基于簇相似度的网络社团结构探测算法*

袁超 柴毅†

(重庆大学自动化学院, 重庆 400030)

(2012 年 3 月 11 日收到; 2012 年 5 月 3 日收到修改稿)

社团结构对复杂系统的结构特性和动力学特性有重要影响. 提出了一个度量社团相似度的模型, 称为簇相似度. 该模型能够度量两个社团的相似度大小, 为研究社团间的作用机制提供帮助. 而且基于该模型, 设计了一个社团划分算法. 算法采用层次聚类的思想, 每次合并两个相似度最大的社团, 并通过一个评价函数选择最优社团划分. 数值实验以及与 CNM, GN, EigenMod 等主流算法做比较, 表明本算法的精度和效率都比较高, 尤其对于边密度较高的网络, 性能非常理想.

关键词: 复杂网络, 社团结构, 相似度, 聚类

PACS: 89.75.Fb, 89.75.Hc, 02.10.Ox

1 引 言

社团结构划分问题是复杂网络理论研究的重要内容, 也是现代网络科学领域面临的一项长期挑战. 所谓网络社团结构, 即基于不同概念如节点相似性等形成的一些节点组 [1]. Newman 给出的定义是社团内的节点连接紧密, 而社团之间的节点连接稀疏 [2−4]. 最近的研究表明, 复杂网络在社团尺度上所反映的结构特性与网络整体所反映的结构特性有很大不同, 忽略网络的社团结构会漏掉许多有意义的结构特性 [5,6]. 因此, 复杂网络社团结构的研究越来越受到人们的关注, 成为一个热门的研究领域.

多年来, 研究者们提出了许多经典的社团探测算法, 如 Kernighan-Lin 算法 [7]、GN 算法 [2]、Newman 快速算法 [3] 以及标号传播法 [8] 等. 归纳起来, 这些算法大致分为两类, 即非重叠社团探测算法和重叠社团探测算法 [9]. 非重叠社团探测算法主要包括模块度优化算法 (大多基于层次聚类的思想和极值优化思想, 如文献 [2,3])、标号传播法 [8]、谱分析法 [10]、基于信息论的方法 [11]、基于网络动力学模型的算法 [12,13]、投影聚类算法 [14] 以及基于节点相似度的算法 [15] 等. 重叠社团探测算法主要包括派系过滤算法 [16]、模糊聚类算法 [17]、基于极值优化思想的算法 [18]、基于概率模型的算法 [19] 以及基于对偶图的算法 [20] 等. 此外, 文献 [21] 提出了一种时间演化网络的社团探测算法, 相对于上述的静态网络社团探测算法, 该算法属于动态网络社团探测算法. 总之, 上述算法有的精度较好, 有的效率较高, 但总体上还不甚理想. 到目前为止, 也没有一个十分有效的社团探测方法. 因此, 研究高精度、高效率、并行计算的社团探测算法仍是未来追求的目标.

基于节点相似度的社团探测算法是一类重要的算法. 比较著名的有 single linkage 算法和 complete linkage 算法 [22]. single linkage 算法将社团称作组件 (component), 并以节点相似度由大到小的顺序在节点间加边. 假设当前连接的两个节点相似度为 x, 则对于之前连接的那些点, 如果任意两个节点间的相似度值大于或等于 x, 则它们属于同一个社团. 算法最后给出一个层次树, 由用户自己选择最优社团划分. complete linkage 算法则基于团 (clique) 的概念来进行社团划分, 将社团定义为

* 国家自然科学基金 (批准号: 60974090) 和高等学校博士学科点专项科研基金 (批准号: 200806110016) 资助的课题.

† E-mail: cqchaiyi@yahoo.com

网络中的最大团. 以上两种算法由于性能的原因, 较少应用. 文献 [15] 则给出了一个节点相似度模型, 并采用一种迭代的思想进行社团划分. 算法首先任选一个节点作为当前节点, 然后选择与该节点相似度最大的节点进行合并, 同时令新加入的节点为当前节点, 并重复前面的操作. 合并过程中, 如果与当前节点最相似的点已经在社团中, 则另选一个没有遍历过的节点重复以上操作. 如此进行下去, 直到遍历完所有节点, 即完成社团结构划分. 该算法的缺点是容易过早收敛, 难以发现最优社团结构. 还有其他一些算法, 如基于高斯核函数的节点相似度算法等. 总的来说, 基于节点相似度的算法存在以下几点问题: 1) 大部分算法需要借助模块度 Q 作为社团划分的评价函数, 而这又带来两个问题, 一是模块度 Q 本身有分辨率的问题, 二是模块度 Q 并不一定对基于节点相似度的划分方式有效; 2) 算法精度不高; 3) 有的算法复杂度较高.

事实上, 基于节点相似度划分社团结构的思想还是很有道理的. 而且如果模型和算法设计合理的话, 精度还是比较高的. 目前, 虽然度量节点相似度的模型很多, 但是度量社团相似度的模型则极为少见. 社团相似度的度量不仅有助于设计社团探测算法, 而且有助于研究复杂系统的结构特性. 本文基于 Dice 模型, 推导出节点相似度模型, 进而提出了一个用于度量社团相似度的模型, 即簇相似度. 并以该模型为基础, 设计了一个精度和效率都较高的算法. 算法采用层次聚类的思想, 按照簇相似度由大到小的顺序依次合并各子社团, 并计算整个网络的平均簇相似度 (文中称为 S 值). 最后, 选择最小 S 值所对应的划分即为最优社团划分. 我们将算法应用于一系列计算机生成网络和实际网络, 并与一些主流算法做了比较, 其性能令人满意.

本文第 2 节介绍簇相似度模型; 第 3 节介绍基于该模型的社团探测算法; 第 4 节给出数值测试和比较结果; 第 5 节对全文进行总结.

2 簇相似度模型

簇相似度用于度量两个社团的相似度大小, 下面给出其推导过程. 我们首先基于向量相似度模型推导节点相似度模型. 目前向量相似度模型很多, 如 Pearson 相关系数 [23]、余弦相似度 [24]、Dice 系数 [25] 等. 由于 Dice 系数属于二元向量的相似性函数, 其元素只有 0 和 1, 与无权网络邻接矩阵中的元素相似, 所以本文基于该模型来推导节点相似度模型.

Dice 系数定义如下: 给定两个向量 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$, $w_k(\boldsymbol{D}_1)$ 和 $w_k(\boldsymbol{D}_2)$ 分别表示向量 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$ 的第 k 个特征项的权值 (此处为 0 或 1), 则两个向量的 Dice 相似度为

$$\mathrm{Sim}(\boldsymbol{D}_1,\boldsymbol{D}_2)=\frac{2\times\sum\limits_{k=1}^{n}(w_k(\boldsymbol{D}_1)\times w_k(\boldsymbol{D}_2))}{\sum\limits_{k=1}^{n}w_k^2(\boldsymbol{D}_1)+\sum\limits_{k=1}^{n}w_k^2(\boldsymbol{D}_2)}, \quad (1)$$

假设网络的邻接矩阵为 $\boldsymbol{A}$, 则我们可将 $\boldsymbol{A}$ 的每行看作相应节点的特征向量. 特征向量中的每个特征项表征节点与其他节点之间是否有连边. 这样, 对于两个节点向量, (1) 式中的分子即为它们的点积乘以 2. 观察分母, 我们可以发现节点特征项的平方和即为节点的度 (因为节点特征向量中的元素为 0 和 1, 所以平方值不变). 因此, 对于任意两个节点向量 $\boldsymbol{i}$ 和 $\boldsymbol{j}$, 将它们代入 (1) 式, 可得节点相似度计算公式如下:

$$\mathrm{Sim}(i,j)=\frac{2\times\sum\limits_{k=1}^{n}(\boldsymbol{A}(i,k)\times\boldsymbol{A}(j,k))}{d_i+d_j}, \quad (2)$$

(2) 式中 $\mathrm{Sim}(i,j)$ 表示节点 i 和 j 之间的相似度, n 为网络的节点数, $\boldsymbol{A}$ 为网络的邻接矩阵, d_i 和 d_j 分别为节点 i 和节点 j 的度.

然而, 如图 1 所示, 我们需要考虑两类情况: 1) 两个节点之间没有连边 (图 1(a)); 2) 两个节点之间存在一条连边 (图 1(b)). (2) 式中的分子项仅满足第 1) 类情况. 而对于第 2) 类情况, (2) 式中的分子项反映不出节点 i 和节点 j 之间存在一条连边这一相似特征. 不过, 我们可以这样考虑, 即认为该边是由一个虚拟节点连接的两条边组成, 如图 1(c) 所示. 可认为边 ij 是由虚拟节点 f 连接的两条边 if 和 jf 组成. 这样一来, 节点 i 和节点 j 之间的连接就会断开. 反映在邻接矩阵 $\boldsymbol{A}$ 中, 那就是 i 行 j 列或者 j 行 i 列的元素由 1 变为 0, 并且需要新加入一行一列, 与虚拟节点 f 相对应. 假设图 1(c) 对应的邻接矩阵为 $\boldsymbol{A'}$, 那么新添加的行和列上仅有 $\boldsymbol{A'}(i,f)$, $\boldsymbol{A'}(f,i)$, $\boldsymbol{A'}(j,f)$ 和 $\boldsymbol{A'}(f,j)$ 四个元素为 1, 其余为 0.

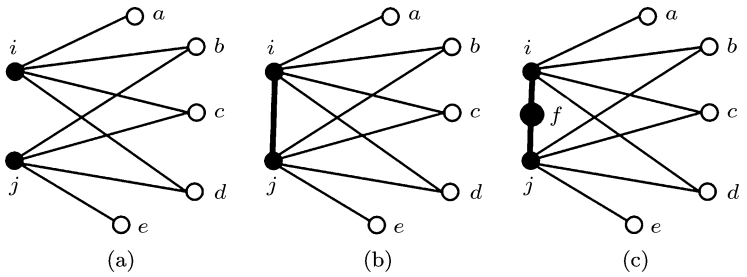

图 1 复杂网络示意图 (a) 节点 i 和节点 j 之间没有连边; (b) 节点 i 和节点 j 之间存在一条连边; (c) 节点 i 和节点 j 之间的连边可以被看成两个节点分别与一个虚拟节点 f 连接

将 $\boldsymbol{A'}$ 代入 (2) 式, 即得

$$\begin{aligned}\mathrm{Sim}(i,j) &= \frac{2\times\sum\limits_{k=1}^{n+1}(\boldsymbol{A'}(i,k)\times\boldsymbol{A'}(j,k))}{d_i+d_j}\\ &= \frac{2\times\Big(1+\sum\limits_{k=1}^{n}(\boldsymbol{A}(i,k)\times\boldsymbol{A}(j,k))\Big)}{d_i+d_j}.\end{aligned} \tag{3}$$

这样, 第 2) 类情况就可用 (3) 式进行计算. 归纳起来, 基于 Dice 相似度模型的节点相似度模型计算公式如下:

$$\begin{aligned}&\mathrm{Sim}(i,j)\\ =&\frac{2\times\Big(\boldsymbol{A}(i,j)+\sum\limits_{k=1}^{n}(\boldsymbol{A}(i,k)\times\boldsymbol{A}(j,k))\Big)}{d_i+d_j}.\end{aligned} \tag{4}$$

本文以该模型为基础, 构建簇相似度模型. 首先, 我们认为相同社团中节点的平均相似度大于两个不同的社团间节点的平均相似度. 现假设有两个社团 C_1 和 C_2, 那么 C_1 中每个节点分别与 C_2 中每个节点的相似度的平均值 $\mathrm{Avg}_{\mathrm{similar}}(C_1,C_2)$ 可按下式计算:

$$\mathrm{Avg}_{\mathrm{similar}}(C_1,C_2)=\frac{\sum\limits_{i\in V_{C_1},j\in V_{C_2}}\mathrm{Sim}(i,j)}{N_{C_1}\times N_{C_2}}, \tag{5}$$

其中, C_1 和 C_2 为两个社团的编号, V_{C1} 和 V_{C2} 分别为两个社团的节点集合, N_{C1} 和 N_{C2} 分别为两个社团中节点的个数. 当 C_1 和 C_2 相等时, 则 (5) 式就退化为计算某个社团内节点平均相似度的公式, 即

$$\mathrm{Avg}_{\mathrm{similar}}(C_1,C_1)=\frac{\sum\limits_{i\in V_{C_1},j\in V_{C_1}}\mathrm{Sim}(i,j)}{N_{C_1}\times N_{C_1}}, \tag{6}$$

其中, C_1 为社团的编号, V_{C1} 为社团的节点集合, N_{C1} 为社团内节点的个数.

于是, 簇相似度可以这样定义: 给定两个社团 C_1 和 C_2, 簇相似度即为两个社团之间节点的平均相似度分别与它们内部节点的平均相似度的比值之和. 即

$$\begin{aligned}S(C_1,C_2)=&\frac{\mathrm{Avg}_{\mathrm{similar}}(C_1,C_2)}{\mathrm{Avg}_{\mathrm{similar}}(C_1,C_1)}\\ &+\frac{\mathrm{Avg}_{\mathrm{similar}}(C_1,C_2)}{\mathrm{Avg}_{\mathrm{similar}}(C_2,C_2)},\end{aligned} \tag{7}$$

$S(C_1,C_2)$ 越大, 则两个社团 C_1 和 C_2 越相似, 越有可能属于同一个社团. 一个最优的社团划分, 可以认为是网络中所有社团间的簇相似度平均值最低. 假设这个值为 S, 则其计算公式为

$$S=\frac{\sum\limits_{C_1,C_2\subset C,C_1\neq C_2}S(C_1,C_2)}{num}, \tag{8}$$

上式中, C 为网络中所有子社团的集合, num 为簇相似度不为 0 的社团对个数.

3 算 法

基于簇相似度模型, 本文提出了一种快速的社团划分算法. 算法首先基于节点相似度模型 (即 (4) 式) 对网络进行初始化. 然后采用层次聚类的思想, 反复合并相似度最大的两个社团, 并计算相应划分的 S 值. 如此进行, 直到整个网络还剩两个社团为止. 最后选择 S 值最小的划分即为最优社团划分.

需要指出, 社团内部节点的平均相似度以及社团之间节点的平均相似度只需要在初始化时计算一次即可. 在后面的合并过程中, 这些值可以通过上一次的值直接得到. 这样可以避免循环操作所带

来的巨大计算量, 使合并过程的时间复杂度下降到线性时间复杂度. 其计算原理如下.

如图 2 所示, 首先假设社团 i 中节点的平均相似度为 $\mathrm{Avg}_{\mathrm{similar}}(i,i)$, 社团 j 中节点的平均相似度为 $\mathrm{Avg}_{\mathrm{similar}}(j,j)$, 社团 k 中节点的平均相似度为 $\mathrm{Avg}_{\mathrm{similar}}(k,k)$, 社团 i 和社团 j 之间节点的平均相似度为 $\mathrm{Avg}_{\mathrm{similar}}(i,j)$, 社团 i 和社团 k 之间节点的平均相似度为 $\mathrm{Avg}_{\mathrm{similar}}(i,k)$, 社团 j 和社团 k 之间节点的平均相似度为 $\mathrm{Avg}_{\mathrm{similar}}(j,k)$, 社团 i 中节点的个数为 N_i, 社团 j 中节点的个数为 N_j, 社团 k 中节点的个数为 N_k. 现欲在当前划分的基础上将社团 i 和社团 j 合并为社团 ij. 那么合并后社团内节点的平均相似度值以及社团间节点的平均相似度值可按下面的公式计算.

1) 社团 ij 内部节点的平均相似度 $\mathrm{Avg}_{\mathrm{similar}}(ij,ij)$ 可按下式计算:

$$\begin{aligned}&\mathrm{Avg}_{\mathrm{similar}}(ij,ij)\\ =&\frac{\mathrm{Avg}_{\mathrm{similar}}(i,i)\times(N_i\times N_i)+\mathrm{Avg}_{\mathrm{similar}}(j,j)\times(N_j\times N_j)+2\times\mathrm{Avg}_{\mathrm{similar}}(i,j)\times(N_i\times N_j)}{(N_i+N_j)\times(N_i+N_j)}.\end{aligned}\tag{9}$$

2) 社团 k 内部节点的平均相似度 $\mathrm{Avg}_{\mathrm{similar}}(k,k)$ 不变.

3) 社团 ij 和社团 k 之间节点的平均相似度 $\mathrm{Avg}_{\mathrm{similar}}(ij,k)$ 可按下式计算:

$$\mathrm{Avg}_{\mathrm{similar}}(ij,k)=\frac{\mathrm{Avg}_{\mathrm{similar}}(i,k)\times(N_i\times N_k)+\mathrm{Avg}_{\mathrm{similar}}(j,k)\times(N_j\times N_k)}{N_k\times(N_i+N_j)}.\tag{10}$$

这样, 合并后社团 ij 和社团 k 之间的簇相似度可按照 (7) 式进行计算. 如果还有其他社团与社团 i 和社团 j 相连接, 也按照这种方法计算.

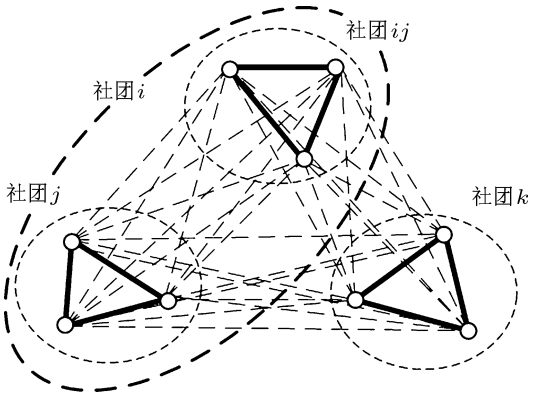

图 2 社团合并示意图 (粗虚线框表示将社团 i 和社团 j 合并为社团 ij, 粗实线表示计算社团内部节点两两之间的相似度, 细虚线表示计算社团间节点两两之间的相似度)

3.1 算法描述

本算法包括六个步骤, 具体描述如下.

1) 网络初始化, 计算节点相似度矩阵并合并最相似的节点. 具体方法为: 根据 (4) 式计算节点相似度矩阵. 每计算完一行, 就将该行最大值所对应的行号和列号 (列号表示与行号所对应节点最相似的节点) 存储到一个三元组 SV(行号, 列号, 社团编号) 中, 其中社团编号为 0 表示该行对应的节点对尚未合并. 待节点相似度矩阵计算完, 采用下面的方法合并最相似节点:

① 从 SV 中任选一对没有合并的节点对 (社团编号为 0), 如果该节点对不存在, 则表示已经完成合并, 生成初始子社团并转步骤 2); 否则将该节点对合并为新社团, 并指定节点对中第二个节点 (列号) 为当前节点;

② 选择当前节点在 SV 中所对应的行, 如果该行社团编号为 0, 则将该行第二个节点 (列号) 并入当前社团, 并指定其为当前节点; 如果社团编号不为 0, 则转步骤 ④;

③ 重复执行步骤 ②;

④ 如果社团编号为当前社团编号, 则表示节点对中第二个节点 (列号) 已经包含在当前社团中, 则转步骤 ①; 如果社团编号不为当前社团编号, 则将当前社团中所有节点并入该行所对应的社团, 并转步骤 ①.

2) 根据 (7) 式计算两两社团之间的簇相似度, 并根据 (8) 式计算当前划分的 S 值.

3) 合并簇相似度最大的两个社团. 并根据 (9), (10) 式更新相关社团的社团内和社团间节点平均相似度值.

4) 根据 (7) 式更新与合并社团有连接的社团之间的簇相似度, 并根据公式 (8) 计算当前划

分的 S 值.

5) 重复执行步骤 3) 和步骤 4), 直到整个网络还剩两个社团为止.

6) 找出 S 值最小的划分, 即为最优社团划分.

3.2 算法复杂度分析

算法的时间复杂度整体较低. 其中第 1) 步的时间复杂度最高, 主要包括两部分, 即计算节点相似度矩阵以及合并最相似节点. 计算节点相似度矩阵的时间复杂度为 $O(n^2)$; 最相似节点合并的过程需要遍历完三元组的所有行, 所以时间复杂度为线性, 约为 $O(n)$. 所以第 1) 步的时间复杂度为 $O(n^2+n)$. 第 2) 步的时间复杂度主要包括计算社团内和社团间节点平均相似度. 其复杂度小于 $O\left(\dfrac{n^2}{k}+C_k^2\times\dfrac{n^2}{k^2}\right)$, 即 $O\left(\dfrac{(k+1)\times n^2}{2k}\right)$ (k 为初始社团个数). 步骤 3), 4), 5) 重复计算四个公式, 共循环 $k-2$ 次, 所以时间复杂度为 $O(4k-8)$. 由于在合并过程中实时记录最小 S 值, 所以第 6) 步时间复杂度忽略不计. 因此算法总时间复杂度为 $O\left(\dfrac{(3k+1)}{2k}\times n^2+n+4k-8\right)$ (n 为节点个数, k 为初始社团个数), 大致为 $O(n^2)$ 水平.

4 数值实验

下面将本文算法应用于几个已知社团结构的现实网络和人工网络, 以测试算法的性能. 其中现实网络包括宽吻海豚网络 [26]、美国大学橄榄球联赛网络 [2] 和小说 << 悲惨世界 >> 中人物关系网络 [27]. 人工网络为 GN 基准图 [2,3].

4.1 宽吻海豚网络

宽吻海豚网络描述了居住在新西兰神奇湾的 62 头宽吻海豚的社会关系. 该网络由 62 个节点和 159 条边组成, 其中节点代表海豚, 边表示两头海豚之间经常保持联系. 在研究过程中, 由于一个关键成员的离开, 海豚网络分裂为两个小的群落.

将本文算法应用于该网络, 其划分过程如图 3 所示. 其中图 3(a) 为网络初始化阶段输出的结果, 共生成了 12 个社团. 后面的 10 幅图则是对这些社团进行合并的过程. 依据簇相似度, 各子社团的合并顺序如图 3(b)—(k) 所示. 其中参与合并的两个社团在每幅图中用虚线标出. 最后图 3(k) 还剩两个社团, 所以合并停止. 每次合并对应的 S 值在图 4 中列出.

由图 4 可知, 图 3(k) 所对应的 S 值最小, 为 0.0651. 所以该划分为最优社团划分. 划分结果符合真实情况, 仅有一个节点 40 划分错误 (在图中用带阴影的圆圈标出). 而对于该网络, CNM[28] 算法将其划分为 5 个社团, EigenMod[29] 算法将其划分为 4 个社团. 因此, 本文算法对该网络的划分效果更好.

4.2 美国大学橄榄球联赛网络

该网络数据来自 2000 年美国 115 所大学橄榄球队之间的比赛, 由 Girvan 和 Newman 收集整理. 网络包括 115 个节点和 613 条边. 其中节点代表球队, 边表示两支球队之间进行过一场比赛. 这些球队被分为 12 个联盟, 联盟内部的比赛次数相对较多, 而联盟之间的比赛次数相对较少.

图 5 给出了算法的测试结果. 由于划分过程与宽吻海豚网络类似, 所以中间的步骤省略, 图中只给出了网络初始化的结果 (图 5(a)) 和最优社团划分结果 (图 5(b)). 为了清楚起见, 我们将图 5(b) 整理成图 5(c) 的形式, 同一个社团中的节点被放入一个方格中, 节点的形状、颜色与原图保持一致. 划分错误的节点在图中用白色边框标出, 共有 11 个节点划分错误, 它们分别是 29, 37, 43, 59, 60, 64, 81, 83, 91, 98 和 111. 因此该划分的正确率为 90.4%. 图 6 给出了划分过程所对应的 S 值. 由图中可见, 网络初始化时共生成 32 个子社团, 与之对应的 S 值为 0.2192. 同时, 我们可以看到横坐标为 11 时所对应的 S 值最小, 为 0.1055. 因此该划分为最优社团划分 (与图 5(b) 相对应).

对于该网络, CNM 算法将其划分为 6 个社团, GN 算法将其划分为 11 个社团 (准确率仅为 78%), FCM 算法将其划分为 10 个社团 (准确率为 90%), 谱分解算法将其划分为 12 个社团 (准确率为 70%). 相比而言, 本文算法划分结果更好.

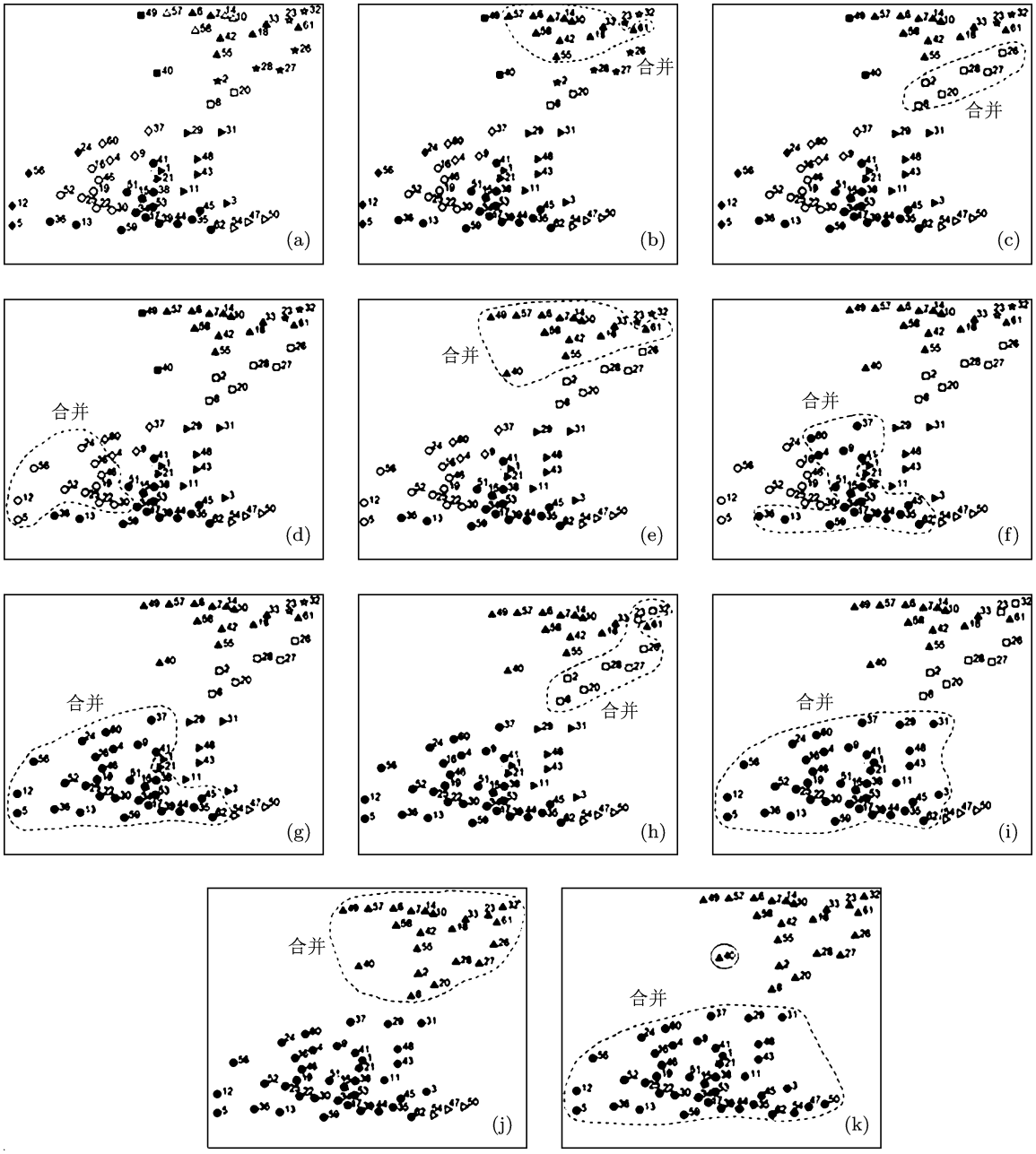

图 3 应用本文算法划分宽吻海豚网络的过程 (a) 网络初始化所生成的社团; (b)—(k) 表示对 (a) 中社团的 10 次合并过程; 当前合并的社团在每幅图中用虚线标出

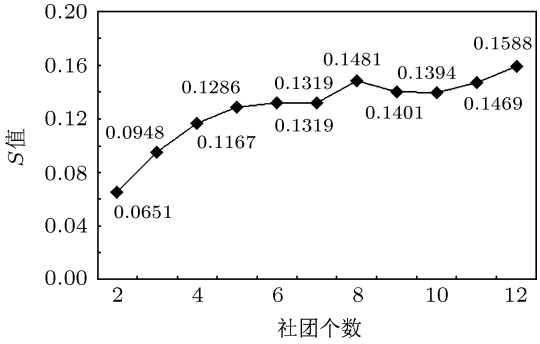

图 4 宽吻海豚网络划分过程所对应的 S 值 (横轴为社团个数, 纵轴为相应的 S 值)

4.3 小说《悲惨世界》中人物关系网络

该网络是一个加权网络, 由 Knuth 提供. 网络统计了雨果的小说《悲惨世界》中的人物在同一章中出现的次数. 其中节点表示小说中的人物, 边表示两个人在同一章中出现. 边上的权值表示出现的次数.

由于本文的节点相似度模型是无权网络的模型. 所以对该网络稍做修改, 将边上的权值统一置为 1. 图 7 为网络的划分结果, 共分成 5 个社团. 其

中, 以冉阿让 (Valjean) 为核心的中间社团最大 (黑色圆点). 这与事实相符, 因为主人公是小说人物的核心, 接触的人最广泛, 所以该社团规模最大, 而且处于中心位置. CNM 算法将该网络划分为 5 个社团, 与本文算法的划分结果大致相符. 而谱分解算法将其划分为两个社团, 差距较大.

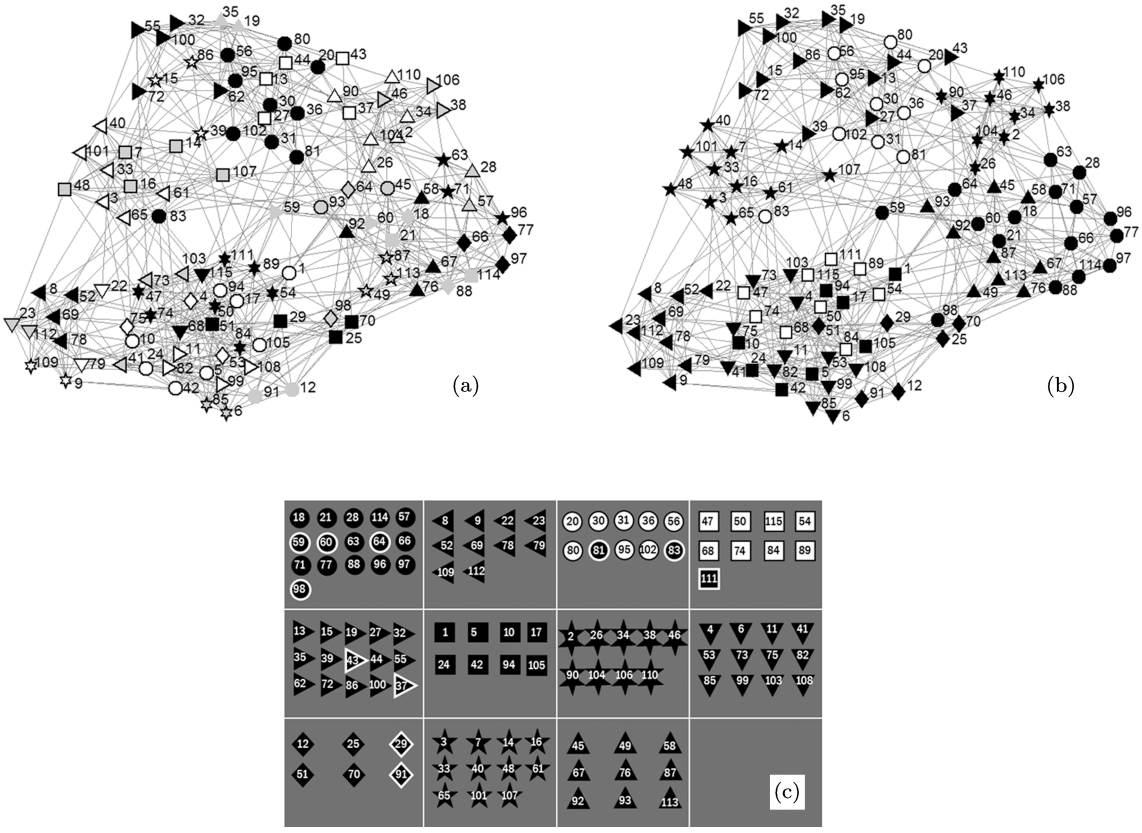

图 5 应用本文算法对美国大学橄榄球联赛网络的划分结果展示 (a) 网络初始化阶段生成的社团; (b) 最优社团结构划分; (c) 对 (b) 中输出结果的整理. 11 个社团被放入 11 个方格中, 节点的形状、颜色与图 (b) 中相对应; 其中 11 个被错误划分的节点用白色边框标出

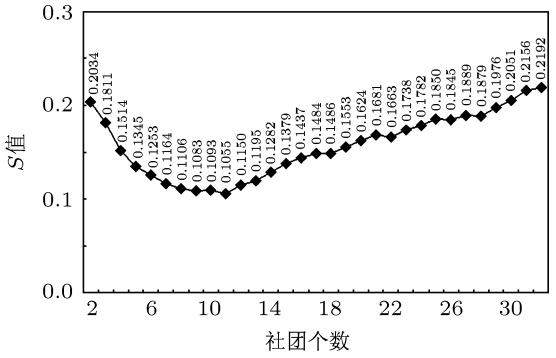

图 6 美国大学橄榄球联赛网络划分过程所对应的 S 值 (横轴为社团个数, 纵轴为相应的 S 值)

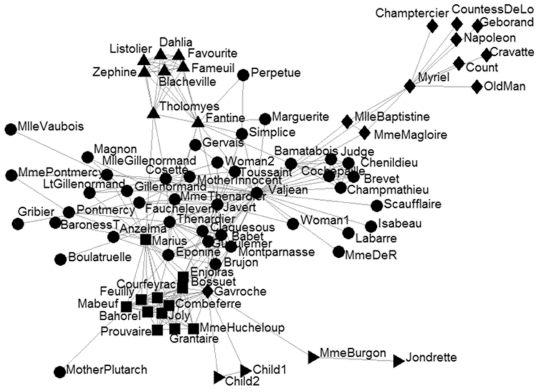

图 7 小说《悲惨世界》中人物关系网络的最优社团划分

4.4 GN 基准图

GN 基准图是一个广泛应用的人工网络, 用于测试社团探测算法的性能. 该网络由 Girvan 和 Newman 提出, 包括 128 个节点和 4 个社团, 每个社团又分别包含 32 个节点. 节点以概率 $P_{\rm in}$ 连接社团内部节点, 以概率 $P_{\rm out}$ 连接外部节点. 其中, 节点内部

度的期望值为 $Z_{\rm in}$, 外部度的期望值为 $Z_{\rm out}$. 整个网络节点的平均度为 16. 其中 $Z_{\rm out}$ 越小, 则网络社团结构越明显. 反之, 则不明显. 所以可通过调节 $Z_{\rm out}$ 的值来测试算法的性能.

图 8 为本文算法 (方形) 与 CNM 算法 (圆形) 关于该网络的社团划分性能比较. 其中图 8(a) 是典型的 GN 基准图, 节点平均度为 16. 从图中可以看出, 当 $Z_{\rm out}$ 小于 6 时, 两种算法表现都较好. 当 $Z_{\rm out}$ 等于 6 时, 本文算法性能变化不大, 表现稍好. 然而当 $Z_{\rm out}$ 为 7 时, 本文算法性能下降幅度较大. 即在 $Z_{\rm out} < 7$ 时, 本文算法性能好于 CNM 算法; 而当 $Z_{\rm out} \geqslant 7$ 时, 本文算法性能略低于 CNM 算法. 其原因是当 $Z_{\rm out}$ 大于 7 时, 社团结构模糊, 社团内外节点的相似度比较接近, 因此计算误差较大.

我们将节点的平均度增大到 28, 再次测试两种算法的性能. 这时, 本文算法表现极好. 如图 8(b) 所示, 本文算法整体优于 CNM 算法. 而且在 $Z_{\rm out} \leqslant 13.5$ 时, 本文算法性能比较稳定. 这也说明本文算法在网络边密度较高时表现更好. 因为此时节点相似度的计算更加精确. 所以, 节点相似度模型的精度对于算法有较大的影响.

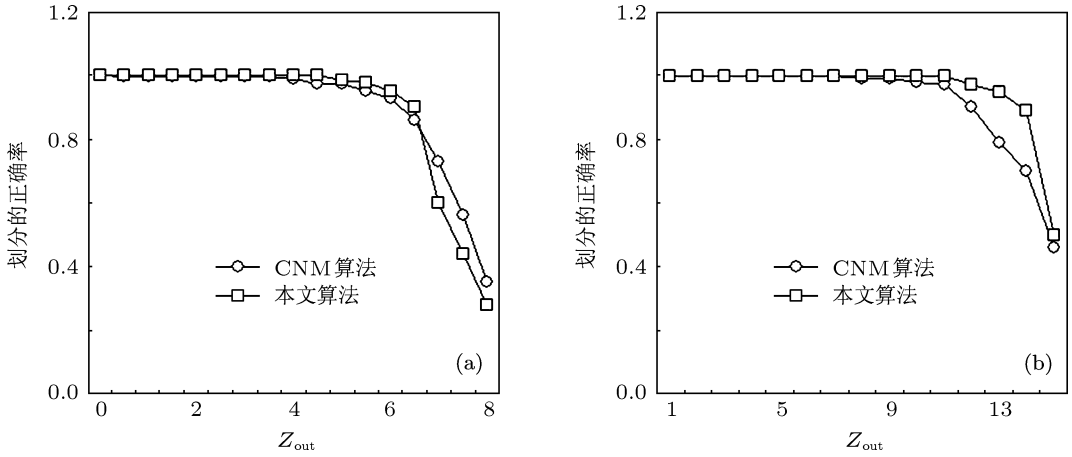

图 8 本文算法 (方形) 与 CNM 算法 (圆形) 关于 GN 基准图的社团划分性能比较 (a) 网络节点平均度为 16; (b) 网络节点平均度为 28

5 结 论

目前, 能够比较精确地度量两个社团相似度的模型比较少见. 本文在节点相似度模型的基础上, 提出了簇相似度模型, 用于度量两个社团的相似度. 该模型有助于研究社团间的作用机制, 从而帮助理解复杂系统的结构和功能. 同时, 以该模型为基础, 本文设计了一个社团探测算法. 算法以社团为单位进行层次聚类, 并通过一个评价函数来选择最优社团划分. 由于在合并过程中只需要计算两个简单的数学公式, 所以算法速度较快.

我们用一系列现实网络和人工网络进行测试, 同时与一些主流算法做了比较. 测试和比较结果表明本文所提的簇相似度模型和算法精度都较高, 尤其对于边密度较高的网络. 不足之处是当社团结构比较模糊时, 节点相似度模型的精度受到限制. 在后续研究中, 我们将寻找更精确的节点相似度模型, 以提高簇相似度模型和社团探测算法的精度.

[1] Shen Y, Xu H L 2010 *Acta Phys. Sin.* **59** 6022 (in Chinese) [沈毅, 徐焕良 2010 物理学报 **59** 6022]

[2] Girvan M, Newman M E J 2002 *Proc. Natl. Acad. Sci. USA* **99** 7821

[3] Newman M E J 2004 *Phys. Rev.* E **69** 066133

[4] Wang G X, Shen Y 2010 *Acta Phys. Sin.* **59** 842 (in Chinese) [王高峡, 沈铁 2010 物理学报 **59** 842]

[5] Newman M E J 2006 *Phys. Rev.* E **74** 036104

[6] Shao P, Jiang G P 2011 *Acta Phys. Sin.* **60** 078902 (in Chinese) [邵裴, 蒋国平 2011 物理学报 **60** 078902]

[7] Kernighan W, Lin S 1970 *Bell. Syst. Tech. J.* **49** 291

[8] Raghavan U N, Albert R, Kumara S 2007 *Phys. Rev.* E **76** 036106

[9] Luo Z G, Ding F, Jiang X Z, Shi J L 2011 *Guofang Keji Daxue Xuebao* **33** 47 (in Chinese) [骆志刚, 丁凡, 蒋晓舟, 石金龙 2011 国防科技大学学报 **33** 47]
[10] Ma X, Gao L 2011 *J. Stat. Mech.-Theory Exp.* **5** P05012
[11] Rosvall M, Bergstrom C T 2007 *Proc. Natl. Acad. Sci. USA* **104** 7327
[12] Stanoev A, Smilkov D, Kocarev L 2011 *Phys. Rev.* E **84** 046102
[13] Reichardt J, Bornholdt S 2006 *Phys. Rev.* E **74** 016110
[14] Li W, Yang J Y, Hadden W C 2009 *Europhys. Lett.* **88** 68007
[15] Pan Y, Li D H, Liu J G, Liang J Z 2010 *Physica* A **389** 2849
[16] Palla G, Derenyi I, Farkas I, Vicsek T 2005 *Nature* **435** 814
[17] Sun P G, Gao L, Han S S 2011 *Inform. Sciences* **181** 1060
[18] Lancichinetti A, Fortunato S, Kertész J 2009 *New J. Phys.* **11** 033015
[19] Newman M E J, Leicht E A 2007 *Proc. Natl. Acad. Sci. USA* **104** 9564
[20] Ahn Y Y, Bagrow J P, Lehmann S 2010 *Nature* **466** 761
[21] Mucha P J 2010 *Science* **328** 876
[22] Newman M E J 2004 *Eur. Phys. J.* B **38** 321
[23] Ahlgren P, Jarneving B, Rousseau R 2003 *J. Am. Soc. Inf. Sci. Tech.* **54** 550
[24] Bhattacharyya A 1946 *SANKHYA* **7** 401
[25] Egghe L, Rousseau R 2006 *Inform. Process. Manag.* **42** 106
[26] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E, Dawson S M 2003 *Behav. Ecol. Sociobiol.* **54** 396
[27] Knuth D E 1993 *The Stanford Graph Base*: *A Platform for Combinatorial Computing* (1st Edn.) (New Jersey: Addison-Wesley Professional) p4
[28] Clauset A, Newman M E J, Moore C 2004 *Phys. Rev.* E **70** 066111
[29] Newman M E J 2006 *Proc. Natl. Acad. Sci. USA* **103** 8577

Group similarity based algorithm for network community structure detection*

Yuan Chao Chai Yi†

(*Institute of Automation, Chongqing University, Chongqing* 400030*, China*)

(Received 11 March 2012; revised manuscript received 3 May 2012)

Abstract

Community structure has an important influence on the structural and dynamic characteristics of the complex system. In the present study, a group similarity model is proposed for the measurement of similarity between two communities. So it can help us understand the mechanism of inter action between these communities. Moreover, based on this model, a hierarchical clustering based algorithm for network community structure detection is put forward. By this algorithm, one pair of communities with the largest similarity is merged in each iteration. And then an evaluation function is adopted for choosing the optimal partition. The algorithm gives a higher performance than many state-of-the-art community detection algorithms when tested on a series of real-world and synthetic networks. Especially, it performs better when the edge density of the network is high.

Keywords: complex network, community structure, similarity, clustering

PACS: 89.75.Fb, 89.75.Hc, 02.10.Ox

* Project supported by the National Natural Science Foundation of China (Grant No. 60974090) and the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 200806110016).

† E-mail: cqchaiyi@yahoo.com