局部搜索与遗传算法结合的大规模复杂网络社区探测

金 弟 [1,2] 刘 杰 [1,2] 杨 博 [1,2] 何东晓 [1,2] 刘大有 [1,2]

摘 要 基于遗传算法的复杂网络社区探测是当前的研究热点. 针对该问题, 本文在分析网络模块性函数 Q 的局部单调性的基础上, 给出一种快速、有效的局部搜索变异策略, 同时为兼顾初始种群的精度和多样性以达到进一步提高搜索效率的目的, 采用了标签传播作为初始种群的产生方法; 综上, 提出了一个结合局部搜索的遗传算法 (Genetic algorithm with local search, LGA). 在基准网络及大规模复杂网络上对 LGA 进行测试, 并与当前具有代表性的社区探测算法进行比较, 实验结果表明了文中算法的有效性与高效性.

关键词 复杂网络, 社区探测, 网络聚类, 遗传算法, 局部搜索

DOI 10.3724/SP.J.1004.2011.00873

Genetic Algorithm with Local Search for Community Detection in Large-scale Complex Networks

JIN Di[1,2] LIU Jie[1,2] YANG Bo[1,2] HE Dong-Xiao[1,2] LIU Da-You[1,2]

Abstract Detecting communities from complex networks by genetic algorithm has triggered a great common interest. For this problem, a genetic algorithm with local search (LGA) which employs network modularity Q as objective function is given in this work. An effective as well as efficient mutation method combined with a local search strategy is proposed based on our profound analysis on local monotonicity of function Q, meanwhile, a label propagation based method is adopted to produce the accurate and diverse initial population, which can further improve the search efficiency of LGA. The proposed LGA has been tested on both benchmark networks and some large-scale complex networks, and compared with some competitive community detection algorithms. Experimental result has shown that LGA is highly effective and efficient for discovering community structure.

Key words Complex network, community detection, network clustering, genetic algorithm, local search

现实世界中的许多复杂系统都以复杂网络的形式存在, 或者能被转化为复杂网络进行处理, 如社会网、生物网、Web 网络、科技网络等. 目前对复杂网络基本统计特性的研究已吸引了不同领域的众多研究者, 复杂网络分析已成为当前最重要的多学科交叉研究领域之一[1−4]. 其中, "小世界效应" 是指复杂网络具有短路径长度和高聚类系数的特点[1]; "无标度特性" 是指复杂网络中的结点的度服从幂率分布特征[2]; 而本文所涉及的 "聚类特性" (也称社区结构特性) 是指复杂网络中普遍存在着 "同一社区内之结点相互连接紧密、而不同社区间之结点相互连接稀疏" 的特点[3].

随着应用领域的不同, 社区结构具有不同的内涵, 譬如: 社会网中的社区代表了具有某些相近特征的人群, 生物网络中的功能组揭示了具有相似功能的生物组织模块, Web 网络中的文档类簇包含了大量具有相关主题的 Web 文档, 等等[5]. 复杂网络社区探测 (又称网络聚类) 的目的就是要探测并揭示出异构复杂网络中固有的社区结构. 该问题的研究具有十分重要的理论及现实意义, 它不仅吸引了大量不同学科的研究工作者, 而且已被应用于如恐怖组织识别、蛋白质功能预测、新陈代谢途径 (Pathway) 预测、Web 社区挖掘、链接预测等众多领域当中[5].

2004 年, Newman 等[6] 提出了一个可定量评价网络社区结构优劣的度量标准, 被称为网络模块性函数 (Q). 此后, 以 Q 函数为目标函数的组合优化方法成为探测网络社区结构的主流方法之一, 譬如 GN (Girvan-newman)[3], FN (Fast newman)[7], SA (Simulated annealing)[8], MSA (Modularity spec-

收稿日期 2010-07-19 录用日期 2011-01-12
Manuscript received July 19, 2010; accepted January 12, 2011
国家高技术研究发展计划 (863 计划) (2006AA10Z245), 国家自然科学基金 (60773099, 60703022, 60873149, 60973088), 模式识别国家重点实验室开放课题 (09-1-1), 中央高校基本科研业务费专项资金 (200903177), 复旦大学智能信息处理上海市重点实验室开放课题 (IIPL-09-007) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2006AA10Z245), National Natural Science Foundation of China (60773099, 60703022, 60873149, 60973088), the Open Project Program of the National Laboratory of Pattern Recognition (09-1-1), the Fundamental Research Funds for the Central Universities (200903177), and the Open Project Program of Shanghai Key Laboratory of Intelligent Information Processing (IIPL-09-007)
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

tral algorithm)[9], ITS (Iterated tabu search)[10] 以及 LPAm (Modularity-specialized label propagation algorithm)[11−12] 等算法. 因最大化 Q 函数是 NP 完全的[13], 故上述方法都是近似优化算法.

近年来, 作为一种求解 NP 难问题的有效方法, 遗传算法 (Genetic algorithm, GA) 被应用于网络社区探测领域[14−21]. 然而, 当前基于 GA 的社区探测算法或因时间复杂度较高或因搜索能力偏弱, 还不能对大规模复杂网络进行有效聚类. 从对文献 [14−21] 的分析可看出, 一方面改进全局搜索算子 (选择、交叉) 或提出新的全局算子都很难有效解决大规模网络社区探测问题, 另一方面当前基于 GA 的社区探测算法的局部搜索能力不足, 所以从局部搜索角度出发, 探索新的兼顾效率和质量的变异策略可能是富有前景的. 由此, 本文通过分析出 Q 函数的一种局部梯度特征, 提出一种快速、有效的局部搜索变异策略; 同时, 为使初始种群能更好兼顾精度和多样性, 采用了基于标签传播的个体生成方法; 进而给出了一个结合局部搜索的遗传算法 (Genetic algorithm with local search, LGA).

1 算法 LGA

文中算法 LGA 以 Q 函数作为目标函数, 采用字符串编码方式. 其首先通过给出一个标签传播方法产生兼顾精度与多样性的初始种群; 然后通过迭代执行单路交叉 (One-way crossing over)[14]、局部搜索变异和 $\mu + \lambda$ 选择[22] 三个遗传算子来探测网络社区结构, 其中结合局部搜索的变异算法是本文的核心工作.

1.1 问题定义

2004 年, Newman 等[6] 基于对 “网络社区结构越明显, 它与随机网络之间的差异也就应该越大” 这一直观现象的思考, 提出了一个可定量评价网络社区结构优劣的度量标准, 被称为网络模块性函数 (Q), 其目前已被大多数相关领域的学者广泛采纳. Q 函数的定义为: 网络社区内实际存在的边数与完全随机的连接情况下社区内期望的边数之差.

给定一个无向无权网络 $N(V, E)$, 假设点集 V 被划分 (聚类) 为若干个社区. 若网络中任一结点 i 的标签为 $r(i)$, 它所属的社区为 $c_{r(i)}$, 则 Q 函数可被定义为

$$Q = \frac{1}{2m}\sum_{ij}\left(\left(A_{ij} - \frac{k_i k_j}{2m}\right) \times \delta\left(r(i), r(j)\right)\right) \tag{1}$$

其中, $A = (A_{ij})_{n\times n}$ 表示网络 N 的邻接矩阵, 如果结点 i 与 j 之间存在边连接, 则 $A_{ij} = 1$, 否则 $A_{ij} = 0$; 对于函数 $\delta(u, v)$, 如果 $u = v$, 其取值为 1, 否则取值为 0; k_i 表示结点 i 的度, 被定义为 $k_i = \sum_j A_{ij}$; m 表示网络 N 中总的边数, 被定义为 $m = \frac{1}{2}\sum_{ij} A_{ij}$.

尽管存在着 “分辨率限制” 这一问题[23], 但目前的多数工作还是将 Q 函数作为评价网络社区结构优劣的衡量标准[7−17, 20−21]. 文中遗传算法 LGA 也采用 Q 函数作为目标函数和适应度函数. 因此, 该算法虽然也存在分辨极限的问题, 但它仍普遍适用于具有模块性 (Modularity) 结构的复杂网络.

1.2 编码方式

目前基于 GA 的社区探测算法主要采用字符串编码方式[14−17] 和基于图的编码方式[18−21] 两种策略. 与基于图的编码方式相比, 字符串编码在表示网络社区结构方面更加直观、高效, 因此文中遗传算法 LGA 采用了字符串编码方式. 若给定某复杂网络 N, 对该网络的任意划分都可表示为字符串编码 (染色体) $R = \{r(1), r(2), \cdots, r(n)\}$, 其中 n 为网络 N 中的结点数, $r(i)$ 为结点 i 的标签, 可用某整数来表示. 对于 R 中的任意结点对 i 和 j, 如果 $r(i) = r(j)$, 则说明这两个结点位于同一社区内; 否则, 它们位于不同社区. 很明显, 字符串编码可以非常方便地表示网络社区探测问题中的任一候选解 (网络聚类结果).

1.3 初始群体生成

一般来说, 在遗传算法的初始种群中, 如果每个个体都具备一定的精度且个体间多样性较强时, 则可以提高算法搜索效率, 加速算法收敛. 不同于随机生成初始群体的方法, 本文借鉴文献 [24] 的主要思想, 提出了一个基于标签传播的种群初始化方法 (Individual generation via label propagation, IGLP), 试图能够快速产生具有一定精度和较强多样性的初始种群.

该算法的主要思想来源于如下直观现象: “在具有明显社区结构的复杂网络中, 如果该网络中每个结点都与其大多数邻居位于同一社区内, 那么一般来说该社区结构是较合理的”. 基于此, 我们给出算法 IGLP 的主要步骤如下: 首先将染色体 R 中每个结点 i 都赋值为唯一的标签, 即 $r(i) \leftarrow i$; 然后在每次迭代中, 每个结点 i 都将自身标签更新为其大多数邻居结点所在社区的标签, 即 $r(i) \leftarrow \arg\max_r \sum_{j\in\sigma(i)} \delta(r(j), r)$, 其中, $\sigma(i)$ 表示 i 的邻居结点集合; 如果达到迭代次数限制或下一代候选解对应的 Q 函数值小于当前代, 则算法终止. 由文献 [24] 分析可知, 当标签传播次数 l 到达 5 时, 产生的候选解一般都具有一定的聚类精度, 且个体 (候选解) 间的多样性较强. 此外, 算法 IGLP 采用了异步的结点标签更新机制, 即每次迭代中每个结点对自身

标签的更新不受其他结点更新标签的影响, 从而可进一步增强初始群体的多样性.

1.4　交叉算子

本文采用字符串编码方式, 而该编码存在着多个不同染色体对应同一网络划分的问题. 举个例子, 假设网络中的结点集合为 $V=\{v1,v2,v3,v4\}$, 那么染色体 $A=(1,2,1,3)$, $B=(2,1,2,4)$ 都对应了相同的社区划分结果 $P=\{\{v1,v3\},\{v2\},\{v4\}\}$. 很显然, 在字符串编码的染色体中, 每个社区标签都用一个随意的整数来表示, 所以不同染色体中相同的标签值一般并不代表同一个网络社区, 也就是说不同染色体中的社区标签一般是互不兼容的. 在这种情况下, 普通交叉算子 (譬如: 单点交叉、多点交叉等) 很容易破坏已有的良好积木, 从而导致遗传算法失效. 因此, 文中遗传算法无法采用传统的交叉算子.

针对上述问题, 我们采用 2007 年 Tasgin 等[14] 为解决网络社区探测问题而提出的单路交叉 (One-way crossing over) 策略. 给定任意两个体 A 和 B, 其中 A 作为源染色体, B 作为目的染色体, 单路交叉操作可被定义如下. 首先任意选择网络中的某结点 i, 获取其在个体 A 中所对应的社区 $c_{a(i)}$ 及社区标签 $a(i)$; 然后将个体 B 中所有属于集合 $c_{a(i)}$ 的结点之标签赋值为 $a(i)$, 即 $b(j)\leftarrow a(i)$, $\forall j\in c_{a(i)}$. 很显然, 该交叉操作能够将源染色体 A 中关于社区结构的一部分信息提供给目的染色体 B, 它很适合被用于网络社区探测问题. 文中通过表 1 对单路交叉操作进行了直观说明如下.

表 1　选取结点 $v4$ 时对单路交叉操作的一次演示

Table 1　An illustration of one-way crossing over when $v4$ is selected

V		A (Source)		B (Destination)		B (New)
1		3		4		4
2		2		1		1
3		4	→	3	→	4
4	→	4	→	4	→	4
5		2		1		1

1.5　变异算子

变异算子是本文研究的核心内容. 面向网络社区探测问题, 当前遗传算法采用的变异策略大都存在局部搜索能力偏弱的缺陷[14−21], 导致其难于有效的聚类大规模复杂网络. 针对该问题, 本文试图通过对网络社区探测问题的深入研究, 提出一个高效的局部搜索变异算法.

由于本文采用 Q 函数作为目标函数, 所以我们首先从每个结点的局部观点出发来对 Q 函数进行理论分析. 文中将式 (1) 转化为式 (2), 即把 Q 函数表示为网络中所有结点的 f 函数之和. 很显然, f 函数可理解为: 从网络中任一结点的局部观点来看, 社区内实际连接数目与随机连接情况下社区内期望连接数目之差, 所以网络中每个结点的 f 函数都可以从一个局部角度来度量网络社区结构的优劣. 下面给出 f 函数的相关性质和定理.

$$Q=\frac{1}{2m}\sum_i f_i,\quad f_i=\sum_{j\in c_{r(i)}}\left(A_{ij}-\frac{k_ik_j}{2m}\right) \tag{2}$$

性质 1. 对于 $\forall i\in V$, 复杂网络中结点 i 的局部函数 f_i 仅与它所在的社区 $c_{r(i)}$ 相关.

由式 (2) 可知, 性质 1 显然成立.

定理 1. 对于 $\forall i\in V$, 复杂网络的 Q 函数值随 f_i 单调递增.

证明. 给定网络 N 及其当前社区结构 C, 取 $\forall i\in V$, 不妨设结点 i 的标签由 $r(i)$ 变为 $r(j)$, 使得当前网络社区结构变为 C'. 若 $r(i)\neq r(j)$, 那么在 C' 中, 结点 i 原来所在的社区将变为 $c'_{r(i)}=c_{r(i)}-\{i\}$, 结点 i 当前所在的社区将变为 $c'_{r(j)}=c_{r(j)}\cup\{i\}$.

由式 (2) 可知, 对于任一结点, 如果它所在的社区发生了变化, 那么它所对应的 f 函数值也会随之发生变化, 所以结点 i 标签的变化将会导致集合 $c=c_{r(i)}\cup c_{r(j)}$ 中所有结点 f 函数值的变化. 下面将集合 c 中的结点分为三类, 并针对每一类结点, 分别给出其 f 函数值变化的计算公式.

1) 取 $\forall s\in c'_{r(i)}$, 其 f 函数值的变化 Δf_s 可被计算为

$$\Delta f_s=f_s(C')-f_s(C)=\sum_{t\in c'_{r(i)}}\left(A_{st}-\frac{k_sk_t}{2m}\right)-\sum_{t\in c_{r(i)}}\left(A_{st}-\frac{k_sk_t}{2m}\right)=-\left(A_{si}-\frac{k_sk_i}{2m}\right) \tag{3}$$

2) 取 $\forall p\in c_{r(j)}$, 其 f 函数值的变化 Δf_p 可被计算为

$$\Delta f_p=f_p(C')-f_p(C)=\sum_{q\in c'_{r(j)}}\left(A_{pq}-\frac{k_pk_q}{2m}\right)-\sum_{q\in c_{r(j)}}\left(A_{pq}-\frac{k_pk_q}{2m}\right)=A_{pi}-\frac{k_pk_i}{2m} \tag{4}$$

3) 结点 i 的 f 函数值的变化 Δf_i 可被计算为

$$\Delta f_i = f_i(C') - f_i(C) = \sum_{e\in c'_{r(j)}}\left(A_{ie} - \frac{k_i k_e}{2m}\right) - \sum_{e\in c_{r(i)}}\left(A_{ie} - \frac{k_i k_e}{2m}\right) \tag{5}$$

将 "由结点 i 标签的变化而导致的整个网络 Q 函数值的变化" 表示为 ΔQ, 那么对 ΔQ 的推导如下.

$$\Delta Q = \frac{1}{2m}\left(\sum_{s\in c'_{r(i)}}\Delta f_s + \sum_{p\in c_{r(j)}}\Delta f_p + \Delta f_i\right) \tag{6}$$

$$\Delta Q = \frac{1}{2m}\left(\sum_{p\in c'_{r(j)}}\left(A_{pi} - \frac{k_p k_i}{2m}\right) - \sum_{s\in c_{r(i)}}\left(A_{si} - \frac{k_s k_i}{2m}\right) + \Delta f_i\right) \tag{7}$$

$$\Delta Q = \frac{1}{2m}\left(\sum_{p\in c'_{r(j)}}\left(A_{pi} - \frac{k_p k_i}{2m}\right) - \sum_{s\in c_{r(i)}}\left(A_{si} - \frac{k_s k_i}{2m}\right) + \Delta f_i\right) \tag{8}$$

$$\Delta Q = \frac{1}{2m}\left(\Delta f_i + \Delta f_i\right) = \frac{1}{m}\Delta f_i \tag{9}$$

即 $Q(C') - Q(C) = \dfrac{1}{m}\left(f_i(C') - f_i(C)\right)$. 由此可知, 若 $f_i(C') > f_i(C)$, 则 $Q(C') > Q(C)$, 故定理成立. □

由定理 1 可知, 如果网络中某结点标签的变化使得其 f 函数值变大 (在其他结点标签不变的前提下), 那么该变化将会导致 Q 函数值的增大. 又由于在具有社区结构的复杂网络中普遍存在如下直观现象: "网络中任一结点都会与它的某些邻居结点位于同一社区内, 或其自身形成一个社区". 所以, 我们采用如下变异策略: 不妨设任一染色体 R 中的某结点 i 被选中进行变异, 且变异点 i 的邻居结点集合为 $\sigma(i)$. 现在我们就不需要再考虑 R 中所有结点的标签, 而仅需从中选择使结点 i 的局部函数值 f_i 最大的某结点 j 的标签 $r(j)$, 使其作为 i 的新标签, 即 $r(i) \leftarrow \arg\max_r\{f_i(r),\, r \in \{r(j)|j \in \sigma(i)\}\}$, 其中 $f_i(r)$ 表示结点 i 取标签 r 时对应的 f 函数值. 基于此, 本文提出一个高效并有效的局部搜索变异算法 (Mutation algorithm with local search, LMA). 在给定变异率 $\beta \in (0\sim 1)$ 的前提下, 对 LMA 进行描述如下.

算法 1. 变异算法 LMA 的基本流程

```
Procedure LMA
Global R, β    // R 为待变异染色体, β 为变异率
Begin
1   For i = 1 : n    // n 为 R 中包含的基因数
2     If rand( ) < β
3       σ(i) ← 取变异点 i 的所有邻居结点
4       labels ← 取 σ(i) 中所有结点对应的社区标签
5       max ← −∞
6        For each r ∈ labels
7          f_i ← 计算结点 i 标签取 r 的 f 函数值
8          If f_i > max
9             max ← f_i
10            label_i ← r
11         End If
12       End For
13       r(i) ← label_i   //更新染色体 R, 对结点 i
                            进行变异
14    End If
15  End For
End
```

为了说明算法 LMA 的有效性与高效性, 我们给出一些性质如下.

性质 2. 任意染色体 R 经 LMA 变异后的适应度值都不会减小.

证明. 由算法流程可知, 任意染色体 R 中的某基因 i 经 LMA 变异后一定会取值为使自身 f 函数值最大的某邻居结点 j 的标签, 而其他结点的标签都不会发生变化. 又由定理 1 可知, 在其他结点标签不变的前提下, 如果某结点标签的变化使得其 f 函数值变大, 那么该变化将会导致 Q 函数值的增大. 所以, 染色体 R 中任一基因的变异都不会使该染色体的适应度函数 Q 值变小. 因此性质 2 成立. □

不妨设网络 N 中结点总数为 n, 结点的平均度为 k, 待变异染色体 R 的平均社区规模为 c. 由于复杂网络一般为稀疏图, 为提高效率, 文中所有算法都是通过稀疏矩阵 (或称为图链表) 来实现的. 下面我们给出算法 LMA 的时间复杂度分析.

性质 3. 算法 LMA 的时间复杂度为 $\mathrm{O}(cn)$.

证明. 很显然, 算法 LMA 中时间复杂度最高的为第 7 步 (即变异点 i 取遍其邻居的所有标签, 同时计算对应的 f 函数值). 由于在算法执行过程中, 网络中任一结点 i 之邻居的标签都有可能存在重复的现象, 所以结点 i 计算 f 函数的平均次数要小于等于 k 次. 由性质 1 可知, 网络中的每个结点对其 f 函数的计算都仅需要用到它所在社区的信息, 因此结点 i 计算一次 f 函数的平均时间为 c. 又由于变异结点的数目大约为 βn 个, 所以该算法的时间复杂度不会超过 $\mathrm{O}(\beta nkc)$. 因为复杂网络一般为稀疏图 (即 k 为常数), 而 β 也为常数, 所以算法 LMA 的时

间复杂度也可表示为 O(cn). □

值得指出的是, 上述算法 LMA 采用了标签同步更新机制. 该机制虽然使 LMA 满足性质 2 (适应度函数 Q 的递增性), 但也有使文中遗传算法 LGA 陷入局部最优解的风险. 若 LMA 采用标签异步更新机制 (即把第 13 步提到 For 循环的外面), 那么就相当于为遗传算法 LGA 引入了一种自适应变异机制. 即: 在算法 LGA 运行初期, 由于每个结点更新自身标签的机会都很大, 使得网络社区结构变化较大, 算法 LMA 的随机性较强, 有利于其跳出局部最优解; 而在算法 LGA 运行后期, 由于每个结点更新自身标签的概率都很小, 网络社区结构的变化明显趋缓, 算法 LMA 几乎又变成了完全的贪婪搜索算法, 有利于其找到更精确的全局最优解. 可以看出, 这一机制和模拟退火算法[8] 中的退火机制有些类似, 它可以改善遗传算法 LGA 的性能. 所以文中算法 LMA 最终采用了标签异步更新机制.

1.6 LGA算法描述

选择算子是遗传算法的另一核心算子, 它决定了遗传算法的基本流程. 为了保存优秀个体并提高算法收敛速度, 本文采用了经常被用于求解组合优化问题的 $\mu+\lambda$ 选择方法[22] 以及该选择算子所对应的遗传算法流程. $\mu+\lambda$ 选择过程就是从父种群 (规模为 μ) 和经交叉、变异产生的子种群 (规模为 λ) 中共同选择适应度最高的 μ 个个体作为下一代父种群. 注意: 这里的父种群是指一般遗传算法中要维持的基本群体, 也就是我们通常所说的种群; 而子种群只是一个暂时生成的中间群体. 算法 LGA 的描述如下.

算法 2. 遗传算法 LGA 的基本流程

Procedure LGA
Input N, L, μ, λ　// N 表示复杂网络, L 表示 LGA 迭代次数, μ 表示父种群规模, λ 表示子种群规模
Output R　//网络社区结构, 或称网络聚类结果
Begin
1　$P \leftarrow$ 运行 μ 次 IGLP 产生初始种群
2　**For** $i=1:\ L$
3　　$P^{(\mathrm{new})} \leftarrow \varnothing$
4　　**For** $j=1:\ \lambda$
5　　$g \leftarrow$ 将单路交叉算子作用于从 P 中任选的两个体
6　　$g \leftarrow$ 将 LMA 变异算子作用于 g //变异率为 β
7　　$P^{(\mathrm{new})} \leftarrow P^{(\mathrm{new})} \cup \{g\}$
8　　**End For**
9　　$P^{(\mathrm{new})} \leftarrow P \cup P^{(\mathrm{new})}$
10　　$P \leftarrow$ 从 $P^{(\mathrm{new})}$ 中选择适应度最高的 μ 个 //$\mu+\lambda$ 选择个体
11　**End For**
12　$P \leftarrow$ 选择 P 中适应度最高的个体
End

可以看出, 算法 LGA 首先采用基于标签传播的方法 IGLP 产生初始群体, 然后通过迭代执行单路交叉、局部搜索变异 LMA 和 $\mu+\lambda$ 选择三个遗传算子来探测网络社区结构. 其中, 单路交叉算子通过使染色体交换部分社区结构信息来实现其全局搜索功能; LMA 变异算子通过一个非常巧妙的局部优化机制来实现其局部搜索功能; $\mu+\lambda$ 选择算子虽未直接作用于单个染色体, 但是它通过 "优胜劣汰, 适者生存" 的选择机制使适应度高的染色体进入下一代种群, 从而实现其全局搜索功能. 此外, 由于算法 IGLP 产生的初始群体具有精度较高且多样性较强的优点, 所以它可以有效地缩减算法的搜索范围, 从而进一步提高算法搜索效率.

下面我们给出算法 LGA 的时间复杂度分析. 不妨设网络 N 中结点总数为 n, 结点的平均度为 k, 算法 LGA 运行过程中所有染色体对应的平均社区规模为 c.

性质 4. 算法 LGA 的时间复杂度为 O(cn).

证明. 算法 LGA 中时间复杂度最高的为第 1 步和第 6 步, 而其他步骤的运行时间均小于或等于 O (n). 很显然, LGA 的第 1 步执行时间为 O (μlkn); LGA 的第 6 步至多会执行 $L\lambda$ 次 LMA 算法, 而执行一次 LMA 的时间为 O (cn) (见性质 3), 所以 LGA 执行第 6 步的时间不会超过 O ($L\lambda cn$). 因此, LGA 的时间复杂度为 O ($\mu lkn + L\lambda cn$). 又因为复杂网络一般为稀疏图 (即 k 为常数), 且文中所有参数均被视为常数, 所以 LGA 的时间复杂度也可表示为 O (cn). □

值得指出的是, 当前大多网络社区探测算法的时间复杂度都不小于 O (n^2), 即使 Newman 快速算法 (FN)[7] 的时间复杂度也仅为 O (n^2). 而本文算法 LGA 时间复杂度为 O (cn), 且其中的网络平均社区规模 c 要远小于整个网络规模 n. 因此, LGA 很适合于聚类真实世界中的大规模复杂网络.

2 实验

为了定量分析算法 LGA 的性能, 我们利用基准测试网络和真实大规模网络对其进行验证, 之后给出参数分析. 算法实验环境为: 处理器 Intel (R) Xeon (R) CPU 5130 @ 2.00 GHz 2.00 GHz, 内存 4.00 GB, 硬盘 160 G, 操作系统 Microsoft windows server 2003. 编程环境为 Matlab 7.3.

算法 LGA 中共有五个参数, 分别为: LGA 迭代次数 L、父种群规模 μ、子种群规模 λ、变异率 β 和 IGLP 标签传播次数 l. 其中前四个参数是遗传算法的标准参数, 我们在文献 [14−15, 22] 的基础上,

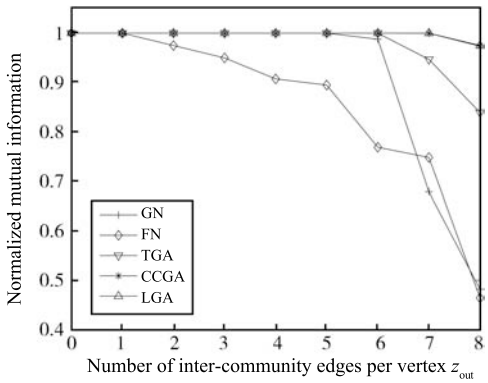
并根据经验将其分别设置为 $L = 200$, $\mu = 80$, $\lambda = 60$, $\beta = 0.2$; 对于 IGLP 的标签传播次数 l, 我们参考文献 [24] 将其设置为 $l = 5$, 并在后面实验中对该参数设置的合理性进行了详细分析.

2.1 计算机生成的网络

2002 年, Newman 等[3] 提出了一个用于测试网络社区探测算法性能的随机网络模型 $RN(a, s, d, z_{\text{out}})$. 该模型的社区结构已知, 其中 a 代表网络中社区的个数, s 代表每个社区内的结点数目, d 代表每个结点的度, z_{out} 代表每个结点与社区外结点构成的边数. 它目前作为测试社区探测算法性能的基准数据集, 已被广泛应用于相关的工作当中.

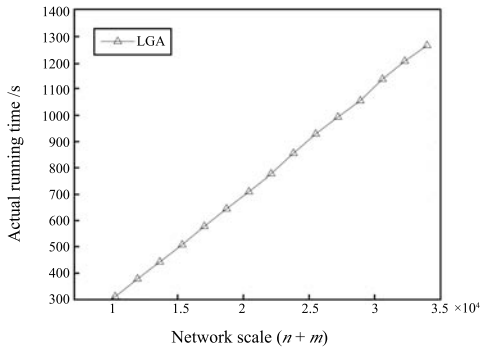
由于不同算法得到的社区数是不同的, 未必等于网络的真实社区数目. 有的算法倾向于将真实的网络社区进一步划分, 有的算法会将若干真实的网络社区分到同一类. 文献 [25] 认为, 在该情况下基于信息理论的精度度量标准 NMI (Normalized mutual information) 较其他精度度量标准更加公平合理. 所以, 文中采用 NMI 方法对不同网络社区探测算法的聚类精度进行评估. 按照该精度度量标准, 我们将算法 LGA 与算法 GN[3], FN[7], TGA (Tasgin′s genetic algorithm)[14], CCGA (Clustering combination based genetic algorithm)[15] 进行比较. 其中算法 GN 和 FN 由 Newman 提出, 是以 Q 函数作为目标函数的经典网络社区探测算法; 算法 TGA 虽仅发表于 arXiv.org 上, 但这并不影响其成为当前最为经典且被引用率最高的遗传型网络社区探测算法; 算法 CCGA 也是目前一个优秀的遗传型网络社区探测算法. 图 1 (a) 给出了实验结果, 这里所采用的随机网络是被普遍采用的基准随机网络 $RN(4, 32, 16, z_{\text{out}})$. 很显然, 当 z_{out} 越大时, 网络对应的社区结构就越模糊, 它给网络社区探测算法带来的挑战也就越大. 当 z_{out} 大于 8 时, 即网络社区间的边数大于社区内的边数, 该网络被认为不具有社区结构[3]. 图 1 (a) 中, y 轴表示聚类精度, x 轴表示 z_{out}, 曲线上的每个数据点是采用不同算法聚类 50 个随机网络得到的平均准确率. 可以看出, 文中算法 LGA 的聚类精度与算法 CCGA 相当, 但要明显高于算法 GN, FN 和 TGA, 且 z_{out} 越大 (网络社区结构越模糊) LGA 对于这三个算法的优越性就越明显. 此外, 即使当 $z_{\text{out}} = 8$ 时 (网络社区内部的边数和社区间的边数相等), 算法 LGA 仍能够正确划分 97.48 % 的网络结点, 而此时算法 GN, FN 和 TGA 的聚类精度已经比较低了. 值得指出的是, 这里算法 CCGA 虽然具有与文中算法 LGA 相当的聚类精度, 但是其时间复杂度高, 收敛速度慢, 无法聚类真实世界的大规模网络, 其详细分析见第 2.2 节.

计算速度是另一个评价网络社区探测算法性能的重要指标. 第 1.6 节中性质 4 已给出对算法 LGA 的时间复杂性分析, 本节从实验角度来进一步评价该算法的运行效率. 图 1 (b) 给出了文中算法实际运行时间随网络规模变化的趋势. 实验中采用了随机网络 $RN(a, 100, 16, 5)$ 进行测试, 该网络的社区结构确定, 但其网络社区的个数可由 a 值调节, 共包含 $100\,a$ 个网络结点, $1\,600\,a$ 条网络连接. 图 1 (b) 中, y 轴表示算法实际运行时间 (秒), x 轴表示网络规模 (结点数 + 连接数). 可以看出, 在网络平均社区规模一定的前提下, 算法 LGA 的运行时间与复杂网络的规模近似成正比. 因此, 该实验不仅验证了性质 4 (算法 LGA 的时间复杂度为 $\mathrm{O}(cn)$) 的正确性, 而且表明算法 LGA 运行过程中所有 LAR 染色体的平均社区规模 c 与网络的真实社区结构所对应的平均社区规模 s 成正比. 实际上, 在真实世界的大规模复杂网络中, 紧凑社区结构所对应的平均社区规模一般都要远小于整个网络的规模[26].



(a) 算法 LGA 与算法 GN, FN, TGA, CCGA 聚类精度的比较

(a) Comparing LGA with GN, FN, TGA and CCGA in terms of NMI accuracy



(b) 算法 LGA 的实际运行时间随网络规模增大的变化趋势

(b) The actual running time of LGA on networks with different scales

图 1 采用随机网络测试算法 LGA 的性能

Fig. 1 Testing the performance of LGA on random networks

2.2 真实的复杂网络

由于真实世界的复杂网络或许与计算机生成的网络具有一些不同的拓扑属性, 本文采用了七个目前已被广泛使用的真实网络作为测试数据集, 试图来进一步验证 LGA 的性能. 这些数据不仅有包含几十个结点、几百条边的小规模网络, 也有包含上万个结点、数十万条边的大规模网络. 表 2 给出了对这些网络的简单描述.

表 2　实验中用到的真实网络

Table 2　Real-world networks used in our experiments

Networks	V (G)	E (G)	Description
Karate	34	78	Zachary's karate club[27]
Dolphin	62	160	Dolphin social network[28]
Polbooks	105	441	Books about US politics[9]
Football	115	613	American College football[3]
Jazz	198	5 484	Jazz musicians network[29]
World	7 207	31 784	Semantic network[30]
Arxiv	56 276	315 921	scientific collaboration network[31]

针对表 2 中的真实复杂网络, 我们将文中算法 LGA 与算法 GN, FN, TGA, CCGA 进行比较, 其中算法 TGA 和 CCGA 的遗传参数分别按文献 [14−15] 进行设置, 算法 GN 和 FN 是免参数的. 由于上述 5 个算法都是以 Q 函数作为目标函数, 在此我们以 Q 函数作为网络聚类结果优劣的评价标准. 表 3 给出了上述算法分别运行 50 次得到的平均 Q 函数值, 其中 "—" 表示算法内存溢出或 48 小时未运行出结果. 可以看出, 算法 LGA 对于真实复杂网络的社区探测结果也要优于其他算法. 值得指出的是, 虽然对于小规模网络, 算法 TGA 和 CCGA 得到的 Q 函数值与文中算法 LGA 得到的结果较为接近, 但是对于两个大规模网络, 算法 TGA 得到聚类结果的质量很差, 其对应的社区结构无意义, 而算法 CCGA 由于时间复杂度高、收敛速度缓慢, 根本就无法在限定时间内 (48 小时) 得到大规模网络的聚类结果.

在这七个真实网络中, 只有 Karate, Dolphin 和 Football 三个网络的社区结构是已知的. 因此, 下面我们针对这三个网络, 参照其真实社区结构对算法 LGA 的运行结果进行进一步分析. 值得指出的是, 对于其中每个复杂网络, 算法 LGA 的每次运行结果几乎都是相同的.

空手道俱乐部网络 (Karate) 是 Zachary 通过对一个美国大学空手道俱乐部历时两年的观测而构建的社会网[27]. 它以俱乐部中的 34 个成员作为结点, 如果两个成员之间存在友谊关系, 那么他们对应的顶点之间就会有一条边相连. 由于一些分歧, 这个俱乐部后来分裂为两个独立俱乐部, 它们分别以管理员和教练作为领导者. 该网络的真实社区结构如图 2 (a) 所示: 其中结点 1 代表俱乐部教练, 结点 33 代表俱乐部的管理者; 左侧所有方形结点代表俱乐部分裂后与教练在同一组的成员, 右侧所有三角形结点代表俱乐部分裂后与管理者在同一组的成员. 我们以该网络为例随机运行一次算法 LGA, 得到的结果包含了四个社区, 如图 2 (a) 所示. 可以看出, 算法 LGA 不仅完全正确地探测出了 Karate 网络的真实社区结构, 而且还将两个真实社区分别又细分成了两个紧凑的子社区. 此外, 我们针对 Karate 网络运行 50 次算法 LGA 得到的平均 Q 函数值为 0.4198, 它比该网络真实社区结构所对应的 Q 值 0.3715 还要大.

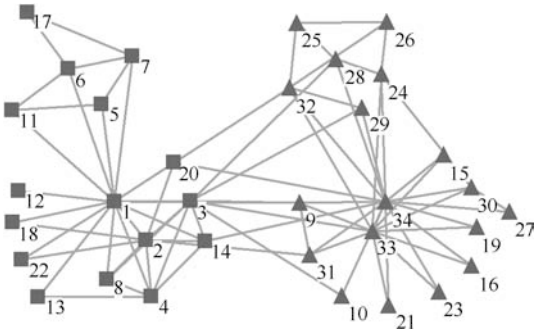
表 3　算法 LGA 与算法 GN, FN, TGA, CCGA 聚类质量的比较

Table 3　Comparing LGA with GN, FN, TGA and CCGA over 50 runs

Q-values	GN	FN	TGA	CCGA	LGA
Karate	0.4013	0.2528	0.4039	0.4198	0.4198
Dolphin	0.4706	0.3715	0.5241	0.5273	0.5280
Polbooks	0.5168	0.5020	0.5245	0.5269	0.5272
Football	0.5996	0.4549	0.5937	0.6005	0.6046
Jazz	0.4051	0.4030	0.4406	0.4445	0.4449
World	—	0.3821	0.1450	—	0.4339
Arxiv	—	0.5953	0.3883	—	0.6294

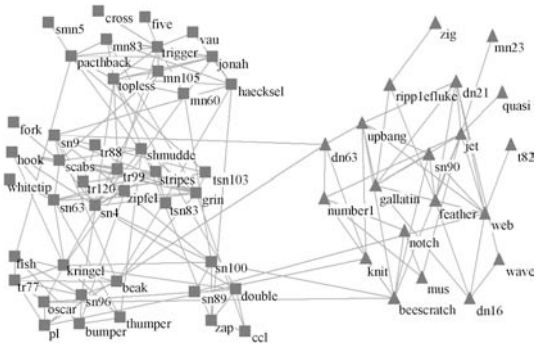
海豚网络 (Dolphin) 是 Lusseau 通过对新西兰神奇湾中 62 个不同性别海豚的观测而构建的动物社会网[28]. 每个海豚代表一个顶点, 如果两个海豚间联系频繁, 那么它们对应的顶点之间就会有一条边相连. 这些海豚被天然的分为雄性海豚组和雌性海豚组两个社区. 该网络的真实社区结构如图 2 (b) 所示: 其中左侧所有方形结点代表雌性海豚社区, 右侧所有三角形结点代表雄性海豚社区. 我们以该网络为例随机运行一次算法 LGA, 得到的结果包含了五个社区, 如图 2 (b) 所示. 可以看出, 算法 LGA 不仅完全正确地探测出了 Dolphin 网络的真实社区结构, 而且还将雌性海豚社区细分为了 4 个紧凑的子社区. 此外, 我们针对 Dolphin 网络运行 50 次算法 LGA 得到的平均 Q 函数值为 0.5280, 它比该网络真实社区结构所对应的 Q 值 0.3722 还要大.

足球联盟网络 (Football) 是 Newman 根据美国大学足球队在 2000 年秋季的常规赛计划而构建的社会网[3]. 该网络的每个结点表示一支球队, 每条



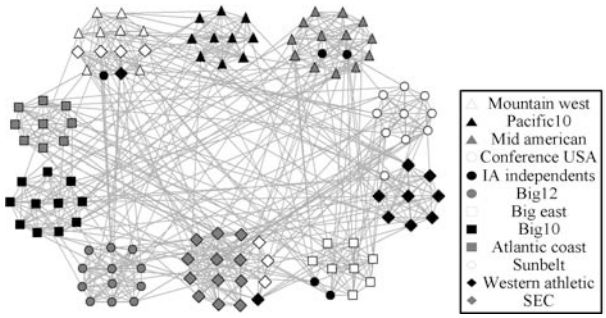
(a) LGA 对空手道网络的聚类结果

(a) Clustering result of karate network



(b) LGA 对海豚网络的聚类结果

(b) Clustering result of dolphin network



(c) LGA 对足球联盟网络的聚类结果

(c) Clustering result of football network

图 2 算法 LGA 对真实网络的聚类结果 (其中相同颜色相同形状的结点属于同一个真实社区, 而每一个紧凑的结点堆是算法 LGA 得到的一个社区)

Fig. 2 Clustering results of LGA against three real-world networks (The nodes with the same color and shape belong to a same (real) community, while each compact node set denotes a community which is got by our algorithm LGA.)

边的权重表示两球队将要比赛的频度, 共有 115 个结点、616 条边, 并被划分成 12 个联合会. 由于同联合会内球队之间的比赛频度一般会高于不同联合会间球队的比赛频度, 因此联合会即可表示为网络的真实社区结构. 该网络的真实社区结构如图 2 (c) 所示: 其中同一类结点 (相同颜色相同形状的结点) 表示同一联合会中的球队, 右下角方框中给出了每类结点所代表的联合会名称. 我们以该网络为例随机运行一次算法 LGA, 得到的结果包含了 10 个社区, 如图 2 (a) 所示. 可以看出, 除 Sunbelt 和 IA Independents 两个联合会之外, 其他 10 个联合会中的球队几乎完全被正确分类. Sunbelt 中 7 只球队被分为两类, 并分别依附于与其联系紧密的联合会中, 文献 [3] 认为这种划分是合理的; IA Independents 中的 5 支球队被分为三类, 并分别依附于三个与其联系紧密的联合会中, 这主要是由于这些球队是独立球队, 它们与其他联合会中球队的比赛要比与联合会内球队的比赛还要多[3]. 总之, 算法 LGA 将 Football 网络中的所有 115 支球队都划分在了与其联系紧密的社区之中. 此外, 我们针对 Football 网络运行 50 次算法 LGA 得到的平均 Q 函数值为 0.6046, 它比该网络真实社区结构所对应的 Q 值 0.5518 还要大.

2.3 参数分析

文中算法共有 5 个参数, 其中前 4 个是遗传算法的标准参数, 比较容易确定; 而第 5 个参数 (IGLP 的标签传播次数 l) 决定了算法 IGLP 是否能够产生具有一定聚类精度且多样性较强的初始种群, 它是本节讨论的重点.

采用基准的随机网络 $RN(4, 32, 16, 7)$ 作为测试数据集. 使算法 IGLP 随机运行 10 次, 产生 10 个独立个体, 它们的聚类精度 (采用 NMI 精度度量) 由表 4 给出. 可以看出, 这些个体的平均 NMI 精度为 52.60 %, 其具有一定的聚类精度.

为了表明由 IGLP 生成的初始种群多样性较强, 我们计算任意两个个体之间的相似度. 这里我们采用了两个非常著名的相似度度量方法, 它们分别是 Jaccard 相似度系数 (Similarity coefficient)[24] 和基于信息理论的相似度度量标准 NMI [25]. 表 5 给出了计算结果, 其中表 5 的下三角为上述 10 个个体间的 Jaccard 相似度, 而它的上三角为这 10 个个体间的 NMI 相似度. 很显然, 由 IGLP 随机生成的 10 个个体中, 任意两个个体间的相似程度都很低 (小于 0.5), 而且不管是采用 Jaccard 还是 NMI 作为相似度度量标准都是如此. 由此可见, 算法 IGLP 生成的初始种群多样性很强.

3 结论

本文以 Q 函数作为目标函数, 采用字符串编码方式, 提出了一个结合局部搜索的遗传算法 LGA. 该算法首先采用标签传播方法 IGLP 产生初始种群, 然后通过迭代执行单路交叉、局部搜索变异 LMA 和 $\mu + \lambda$ 选择三个遗传算子来探测网络社区结构. 最

表 4　针对网络 $RN(4, 32, 16, 7)$, 算法 IGLP 随机产生的 10 个体的 NMI 精度

Table 4　NMI accuracies of the ten individuals by IGLP on $RN(4, 32, 16, 7)$

个体编号	1	2	3	4	5	6	7	8	9	10	平均精度 (%)
准确率 (%)	80.29	66.23	73.58	26.49	40.93	73.25	40.43	59.02	32.80	32.89	52.60

表 5　表 4 中 10 个个体两两间的 Jaccard 相似度及 NMI 相似度

Table 5　Similarities between the ten individuals of table 4 in terms of Jaccard and NMI

个体编号	1	2	3	4	5	6	7	8	9	10
1	—	0.5003	0.6321	0.2391	0.3485	0.5447	0.2715	0.5311	0.3346	0.2927
2	0.4246	—	0.4665	0.2207	0.3375	0.5551	0.4284	0.3014	0.1978	0.2525
3	0.5715	0.3924	—	0.2252	0.3419	0.4634	0.2562	0.5352	0.2939	0.2697
4	0.1593	0.1822	0.1578	—	0.2583	0.1766	0.1202	0.2434	0.2427	0.2291
5	0.2668	0.3195	0.2687	0.1329	—	0.3263	0.3150	0.2495	0.3392	0.2226
6	0.4362	0.5088	0.3638	0.1792	0.2628	—	0.4044	0.3206	0.2219	0.2819
7	0.2909	0.4645	0.2882	0.1927	0.3962	0.4224	—	0.2066	0.1134	0.1732
8	0.4224	0.3073	0.4237	0.2084	0.1765	0.3258	0.2889	—	0.2747	0.2618
9	0.2142	0.1657	0.1962	0.0901	0.2113	0.1412	0.1679	0.1297	—	0.2464
10	0.2531	0.2612	0.2587	0.1225	0.1279	0.2369	0.1972	0.2793	0.0877	—

后通过实验表明, 算法 LGA 具有收敛速度快、搜索能力强的特点, 对于包含上万个结点、数十万条边的大规模复杂网络仍能够较快的获得高质量聚类结果.

下面我们将文中算法 LGA 与经典网络社区探测算法 GN 进行比较. GN 算法是通过反复地计算边界数 (Edge betweenness)、识别社区间连接和删除社区间连接, 以自顶向下的方式建立一棵层次聚类树, 从而探测网络社区结构. 该算法具有很高的时间复杂性 ($O(m^2n)$), 只适合处理包含几百个结点的中小规模复杂网络. 而文中算法 LGA 与 GN 完全不同, 它是以网络模块性函数 Q 作为目标函数, 通过一种结合局部搜索的遗传算法机制来探测网络社区结构. 该算法的主要贡献为: 在分析 Q 函数局部单调性的基础上, 提出一种快速、有效、结合局部搜索的变异策略; 同时, 采用标签传播方法产生兼顾精度和多样性的初始种群, 进一步提高算法搜索效率. 经分析, 算法 LGA 的时间复杂度为 $O(cn)$, 其中 c 要远小于 n. 很显然, LGA 的时间复杂度要远小于算法 GN 的时间复杂度. 最后, 通过在基准网络及大规模复杂网络上进行实验, 结果表明算法 LGA 的聚类质量和运行效率都要明显优于算法 GN.

我们以后的工作主要是将算法 LGA 应用于生物网络分析或 Web 社区挖掘等研究领域, 并试图从中探测和揭示出具有重要意义的真实社区结构.

References

1 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. *Nature*, 1998, **393**(6638): 440−442

2 Adamic L A, Huberman B A, Barabasi A L, Albert R, Jeong H, Bianconi G. Power-law distribution of the world wide web. *Science*, 2000, **287**(5461): 2115a

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

4 Yan G, Chen G, Lv J, Fu Z Q. Synchronization performance of complex oscillator networks. *Physical Review E*, 2009, **80**(5): 056116

5 Fortunato S. Community detection in graphs. *Physics Reports*, 2010, **486**(3−5): 75−174

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

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

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

9 Newman M E J. Modularity and community structure in networks. *Proceedings of National Academy of Sciences of the United States of America*, 2006, **103**(23): 8577−8582

10 Lv Z, Huang W. Iterated tabu search for identifying community structure in complex networks. *Physical Review E*, 2009, **80**(2): 026130

11 Barber M J, Clark J W. Detecting network communities by propagating labels under constraints. *Physical Review E*, 2009, **80**(2): 026129

12 Liu X, Murata T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A*, 2010, **389**(7): 1493−1500

13 Brandes U, Delling D, Gaertler M, Goerke R, Hoefer M, Nikoloski Z, Wagner D. Maximizing modularity is hard. arXiv: physics/0608255, 2006

14 Tasgin M, Herdagdelen A, Bingol H. Community detection in complex networks using genetic algorithms. arXiv: 0711.0491, 2007

15 He Dong-Xiao, Zhou Xu, Wang Zuo, Zhou Chun-Guang, Wang Zhe, Jin Di. Community mining in complex networks-clustering combination based genetic algorithm. *Acta Automatica Sinica*, 2010, **36**(8): 1160−1170
(何东晓, 周栩, 王佐, 周春光, 王喆, 金弟. 复杂网络社区挖掘 — 基于聚类融合的遗传算法. 自动化学报, 2010, **36**(8): 1160−1170)

16 He Dong-Xiao. Research on Intelligent Algorithms for Network Community Mining [Master dissertation], Jilin University, China, 2010
(何东晓. 网络社区智能挖掘算法的研究 [硕士学位论文], 吉林大学, 中国, 2010)

17 Li S, Chen Y, Du H, Feldman M W. A genetic algorithm with local search strategy for improved detection of community structure. *Complexity*, 2010, **15**(4): 53−60

18 Pizzuti C. Community detection in social networks with genetic algorithms. In: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation. New York, USA: ACM, 2008. 1137−1138

19 Pizzuti C. A multi-objective genetic algorithm for community detection in networks. In: Proceedings of the 21st IEEE International Conference on Tools with Artificial Intelligence. New Jersey, USA: IEEE, 2009. 379−386

20 Shi C, Yan Z, Wang Y, Cai Y, Wu B. A genetic algorithm for detecting communities in large-scale complex networks. *Advances in Complex Systems*, 2010, **13**(1): 3−17

21 Jin D, He D, Liu D, Baquero C. Genetic algorithm with local search for community mining in complex networks. In: Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence. Arras, France: IEEE, 2010. 105−112

22 Mezura-Montes E, Coello C A C. A simple multimembered evolution strategy to solve constrained optimization problems. *IEEE Transactions on Evolutionary Computation*, 2005, **9**(1): 1−17

23 Fortunato S, Barthelemy M. Resolution limit in community detection. *Proceedings of National Academy of Sciences of the United States of America*, 2007, **104**(1): 36−41

24 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

25 Danon L, Diaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, **2005**(9): P09008−09008

26 Leskovec J, Lang K J, Dasgupta A, Mahoney M W. Statistical properties of community structure in large social and information networks. In: Proceedings of the 17th International Conference on World Wide Web. Beijing, China: ACM, 2008. 695−704

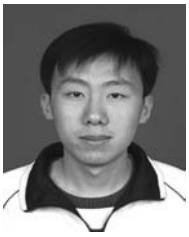
27 Zachary W W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 1977, **33**(4): 452−473

28 Lusseau D. The emergent properties of a dolphin social network. *Proceedings of the Royal Society B: Biological Sciences*, 2003, **270**(S2): 186−188

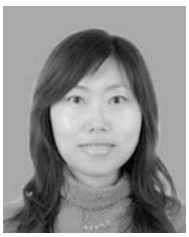
29 Gleiser P M, Danon L. Community structure in jazz. *Advances in Complex Systems*, 2003, **6**(4): 565−573

30 Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structures of complex networks in nature and society. *Nature*, 2005, **435**(7043): 814−818

31 Newman M E J. The structure of scientific collaboration networks. *Proceedings of National Academy of Sciences of the United States of America*, 2001, **98**(2): 404−409



金 弟 吉林大学博士研究生. 主要研究方向为数据挖掘, 复杂网络分析和多 Agent 系统.
E-mail: jindi.jlu@gmail.com
(**JIN Di** Ph. D. candidate at Jilin University. His research interest covers data mining, complex networks analysis, and multi-agent systems.)



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

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

何东晓 吉林大学博士研究生. 主要研究方向为数据挖掘和复杂网络分析.
E-mail: hedongxiaojlu@gmail.com
(**HE Dong-Xiao** Ph. D. candidate at Jilin University. Her research interest covers data mining and complex networks analysis.)

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.)