计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900170

基于耦合强度的多项式时间社团探测算法

杨卓璇[1]　马源培[1]　严　冠[2]
1 华北电力大学经济与管理学院　北京 102206
2 中央财经大学管理科学与工程学院　北京 100081
(yangzhuoxuan@ncepu.cn)

摘　要　在资本市场中,根据交易者联系的密切程度,可以划分出众多团体,从而产生特定的社团结构。社团结构探测是一项非常重要而具有挑战性的工作,已经引起来自不同领域学者的广泛关注。然而,极少有多项式时间算法能够快速、准确地探测社团结构。基于著名的模块化设计优化理论,用新颖的 k 强度关系代表两个节点之间的耦合距离这一想法随之产生。社团结构探测算法使用基于 k 强度矩阵的广义模块度测量。为了得到最优社团数量,一种新颖的无参数结构得以使用,该结构使用特定转移矩阵的特征值之差作为社团划分边界。最后,将此算法应用于基准网络和实际网络,以评估其有效性。理论分析和实证结果表明,该算法可以快速、准确地探测社团,且易于扩展至大型实际网络。

关键词:社团结构;社会网络;耦合距离;k 强度关系;最优社团数量;多项式时间
中图法分类号　TP391

Polynomial Time Community Detection Algorithm Based on Coupling Strength

YANG Zhuo-xuan[1], MA Yuan-pei[1] and YAN Guan[2]
1 School of Economics and Management, North China Electric Power University, Beijing 102206, China
2 School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100081, China

Abstract　In capital markets, groups can be divided according to how closely traders are connected, resulting in specific community structures. Community structure detection is one of the most interesting issues in the study of social networks. However, there are few polynomial time algorithms that can detect the community structure quickly and accurately. Inspired by the famous theory of modularity design optimization, in this paper, the idea of using a novel k-strength relationship to represent the coupling distance between two nodes is proposed. Community structure detection algorithm is presented using a generalized modularity measure based on the k-strength matrix. To obtain the optimal number of communities, a new parameter-free structure is adopted, which uses the difference of eigenvalues of specific transition matrix as the boundary of community classification. Finally, the algorithm is applied on both benchmark network and real network. Theoretical analysis and experiments show that the algorithm can detect communities quickly and accurately, and is easy to be extended to large scale real networks.

Keywords　Community structure, Social network, Coupling distance, k-strength relationship, Optimal number of communities, Polynomial time

1　概述

社团结构探测[1-3]是复杂网络研究中非常重要的部分,已经吸引了来自不同学科的科学家的广泛关注。众所周知,社会网络已被广泛应用于社会中的各个领域,诸如营销、生产、运作、理财、公司治理等各个方面,是每一社会组织正常运行的必要因素。以财务学的知识举例,在同一社会网络中,各个公司之间存在着疏密不等、多少不均的联系,这些联系使得公司之间可以相互帮助、取长补短。通过公司之间的信息交流,有些公司可以实现降低运营成本的目的,有些公司可以实现提高盈利能力的目标,有些则可以获得降低资本成本的效果,也有些可以达到上市、合并分立的要求,从而构成更加密切的团体。直观上来讲,社团是网络中的一些子集,子集内部的边密度要比子集间边密度大。由于社团内部的节点更有可能具有相同的属性而形成功能团,因此社团探测研究在实际社会网络中起着重要的作用。社会网络中的社团探测方法与图论[4-5]中的图分割类似。例如,在并行计算中,通信模式可以用图或网络表示,其中点表示过程,边表示两个过程之间的相互通信关系。问题在于如何把过程分配给中央处理器,以使每个过程上的载荷大致平衡,同时还要使得边的数量最少,且过程内部的通信数量最大。总体而言,如何找到一个完美的解决分割问题的方案属于 NP 问题,因此要将之用于解决大型网络问题尤为困难。受此启发,许多启发式算法已经为众多案例提供了可接受方案,其中最著名的是在稀疏图上的

基金项目:国家自然科学基金(71401233);北京市自然科学基金(9182015)
This work was supported by the National Natural Science Foundation of China (71401233) and National Natural Science Foundation of Beijing (9182015).
通信作者:严冠 (ygcufe@126.com)

Kernighan-Lin 算法，其时间复杂性为 $O(n^3)$。

近年来许多研究都提出了一些社团探测的算法[6-8]，其中一些基于网络参数设计，例如图矩阵的特征向量、最优化模块度、相关性聚类等；还有一些算法基于网络的动态特征设计，例如随机漫步和传播机制等。然而，这些算法只解决某些特定案例。例如，根据 Laplace 矩阵的次小特征向量，一个网络被划分为两个社团，然而次小特征向量并不适用于超过两个社团的网络。根据最优化模块度算法[9]，网络被划分成多个社团，不幸的是，最优模块度 Q 被证明是 NP 完全的[10]。这意味着即使存在多种启发性算法，也并非所有的社团探测都依赖于计算 Q 值。随机漫步[12-13]的每个节点都是一位随机游走者，它每次随机地选择下一个节点，因此可能到达任何一个其他节点，由此得到一个系统树图，利用最大模块度 Q 进行社团探测。但是，要得出随机漫步的最优时间是困难的。信号传递[14]把节点间的拓扑关系转换成 n 维欧氏空间的几何关系，即使 F 统计量显著，但如何选择一个合适的 p 并把所有节点聚合成 p 类仍旧是困难的。其他方法则依赖于动态社会网络[15-16]中的社团概率和模块度值[7-8]来探测社团。

本文以经济生活中广泛存在的现象为切入点，从人们熟悉的证券交易市场开始，深入思考社团结构的作用和深层意义。由于极少有多项式时间算法能精确地探测社区，一些有价值的研究转而专注于如何获得更低的计算复杂度而非更高的精确度。为了得到快速、精确的赋权网络社团探测算法，一个新颖的定义被提出，即 k 强度关系，它代表两个节点之间的耦合距离。社团结构探测算法使用基于 k 强度矩阵对不同类型社会网络的广义模块度进行测量。此外，为了获得最优社团数量，基于特定的转移矩阵特征值之差，一种新的无参数算法被提出。最后，将该算法应用于基准网络和实际网络，以评估其有效性。理论分析和实证结果表明，该算法可以快速、准确地探测社团，且易于扩展至大型实际网络。

本文第 2 节介绍了一些基本定义，如 k 强度关系及广义模块度测量；第 3 节介绍框架的细节，包括算法的具体过程和计算复杂度分析；第 4 节描述了一种新颖的方法，以确定最优自然社团数量；第 5 节给出一些基准网络和实际网络中的代表性实验，以验证算法的有效性和效率；最后总结全文。

2 定义

在许多社会关系（如经济系统中）经纪人相互影响：急于买进或卖出某一特定资产可以促使其他人做相同的事情。正如在资本市场当中，参与股票交易的人们也类似地分为买方和卖方两大类，先知先觉的买入交易者对股票价格产生拉升的动力和趋势，向股票市场中传递积极信号，从而吸引大量后知后觉的交易者做出相同的买入行为；同样地，先知先觉的卖出交易者对股票价格产生拉低的动力和趋势，向股票市场中传递消极信号，从而吸引大量后知后觉的交易者做出相同的卖出行为。在大多数情况下，这些经纪人和交易者只影响与他们有直接关联的人。股票的交易永远只存在于股票市场中，其中的每一名交易者都会与其他的交易者产生或强或弱、或多或少的联系，而与外部事物关联较弱，从而构成一个与外部近了乎隔绝的团体。也就是说，所有的买方和卖方拥有固定的活动范围，形成一个不可分割的结构，他们和结构外部的人几乎没有交流。这种结构被称为社会网络中的社团。网络中的节点表示交易者或经纪人，边的权重表示两个交易者或经纪人相互影响的程度。给定一个网络 $G=(V,E)$，其中包含 n 个节点与 m 条边，G 是一个无向简单图。G 的邻接矩阵是 $N\times N$ 的 0-1 矩阵，表示为 $A=(a_{ij})_{n\times n}$。如果节点 i 与节点 j 连通，$a_{ij}=1$，否则 $a_{ij}=0$。如果网络是赋权图，则 $\omega_{i,j}$ 表示每条边的权，$W=(\omega_{i,j})_{N\times N}$ 表示权重矩阵。给定正整数 k，路径 k 表示在节点数为 $k+1$ 的无环网络中节点 i 到 j 的第 k 条路径。$A^k=(a_{ij}^k)_{n\times n}$，其中 $a_{ij}^k=\sum_{l=1}^{N}a_{il}^{k-1}\times a_{lj}$ 表示从节点 i 到 j $(i\neq j)$的路径数目，当 $i=j$ 时 $a_{ij=0}$。$S^k=(s_{ij}^k)_{N\times N}$ 被定义为 G 的 k^{th} 强度矩阵，其递归定义如下：

$k=0$ 时；

$$S^0=A \tag{1}$$

$k=1$ 时；

$$S^1=(\omega_{ij})_{N\times N} \tag{2}$$

$k\geqslant 2$ 时；

$$S^k=(s_{i,j}^k)_{N\times N},s_{i,j}^k=\sum_{s=1}^{a_{i,j}^k}\frac{1}{k}\sum_{l=1}^{k}\omega_{i_{l-1}^s i_l^s} \tag{3}$$

其中，$i=i_0^s,i_1^s,\cdots,i_{k-1}^s,i_k^s=j$ 是 $s=1,2,\cdots,a_{i,j}^k$ 的 k 路径。对于给定的网络，$S^0=A$ 始终成立。矩阵 A^k 包含了所有 k 路径，$S^{i,j}$ 是递增多项式。因此，$S_{i,j}^k$ 的值可以被准确计算。

每个 k 强度矩阵表示一个 k 强度关系 $R_k=\{(i,j,s_{i,j})\mid s_{i,j}=\sum_{l=1}^{k}s_{i,j}^l\}$。也就是说，$R_k$ 中的 $s_{i,j}$ 是 $S=S^1+S^2+\cdots S^k=(\sum_{l=1}^{k}s_{i,j}^l=s_{i,j})_{N\times N}$ 中的元素。其中，S 表示 G 的 k 强度矩阵，G 是一个 k 强度关系网络并涉及到平均场理论，每个节点都知道其他所有节点的信息（节点的权），这在许多实际系统中是合理的。例如，我国上海证券交易所的交易员可能被同一层楼的其他人影响，也可能被我国深圳证券交易所的交易员的交易模式影响，更甚至，他们也可能被伦敦或巴黎的交易模式影响。由此，一些成熟的交易行为模式在整个经济系统中形成。在社交网络中，人们用强度关系衡量友谊也是很常见的。例如，在熟人网络中，关系指关联的紧密程度，关系值越高，关联越紧密，交流也就越多。框架中另一个有用的定义是最小 q 分割，即分割边界取决于权值最小和。

这里 q 是正整数，对于集合 $\{C_1,C_2,\cdots,C_q\}$，总有 $|C_i|=k_i$，且 $\bigcup_{i=1}^{q}C_i\subseteq V(G)$ 是 G 的顶点子集。删除所有 C_i 后的 G 是不连通的，且删除后的 G 中连接权值达到最小。当 $\bigcup_{i=1}^{q}C_i\subseteq V(G)$ 成立时，最小分割是 G 的子集。

Guttmann 设计了一种算法来探测完全图的最小分割[17]，基于他的算法，社团用强度关系来探测。框架基于 Newman 最先提出的最优化模块度算法，被应用于 G 的强度矩阵。假定 G 中有 $q(q\leqslant N/2)$ 个社团（怎样确定 q 值是一项艰巨的任务，这个问题将由第 4 节提出的一个框架解决），表示为 $C=\{C_1,C_2,\cdots,C_q\}$，强度矩阵中的广义模块度 Q 则被定义为：

$$Q=\max_{q}\sum_{i=1}^{q}(c_{i,i}-c_i^2) \tag{4}$$

其中，$c_i=\sum_j s_{i,j}$，$\Delta=\sum_{i,j}s_{i,j}$。如果节点 i 和节点 j 属于同一个社团，那么 $\delta_{i,j}=1$；否则 $\delta_{i,j}=0$。$c_{i,i}$ 表示强度关系的两个节点都在 C_i 中，c_i 表示只有一个节点在 C_i 中。如果网络不含权（二值网络），那么 Q 等于 *Newmann* 的模块度。可以发现，两个

节点之间直接与间接的信息都可以应用于此框架。当社会网络中的两个节点交换信息时，消息链将在同一个社团中形成。因此，用强度描述两个节点之间的关系更加合理，社团中相同信息的关联更加紧密，传播也更加迅速。这意味着相比稀疏的社团，关联紧密的社团具有更快的信息传播速率和谣言传播率，因为路径越多，传播速率越快。通俗地讲，仍以股票交易为例，相互之间联系越紧密的交易者们所形成的团体具有更快的信息传递速度，每时每刻发生的预期对股票市场产生影响的新闻会更快地在这些人之间扩散开，使其更快地做出买入或者卖出的行为，这正是因为他们联系密切，信息的传递路径多，传递速度快。

3 算法框架

3.1 确定 k 强度关系矩阵

如上所述，k 强度关系是整个框架的基础。本部分主要确定 k 强度关系矩阵。下面的定理不仅提供了计算所有元素的过程，也揭示了重要的时间复杂性信息。

定理 1 k 强度关系矩阵可在多项式时间内得到。

证明：假设网络 G 的邻接矩阵为 A，节点 i 到 j 的 k 长度路径数量为 $a_{i,j}^k$，则 $A^k=A^{k-1}\times A=(a_{i,j}^k)$[16]，其中 k 长度路径表示路径的长度为 k。用 $\{i_0,i_1,\cdots,i_{k-1},i_k\}$ 表示有 $k+1$ 个节点的 k 路径集，对所有的 s 和 j，都有 $i_s\neq i_j$，即路径中没有闭环。

为了得到 k 强度关系矩阵，研究方法在权值矩阵 G 上定义运算 $\oplus$：$W^k=W^{k-1}\oplus W=(\omega_{i,j}^k)_{N\times N}$。其中，$\omega_{i,j}^k$ 被定义为：如果 $\sum_{l=1}^{N}(\omega_{i,l}^{k-1}\times\omega_{l,j}\neq 0)$，则节点 i 和节点 j 之间有 k 个连接。也就是说，多项式 $\sum_{l=1}^{N}(\omega_{i,l}^{k-1}\times\omega_{l,j}\neq 0)$ 中至少包含一项。非广义地讲，假定有 h 项不为 0，$\omega_{i,l^1}^{k-1}\times\omega_{l^1,j}\neq 0$，$\omega_{i,l^2}^{k-1}\times\omega_{l^2,j}\neq 0$，$\omega_{i,l^h}^{k-1}\times\omega_{l^h,j}\neq 0$，那么 $\omega_{i,j}^k=\sum_{s=1}^{h}(\omega_{i,l^s}^{k-1}+\omega_{l^s,j})$。反之，如果节点 i 与 j 之间没有连接，则 $\omega_{i,j}^k=0$。

$\omega_{i,j}^k$ 是节点 i 到 j 的 k 路径的权重之和。不难得到 $s_{i,j}^k=\omega_{i,j}^k/k$，即 $s_{i,j}^k=\sum_{s=1}^{a_{i,j}^k}\frac{1}{k}\sum_{l=1}^{k}\omega_{i_{l-1}^s,i_l^s}^s=\omega_{i,j}^k/k$。

一旦 $\sum_{l=1}^{N}(\omega_{i,l}^{k-1}\times\omega_{l,j})$ 确定下来，所有的 k 路径都可以确定。如果 $\omega_{i,l^1}^{k-1}\times\omega_{l^1,j}\neq 0$，$\omega_{i,l^2}^{k-1}\times\omega_{l^2,j}\neq 0$，$\omega_{i,l^h}^{k-1}\times\omega_{l^h,j}\neq 0$ 对所有 $k\geqslant 2$ 的正整数都成立，则用 $P_{i,j}^k$ 表示从节点 i 到节点 j 的一条 k 路径。容易得到从节点 i 到节点 j 有 $a_{i,j}^k$ 条 k 路径，于是得到 $P_{i,j}^k=\{P_{i,l^1}^{k-1}\vee(l^1,j),P_{i,l^2}^{k-1}\vee(l^2,j),\cdots,P_{i,l^h}^{k-1}\vee(l^h,j)\}$，其中 $P_{i,l}^{k-1}\vee(l,j)$ 指由 $P_{i,l}^{k-1}$ 中 $k-1$ 路径形成的所有 k 路径。

最后，归纳出所有的 k 路径。也就是说，$\oplus$ 是一种多项式时间算法。强度矩阵由计算权值矩阵得到，其时间复杂性为 $O(n^2m)$，其中每对 N 列矩阵最多需要计算 $N\times N$ 个数，一共有 m 对矩阵。对于给定的 k，结果为 $S^k=(s_{i,j}^k)N\times N$，由此列出节点 i 到 j 的所有长度为 k 的路径，其中 $i=i_0^s,i_1^s,\cdots,i_{k-1}^s,i_k^s=j(s=1,2,\cdots,a_{i,j}^k)$。

证明完毕。

3.2 社团探测算法

G 的最小 q 分割 $\widetilde{E}$ 是删除边集后权重和最小的边集，$G-\widetilde{E}$ 是一个孤立的图。一种直接确定分区的方法是调查剩下的图 $C=\{C_1,C_2,\cdots,C_c\}$ 中的所有元素。根据最大流和最小割原理，需要选择元素以使 $\sum_{i<j}\omega(C_i,C_j)$ 最小，或者 $\sum_{i=1}^{c}\omega(C_i,C_i)$ 最大。然而，最小 q 分割的问题在于其 NP 完全，且难以找到适应的多项式算法。幸运的是，Guttmann-Beck 和 Hassin 设计了一个完全图算法，证明了最优近似解不到三次方[19]。基于这个绝妙的主意，得到算法 1 的详细过程。

3.3 算法复杂性

对于给定正数 q，0-1 运输问题的时间复杂性为 $O(N)$，因此子集个数为 $C_N^qO(qN)$ 的 0-1 运输问题的时间复杂性为 $O((q+1)N)$。以下两个重要的声明对分析极为有用。

条件 1 当且仅当 $\{C_1,C_2,\cdots,C_q\}$ 是 $\widetilde{G}$ 的子集。

证明：这一说法很容易得到验证，因为 $\widetilde{G}$ 和 G 有相同的顶点集。

条件 2 设 $\{C_1,C_2,\cdots,C_q\}$ 是 $\widetilde{G}$ 的子集，则 $\sum_{i<j,C_i,C_j\subset\widetilde{G}}^{q}\omega(C_i,C_j)$ 达到最小。那么存在 G 的最小子集 $\{\overline{C}_1,\cdots,\overline{C}_q\}$ 也是 $\widetilde{G}$ 的最小子集，$\sum_{i<j,\overline{C}_i,\overline{C}_j\subset\widetilde{G}}^{q}\omega(\overline{C}_i,\overline{C}_j)$ 达到最小。

证明：根据 $\widetilde{G}$ 的定义，$\Delta=\sum_{i<j,C_i,C_j\subset\widetilde{G}}^{q}\omega(C_i,C_j)$，其中 $\{C_1,C_2,\cdots,C_q\}$ 是 $\widetilde{G}$ 的最小子集，$\omega(C_i,C_j)=\sum_{i\in C_i,j\in C_j}s_{i,j}^1+\sum_{i\in C_i,j\in C_j}\sum_{k=2}s_{i,j}^k$，且 Δ 的值不变。根据 k 强度关系的定义，已知 $\sum_{i\in C_i\subset\widetilde{G},j\in C_j\subset\widetilde{G}}s_{i,j}^1=\sum_{i\in C_i\subset G,j\in C_j\subset G}s_{i,j}^1$，用算法 1 构造一个最小子集 $\{\overline{C}_1,\cdots,\overline{C}_q\}$ 使得 $\sum_{i\in\overline{C}_i\subset G,j\in\overline{C}_j\subset\widetilde{G}}s_{i,j}^1$ 最小。

算法 1 社团探测算法

输入：社会网络 G=(V,E)

输出：最大模块度下的最小 q 分割。

1. 把深度为 1 的节点与其相邻节点合并，直到 G 中不存在深度为 1 的节点。这个操作并不影响社团探测，因为深度为 1 的节点只能与相邻节点合并，因此仍和相邻节点处在同一个社团。此处仍把网络写作 G。
2. 根据定理 2(定理 2 将在下一部分介绍)确定最优社团数量 q。
3. 对于固定值 q，最小 q 分割问题的多项式时间复杂性为 $O(|V|^{q^2})$[17]。假定 $\widetilde{G}=(V,\widetilde{E})$ 的一个分割为 $C=(C_1,C_2,\cdots,C_q)$，其中 $|C_i|=k_i$，$\sum_{i=1}^{q}k_i=N$，$C_i\cap C_j=\emptyset$。接下来将探测 $\widetilde{G}$ 的最小 q 分割，并进行证明。令 $v_i\in C_i(i=1,2,\cdots,q)$，$x_{i,j}=\begin{cases}1, & u_j\in C_i\\ 0, & \text{otherwise}\end{cases}$。

Begin

For $\{v_1,v_2,\cdots,v_q\}\subset V,v_i\in C_i$

For $u_j\in V-\{v_1,v_2,\cdots,v_q\}$

//下面的转移问题是最优的

$$\min:\sum_{i=1}^{q}\sum_{j=1}^{N-q}\omega(C_i,u_j)(1-x_{i,j})$$

$$\text{Subject to}\begin{cases}\sum_{j=1}^{N-q}x_{i,j}=k_i-1,i=1,2,\cdots,q\\ \sum_{i=1}^{q}x_{i,j}=1,j=1,2,\cdots,N-q\\ x_{i,j}\in\{0,1\},i=1,2,\cdots,q;j=1,2,\cdots,N-q\end{cases}$$

End.

$C_i=C_i\cup\{u_j\mid x_{i,j}^*=1,1\leqslant j\leqslant N-q\},1\leqslant i\leqslant q$

End

Back to begin.

4. 输出结果$\{C_1,C_2,\cdots,C_q\}$是$\widetilde{G}$的最小 q 分割集，$\{\overline{C}_1,\overline{C}_2,\cdots,\overline{C}_q\}$是$\widetilde{G}$的最小 q 分割，其中$v_i\in C_i$。

证明完毕。

4 确定社团数量

不难发现，如果一条信息掉进一个真实的社团，它在关联紧密的社团里停留的时间更长。仍用财务学的相关知识进行举例，在诸如股票市场等证券交易市场中，联系越密切、关系越亲近的交易者们，互相之间分享信息的速度越快，因此信息得以在全部交易者即整个团体中扩散、流传开，因此，这一信息会更久地影响这一社团；相反，当信息出现在联系不密切的交易者中间时，信息往往不会被快速地扩散开，因此在这一群交易者中停留的时间更短，对这一社团做出买入或者卖出举动的影响力也就越低。根据随机理论，马尔可夫过程的光谱性质自然地揭示了一个特定的“稳定”分区。受此启发，一种新颖的确定社会网络中最优社团数量的方法被提出。

定理 2 令 P 是 G 的广义转移矩阵，则最优社团数量为：

$$opt=\arg\min_{z}\left(\frac{\log|\lambda_{z-1}|}{\log|\lambda_z|}\right)$$

其中，λ_z 是 G 的第 z 特征值。

证明：转移概率表示某个节点将信息或疾病传播给其他节点的能力，它与连接的紧密程度正相关。转移概率矩阵 $P=(p_{i,j})$ 被定义为：

$$p_{i,j}=\frac{r_{i,j}}{\sum_{j=1}^{N}r_{i,j}} \tag{5}$$

$r_{i,j}=\langle s_{i,j}\rangle_k$ 是所有 k 强度关系值的平均。通过这种表示，框架可被用于社团探测。令 P 为转移概率矩阵，从而得到：

$$P=D_C^{-1}C \tag{6}$$

D_C 是 $R=(r_{i,j})$ 的对角阵。令 $p_{i,j}^{(\tau)}$ 表示从节点 i 开始经过 τ 个步骤后到达节点 j 的概率，得到：

$$p_{i,j}^{(\tau)}=(P^\tau)_{i,j} \tag{7}$$

对于这个 Markov 遍历过程，P^τ 对应于在 τ 时间内状态转移的概率矩阵。为了计算转移矩阵 P^τ，需要对 P 进行特征值分解。如果 $\lambda_k(k=0,\cdots,N-1)$ 是 P 的特征值，扩展它的左右特征值 f_k 和 h_k 以满足标准正交性[21]：

$$f_k h_l=\delta_{kl} \tag{8}$$

P 的频谱图表示如下：

$$P=\sum_k\lambda_k f_k h_k \tag{9}$$

于是：

$$P^\tau=\sum_k\lambda_k^\tau f_k h_k \tag{10}$$

假设 P 的特征值按照 $\lambda_0=1>|\lambda_1|\geqslant|\lambda_2|\geqslant\cdots\geqslant|\lambda_{N-1}|$ 排列。每个初始分布的收敛稳态分布 $P^{(0)}$ 对应于这样一个事实：随着时间的推移，整个系统最终达到完全相同的自旋值。在时间 $\tau\to\infty$ 时，最大特征值 $\lambda_0^\tau=1$，其余特征值 λ_k^τ 随时间趋近于 0。在极端情况 $\tau=0$ 时，P^τ 变成单位矩阵。它的所有列都是不同的，系统分裂成和元素个数一样数目的社团。

进行社团识别时，重要的是中间时间标度，但并非所有的特征值都为 0。如果想要识别 z 个社团，需要在时间维度上找到 P^τ，特征向量 λ_k^τ 只有在 $k=0,\cdots,z-1$ 时不为 0。这是通过确定 τ 以使 $|\lambda_z|^\tau\approx0$ 得到的。使用一个趋近于 0 的变量 $\zeta\ll1$，要求 $|\lambda_z|^\tau=\zeta$，整个系统进入 z 个社团的亚稳定状态需要的时间为：

$$\tau(z)=\frac{\log\zeta}{\log|\lambda_z|} \tag{11}$$

较小特征值在给定时间标度上的坍缩描述了不同自旋状态和较弱特征向量的结构特征的消失。接下来定义第 z 个社团的稳态结构为 N_z，状态 z 的进入时间 $\tau(z)$ 和退出时间 $\tau(z-1)$ 之间的比率为：

$$N_z=\frac{\log\zeta/\log|\lambda_z|}{\log\zeta/\log|\lambda_{z-1}|}=\frac{\log|\lambda_{z-1}|}{\log|\lambda_z|} \tag{12}$$

由于 $\log\zeta/\log|\lambda_z|\leqslant\log\zeta/\log|\lambda_{z-1}|$，不难发现更小的 N_z 表明更优的社团结构。在实际网络中，可以用最小的 N_z 来估计社团数量 opt。在给定网络中：

$$opt=\arg\min_{z}(N_z) \tag{13}$$

在 N_z 最小的条件下，arg 表示最优的 z。

证明完毕。

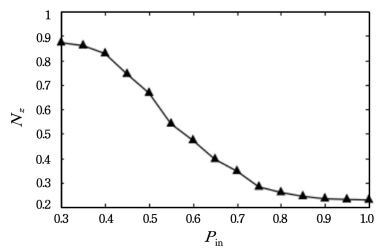
5 实验

接下来将测试所提出算法的性能。为此设计并实现了 3 个实验，主要有两个目的：1）评估该算法的划分精度；2）将其应用到实际的大型网络。

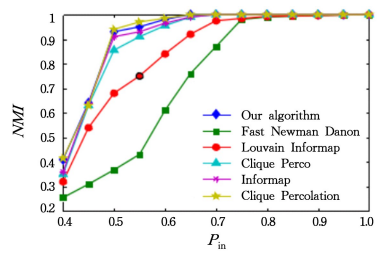
5.1 人工网络

为了验证算法的有效性，将其与其他 5 个著名的算法在人工网络上进行比较。这些算法包括：Newmann 快速算法[1]、Danon 方法[22]、Louvain 算法[23]、Informap 方法[24]和 Clique Percolation 算法[25]。广泛使用的 Ad-Hoc 网络模型可以构建一个随机合成网络，该网络包含 4 个预定义社团，每个社团有 32 个节点。节点的平均度为 16，社团内部边的比例用 P_{in} 表示。随着 P_{in} 的下降，Ad-Hoc 网络的社团结构变得越来越模糊，相应地，N_4 值从 0 上升至 1，如图 1(a)所示。

每个算法的优劣用归一化互信息测度(NMI)[27]进行评估，以验证网络社团的内在尺度能否被正确发现。实验结果如图 1(b)所示，y 轴表示归一化互信息测度值，曲线上各点都是 50 个 Ad-Hoc 网络的平均值。



(a) N_z($z=4$)随 P_{in} 变化的情况



(b) 6 种算法精确度的比较

图 1 不同算法性能的比较

Fig. 1 Comparison of different algorithms performance

如图所示，当 $P_{in}>0.7$ 时，所有算法都是有效的，NMI 值都大于 0.85。与其他 5 种算法相比，所提算法性能最优。只有在 $0.5\leqslant P_{in}\leqslant 0.65$ 时，其精度略低于 Clique Percolation 算法。但是 Clique Percolation 算法的时间复杂性超过 $O(n^3)$，其耗时几乎与广度优先搜索(BFS)算法相当。相比之下，所提算法的时间复杂性非常低，仅为 $O(n^2)$，且易于在各种网络中实现。

5.2 美国大学足球队网络

由于具有非常清晰的自然社团结构，美国大学足球队网络作为一个基准例子已经得到了广泛的应用。其中的数据[1]是使用 Girvan 和 Newman 收集的，它是 2000 赛季美国第一等级美式足球比赛的赛程表。网络中的节点代表 115 支球队，边代表 613 场比赛。整个网络可以自然地划分为 12 个组。相对于不同组球队之间，同一个组的球队之间的比赛更加频繁。

首先计算 N_z，结果如图 2 所示，最优社团数量 $opt=12$，与实际情况一致。

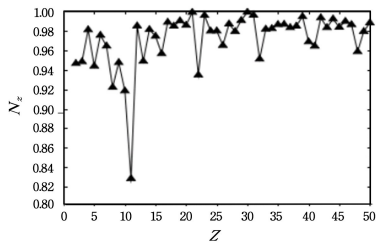
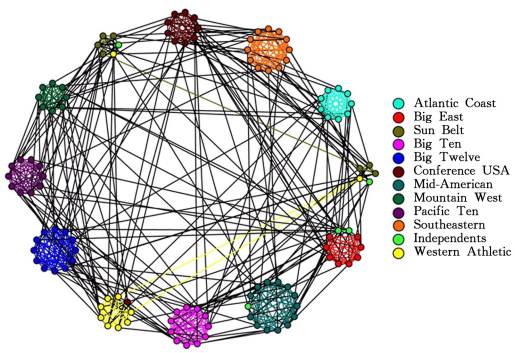


图 2 不同 z 值对应的 N_z 值

Fig. 2 N_z corresponding to different z

然后把算法应用到美国大学足球队网络，将网络划分为 12 个社团，结果如图 3 所示。算法划分的准确率超过 93%，说明划分出的社团结构与社团真实结构高度一致。实际上，基于最优模块化 Q 的算法只能探测到 11 个社团，而且由于网络的模糊性，划分的准确率也较低。由此可见，所提算法对揭露真实网络的自然特征是十分有价值的。此外，它识别了 5 个有趣的重叠节点，如图中 3 条黄色线所连接的 5 个节点所示。这些节点都位于社团之间的边界上，可以被看作一些相对独立的俱乐部。



形状相同的节点属于同一个社团，右侧列出的稠密子集是算法检测到的社团

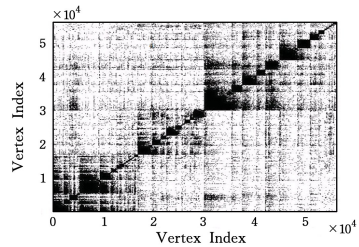
图 3 美国大学足球队网络的划分结果(电子版为彩色)

Fig. 3 Division of network of college football teams in United States

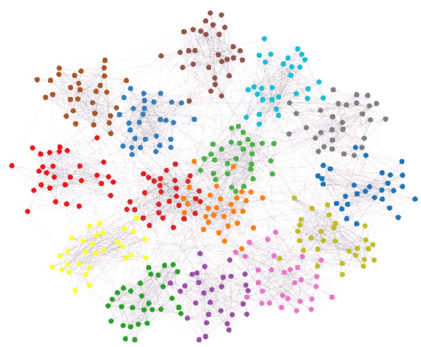
5.3 科学家合作网络

最后，将本算法在大型网络[20]上进行测试，选用的是 Girvan 和 Newman[6]使用的科学家合作网络。该网络通过研究 56276 名科学家在物理电子文档网站 arxiv.org 上发表的合著论文，揭示科学家之间的研究合作关系，网络总共包含 315810 条赋权边。为了可视化，算法输出了一个转换后的具有层次结构的邻接矩阵(即同一个社团的节点组合在一起)。从转移后的邻接矩阵(见图 4(a))，可以观察到一个有效的社团结构，其具有集体倾向的合作模式。最大的 3 个研究社团是关于 3 个研究领域而自发形成的：凝聚态物质，高能物理(包括理论、现象学和核物理学)，以及天体物理学。

科学家合作网络是典型的无标度分布，广泛存在于实际社会。总体上，此算法探测到了 694 个社团，最大的社团规模为 201，最小的为 3，平均社团规模为 77。规模前 5.7%的社团包含了 23.4%的节点，其他的社团规模都相对较小。3 个最大的社区紧密对应着学科研究的分学科，它们分别是固体物理学、超核物理和理论天体物理。此外，一个包含 15 个社团的子网络如图 4(b)所示。划分结果与参考文献[6]和文献[26]完全一致。算法在大型实际网络的有效性表明其适用于在各个领域进一步的研究。



(a)科学家合作网转换后的邻接矩阵



(b)包括 15 个社团的子图，每个社团用不同的颜色表示

图 4 科学合作网结构图

Fig. 4 Structure of scientific cooperation network

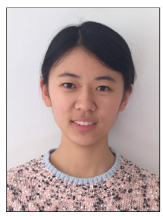
结束语 本文用 k 强度关系探测社会网络中的社团，其中 k 强度关系指两个节点之间的耦合距离。理论分析表明，所提算法的模块化时间远少于大多数已有的算法。为了得到最优社团数量，提出了无参数算法，即使用特定的转移矩阵特征值之差作为社团划分边界。最后，将算法应用于基准网络和实际网络以评估其有效性。理论分析和实证结果表明，该算法可以快速、准确地探测社团，且易于扩展至大型实际网络。

参 考 文 献

[1] NEWMAN M E J. Fast Algorithm for Detecting Community Structure in Networks[J]. Physical Review E, 2004, 69: 066133.

[2] NEWMAN M E J, GIRVAN M. Finding and Evaluating Community Structure in Networks[J]. Physical Review E, 2004, 69: 026113.

[3] NEWMAN M E J. Modularity and Community Structure in Networks[J]. Proc Natl Acad Sci,2006,103:8577-8582.
[4] GAREY M R,JOHNSON D S. Computers and Intractability:A Guide to the Theory of NP-Completeness[M]. San Francisco:Freeman,1979.
[5] SCOTT J. Social Network Analysis:A Handbook[M]. London:Sage Publications,2000.
[6] GIRVAN M,NEWMAN M E J. Community Structure in Social and Biological Networks[J]. Proc Natl Acad Sci,2002,99:7821-7826.
[7] ZHANG X S,WANG R S,WANG Y,et al. Modularity Optimization in Community Detection of Complex Networks[J]. Europhysics Letters,2009,87:38002.
[8] ZHANG X S,LI Z,WANG R S,et al. A Combinatorial Model and Algorithm for Globally Searching Community Structure in Complex Networks[J]. J. Comb. Optim.,2012,23(4):425-442.
[9] RADICCHI F,CASTELLANO C,CECCONI F,et al. Defining and Identifying Communities in Networks[J]. Proc Natl Acad Sci,2004,101(9):2658-2663.
[10] LATAPY M,PONS P. Proceedings of the 20th International Symposium on Computer and Information Sciences[J]. Lecture Notes in Computer Science,2005,3733:284-293.
[11] KERNIGHAN B W,LIN S,BELL. An Efficient Heuristic Procedure for Partitioning Graphs[J]. System Technical Journal,1970,49:291-307.
[12] LI H J,ZHANG X S. Analysis of Stability of Community Structure Across Multiple Hierarchical Levels[J]. Europhysics Letters,2013,103:58002.
[13] LI H J,WANG Y,WU L Y,et al. Potts Model Based on a Markov Process Computation Solves the Community Structure Problem Effectively[J]. Physical Review E,2012,86:016109.
[14] BRANDES U. An Algorithm to Detect Community by Geodesic Line in Social Networks[J]. International Journal on Advances in Information Sciences and Service Sciences,2006,3:0608255.
[15] HUANG L C,YEN T J,CHOU S C T. International Conference on Advances in Social Networks Analysis and Mining[C]// IEEE Computer Society. 2011:110-117.
[16] MUCHA P J. Community Structure in Time-Dependent,Multiscale,and Multiplex Networkset[J]. Science,2010(328):876-878.
[17] GUTTMANN-BECK N,HASSIN. Approximation Algorithms for Minimum K-Cut[J]. Algorithmica,2000(27):198-207.
[18] GAREY M R,JONSON D S. Computers and Intractability:A Guide to the Theory of NP-Completeness[M]. San Francisco:Freeman,1979.
[19] MAKI D P. Mathematical Models and Applications,with Emphasis on social,life,and Management sciences[M]. San Francisco:Prentice Hall College Div,1973.
[20] XIA Z Y,BU Z. Community Detection Based on a Semantic Network[J]. Knowledge-Based Systems,2012(26):30-39.
[21] W N E,LI T,VANDEN-EIJNDEN E. Optimal Partition and Effective Dynamics of Complex Networks[J]. Proc Natl Acad Sci,2008,105(23):7907-7912.
[22] DANON L,DUCH J,GUILERA D. Comparing Community Structure Identification[J]. J. Stat. Mech,2005,29:P09008.
[23] BLONDEL V D,GUILLAUME J L,LAMBIOTTE R,et al. Fast Unfolding of Communities in Large Networks[J]. J. Stat. Mech,2005,10:P10008.
[24] ROSVALL M,BERGSTROM C T. Maps of Random Walks on Complex Networks Reveal Community Structure[J]. Proc Natl Acad Sci,2008,105(4):1118-1123.
[25] PALLA G,DERENYI I,FARKAS I. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society[J]. Nature,2005,435:814-818.
[26] LI Z P,ZHANG S H,WANG R S,et al. Quantitative Function for Community Detection[J]. Physical Review E,2008,77:036109.
[27] LANCICHINETTI A,FORTUNATO S. Community Detection Algorithms:A Comparative Analysis[J]. Physical Review E,2009,80:056117.



YANG Zhuo-xuan, born in 1998, postgraduate. Her major research interests include financial management and computer science and technology.

YAN Guan, Ph.D. Her research interests include data mining and management science.