n}\sum_{k=1}^{K} d_k^i\right) = O\left(\sum_{i=1}^{n} d_i^{\text{in}}\right) = O(m)$$

其中, d_i^{in} 表示结点 i 的入度, m 表示网络中有向边的条数. 计算式 (8) 需要 $O(n)$ 时间, 因此步骤 3.4 中计算全部 n 个结点的局部似然函数值需要 $O(n^2)$ 时间. 步骤 3.5 更新全局似然函数值需要 $O(n)$ 时间. 综上, 在局部搜索过程中, 一次迭代需要的时间是:

$$O\left(r_1 n + (r_2 + 1)\frac{Kn}{2} + m + n^2\right) = O(n^2)$$

因此 G-NCMA 算法的时间复杂性为 $O(In^2)$, 其中 I 表示规定的最大迭代次数.

对于输入类型未知的网络, 采用 G-NCMA 算法进行极大似然意义下的最优 K–划分, 所得划分中的每个随机子网络对应网络中的一个社区. 根据参数 θ 计算该划分对应的子网络耦合矩阵 Δ_c, 进而分析出每个社区及整个网络的类型.

作为示例, 分析图 1 给出的两个网络示例. 图 1 (a) 是一个具有 3 个社区的同配网络, 具有社区内边密集而社区间边稀疏的特点. G-NCMA 算法为该网络计算出的最优 3–划分为 (1, 2, 3, 4, 5)、(6, 7, 8, 9) 和 (10, 11, 12, 13), 对应的似然值为 $\ln L(N) = -56.28$. 对照图 1 (a) 可知, 该划分的 3 个随机子网恰好对应 3 个网络社区. 该划分对应的子网络耦合矩阵为

$$\Delta_c = \begin{bmatrix} 0.64 & 0.05 & 0.05 \\ 0.05 & 0.625 & 0.1875 \\ 0.05 & 0.1875 & 0.75 \end{bmatrix}$$

阈值 $\delta' = 0.288$. 3 个社区的自耦合度均高于阈值, 因此都为同配社区. 相应的, 该网络为同配网络.

图 1 (b) 是一个带有噪声的有向二分图 (Bipartite graph), 该网络中边是从左边子图指向右边子图, 并且除了由结点 2 指向结点 1 的边, 子图内没有互相连接. G-NCMA 算法为该网络计算出的最优 2–划分为 (1, 2, 4, 6) 和 (3, 5, 7), 在该划分下网络出现的似然值为 $\ln L(N) = -8.99$, 该划分对应的 2 个随机网络恰好是该异配网络的 2 个社区. 该划分

对应的子网络耦合矩阵为

$$\Delta_c = \begin{bmatrix} 0.0625 & 0.75 \\ 0 & 0 \end{bmatrix}$$

阈值 $\delta' = 0.203$. 两个社区的自耦合度均低于阈值, 因此都为异配社区. 相应的, 该网络为异配网络.

3　实验

本节选择了 3 个实际同配网络、1 个实际异配网络和 1 个实际混合网络, 以及由计算机生成的随机网络对 G-NCMA 算法进行了测试, 并与几个有代表性的网络社区挖掘算法进行了比较, 实验结果和分析如下.

3.1　测试基准同配网络

图 2 的第 1 行分别给出了用于测试同配网络社区挖掘算法的三个常用基准网络: 空手道俱乐部网络 (Karate 网络)、海豚社会网络 (Dolphin 网络) 和美国大学足球联盟网络 (Football 网络). 表 1 给出了有关这些网络的统计信息.

表 1　实验同配网络的统计信息

Table 1　The statistics of the experimental assortative networks

网络	Karate	Dolphin	Football
结点数	34	62	115
边数	88	160	616
社区数	2	2	12

Karate 网络刻画的是美国某大学空手道俱乐部成员之间的交往关系. 在多种社会因素作用下, 俱乐部最终分裂成两个互不相交的独立团体, 分别由行政主管和教练领导, 如图 2 (a) 所示, 整个网络被实线分开形成 2 个社区结构.

图 2 第 2 行是采用 2–划分模型对 Karate 网络的实验结果. 其中, 图 2 (d) 给出 Karate 网络的邻接矩阵. 图 2 (f) 给出了 G-NCMA 算法的搜索过程, 经过 352 步迭代, 得到 $\ln L(N)$ 的一个局部最优解 -348.77. 通过重新安排原邻接矩阵的行和列, 将同社区结点排列在一起, 得到转换后的邻接矩阵如图 2 (e) 所示, 矩阵中的两条实线表示两个社区的 "边界". 该矩阵能够清晰地表示出隐藏在网络中的社区结构. 如果网络具有分明的社区结构, 其对应的转换矩阵应该是一个近似的对角矩阵, 分布在主对角线区域的非零元素 (对应社区内边) 远远多于散落在主对角线区域之外的非零元素 (对应社区间边). 主对角线上的每一个块矩阵 (由实线区分) 恰好对应一个社区结构. 通过分析可知, 由算法 G-NCMA 得出的 2–划分与 Karate 网络实际的分裂结果一致. 该 2–划分对应的子网络耦合矩阵为

$$\Delta_c = \begin{bmatrix} 0.215 & 0.035 \\ 0.035 & 0.28 \end{bmatrix}$$

阈值 $\delta' = 0.136$. 两个社区的自耦合度均高于阈值, 因此都为同配社区. 相应的, 该网络为同配网络.

Dolphin 网络刻画了位于新西兰海域中 62 只海豚 7 年间的社会关系. 由于维系整个群体稳定性的海豚走失, 这些海豚最终分裂为如图 2 (b) 所示的 2 个群体, 对应 2 个社区结构. 图 2 第 3 行是采用 2–划分模型对 Dolphin 网络的测试结果. 其中图 2 (g) 给出了该网络的邻接矩阵. 图 2 (i) 给出了 G-NCMA 算法的搜索过程, 经过 361 步迭代, 找出了该网络 $\ln L(N)$ 的一个局部最优解 -891.61. 通过重新安排原邻接矩阵的行和列, 将同区结点排列在一起, 得到转换后的邻接矩阵如图 2 (h) 所示, 矩阵中的两条实线表示两个社区的 "边界". 分析可知, 由算法 G-NCMA 得出的 2–划分与 Dolphin 网络实际的分裂结果基本一致, 仅有处于社区边界的两只海豚被错分. 该 2–划分对应的子网络耦合矩阵为

$$\Delta_c = \begin{bmatrix} 0.134 & 0.008 \\ 0.008 & 0.190 \end{bmatrix}$$

阈值 $\delta' = 0.085$. 两个社区的自耦合度均高于阈值, 因此都为同配社区. 相应的, 该网络为同配网络.

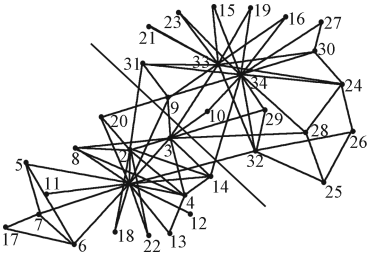
Football 网络中的每个结点表示一支大学足球队, 每条边表示两个球队间的一场比赛. 根据地理位置, 全部球队组织成 12 个联盟. 根据赛制安排, 联盟内的比赛远多于联盟间的比赛. 因此, 按照比赛的关系, 12 个联盟对应 12 个社区结构. 图 2 (j) ∼ (l) 给出了采用 12–划分模型对 Football 网络分析的实验结果.

图 2 (j) 给出了 Football 网络的邻接矩阵. 图 2 (l) 给出了 G-NCMA 算法的搜索过程, 经过 512 步迭代找到了该网络 $\ln L(N)$ 的一个局部最优解 $-2\,066.9$. 通过重新安排原邻接矩阵的行和列, 将同区结点排列在一起, 得到转换后的邻接矩阵如图 2 (k) 所示, 矩阵中的实线表示社区的 "边界", 矩阵对角线上的 12 个 "块矩阵" 依次对应算法所得的 12 个社区.

设 G-NCMA 算法所得社区 i 对应足球联盟 j, 定义 G-NCMA 算法对联盟 j 的查准率 (A_j) 和查全率 (C_j) 如下:

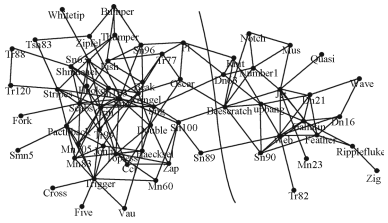
$$A_j = \frac{\text{社区 } i \text{ 中包含联盟 } j \text{ 球队的个数}}{\text{社区 } i \text{ 中球队的总数}} \tag{10}$$

$$C_j = \frac{\text{社区 } i \text{ 中包含联盟 } j \text{ 球队的个数}}{\text{联盟 } j \text{ 中球队的总数}} \tag{11}$$



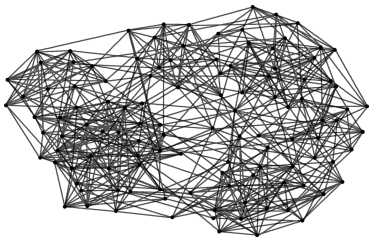
(a) Karate 网络

(a) Karate network



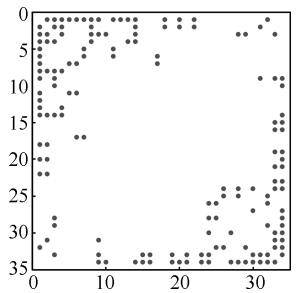
(b) Dolphin 网络

(a) Dolphin network



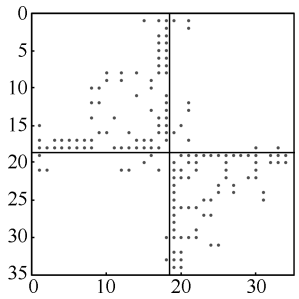
(c) Football 网络

(c) Football network



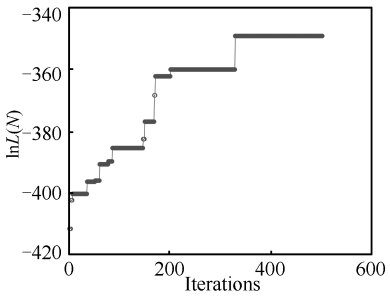
(d) Karate 网络的邻接矩阵

(d) The adjacency matrix of the Karate network



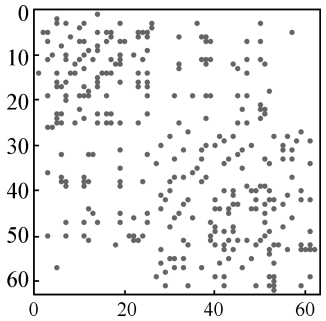
(e) Karate 网络重排的邻接矩阵

(e) The reordered adjacency matrix of the Karate network



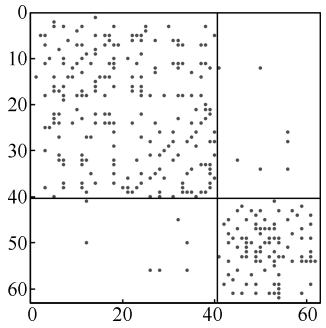
(f) Karate 网络搜索过程

(f) The searching process of the Karate network



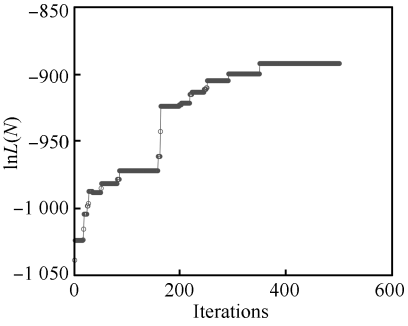
(g) Dolphin 网络的邻接矩阵

(g) The adjacency matrix of the Dolphin network



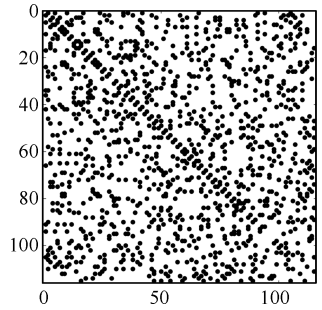
(h) Dolphin 网络重排的邻接矩阵

(h) The reordered adjacency matrix of the Dolphin network



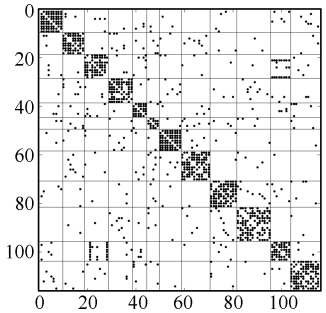
(i) Dolphin 网络的搜索过程

(i) The searching process of the Dolphin network



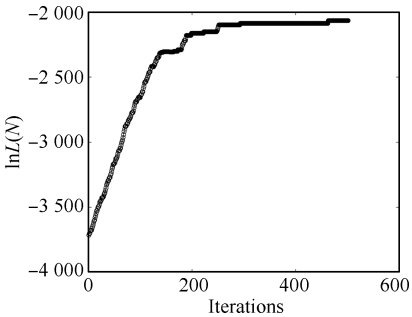
(j) Football 网络的邻接矩阵

(j) The adjacency matrix of the Football network



(k) Football 网络重排的邻接矩阵

(k) The reordered adjacency matrix of the Football network



(l) Football 网络的搜索过程

(l) The searching process of the Football network

图 2 使用 G-NCMA 算法对三个基准同配网络的实验结果

Fig. 2 The experimental results obtained by the G-NCMA algorithm applied to three benchmark assortative networks

表 2 给出了 G - NCMA 对 12 个足球联盟的查准率和查全率. 其中, 联盟 5 (IA independents) 被分散到不同联盟中, 没有对应的社区. 这是因为, 独立联盟中球队与联盟外球队的比赛远多于联盟中相互之间的比赛. 联盟 10 被分为两个社区, 原因是这两个社区内球队相互比赛多, 各自构成小的社区.

表 2　各个联盟的查准率和查全率

Table 2　The precision and recall of respective associations

联盟编号	联盟名称	A (%)	C (%)
0	Atlantic Coast	100	100
1	Big East	80	100
2	Big 10	100	100
3	Big 12	100	100
4	Conference USA	100	90
5	IA Independents	—	—
6	Mid American	92.9	100
7	Mountain West	100	100
8	Pac 10	100	100
9	SEC	100	100
10	Sunbelt	66.7	57.1
11	Western Athletic	72.7	88.9

算法所得 12 个社区的自耦合度分别为: 0.89, 0.8, 0.77, 0.61, 0.72, 0.88, 0.89, 0.60, 0.70, 0.52, 0.83, 0.67. 阈值 $\delta' = 0.093$. 各社区的自耦合度均远大于阈值, 因此都为同配社区. 相应的, 该网络为同配网络.

3.2　测试异配网络

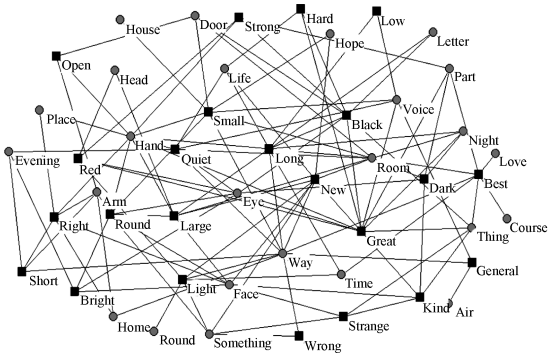
图 3 (a) 给出了一个包含 48 个英文单词的网络, 网络中的每个结点是一个名词 (用圆形结点表示) 或者形容词 (用方形结点表示). 这些单词是在查尔斯 · 狄更斯 (Charles Dickens) 小说《David Copperfield》中出现频率最高的单词. 如果两个单词经常同时出现在一个语句中 (出现的频率超过规定的阈值), 则在它们之间建立一条链接. 在自然语言中, 相比于同性单词, 异性单词更加倾向于同时出现在相同的上下文中. 例如, 形容词和被其修饰的名词经常先后出现在临近位置. 因此, 图 3 (a) 所示的单词网络是一个异配网络, 网络中词性相异的结点具有较高的连接概率, 而词性相同的结点则具有较低的连接概率.

G - NCMA 算法得到最优 2 – 划分如图 3 (b) 所示. 其中, 左边社区包含大部分的形容词和一个错分的名词; 右边社区包含大部分的名词和 5 个被错分的形容词. 由图 3 (b) 可以看出, 社区之间的链接密度远远大于社区内的链接密度.

该 2 – 划分对应的子网络耦合矩阵为

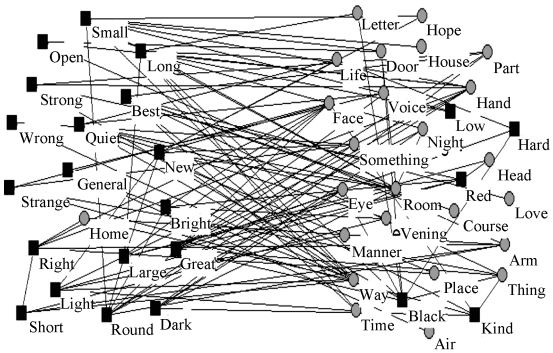
$$\Delta_c = \begin{bmatrix} 0.031 & 0.169 \\ 0.169 & 0.044 \end{bmatrix}$$

阈值 $\delta' = 0.103$. 两个社区的自耦合度均低于阈值, 因此都为异配社区. 相应的, 该网络为异配网络.



(a) 由名词和形容词及其共现关系构成的网络

(a) The phrase network in terms of the co-occurrence of nouns and adjectives



(b) 由 G-NCMA 算法得到的一个 2-划分

(b) The 2-partition obtained by the G-NCMA

图 3　使用 G-NCMA 算法对一个异配网络的分析结果

Fig. 3　The experimental results obtained by the G-NCMA algorithm applied to a disassortative network

根据划分可计算出 G - NCMA 算法对网络中两类单词的查准率 (由式 (10) 定义) 和查全率 (由式 (11) 定义), 如表 3 所示.

表 3　两种词性的查准率和查全率

Table 3　The precision and recall of two types of words

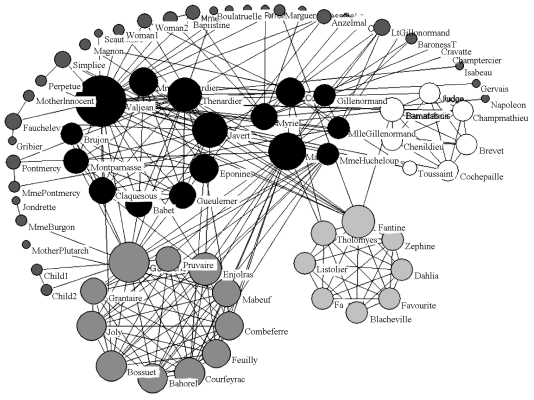
词性	A (%)	C (%)
形容词	94.7	78.3
名词	82.76	96.0

被 G - NCMA 算法错分的单词包括 1 个名词 “Home” 和 5 个形容词 “Red”, “Black”, “Hard”, “Kind” 和 “Low”. 其中被错分的前 4 个形容词既

可以作为形容词使用, 又可以作为名词使用, 词义的“二义性” 是导致这些词被划分错误的主要原因.

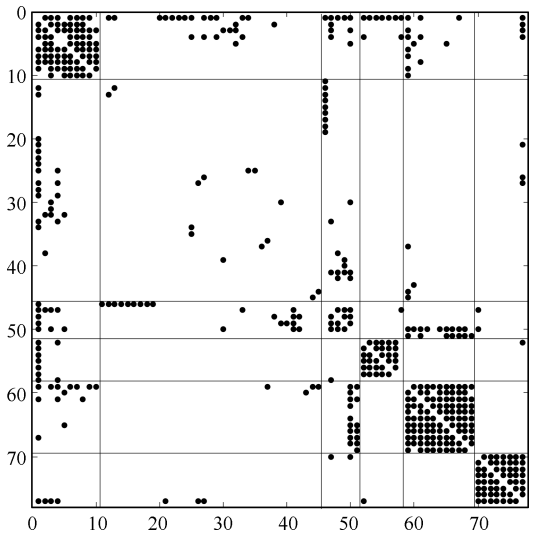
3.3 测试混合型网络

图 4 (a) 给出了刻画名著《悲惨世界》中人物关系的社会网络. 网络中的 77 个节点表示小说中的 77 个人物角色, 网络中的无向边表示他们之间的交互关系. 图 4 (a) 中, 网络结点的大小正比于结点的度. 度大的结点对应于小说中的主要角色, 度小的结点对应于小说中的次要角色.



(a) 小说《悲惨世界》的人物关系网

(a) The social network of the novel “Les Miserables”



(b) 网络的邻接矩阵, 其中包含一个由算法 G-NCMA 得到的 6 – 划分

(b) The adjacency matrix of the network, which contains a 6-partition obtained by the G-NCMA

图 4 使用 G-NCMA 算法对《悲惨世界》人物关系网的分析结果

Fig. 4 The experimental results obtained by the G-NCMA algorithm applied to the social network of “Les Miserables”

G - NCMA 算法得到的最优 6 – 划分如图 4 (b) 所示, 对应的 6 个网络社区如图 4 (a) 所示 (其中不同灰度表示不同社区). 6 个社区的自耦合度分别为: 0.78, 0.01,0.33, 0.61, 0.88, 0.88. 阈值 $\delta' = 0.15$. 第 2 个社区 (包含 35 个结点) 的自耦合度远低于阈值, 为异配社区. 其余社区的自耦合度远高于阈值, 为同配社区. 因此, 该网络是同时包含同配和异配社区的混合型网络.

图 4 (a) 可视化了 6 个社区, 从中可知, 小说的主要角色形成了 4 个同配社区, 主角之间倾向于社区内部的相互交互. 小说的次要角色形成了一个大的异配社区, 配角之间交互很少, 他们倾向于和不同的主角进行少量交互. 主角和配角间的同配、异配交互关系共同形成了整部小说的复杂人物关系.

通过该例子, 一方面说明现实网络的复杂性 (可能同时包含同配和异配社区); 另一方面也说明, 广义社区挖掘方法更有助于分析复杂的社会关系网络.

3.4 基于基准网络的算法性能比较

本节将采用以上讨论的基准网络比较本文提出算法和现有社区挖掘算法的识别精度. 该实验中, 选择的对比算法包括: GN (Girvan-Newman)[1]、FN (Fast Newman)[4]、GA (Guimera-Amaral)[2] 和 RB (Rosvall-Bergstrom)[11] 等 4 个代表性社区挖掘算法, 以及 Newman 等提出的基于混合模型的社区挖掘算法 MM[12]. 由于已知基准网络的真实社区划分, 通过比较算法所得社区结构与真实社区结构, 可计算出各个算法的社区识别精度. 表 4 给出了实验结果, 其中社区识别精度定义为被正确划分结点的百分比. G-NCMA 算法中的参数设置如下: $r_1 = 0.6, r_2 = 0.8, \alpha = 0.9$, 搜索步长为 600.

表 4 对比不同算法对基准网络的社区识别精度 (%)

Table 4 Accuracy comparison of different algorithms in terms of benchmark (%)

	Karate	Dolphin	Football	Word
G-NCMA	100	96.77	90.43	87.50
GN[1]	97.06	98.39	90.43	52.08
FN[4]	97.06	96.77	63.48	58.33
GA[2]	100	100	81.74	58.33
RB[11]	100	98.39	86.96	52.08
MM[12]	97.06	91.94	72.33	69.38

分析表 4 可得: 1) 与 GN、FN、GA 和 RB 4 个代表性社区识别算法相比, G-NCMA 在处理同配网络时识别精度没有降低, 在某些情况下还占优; 在处理异配网络时, G - NCMA 的识别精度远高于其他三种算法. 2) 与 MM 算法相比, G - NCMA 算法对同配和异配网络的识别精度均优于 MM 算法. 主要原因在于: a) 在以上被测试的真实网络中, 社区结构往往受到噪声链接的影响, 并且这些网络的稀疏

度不是很高, 因此 MM 算法对真实网络似然近似造成的误差不能被忽视, 从而降低了识别精度; b) MM 算法采用 EM 算法估计参数, EM 是一种点估计方法, 受噪声链接影响较大, 且 EM 方法得到的是待优化目标的局部最优解, 而 G-NCMA 采用模拟退火的控制策略可得到一个近似的全局最优解.

3.5　基于随机网络的算法性能比较

能够生成指定社区结构的随机网络模型由 Newman 提出[4], 并被广泛使用, 已成为比较不同网络社区挖掘算法性能的基准模型. 对该模型进行适当改进, 使之可以生成已知社区结构的同配/异配有向随机网络, 用于测试广义社区挖掘算法的性能.

已知社区结构的有向随机网络定义为 $RN(K, s, d, p_{\text{in}})$, 其中 K 表示网络社区的个数, s 表示每个社区包含的结点个数, d 表示网络中结点的平均出度, p_{in} 表示社区内连接密度 (即社区内连接总数与网络连接总数的比值). 通过控制参数 p_{in}, 可得到同配网络或者异配网络. 当 $p_{\text{in}} > 0.5$ 时, 得到的随机网络是同配网络; 当 $p_{\text{in}} < 0.5$ 时, 得到的随机网络是异配网络.

需要指出的是: 现有的大多数网络社区挖掘算法只有在 $p_{\text{in}} > 0.5$ 时有效, 这些算法认为 $p_{\text{in}} < 0.5$ 时的随机网络不具有社区结构. 在广义社区结构的概念下, p_{in} 在整个 [0, 1] 区间上都存在社区结构.

社区识别精度定义为: 被正确划分结点的百分比. 一个随机网络的社区识别率为 100 %, 当且仅当预定义的 K 个网络社区被全部正确识别. 图 5 给出了实验结果. 本实验使用的随机网络是 $RN(2, 16, 16, p_{\text{in}})$.

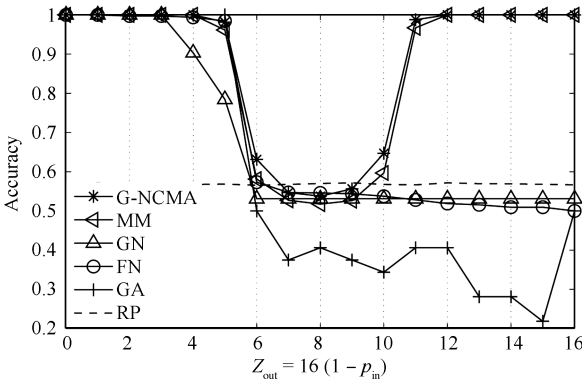


图 5　基于随机网络的社区挖掘算法精度比较

Fig. 5　Accuracy comparison of different community mining algorithms in terms of random network

该实验比较了算法 G - NCMA 与 GN[1], FN[4], GA [2] 和 MM[12] 的社区识别精度. 实验结果如图 5 所示, 图 5 中的 RP 表示将网络中的结点随机分配到其中一个社区的方法.

根据 Z_{out} 值, 可将网络结构分 4 种情况讨论.

情况 1. 对于 $0 \le Z_{\text{out}} \le 5$, 具有明显社区结构的同配网络. 此时, 5 种算法的识别精度都比较高; 除 GN 算法外, 其余 3 种算法的识别精度都接近 1.

情况 2. 对于 $5 < Z_{\text{out}} \le 8$, 不具有明显社区结构的同配网络. 此时, 5 种算法的识别精度都很低. 只有在 $Z_{\text{out}} = 6$ 时, 算法 G - NCMA、算法 FN 和算法 MM 的识别精度略高于随机分配下的识别精度. 其他情况下, 5 种算法的识别精度都低于随机分配下的识别精度.

情况 3. 对于 $8 < Z_{\text{out}} \le 10$, 不具有明显社区结构的异配网络. 此时, GN, FN 和 GA 3 种算法的识别精度都低于随机分配情况下的识别精度. 当 $Z_{\text{out}} = 10$ 时, 算法 G - NCMA 和算法 MM 的精度略高于随机分配情况下的识别精度.

情况 4. 对于 $10 < Z_{\text{out}} \le 16$, 具有明显社区结构的异配网络. 此时, 算法 G - NCMA 和算法 MM 的识别精度接近 1, 其余 3 种算法的识别精度都低于随机分配情况下的识别精度.

由该实验结果可以看出, 针对具有较少噪声的随机网络, 算法 G - NCMA 和算法 MM 在处理同配和异配网络时都表现出很好的识别精度.

4　结论

本文提出了复杂网络广义社区挖掘问题, 该问题是传统复杂网络社区挖掘问题的推广, 旨在网络类型 (同配或异配网络) 未知的情况下挖掘出有意义的社区结构. 现有的多数网络社区挖掘算法仅适用于处理特定类型的网络. 针对该问题, 本文提出了基于极大网络似然的随机网络集成模型和基于局部搜索策略的模型参数估计方法, 进而提出了广义网络社区挖掘算法 G - NCMA. 采用实际的复杂网络和计算机生成的合成网络对 G - NCMA 算法进行性能测试和分析. 实验结果表明: 基于随机网络集成模型的广义网络社区挖掘算法 G-NCMA 能在网络类型未知的前提下准确地挖掘出网络中有意义的社区结构, 并能分析出这些被挖掘社区的类型特征.

本文提出的随机网络集成模型不仅可用于社区挖掘, 还可用于网络链接预测, 作为参数估计出的链接概率可用于预测缺失链接或者噪音链接. 在该工作的基础上, 我们将进一步研究基于随机集成模型的复杂网络链接预测方法.

References

1 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

2 Guimerà R, Amaral L A N. Functional cartography of complex metabolic networks. *Nature*, 2005, **433**(7028): 895−900

3 Mucha P J, Richardson T, Macon K, Porter M A, Onnela J P. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 2010, **328**(5980): 876−878

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

5 Jin Di, Liu Jie, Yang Bo, He Dong-Xiao, Liu Da-You. Genetic algorithm with local search for community detection in large-scale complex networks. *Acta Automatica Sinica*, 2011, **37**(7): 873−882
(金弟, 刘杰, 杨博, 何东晓, 刘大有. 局部搜索与遗传算法结合的大规模复杂网络社区探测. 自动化学报, 2011, **37**(7): 873−882)

6 Yang B, Liu J M, Feng J F. On the spectral characterization and scalable mining of network communities. *IEEE Transactions on Knowledge and Data Engineering*, 2012, **24**(2): 326−337

7 Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2010, **466**(7307): 761−764

8 Yang T B, Chi Y, Zhu S H, Gong Y H, Jin R. Detecting communities and their evolutions in dynamic social networks — a Bayesian approach. *Machine Learning*, 2011, **82**(2): 157−189

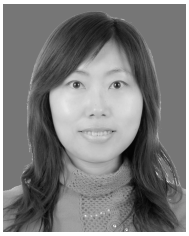
9 Newman M E J. Mixing patterns in networks. *Physical Review E*, 2003, **67**(2): 026126

10 Newman M E J. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 2006, **74**(3): 036104

11 Rosvall M, Bergstrom C T. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**(18): 7327−7331

12 Newman M E J, Leicht E A. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**(23): 9564−9569

杨 博 博士, 吉林大学教授. 主要研究方向为复杂网络分析, 数据挖掘, 多 Agent 系统. E-mail: ybo@jlu.edu.cn
(**YANG Bo** Ph. D., professor at Jilin University. His research interest covers complex network analysis, data mining, and multi-agent system.)



刘 杰 博士, 吉林大学副教授, 主要研究方向为数据挖掘, 模式识别.
E-mail: liu_jie@jlu.edu.cn
(**LIU Jie** Ph. D., associate professor at Jilin University. Her research interest covers data mining and pattern recognition.)

刘大有 吉林大学计算机科学与技术学院教授. 主要研究方向为知识工程, 专家系统与不确定性推理, 时空推理, 分布式人工智能, 多 Agent 和移动 Agent 系统, 数据挖掘与多关系数据挖掘, 数据结构与计算机算法. 本文通信作者.
E-mail: dyliu@jlu.edu.cn
(**LIU Da-You** Professor at the College of Computer Science and Technology, Jilin University. His research interest covers knowledge engineering, expert system and uncertainty reasoning, spatio-temporal reasoning, distributed artificial intelligence, multi-agent systems and mobile agent systems, data mining and multi-relational data mining, data structures, and computer algorithms. Corresponding author of this paper.)