DOI:10.13306/j.1672-3813.2008.03.002
第5卷第3期 复杂系统与复杂性科学 Vol.5 No.3
2008年9月 COMPLEX SYSTEMS AND COMPLEXITY SCIENCE Sep. 2008

文章编号:1672-3813(2008)03-0019-24

复杂网络中的社团结构

李晓佳,张 鹏,狄增如,樊 瑛
(北京师范大学管理学院系统科学系,北京 100875)

摘要:对复杂网络社团结构问题进行了综述。介绍了无权无向网络中社团结构的定义、探索社团结构的算法及算法的评价标准和检验网络。重点总结与类比了具有代表性的算法及其在检验网络上得到的结果,并依据这些结果和评价标准对算法进行了评述。部分地概括了原有算法在加权无向网络中的推广方法。最后对部分社团结构算法的特点进行了横向的比较,对社团结构与网络功能的研究进行简略介绍,并对社团结构研究的发展做出展望。

关键词:复杂网络;社团结构;聚类
中图分类号:N94 文献标识码:A

Community Structure in Complex Networks

LI Xiao-jia ZHANG Peng DI Zeng-ru FAN Ying
(Department of Systems Science, School of Management, Beijing Normal University, Beijing 100875, China)

Abstract: Community structure exists widely in most of actual systems and networks. Investigation on community structure is an important way to understand both the structure and function of networks. In this paper, we review main results in the study of community structure in complex networks. Firstly, we focus on the unweighted and undirected networks. Definitions of community structure and algorithms that detect communities are introduced. Meanwhile some measurements on detecting algorithms and classical networks are listed. The emphasis of our work is evaluating algorithms using measurements and the results for classical networks. Secondly we extend study to weighted and undirected networks. Finally, the comparison of some algorithms and a brief introduction to the relationship between community structure and network function are given, and prospect of study on community structure in the future is outlined.

Key words: complex networks; community structure; clustering

1 引言

近些年来,复杂网络逐渐成为受人关注的研究领域,越来越多的科研工作者投身其中,使得新思想新成果不断涌现[1-6]。在此过程中,新的研究方向也被开拓出来,网络结构的社团划分就是其中之一。通过近几年的发展,在这一方向上已经积累了一些优秀的思想和成果,有必要对其进行全面系统的整理,为今后的发展提供帮助和借鉴。

收稿日期:2008-01-23
基金项目:国家自然科学基金(70771011)
作者简介:李晓佳(1984-),女,江苏人,硕士研究生,主要研究方向为系统工程及复杂网络。

网络由大量顶点 (或称为节点)和连接顶点的边组成。许多实际系统都可以从网络的角度进行刻画, 例如: Internet网、万维网[7—8]、食物链[9]、人际交往关系[10—11]等等。用网络对实际系统进行抽象时, 系统中的个体对应网络中的顶点, 个体间的相互关系对应网络中的边, 从而这些系统都可以表现为由点和边构成的图。这种抽象过滤了系统纷繁复杂的背景信息, 只保留了系统的基本结构, 有助于对系统内在共同特征和性质的研究。

近年来对众多实际网络的研究发现, 它们存在一个共同的特征, 称之为网络中的社团结构。它是指网络中的顶点可以分成组, 组内顶点间的连接比较稠密, 组间顶点的连接比较稀疏[12], 如图 1。社团结构在实际系统中有着重要的意义: 在人际关系网中, 社团可能基于人的职业、年龄等因素形成; 在引文网[13]中, 不同社团可能代表了不同的研究领域; 在万维网中, 不同社团可能表示了不同主题的主页[14—16]; 在新陈代谢网、神经网中, 社团可能反映了功能单位; 在食物链网中, 社团可能反映了生态系统中的子系统。在网络性质和功能的研究中, 社团结构也有显著的表现。例如: 在网络动力学的研究中, 当外加能量处于较低水平时同一社团的个体就能达到同步状态[17]; 在网络演化的研究中, 相同社团内的个体可能最终连接在一起。总之, 对网络中社团结构的研究是了解整个网络结构和功能的重要途径。

本文着重关注网络社团结构研究中以下 4方面问题: 1)社团结构的定义; 2)探索社团结构的算法; 3)划分方法的检验与评价; 4)社团结构及划分算法在加权网上的推广。

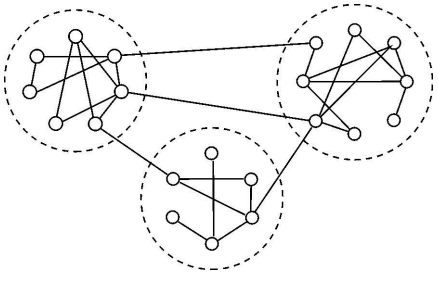


图 1 社团结构网络示意图[22]

在这里不过多涉及社团结构与网络功能之间的关系。因此本文的基本结构为: 第 2节介绍社团结构的各种定义, 以及社团结构的描述方法。第 3节在概括介绍一种社团结构划分思路的基础上, 讨论社团结构划分方法的评价标准, 包括算法复杂度及一些作为检验工具的人造网和实际网, 并进一步介绍划分结果的比较方法。第 4节具体介绍各种社团结构的划分算法, 及应用于经典网络的划分结果。第 5节将上述讨论推广到加权网, 展现权重对网络社团结构划分的影响。最后对社团结构划分问题进行总结与展望。

2 社团结构的定义与模块化函数

2 1 社团结构的定义

关于网络中的社团结构目前还没有被广泛认可的唯一的定义, 较为常用的是基于相对连接频数的定义: 网络中的顶点可以分成组, 组内连接稠密而组间连接稀疏[12, 22]。这一定义中提到的“稠密”和“稀疏”都没有明确的判断标准, 所以在探索网络社团结构的过程中不便使用。因此人们试图给出一些定量化的定义, 如提出了强社团和弱社团的定义。强社团的定义[18—19]为: 子图 V中任何一个顶点与 V内部顶点连接的度大于其与 V外部顶点连接的度。弱社团的定义[18—19]为: 子图 V中所有顶点与 V内部顶点的度之和大于 V中所有顶点与 V外部顶点连接的度之和。此外, 还有比强社团更为严格的社团定义 —— LS集[24], 一个 LS集是一个由顶点构成的集合, 它的任何真子集与该集合内部的连边都比与该集和外部的连边多。

另一类定义则是以连通性为标准定义社团, 称之为派系[21]。一个派系是指由 3个或 3个以上的顶点组成的全连通子图, 即任何两点之间都直接相连。这是要求最强的一种定义, 它可以通过弱化连接条件进行拓展, 形成 n—派系。例如: 2—派系是指子图中的任意两个顶点不必直接相连, 但最多通过一个中介点就能够连通。3—派系是指子图中的任意两个顶点, 最多通过两个中介点就能连通。随着 n值的增加, n—派系的要求越来越弱。这种定义允许社团间存在重叠性[21]。所谓重叠性是指单个顶点并非仅仅属于一个社团, 而是可以同时属于多个社团。社团与社团由这些有重叠归属的顶点相连。有重叠的社团结构问题有研究的价值, 因为在

实际系统中, 个体往往同时具有多个群体的属性。

除上述提到的社团定义以外, 还有多种其他定义方式, 文献[24] 进行了更为详细的介绍。本文重点关注于可以完全分离的社团结构问题, 并采用较为常用的基于相对连接频数的社团结构定义。

2.2　社团结构的定量描述 —— 模块化 Q函数

在探索网络社团结构的过程中, 描述性的定义无法直接应用。因此 Girvan和 Newman定义了模块化函数[25], 定量地描述网络中的社团, 衡量网络社团结构的划分。所谓模块化是指网络中连接社团结构内部顶点的边所占的比例与另外一个随机网络中连接社团结构内部顶点的边所占比例的期望值相减得到的差值。这个随机网络的构造方法为: 保持每个顶点的社团属性不变, 顶点间的边根据顶点的度随机连接。如果社团结构划分得好, 则社团内部连接的稠密程度应高于随机连接网络的期望水平。用 Q函数定量描述社团划分的模块化水平。

假设网络已经被划分出社团结构, ς_i为顶点 i所属的社团, 则网络中社团内部连边所占比例可以表示成

$$\frac{\sum_{ij} A_{ij}(\varsigma_i, \varsigma_j)}{\sum_{ij} A_{ij}} = \frac{1}{2m}\sum_{ij} A_{ij}\delta(\varsigma_i, \varsigma_j)$$

其中, A_{ij}为网络连接矩阵中的元素, 如果 i, j两点有边相连则 $A_{ij}=1$, 否则等于 0, $\delta(\varsigma_i, \varsigma_j)$为 δ函数, 即 $\varsigma_i=\varsigma_j$时值等于 1, 否则等于 0, $m=\frac{1}{2}\sum_{ij} A_{ij}$为网络中边的数目。在社团结构固定, 边随机连接的网络中, i, j两点存在连边的可能性为$\frac{k_i k_j}{2m}$, k_i为顶点 i的度。所以 Q函数的表达式[26] 为

$$Q=\frac{1}{2m}\sum_{ij}\left[A_{ij}-\frac{k_i k_j}{2m}\right]\delta(\varsigma_i, \varsigma_j)$$

Q函数还有另一种表达方法[25]。如果网络被划分为 n个社团, 那么定义 $n\times n$的对称矩阵 e, 其中的元素 e_{ij}表示连接社团 i与社团 j中的顶点的边占所有边的比例。这个矩阵的迹 $\mathrm{Tr}e=\sum_i e_{ii}$表示网络中所有连接社团内部顶点的边占总边数的比例。定义行 (或列) 的加总值 $a_i=\sum_j e_{ij}$表示所有连接了社团 i中的顶点的边占总边数的比例。由 e_{ij}和 a_i的定义可知 $e_{ij}=a_i a_j$, 从而, Q函数可以表达为

$$Q=\sum_i (e_{ii}-a_i^2)=\mathrm{Tr}e-\|e^2\|$$

其中, "$\|e^2\|$"为矩阵 e^2的模, 即 e^2中元素的加总。同时, Q函数还可以表达为[23]

$$Q=\sum_{v=1}^{n}\left[\frac{l_v}{L}-\left(\frac{d_v}{2L}\right)^2\right]$$

其中, l_v为社团 v中内部连边的数目, d_v为社团 v的总度值, L为网络中的总边数。

如果社团内部顶点间的边没有随机连接得到的边多, 则 Q函数的值为负数。相反, 当 Q函数的值接近 1时, 表明相应的社团结构划分得很好。实际应用中, Q的最大值一般在 0.3 ~ 0.7的范围内, 更大的值很少出现[25]。在社团结构的划分过程中, 计算每一种划分所对应的 Q值, 即模块化值, 并找出数值尖峰所对应的划分 (通常会有一两个), 这就是最好或最接近期望的社团结构划分方式 (图 2)。

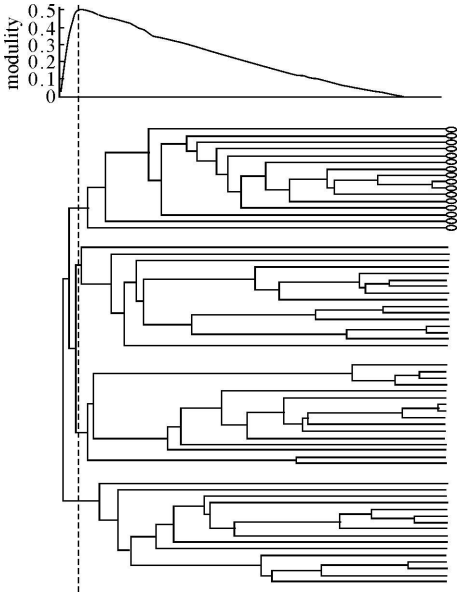


图 2　不同社团结构划分对应不同模块化函数值[25]

模块化 Q函数自提出以来得到了广泛的认可，不但完善了一些旧有的探索社团结构的算法，而且发展了众多以 Q函数为目标函数的新算法。然而，最近有些科研工作者对 Q函数的有效性提出了质疑[23]。他们指出，根据模块化的定义，包含 l条内部连边且总度值为 d的子图 V满足$\frac{l}{L}-\left(\frac{d}{2L}\right)^2>0$，则这个子图便可视为一个社团。但在最大化 Q函数的过程中，若这种社团内部连边的数目小于 $\sqrt{2L}$，即使社团间的连接十分稀疏，它们也会被合并成一个大社团。因此 Q函数最大值所确定的社团，可能是多个满足模块化社团定义的小社团的结合。

3 社团划分思路及社团划分的相关问题

3.1 社团划分思路

按照复杂网络中社团形成的过程，网络中社团结构的划分思路大体可以分成 4类：凝聚过程、分裂过程、搜索过程和其他过程。

凝聚过程是以顶点为基础，通过逐步凝结形成社团。其主要步骤为：1）设定某种标准可以衡量社团与社团之间的距离或相似度；2）将网络中的每一个顶点视为一个社团，所以网络中有多少顶点就有多少初始社团，并且每个社团只包含一个顶点；3）根据设定的衡量标准，计算社团与社团间的距离或相似度，并将距离最近的社团或相似度最高的社团合并在一起形成新的社团；4）重新计算每对社团间的距离或相似度；5）不断重复合并及重新计算的步骤，直到所有顶点都聚集成一个社团。整个过程可以用一个倒立的树状图表示，如图 3，第 4节将介绍的层次算法就是一个典型的例子。

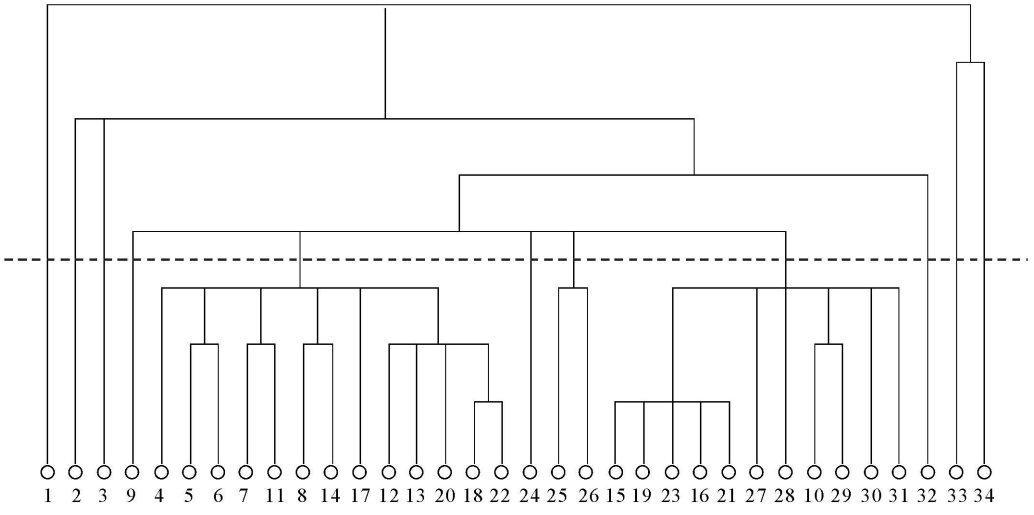


图 3 树状图[22]

分裂过程则相反，它首先将网络中的所有顶点都视为一大社团，通过逐步分割这个大社团形成小社团。其主要步骤为：1）设定某种衡量顶点间密切程度或边对网络结构影响程度的指标；2）按照一定标准进行断边；3）不断重复计算和断边的过程，网络将被划分成一个个越来越小的连通社团，这些连通社团就是对应某一阶段的社团；4）全部过程以每个顶点被独立地分成一个社团为终点。整个过程可以用一个直立的树状图表示，过程方向与凝聚过程相反。其代表之一就是第 4节中将介绍的 GN算法。

搜索过程不拘泥于统一的凝结或分裂，而是建立一个逐步优化目标的探索过程，社团结构直接由最后的优化结果给出。搜索的方法可以应用成型的算法，Potts模型算法中就应用了模拟退火算法进行搜索。

其他过程包含不归属于上述 3个过程中的其他社团形成过程。如第 4节将介绍的谱分析算法。

3.2 划分方法的复杂度研究

由于各种划分方法在判断标准, 优化思路、搜索步骤等方面的不同, 每种划分算法都有其自身的复杂度。复杂度往往与被研究网络的顶点数目、边的数目、层次的数目、社团的数目有关。划分方法的复杂度越高划分所需的时间越长。对于较小的网络, 划分方法所需时间的长短不会产生本质的影响。然而由于受到计算机技术的制约, 一种划分方法的复杂度决定该划分方法能否运用于大规模的网络。可见, 算法复杂度既是判断划分方法速度的标准, 又是判断划分方法适用范围的标准。复杂度越小的划分方法, 划分的速度越快, 适用的网络规模越大。因此, 构造复杂度低的算法, 是科研工作者的目标之一。算法复杂度的具体分析方法将在算法介绍中作简单讨论。

3.3 检验划分方法的经典网络

检验划分方法的网络有两大类: 人造网和实际网。之所以要构建人造网, 是因为人造网的结构可以人为给定, 在分析之前就拥有较多的已知信息, 从而可以用来检验划分方法的有效性及正确率。另外, 人造网的参数可以调控, 从而可以研究划分方法的适用范围, 以及划分正确率与参数的联系。常用的人造网是由 128个顶点构成的网络[12], 这 128个顶点被平均分成 4份, 形成 4个社团, 每个社团包含 32个顶点。顶点之间相互独立的随机连边, 如果两顶点属于一个社团, 则以概率 p_{in}相连, 如果两点属于不同的社团, 则以概率 p_{out}相连。p_{in}和 p_{out}的取值, 保证每个顶点的度的期望值为 16。记 Z_{in}为顶点与社团内部顶点连边数目的期望值, Z_{out}为顶点与社团外顶点连边数目的期望值, 从而 $Z_{in}+Z_{out}=16$。Z_{out}越小, 说明顶点与社团外顶点的连边越少, 网络的社团结构越明显; Z_{out}越大, 说明顶点与社团外顶点的连边越多, 网络越混乱, 社团结构越不明显。对于 Z_{out}值大的网络还能够基本正确对网络进行划分的方法, 在实际应用中适用范围更广, 价值更大。众多方法的实践表明, 当 Z_{out}的取值在一定范围内时, 其值对顶点划分正确率没有影响, 并且正确率都保持在 100%, 然而当 Z_{out}的取值超过这一临界值之后, 网络中顶点被正确划分的比率与 Z_{out}的取值呈现负相关关系, 即 Z_{out}越大, 顶点被正确划分的比例越低。

人造网的检验在一定程度上反映了划分方法的有效性。然而, 由于人们感兴趣的问题大多是实际网络, 所以需要用实际网络对划分方法进行再检验。选择用作检验的实际网络时, 首先要保证构建网络的数据是方便易得的; 其次要保证网络有实际的意义, 从而可以判断社团划分的结果是否具有可解释性; 另外为了方便划分方法间的比较, 宜采用已被广泛使用的实际网络。以下介绍几个常用的实际网络, 方便对第 4节所介绍的划分方法进行比较。

空手道俱乐部网[12]: 20世纪 70年代初, Wayne Zachary观察了美国大学空手道俱乐部成员间的人际关系, 并依据俱乐部成员间平时的交往状况建立了一个网络。这个网络包含 34个顶点, 代表了俱乐部成员; 包含了 78条边, 代表他们之间的人际关系。由于突发的原因, 俱乐部管理者与俱乐部主要教师之间针对是否提高收费这一问题产生了激烈的争论, 并最终导致俱乐部分裂成两部分。图 4为空手道俱乐部网, 其中方形顶点代表归于俱乐部管理者 (1号顶点) 的成员, 圆形顶点代表归于俱乐部主要教师 (33号顶点) 的成员。

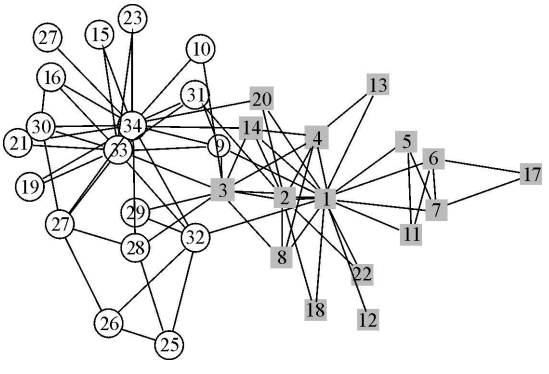


图 4　空手道俱乐部人际关系[25]

科学家合作网: 科学家之间合作的表现方式是广泛的, 如文章的合作、引用、致谢。这里介绍 3个科学家合作网: 1) 物理学家合作网[26], 它是收集了 arxiv.org网上的关注于物理研究的科学家的文章, 并据此构建的科学家合作网。其中顶点表示发表过文章的科学家, 如果两位科学家共同发表过文章就将他们用边连接起来; 2) 桑塔菲研究所科学家合作网[12], 是收集了 1999年、2000年研究所内 271位科学家的合作情况构建的网络, 其

中的顶点是桑塔菲研究所内的科学家,如果他们合作发表过文章,就将他们用边连接起来; 3)经济物理学家合作网[27],收集了经济物理学主页(www. econophysics. org)上发表的文章、Physica A发表的经济物理学文章、ISI(www. isi knowledge. com)上发表的文章,根据文章的合作、引用、致谢建立经济物理学者之间的关系。

类似的还有根据美国大学橄榄球队 2000年一个赛季的赛程建立的网络[12],其中顶点代表大学的橄榄球队,共有 115支大学橄榄球队,连边表示两队之间进行了常规赛,这一赛季共包含 616场常规赛;根据 Linda Wolf收集的猴子 3个月的活动数据,通过猴子之间相互刷毛的关系抽象出的网络[28],其中 16个顶点代表 16只猴子,若两只猴子相互刷毛则用边连接起来,共有 69条边。在这些实际网络中,除了空手道俱乐部网和美国大学橄榄球比赛网有可以判断划分准确性的已知社团结构,其他网络并没有这样的标准,因此用来判断划分方法优劣的标准是划分结果是否具有可解释性。虽然这样的判断标准是不严谨的,然而社团结构划分的主要对象正是这样未知结果的网络,因此结果更具解释性的划分算法更有价值。

3.4 划分结果的比较方法

不同的算法往往会将同一网络划分出不同的社团结构。对于社团结构已知的网络,划分结果与网络真实社团的比较可以得到划分方法的准确性;对于社团结构未知的网络,多种划分方法所得结果间的比较同样可以加深对各种算法的理解及对网络的了解。划分结果的比较方法主要有 3种。

3.4.1 正确划分率比较法[12]

这种方法多用在社团结构已知的网络研究中,比较对象是划分得到的社团结构与网络实际的社团结构。具体方法是在划分得到的所有社团中,找到能够被真实社团结构中的任一社团所包含的规模最大的社团。并以此社团为标准,顶点数超过此标准的社团中的顶点都被视为错误划分了,小于此标准的社团,则只将不在真实社团中的顶点视为错误划分。以 128个顶点的人造经典网为例,如果划分得到的 3个社团,其中两个社团与真实的社团完全相同,第 3个社团由另外两个真实社团组成,则正确划分的比例是 50%。这种方法过于严厉,会将一些人们主观认为被正确划分的顶点归于错误划分。然而,这种比较方法的应用十分广泛,大多数研究者在研究经典人造网时使用它。

3.4.2 共同信息比较法

Danon L等人将文献[30-31]介绍的标准共同信息的衡量引入到社团结构的比较中来[29],并认为这种方法较之 3.4.1中介绍的方法更具判别力。具体过程为:首先定义一个混乱矩阵 N,其中行为真实的社团,列为划分得出的社团。矩阵 N中的元素 N_{ij}为既在真实社团 i中出现又在划分出的社团 j中出现的顶点的个数。基于信息理论得到的两种社团结构 A, B的相似程度为

$$I(A,B)=\frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}N_{ij}\log\left(\frac{N_{ij}N}{N_{i\cdot}N_{\cdot j}}\right)}{\sum_{i=1}^{c_A}N_{i\cdot}\log\left(\frac{N_{i\cdot}}{N}\right)+\sum_{j=1}^{c_B}N_{\cdot j}\log\left(\frac{N_{\cdot j}}{N}\right)}$$

其中,真实社团的个数用 c_A表示,划分所得结果中的社团个数用 c_B表示, $N_{i\cdot}$为 N_{ij}的第 i行的加总, $N_{\cdot j}$为 N_{ij}的第 j列的加总。

如果划分结果与真实的社团结构完全一致,则 $I(A,B)$达到最大值 1;当划分结果与真实社团结构没有任何重叠时 $I(A,B)$达到最小值 0。可见, $I(A,B)$值越大说明社团划分的准确性越高。采用这种比较方法衡量 3.3节中提到的人造网的例子,得到划分的正确率为 0.858,比原来的结果更加符合人们的感受。

3.4.3 D函数比较法[32]

两种划分结果差异性可以分解成社团对之间差异的总和, D函数法就是采用这一思路讨论两种划分结果间的差异性。划分得到的社团可以视为集合,网络划分的结果就是一组集合,社团间的差异表现为集合中的不同元素。设 A, B是任意两个集合,定义 $A\cap B$为两个集合的相似度,而 $(A\cap\bar{B})\cup(\bar{A}\cap B)$($\bar{A}$和 $\bar{B}$的全空间是 $A\cup B$)是两个集合的相异度。从而,集合 A, B标准化后的相似度(s)和相异度(d)为

$$\begin{cases} s = \dfrac{|A \cap B|}{|A \cup B|} \\ d = \dfrac{|(A \cap \bar{B}) \cup (\bar{A} \cap B)|}{|A \cup B|} \end{cases}$$

两种划分结果就是两组不同的社团, 对它们进行比较时有多种配对的方法, 这里采用的比较规则为:

1) 建立不同划分得到的两组社团之间的匹配关系: 将两个集合组中的集合进行对比, 相似度最大的两个集合组成一对, 然后根据相似度排序把各个集合配对。若两组集合所包含的集合数目不相等, 则多出的集合与空集配对。

2) 根据配对, 计算每对集合的相异度。

3) 综合每对集合的相异性, 得到两种划分的相异度的数值:

$$D = \frac{\sum d_{XY}}{k}$$

其中, XY为配对的集合, k为集合对的总数。

D函数的取值范围是 [0, 1], 取值为 1表示两种划分完全不同, 取值为 0表示两种划分完全相同, 可见取值越大说明两种划分之间的差异越大。用这种方法考察上述两小节中的例子, 得到两种划分的相似度或者说划分的正确率为 0. 625, 介于前两种指标之间。

4　社团划分的具体算法

随着关注网络社团结构问题的科研工作者不断增加, 众多划分网络社团结构的算法被设计出来。根据不同的标准, 这些算法可以被分成不同的种类。例如: 根据第 3节提到的社团结构的形成过程, 算法可以分为凝聚算法、分裂算法, 搜索算法及其他算法 4大类。从算法的物理背景上考虑, 又可以将其分为基于网络拓扑结构的算法, 基于网络动力学的算法, 基于 Q函数优化的算法及其他算法。在本节中将根据这一分类, 对其中的几种算法分别予以具体介绍。

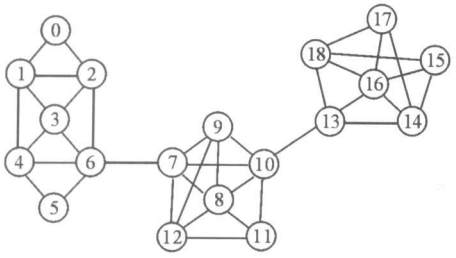
4. 1　基于网络拓扑结构的算法

这类算法的特点在于关注网络连通形成的拓扑结构, 并应用拓扑结构的特性刻画顶点和连边, 划分网络中的社团。应用这类算法时往往并不需要额外的信息。

4. 1. 1　谱分析思想的算法

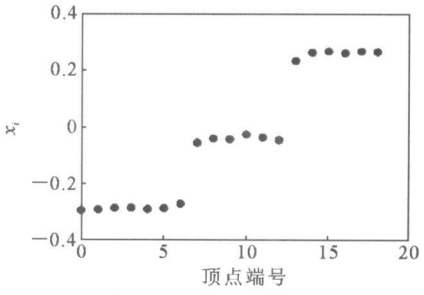
谱分析早在 20世纪 70年代就已经有所发展[33-34], 到 90年代变得普及。它的主要思想是通过对连接矩阵形成的拉普拉斯矩阵或标准矩阵的特征值、特征向量的分析, 挖掘网络中的社团结构[35]。以标准矩阵的分析为例, 具体介绍这种算法[36]。所谓标准矩阵 N, 是由网络的连接矩阵 A和一个对角矩阵的逆矩阵 K^{-1} 构成的, $N = K^{-1}A$, 对角矩阵 K中的元素是每个顶点的度值 $k_i = \sum_{j=1}^{S} a_{ij}$, S为网络中顶点的个数。由于标准矩阵行的标准化, 标准矩阵总有最大的特征值等于 1, 以及与之对应的特征向量 (1, 1, 1……)。在对社团化明显的网络如图 5的分析中发现, 如果网络自然呈现 m个社团, 则标准矩阵 N就有 m-1个十分接近 1的特征值, 而其余的特征值则有较大的距离。这 m-1个特征值中, 最大的特征值所对应的特征向量有一个特性: 在同一个社团中的顶点所对应的值较为接近。因此, 特征向量元素的值呈现阶梯状分布, 如图 6所示, 并且阶梯的级数与社团的个数相匹配。

这种方法对社团结构比较清晰的网络十分有效, 然而实际网络的社团结构往往并非这般显著。在由众多顶点构成的连接混乱的网络中, 社团间的过渡是平滑的, 第一大非平庸特征值对应的特征向量中的元素没有呈现明显的阶梯状分布, 取而代之的是几乎光滑的曲线。可见, 仅仅参考这一指标无法进行社团结构的划分。因此需要对上述方法进行拓展, 使之适用范围更广。



顶点 0 ~ 6 号为一个社团,顶点 7 ~ 12 号为一个社团,顶点 13 ~ 18 号为一个社团

图 5 一个社团结构清晰的网络[36]



0 ~ 6 号的数值比较接近,7 ~ 12 号的数值比较接近,13 ~ 18 号的数值比较接近

图 6 特征向量中顶点对应数值[36]

谱分析法根据特征值、特征向量对顶点进行划分的过程可以理解为:依据特征值设立众多的标准,并根据这些标准对顶点进行划分。根据平庸特征值 1,无法对顶点做出任何的区分;对于社团结构明显的网络,只需要采用最大的非平庸特征值,就可对顶点进行划分,而对于混乱的网络仅用这一个标准无法实现。划分混乱的网络,需要综合考虑多个标准,即同时考虑多个特征值对应的特征向量。其依据是同属一个社团的顶点在各个标准上都更为相近。通过对两个顶点在各个标准下取值的综合考察,得到它们的紧密程度,用以表明它们同属于一个社团的倾向[36]。,i 两顶点间的紧密程度表示为

$$r_{ij}=\frac{\langle x_i x_j\rangle-\langle x_i\rangle\langle x_j\rangle}{[(\langle x_i^2\rangle-\langle x_i\rangle^2)(\langle x_j^2\rangle-\langle x_j\rangle^2)]}$$

其中,〈 xi〉为几个较大的特征值所对应的特征向量中顶点 对应的元素的平均值。虽然多考虑一些特征值会使精确性有所提高,但是由于过多的特征值和特征向量的计算会大大提高计算量。

同样的方法也可以对拉普拉斯矩阵进行分析。差别在于,拉普拉斯矩阵总存在平庸的特征值 0,考察的标准是大于 0的最小的特征值及其对应的特征向量。

这类方法对于社团结构显著的网络是高效的,然而对于规模大连接混乱的网络却没有明显的优势。因为对于大规模的网络而言,求特征值和特征向量的计算相当复杂耗时。并且即便得到两顶点间的紧密程度 rij 也需要人为的设定标准,判断它们是否归为一类。从而,使得划分的结果在很大程度上受到了人为因素的影响。

4.1.2 层次聚类算法

层次聚类法在社会科学中被广泛应用[37-38],其核心思想是:由距离最近、相似度最高的社团开始合并,直到所有元素都归于一个社团为止。可见这算法的核心在于对距离及相似度的定义。针对复杂网络社团划分问题,已有一些距离和相似度的定义。点与点之间的距离可以定义为两点间的最短路径,它们的相似度定义为最短路径的倒数。这样最短路径近的顶点相似度较高,最短路径远的顶点相似度较低。一种较为科学的方法则是用结构等价的程度来衡量两顶点的相似度。结构等价的概念在 1971年由 Lorrain和 White引入社会网络。如果一个顶点与网络中其余顶点的连接方式和另一顶点与网络中其余顶点的连接方式完全相同,则这两个顶点结构等价。例如在人际关系网中,如果两个人的朋友完全相同,则这两个人结构等价。Burt首次引入欧几里德距离衡量结构等价,顶点 ,i 之间的欧几里德距离为

$$D_{ij}=\sqrt{\sum_{k=1,k\neq i,j}^{S}(a_{ik}-a_{jk})^2}$$

其中, aik为连接矩阵中的元素,表示顶点 与顶点 k的连接状况。如果 ,i 两点结构完全等价,则它们的连接矩阵完全相同,所以它们的距离 Dij等于 0。网络中的每一对顶点都可以计算出距离,然后按照凝聚思想划分社团结构。首先选择距离最小的归为一个社团,在进一步的凝聚过程中,由于每个社团所包含的元素不再唯一,

因此要定义包含多个元素的社团间的距离。常用的方法有 3种: 1)最短距离法, 两个社团的距离等于两个社团间所有顶点对的距离中最短的值; 2)最长距离法, 两个社团的距离等于两个社团间所有顶点对的距离中最长的值; 3)平均距离法, 两个社团的距离等于两个社团间所有顶点对距离的平均值。任意选取一种定义将凝聚步骤进行下去都可以将网络中顶点间的关系用树状图表示出来, 并且可以通过 Q函数确定最优的社团划分。

4.1.3　GN算法

Girvan和 Newman提出的分裂算法已经成为探索网络社团结构的一种经典算法, 简称 GN算法[12, 22, 25]。由网络中社团的定义可知, 所谓社团就是指其内部顶点的连接稠密, 而与其他社团内的顶点连接稀疏。这就意味着社团与社团之间联系的通道比较少, 从一个社团到另一个社团至少要通过这些通道中的一条。如果能找到这些重要的通道, 并将它们移除, 那么网络就自然而然地分出了社团。Girvan和 Newman提出用边介数来标记每条边对网络连通性的影响。某条边的边介数是指网络中通过这条边的最短路径的数目。两顶点间的最短路径在无权网中为连接该顶点对的边数最少的路径。由此定义可知, 社团间连边的边介数比较大, 因为社团间顶点对的最短路径必然通过它们; 而社团内部边的边介数则比较小。这种算法的具体过程是: 1)计算网络中各条边的边介数; 2)找出边介数最大的边, 并将它移除 (如果最大边介数的边不唯一, 那么既可以随机挑选一条边断开也可以将这些边同时断开); 3)重新计算网络中剩余各条边的边介数; 4)重复第 2)、3)步, 直到网络中所有的边都被移除。

算法中包括了重复计算边介数值的环节是十分必要的。因为当断开边介数值最大边后, 网络结构发生了变化, 原有的数值已经不能代表断边后网络的结构, 各条边的边介数需要重新计算。举一个形象的例子: 假如网络中有两个社团, 它们之间只有两条边相连。起初其中一条边的边介数最大, 而另外一条边介数较小, 则第一条边被断开。如果不重新计算各条边的边介数, 那么第二条边依据其原有边介数值可能不会被立即断开。如果重现计算各条边的边介数, 那么第二条边的边介数可能成为最大值, 会被立即断开。这显然会对社团结构的划分产生重大的影响。

GN算法分析网络的整个过程也可以用树状图表示, 网络的最优划分要通过 Q函数进行判断。对于由 n个顶点 m条边构成的网络, 按照广度优先的法则, 计算某个顶点到其他所有顶点的最短路径对网络中每条边边介数的贡献最多耗时 $O(m)$, 由于网络中共有 n个顶点, 所以计算网络中每条边的边介数总共耗时 $O(mn)$, 又因为每次断边后需要重新计算每条边的边介数, 因此总体上讲这种算法的复杂度为 $O(m^2n)$; 对于稀疏网, 算法的复杂度为 $O(n^3)$。复杂度较高是 GN算法的显著缺点。

应用 GN算法分析人造经典网络, 得到的结果如图 9所示。当 Z_{out}小于等于 6时, 有 90% 以上的顶点被正确划分; Z_{out}继续增加时, 正确划分的比例迅速下降, 当 Z_{out}等于 8时, 正确划分的比较仅为 30% 左右。所以 GN算法不适用于较为混乱的网络。对空手道俱乐部人际关系网的划分, GN算法的准确率很高, 仅有标号为 3的顶点被划分到错误的社团。结果的树状图及与之对应的 Q函数变化如图 7所示。用 GN算法对物理学家合作网的最大连通社团进行研究。科学家之间的合作关系如图 8所示, 其中每一个矩形代表一位科学家。分析的结

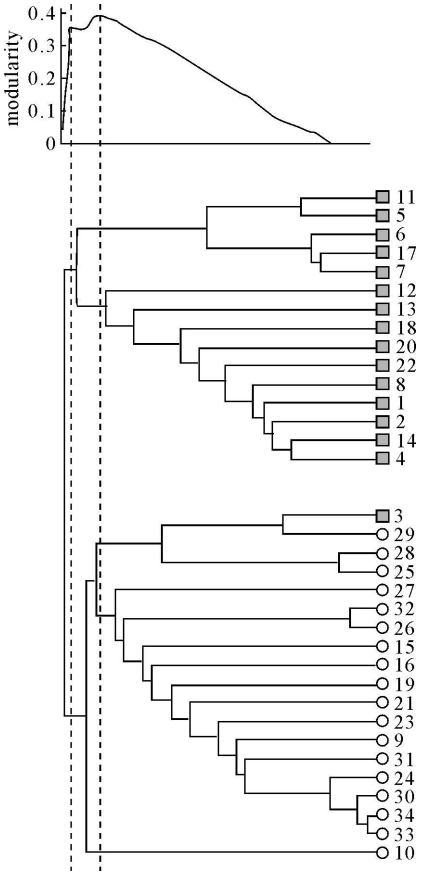
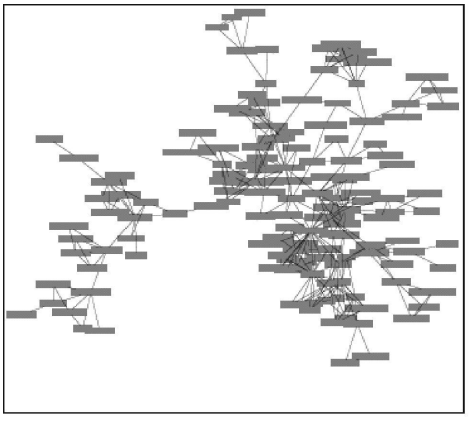


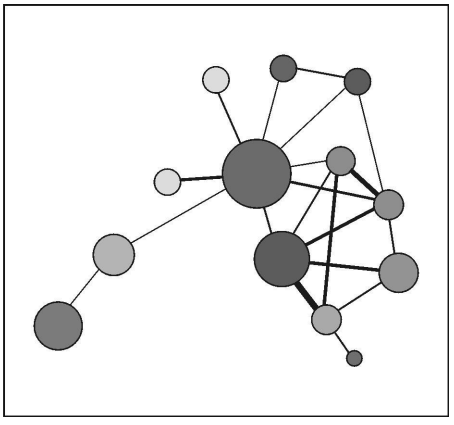
图 7　GN算法得到的空手道俱乐部社团结构划分的树状图及对应的模块化函数值[25]

果是: 当网络划分成 13个社团时, Q= 0. 72 ±0. 02为峰值。用图 8 b形象地表示出来, 其中圆形代表社团, 圆形的大小代表包含的科学家的多少; 连边表示社团间的合作关系, 边的粗细代表合作的密切程度。划分出的社团具有可解释性, 如处于中心位置的社团所包含的科学家基本来自西欧国家。



矩形表示科学家, 连边表示合作关系

a 网络示意图



圆形代表社团, 圆形的大小代表包含的科学家的多少; 连边表示社团间的合作关系, 边的粗细代表合作的密切程度

b 社团划分结果

图 **8** 科学家合作关系网[25]

4. 1. 4 边集聚系数法

GN算法的核心概念最短路径边介数由网络的全局结构决定, Radicch等人提出基于网络局部结构的边集聚系数的定义[18], 并以此寻找社团间的连边。一条边的边集聚系数定义为网络中包含该边的实际三角形数目与包括该边的所有三角形的数目之比。其中包含该边的所有三角形包括实际三角形和潜在三角形。, i j 两顶点间连边的边集聚系数为

$$C_{i,j}^{(3)} = \frac{z_{i,j}^{(3)}}{\min[(k_i-1),(k_j-1)]}$$

其中, $z_{i,j}^{(3)}$ 为包含该边的实际三角形个数, $\min[(k_i-1),(k_j-1)]$ 为包含该边的所有三角形数目。由于社团内部顶点间的连接比较稠密, 所以处于社团内部的连边被较多的实际三角形所包含; 而社团间的连边被较少的实际三角形包含, 甚至不被任何实际三角形包含。因此根据边集聚系数可以挖掘网络中社团间的连边。然而当包含某边的实际三角形个数 $z_{i,j}^{(3)} = 0$时, 无论构成这条边的两个顶点的度 k_i k_j为何值, $C_{i,j}^{(3)}$都等于 0 无法反映出结构的差异。为克服这一缺陷, 对原式做微小的调整:

$$C_{i,j}^{(3)} = \frac{z_{i,j}^{(3)} + 1}{\min[(k_i-1),(k_j-1)]}$$

边集聚系数的定义可以进一步推广到更大的环, 如考虑网络中的四边形、五边形 ……, 从而边集聚系数的通式为

$$C_{i,j}^{(g)} = \frac{z_{i,j}^{(g)} + 1}{s_{i,j}^{(g)}}$$

其中, g为研究包含边的 g边形, $z_{i,j}^{(g)}$ 为包含该边的实际 g边形的个数, $s_{i,j}^{(g)}$ 为包含该边的所有 g边形个数。

算法的具体过程为: 1)确定研究环的种类 (三角形、四边形 ……); 2)根据定义计算每条边的边集聚系数, 断开边集聚系数最小的边; 3)重复计算和断边的操作, 直到网络中所有的边都被断开为止。该算法的复杂度大致为 $O(m^2)$, 比 GN算法有显著的降低。然而, 这种方法的局限在于只能分析包含环的网络, 以三角形

为例: 该方法对于包含三角形较多的社会网络有着良好的效果, 但对于非社会网的效果则较差。

4.1.5　信息集中性算法

Latora, Marchiori等为方便分析非连通图的通讯效率提出了信息效率的概念[39—40]。并进一步提出了信息集中性的概念, 用以衡量网络中边的重要性[41]。网络中两顶点 i, j之间信息传递的效率 ε_{ij}等于它们最短路径 d_{ij}的倒数。整个网络的信息传递效率 $E[G]$等于所有顶点对信息传递效率的平均值:

$$E[G]=\frac{\sum_{i\neq j\in G}\varepsilon_{ij}}{n(n-1)}=\frac{1}{n(n-1)}\sum_{i\neq j\in G}\frac{1}{d_{ij}}$$

其中, G为整个网络, n为网络中顶点的数目。若顶点 i, j之间不连通, 即 $d_{ij}=+\infty$, 则它们之间的信息传递效率等于 0。网络边 k的信息集中性定义为由于 k边的移除导致的网络整体信息传递效率的相对变化:

$$C_k=\frac{\Delta E}{E}=\frac{E[G]-E[G'_k]}{E[G]},\quad k=1,\ \cdots,\ m$$

其中, m为网络中包含的边的数目, G'_k为将 G网中 k边移除后形成的包含 m—1条边的新网。Fortunato将这些概念引入到划分社团结构的问题之中[42]。信息集中性越高的边对网络连通性的影响越大, 而社团间的连边往往对网络的连通性有着重要的影响, 因此考虑使用信息集中性探索网络中社团间的连边。算法的具体过程为: 首先计算每条边的信息集中性; 然后将信息集中性最高的边移除; 重新计算网络的信息传递效率; 重复以上的过程直到网络中所有的边都被断开为止; 最终使用 Q函数判断最优的社团结构划分。

应用这种算法分析经典人造网, 得到的结果如图 9所示, 表明该算法适用的网络的混乱程度与 GN算法所能适用的网络的混乱程度相当, 只有微小的优势。然而对于众多实际网络, 该算法的分析结果并不令人满意。问题在于总是在划分初期就产生孤立的顶点, 致使得到的社团结构与实际情况的拟合程度不高或结果的可解释性不强。例如: 对空手道俱乐部网络, 该算法首先分离出两个由单点构成的社团。通过分析发现, 最先被分离的孤立点通常处于网络的边缘如图 10中的 k点, 因为当网络没有明显社团结构时, 移除与该点相连的边会导致网络传递效率的较大变化, 分析结果树状图如图 11。信息集中性算法有这一显著缺陷, 并且它的复杂度较高为 $O(n^4)$, 所以较之其他算法它没有显著的优势。

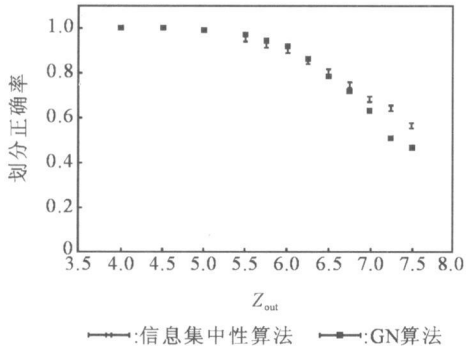


图 9　GN 算法和信息集中性算法划分正确率检验图[24]

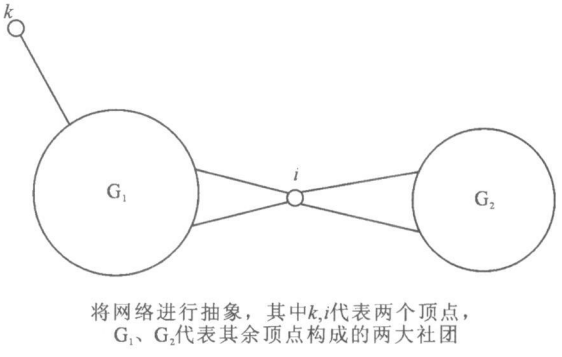


图 10　信息集中性算法易划分出孤立点的情况[42]

4.2　基于网络动力学的算法

4.2.1　Potts模型算法

Jorg Reichardt和 Stefan Bornholdt将物理学的 Potts模型引入到确定网络社团最优划分的问题之中[43—44]。网络中的顶点与模型中的粒子相对应, 并对模型的哈密顿量进行修正。当修正后的哈密顿量处于基态时, 具有相同自旋值的粒子归为一个社团, 从而得到网络的最优划分。可见, 问题的关键在于根据划分社团结构这一目标, 修正哈密顿量。修正的依据是网络社团内部连接稠密, 社团间连接稀疏的性质。因此得出 4项标准: 1)对连接社团内部顶点的边进行奖励; 2)对社团内部可以连接却没有连接的边进行处罚; 3)对连接

社团间顶点的边进行处罚；4）对社团间可以连接却没有连接的边进行奖励。由这 4项标准得到的修正哈密顿量为

$$H(\{\sigma\}) = -\sum_{i\neq j} a_{ij} A_{ij} \delta(\sigma_i, \sigma_j) + \sum_{i\neq j} b_{ij} (1 - A_{ij}) \delta(\sigma_i, \sigma_j)$$
$$\text{internal links} \qquad \text{internal non-links}$$
$$+ \sum_{i\neq j} c_{ij} A_{ij} [1 - \delta(\sigma_i, \sigma_j)] - \sum_{i\neq j} d_{ij} (1 - A_{ij}) [1 - \delta(\sigma_i, \sigma_j)]$$
$$\text{external links} \qquad \text{external non-links}$$

其中，i，j 为网络中的任意两个顶点；A_{ij}为连接矩阵中的元素，如果 i，j 两点有边相连则 $A_{ij}=1$，否则等于 0；σ_i，σ_j在原模型中表示粒子的自旋值，在这里表示 i，j 顶点所属的社团的编号；δ为一函数，当 $\sigma_i=\sigma_j$即 i，j 两顶点属于一个社团时，$\delta(\sigma_i, \sigma_j)=1$，否则等于 0；$a_{ij}$，$b_{ij}$，$c_{ij}$，$d_{ij}$分别为奖励和惩罚的力度。值得指出的是，因为目标是使得哈密顿量最小，所以这里奖励的部分为负号，惩罚的部分为正号。以表达式中的第一项为例：如果顶点 i，j 属于同一社团即 $\delta(\sigma_i, \sigma_j)=1$，且它们之间有边相连即 $A_{ij}=1$，则以 a_{ij}的力度进行奖励，遍历所有不同的顶点对并取和，从而得到由于社团内顶点相连对哈密顿量的贡献。如果两顶点间是否有边相连对哈密顿量影响力度相同，那么有 $a_{ij}=c_{ij}$，$b_{ij}=d_{ij}$，从而将 4个参量压缩成 2个。进一步，需要找出一个适当的参量将 a_{ij}，b_{ij}共同表示出来。一个显见的参数为 p_{ij}，它为顶点 i，j 间存在边的可能性。从而 a_{ij}，b_{ij}可以表达为 $a_{ij}=1-\gamma p_{ij}$，$b_{ij}=\gamma p_{ij}$，其中 γ为 a_{ij}，b_{ij}对 p_{ij}的依赖程度。用顶点间存在边的可能性作为参数是合理的，例如：若 i，j 两顶点间存在边的可能性 p_{ij}比较小，则 a_{ij}比较大，当 i，j 间存在边时，得到的奖励就比较大。进而，修正后的哈密顿量表示为：$H(\{\sigma\}) = -\sum_{i\neq j} (A_{ij} - \gamma p_{ij}) \delta(\sigma_i, \sigma_j)$。使得哈密顿量最小的网络社团划分是最优划分。

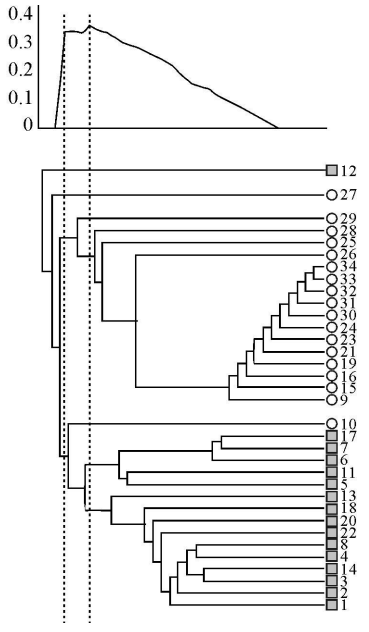


图 11 信息集中性算法得到的空手道俱乐部社团结构划分的树状图及对应的模块化函数值[42]

两顶点间存在边的可能性可以在应用时自行选择，常用的取法有两种。第一种方法是选择一个固定 p 值，即任意两顶点 i，j 之间存在边的可能性 p_{ij}都等于 p，这是最简便的方法。第二种方法要考虑网络的分布，若两顶点的度值都比较大，则它们之间存在连边的可能性就比较大，即任意两顶点 i，j 之间存在边的可能性表示为 $p_{ij}=\frac{k_i k_j}{2M}$，其中 k_i，k_j分别为顶点 i，j 的度值，M为网络中的总边数。

确定目标函数后，可以采用模拟退火算法进行搜索。假设网络中的顶点初始有 q种自旋态，该算法的具体过程为：

1）给定系统一个初始温度，网络中每个顶点都被赋予一个从 q个自旋态中随机选择的状态。

2）随机挑选一个顶点改变它的自旋态。

3）如果新状态产生的系统修正的哈密顿量的变化值 $\Delta H = H_{new} - H_{old} < 0$，那么该顶点就接受这个新的自旋态；如果 $\Delta H = H_{new} - H_{old} > 0$，则在（0，1）间随机选择一个数 ε，若 $\varepsilon < \exp(-\beta \Delta H)$，也接受这个新的自旋态，其中 $\beta = \frac{1}{T}$，否则保持原有状态。

4）回到第 2）步，遍历网络中的所有顶点。

5）降低系统温度，重复以上所有操作。当系统温度接近 0时，停止计算。根据此时每个顶点所处的自旋

态, 对它们进行社团划分。

这种算法的复杂度与计算停止的温度有密切的关系。

在文献 [44] 中, Jorg Reichardt和 Stefan Bornholdt指出, 采用第 2种 P_{ij}的选择方式并且令 $\gamma = 1$时, 哈密顿量与 Q函数有负相关关系: $Q = -\frac{1}{M}H(\{\sigma\})$。此时最小化哈密顿量与最大化 Q函数值是等价的。

4.2.2　随机行走

随机行走算法建立在层次算法之上, 其特别之处在于用随机行走粒子的跃迁行为定义顶点间的距离[45—47]。假设网络上有一个可以任意跳跃到其邻居位置上的粒子, 它每一步跳跃都只与其当时所处的位置有关, 而与之前的状态没有关系, 即一系列跳跃形成一个马尔科夫链。在每一步中, 由顶点 i跳跃到其邻居顶点 j的概率为 $P_{ij} = \frac{A_{ij}}{d(i)}$, 其中 A_{ij}为连接矩阵中的元素, $d(i)$为顶点 i的度。从而得到顶点间的一步转移概率矩阵 P。由顶点间的一步转移概率矩阵可以得出顶点间的 t步转移概率矩阵 P^t, 其中的元素 P^t_{ij}为由顶点 i通过 t步转移到顶点 j的概率。这里的顶点 j可以是网络中的任何顶点, 不局限于顶点 i的邻居顶点。如果两个顶点同属于一个社团, 那么分别从两个顶点透视整个网络得到的结果应该相近, 即如果顶点 i和顶点 j在同一社团, 则对于任意顶点 k有 $P^t_{ik} \cong P^t_{jk}$。两顶点结构等价的程度由距离 r_{ij}衡量, $r_{ij} = \sqrt{\sum_{k=1}^{S}\frac{(P^t_{ik} - P^t_{jk})^2}{d(k)}}$。因为这个值依赖 t的取值, 所以也可以表示成 r^t_{ij}。t值的选取不宜过大, 因为当 t趋于无穷时 P^t_{ij}只与顶点 j的度有关, 而与顶点 i无关。应用这种距离也可以将网络梳理成树状图的形式, 并应用 Q函数得到最优的社团结构。

4.2.3　电流算法

F. Wu和 B. A. Huberman将网络类比成电路, 形成一种算法[48]。其主要思想是: 将网络中的边视为阻值相等的电阻, 在两顶点 i, j上施加一个固定的电压, 比如顶点 i的电势为 1, 顶点 j的电势为 0, 从而整个网络就成为了一个电路, 每个顶点上都会有相应的电势值, 进而按照电势值划分顶点的社团。同属一个社团的顶点, 其电势值应当比较接近, 而与其他社团顶点的电势值相差较多。

顶点电势的计算应用基尔霍夫等式, 它是指流入顶点的电流净值为 0。若顶点 i有 n个与它相连的顶点, 则根据基尔霍夫等式, 流入 i点的净电流为

$$\sum_{k=1}^{n} I_k = \sum_{k=1}^{n}\frac{V_k - V_i}{R} = 0$$

其中, I_k为由顶点 k流向顶点 i的电流, V_k为顶点 k的电势, R为每条边的阻值。从而顶点 i的电势可以表示为 $V_i = \frac{1}{n}\sum_{k=1}^{n} V_k$。即每一顶点的电势等于其所有邻居的电势的平均值。

一般的想法是用谱分析法计算各个顶点的电势值, 然而这种方法消耗的时间比较长。F. Wu和 B. A. Huberman推荐了一种复杂度为线性的算法。在拥有 N个顶点的网络中, 首先将 i, j两点的电势设定, 如 $V_i = 1$, $V_j = 0$, 其余顶点的电势也赋予初始值 0, 然后按照上式更新除 i, j外每个顶点的电势, 称这一过程为一轮; 多次重复这一过程, 在一定精度上可以得到电势值的稳定解。其精确度并不取决于网络的规模, 而是由重复的次数决定的。得到每个顶点的电势值后, 通过设立划分标准, 得到顶点的社团归属。通常选择基本将顶点等分的电势值最大跳跃处作为两社团的划分标准。

因为 i, j两点分别被赋予 1, 0两个电势值, 其他顶点的电势则介于 1和 0之间, 这就意味着 i, j两点一定不会在同一社团当中。在不了解网络的前提下, 如何确定两个不在同一社团中的顶点, 成为需要解决的问题。在同一社团中由于连接稠密, 所以两点之间的距离通常比较近, 而社团间顶点的距离相对比较远。因此距离越大的顶点属于不同社团的可能性越大。根据这一特点, 只需找出距离相对较远的顶点施加电压就基本上可以保证算法的正确性。通过统计发现, 只要不选择相邻的两个顶点, 社团划分的正确率便会大大提高。因此, 另一种解决方法便是对比多次排除相邻两点的随机选择电极得到的社团划分的结果, 确定顶点的归属, 进而

划分网络社团。

这种算法虽然可以推广到多个社团的划分，然而更适合于网络两社团的划分问题。F. Wu和 B. A. Huberman将这一算法运用到空手道俱乐部网，第 1次将电极安放于顶点 1和顶点 34处，第 2次将电极安放在顶点 16和顶点 17处，第 3次将电极安放顶点 12和顶点 26处，第 4次将电极安放在顶点 32和顶点 33处。得到的结果如图 12所示，可见第 1, 2, 3次当电极安放在不同社团的顶点间时，划分的结果很好；而第 4次将电极安放在同一社团的顶点之间，社团没有被正确划分。

4.3 优化 Q函数算法

4.3.1 Newman贪婪算法[49]

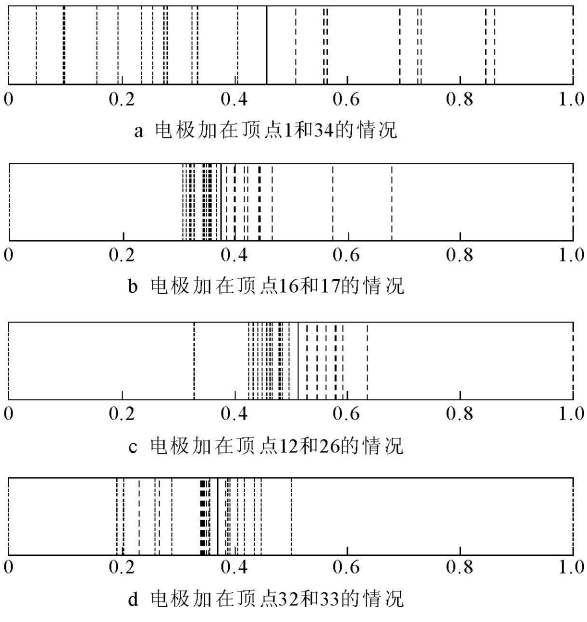
这类算法的共同之处在于都是以最大化 Q函数值为目标，区别在于最大化的途径不同。Newman贪婪算法的过程为：

1）初始时将网络中的每 1个顶点都视为 1个社团，每个社团内只有 1个顶点。即如果网络中共有 n个顶点，则初始有 n个社团。

2）两两合并社团，并计算社团合并所产生的 Q值的变化量：

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$$

选择使得 Q值增加最大（或减少最小）的方式进行合并。需要指出的是，如果两个社团间不存在任何连边，那么它们的合并不能对 Q值产生正向的影响。因此在计算 Q值变化时，只需考虑存在连边的社团对。当网络中包含 m条边时，这步算法的复杂度为 O(m)。社团合并后必然对 e矩阵产生影响（e矩阵的含义参见 Q函数的第 2种表达式），因此将合并的两个社团所对应的行和列相加，对 e_{ij}进行更新。这一步的复杂度为 O(n)。因此这一步最多耗时 O(m + n)。



a 电极加在顶点1和34的情况

b 电极加在顶点16和17的情况

c 电极加在顶点12和26的情况

d 电极加在顶点32和33的情况

线状虚线和点状虚线表示顶点的实际社团属性，
实线表示按照顶点基本等分电压最大跳跃处社团的划分值

图 12 电流算法分析空手道俱乐部网[48]

3）重复步骤 2）的操作，不断对社团进行合并，直到所有顶点被凝聚到一个社团中为止。这样的操作最多进行 n - 1次。

因此，这种算法总体的复杂度为 O((m + n)n)，对于稀疏网则为 $O(n^2)$。这种方法将网络中的社团用树状图的形式表现，使得 Q函数值最大的社团划分方式就是网络的最优划分。这种方法可以直接推广到加权网的分析，只需在初始对 e_{ij}赋值时用权重替代无权网中的 0, 1赋值，并且这种推广不会改变算法的复杂度。

应用 Newman贪婪算法分析人造经典网络，即 2.3小节介绍的包含 128个顶点的网络。随着 Z_{out}值逐渐增大，顶点被正确划分的比例在不断减少，如图 13所示。图 13中横坐标为 Z_{out}值，纵坐标为被正确划分的顶点的比例。当 Z_{out}比较小时顶点的划分完全正确，即便当 $Z_{out} = 6$时，顶点被正确划分的比例也大于 90%。但当 $Z_{out} = 8$即顶点有一半的边与社团外的顶点相连时，正确划分的比例很低，所以这种方法不适宜过于混乱的网络。

应用 Newman贪婪算法分析空手道俱乐部网，Q函数的峰值为 0.381，其对应的网络树状图为图 14。它将网络平均分成两个社团，每个社团包含 17个成员，除 10号成员被划分错误外，其他所有成员的划分都与实际结果相同。

这种方法的突出优势在于复杂度较低，因此适用于规模较大的网络。已经用它对包含 5万多个顶点的物理学家合作网进行了分析。在相同硬件设备上，这种算法所消耗的时间显著少于 GN算法所消耗的时间。Q函数的峰值为 0.713，对应的结果包含 600余个社团，其中有 4个大社团，它们包含的顶点占所用顶点的 77%，

并且这 4个社团具有良好的解释性, 它们与研究领域有着密切的联系, 一个社团对应天体物理学, 一个社团对应高能物理学, 两个社团对应凝聚态物理学。

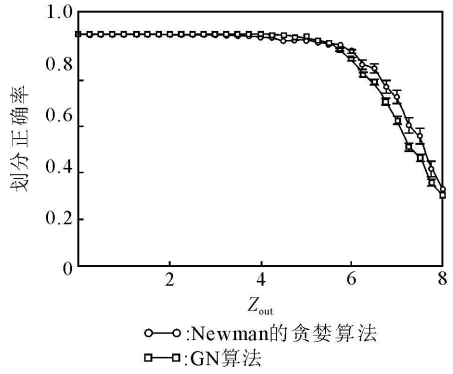


图 13　贪婪算法和 GN算法划分正确率检验图[49]

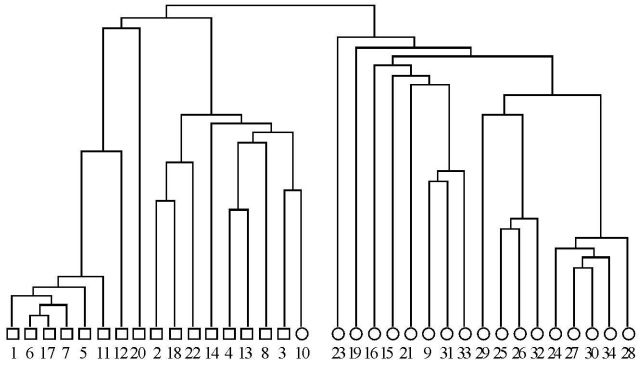


图 14　贪婪算法划分空手道俱乐部网络的树状图[49]

4.3.2　改进的贪婪算法

Clauset等人对 Newman的贪婪算法进行了改进[72], 采用堆的数据结构来存储和运算 Q函数, 使得算法的复杂度进一步降低为 $O(n\log^2 n)$, 接近线性复杂度。Clauset贪婪算法的核心是建立了存储模块化增加量的矩阵 ΔQ, 通过对这个矩阵元素的更新得到 Q值最大的社团划分方式。由于合并没有边相连的社团不会产生 Q值的正向变化, 因此只需要存储有边相连的社团的信息, 这样既节省了存储空间又缩短了运算时间。这种算法应用到 3种数据结构: 1)存储模块化增加量的稀疏矩阵 ΔQ, 其中的元素只包含存在边相连的社团对。并且矩阵的每一行都以平衡二叉树的方式存储; 2)最大堆 H, 其中储存了 ΔQ矩阵中每一行的最大元素, 以及这个元素对应的两个社团的编号; 3)一个辅助向量 a。

方法的具体过程:

1)将网络中的每一个顶点都视为一个社团。在这种前提下, 如果顶点 v_i 有边相连则 $e_{ij}=1/2m$(m为网络中边的数目), 否则为 0, 而 $a_i=k_i/2m$。初始化 ΔQ矩阵:

$$\Delta Q_{ij}=\begin{cases}1/2m-k_ik_j/(2m)^2 & \text{如果 } v_i \text{ 间有边相连}\\ 0 & \text{其他}\end{cases}$$

由初始化的 ΔQ矩阵可以得到每行的最大元素, 从而构成最大堆 H。

2)最大堆 H中找出最大的 ΔQ_{ij}, 合并与之对应的社团 v_i, v_j合并后的社团标记为 v_j, 更新矩阵 ΔQ, 最大堆 H及辅助向量 a, 具体方法为:

(1)对于矩阵 ΔQ, 移除第 i行和第 i列, 更新第 j行和第 j列的元素: 如果社团 k与社团 v_i 都相连, 则 $\Delta Q'_{jk}=\Delta Q_{ik}+\Delta Q_{jk}$; 如果社团 k只与社团 i相连而与社团 j不相连, 则 $\Delta Q'_{jk}=\Delta Q_{ik}-2a_ja_k$; 如果社团 k只与社团 j相连而与社团 i不相连, 则 $\Delta Q'_{jk}=\Delta Q_{jk}-2a_ia_k$。

(2)根据更新后的 $\Delta Q'$ 更新最大堆 H。

(3)辅助向量 a的更新: $a'_j=a_j+a_i$　$a_i=0$。

3)重复 2), 直到所有的顶点都归为一个社团为止。

这种计算过程使得 Q函数有唯一的峰值, 因为当最大的 Q值增量成为负值后, 所有的 Q值增量便都为负值, Q函数的值只能逐渐减小。由此可知, 只要最大的 ΔQ_{ij}由正值变成负值, 就不需要再继续合并社团。因为此时的 Q函数值最大, 所以其对应的社团划分就是最优的划分方式。

这种贪婪算法比 Newman贪婪算法在复杂度方面有显著的降低, 因此适用范围进一步拓宽, 可以分析规模更为巨大的网络。应用这种方法分析包含 40多万个顶点 200多万条边的由 Amazon.com网上书店网页连

接关系构成的网络,得到的结果具有良好的解释性。

4.3.3 极值优化算法

极值优化算法的思想类似于生物系统演化中的断续平衡问题[50],之后用于离散和连续的 NPC问题[51—52],解决如图分割,伊辛模型,原子最优团簇结构等问题。Duch和 Arenas将该思想引入网络社团结构划分问题当中[53],以最大化 Q函数为目标,判断网络中的连边是否被断开。首先定义极值优化算法中的局部变量,它被定义为在一种社团划分下顶点 对总体 Q函数值的贡献,表达式为

$$q_i = \kappa_{i(j)} - k_i a_{i(j)}$$

其中,$\kappa_{i(j)}$为社团 r中的顶点 i与社团 r内的顶点构成连边的数目,k_i为顶点 i的度,$a_{i(j)}$为至少一端在顶点 i所属的 r社团中的边的比例。若用 m表示网络中的总边数,则全局变量 Q与局部变量 q的关系为 $Q = (1/2m)\sum_i q_i$。因为 Q函数的取值范围为 [-1, 1],所以对 q进行标准化,使其取值范围与 Q函数相同,从而得到更为合理的变量,表示顶点 对 Q函数的贡献:

$$\lambda_i = \frac{q_i}{k_i} = \frac{\kappa_{i(j)}}{k_i} - a_{i(j)}$$

λ_i越大表明顶点 对 Q函数的贡献越大,λ_i越小表明顶点 对 Q函数的贡献越小。针对最大化 Q函数这一目标而言,λ_i也反映出顶点 归于 r社团的适合性,λ_i值小说明顶点被归于 r社团不太合适。根据这一理解,极值优化算法的具体过程为:

1)任意将网络中的顶点分成等大的两部分,每部分中相互连通的顶点形成一个社团,从而形成一个初始的社团结构。如图 15 a分别用圆圈和方形代表随机分成的两部分,图 15 b表示这种任意等分得到的初始社团结构,其中一种灰度代表一个社团。可见,初始的等分并不将网络划分成等大的两个社团。

2)根据社团结构计算每个顶点的适合度 λ,并将适合度最低的顶点归入另外一部份中去 (如从初始的圆圈变为方形),这可能使得社团结构发生巨大的变化。计算新社团结构的 Q函数值,并按照新的社团结构重新计算每个顶点的适合度。

3)重复 2)过程,直到得到最大的 Q函数值为止。断开两部分之间的所有连边,从而将网络划分成了两个社团,即由圆圈和方形代表的两个社团,如图 15 c中的左图。

4)对得到的社团递归地重复上述 1)~3)操作,当 Q函数值不能被进一步增大时,就得到了网络社团的最优划分。如图 15 c所示,逐步将网络划分成 4个社团。

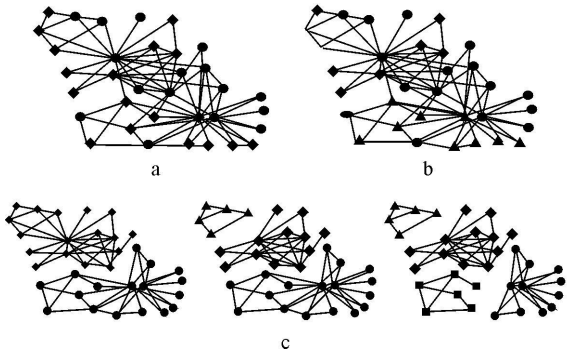


图 15 EO算法的划分过程示意图[53]

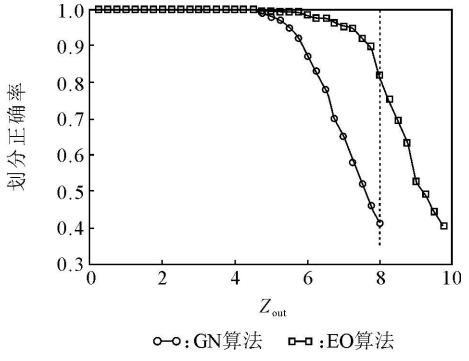


图 16 EO算法和 GN算法划分正确率检验图[53]

上述方法可能导致搜索陷入局部极值,因此用 τ-EO方法进行改进。每个顶点被选到的概率为:$P(q) \propto q^{-\tau}$。其中 q值为顶点按照适合度排列的序号,$\tau \sim 1 + 1/\ln(N)$与网络的规模有关。

本方法的基本思想是:逐步达到最优划分。首先以将网络划分成两个社团为目标,并以顶点的适合度值判别需要调整的顶点,通过调整顶点的属性,达到将网络划分成两个社团的最优划分;再逐步增加社团数目,

并分别达到对应社团数的最优划分; 直到 Q值不能被进一步提高为止, 便确定了最优社团数目及对应的划分方法。

应用极值优化算法对经典人造网进行分析, 顶点划分正确率与 Z_{out}取值的关系如图 16所示。当 Z_{out}小于等于 6时, 顶点划分的正确率都为 100%; 当 Z_{out}等于 8时, 划分的正确率仍有 80%; 但 Z_{out}进一步增大时, 划分的正确率迅速下降, 当 Z_{out}等于 10时, 正确率仅为 40%。由此可知, 若社团间的连边平均占所有连边 50%或以下, 极值优化算法都能较好地进行分析, 说明此算法也适用于混乱的网络。这种算法的复杂度为 $O(n^2 \log n)$。推广到加权网时, 只需相应调整成加权网中的 Q函数。

4.4　其他算法

有些研究者对已有的算法进行适当的综合, 提出了一些新的算法, 例如: Newman将 Q函数与谱分析算法相结合[54—55]。Josep M P等人将随机行走思想、贪婪算法和 Q函数相结合[56], 张世华等人将谱分析法、模糊聚类法和 Q函数相结合[57]等等; 有的学者通过网络演化的思想对社团结构进行分析[58]; 有的学者则通过视觉大致分析网络社团结构[59]。另外, 有些研究者通过挖掘网络的局部信息, 划分社团结构[60—61]。这种算法可以运用于网络全局信息无法得到的问题, 且得到的结果与基于全局信息的方法相近。下面具体介绍 Newman提出的综合法及 Clauset A提出的基于局部信息的方法。

4.4.1　Q函数与谱分析算法的结合

Newman将 Q函数与谱分析算法相结合, 建立 Q函数矩阵并根据 Q函数矩阵的特征值特征问题的解划分网络中的社团。

假设含有 n个顶点的网络包含两个社团, 如果顶点 i属于第一个社团, 则 $s_i = 1$, 如顶点 i属于第二个社团, 则 $s_i = -1$。A_{ij}表示 i, j之间连接边的条数, 由 A_{ij}组成的矩阵为连接矩阵。k_i, k_j为顶点的度, $m = \frac{1}{2}\sum_i k_i$, i, j之间连边的期望值为 $k_i k_j / 2m$。因此模块化表示为

$$Q = \frac{1}{4m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right) s_i s_j = \frac{1}{4m} s^{\mathrm{T}} B s$$

其中, s是由 s_i组成的向量。

定义一个新的矩阵 B, 称为模块化矩阵, 其中包含的元素为 $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$, 这个矩阵各行各列元素相加之和为 0, 所以这个矩阵一定有特征向量 (1, 1, 1…)以及与之对应的特征值 0, 与拉普拉斯矩阵相同。用矩阵 B的标准化特征向量 u_i的线性组合表示向量 s: $s = \sum_{i=1}^{n} a_i u_i$, 进而模块化函数可以表示成

$$Q = \sum_i a_i u_i^{\mathrm{T}} B \sum_j a_j u_j = \sum_{i=1}^{n} (u_i^{\mathrm{T}} \cdot s)^2 \beta_i$$

其中 β_i是矩阵 B对应特征向量 u_i的特征值, 原式中常数 $1/4m$不影响后续的计算, 因此先不考虑。假设特征值按照递减次序排列 $\beta_1 \geqslant \beta_2 \geqslant \cdots \geqslant \beta_n$, 那么最大化 Q的方法就是选择适当的 $(u_i^{\mathrm{T}} \cdot s)^2$, 使其中的大值与特征值中的大值相对应。如果对 s没有限制, 只需使 s与模块化矩阵的最大特征向量 u_1成比例就可以达到目的。然而事实上 s中的元素 s_i只能取 1或 -1, 无法做到与 u_1成比例。那么只能尽可能使 s与 u_1成平行关系: 当 u_1中的元素 $u_1^{(1)}$大于 0时, 对应的 s_i等于 1; 当 u_1中的元素 $u_1^{(1)}$小于 0时, 对应的 s_i等于 -1。至此就将网络划分出了两个社团结构。

众多的实际应用证明这种方法的优越性。它在有效划分社团结构的同时, 还提供了十分有用的网络信息。向量 u_1中元素的数值表示出了该顶点属于某个社团结构的强度, 某顶点对应值为 0或接近 0, 表示它在两个社团结构的边界线上。

这种算法并非只能将网络划分成两个社团结构。将其拓展到划分多个社团结构的思路是重复上述划分成两个社团结构的过程。需要注意的是, 在第 1次社团结构划分完毕后, 不要将网络中两个社团结构间的连

边删除。因为边的删除会导致 Q函数中度值的变化，从而使得最大化的目标不是初始目标。正确的做法是定义子图模块化矩阵，如果子图中有 n_g个顶点，那么其模块化矩阵 $B^{(g)}$就是 $n_g \times n_g$的矩阵。模块化矩阵的元素 $B_{ij}^{(g)} = A_{ij} - \frac{k_i k_j}{2m} - \delta_{ij}[k_i^{(g)} - k_i \frac{d_i}{2m}]$，其中 $k_i^{(g)}$为顶点 i在子图 g中的度，d_i为子图中各顶点的度（在整个网络中的度）的总和。子图的模块化 $Q_g = \sum_{ij} B_{ij}^{(g)}$的值是重复划分对原有模块化值的增加。重复划分何时为止？这种算法有明确的标准：如果子图的任何划分都不能增加网络的模块化，或得不到正的 Q_g，那么就无需进一步的划分。

总体上这种方法可以概括如下：首先计算模块化矩阵的最大特征值和特征向量，根据这个向量的符号将网络中的顶点划分成两部分，然后对每部分重复上述过程。如果某次划分对总体模块化值没有贡献或贡献为负数，就取消这个划分，将这部分视为不可进一步划分的子图。当整个网络完全由不可划分子图组成时，就得到了社团结构划分的结果。

4.4.2 基于局部信息的方法

以上具体介绍的探索网络社团结构的算法都要求在进行分析前了解网络的整体结构。然而，这一要求对于规模巨大且动态性强的网络是难以实现的，例如万维网。因此有学者提出了基于局部信息的探索网络社团结构的方法。这里具体介绍 Clauset A提出的方法[60]。

网络社团结构划分问题可以转化为对某个刻画网络社团结构函数的优化问题，如 4.3小节介绍的算法，都以 Q函数为优化目标。当缺乏全局信息时，目标函数应当独立于全局性质。由于 Q函数依赖网络中的总边数，因此不能作为基于局部信息划分社团结构的目标函数，需要设立新的仅依赖于部分连接的局域模块函数。

考虑一个网络 G，假设只了解其中一部分 C的全部信息。那么就必然存在一部分 U，只知道这些顶点与 C内部顶点的连接状况，而不知道这些顶点的全部信息。进一步假设获得更多关于 C的信息的唯一方法是访问 U中的顶点 v，获得它的全部连接状况，这样 v成为 C内的顶点，同时可能有一些初始时完全不了解的顶点加入 U部分中。对于这类仅仅了解部分信息的网络，定义局部连接矩阵为

$$A_{ij} = \begin{cases} 1, & \text{顶点 } i,j \text{ 间存在连接且任意一个顶点属于 } C \\ 0, & \text{其他} \end{cases}$$

若将 C视为 G的一个局部模块，对于这种划分的一个简便衡量方式是 C内部连边占总体已知连边的比例，表示为 $\frac{\sum_{ij} A_{ij}\xi(i,j)}{\sum_{ij} A_{ij}} = \frac{1}{2m^*}\sum_{ij} A_{ij}\xi(i,j)$。其中 $m^* = \frac{1}{2}\sum_{ij} A_{ij}$为局部连接矩阵中的总边数，当顶点 i,j都属于 C时，$\xi(i,j)$等于 1，否则等于 0。当 C内部连边众多并且与 G中未知部分连接较少时，该值便会较大。依据该标准，当 $|C| \geq |U|$时，将 C视为局部模块的划分总是好的。

C中有这样一些顶点，它们至少有一条边与 U内的顶点相连。它们构成了 C的边界 B，如图 17所示。边界的连接矩阵表示为

$$B_{ij} = \begin{cases} 1, & \text{顶点 } i,j \text{ 间存在连接且任意一个顶点属于 } B \\ 0, & \text{其他} \end{cases}$$

对边界清晰度的衡量标准应当独立于边界所包围的社团的规模。直观上，划分恰当的社团其边界应当比较明显，即边界与网络未知部分的连边少而与社团内顶点的连边多，如图 17所示。因此，定义局部模块衡量标准 R为

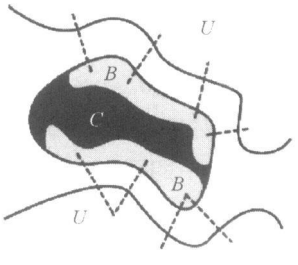


图 17 边界 B含义的形象表明[60]

$$R = \frac{\sum_{ij} B_{ij}\delta(i,j)}{\sum_{ij} B_{ij}} = \frac{I}{T}$$

其中, 当 $v_i \in B$, $v_j \in C$或 $v_j \in B$, $v_i \in C$时, $\delta(i, j)$等于 1, 否则等于 0. T为至少有一端属于 B的边的数目, I为两端都不属于 U的边的数目。

因为衡量局部模块的标准是针对单一模块, 所以网络中的社团是逐一划分出来的。算法的具体过程是:

1) 给定已知信息的顶点 v_0作为某一社团的源顶点。令 $v_0 = C$, 与之相连的顶点构成 U.

2) 将使 U值增大最多 (或减小最少) 的邻居顶点加入到 C中。

3) 向 U中加入新增的顶点, 更新 R值。

4) 重复 2)、3) 操作, 直到达到指定次数 k或完成了全体连通的部分。

Clauset A将该方法应用于经典人工网和亚马逊购物网。结果显示, 其分析经典人工网的能力与基于全局信息的 GN算法、边集聚系数算法及 Newman贪婪算法相近; 并能够从亚马逊网中摘录出一些性质完全不同的物品社团。

5　加权网上的社团结构问题

以往的研究大多都是建立在无权无向网的基础之上, 然而随着复杂网络研究的深入, 人们发现现实系统构成的网络其连接总是有权重中的, 并不是非是既否的关系。因此富含权重的网络才能反映出系统的本质。网络的权重对社团划分产生影响, 也成为值得探讨的问题。

5.1　加权网

实际系统形成的网络往往包含权重关系, 这一事实已经得到广泛认同, 比如在人际关系网中, 人与人的关系并非完全相同, 有些人是推心置腹的朋友, 而有些人只是泛泛之交; 在科学家合作关系网中, 科学家之间合作的密切程度也存在很大的差异, 有些是经常合作, 有些仅仅合作过一次。边的权重刻画了这些差异。较之无权网的抽象方式而言, 加权网的抽象方式更大限度上保留了系统的信息。

网络权重的增加使得在讨论网络中社团时不能仅仅考虑连接强度, 必须将边上的权重考虑进来。因此定量描述社团结构的模块化 Q函数应当包含边的权重[62]。含权重的 Q函数表达为

$$Q^w = \frac{1}{2T}\sum_{ij}\left[w_{ij} - \frac{T_i T_j}{2T}\right]\delta(\varsigma_i, \varsigma_j)$$

其中, w_{ij}为顶点 i, j间连边的权重, T_i为顶点 i的点权, $T_i = \sum_j w_{ij}$, $T = \frac{1}{2}\sum_{ij} w_{ij}$为网络中所有连边的权重之和, ς_i为顶点 i所属的社团。

5.2　算法推广

针对探索网络社团结构的研究, 网络权重的引入会影响到网络社团的定义。但权重的引入并不会改变探索网络社团结构的主要思路, 只需要在细节之处作适当的调整, 反映出加权网的特性即可。在谱分析算法和 Potts模型算法中, 用权重矩阵替代连接矩阵; 凝聚思想和分裂思想都与距离和相似度的定义有紧密的联系, 因此采用这种思想的算法要在定义距离和相似度时添加边的权重的影响; 优化算法的推广更为简便, 只需将优化目标替换为含权重的 Q函数。以下详细介绍 GN算法和 EO算法在加权网上的推广。

GN算法在加权网络上推广的关键在于如何计算加权网中的边介数[62]。边介数的定义是不变的, 区别在于顶点间距离的定义。一种显而易见的方法就是将距离看作权重的倒数, 从而两个顶点连接边的权重越大, 它们之间的距离也就越近, 这与人们的直观感受是相符合的。然而进一步研究会发现这种方法是不可取的。两点间权重越大, 它们连线的距离就越短, 就有越多的最短路径通过它们的连线, 因此它们连线的最短路径边介数就越大, 从而它们的连线就越早被移除。这就意味着连接越紧关系越密切的会被越早地断开, 与聚类的初衷完全相反。

正确的方法是将加权网转化为无权多图。加权网是用边的权重代表点与点之间的紧密程度; 而在无权多图中每条边的权重都是相同的, 点与点之间的紧密程度是由边的个数表示的。如图 18 它们的连接矩阵是相同的。

权重为 n的边与 n条权重为 1的平行边等价，加权网和无权多图可以相互替代。GN法运用于无权多图，与运用于与之相应的普通无权网相比，任意两点间的最短路径是不变的，然而由于重复边的存在，边介数的值发生了变化。如果两顶点间有两条边，则每条边的边介数是原值的一半；如果两顶点间有三条边，则每条边的边介数是原值的三分之一 …… 随后找出边介数最大的边将其断开，重新计算边介数，断开边介数最大的边，并重复此过程，便可划分出网络的社团结构。选择最优划分时，只需采用包含权重的 Q函数即可。

将 EO算法推广到加权网[63]时，用含权的 Q^w函数代替 Q函数作为全局目标。并将 Q^w函数改写为

$$Q^w = \sum_r \left(e_{rr}^w - (a_r^w)^2 \right)$$

其中，$e_{rr}^w = \frac{1}{2T_{ij}} \sum w_{ij}\delta(\sigma_i, r)\delta(\sigma_j, r)$为连接 r社团内部顶点的连边的权重之和占总权重的比例，$a_r^w = \frac{1}{2T_i} \sum T_i \delta(\sigma_i, r)$为社团 r内所有顶点的权重之和所占的比例。局部变量，即每个顶点对 Q^w的贡献变为：$q_i^w = T_{r(i)} - T_i a_{r(i)}^w$。其中 $T_{r(i)}$表示如果顶点 i属于 r社团，则其与 r社团内顶点构成连边的权重的总和，T_i为顶点 i的点权。将局部变量对全局变量的贡献标准化，得到顶点属于某个社团的适合度 $\lambda_i^w = \frac{q_i^w}{T_i} = \frac{T_{r(i)}}{T_i} - a_{r(i)}^w$。具体优化过程与无权网 EO算法相同。

5.3 权重对社团结构的影响

权重的引入会对网络社团结构产生多大的影响？一些学者以 GN算法、Potts模型算法和 EO算法在加权网中的推广为工具，对此进行了研究[63]。将经典人造网进行拓展，使其成为一个加权网，w_{inter}表示与社团外顶点连边的权重，w_{intra}表示与社团内顶点连边的权重。在顶点度值固定的情况下，检验权重变化对社团划分正确率的影响。如图 19，其中顶点向内部与向外部连边的平均数都等于 8，即 $\langle z_{in} \rangle = \langle z_{out} \rangle = 8$，连边的总权重给定 $\langle w_{inter} \rangle + \langle w_{intra} \rangle = 2$，但分布逐渐变化，竖轴为顶点被划分正确的比例。

第 4节具体介绍算法时大都进行了经典人造网的检验，当 $\langle z_{in} \rangle = \langle z_{out} \rangle = 8$时，除 EO算法的准确率还能达到 80%以外，其他算法准确率较低。然而加入权重后可以发现，各种算法的准确性都有所提高，当边权主要集中在每部连边上时，社团结构仍能被较为准确地划分出来。可见，权重的引入会对网络社团结构的划分产生较大影响。

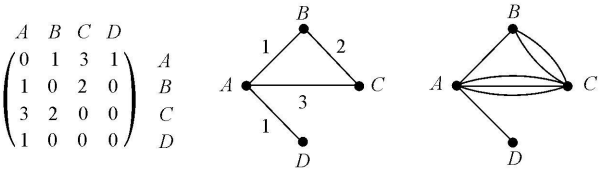


图 18 对应相同连接矩阵的加权网和无权多图[62]

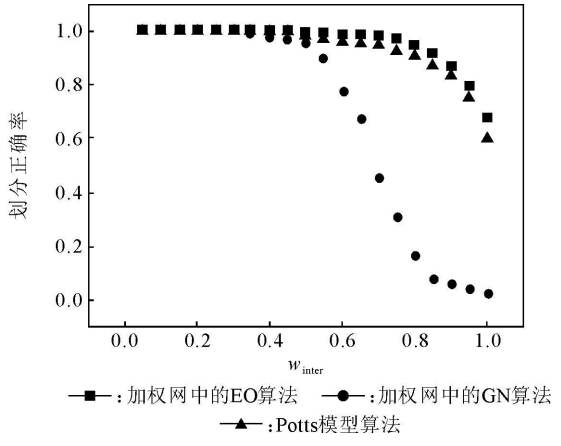


图 19 权重变化导致的社团结构划分正确率的变化[63]

6 总结与展望

本文对复杂网络社团结构研究的成果进行归纳，讨论了网络中社团的定义，着重介绍了有代表性的探索

网络社团结构的算法及网络社团结构的相关问题, 并关注了权重对网络社团结构的影响。

纵观众多划分网络中社团结构的算法, 有些算法的准确率比较高, 有些算法的复杂度比较低, 有些算法可以处理比较混乱的网络, 而有些算法只能应用于具有特定结构的网络。为了对算法特点有更为深刻的认识, 一些科研工作者对部分算法进行了横向的比较[29 64]。这里对文章第 4部分重点介绍的算法进行小结: 谱分析思想的算法适用于社团结构显著的网络, 且由于求特征值和特征向量的计算复杂耗时, 不适合分析规模大的网络; 层次距离算法的分析能力严重依赖于顶点间相似度及社团间相似度的定义; GN算法精确性高, 即算法包含的随机因素少, 但它的复杂度高且不适宜用于混乱的网络; 边集聚系数法较之 GN算法, 复杂度有显著降低, 但只能应用于簇系数比较大的网络; 信息集中性算法复杂度高, 不能应用于分析混乱的网络, 且易在划分初期产生孤立点; Potts模型算法中划分社团的操作采用模拟退火算法, 因此该算法的复杂度与系统降温速度及中止温度关系密切; 随机行走算法建立在层次算法的基础上, 其特点与层次算法相仿; 电流算法更适合于将网络划分成两个社团的问题; Newman贪婪算法较之 GN算法复杂度有显著降低, 但依然不适宜分析过于混乱的网络; 改进的贪婪算法进一步降低了复杂度, 使之接近于线性复杂度, 因此可以分析规模较为巨大的网络, 但分析混乱网络的能力没有提高; 极值优化算法的复杂度较之 GN算法有显著下降, 且适用于连接混乱的网络; Q函数与谱分析相结合的算法在有效划分网络社团结构的同时, 更细致地提供了顶点属于某个社团的倾向; 基于局部信息的方法适用于不了解全局信息的大规模网络。综上所述, 已有的社团结构划分方法各具特色, 仍有发展的空间。因此, 发展划分复杂网络社团结构的算法依然是值得关注的方向。而且, 针对一个具体的网络, 如何高效选取精确性和准确性高, 复杂度低的算法也是需要解决的问题。

网络社团结构与网络功能间的联系也是一个有价值的研究方向。Oh, E等学者在酵母网和 Internet网两个社团结构差异明显的无标度网络上研究了 Kuramoto改良模型的同步问题[65]。结果表明, 社团间的连接状况对同步具有影响: 当社团间的连接分散时同步过渡陡峭, 达到全局同步; 当社团间的连接集中时同步过渡平滑, 达到社团内部的同步。Park等学者为了探索社团结构对网络同步的影响, 构建并研究了社团间连接方式不同的多个网络[66]。他们发现社团间的随机连接和长程连接能够增强网络的同步能力, 而社团内部顶点间的连接对网络同步能力的影响不大。因此他们认为在社会网络中实现全局同步的策略是建立和强化远距离社团间的连接。严钢等学者应用 SIRS传染病模型探讨了网络社团结构对集群同步的影响[67]。他们应用 Q函数衡量网络的社团结构水平, 提出一种生成不同 Q值网络的方法。对这些生成网的研究发现, Q值小的网络有强同步能力, 该结论与 Donetti, L等人的结论[68]一致。周涛等学者也提出了一种社团结构强度可变的无标度网络增长模型, 并用社团间连边数目与社团内部连边数目的比值衡量社团结构的强度。基于这种网络对 Kuramoto模型相同步的研究发现, 在一定的社团强度之内, 网络的同步能力低于社团间没有连接的网络; 而社团强度超出某一域值后, 社团结构对网络同步能力的影响便会消失; 社团强度介于两个特殊值之间时, 社团结构越强网络的全局同步能力越弱[69]。另外, 有的研究中[70-71]还发现网络中拥有相同功能的顶点往往被归于同一个社团。这些研究结果不但指出了探索划分社团结构算法的新思路, 而且丰富了网络社团结构的意义。

有向网中社团结构的研究有着广泛的空间, 如何定义和划分有向网中的社团结构, 有向网中的环型结构与集团结构的关系, 以及如何探索有向网中的层次问题 (如食物链网络中的植食性动物层和肉食性动物层) 都是值得关注的问题。另外, 复杂网络社团结构问题的研究也可以推广到包含两类顶点的二分网。用二分网描述系统的优势在于: 减少了描述过程中的信息损失, 保留了更多的系统信息。并且众多的实际系统往往具有二分性, 例如图书借阅系统, 包含读者和图书这两类主体。对二分网社团结构的研究不但需要重新定义社团结构的概念, 还需要开发两类顶点协同考虑的搜索算法, 使得两类顶点被同时划分出社团结构。

复杂网络社团结构分析在实际问题中还没有得到广泛的应用, 仅仅涉及 Internet网、科学家合作网等领域, 但其在大脑系统、生物系统、经济系统、管理系统等领域的应用可能会揭示出以往方法未发掘的信息。不仅如此, 网络社团结构对实际问题还有着指导意义。因此, 社团结构对分析解决实际问题的价值需要进一步的探讨。

参考文献:

[1] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167—256.

[2] 吴金闪, 狄增如. 从统计物理看复杂网络研究[J]. 物理学进展, 2004, 24(1): 18—45.
Wu Jinshan, Di Zengru. Complex networks in statistical physics[J]. Progress in Physics, 2004, 24(1): 18—45.

[3] 汪秉宏, 周涛, 何大韧. 统计物理与复杂系统研究最近发展趋势分析[J]. 中国基础科学, 2005, 7(3): 37—43.
Wang Binghong, Zhou Tao, He Daren. The trend of recent research on statistical physics and complex systems[J]. China Basic Science, 2005, 7(3): 37—43.

[4] 周涛, 柏文洁, 汪秉宏, 等. 复杂网络研究概述[J]. 物理, 2005, 34(1): 31—36.
Zhou Tao, Bai Wenjie, Wang Binghong, et al. A brief review of complex networks[J]. Physics, 2005, 34(1): 31—36.

[5] 方锦清, 汪小帆, 郑志刚, 等. 一门崭新的交叉科学: 网络科学(上)[J]. 物理学进展, 2007, 27(3): 239—343.
Fang Jinqing, Wang Xiaofan, Zheng Zhigang, et al. New interdisciplinary science: network science(I)[J]. Progress in Physics, 2007, 27(3): 239—343.

[6] 方锦清, 汪小帆, 郑志刚, 等. 一门崭新的交叉科学: 网络科学(下)[J]. 物理学进展, 2007, 27(4): 361—448.
Fang Jinqing, Wang Xiaofan, Zheng Zhigang, et al. New interdisciplinary science: network science(II)[J]. Progress in Physics, 2007, 27(4): 361—448.

[7] Albert R, Jeong H, Barabasi A-L. Diameter of the world-wide web[J]. Nature, 1999, 401(6749): 130—131.

[8] Andrei B, Ravi K, Farzin M, et al. Graph structure in the Web[J]. Compuer Networks, 2000, 33: 309—320.

[9] Williams R J, Martinez N D. Simple rules yield complex food webs[J]. Nature, 2000, 404(6774): 180—183.

[10] Amaral L A N, Scala A, Barthelemy M, et al. Classes of small-world networks[J]. Proc Natl Acad Sci USA, 2000, 97(21): 11149—11152.

[11] Gleiser P, Danon L. Community structure in jazz[J]. Advances in Complex System, 2003, 6(4): 565—573.

[12] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc Natl Acad Sci USA, 2002, 99: 7821—7826.

[13] Redner S. How popular is your paper? An empirical study of the citation distribution[J]. Eur Phys J B, 1998, 4: 131—134.

[14] Gibson D, Kleinberg J, Raghavan P. Inferring web communities from link topology[C]. Proceedings of the 9th ACM Conference on Hypertext and Hypermedia. Pittsburgh: ACM Press, 1998: 225—234.

[15] Flake G W, Lawrence S R, Giles C L, et al. Self-organization and identification of web communities[J]. IEEE Computer, 2002, 35(3): 66—71.

[16] Adamic A L, Adar E. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3): 211—230.

[17] Arenas A, Diaz-Guilera A, Perez-Vicente C J. Synchronization reveals topological scales in complex networks[J]. Phys Rev Lett, 2006, 96(11): 114102.

[18] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. Proc Natl Acad Sci USA, 2004, 101: 2658—2663.

[19] Castellano C, Radicchi F. Self-contained algorithms to detect communities in networks[J]. Eur Phys J B, 2004, 38: 311—319.

[20] 解雪, 汪小帆. 复杂网络中的社团结构分析算法研究综述[J]. 复杂系统与复杂性科学, 2005, 2(3): 1—12.
Xie Zhou, Wang Xiaofan. An overview of algorithms for analyzing community structure in complex networks[J]. Complex Systems and Complexity Science, 2005, 2(3): 1—12.

[21] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814—818.

[22] Newman M E J. Detecting community structure in networks[J]. Eur Phys J B, 2004, 38: 321—330.

[23] Fortunato S, Barthelemy M. Resolution limit in community detection[J]. Proc Natl Acad Sci USA, 2007, 104: 36—41.

[24] Wasserman S, Faust K. Social Network Analysis[M]. Cambridge, UK: Cambridge Univ Press, 1994.

[25] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69(2): 026113.

[26] Park J and Newman M E J. The origin of degree correlations in the Internet and other networks[J]. Phys Rev E, 2003, 68(2): 026112.

[27] LiM H, Fan Y, Chen JW, et al. Weighted networks of scientific communication: the measurement and topological role of weight[J]. PhysicaA, 2005, 350: 643—656.
[28] Sade D S. Sociometrics of macacamulatta linkages and cliques in grooming matrices[J]. Folia Primatol, 1972, 18: 196—223.
[29] Danon L, Guilera A, Duch J, et al. Comparing community structure identification[J]. J StatMech, 2005, P09008.
[30] Kuncheva L I, Hadjitodorov S T. Using diversity in cluster ensembles[C]. 2004 IEEE International Conference Systems, Man and Cybernetics, 2004, 2: 1214—1219.
[31] Fred A L N, Jain A K. Robust data clustering[C]. Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, USA: IEEE, 2003, 2: 128—133.
[32] Zhang P, LiM H, Wu J S, et al. The analysis and dissimilarity comparison of community structure[J]. PhysicaA, 2006, 367: 577—585.
[33] Hall K M. An r-dimensional quadratic placement algorithm[J]. Management Science, 1970, 17: 219—229.
[34] Fiedler M. Algebraic connectivity of graphs[J]. Czech Math J, 1973, 23(98): 298—305.
[35] Pothen A, Simon H, Liou K-P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM J Matrix Anal Appl, 1990, 11(3): 430—452.
[36] Capocci A, Servedio V D P, Caldarelli G, et al. Detecting communities in large networks[J]. PhysicaA, 2005, 352: 669—676.
[37] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: structure and dynamics[J]. Phys Rep, 2006, 424: 175—308.
[38] Scott J. Social Network Analysis: A Handbook[M]. 2nd ed. London: Sage Publications, 2002.
[39] Latora V, Marchiori M. Efficient behavior of small-world networks[J]. Phys Rev Lett, 2001, 87: 198701.
[40] Latora V, Marchiori M. Economic small-world behavior in weighted networks[J]. Eur Phys J B, 2003, 32: 249—263.
[41] Latora V, Marchiori M. A measure of centrality based on the network efficiency[DB/OL]. [2007—12—18]. http://arxiv.org/abs/cond-mat/0402050.
[42] Fortunato S, Latora V, Marchiori M. Method to find community structures based on information centrality[J]. Phy Rev E, 2004, 70(5): 056104.
[43] Reichardt J, Bornholdt S. Detecting fuzzy community structures in complex networks with a Potts mode[J]. Phys Rev Lett, 2004, 93(21): 218701.
[44] Reichardt J, Bornholdt S. Statistical mechanics of community detection[J]. Phys Rev E, 2006, 74(1): 016110.
[45] Zhou H J. Network landscape from a Brownian particle's perspective[J]. Phys Rev E, 2003, 67(4): 041908.
[46] Zhou H J. Distance, dissimilarity index, and network community structure[J]. Phys Rev E, 2003, 67(6): 061901.
[47] Pons P, Latapy M. Computing communities in large networks using random walks[J]. LNCS, 2005, 3733: 284—293.
[48] Wu F, Huberman B A. Finding communities in linear time: a physics approach[J]. Eur Phys J B, 2004, 38: 331—338.
[49] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Phys Rev E, 2004, 69(6): 066133.
[50] Bak P, Sneppen K. Punctuated equilibrium and criticality in a simple model of evolution[J]. Phys Rev Lett, 1993, 71: 4083—4086.
[51] Boettcher S, Percus A G. Optimization with extremal dynamics[J]. Phys Rev Lett, 2001, 86: 5211—5214.
[52] Zhou T, Bai W J, Chen L J, et al. Continuous extremal optimization for Lennard-Jones clusters[J]. Phys Rev E, 2005, 72(1): 016702.
[53] Duch J, Arenas A. Community detection in complex networks using extremal optimization[J]. Phys Rev E, 2005, 72(2): 027104.
[54] Newman M E J. Modularity and community structure in networks[J]. Proc Natl Acad Sci USA, 2006, 103: 8577—8582.
[55] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Phys Rev E, 2006, 74(3): 036104.
[56] Josep M P, Javier B, Jordi D. Clustering algorithm for determining community structure in large networks[J]. Phys Rev E, 2006, 74(1): 016107.
[57] Zhang S H, Wang R S, Zhang X S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering[J]. Physica A, 2007, 374(1): 483—490.

[58] Yang B. Self-organizing network evolving model for mining network community structure[J]. LNCS 2006, 4093: 404—415.
[59] Yang S Z, Luo S W, Li J Y. A novel visual clustering algorithm for finding community in complex network[J]. LNCS 2006, 4093: 396—403.
[60] Clauset A. Finding local community structure in networks[J]. Phys Rev E, 2005, 72(2): 026132.
[61] Bagrow J P, Bollt E M. Local method for detecting communities[J]. Phys Rev E, 2005, 72(4): 046108.
[62] Newman M E J. Analysis of weighted networks[J]. Phys Rev E, 2004, 70(5): 056131.
[63] Fan Y, Li M H, Zhang P, et al. Accuracy and precision of methods for community identification in weighted networks[J]. Physica A, 2007, 377(1): 363—372.
[64] Mika G, Michael H, Anna L. Comparison and validation of community structures in complex networks[J]. Physica A, 2006, 367: 559—576.
[65] Oh E, Rho K, Hong H, et al. Modular synchronization in complex networks[J]. Phys Rev E, 2005, 72(4): 047101.
[66] Park K, Lai Y C, Gupte S, et al. Synchronization in complex networks with a modular structure[J]. Chaos, 2006, 16: 015105.
[67] Yan G, Fu Z Q, Ren J, et al. Collective synchronization induced by epidemic dynamics on complex networks with communities[J]. Phys Rev E, 2007, 75(1): 016108.
[68] Donetti L, Hurtado P I, Munoz M A. Entangled networks, synchronization, and optimal network topology[J]. Phys Rev Lett, 2005, 95(18): 188701.
[69] Zhou T, Zhao M, Chen G R. Phase synchronization on scale-free networks with community structure[J]. Phys Lett A, 2007, 368: 431—434.
[70] Zhou C S, Zemanova L, Zamora G, et al. Hierarchical organization unveiled by functional connectivity in complex brain networks[J]. Phys Rev Lett, 2006, 97(23): 238103.
[71] Zemanova L, Zhou C S, Kurths J. Structural and functional clusters of complex brain networks[J]. Physica D, 2006, 224: 202—212.
[72] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Phys Rev E, 2004, 70(6): 066111.