北京大学

博士研究生学位论文

题目：复杂网络社团结构的动力学方法研究

姓　　名：刘　健

学　　号：10601823

院　　系：数学科学学院

专　　业：计算数学

研究方向：科学计算与随机 PDE

导师姓名：鄂维南 教授，李铁军 教授

二〇一一 年 五 月

版权声明

复杂网络社团结构的动力学方法研究

刘　健　计算数学

导师姓名：　鄂维南 教授, 李铁军 教授

摘　要

复杂网络的现代科学理论为人们更好地理解复杂系统带来了重大变革. 近年来关于复杂网络结构和动力学方面的的研究工作大量涌现. 复杂网络模型已经成为物理学、生物学、计算机科学、社会学等很多领域中的流行工具. 网络可以表现真实系统的特征之一就是社团结构, 即整个网络由若干个社团构成, 属于相同社团的节点之间具有许多边相连接, 而属于不同社团的节点之间具有较少的边相连接. 这样的社团可以被看作是复杂网络的相当独立的分量, 起着类似于人体中的组织或者器官的作用. 由于许多真实的系统通常表示为网络的形式, 从而社团检测在很多学科中具有重要意义. 这一问题十分复杂, 尽管在过去几年中来自不同领域的科学家致力于此问题的研究, 并做出了大量相关工作, 但是至今仍未圆满解决.

在社团结构检测中, 动力学方法是一类强而有力的方法, 它的构造是基于研究在复杂网络上某种与结构相联系的动力学过程而建立的. 在这个分支中, 通常采用的模型是网络上的随机游动. 如果网络具有很强的社团结构, 那么随机游动者将花费很长的时间待在一个社团中, 这是由于社团内部的边的高密度进而导致路径数量也很大所引起的. 在本论文中, 作者主要研究的也是一类基于随机游动的动力学方法. 本论文的基石是最近由 E, Li 和 Vanden-Eijnden 发展的基于 Hilbert-Schmidt 度量粗粒化可逆马氏链的理论框架 (*Proc. Natl. Acad. Sci. USA* **105** (2008), 7907–7912), 作者进一步发展并完善了由此理论所建立的复杂网络社团结构的确定性分区方法, 所得到的主要创新成果如下:

(a) 提出了复杂网络社团结构的一个概率性框架, 其中每个节点以某一概率从属于某一个社团. 这可以看作是统计中的 fuzzy c-means 算法向网络分区问题的自然扩展, 也可以看做是之前的网络最优分区的确定性框架的推广. 提出的算法成功地应用于几个具有代表性的算例. 概率性框架为网络分区问题的研究提供更详细的信息. 更重要的是, 它比传统的网络确定性分区方法更具有预测性能.

(b) 设计了一个基于有效性指标 (validity index) 的方法来实现确定性分区的自动模型选择. 提出的有效性指标函数可以为社团结构的优良程度提供一种度量, 它是由两个因素的乘积所定义的, 分别是每个分区的社团内部紧密程度 (compactness) 与社团间分离程度 (separation). 数值试验表明算法在降温过程中可以有效找出社团结构, 并且无需任何关于社团结构的先验信息就可以自动确定社团的个数. 算法的 matlab 程序可以从网上免费下载使用, 下载链接为:

http://dsec.pku.edu.cn/~tieli/software/SAVI.zip.

(c) 分别利用结合了两种 k-means 迭代的模拟退火方法来最大化模量 (modularity), 以实现确定性分区的自动模型选择. 这两种 k-means 分别基于相异性指标和扩散距离. 算法可以得到较之许多已有方法更大的模量的值, 从这个意义上来说胜过了许多已有的方法. 算法不仅可以确定社团的个数以及社团结构, 还可以给出每个社团的中心节点.

(d) 构造了实现网络概率性分区的自动模型选择的方法. 提出了模糊模量 (fuzzy modularity) 函数, 它可以看作是传统模量的一个推广, 并为网络模糊社团结构的优良性提供了度量. 算法可以有效确定每个节点属于不同社团的概率, 并且初始的模糊分区可以随机选取, 社团的个数也可以自动确定而不再是将其固定为已知的模型参数.

本文提出的社团检测方法具有一般性, 可以推广到许多其它的复杂网络和数据集, 并且可以应用到更广泛的实际问题中.

关键词： 复杂网络, 社团结构, 动力学方法, 随机游动, 最优预测, Hilbert-Schmidt 范数, 有效性指标, 模量, 模糊模量, 模拟退火, k-means, fuzzy c-means, 自动模型选择

Dynamic Methods for Detecting Community Structure in Complex Networks

Jian Liu (Computational Mathematics)

Supervised by **Professors Weinan E and Tiejun Li**

Abstract

The theory of network science has significantly improved our understanding of complex systems. An explosive growth of interests and activities on the structure and dynamics of complex networks is appearing during recent years. Network models have also become popular tools in physics, biology, computer science, sociology, etc. One of the most relevant features of networks representing real systems is the community structure, i.e. the network consists of a number of communities, with many edges joining vertices in the same community and comparatively few edges joining vertices in different communities. Such communities can be considered as fairly independent compartments of a network, playing the similar role like, e.g., the tissues or the organs in the human body. Detecting communities is of great importance in disciplines where systems are often represented as networks. This problem is very hard and not yet satisfactorily solved, despite the huge efforts of the scientists in different research fields over the past few years.

Dynamic methods are powerful techniques for community structure detection and constructed by employing processes running on the network. Random walk is among the most popular models in this branch. If a network has strong modular structure, a random walker will spend a long time inside one community due to the

high density of internal edges and consequently large number of connecting paths. We will focus on such dynamic methods based on random walks in this thesis. The basis of this thesis is the theoretical framework for coarse graining irreversible Markov chain based on the Hilbert-Schmidt metric recently developed by E, Li and Vanden-Eijnden (*Proc. Natl. Acad. Sci. USA* **105** (2008), 7907–7912). We extend and develop the deterministic partitioning method within this framework. The main contributions of the thesis are as follows:

(a) We develop a probabilistic framework for network partition, in which each node has a certain probability of belonging to a certain community. It can be considered as a natural extension of the fuzzy c-means algorithm in statistics to the previous deterministic framework for optimal network partition. Algorithms are proposed and successfully applied to several representative examples. The probabilistic framework provides more detailed information to the network partitions, and it has more predictive power than the deterministic way.

(b) We design a validity index approach to resolve the model selection issue of deterministic clustering. The proposed validity index function can provide a measure of goodness for the community structure. It is defined as a product of two factors, involving the compactness and separation for each partition. It is demonstrated by experiments that the algorithm can efficiently find the community structure during the cooling process. The number of communities can be automatically determined without any prior knowledge about the community structure. A free matlab code can be downloaded from the website: http://dsec.pku.edu.cn/~tieli/software/SAVI.zip.

(c) We present new methods for automated model selection of deterministic clustering, in which the simulated annealing strategy is used to maximize the modularity function, associating with a dissimilarity-index-based and a diffusion-distance-based k-means iterative procedure, respectively. The proposed algorithms outperform most existing methods in the literature as regards the optimal modu-

larity found. They can not only identify the community structure, but also give the central node of each community.

(d) We propose a new method for automated model selection of probabilistic clustering. The fuzzy modularity, which can be considered as an extended version of traditional modularity, is developed to provide a measure of goodness for the fuzzy community structure in networks. The present algorithm can efficiently identify the probabilities of each node belonging to different communities with random initial fuzzy partition. An appropriate number of communities can be automatically determined without fixing it as a known model parameter.

The community detection methods proposed in this thesis are quite general, and can be extended to many other complex networks and data sets in real-world problems.

Keywords: complex networks, community structure, dynamic methods, random walk, optimal prediction, Hilbert-Schmidt norm, validity index, modularity, fuzzy modularity, simulated annealing, k-means, fuzzy c-means, automated model selection

目　录

表　格

插　图

第一章　绪　论

复杂网络理论为人们研究复杂系统提供了重要意义. 复杂网络 (complex network) 是复杂系统的主要表现形式, 而真实的复杂系统的特征之一就是社团结构 (community structure), 它们是复杂网络的独立的组成部分, 例如人体的组织或者器官. 研究复杂网络的社团结构在物理学、社会学、生物学、计算机科学以及其他的学科都有重要的意义.

目前关于复杂网络的结构和动力学的研究工作呈现出突飞猛进的增长, 这主要是由于受到统计力学的观点的启发以及来自新型复杂网络模型研究的需要[4, 16, 22, 138, 143]. 复杂网络已经成为计算机网络 (www 网络, Internet 网络), 电网, 交通网络, 通讯网络, 社会学网络 (疾病传播, 科学文章引用), 生物学网络 (细胞网络, 生态学网络, 蛋白质折叠, 神经网络), 经济学系统以及银行系统的重要模型工具[1, 13, 63, 68, 162]. 另一方面, 计算机视觉和数据挖掘的近期进展表明, 将数据集或图像视为一个复杂网络可以较好的从中析取出关于此数据集或图像的某些重要特征的信息[107, 132, 182]. 从数学的观点看, 复杂网络的研究本质上是关于图上的动力系统的研究. 这里的图可以是确定性的, 也可以是随机图; 可以是大小固定的, 也可以是随时间演化的. 这里的动力系统可以是确定性的, 也可以是随机的. 由于真实网络十分复杂, 故如何将其约化成较为简单的系统, 即社团结构的检测, 具有重要意义[70]. 从广义的角度看, 这与微分方程的模型约化理论具有紧密联系[174].

1.1　复杂网络的研究背景及意义

地球上任意两个人之间要通过多少个朋友才能互相认识? 万维网 WWW 上从一个页面到另一个页面平均需要点击多少次鼠标? 层出不穷的计算机病毒是如何在互联网 Internet 上传播的? 各种传染病 (艾滋病, 非典型性肺炎和禽流感等) 是如何在人类和动物中流行的? 为什么流言蜚语会散布得很快? 全球或地区性金融危机是如何发生的? 局部故障时如何触发大面积停电事故的? 大城市的交通堵塞

问题是如何引起的? 应该如何建立合理的公共卫生与安全网络? 为什么大脑能够具有思维的功能? 这些问题尽管看上去各不相同, 但是每一个问题中都涉及很复杂的网络, 包括 WWW, Internet, 社会关系网络, 经济网络, 电力网络, 交通网络, 神经网络等等. 更为重要的是, 越来越多的研究表明, 这些看上去各不相同的网络之间有着许多惊人的相似之处[199].

20 世纪 90 年代以来, 以 Internet 为代表的信息技术的迅猛发展使人类社会大步迈入了网络时代. 从 Internet 到 WWW, 从大型电力网络到全球交通网络, 从生物体中的大脑到各种新陈代谢网络, 从科研合作网络到各种经济, 政治, 社会关系网络等, 可以说, 人们已经生活在一个充满着各种各样的复杂网络的世界中. 人类社会的网络化是一把“双刃剑”: 它既给人类社会生产与生活带来了极大的便利, 提高了人类生产效率和生活质量, 但也给人类社会生活带来了一定的负面冲击, 如传染病和计算机病毒的快速传播以及大面积的停电事故等[199]. 因此, 人类社会的日益网络化需要人类对各种人工和自然的复杂网络的行为有更好的认识. 长期以来, 通信网络, 电力网络, 生物网络, 和社会网络等分别是通信科学, 电力科学, 生命科学, 和社会学等不同学科的研究对象, 而复杂网络理论所要研究的则是各种看上去互不相同的复杂网络之间的共性和处理它们的普适方法[199]. 从 20 世纪末开始, 复杂网络研究正渗透到数理学科, 生命学科和工程学科等众多不同的领域, 对复杂网络的定量与定性特征的科学理解, 已成为网络时代科学研究的一个极其重要的挑战性课题, 甚至被称为“网络的新科学 (new science of networks)”[11, 200].

以生命科学为例, 20 世纪的生命科学研究主流是建立在还原论基础上的分子生物学. 还原论的基本前提是, 在由不同层次组成的系统内, 高层次的行为是由低层次的行为所决定的. 具有还原论观点的生物学家通常认为, 只要认识了构成生命的分子基础 (如基因和蛋白质) 就可以理解细胞或个体的活动规律, 而组分之间的相互作用常常被忽略不计. 尽管基于还原论的分子生物学极大地促进了人类对单个分子功能的认识, 然而绝大多数生物特征都来自于细胞的大量不同组分, 如蛋白质, DNA, RNA 和小分子之间的交互作用[199]. 对这些极其复杂的交互作用网络的结构和动力学的理解已成为 21 世纪生命科学的关键性研究课题和挑战之一[95, 96, 128].

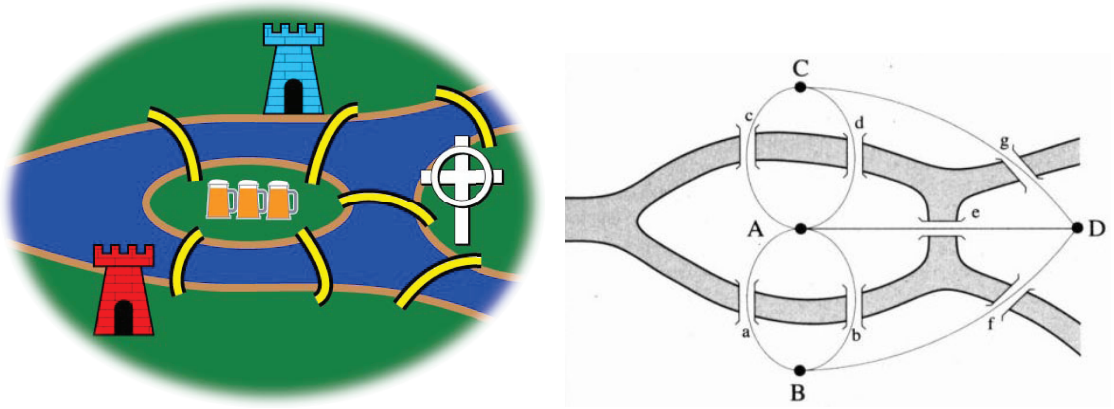
许多真实系统都可以用网络的形式加以描述, 一个典型的网络是由许多节点与链接节点之间的边组成的. 节点代表系统中的个体, 边则表示节点之间的作用关系. 如 WWW 网络可以看成是网页之间通过超链接构成的网络; Internet 网络可以看作不同的计算机通过光缆链接构成的网络; 科学家合作网络可以看作不同的科学家合作关系构成的网络; 基因调控网络可以看作是不同的基因通过调控与被调控关系构成的网络. 这些真实网络的普遍存在, 促使来自不同学科领域的科学家共同致力于复杂网络的研究. 这些学科领域包括复杂性科学, 数学, 物理, 生物, 计算机等. 复杂网络的研究可以使人们更好的了解现实世界的复杂系统, 为设计具有良好性能的网络提供依据[4, 16, 22, 138, 143].

1.2 复杂网络的研究简史

近年来复杂网络研究的兴起, 使得人们开始广泛关注网络结构复杂性及其网络行为之间的关系. 要研究各种不同的复杂网络在结构上的共性, 首先需要有一种描述网络的统一工具. 这种工具在数学上称为图 (graph). 任何一个网络都可以看作是由一些节点按某种方式连接在一起而构成的一个系统. 具体网络的抽象图表示, 就是用抽象的点表示具体网络中的节点, 并用节点之间的连线来表示具体网络终结点之间的连接关系.

实际的网络的图表示方法可以追溯到 Euler 对著名的七桥问题的研究[65]. Konigsberg 镇是东普鲁士的一个城镇, 城中有横贯城区的河流, 河中有两个岛, 两岸和两岛之间共架七座桥, 如图 1.1(a) 所示. 人们常常议论这样一个问题: 一个人能否再一次散步中走过所有的七座桥, 而且每座桥只经过一次, 最后返回原地? 1736 年, Euler 利用数学抽象法, 将七桥问题转化成如下数学问题: 从图 1.1(b) 中任意一点出发, 经过每条边一次而后返回原点的回路是否存在? 他给出了存在这样一条回路的充要条件, 并由此推得上述七桥问题无解. Euler 对于七桥问题的论证开创了图论 (graph theory) 的研究, 图 1.1(b) 也称为 Euler 图.

20 世纪 60 年代, 由 Erdös 和 Rényi 建立的随机图理论 (random graph theory) 在数学上开创了复杂网络理论的系统性研究[64]. Erdös 和 Rényi 研究的随机图模型 (ER 随机图) 中, 任意两个节点之间有一条边相连接的概率都为 p. 因此, 一个含



(a) (b)

图 1.1: (a) 1736 年的 Konigsberg 镇. Konigsberg 镇是东普鲁士 (现俄罗斯) 的一个城镇, 城中有横贯城区的河流, 河中有两个岛, 两岸和两岛之间共架七座桥. (b) Konigsberg 七桥问题的图示. Euler 将被河流分隔开的四块陆地抽象为四个点, 分别用 A, B, C, D 表示, 而将连接这四块陆地之间的七座桥抽象为连接四个点的七条线, 分别用 a, b, c, d, e, f, g 表示, 这样就得到了由四个点和七条线构成的一个图.

n 个节点的 ER 随机图中边的总数的期望为 $pn(n-1)/2$. 由此可得, 产生一个有 n 个节点和 m 条边的 ER 随机图的概率为 $p^m(1-p)^{n(n-1)/2-m}$. Erdös 和 Rényi 系统性地研究了当 $n \to \infty$ 时 ER 随机图的性质 (如连通性等) 与概率 p 之间的关系. 设集合每一个 ER 随机图都具有某种性质, 如果当 $n \to \infty$ 时产生具有这种性质的 ER 随机图的概率为 1. Erdös 和 Rényi 的最重要的发现是: ER 随机图的许多重要性质都是突然涌现的. 也就是说, 对于任意的概率 p, 要么几乎每一个图都具有某种性质 (例如连通性), 要么几乎每一个图都不具有该性质.

20 世纪 60 年代美国哈佛大学的社会心理学家 Stanley Milgram 在 [135] 中描述了一份信件通是如何仅用 3 步就从堪萨斯州的一位农场主手中转交到马萨诸塞州的一位神学院学生的妻子手中的. 尽管并不是每一个实验对象都如此成功, 但 Milgram 根据最终到达目标者手中的信件的统计分析发现, 从一个志愿者到其目标对象的平均距离只是 6. 实验结果在某种程度上反映了人际关系的“小世界”特征. 这就是复杂网络的小世界效应 (small-world effect), 也就是著名的六度分离 (six degrees of separation) 推断, 如图 1.2 所示. 20 世纪 60 年代末, 哈佛大学研究

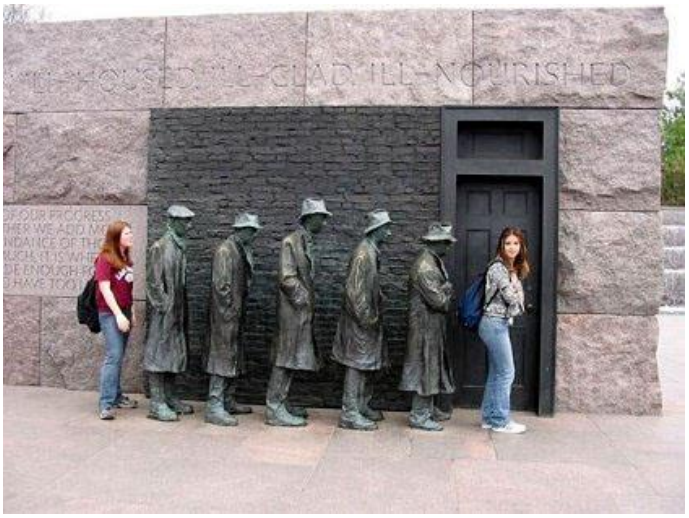


图 1.2: 小世界效应 (六度分离) 的图示. Milgram 给出的推断是: 地球上任意两个人之间的平均距离是6. 也就是说, 平均中间只要通过5个人, 你就能与地球上任何一个角落的任何一个人发生联系[135].

生 Mark Granovetter 通过研究发现, 人们在找寻工作时, 那些关系紧密的朋友 (强连接) 反倒没有那些关系一般的甚至只是偶尔见面的朋友 (弱连接) 更能够发挥作用, 事实上, 关系紧密的朋友也许根本帮不上忙. 于是他提出了著名的弱连接强度的理论[81], 现已被认为是最有影响的社会学贡献之一.

在 20 世纪末, 对复杂网络的科学探索发生了重要的转变, 复杂网络理论研究不再局限于数学领域. 人们开始考虑节点数量众多, 连接结构复杂的实际网络的整体特性, 在从物理学到生物学的众多学科中掀起了研究复杂网络的热潮. 有两个开创性的工作可以看作是复杂网络研究新纪元开始的标志: 一个是 Watts 和 Strogatz 发表于 *Nature* 的工作 [201], 另一个是 Barabási 和 Albert 发表于 *Science* 的工作 [12]. 这两个工作分别揭示了复杂网络的小世界特征和无标度性质, 并建立了相应的模型以阐述这些特性的产生机理. 复杂网络研究的简单历史见表 1.1.

1.3 复杂网络及其基本性质

一个具体的网络可抽象为一个由节点 (vertex, 或 node) 集合 V 和边 (edge) 集合 E 组成的图 $G = (V, E)$, 如图 1.3 所示. 节点数记为 $n = |V|$, 边数记为 $m = |E|$.

表 1.1: 复杂网络研究的简史[199].

时间 (年)	人物	事件
1736	Euler	七桥问题
1959	Erdös 和 Rényi	随机图理论
1967	Milgram	小世界实验
1973	Granovetter	弱连接强度
1998	Watts 和 Strogatz	小世界模型
1999	Barabási 和 Albert	无标度网络

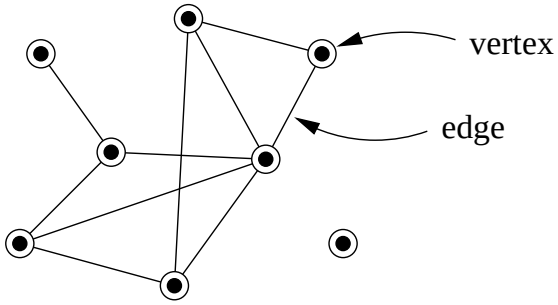


图 1.3: 由 8 个节点和 10 条边构成的小网络的例子[138].

E 中每条边都有 V 中一对点与之相对应. 如果任意点对 (i,j) 与 (j,i) 对应同一条边, 则该网络称为无向网络 (undirected network), 否则称为有向网络 (directed network). 如果给每条边都赋予相应的权值, 那么该网络就称为加权网络 (weighted network), 否则称为无权网络 (unweighted network). 当然无权网络也可以看作是每条边的权重都为 1 的等权网络. 此外一个网络中还可能包含多种不同类型的节点. 例如, 在社会关系网络中可以用权重表示两个人的熟悉程度, 而不同类型的节点可以代表具有不同国籍, 地区, 年龄, 性别和收入的人. 图 1.4 给出了几个不同类型的网络的例子. 本文主要介绍的是无向网络, 并且假设没有重边和自环 (即任意两个节点之间至多只有一条边, 且没有以同一个节点为起点和终点的边). 在图论中, 没有重边和自环的图称为简单图 (simple graph). 关于图论中基本要素的较为详细的介绍参见附录 A. 这里仅介绍一些基本的概念和性质.

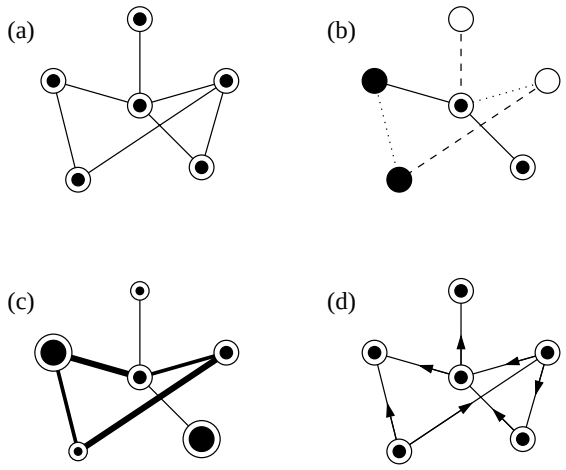


图 1.4: 不同类型网络的例子[138]. (a) 单一类型节点和和边构成的无向网络. (b) 不同类型节点和边构成的无向网络. (c) 节点和边权重变化所构成的无向网络. (d) 有向网络.

节点 i 的度 (degree) 定义为与该节点连接的其它节点的数目, 记为 $d(i)$. 有向网络中的一个节点的度分为出度 (out-degree) 和入度 (in-degree). 节点的出度是指从该节点指向其它节点的边的数目, 节点的入度是指从其它节点指向该节点的边的数目. 直观上看, 一个节点的度越大就意味着这个节点在某种意义上越重要. 网络中所有节点 i 的度 $d(i)$ 的平均值称为网络的 (节点) 平均度, 记为 $\langle d\rangle$. 网络中的两个节点 i 和 j 之间的最短路径 (shortest path, 或 geodesic) 定义为它们之间长度最短的路径, 记为 s_{ij}. 网络中任意两个节点之间的最短路径的最大值称为网络的直径 (diameter), 记为 $D = \max_{ij} s_{ij}$.

复杂网络的一个重要的局部定义是聚集系数 (clustering coefficient), 它描述了与一个节点相连的节点之间也相连的属性[201]. 一般地, 假设节点 i 的度为 $d(i)$, 则称与之相连的 $d(i)$ 个节点为节点 i 的邻居. 显然在这 $d(i)$ 个节点之间至多有 $d(i)(d(i)-1)/2$ 条边, 而它们之间实际存在的边数 $E(i)$ 和总的可能的边数之比就定义为节点 i 的聚集系数 $C(i)$, 即

$$C(i) = \frac{2E(i)}{d(i)(d(i)-1)} = \frac{\text{与节点 } i \text{ 相连的三角形数量}}{\text{与节点 } i \text{ 相连的三元组数量}}, \tag{1.1}$$

其中与节点 i 相连的三元组是指包括节点 i 的三个节点, 并且至少存在从节点 i 到

其它两个节点的两条边. 整个网络的聚集系数 C 定义为

$$C = \frac{1}{n}\sum_{i=1}^{n} C(i). \tag{1.2}$$

显然有 $0 \le C \le 1$. $C = 0$ 当且仅当所有的节点均为孤立节点, 即没有任何连接边; $C = 1$ 当且仅当网络是完全图, 即网络中任意两个节点都有边连接. 对于一个 n 个节点的完全随机网络, 当 n 很大时, $C = O(n^{-1})$. 而许多大规模的实际网络都具有明显的聚集效应, 它们的聚集系数尽管远小于 1, 但却比 $O(n^{-1})$ 要大很多. 事实上, 对于很多实际中的网络, 当 $n \to \infty$ 时, $C = O(1)$. 这意味着实际的复杂网络在某种程度上具有"物以类聚, 人以群分"的特性[138].

下面将主要介绍真实世界中复杂网络的三个基本性质, 包括小世界性质, 无标度性质以及社团结构性质.

1.3.1 小世界性质

一个社会网络就是一群人或团体按某种关系连接在一起而构成的一个系统. 这里的关系可以多种多样, 如个人之见的朋友关系, 同事之间的合作关系, 家庭之间的联姻关系和公司之间的商业关系等等. 以朋友关系为例, 很多人可能都有这样的经历: 偶尔碰到一个陌生人, 同他聊了一会儿后发现你认识的某个人居然他也认识, 然后一起发出"这个世界真小"的概叹. 那么对于地球上任意两个人来说, 借助第三者, 第四者这样的间接关系来建立起他们两人的联系, 平均需要通过多少人呢? 20 世纪 60 年代美国哈佛大学的社会心理学家 Stanley Milgram 的通过一些社会调查后给出的推断是: 地球上任意两个人之间的平均距离是 6. 也就是说, 平均中间只要通过 5 个人, 你就能与地球上任何一个角落的任何一个人发生联系. 这就是小世界效应 (small-world effect), 也就是著名的六度分离 (six degrees of separation) 推断[135].

现代版本则是哥伦比亚大学用 E-mail 进行的类似的小世界实验[51]. 2001 年, 哥伦比亚大学社会学系的 Watts 主持了一项最新的对六度分离理论的验证工程, 并建立了一个称为小世界项目 (small world project) 的网站①. 166 个不同国家的 6

① http://smallworld.columbia.edu/

万多名志愿者参加了该研究. Watts 随机选定 18 名目标, 要求志愿者选择其中的一名作为自己的目标, 并发送 Email 给自己认为最有可能发送邮件给目标的亲友. 最终的实验结果表明邮件要达到目标, 平均也只要经历 5 至 7 个人左右.

在 [138] 中, 定义网络的平均路径长度 l 为任意两个节点之间的最短路径的平均值, 也称网络的特征路径长度, 即

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} s_{ij}. \tag{1.3}$$

为了便于处理, 上式包含了节点到自身的最短路径 (为 0); 如果不考虑节点到自身的最短路径, 需在上述右端乘以 $(n+1)/(n-1)$. 在实际应用中, 这么小的差别完全可以忽略不计. 一个含有 n 个节点和 m 条边的网络的平均路径长度可以用计算量为 $O(mn)$ 的广度优先搜索算法来确定[2]. 尽管许多实际的复杂网络的节点数巨大, 但是网络的平均路径长度却小的惊人. 具体地说, 一个网络称为是有小世界效应的, 如果对于固定的网络节点平均度 $\langle d \rangle$, 平均路径长度 l 的增加速度至多与网络规模 n 的对数成正比.

小世界效应对于网络上发生的过程的动力学具有显而易见的含义. 例如, 如果考虑网络上信息或任何事物的传播, 小世界效应意味着在多数真实世界网络中传播是非常快速的. 如果一个谣言从一个人传播到另一个人仅需 6 步, 那么这将比传播途中需要一百步或者一百万步的情况要快得多. 这影响了 Internet 上一个信息包从一台计算机到另一台所需的跳跃数, 一个飞机或火车乘客旅程的段数, 一种疾病传播遍及整个人口的时间等等. 小世界效应也成为了一些著名的室内游戏的基础, 特别是 Erdös 数[46]和 Bacon 数②的计算.

1.3.2 无标度性质

网络中节点的度的分布情况可用度分布函数 (degree distribution function) $P(d)$ 来描述, 它表示的是一个随机选定的节点的度恰好为 d 的概率. 规则的格子具有简单的度序列, 因为所有节点具有相同的度, 所以其度分布为 Delta 分布, 它是个单个尖峰. 网络中任何随机化倾向都将使这个尖峰的形状变宽. 完全随机图的度分

②http://www.cs.virginia.edu/oracle/

布近似为 Poisson 分布, 其形状在远离峰值 $\langle d\rangle$ 处呈指数下降. 这意味着当 $d \gg \langle d\rangle$ 时, 度为 d 的节点实际上是不存在的. 因此, 这类网络也称均匀网络 (homogeneous network). 然而近几年的研究表明, 许多实际网络的度分布明显不同于 Poisson 分布, 特别是许多网络的度分布可以用幂律 (power law) 形式 $P(d) \propto d^{-\gamma}$ 来更好地描述[4, 138, 187]. 幂律分布曲线比 Poisson 分布曲线下降要缓慢得多.

幂律 (power law) 分布也称为无标度 (scale-free) 分布, 具有幂律度分布的网络也称为无标度网络, 这是由于幂律分布函数具有无标度性质③. 在一个度分布为具有适当幂指数 (通常为 $2 \leq \gamma \leq 3$) 的幂律形式的大规模无标度网络中, 绝大部分的节点的度相对很低, 但存在少量的度相对很高的节点. 因此, 这类网络也称非均匀网络 (inhomogeneous network)[5], 而那些度相对很高的节点称为网络的集线器 (hub)[4, 187].

另外一种表示度数据的方法是绘制累积度分布函数 (cumulative degree distribution function)

$$P_d = \sum_{d'=d}^{\infty} P(d'), \tag{1.4}$$

它表示的是度不小于 d 的节点的概率分布. 如果度分布为幂律分布, 即 $P(d) \propto d^{-\gamma}$, 那么累积度分布函数服从 $\gamma - 1$ 的幂律

$$P_d \propto \sum_{d'=d}^{\infty} d'^{-\gamma} \propto d^{-(\gamma-1)}. \tag{1.5}$$

如果度分布为指数分布, 即 $P(d) \propto \mathrm{e}^{-d/\kappa}$, 其中 $\kappa > 0$ 是一常数, 那么累积度分布函数也是指数分布, 且具有相同的指数

$$P_d \propto \sum_{d'=d}^{\infty} \mathrm{e}^{-d'/\kappa} \propto \mathrm{e}^{-d/\kappa}. \tag{1.6}$$

幂律分布在对数坐标系中对应于一条直线, 而指数分布在半对数坐标系中对应于一条直线, 因此分别通过采用对数坐标和半对数坐标就可以很容易识别幂律和指

③幂律分布函数的无标度性质: 考虑一个概率分布函数 $f(x)$, 如果对任意给定常数 a, 存在常数 b 使得函数 $f(x)$ 满足如下的无标度条件 $f(ax) = bf(x)$, 那么必有 (假定 $f(1)f'(1) \neq 0$)

$$f(x) = f(1)x^{-\gamma}, \quad \gamma = -\frac{f(1)}{f'(1)}.$$

也就是说, 幂律分布函数是唯一满足无标度条件的概率分布函数[4, 187].

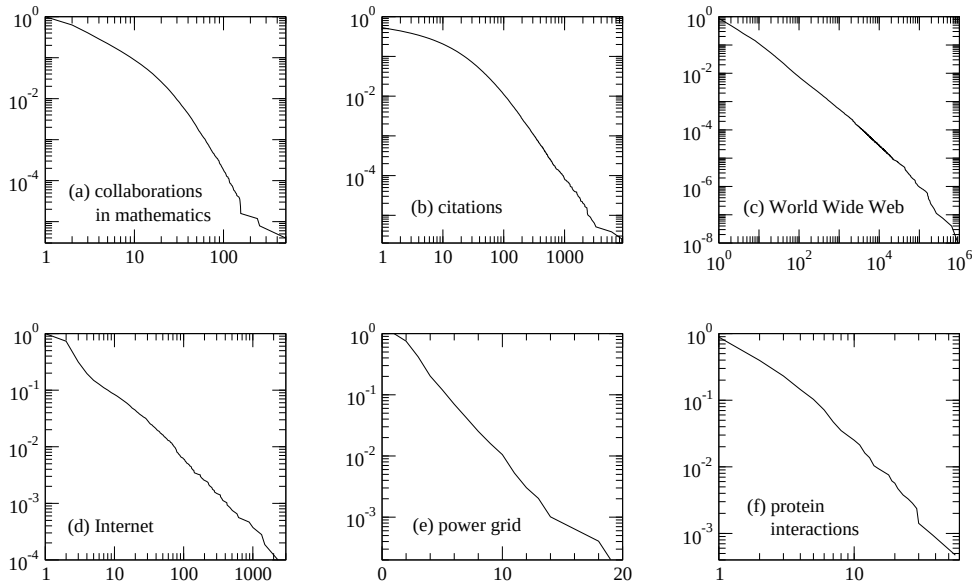


图 1.5: 六种网络的累积度分布曲线[138]. (a) 数学合作网络. (b) 引用网络. (c) WWW 子网. (d) Internet. (e) 电力网络. (f) 蛋白质相互作用网络.

数分布. 图 1.5 给出了一些网络的累积度分布曲线, 横轴是节点的度 (对于有向的引用网络和 WWW 子网, 表示入度), 纵轴是累计度分布. 其中 (a) 对应于数学合作网络, (b) 是 1981 至 1997 年间 *Institute for Scientific Information* 上发表的文献之间的引用网络, (c) 为1999 年的 WWW 的一个拥有 3 亿个节点的子网, (d) 对应于 1999 年 4 月的自治 (autonomous system, AS) 层的 Internet, (e) 表示美国西部电力网络, (f) 表示酵母菌代谢网络中的蛋白质相互作用网络. 曲线 (c), (d) 和 (f) 服从幂律, 分布曲线在对数坐标系中基本为直线形式; (b) 只在末端服从幂律; (e) 服从指数分布 (半对数坐标); (a) 看上去像两个不同指数的幂律曲线的组合.

1.3.3 社团结构性质

随着对网络性质的物理意义和数学特性的深入研究, 人们发现许多实际网络都具有一个共同性质, 即社团结构 (community structure). 也就是说, 整个网络是由若干个“群 (group)”或“团 (cluster)”构成的. 每个社团内部的节点之间的连接相对非常紧密, 但是各个社团之间的连接相对来说却非常稀疏, 如图 1.6 所示. 图中的网络包含三个社团, 分别对应于三个虚线圆圈包围的部分. 在这些社团内部, 节

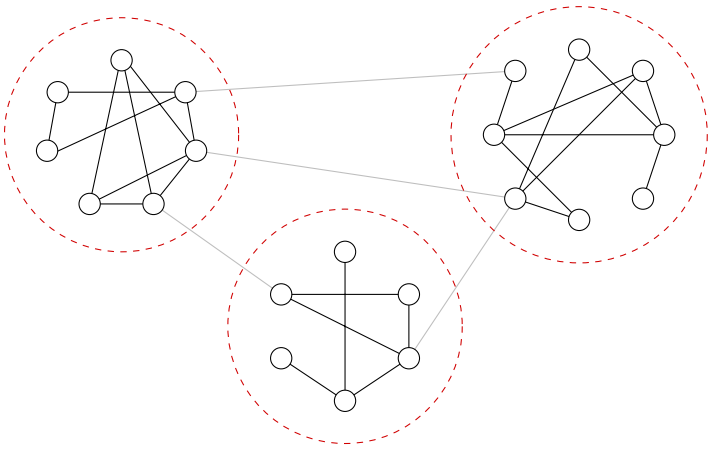


图 1.6: 一个小型的具有社团结构性质的网络示意图[139]. 图中的网络包含三个社团, 分别对应于三个虚线圆圈包围的部分. 在这些社团内部, 节点之间的联系非常紧密, 而社团之间的联系就稀疏得多.

点之间的联系非常紧密, 而社团之间的联系就稀疏得多.

一般而言, 社团可以包含模块 (module), 类 (class), 群 (group), 团 (cluster) 等各种含义. 例如, WWW 可以看成是由大量网站社团组成的, 其中同一个社团的各个网站所讨论的都是一些具有共同兴趣的话题[1, 68]. 类似地, 在生物网络或者电路网络中, 同样可以将各个节点根据其不同的性质划分为不同的社团[136, 181]. 揭示复杂网络中的社团结构, 对于了解网络结构与分析网络特性都是很重要的. 社团结构分析在生物学, 物理学, 计算机图形学和社会学中都有广泛的应用[70, 77, 139, 144].

真实世界中的复杂网络社团结构如图 1.7, 1.8, 1.9 和 1.10 所示. 图 1.7 给出了蛋白质之间相互作用网络的社团结构, 这是一只老鼠的癌细胞中蛋白质之间的相互作用[97], 由 Palla 等人提出的派系过滤算法[151]所得到的社团标记为不同的颜色. 图 1.8 展示了进行网络研究的物理学家之间的合作网络的最大分量的社团结构, 这个网络由出现在 [138] 的冗长的参考文献里的作者名字所构成, 不同的颜色表示用 [144] 中的算法的最短路径介数 (betweenness) 形式所得到的最优分区的社团. 图 1.9 描述的是查尔斯 • 狄更斯 (Charles Dickens) 的小说《大卫 • 科波菲尔》(*David Copperfield*) 中通常出现的英语词汇之间的邻接网络应用 [141] 中的算法得到的社团结构[141], 其中圆形表示小说中的形容词, 方形表示名词. 图 1.10 画出了一个网

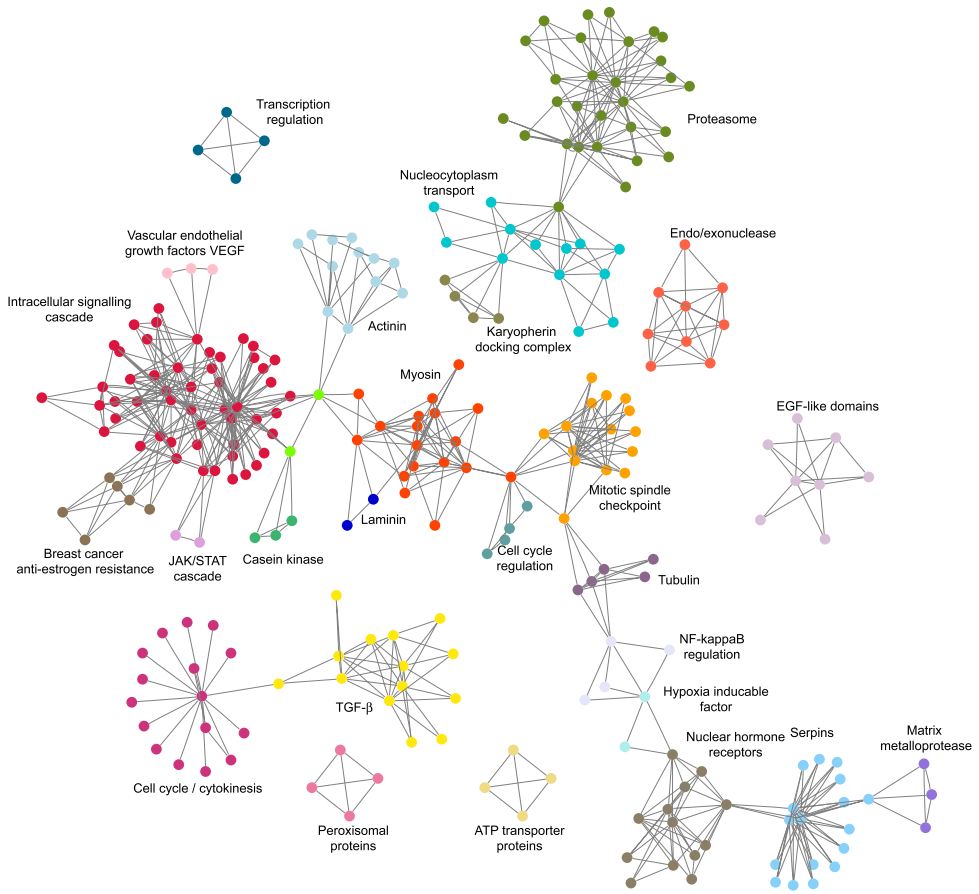


图 1.7: 蛋白质之间相互作用网络的社团结构[97]. 图中画出了一只老鼠的癌细胞中蛋白质之间的相互作用. 由 Palla 等人提出的派系过滤算法[151]所得到的社团标记为不同的颜色.

站的网页与它们之间的超链接所构成的网络的社团结构, 这里也采用 [144] 中的算法的最短路径介数 (betweenness) 形式所得到的结果.

本文即将探索研究诸如上述例子的复杂网络社团结构的算法.

1.4 研究复杂网络社团结构的主要算法

历史中研究复杂网络社团结构的算法层出不穷, 如图 1.11 所示, 基本问题包括传统方法, 模量最大化方法, 动力学方法以及基于统计推断的方法. 其它类型的问题的方法, 如检测重叠社团的方法, 多尺度分级聚类方法, 检测动态社团的方法等等在此不再赘述[70]. 本节内容主要参考 [70].

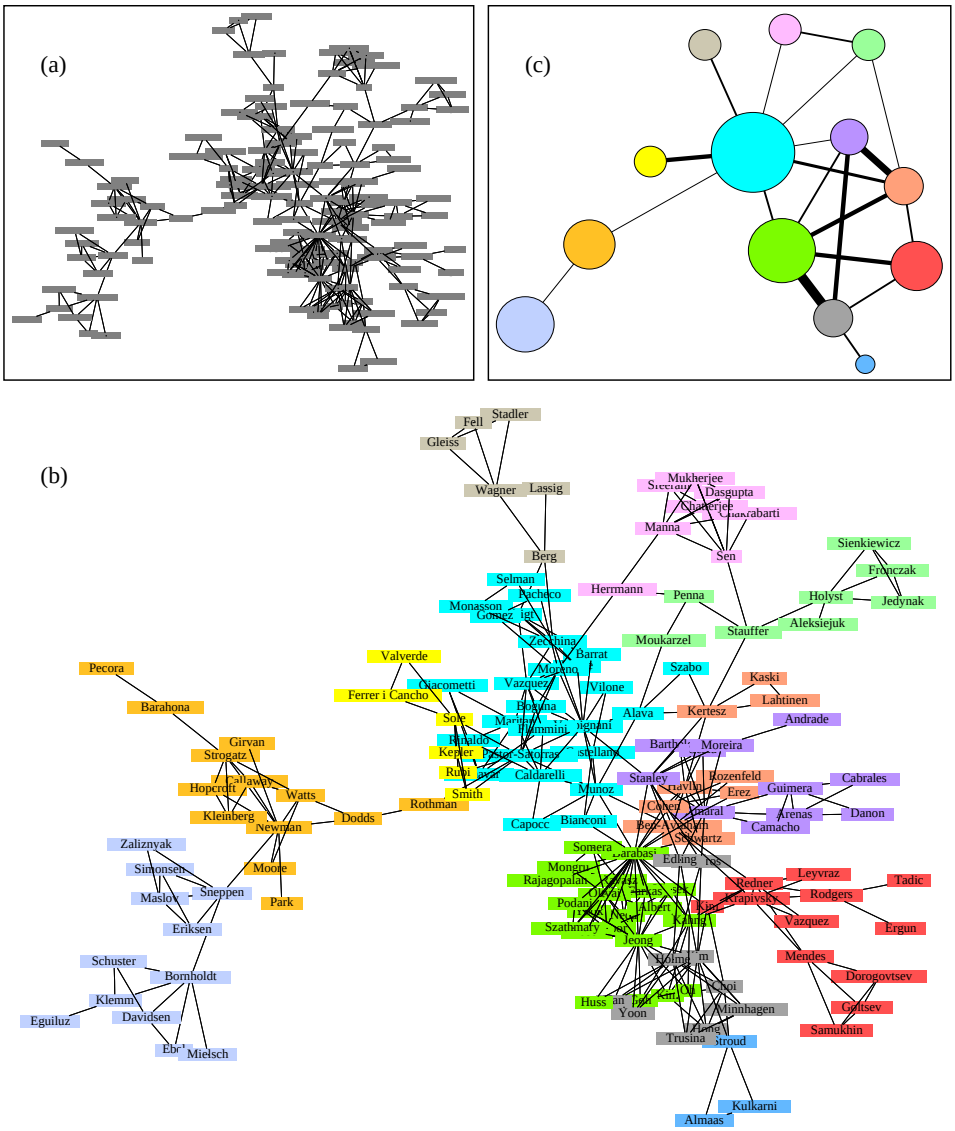


图 1.8: 进行网络研究的物理学家之间的合作网络的最大分量的社团结构[144]. 这个网络由出现在 [138] 的冗长的参考文献里的作者名字构成. (a) 物理学家之间的合作网络最大分量的原始网络. (b) 用 [144] 中的算法的最短路径介数形式所得到的最优分区的社团, 以不同的颜色表示. (c) 将每个社团表示成节点且将社团之间的合作表示成边所得到的粗粒化网络, 其中边的粗细与社团之间合作的对数成正比. 显然, (c) 揭示出了许多在原始网络 (a) 中不容易看出的信息.

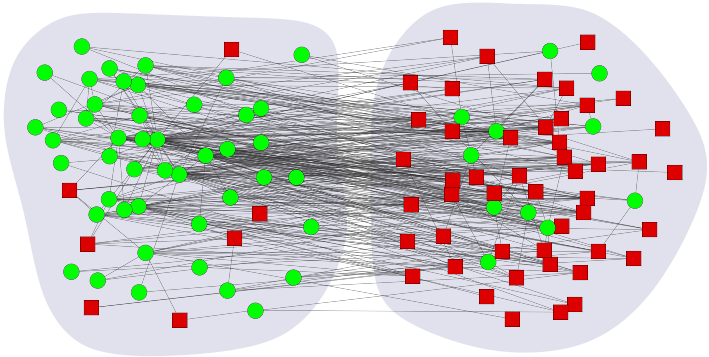


图 1.9: 查尔斯 • 狄更斯的小说《大卫 • 科波菲尔》中通常出现的英语词汇之间的邻接网络的社团结构[141]. 其中圆形表示小说中的形容词, 方形表示名词. 这是用 [141] 中的算法得到的结果.

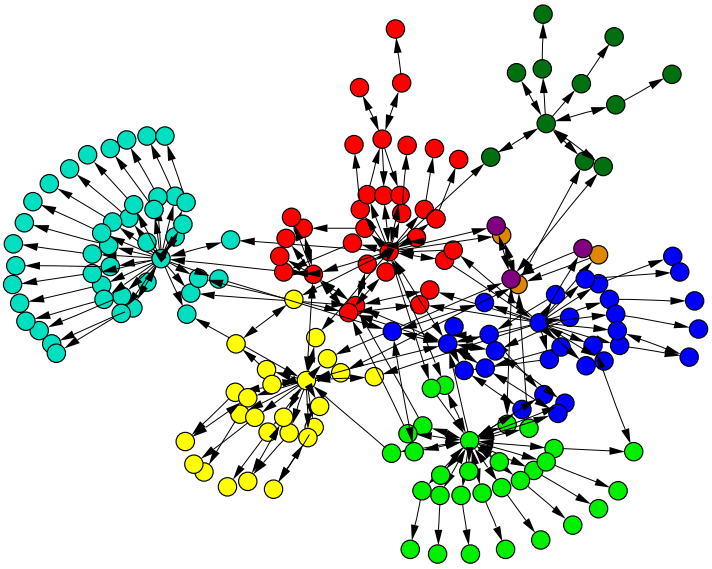


图 1.10: 一个网站的网页与它们之间的超链接所构成的网络的社团结构[144]. 用 [144] 中的算法的最短路径介数形式所得到的最优分区的社团, 以不同的颜色表示.

1.4.1 传统方法

图形分割方法. 图形分割方法包括著名的 Kernighan-Lin 算法[100]及其推广[188], 谱平分法[15]. Ford 和 Fulkerson 提出最大流量最小切割定理[69], Goldberg

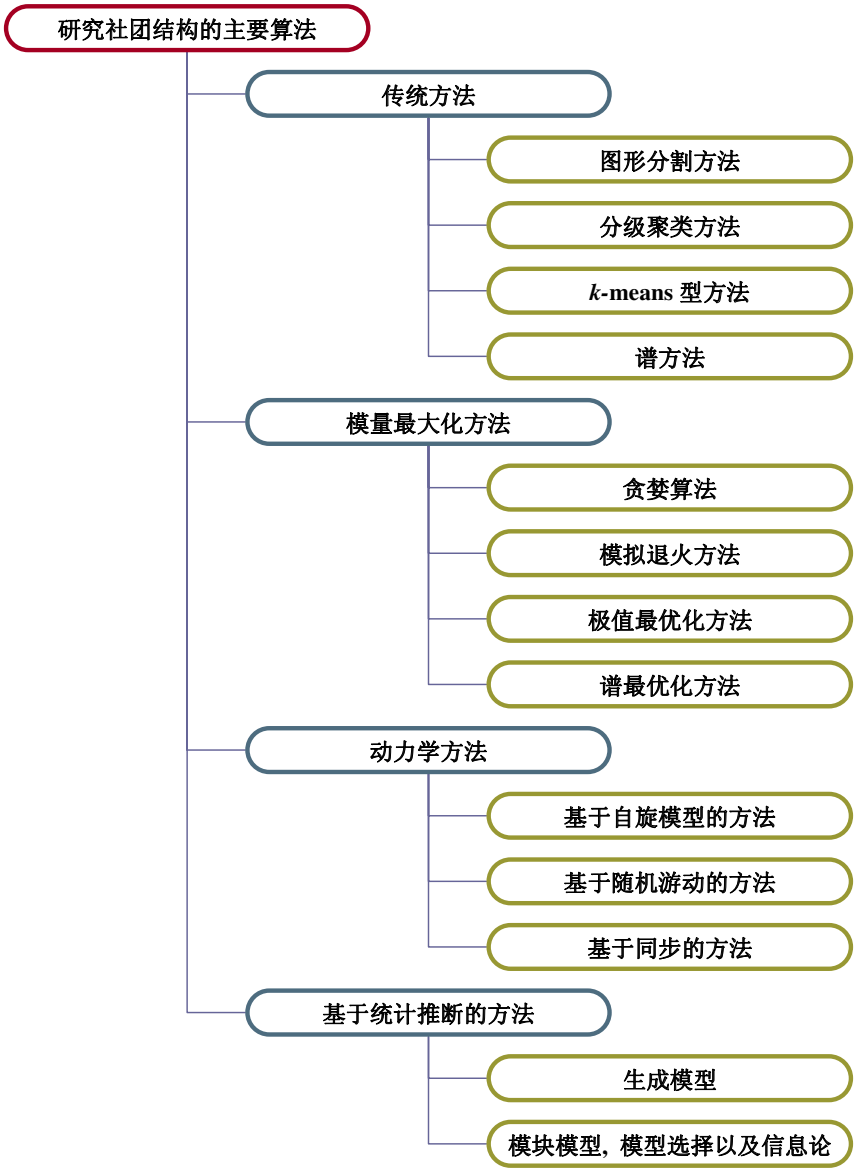


图 1.11: 研究复杂网络社团结构的主要算法[70]的组织结构图.

和 Tarjan 提出计算图中最大流量的算法[78], Flake 等人利用最大流量来确定社团结构[68]. 其它的图形分割方法包括能级结构分割, 几何算法, 多层算法等等[156].

分级聚类方法. 又分为凝聚算法和分裂算法两类[86]. 分裂算法包括著名的 Girvan 和 Newman 提出的 GN 算法[77, 144], Tyler 等人提出的算法[192, 203], Rattigan 等人提出的算法[161], Holme 等人提出的算法[90], 以及 Radicchi 等人提出了自包

含 GN 算法[158], Fortunato 等人提出的算法[72]等等. 凝聚算法包括 Newman 快速算法[140], Clauset 等人提出的算法[39], Latapy 和 Pons 提出的算法[155], Bagrow 和 Bollt 提出的算法[10], 以及 Donetti 和 Muñoz 提出的结合谱分析的凝聚算法[53, 54]等等.

k-means 型方法. 这是研究数据点集合中聚类的一类流行的方法, 常用的方法包括 minimum k-clustering, k-clustering sum, k-center, k-median 等等[70, 86]. 最流行的方法莫过于 k-means 算法[127], 及其向图中的推广[88, 173]. 另一个流行的方法 fuzzy c-means 算法[19, 58]考虑到一个样本点可能同时属于两个或者更多的分区.

谱方法. 这是一种利用图矩阵 (关于图矩阵的介绍见附录 A.2) 的特征向量来进行分区的方法 [185, 195]. 首次根据邻接矩阵的特征向量来提出谱分割方法的是 Donath 和 Hoffman[52], 同一年 Fiedler 根据 Laplace 矩阵的第二小特征值的特征向量提出的平分法[67]. 此外还有 Shi 和 Malik 提出基于非归一化 Laplace 矩阵的算法[132, 182], Ng 等人提出基于归一化 Laplace 矩阵的算法[146]. Donetti 设计了基于 Laplace 矩阵特征向量的方法[53], Wu 和 Huberman 提出了一种基于电阻网络电压谱的快速谱分割法[205], Capocci 等人提出的方法[26], Yang 和 Liu 提出的递归平分法[209]等等.

1.4.2 模量最大化方法

贪婪算法. 包括著名的 Newman 快速算法[140], Clauset 等人提出的算法[39], Danon 等人提出的算法[42], Wakita 和 Tsurumi 提出的算法[196], Blondel 等人提出的算法[20], 以及 Schuetz 和 Caflisch 提出的算法[175, 176]等等.

模拟退火方法. Guimerà 等人首次将模拟退火[103, 133]运用到模量最优化问题[82, 83], 并在后来得到进一步发展和应用[129, 131]. 本文中的部分工作[117–119]就利用了模拟退火结合迭代法的策略来最优化目标函数, 包括有效性指标和模量以及模糊模量的形式.

极值最优化方法. 这是一种由 Boettcher 和 Percus 提出的启发式搜索方法[23], 这个方法并由 Duch 和 Arenas 应用于模量最优化问题[57], 其基本思想是通过调整局部极值来优化全局的变量, 从而提高运算效率.

谱最优化方法. 首次由 Newman 提出[141, 142], 将 Laplace 矩阵替换为模量矩阵并通过谱平分法得到最优化模量的两部分, 后来又得到进一步推广和应用, 包括 Wang 等人提出的算法[198], Sun 等人提出的算法[190], 以及 Richardson 等人提出的算法[165]等等.

1.4.3 动力学方法

基于自旋模型的方法. 基于自旋模型[206]的方法主要包括 Reichardt 和 Bornholdt 提出的方法[163], Ispolatov 等人提出的方法[94], Son 等人提出的方法[184], 以及 Ronhovde 和 Nussinov 提出的方法[167]等等.

基于随机游动的方法. 基于随机游动[93]的方法主要包括 Zhou 等人提出的方法[213–215], Latapy 和 Pons 提出的方法[155], Hu 等人提出的方法[92], Delvenne 等人提出的方法[48], E 等人提出的方法[60], 以及 Van Dongen 提出的方法[55]等等.

基于同步的方法. 首次提出基于同步[154]的方法的是 Arenas, Díaz-Guilera 和 Pérez-Vicente[6, 7], 此外还有 Boccaletti 等人提出的方法[21], 以及 Li 等人提出的方法[113]等等.

1.4.4 基于统计推断的方法

生成模型. 基于 Bayesian 推断 (生成模型)[204]的方法主要包括 Hastings 提出的方法[87], Newman 和 Leicht 提出的方法[145], Vazquez 提出的方法[193], Ramasco 和 Mungan 提出的方法[159], Čopič 等人提出的方法[194], Zanghi 等人提出的方法[212], 以及 Hofman 和 Wiggins 提出的方法[89].

模块模型, 模型选择以及信息论. 基于模块模型[56]的方法有 Reichardt 和 White 提出的方法[164]; 基于模型选择[25]及相关准则[3, 166, 178, 197]的方法有 Rosvall 和 Bergstrom 提出的方法[168–170], 以及 Chakrabarti 提出的方法[27]; 基于信息论[126]的方法有 Ziv 等提出的方法[216].

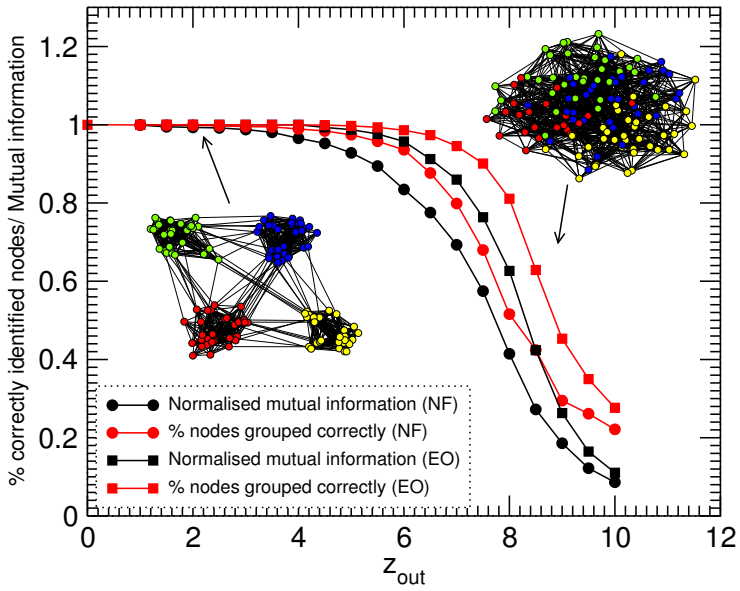


图 1.12: 具有 4 个社团的 ad hoc 网络: 对于较低的 $z_{\rm out}$, 社团很容易被区分出来; 对于较高的 $z_{\rm out}$, 社团结构变得较为复杂[43]. 在识别社团过程中, 归一化互信息看起来比节点识别的正确率对于误差更为敏感. 这里给出的是 Newman 快速算法[140]和极值最优化算法[57]的结果.

1.5 研究复杂网络社团结构的主要算例

1.5.1 人工生成的网络

128 个节点的 ad hoc 网络. 这是一个在很多文章中考虑过的典型的基准网络[43, 70, 77, 144], 它具有已知的社团结构并构造如下. 假设选取 $n = 128$ 个节点, 分成 4 个社团, 每个社团包含 32 个节点. 假设属于相同社团的节点对以概率 $p_{\rm in}$ 相连接, 而属于不同社团的节点对以概率 $p_{\rm out}$ 相连接. 这些值得选取要使得平均节点度 $\langle d\rangle$ 固定为 $\langle d\rangle = 16$. 换句话说, $p_{\rm in}$ 和 $p_{\rm out}$ 有如下关系

$$31p_{\rm in} + 96p_{\rm out} = 16. \tag{1.7}$$

这里自然地选择节点组 $S_1 = \{1:32\}, S_2 = \{33:64\}, S_3 = \{65:96\}, S_4 = \{97:128\}$. 通常定义 $z_{\rm out}$ 为某个节点与属于其它社团节点之间连接的平均数, 即 $z_{\rm out} = 96p_{\rm out}$, 并用这个量作为一个控制参数. $z_{\rm out}$ 越大, 社团就变得越模糊 (diffusive). 为比较固定模块结构与算法得到的结构, 可以将 $z_{\rm out}$ 从小到大地变化,

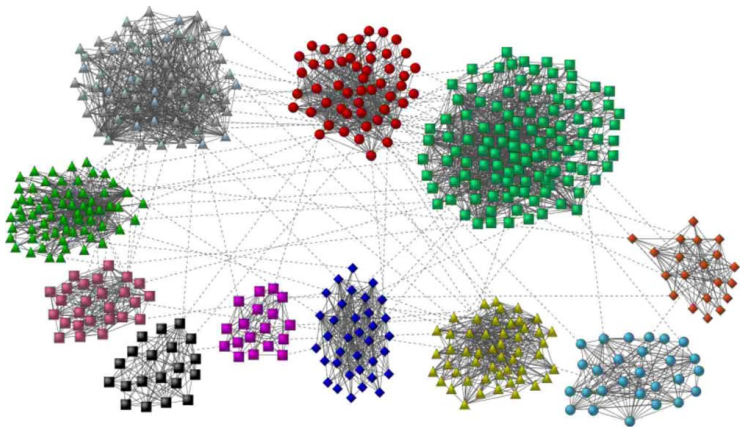


图 1.13: $n = 500$ 个节点的 LFR 基准网络的一个实现[111].

并考察两个被证实是可靠的量: 节点识别的正确率 (fraction of correctly identified nodes)[43, 77, 144], 以及归一化互信息 (normalized mutual information)[43, 109, 111]. 归一化互信息的定义基于模糊矩阵 M, 其行对应于真实社团, 其列对应于找到的社团. M 的分量 M_{kl} 为出现在找到的社团 S_l 中的真实社团 S_k 中的节点数目. 真实社团数目记为 N_r, 而找到的社团数目记为 N_f. 矩阵 M_{kl} 对第 k 行求和记为 M_k, 对第 l 列求和记为 M_l. 基于信息论[126]的分划之间相似性的度量定义为

$$I(\mathbb{S}_r, \mathbb{S}_f) = \frac{-2\sum_{k=1}^{N_r}\sum_{l=1}^{N_f} M_{kl}\log(\frac{nM_{kl}}{M_k M_l})}{\sum_{k=1}^{N_r} M_k \log(\frac{M_k}{n}) + \sum_{l=1}^{N_f} M_l \log(\frac{M_l}{n})}. \tag{1.8}$$

图 1.12 给出了 Newman 快速算法[140]和极值最优化算法[57]测试于 ad hoc 网络的结果.

LFR 基准网络. LFR 基准网络[108, 109, 111]是为研究社团结构而构造的一个现实主义的网络, 它同时要求节点度和社团规模的非均匀性. 节点度服从指数为 γ 的幂律分布, 而社团规模服从指数为 β 的幂律分布. 在 LFR 基准网络的构造中, 每个节点坚决地接收它的度并保持它固定直到最后. 更为实际的做法是选取混合参数 μ 作为独立参数, 它表示一个节点关于它所在社团的外面的度与全部度之间的比率[111]. 图 1.13 给出了 $n = 500$ 个节点的 LFR 基准网络的一个实现. LFR 基准网络可进一步地推广到具有重叠社团的情形[108], 相应的对于重叠社团的广义化的归

一化互信息被提出并用来实现测试算法的目的[110].

Gauss 混合模型生成的样本网络. 这个模型与 Penrose 提出的随机几何图的概念[153]有关, 只是在本文中选取 Gauss 混合模型, 而不再是 [153] 中的均匀分布. 首先, 在二维欧式空间中生成 n 个样本点 $\{\boldsymbol{x}_i\}$ 其服从 K-Gauss 混合分布

$$\sum_{k=1}^{K} q_k G\left(\boldsymbol{\mu}_k, \Sigma_k\right), \tag{1.9}$$

其中 $\{q_k\}$ 是混合比例且满足 $0 < q_k < 1$, $\sum_{k=1}^{K} q_k = 1$, 这里 $\boldsymbol{\mu}_k$ 和 Σ_k 分别是每个分量的均值和协方差矩阵. 然后, 根据阀值策略生成网络, 即若 $|\boldsymbol{x}_i - \boldsymbol{x}_j| \leq dist$, 在第 i 个和第 j 个节点之间赋一条边; 否则它们不相连. 根据这个策略, 网络的拓扑由这个度量所诱导. 因此这个网络的某些性质, 例如聚类的性质, 由这个度量得到了继承. 这是本文中利用这个模型的基本动机. 在后面的第四,五,六章将使用这个模型网络来测试算法.

Mueller 势生成的样本网络. 这是一个 Langevin 轨道点与阀值准则结合起来而形成的网络, 类似于 Gauss 混合模型生成的样本网络. 考虑 Langevin 动力学

$$d\mathbf{X}_t = -\nabla V(\mathbf{X}_t)dt + \sqrt{\varepsilon}d\mathbf{W}_t, \tag{1.10}$$

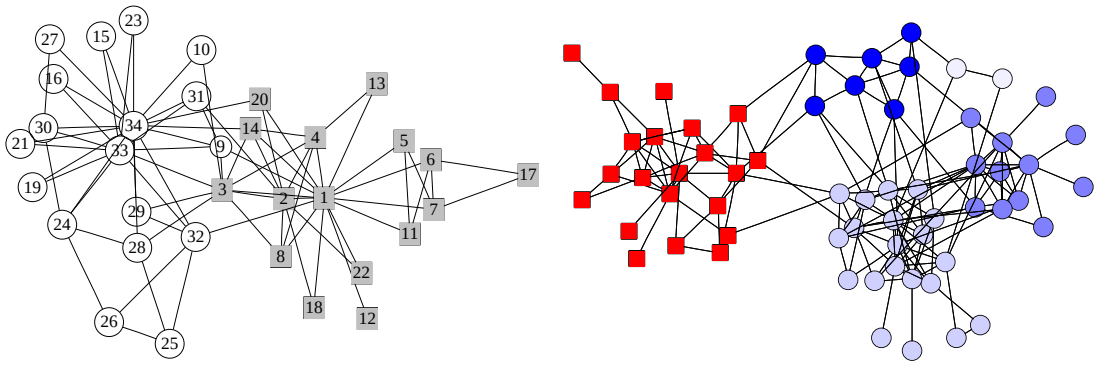
这里选取的 Mueller 势 $V(x, y)$ 具有如下形式

$$V(x,y) = \sum_{i=1}^{4} A_i \exp(a_i(x-x_i)^2 + b_i(x-x_i)(y-y_i) + c_i(y-y_i)^2) \tag{1.11}$$

其中参数为

$$\begin{aligned} A &= (-200, -100, -170, 15), \\ a &= (-1, -1, -6.5, 0.7), \\ b &= (0, 0, 11, 0.6), \\ c &= (-10, -10, -6.5, 0.7), \\ x &= (1, 0, -0.5, -1), \\ y &= (0, 0.5, 1.5, 1). \end{aligned}$$

如第四章中图 4.8 和图 4.9 所示, 它有三个局部极小值点, 分别标记为 A, B 和 C; 两个鞍点, 分别标记为 D 和 E. 由弦方法[61, 62]得到的从 A 到 C 得最小能量路径也



(a) (b)

图 1.14: (a) Zachary 空手道俱乐部网络的社团结构[210]. 节点 1 和节点 33 分别表示俱乐部的管理者和主教练. 深色方形代表在俱乐部分裂后跟随俱乐部管理者的成员, 浅色圆形代表跟随俱乐部教练的成员. (b) 新西兰道尔福峡湾的宽吻海豚网络的社团结构[121, 122]. 方形和圆形代表网络主要分裂成两个社团, 圆形进一步细分为四个较小的社团, 由不同颜色深度的节点表示. 这两个图均是用 [144] 中的算法的最短路径介数形式所得到结果.

在这两个图中绘出. 作为反应路径中的瓶颈的鞍点 D 和E 起到了不同能量盆地之间的转移状态的作用. 利用如下的 Euler-Maruyama 格式[105]可得到样本点

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \delta t \nabla V(\mathbf{X}_n) + \sqrt{\varepsilon}\delta \mathbf{W}_n, \tag{1.12}$$

其中 $\delta \mathbf{W}_n$ 为标准 Gauss 随机变量 $G(0, \delta t \boldsymbol{I})$. 于是可选取 n 个样本点. 然后类似于 Gauss 混合模型, 根据阀值 $dist$ 生成网络. 由于三个极小值点 A, B 和 C, 故社团结果数目的选取为 $N = 3$.

1.5.2 真实世界中的网络

空手道俱乐部网络. 这个网络是由 Wayne Zachary 在观察一所美国大学空手道俱乐部成员之间的社会联系而构建的[210]. 不久, 俱乐部的管理者和主教练之间发生争吵, 于是俱乐部分裂成两个小俱乐部. 如图 1.14(a) 所示, 在空手道俱乐部网络中有 34 个节点, 每个节点表示俱乐部中的一个成员. 节点 1 和节点 33 分别表示俱乐部的管理者和主教练. 深色方形代表在俱乐部分裂后跟随俱乐部管理者的成员, 浅色圆形代表跟随俱乐部教练的成员. 在 [210] 中, Zachary 给出分区

$S_1 = \{1:8, 11:14, 17, 18, 20, 22\}$ 和 $S_2 = \{9, 10, 15, 16, 19, 21, 23:34\}$. 这个网络广泛地应用于验证研究网络社团结构的算法[70, 77, 144].

宽吻海豚网络. 宽吻海豚网络由生活在新西兰道尔福峡湾 (神奇湾) 的一个组织中的 62 只宽吻海豚之间的频繁联系所构成的无向社会网络[121, 122]. 这个网络是由 Lusseau 对这些海豚七年的现场研究而构造的, 通过对于统计上的重要且频繁的联系的观察而建立每对海豚之间的边[121]. 如图 1.14(b) 所示, 网络分裂成两个较大的社团, 分别用圆形和方形表示, 并且这两个社团中较大的那个继续分裂成四个较小的子社团, 由不同的颜色深度表示. 分裂为两个较大社团对应于海豚组织的根据年龄一个已知的分区[122]. Lusseau 指出, 在为期两年的对于这些海豚的观察的时间里, 这些海豚依照 [144] 中算法的结果分成两个社团, 这显然因为在两个社团之间的沿线上的个体的消失. 当这些消失的部分个体后来又再次出现的时候, 网络的两半又重新聚在了一起. 正如 Lusseau 指出, 这种形式的发展说明研究海豚的网络不仅是出于对科学本身的好奇心, 而是像人类社会网络那样, 紧密的与社会进化连接在一起. 网络中较大一半的社团也似乎相对应动物中的真实分区: 最大的子社团几乎完全为雌性海豚组成, 而其它子社团几乎全部为雄性海豚, 从而可以推测雄性海豚社团之间的分区由母系家族所主导.

美国政治书籍网络. 这个网络是由 V. Krebs 编制的关于美国政治的书籍的网络④. 网络中的节点表示从在线书商 Amazon.com 上购买的最近的 105 本关于美国政治的书籍, 连接书籍对的边表示这两本书频繁地由相同顾客购买. 书籍的分类是按照它们所陈述的明显的政治立场, 自由党或者保守党, 除了一小部分书籍是明确的两党派或中立者, 或者没有明确的从属关系[142]. 图 1.15(a) 展现了不同形状代表书籍的政治上的组合: 圆形 (蓝色) 为自由党, 方形 (红色) 为保守党, 三角形 (紫色) 为中立或无党派. 并给出了由 [142] 中算法得到的结果. 这种算法发现网络的四个社团, 由图中的虚线标出并分开. 可以看到其中的一个社团几乎完全由自由党书籍组成, 还有一个几乎完全由保守党书籍组成. 大多数的中立的书籍在其余的两个社团中. 因此这些书似乎形成了与政治观点密切相关的联合购买的社团, 结果支持 [142] 中算法能够从原始数据网络中提取有意义的结果. 特别有趣的是, 中立的书籍

④未出版, 但是可见于 http://www.orget.com/

具有属于他们自己的社团, 而不是像多数情况下那样并归于自由党社团或保守党社团, 这或许指出政治温和派形成他们自己的购买社团.

小说《悲惨世界》人物关系网络. 这是维克多·雨果 (Victor Hugo) 创作的关于法国恢复后的犯罪与救赎的长篇巨著《悲惨世界》(*Les Misérables*) 中的主要人物之间的相互关系所构成的网络, 它是由 Knuth 根据戏剧的场次中出现的人物列表而构造的[106]. 网络中的节点代表人物, 两个节点之间的边代表与相关人物共同出现在一场或多场中. 图 1.15(b) 中给出了由 [144] 中算法的最短路径形式得到的 11 个社团, 分别用不同的颜色表示. 社团清楚地反映了书中次要情节的结构: 主角 Jean Valjean 和他的复仇者, 警务人员 Javert 均是网络的重要成员, 并形成由他们的拥护者组成的社团的中心. 其它集中在 Marius, Cosette, Fantine 和主教 Myrial 的次要情节也在图中表现出来.

圣达菲研究所科学家合作网络. 这是美国新墨西哥州圣达菲 (Santa Fe) 的一个交叉学科研究中心: 圣达菲研究所中的科学家之间的合作网络[77]. 在这个网络中的 271 节点代表在 1999-2000 年居住在圣达菲研究所的科学家以及他们的合作者. 如果在同样的时间段内, 两个科学家之间合作过一篇或者更多的论文, 则他们之间就画上一条带权重的边. 这个网络包含了上述科学家出版的所有杂志和书籍, 连同所有出现在研究所的技术报告系列中的文章. 平均起来, 每个科学家近似地与其他五个人合作文章. 图 1.15(c) 中给出了将 [77] 中算法应用于合作网络的最大分量所得到的结果. 这个网络由 118 个科学家组成, 根据分区结果将 4 个社团的节点表示为不同的形状. 网络分成了几个强大的社团, 并且主要根据学科之间的差异区分开来. 图中位于顶部的社团 (菱形) 是最差良定义的, 它表示利用基于智能体模型来研究经济和交通流量问题的科学家组成的社团. 下面一个社团 (圆形) 表示研究生态学中数学模型的科学家组成的社团, 形成了一个相当凝聚的结构. 最大的社团 (方形) 是主要研究统计物理的科学家组成的社团, 并且进一步细分区成几个良定义的更小的社团, 用不同颜色深度表示. 在这个情形下, 每个子社团看起来围绕在一个主导成员的研究兴趣周围. 最后的位于图中底部的社团 (三角形) 是主要研究 RNA 结构的科学家组成的社团.

美国足球队网络. 这个网络表示美国大学生足球联联赛 2000年 第一季度的比

赛日程[77]. 网络中的节点表示 115 个由学校名字命名的足球队, 连接两个节点的边表示他们之间的规则季度赛. 使这个网络有趣的是它包含了一个已知的社团结构. 这些足球队被分成一些联盟, 每个联盟包含 8 到 12 个足球队. 同一个联盟中的球队之间的比赛比不同联盟球队之间的比赛要频繁, 这些球队在 2000 赛季要参加平均 7 场的联盟内部比赛和 4 场联盟之间比赛. 联盟之间比赛不是均匀分布的, 一支球队在地理位置上靠近另一支球队但是属于不同的联盟比相隔遥远的地理距离的球队比赛频繁. 图 1.15(d) 中给出了由 [77] 中算法得到的反映这个网络社团结构的层次树, 并很成功地识别了联盟结构. 几乎所有的球队都正确地与他们联盟内的其他球队分组在一起. 有少部分独立球队不属于任何联盟, 他们倾向于同他们密切联系的球队分到一个联盟. 极少数未识别出的情况实际上由于比赛时间表里的细微差别. 例如, Sunbelt 联盟分成两部分, 并和 Western Athletic 联盟的球队组合在一起. 这种情况的发生由于 Sunbelt 联盟的球队与 Western Athletic 联盟的球队的比赛几乎与他们在各自联盟内部比赛的次数几乎相同.

1.6 章节概要

本文共分七章, 除本章外, 其余六章的结构如下文所述.

第二章将介绍本文的动力学分区方法所涉及的理论基础, 以及与确定性分区方法[60]相似的工作. 其中理论基础包括 Chorin 等人提出的最优预测理论[31–35]和 Schütte 等人提出的不变集合体的识别方法[47, 49, 50, 177], 与确定性分区方法具有某些相似思想的工作包括 Meila 和 Shi 提出的图像分割中的 MNCut 算法[132, 182]以及 Lafon 和 Lee 提出的数据挖掘中的扩散映射方法[40, 41, 107, 137].

第三章将介绍基于最优预测理论[31–35]的网络确定性分区的方法[60]. 其基本思想是将网络与随机游动 Markov 动力学[120]联系起来, 然后引入马氏链空间中的一种度量, 即前向算子的 Hilbert-Schmidt 范数, 并且在这个度量下最优约化马氏链. 最终的极小化问题由聚类分析中的传统 k-means 算法[86]的一个变形来求解.

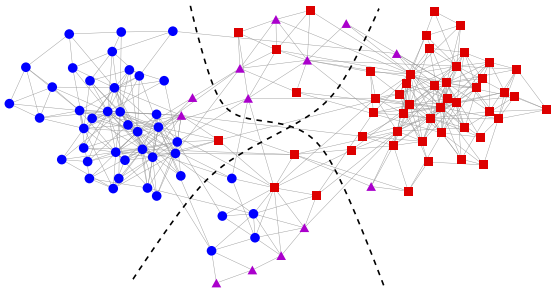
第四章将第三章介绍的工作扩展到概率性的框架中[114, 116]. 此时网络中的每个节点以某一概率从属于某一社团, 而不是将节点分配到确定的社团中. 作者提出一个概率分布空间的自由能函数, 当温度为 $-\infty$ 时, 该自由能函数退化成第三章中

提出的目标函数. 对于这个概率性的框架, 构造了相应的网络概率性分区的算法. 这种扩展是十分自然和有价值的, 特别是对于那些没有显著社团结构的网络, 概率性分区通常包含更多详细的信息.

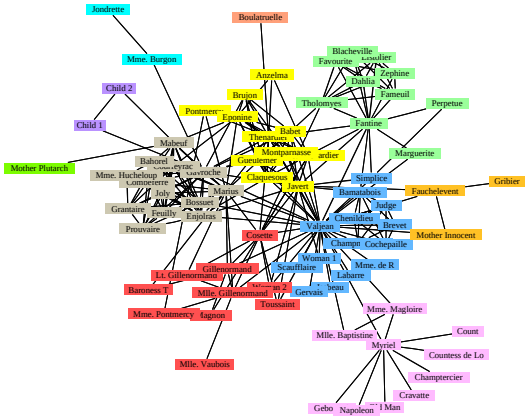
第五章将解决网络确定性分区的自动模型选择[118]. 利用有效性指标 (validity index) 的思想, 构造了一个新的针对于第三章中网络确定性分区的有效性指标函数, 来度量识别出的社团结构的优良程度, 它包含每个分区的紧密程度和分离程度这两个因素. 然后利用模拟退火的策略[103, 133]来得到这个函数的极小值. 这种结合了之前的变形 k-means 的模拟退火方法不仅可以有效得到网络的社团结构, 而且不用任何关于社团结构的先验信息就可以自动确定出社团的数目.

第六章将分别利用模量 (modularity) 和模糊模量 (fuzzy modularity) 来实现复杂网络确定性分区和概率性分区的自动模型选择[117, 119]. 6.1 中算法不仅可以确定社团结构, 还可以确定每个社团的中心节点, 并且最优社团数目在可以被自动地确定, 而不需要任何关于社团结构的先验信息. 6.2 提出了模糊模量函数来衡量网络概率性分区的优良性, 相应的算法可以给出每个节点属于不同社团的概率, 并成功地克服了第四章中算法的弱点, 即此时社团数目可以被自动确定而不再是将它固定为已知的模型参数, 并且初始模糊分区可以随机选取.

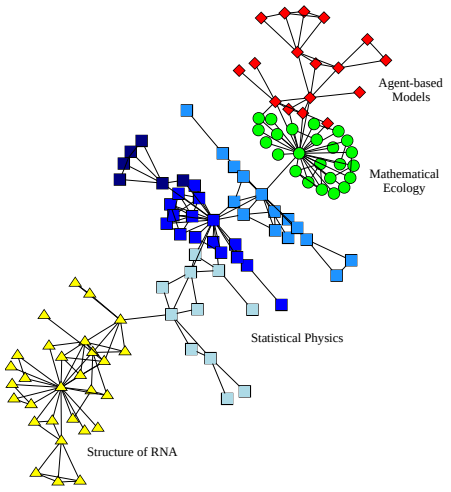
第七章将简要的总结本文的工作, 展现本论文创新点和不足, 并将本论文提出的一些算法与文献中的其它方法进行比较分析, 最后对未来将要继续研究的内容进行了展望.



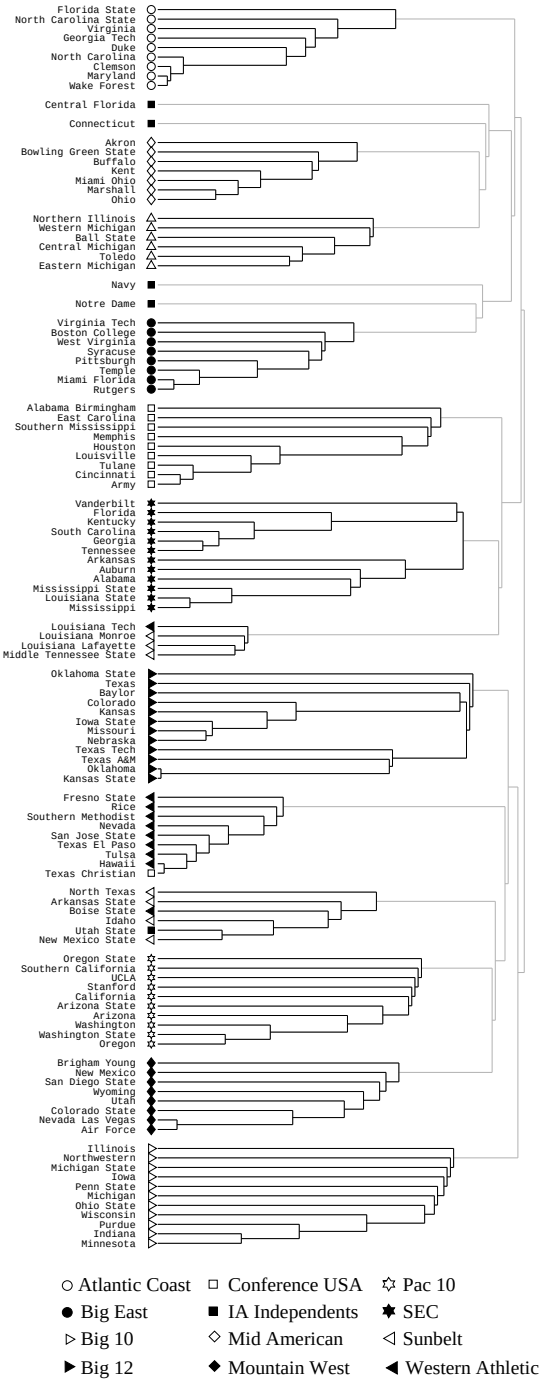
(a)



(b)



(c)



(d)

图 1.15: (a) Krebs 编制的关于美国政治的书籍网络[142]. (b) 维克多·雨果的小说《悲惨世界》的主要人物之间的相互关系网络[144]. (c) 圣达菲研究所科学家合作网络的最大分量[77]. (d) 反映美国大学生足球联联赛 2000 年第一季度的比赛日程的网络的层次树[77].

第二章　基于最优预测的动力学方法的理论基础

本章将介绍本文的动力学分区方法所涉及的理论基础, 以及与确定性分区方法[60]相似的工作. 其中理论基础包括 Chorin 等人提出的最优预测理论[31–35]和 Schütte 等人提出的不变集合体的识别方法[47, 49, 50, 177], 与确定性分区方法具有某些相似思想的工作包括 Meila 和 Shi 提出的图像分割中的 MNCut 算法[132, 182]以及 Lafon 和 Lee 提出的数据挖掘中的扩散映射方法[40, 41, 107, 137]. 本文的动力学分区方法所涉及的理论基础, 以及与确定性分区方法[60]相似的工作.

2.1　最优预测理论

现代科学中有许多问题的解是用一组微分方程来描述的, 然而有些方程的解由于太复杂而不能被数值地求出. 精确的数值解要求问题可以被良好地解决, 即在计算中保持足够多的变量 (自由度) 来表示解的全部相关的特征. 有很多在实践中不能得到良好解决的著名的例子, 包括扰动以及统计物理和经济学中的各种各样的问题. 在这一节中即将讨论考虑这样的不能被良好解决的问题.

本节内容主要参考 [33].

2.1.1　目标与方法概述

考虑具有如下形式的偏微分方程

$$\frac{\partial u}{\partial t} = R(u), \tag{2.1}$$

其中 t 为时间, $R(u) = R(u, \partial u/\partial x, \ldots)$ 为一个关于其自变量的函数 (通常是非线性的). 假设 (2.1) 的解空间中的一种度量在由 (2.1) 诱导的流 (flow) 下是不变量, 而它是什么是已知的. 记关于这个测度的均值为 $\langle\cdot\rangle$. 进一步假设不能解出 u, 但是

可以在少量的网点上发现关于 u 的一些信息; 这些信息可由点值组成, 但在物理上和数学上更合理的假设是, 如果已有"过滤"值, 它们将实际上由真实的物理度量产生. 因此, Chorin 等人假设在网格点上具有值 $\bar{u}_\alpha$, 定义为

$$\bar{u}_\alpha(t) = \int G(x - x_\alpha)u(x,t)dx, \tag{2.2}$$

这里 α 是网格上的指标. 核函数 G 可表示诸如空间平均这样的量. (2.2) 中的粗网格数据确定了在时间方向的每个时刻的一个子集的函数 (这些函数是符合观察值的). 记在这个限制子集上的均值为 $\langle\cdot\rangle_{\bar{\mathbf{u}}}$, 称在这个限制子集上的测度为"约束测度". 如果问题有待求解, 则这个测度可表示为函数的一个非平凡集合. 注意到限制测度不是不变量; 事实上, 如果初始的不变测度是遍历的, 则限制测度在时间上趋于非限制测度; 例如, 如果要求初始时所有函数假设在网格点给定值, 则没有理由相信微分方程的由这些数据所进化出来的解在后面的时间中仍然在网格点上取相同的值.

Chorin 等人的目的是计算关于约束测度的均值; 这些均值表示能够从粗网格中得到的, 关于不能在其上表示的"不可见"自由度的平均. 给定约束测度和过滤值, 可通过插值 (等价于线性回归[152]) 来得到解的均值和矩, 从而出于忽略误差的实践目的可得到这些量在计算点处的平均导数. 余下的问题是表征约束测度的进化, 使得平均解和任意阶矩可以按时间更新. Chorin 等人的假设是由 N 个过滤器限制的约束测度保持不变测度, 其中 N 是网格点的个数. 因此, 过滤器随时间变化. 下面的公式将与适合确定演化过滤器的参数的进化的量联系起来. 这里, Chorin 等人通过假设过滤器由平均解 (或解的高阶矩) 的演化所确定, 将寻找演化过滤器的问题简化. 这等价于假设方程

$$\frac{d\bar{u}_\alpha}{dt}(t) = \left\langle \int G(x - x_\alpha)R(u)dx \right\rangle_{\bar{\mathbf{u}}(t)} \tag{2.3}$$

是由 (2.1) 和 (2.2) 给出的 $\bar{u}_\alpha(t)$ 的真实演化的一个很好的近似. 当然, 这个假设的有效性依赖于过滤器 G 的选择. 下面将进一步假设测度是 Gauss 分布或近似 Gauss 分布的.

两个成功的关键点是: (i) 关于正确的约束测度的均值, (ii) 当解演化时更新约束. 因素 (i) 已在 [30] 中数值地应用. 先前的一些有趣的尝试[130, 179] 通过困难的计

算来填充粗网格上的数据, 从而不需加细网格来就可加强精度, 但是如果没有这两个关键因素, 早期方法的有效性是很有局限的.

2.1.2 Gauss 分布和仿射约束下的条件期望

考虑当变量满足仿射形式的约束时, 如何计算具有 Gauss 分布变量的函数的期望值. 设 $\mathbf{u}=(u_1,\ldots,u_n)^T$ 为联合 Gauss 随机变量组成的实向量; 它具有如下形式的概率密度 $F(\mathbf{u})$

$$\begin{aligned} & P(s_1<u_1\le s_1+ds_1,\ldots,s_n<u_n\le s_n+ds_n) \\ = \;& F(\mathbf{s})ds_1\ldots ds_n \\ = \;& Z^{-1}\exp\left(-\frac{1}{2}s_ia_{ij}s_j+b_is_i\right)ds_1\ldots ds_n, \end{aligned} \tag{2.4}$$

其中 Z 是适当的归一化因子, 重复指标表示求和, $n\times n$ 的矩阵 $A=\{a_{ij}\}$ 是对称的, 假设其逆 A^{-1} 存在, 为逐点协方差矩阵, 其元素为

$$a_{ij}^{-1}=\mathrm{Cov}(u_i,u_j)\equiv\langle u_iu_j\rangle-\langle u_i\rangle\langle u_j\rangle, \tag{2.5}$$

这里 $\langle\cdot\rangle$ 表示关于概率密度的均值, 且向量 $\mathbf{b}=(b_i)$ 与逐点期望值有关

$$a_{ij}^{-1}b_j=\langle u_i\rangle. \tag{2.6}$$

这个分布完全由 n 个均值和协方差矩阵的 $\frac{1}{2}n(n+1)$ 个独立元素所决定; 从而所有观察量的期望值 $\mathscr{O}(\mathbf{u})$ 都可由这些参数表示, 特别是所有高阶矩已由 Wick 定理[104]给出.

下面, 假设随机向量 $\mathbf{u}$ 满足一系列仿射约束, 形式如下

$$g_{\alpha i}u_i=\bar{u}_\alpha,\quad \alpha=1,\ldots,N<n, \tag{2.7}$$

其中指标 α 列举约束, 矩阵 $G=(g_{\alpha i})$ 为 (2.2) 中核函数 $G(\cdot)$ 的离散形式. 这里分别用指标 i 和 α 来区分随机变量 $(u_1,\ldots,u_n)$ 的向量空间和约束 $(\bar{u}_1,\ldots,\bar{u}_N)$ 的向量空间.

Chorin 等人的目标是计算期望值, 即关于满足约束的函数的均值; 形式上, 有

$$\langle\mathscr{O}(\mathbf{u})\rangle_{\bar{\mathbf{u}}}=\frac{\int(\prod_{i=1}^n du_i)\mathscr{O}(\mathbf{u})F(\mathbf{u})\prod_{\alpha=1}^N\delta(g_{\alpha j}u_j-\bar{u}_\alpha)}{\int(\prod_{i=1}^n du_i)F(\mathbf{u})\prod_{\alpha=1}^N\delta(g_{\alpha j}u_j-\bar{u}_\alpha)} \tag{2.8}$$

其中左端引入了约束均值的记号, $F(\mathbf{u})$ 为归一化的概率密度 (2.4). 接下来将利用下述三个引理

引理 2.1 变量 u_i 的条件期望服从仿射关系

$$\langle u_i\rangle_{\bar{\mathbf{u}}} = q_{i\alpha}\bar{u}_\alpha + c_i, \tag{2.9}$$

其中 $n\times N$ 的矩阵 $Q=\{q_{i\alpha}\}$ 和 n 维向量 $\mathbf{c}=\{c_i\}$ 为

$$\begin{aligned} Q &= (A^{-1}G^T)(GA^{-1}G^T)^{-1}, \\ \mathbf{c} &= A^{-1}\mathbf{b} - (A^{-1}G^T)(GA^{-1}G^T)^{-1}(GA^{-1}\mathbf{b}). \end{aligned} \tag{2.10}$$

引理 2.2 条件协方差矩阵的元素为

$$\begin{aligned} \mathrm{Cov}_{\bar{\mathbf{u}}}(u_i,u_j) &= \langle u_iu_j\rangle_{\bar{\mathbf{u}}} - \langle u_i\rangle_{\bar{\mathbf{u}}}\langle u_j\rangle_{\bar{\mathbf{u}}} \\ &= a_{ij}^{-1} - \big((A^{-1}G^T)(GA^{-1}G^T)^{-1}(GA^{-1})\big)_{ij}. \end{aligned} \tag{2.11}$$

引理 2.3 Wick 定理对于约束条件期望成立, 即

$$\left\langle \prod_{k=1}^{K}\left(u_{i_k}-\langle u_{i_k}\rangle_{\bar{\mathbf{u}}}\right)\right\rangle_{\bar{\mathbf{u}}} = \begin{cases} 0, & K\text{为奇数} \\ \sum_{\{i_1,\ldots,i_K\}}\mathrm{Cov}_{\bar{\mathbf{u}}}(u_{i_1},u_{i_2})\cdots\mathrm{Cov}_{\bar{\mathbf{u}}}(u_{i_{K-1}},u_{i_K}) & K\text{为偶数} \end{cases}$$

其中 $\sum$ 是对于 K 坐标所有可能的对的求和.

引理 2.1 和引理 2.2 可从标准线性归回理论中推导出来. 引理 2.3 可通过利用一个 δ 函数是狭窄 Gauss 函数的极限这一事实而证明. 最终, Gauss 测度在满足约束的函数的子空间上的投影可视为近似 Gauss 的, 从而满足 Wick 定理; 于是可取适当的极限.

2.1.3 最优预测理论的应用

在 [33] 中, Chorin 等人通过将方法应用于两个 Schrödinger 类型的方程来验证其有效性. 之所以选择这些问题是由于非线性 Schrödinger 方程是令人最感兴趣的 Euler/Navier-Stokes 问题的一个一维实例: 它是 Hamilton 的且非线性的. 更流行

的实例 Burgers 方程, 将另作分析; 它的奇特的性质 (激波对解的主导以及需要驱动噪声项来得到一个不变测度) 引入额外的复杂性.

在之后的一系列工作中推广了最优预测 (optimal prediction) 理论及其相关的算法[31, 32, 34, 35]. 在 [34] 中提出依赖于时间的偏微分方程当不能数值求解但具有先验的统计信息时的求解方法. 稀疏的数值数据可视为解上的限制, 而提议的核心为随时间更新限制的一些列方法, 使得回归方法可以用来重新构造未来的均值. 在 [32] 中指出了最优预测与不可逆过程的统计力学之间的联系, 并利用了一种 Mori-Zwanzig 形式来构造一种高阶最优预测方法. 在 [35] 中提出了关于基本方法的一个新的推导, 指出了场论的扰动理论为处理拟线性问题提供了有用的方法, 并提出了一个非线性的例子来说明伪谱方法和具有 Fourier 核的最优预测方法之间的差别. 在 [31] 中采用一个小单元 Monte Carlo 重正化群方法在 Hamilton 系统中求解条件期望的问题. 在后面的内容中将借鉴最优预测的观点来处理构造动力学方法解决复杂网络的分区问题.

2.2　不变集合体的识别

这一节中将介绍的内容是受到了近来提出的识别和计算生物分子的亚稳态化学构造的方法的启发. 给定这些分子关于其动能和势能的物理特征, 这些构型 (conformation) 可被看作相关动力系统的几乎不变子集[49, 177]. 经某种 Markov 算子的离散化之后, 出现一个有限维时齐的马氏链, 它具有可逆性 (reversible), 即关于时间反向是对称的. 每个有限状态空间的马氏链与一个随机转移矩阵相联系. 由于马氏链的可逆性, 其转移矩阵在加权 L^2 意义下是对称的. 识别方法包括由数值解来确定构型和对于 Perron 根 $\lambda = 1$ 周围的特征值分区问题的分析.

本节内容主要参考 [50].

2.2.1　马氏链和转移矩阵

首先, Schütte 等人介绍了一些关于有限维马氏链及其转移矩阵之间关系的基本结果, 这包括随机表征及其相应的线性代数部分.

2.2.1.1 转移矩阵的性质

设 $S = \{s_1, \ldots, s_n\}$ 为时齐的马氏链的限离散状态集合, $n \times n$ 的随机矩阵 $P = \{p_{ij}\}$ 为转移矩阵. 给定处于状态 s_i 的动力系统, 每个转移矩阵元素 p_{ij} 表示系统转移到状态 s_j 的概率. 关于马氏链及其解释的更多内容参见 [180]. 在 2.2 中均假设 P 是本原的, 即存在 $m \in \mathbb{Z}^+$, 使得 $P^m > 0$. 本原随机矩阵有具有一些良好的性质.

定理 2.4 *设 P 为本原随机矩阵, 则有*

(a) Perron根 $\lambda = 1$ *为单重根, 且是占优的, 即对其它 $\lambda \neq 1$ 的特征值均有 $|\lambda| < 1$.*

(b) *存在对应于 $\lambda = 1$ 的正的左右特征向量, 在差常数倍的意义下唯一.*

特别地, 对应于 $\lambda = 1$ 的右特征向量为 $e = (1, \ldots, 1)^T$, 左特征向量 $\pi = (\pi_1, \ldots, \pi_n)^T$ 表示平稳分布, 满足归一化条件 $\pi^T e = 1$, 记为矩阵形式

$$\pi^T P = \pi^T, \quad Pe = e. \tag{2.13}$$

在 [177] 中特征向量 π 被先验地给出. 进一步, 马氏链可逆是已知的, 故细致平衡条件

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \forall i, j \tag{2.14}$$

成立, 或等价地, 记 $\mathscr{D} = \mathrm{diag}(\pi(i))$, 则有

$$\mathscr{D} P = P^T \mathscr{D}. \tag{2.15}$$

为后面分析简便, 这里假设离散状态的选取使得 π 的所有元素严格为正, 即权重矩阵 $\mathscr{D}$ 非奇异; 如果这个假设不满足, 则需要限制状态集 S. 一旦 $\pi > 0$ 成立, 则可以诱导内积 $(\cdot, \cdot)_\pi$ 如下

$$(x, y)_\pi = x^T \mathscr{D} y. \tag{2.16}$$

这个内积相应于有限维加权欧式空间 $L^2_\pi(n)$. 两个向量 x, y 若满足 $(x, y)_\pi = 0$ 则称它们为 π 正交.

命题 2.5 *设 P 为可逆本原随机矩阵, 则 P 关于内积 $(\cdot, \cdot)_\pi$ 是对称的.*

证明 由 (2.15) 可得, $(x, Py)_\pi = x^T \mathscr{D} P y = x^T P^T \mathscr{D} y = (Px, y)_\pi$. □

命题 2.6 设 P 为可逆本原随机矩阵, 则 P 满足如下几条性质[①]

(a) 存在 π 正交的右特征向量构成的基, 将 P 对角化.

(b) P 的所有特征值是实值, 且包含在区间 $[-1, 1]$.

(c) 对 P 的每个右特征向量 x, 对应于相同特征值的左特征向量 y 满足 $y = \mathscr{D} x$.

(d) P 相似于对称的但一般非随机的矩阵 $P_{\text{sym}} = \mathscr{D}^{1/2} P \mathscr{D}^{-1/2}$.

2.2.1.2　非耦合马氏链

作为单个状态 s_i 之间的转移概率的推广, 现将定义状态空间非空子集之间的转移概率, 通常称这些非空子集为集合体 (aggregates).

定义 2.7 给定马氏链的转移矩阵 P (不一定本原) 和平稳分布 $\pi > 0$. 对于任意非空指标子集 I, 定义它的特征向量 $e_I = (e_{I,i})_{i=1,\ldots,n}$, 其中当 $i \in I$ 时有 $e_{I,i} = 1$, 否则 $e_{I,i} = 0$. 记指标集 A 和 B 等同于它们相应的集合体 A 和 B. 则从 A 到 B 的关于 π 的 (条件) 转移概率定义为系统由 A 经一步转移到 B 的条件概率, 即

$$w_\pi(A, B) = \frac{\sum_{a\in A, b\in B} \pi_a p_{ab}}{\sum_{a\in A} \pi_a} = \frac{(e_B, P e_A)_\pi}{(e_A, e_A)_\pi}. \tag{2.17}$$

定义 2.8 设 $A_1, A_2, \ldots, A_N$ 表示状态空间两两不交地分解为 N 个集合体, 则联合随机矩阵 W_π 为 $N \times N$ 的矩阵, 其定义为

$$(W_\pi)_{ij} = w_\pi(A_i, A_j), \quad i, j = 1, \ldots, N, \tag{2.18}$$

称其为此分解的耦合矩阵 (coupling matrix).

对于特殊情形 $A = B$, 则称 $w_\pi(A, A)$ 为系统停留在 A 的概率. 满足 $w_\pi(A, A) = 1$ 的集合体 A 称为不变的 (invariant), 它表示一旦系统处于 A 则将停留在 A 的时

[①] 由命题 2.5, P 在 $L^2_\pi(n)$ 的意义下对称, 于是欧式空间中关于 Hermite 矩阵的相关结论可以类似地推广过来, 详细内容见 [79].

间为 ∞. 马氏链称为非耦合的 (uncoupled), 如果它的状态空间可分解成两两不交的不变集合体 $A_1, \ldots, A_N$, 即

$$w_\pi(A_i, A_j) = \delta_{ij}, \qquad 或 \quad W_\pi = I_{N\times N}. \tag{2.19}$$

严格地说, 由于随机矩阵非本原, 故其平稳分布不唯一, 转移概率独立于任何平稳分布. 对于转移矩阵 P, 具有 N 个集合体的非耦合马氏链在假定合适的状态顺序时, 呈现出对角块形式

$$P = D = \begin{pmatrix} D_{11} & 0 & \cdots & 0 \\ 0 & D_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_{NN} \end{pmatrix}, \tag{2.20}$$

其中每个块 D_{ii} 均为方块随机矩阵, 关于相应的平稳分布是对称的. 再假设每个矩阵 D_{ii} 为本原的, 则由 Perron-Frobenius 定理, 每个块 D_{ii} 具有唯一的对应于其 Perron 根 $\lambda_i = 1$ 的维数为 $\dim(D_{ii})$ 的特征向量 $e_i = (1, \ldots, 1)^T$. 因此, 对于整个的转移矩阵 P, 特征值 $\lambda = 1$ 是 N 重的, 相应的特征子空间由向量

$$\chi_{A_i} = (0, \ldots, 0, e_i, 0, \ldots, 0)^T, i = 1, \ldots, N, \tag{2.21}$$

张成. 用反问题的观点来阐述, Schütte 等人的符号特意强调了这些特征向量可以用不变集合体的特征函数来解释, 如图 2.1(a) 所示. 一般地, 对应于 $\lambda = 1$ 的特征子空间的任意基 $\{X_i\}$ 可以写成特征函数 χ_{A_i} 的线性组合, 系数 $\alpha \in \mathbb{R}$, 即

$$X_i = \sum_{j=1}^{N} \alpha_{ij}\chi_{A_j}, \quad i = 1, \ldots, N. \tag{2.22}$$

从而对应于 $\lambda = 1$ 的特征向量在每个集合体上是常数, 如图 2.1(b) 所示. 有了这些准备, Schütte 等人给出了 2.2.3 中算法的关键结论[50].

引理 2.9 *设对角块形式的随机矩阵 P 由可逆本原的块组成, 平稳分布为 $\pi > 0$, 对应于 $\lambda = 1$ 的特征子空间的 π 正交基为 $\{X_i\}_{i=1,\ldots,N}$. 每个状态 s_i 具有符号结构* (sign structure)

$$s_i \longmapsto \Big(\mathrm{sign}\big((X_1)_i\big), \ldots, \mathrm{sign}\big((X_N)_i\big)\Big). \tag{2.23}$$

则有

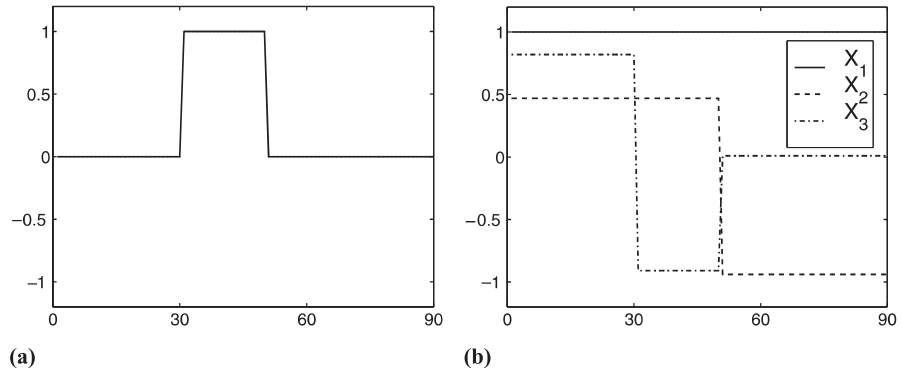


图 2.1: 具有 $N = 3$ 个集合体的非耦合马氏链[50], 状态空间 $\{s_1, \ldots, s_{90}\}$ 被分成集合体 $A_1 = \{s_1, \ldots, s_{29}\}$, $A_2 = \{s_{30}, \ldots, s_{49}\}$, $A_3 = \{s_{50}, \ldots, s_{90}\}$. (a) 特征函数 χ_{A_2}. (b) 对应于 $\lambda = 1$ 的特征子空间的一组基. 可以发现每个特征向量在每个集合体上是常数. 在引理 2.9 的意义下, 状态 s_{69} 的符号结构为 (+, −, 0).

(a) 不变集合体为具有共同符号结构的状态的集合.

(b) 不同的集合体具有不同的符号结构.

证明 为证明结论 (a), 注意到对应于 $\lambda = 1$ 的每个特征向量在每个集合体上均为常数, 这意味着属于相同集合体的状态具有相同的符号结构. 对于结论 (b), 不失一般性, 设每个集合体仅包含一个状态. 第一步, 将证明关于对称矩阵 $P_{\text{sym}} = \mathscr{D}^{1/2} P \mathscr{D}^{-1/2}$ 的一组正交特征向量基 $\{Q_i\}_{i=1,\ldots,N}$ 的结论; 第二步, 将其推广到命题中的结论.

定义 $N \times N$ 的矩阵 $Q = [Q_1 \cdots Q_N]$. 由于 Q 是正交阵, 即 $Q^T = Q^{-1}$, 其转置 Q^T 也为正交阵, 因此 Q 的行正交. 现考虑 P 的一组 π 正交特征向量基 $\{X_i\}_{i=1,\ldots,N}$, 则 $X_i = \mathscr{D}^{-1/2} Q_i$, $i = 1, \ldots, N$. 由于变换矩阵 $\mathscr{D}^{-1/2}$ 有正的对角线元素, 故 X_i 和 Q_i 的符号结构相同, $i = 1, \ldots, N$. 由命题 2.6, 第 m 个集合体的符号结构等于 $X = [X_1 \cdots X_N]$ 的第 m 行的符号结构. 现假设存在两个具有相同符号结构的集合体 A_i 和 A_j, 则 X 的第 i 行和第 j 行符号相同, 故 Q 的第 i 行和第 j 行符号也相同, 这与 Q 的正交性矛盾. □

综上, 引理 2.9 表明对应于 N 重特征值 $\lambda = 1$ 的 N 个右特征向量的集合, 可通过符号结构来识别出 N 个不变集合体. 这种识别是按分量来检验的, 从而独立

于任何 (未知的) 置换. 原则上, 这个检验既可通过左特征向量也可通过右特征向量来实现, 因为它们的符号结构相同. 因为对于每个左特征向量 $y = (y_i)$, 存在相应的右特征向量 $x = (x_i)$ 满足 $y_i = \pi_i x_i$, 故 $\mathrm{sign}(y_i) = \mathrm{sign}(x_i)$. 由于它们的常数水平结构, 右特征向量在处理下述带有扰动的的反问题时更为合适.

2.2.2 几乎非耦合马氏链

在多数实际生活的应用中, 包括分子动力学, 扰动的出现产生几乎非耦合马氏链而不是非耦合马氏链, 相应的分解为几乎不变集合体而不是不变集合体. 粗略地说, 只要动力系统处于几乎不变集合体中, 它将停留于此很长时间而不是无穷时间, 因此描述这种情形用亚稳性而不是稳定性. 在转移矩阵方面, 将出现快对角占优矩阵而不是快对角矩阵. 后面的结果将要表明, 基于一些微妙的扰动分析, P 的右特征向量可再次用来识别这样的集合体.

2.2.2.1 扰动分析

本节的扰动分析理论紧随 Stewart 在一般本原随机矩阵方面的工作 [186]. 但这里 Schütte 等人假设马氏链可逆, 并利用 2.2.1 中的准备知识. 在即将处理的扰动情形中, 随机矩阵的平稳分布 π 唯一, 故关于 π 的内积良定义. 对于本节的扰动分析, 将采用 [99] 中的著名理论, 具体讨论命题 2.5 意义下的对称矩阵情形.

由 (2.19) 知, 不变集合体 A 定义为 $w_\pi(A, A) = 1$. 因此粗略地说, 若 $w_\pi(A, A) \approx 1$, 则称集合体 A 为几乎不变的 (almost invariant). 同样地, 马氏链称为几乎非耦合的 (nearly uncoupled), 如果它的状态空间可分解为 N 个两两不交的几乎不变集合体 $A_1, \ldots, A_N$, 使得

$$w_\pi(A_i, A_j) \approx \delta_{ij} \quad \text{或} \quad W_\pi \approx I_{NN}. \tag{2.24}$$

在这种情形下, 具有 N 个集合体的几乎非耦合马氏链 (NUMC) 的状态可重新排

序, 使得转移矩阵 P 为块对角占优形式

$$P = D + E = \begin{pmatrix} D_{11} & E_{12} & \cdots & E_{1N} \\ E_{21} & D_{22} & \cdots & E_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ E_{N1} & E_{N2} & \cdots & D_{NN} \end{pmatrix}. \tag{2.25}$$

这里, 扰动矩阵 E 满足 $E = O(\epsilon)$, 其中 ϵ 为扰动参数, 将在 2.2.2.2 中将做进一步说明. 目前, 只需分析本质的块对角占优结构作为这个扰动参数的函数. 设 $P(\epsilon)$ 为一族矩阵, 定义 ϵ_* 使得 $P(\epsilon_*) = P$. 注意到在应用中不仅 ϵ_* 的实际大小未知, 而且 (2.25) 中的块的个数N也未知. 这两方面依赖于度量集合体之间弱耦合的标准的选择, 这些将在 2.2.2.2 中阐述. 为完成线性扰动分析, Schütte 等人采用 [99] 中的假设.

正则化条件 *(Regularity conditions, RC)*. 设一族矩阵

$$P(\epsilon) = P(0) + \epsilon P^{(1)} + \epsilon^2 P^{(2)} + \cdots \tag{2.26}$$

在复平面的一个包含原点的区域内解析, 使得对实值 ϵ, $P(\epsilon)$ 为可逆的随机矩阵. 进一步, 设对于实值 $\epsilon \neq 0$, $P(\epsilon)$ 是本原的, 且 $P(0)$ 为具有本原块 $D_{ii}, i = 1, \ldots, N$ 的块对角形式 (2.20). 由定理 2.4, 每个 $P(\epsilon)$ 具有唯一的正的平稳分布 $\pi(\epsilon)$. 假设所有 $\pi(\epsilon)$ 的集合一致有界远离 0 值, 即存在常数 $C > 0$, 使得 $\pi_i(\epsilon) \geq C, i = 1, \ldots, n$, 实值 ϵ 包含 $\epsilon = 0$.

这些正则化条件保证, 对于充分小的 $\epsilon \in \mathbb{R}$, 特征值关于 ϵ 连续, 且 $P(\epsilon)$ 的谱可分为三部分[99, 134, 186]

(1) Perron 根 $\lambda_1(\epsilon) \equiv 1$.

(2) 当 $\epsilon \to 0$ 时, $N-1$ 个特征值 $\lambda_2(\epsilon), \ldots, \lambda_N(\epsilon)$ 的簇接近于 1.

(3) 谱的尾部, 当 $\epsilon \to 0$ 时远离值 1.

换句话说, 对于充分小的实值 ϵ, 在 Perron 根附近存在 N 个特征值的良识别簇, Perron 簇, 可认为是扰动下 N 重 Perron 根的分解. 下述定理给出了对应于 Perron 簇的特征向量 $X_1(\epsilon), \ldots, X_N(\epsilon)$ 的描述[50].

定理 2.10 设 $P(\epsilon)$ 为满足正则化条件 (RC) 的一族矩阵, Π_j 表示由非扰动转移矩阵 $P(0)$ 的特征向量 X_j 所张成的特征子空间上的 π 正交投影, 则对于实值 ϵ, 存在如下形式的 π 正交特征向量 $X_1(\epsilon),\dots,X_N(\epsilon)$

(a) 对应于 Perron 根 $\lambda_1(\epsilon)\equiv 1$ 的特征向量为

$$X_1(\epsilon)\equiv(1,\dots,1)^T, \tag{2.27}$$

(b) 对应于在 $\lambda=1$ 附近的特征值簇 $\lambda_2(\epsilon),\dots,\lambda_N(\epsilon)$ 的 $N-1$ 个特征向量集合具有形式

$$X_i(\epsilon)=\sum_{j=1}^{N}\alpha_{ij}\chi_{A_j}+\epsilon X_i^{(1)}+O(\epsilon^2), \tag{2.28}$$

其中

$$X_i^{(1)}=\sum_{j=1}^{N}\beta_{ij}\chi_{A_j}+\sum_{j=N+1}^{n}\frac{1}{1-\lambda_j}\Pi_j P^{(1)}X_i, \tag{2.29}$$

$\alpha_{ij},\beta_{ij}\in\mathbb{R}$ 为适当系数, 集合体 $A_1,\dots,A_N$ 对应于 $P(0)$ 的块对角形式.

证明 由于对实值 $\epsilon\neq 0$, $P(\epsilon)$ 为本原的, 特征值 $\lambda_1(\epsilon)\equiv 1$ 为单重的, 相应的左特征向量 $\pi(\epsilon)$, 即平稳分布, 为正向量且对于实值 ϵ 解析[99]. 定义变换矩阵 $\mathscr{D}=\mathrm{diag}(\pi_i(\epsilon))$, 由于 $\pi(\epsilon)$ 一致有界远离值 0 值, 则 $\mathscr{D}$ 对于实值 ϵ 是可逆的. 于是变换的矩阵族 $P_{\mathrm{sym}}(\epsilon)=\mathscr{D}^{1/2}P(\epsilon)\mathscr{D}^{-1/2}$ 在 ϵ 解析且对于实值 ϵ 对称.

由 [99] 知, 存在 $P_{\mathrm{sym}}(\epsilon)$ 的对应于特征值 $\lambda_1(\epsilon),\dots,\lambda_N(\epsilon)$ 的右特征向量 $Y_1(\epsilon),\dots,Y_N(\epsilon)$, 它关于实值 ϵ 解析. 用 $\mathscr{D}^{-1/2}$ 作用于这些向量, 得到 $X_i(\epsilon)=\mathscr{D}^{-1/2}Y_i(\epsilon)$, 相应的可逆矩阵 $P(\epsilon)$ 的特征向量对于实值 ϵ 解析, 从而在 ϵ 有展开 $X_i(\epsilon)=X_i+\epsilon X_i^{(1)}+O(\epsilon^2)$.

现令 $\Pi(\epsilon)=\Pi_1(\epsilon)+\cdots+\Pi_N(\epsilon)$ 表示 $P(\epsilon)$ 的相应于特征值 $\lambda_1(\epsilon),\dots,\lambda_N(\epsilon)$ 的特征子空间上的 π 正交投影, 则由 [99] 知, $\Pi(\epsilon)$ 在 ϵ 解析且

$$\Pi(\epsilon)=\Pi(0)+\epsilon\sum_{j=N+1}^{n}\frac{1}{1-\lambda_j}\Big(\Pi(0)P^{(1)}\Pi_j+\Pi_jP^{(1)}\Pi(0)\Big)+O(\epsilon^2). \tag{2.30}$$

将 $X_i(\epsilon)=X_i+\epsilon X_i^{(1)}+O(\epsilon^2)$ 代入恒等式 $X_i(\epsilon)=\Pi(\epsilon)X_i(\epsilon)$, $i=1,\dots,N$, 得到

$$X_i^{(1)}=\sum_{j=1}^{N}\tilde{\beta}_{ij}X_j+\sum_{j=N+1}^{n}\frac{1}{1-\lambda_j}\Pi_jP^{(1)}X_i, \tag{2.31}$$

其中适当选取系数 $\tilde{\beta}_{ij} \in \mathbb{R}$. 再由 (2.22) 则完成证明. □

结合 (2.28) 和 (2.29), 定理 2.10 的一阶扰动结果

$$X_i(\epsilon) = \underbrace{\sum_{j=1}^{N}(\alpha_{ij} + \epsilon\beta_{ij})\chi_{A_j}}_{\text{(I)}} + \underbrace{\epsilon \sum_{j=N+1}^{n} \frac{1}{1-\lambda_j}\Pi_j P^{(1)} X_i + O(\epsilon^2)}_{\text{(II)}} \tag{2.32}$$

表明, (I) 仅变化 (上或下的) 局部常数水平线, 这与几乎不变集合体有关, 这部分误差不会破坏符号结构. (II) 具有形式 $\epsilon B + O(\epsilon^2)$, 可在某种程度上破坏常数水平模式, 但仅仅在很小的程度上影响引理 2.9 中的符号结构, 并对于任何“几乎为零”的水平线的扰动应采取谨慎态度. 上述两项关于“弱模式” X_i, $i = N+1, \ldots, n$, 和“主模式” X_i, $i = 1, \ldots, N$有进一步的解释: (I) 表示“主–主”耦合, 而 (II) 表示“弱–主”耦合. 最后, 可以观察到 (II) 主要依赖于 Perron 根和谱的尾部之间的谱间隙 (spectral gap) $1 - \lambda_{N+1}$, 而不是 Perron 簇和谱的尾部之间的谱间隙 $\lambda_N - \lambda_{N+1}$.

2.2.2.2　集合体间的弱耦合

现在讨论如何定义比模糊假设 (2.24) 更精确的扰动参数 ϵ. 一旦计算出 N 个几乎不变集合体, 那么由定义 2.7 就可计算 (N, N) 耦合矩阵 W_π. 在此基础上, 称马氏链几乎非耦合 (nearly uncoupled), 如果

$$\|W_\pi - \mathrm{diag}(W_\pi)\|_\infty = 1 - \min_i w_\pi(A_i, A_j) = \epsilon_* \tag{2.33}$$

成立, 其中 $\mathrm{diag}(W_\pi) = \mathrm{diag}((w_\pi)_{11}, \ldots, (w_\pi)_{NN})$, $\epsilon_* > 0$ 充分小 (对比前述定义 $P(\epsilon_*) = P$). 由定义 2.7 和关系 (2.33) 可证明转移矩阵 (2.25) 的扰动有上界

$$\|\mathscr{D}E\|_\infty \leq \epsilon_*, \tag{2.34}$$

这里Schütte 等人关于几乎非耦合马氏链的描述与非耦合度量不同[85], 但与马氏链的传导概念有关[183].

基于上述说明, 现在来讨论识别过程. 如 2.2.2.1 结尾所述, Schütte 等人试图在理论基础上研究上述扰动结果的符号结构. 假设确认过程通过符号

结构给出几乎不变集合体 $A_1,\dots,A_N$, 记 $n\times N$ 的矩阵 $\chi=[\chi_{A_1}\cdots\chi_{A_N}]$ 和 $X=[X_1\cdots X_N]=X(\epsilon)$, 扰动结果 (2.32) 在实际计算中可表示为

$$X=\chi\mathscr{A}^{-1}+\epsilon B+O(\epsilon^2), \tag{2.35}$$

其中 $\mathscr{A}=\mathscr{A}(\epsilon)$ 为 $N\times N$ 系数矩阵, B 为 $n\times N$ 矩阵, 表示 (2.32) 中的"弱–主"耦合项 (II). 从本质扰动理论的角度上, 可以通过最小二乘拟合

$$\arg\min\|\chi_{A_i}-\sum_{j=1}^{N}a_{ij}X_j\|_\pi,\quad i=1,\dots,N. \tag{2.36}$$

来确定非奇异系数矩阵 $\mathscr{A}=\{a_{ij}\}$. 记 $\Lambda=\mathrm{diag}(\lambda_1,\dots,\lambda_N)=\Lambda(\epsilon)$ 为特征值的 Perron 簇, 则耦合矩阵可写成

$$W_\pi=(\chi^T)\mathscr{D}\chi)^{-1}(\chi^T\mathscr{D}P\chi)=\mathscr{A}^{-1}\Lambda\mathscr{A}+\epsilon\Delta, \tag{2.37}$$

其中 $\mathscr{A}^{-1}\Lambda\mathscr{A}$ 描述"主–主"相互作用, 而矩阵 Δ 表示如下的"弱–主"相互作用

$$\Delta=X^T\mathscr{D}B\Lambda-\Lambda B^T\mathscr{D}X+O(\epsilon). \tag{2.38}$$

在非耦合马氏链情形中, 有 $B=0$ 和 $\Lambda=I_{NN}$, 从而 $\Delta=0$ 且 $W_\pi=\mathscr{A}^{-1}\Lambda\mathscr{A}=I_{NN}$. 对于要考虑的几乎非耦合马氏链, 自然希望 $\mathscr{A}^{-1}\Lambda\mathscr{A}\approx I_{NN}$, 从而Schütte 等人将 $N\times N$ 的矩阵

$$\mathrm{err}(A_1,\dots,A_N)=\epsilon\Delta=W_\pi-\mathscr{A}^{-1}\Lambda\mathscr{A} \tag{2.39}$$

作为误差指示器. 根据 2.2.2.1 中的扰动分析, 这个指示器度量弱模式和主模式的耦合程度, 即它度量了不能由 Perron 簇的特征子空间表征的耦合部分. 如果 $\mathrm{err}(A_1,\dots,A_N)$ 的一个元素很大, 则是由于下列原因之一:

(1) ϵ_* 的值不是"足够小", 无法保证线性扰动分析.

(2) 正则化条件被破坏.

(3) 识别算法给出"错误"的几乎不变集合, 这个现象可在扰动严重破坏符号结构时出现. 当然, 这与第一条原因有重叠.

2.2.3　识别算法

下面来介绍几乎不变集合体的识别算法的具体执行. 算法的关键思想是: 通过对应于特征值的 Perron 簇的特征向量的符号结构来按分量地识别几乎不变集合体.

首先, 需要确定几乎不变集合体的个数 N. 这需要计算 $\lambda = 1$ 附近的特征值簇, 即 Perron 簇, 它与谱的尾部有谱间隙. Schütte 等人的算法中简单应用直接的特征值求解器计算全部特征值, 并通过检验分离出 Perron 簇. 其次, 一旦 $N-1$ 个右 Perron 特征向量 (除已知的特征向量 e 之外) 已经计算出来, 于是想将状态空间分解为 N 个几乎不变集合体. 如 2.2.1.2 所述, 这可以通过使用特征向量的"分片常数水平线"结构或符号结构来实现. 然而对于几乎非耦合马氏链, 特征向量的扰动和置换都将这些结构掩盖到一个未知程度, 这使得构造有效的识别算法成为艰巨的任务. Schütte 等人所提出的识别算法的主要的三个步骤如下: (1) 选取具有稳定符号结构的状态. (2) 定义符号结构的类. (3) 识别分片常数水平线模式. 详细的描述参见 [50].

几乎不变集合体的识别算法受到了近年来的识别以及计算生物分子的亚稳态化学构型的方法的启发, 在一系列相关工作[47, 49, 177]的基础上演化而成的. 在 [47] 中首次提出了一种复杂动力行为的数值近似的有效方法, 特别是近似 SRB (Sinai-Ruelle-Bowen) 测度和动力系统的几乎循环行为的数值方法, 这种方法是建立在 Frobenius-Perron 算子离散化的基础之上, 并结合了紧算子有限维近似的经典收敛结果和关于由随机扰动系统不变测度近似 SBR 测度的遍历论中的结果. 在 [49] 中则提出了一种新型的分子动力系统中的约化动力学, 即由 Hamilton 微分方程描述的系统的数值计算的方法, 它通过对于 Frobenius-Perron 算子的特征模块应用多层再分算法来直接计算目标. 在 [177] 中提出了另一个相关的概念, 将动力系统方法概念上的优点与适当的统计物理框架合并起来, 定义关于分子系统动力行为的构型 (conformation), 并刻画构型的动力稳定性.

2.3　MNCut 方法

本节所介绍的内容揭示出一种概率性的解释, 它可以作为一种分析工具, 为

所有的谱方法[86]提供见解和服务[132]. Meila 和 Shi 将点对间的相似度看作 Markov 随机游动的边上的流, 并研究其转移矩阵的特征值和特征向量的性质. 利用这个观点, 许多谱方法在某种意义上都可以包含在后面将要介绍的 Normalizied Cut (NCut) 图像分割算法[182]之中. 因此, 下面的内容将集中在 NCut 算法上, 并采用图像分割这一术语 (即数据点为像素, 所有像素的集合为图像), 其介绍的所有结果对于基于相似度的分区算法都是有效的.

本节内容主要参考 [132].

2.3.1 NCut 标准与算法

这里将图像表示为像素集 S, 一个分区是将 S 分为两两不交的子集的分划. 对于每个像素对 $i, j \in S$, 给定相似度 $w_{ij} = w_{ji} \geq 0$. 在 Ncut 框架中, 相似度 w_{ij} 可视为 S 上的图 G 的边 e_{ij} 上的权重, 若 $w_{ij} = 0$ 则 G 没有边 e_{ij}. 矩阵 $W = \{w_{ij}\}$ 为 G 的实值邻接矩阵. 设 $d_i = \sum_{j \in S} w_{ij}$ 为节点 i 的度, $\mathrm{vol}A = \sum_{i \in A} d_i$ 为集合 $A \subset S$ 的体积. 集合 A 与其余集 $\bar{A}$ 之间的边的集合称为边割 (edge cut), 或简称切割 (cut). 正规化切割 (NCut) 标准[182]是将图像分割成两部分的一个图论的标准, 通过对所有的切割 $A, \bar{A}$, 极小化

$$\mathrm{NCut}(A, \bar{A}) = \left(\frac{1}{\mathrm{vol}A} + \frac{1}{\mathrm{vol}\bar{A}} \right) \sum_{i \in A, j \in \bar{A}} w_{ij}. \tag{2.40}$$

极小化 NCut 意味着寻找一个切割, 使得相应的两个子集之间的权重较小而内部连接的权重较大. 最优化 NCut 标准是 NP 难题[182].

在 [182] 中介绍的 NCut 算法是通过特征值和特征向量来求解极小化 NCut 问题的近似方法. 此方法采用了 Laplace 矩阵 $L = D - W$, 其中 D 是节点的度所组成的对角矩阵. 算法的构成是求解广义特征值和特征向量问题

$$Lx = \lambda Dx. \tag{2.41}$$

NCut 算法关注 (2.41) 的第二小特征值及其对应的特征向量, 分别记为 λ^L 和 x^L.

图 2.2 给出了一个具有显著块结构 (Ib) 的相似度矩阵及其前 3 个广义特征向量 (IIIa) 的例子. 从图中可见 x^L 的的元素在每个分区中近似相等. 在 [182] 中指出

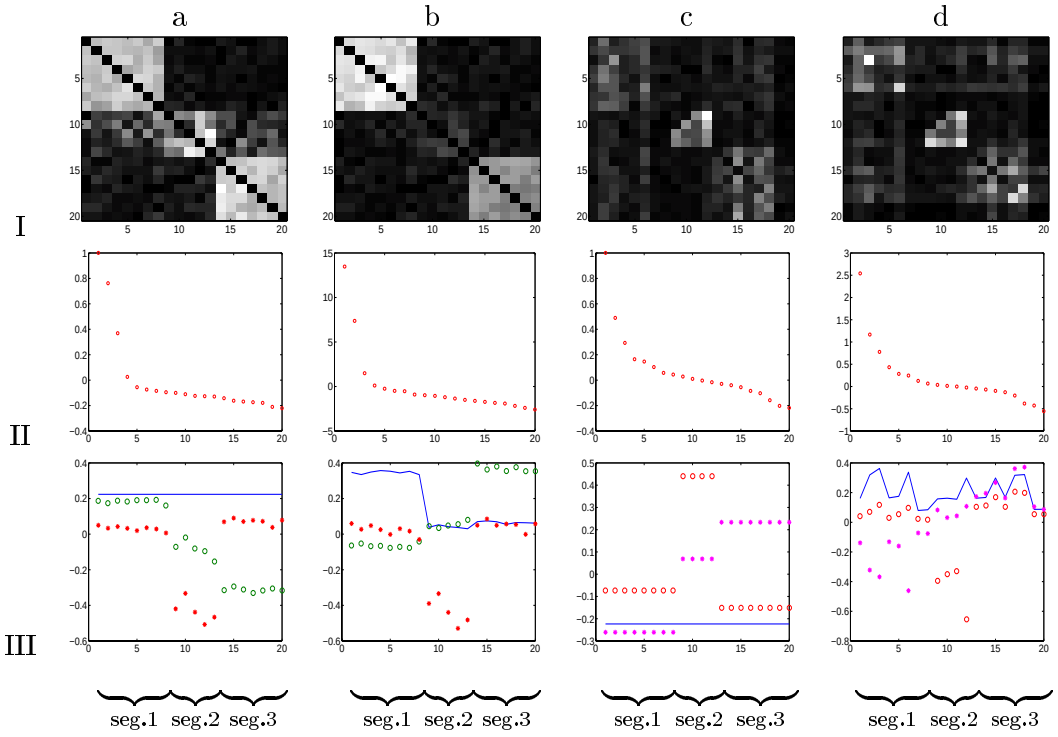


图 2.2: 四个矩阵 (I 行) 及其特征值 (II 行) 和前3个特征向量: – 表示 x^1, ∘ 表示 x^2 (在 b, d 中 $= x^L$), ⋆ 表示 x^3 (III 行)[132]. 所有的矩阵都用灰度表示, 其中黑色表示 0, 颜色越浅表示值越高. 所有的矩阵对应于 20 个像素形成 3 个分割的图像. (a) 近似块对角随机矩阵 P_1. 第二和第三特征向量近似分片常数并且包含关于分割的信息. (b) 生成 P_1 的对称的相似度矩阵. 注意所有三个特征向量均包含关于分割的信息. 求解 (2.41) 所得到的关于这个矩阵的特征向量和 P_1 的特征向量相同. (c) 块随机矩阵 P_2. 第二和第三特征向量分片常数, 反映出正确的分割. (d) 生成 P_2 的对称的相似度矩阵. 前三个特征向量仅仅为粗略的分片常数并导致错误的分割.

了当存在 S 的一个分划 $A, \bar{A}$, 使得

$$x_i^L = \begin{cases} \alpha, & i \in A \\ \beta, & i \in \bar{A} \end{cases} \tag{2.42}$$

则 $A, \bar{A}$ 是最优 NCut, 切割的值为 $\mathrm{NCut}(A, \bar{A}) = \lambda^L$. 这个结果描述了用正规化切割来进行谱分割的基础. 首先求解广义谱问题 (2.41), 然后寻找将 x^L 的元素分成两个包含着几乎相等的值的分割, 这个分割可通过对元素设定阀值来得到. 特征向量的分划诱导出 S 上的一个分区, 这就是期望得到的分区. 若想要得到多于两个分

区只需递归地重复上述过程, 则称这个过程为 NCut 算法. 称满足 (2.42) 的向量关于分区 $(A, \bar{A})$ 为分片常数. 在 2.3.3 中, 将考虑关于将 S 分割成 N 个集合的分区的分片常数的特征向量.

如上所述, NCut 算法缺少令人满意的直观解释, 特别是 NCut 算法和标准关于下列问题提供很少的直观说明: (1) 什么导致特征向量为分片常数? (2) 当存在两个以上的分割时将会发生什么情况? (3) 当 x^L 不是分片常数时, 算法性能将如何衰减? 接下来即将描述的随机游动的解释将回答前两个问题, 并给出关于得到的谱分区的良好理解. 第三个问题将不在此处理, [98] 中的结果反应出了 NCut 算法应用效果.

2.3.2 马氏链和正规化切割

将相似度矩阵 W 归一化, 可以得到随机矩阵 $P = D^{-1}W$, 其行和均为 1. 由马氏链理论知, p_{ij} 表示从节点 i 经一步转移到节点 j 的概率. P 的特征值为 $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq -1$, 对应的特征向量为 $x^1, x^2, \ldots, x^n$. P 的第一个特征向量 $x^1 = \mathbf{1} = (1, \ldots, 1)^T$. 不失一般性, 这里假设没有节点的度为 0.

现考虑矩阵 P 的谱问题, 也就是求解方程

$$Px = \lambda x, \quad x \neq 0. \tag{2.43}$$

命题 2.11 *若 (λ, x) 是 (2.43) 的解, 且 $P = D^{-1}W$, 则 $(1 - \lambda, x)$ 是 (2.41) 的解.*

换句话说, NCut 算法和矩阵 P 具有相同的特征向量, P 的特征值等于 1 与 (2.41) 中广义特征值的差. 命题 2.11 指出了由 NCut 算法抽象出的谱问题和随机矩阵 P 的特征值及特征向量之间的等价性. 这也帮助解释了为什么 NCut 算法利用第二小广义特征向量: (2.41) 最小的特征向量对应于 P 的最大特征向量, 而 P 的最大特征向量多数情况等于 $\mathbf{1}$, 故不包含信息.

NCut 标准也可在此框架下加以理解. 首先定义 $\pi = \{\pi_i\}_{i \in S}$ 如下

$$\pi_i = \frac{d_i}{\mathrm{vol}S}. \tag{2.44}$$

容易证明 $\pi P = \pi$, 因此 π 为马氏链的一个平稳分布. 如果马氏链是遍历的[②], 则 π 是 S 上唯一的平稳分布. 同时注意到马氏链是可逆的, 由于

$$\pi_i p_{ij} = \pi_j p_{ji} = \frac{w_{ij}}{\mathrm{vol}S}. \tag{2.45}$$

设状态集 $A, B \subset S$, 如果当前状态处于 A 且随机游动开始于它的平稳分布, 则定义 $\hat{p}_{AB} = Pr(A \to B|A)$ 为随机游动从集合 A 经一步转移到集合 B 的概率,

$$\hat{p}_{AB} = \frac{\sum_{i\in A, j\in B} \pi_i p_{ij}}{\pi(A)} = \frac{\sum_{i\in A, j\in B} w_{ij}}{\mathrm{vol}A}, \tag{2.46}$$

从而可以得到

$$\mathrm{NCut}(A, \bar{A}) = Pr(A \to \bar{A}|A) + Pr(\bar{A} \to A|\bar{A}) = \hat{p}_{A\bar{A}} + \hat{p}_{\bar{A}A}. \tag{2.47}$$

如果对于某个分区 $A, \bar{A}$, NCut 很小, 则意味着一旦游动处于 A 中从 A 逃出和概率, 和一旦处于 $\bar{A}$ 中从 $\bar{A}$ 逃出的概率, 都很小. 直观上, 将状态集 S 分割成两部分, 使得随机游动移动处于其中一个部分, 则倾向于停留在里面.

NCut 与马氏链中的低传导集有密切联系. 低传导集 A 是 S 的一个子集, 使得 $h(A) = \max(\hat{p}_{A\bar{A}}, \hat{p}_{\bar{A}A})$ 的值很小. 这在与马氏链的混合时间有关的谱图理论中有过研究[36]. 近来 [98] 利用了它们定义了一种分区的新标准, 其中的启发式分析与 NCut 算法十分相似.

2.3.3　随机矩阵特征向量的分片常数性质

接下来, Meila 和 Shi 将利用随机矩阵 P 来得到一个关于 NCut 算法的更好的解释. 由于 NCut 算法关注 P 的第二大特征值, 记为 $x^2 = x^L$, 目的是得到 S 的一个分割. 定义向量 x 关于 S 的分区 $\mathbb{A} = \{A_1, A_2, \ldots, A_N\}$ 为分片常数的, 当且仅当对于在相同集合 $A_k, k = 1, \ldots, N$ 中的像素 i, j, 有 $x_i = x_j$. 注意到 P 的第一个特征向量 $\mathbf{1}$ 总是分片常数的. 由于具有分片常数的特征向量对谱分割是必需的, 故研究 P 何时具有这个特征向量分片常数的性质至关重要. 下面研究 n 个特征向量的前 N 个为分片常数的情形.

[②] 遍历性发生在温和的条件下[36].

命题 2.12 设 P 为行和列的指标索引于 S 的矩阵, 并且具有独立的特征向量, 设 $\mathbb{A} = \{A_1, A_2, \ldots, A_N\}$ 为 S 的一个分割, 则 P 有 N 个非零特征值所对应的特征向量关于 $\mathbb{A}$ 为分片常数的, 当且仅当所有的求和 $p_{il} = \sum_{j \in A_l} p_{ij}$ 为常数, $\forall i \in A_k, \forall k, l = 1, \ldots, N$, 且矩阵 $R = (p_{kl})_{k,l=1,\ldots,N}$ 非奇异, 其中 $p_{il} = \sum_{j \in A_l} p_{ij}, i \in A_k$.

命题 2.13 如果 n 维矩阵 P 具有形式 $P = D^{-1}W$, 其中 W 为对称矩阵, D 为非奇异矩阵, 则 P 具有 n 个独立的特征向量.

命题 2.12 的证明详见 [132]. 称满足命题 2.12 中条件的随机矩阵 P 为块随机矩阵. 直观上, 命题 2.12 指出, 如果随机矩阵从像素 i 转移到分割 A_k 的概率等于所有与 i 在同一分区中的像素转移到 A_k 的概率, 则它具有分片常数的特征向量. 在 [98, 182, 202] 中表明, 对于非连通图 G (导致块对角的 W), NCut 算法和其它算法正确地运行. 块对角的 W 意味着不同分区中的像素非常不同. 这个情况如图 2.2(a,b) 所阐述, 这是目前分区问题最简单的情形. 现在命题 2.12 表明, 谱分区实际上能够通过转移概率的相似性来将像素分组成为 S 上的子集. 这个情形如图 2.2(c,d) 所示. 数值试验 [182] 表明 NCut 在很多连通图上效果良好, 从而用实践证据支持此结论. 然而, 具有分片常数的特征向量仅是研究的一部分. 此外还需要求 R 的对应于分片常数特征向量的特征值大于 P 的其它 $n - N$ 个特征值, 这 $n - N$ 个特征值也称伪特征值 (spurious eigenvalues).

基于上述观点, Meila 和 Shi 定义了一个理论算法, 称为修正 NCut (Modified NCut, MNCut), 来寻找像素集合中所有的 N 个分区, 步骤如下:

(1) 由 W 计算 P 及其特征值和特征向量.

(2) 选取最大的 N 个特征值和对应的特征向量.

(3) 通过寻找在已选取的特征向量中的几乎相等的元素来求出分区.

最后一步可以通过, 例如, 向由行 $(x^2, \ldots, x^N)$ 所定义的 $N - 1$ 维空间上投影, 或在其上运用 k-means 算法 (N 已知), 来实现.

命题 2.14 如果 P 是块随机矩阵, R 的特征值大于伪特征值, 则 MNCut 算法是精确的.

因此, MNCut 同时考虑了不同分区中像素的相异性和相同分区中像素转移的相似性. MNCut 方法具有另一个优点: 如果在 R 的特征值和伪特征值之间存在谱间隙 (如图 2.2(c,d) 所示), 则分区的数目 N 可以自动确定. 这当满足下面条件时很有可能发生: (i) R 接近于单位矩阵, 其特征值趋于 1; (ii) P 的行在相同的分区中趋于相等, 伪特征值趋于 0. 因此, 再一次地, 分区之间的相异性和转移的相似性的混合可描述具有自然分区结构的数据集.

2.4　扩散映射和粗粒化

本节中的内容指出非线性降维, 聚类以及数据集参数化可以由同一框架来解决. 主要的思想是定义具有显式度量的坐标系统来反映给定数据集的连接程度, 并且可以抗噪声. 框架的构造基于数据上的随机游动, 利用内在几何学对图和高维任意形状的数据集的改造和在抽样提供了普遍性的方案.

本节内容主要参考 [107].

2.4.1　作为高维数据分析工具的几何扩散

Lafon 和 Lee 的目标是在任意集合中定义距离度量, 来反映集合中点的连接程度. 假设所处理的数据集以图的形式存在. 当在图中识别社团时, 需要度量节点对之间相互联系的总量. 根据这个想法, 如果两个点由图中很多条较短路径相连接, 则认为它们是相近的. 因此, 高密度区域中的点 (定义为图中具有高节点度的节点的组) 将具有较高的连接程度. 进一步, 连接程度由图中权重决定. 下面主要回顾在 [41] 中首次提出的扩散框架, 并将其置入特征映射, 降维以及图上的随机游动的框架中.

2.4.1.1　扩散距离

设 $G = (S, W)$ 为 n 个节点的有限图, 其中权重矩阵 $W = \{w(x, y)\}_{x,y\in S}$ 满足下面两个条件: (i) 对称性: $W = W^T$; (ii) 逐点正性: $w(x, y) \geq 0, \forall x, y \in S$. 定义权重的的方式完全出于应用的考虑, 唯一的要求是 $w(x, y)$ 需表示 x 和 y 的相似度 (具体应用中定义). 特别地, 期望 $w(x, y)$ 为正数. 例如, 如果处理流形上的数据点,

可首先定义 Gauss 核 $w_\varepsilon = \exp(-\|x-y\|^2/\varepsilon)$, 然后为了调整流形上点的几何的影响和分布将其归一化. 不同的归一化方案和它们与流形上的 Laplace-Beltrami 算子的联系, 当大样本极限 $n \to \infty$ 和 $\varepsilon \to 0$ 时于 [40] 中有详细讨论.

具有权重 W 的图 G 描述了关于集合局部几何的认识. 接下来定义图上的随机游动. 为此引入节点 x 的度 $d(x)$ 如下

$$d(x) = \sum_{z \in S} w(x,z). \tag{2.48}$$

如果定义 $n \times n$ 的矩阵 P, 其元素为

$$p_1(x,y) = \frac{w(x,y)}{d(x)}, \tag{2.49}$$

则 $p_1(x,y)$ 可解释为从 x 经 1 个时间步转移到 y 的概率. 由构造知, 这个量反映图的一阶临近结构. 扩散映射框架中引入的一个新观点是通过取矩阵 P 的幂来捕获更多的邻居的信息, 或等价地, 在时间上向前运行随机游动. 如果 P^t 是 P 的第 t 次重复, 则其元素 $p_t(x,y)$ 表示从 x 经 t 个时间步转移到 y 的概率. 增加 t, 对应于传播每个节点对它的邻居的局部影响. 换句话说, P^t 这个量反映在一个扩散过程中, 由图的连接程度所定义的数据集内在几何性, 并且扩散的时间 t 在分析中起到了尺度参数的作用.

如果图是连通的, 由 [36] 中的介绍, 有

$$\lim_{t \to +\infty} p_t(x,y) = \mu(y), \tag{2.50}$$

其中 μ 是唯一的平稳分布

$$\mu(x) = \frac{d(x)}{\sum_{z \in S} d(z)}. \tag{2.51}$$

这个量正比于图中 x 的度, 节点的度可度量点的密度. 此外, 马氏链可逆, 即满足下面的细致平衡条件

$$\mu(x) p_1(x,y) = \mu(y) p_1(y,x). \tag{2.52}$$

Lafon 和 Lee 主要考虑下述观点: 对于固定的有限的 $t > 0$, 想要定义 S 中点的一个度量, 使得两个点 x 和 y 将临近, 如果它们相应的条件分布 $p_t(x,\cdot)$ 和$p_t(y,\cdot)$ 接近. 一个相似的想法出现在 [191] 中, 其作者考虑 L^1 模 $\|p_t(x,\cdot) - p_t(y,\cdot)\|$. 作为

选择, 可利用 Kullback-Leibler 距离或其它任何 $p_t(x,\cdot)$ 和 $p_t(y,\cdot)$ 间的距离. 然而下面将指出, 条件分布之间的 L^2 度量具有将距离与随机游动的谱性质联系起来的优点, 从而通过特征映射将图上的 Markov 随机游动学习与数据参数化联系起来. 正如 [137] 中所述, Lafon 和 Lee 定义 x 和 y 之间的扩散距离 D_t 为有权重的 L^2 距离

$$D_t^2(x,y) = \|p_t(x,\cdot) - p_t(y,\cdot)\|_{1/\mu}^2 = \sum_{z\in S} \frac{(p_t(x,z) - p_t(y,z))^2}{\mu(z)}, \tag{2.53}$$

其中权重 $1/\mu(x)$ 惩罚低密度区域多于高密度区域的差异. 这个图中点的接近程度的概率反映了在扩散过程中关于数据点连接程度的集合的内在几何性. 两点间的扩散距离将较小如果它们由图中许多路径相连接. 因此这个度量是设计那些对于给定假设下基于证据优势的推断算法中的一个关键的量. 例如, 假设基于少量标示样本来推断数据点的分区标示, 则可以遵循如下方式简单地将标示信息从一个标示样本 x 传播到新的标示样本 y: (i) 最短路径, (ii) 所有连接 x 和 y 的路径. 第二种解决方式 (在扩散框架工作和 [191] 中采用) 通常更为合适, 因为它考虑了所有联系 x 和 y 的证据. 进一步, 由于基于扩散的距离与最短路径不同, 它将一些路径的贡献加总, 故可抗噪声.

2.4.1.2　降维和数据参数化

正如提到的, 上述扩散距离定义的优点是可以与随机游动的谱理论联系起来. 众所周知, 转移矩阵 P 具有特征值 $\lambda_0 \geq |\lambda_1| \geq \cdots \geq |\lambda_{n-1}| \geq 0$, 以及右, 左特征向量

$$P\varphi_j = \lambda_j \varphi_j, \quad \psi_j^T P = \lambda_j \psi_j^T, \tag{2.54}$$

这里可证明 $\lambda_0 = 1, \varphi_0 \equiv 1, \psi_i^T \varphi_j = \delta_{ij}$, 显然 $\psi_0 = \mu$. 事实上, 左右特征向量是对偶的, 并且可以分别认为是广义测度和试验函数. 这两个集合的向量由如下关系式

$$\varphi_j(x) = \frac{\psi_j(x)}{\mu(x)}, \quad \forall x \in S, \tag{2.55}$$

便于记号, 将 P 的左特征向量关于 $1/\mu$ 正规化

$$\|\psi_j\|_{1/\mu} = \sum_{x\in S} \frac{\psi_j^2(x)}{\mu(x)} = 1, \tag{2.56}$$

将其右特征向量关于 μ 正规化

$$\|\varphi_j\|_\mu = \sum_{x\in S} \varphi_j^2(x)\mu(x) = 1. \tag{2.57}$$

如果记第 t 次迭代为 $P^t = \{p_t(x,y)\}$, 则有如下的双正交谱分解

$$p_t(x,y) = \sum_{j=0}^{n-1} \lambda_j^t \varphi(x)\psi(y). \tag{2.58}$$

这对应于 P^t 的加权主成分分析. 前 N 项给出了 P^t 的最佳秩 N 逼近, 这里最佳定义为下述矩阵的加权度量

$$\|A\|^2 = \sum_{x,y\in S} = \mu(x)a(x,y)^2 \frac{1}{\mu(y)}. \tag{2.59}$$

Lafon 和 Lee 的主要想法是: 将 (2.58) 代入 (2.53), 得到

$$D_t^2(x,y) = \sum_{j=1}^{n-1} \lambda_j^{2t}(\varphi_j(x) - \varphi_j(y))^2. \tag{2.60}$$

由于 $\varphi_0 \equiv 1$ 为常数向量, 故不计入上述求和中. 进一步, 因为特征值的衰减[③], 故仅需要上述求和中的几项来满足某种精度. 精确地说, 令 $q(t)$ 为满足 $|\lambda_j|^t > \delta|\lambda_1|^t$ 的最大的指标 j, 于是扩散距离可于精度 δ 下用前 $q(t)$ 个非平凡特征向量和特征值近似地表示

$$D_t^2(x,y) \simeq \sum_{j=1}^{q(t)} \lambda_j^{2t}(\varphi_j(x) - \varphi_j(y))^2. \tag{2.61}$$

注意到如果用带权重 λ_j^t 的右特征向量作为数据的坐标, 则上式可看做 $\mathbb{R}^{q(t)}$ 上的欧氏距离. 换句话说, 这意味着如果引入扩散映射

$$\Phi_t : x \longmapsto \begin{pmatrix} \lambda_1^t\varphi_1(x) \\ \lambda_2^t\varphi_2(x) \\ \vdots \\ \lambda_{q(t)}^t\varphi_{q(t)}(x) \end{pmatrix}, \tag{2.62}$$

[③]衰减速度依赖于图的结构. 例如, 对于完全连通图这一特殊情况, 第一特征值为 1, 余下的特征值均为 0. 另一个极端情况如完全不连通图, 则所有特征值均为 1.

从而得到

$$D_t^2(x,y) \simeq \sum_{j=1}^{q(t)} \lambda_j^{2t}(\varphi_j(x)-\varphi_j(y))^2 = \|\Phi_t(x)-\Phi_t(y)\|^2. \tag{2.63}$$

其中 Φ_t 定义中的因子 λ_j^t 是上式成立的关键. 映射 $\Phi_t : S \to \mathbb{R}^{q(t)}$ 为数据集 S 提供了一个参数化, 或等价地, 图 G 用低维空间 $\mathbb{R}^{q(t)}$ 中的一团点来描述, 用重新标准化的特征向量作为坐标. 降维和特征向量权重由随机游动的时间 t 和特征值的谱衰减来确定. (2.63) 表示 Φ_t 通过用欧氏距离近似扩散距离的方式将整个数据集嵌入到 $\mathbb{R}^{q(t)}$. 因此, Lafon 和 Lee 的方法不同于其它的特征映射方法: 他们的出发点是数据集或图上显式定义的距离度量, 这个距离也是他们希望在非线性降维中保持的量.

2.4.2　图形分割和二次抽样

下面介绍关于上述二次抽样的一个新型方案, 可保持由图中数据点的连接程度定义的内在几何性. 其思想是构造一个原始随机游动于新图 $\hat{G}$ 上的具有相似谱性质的粗粒化形式. 这个新马氏链通过将点分区并且适当平均化这些分区之间转移概率的方式而得到. Lafon 和 Lee 指出为保持原始随机游动的大多数谱性质, 分区的选取至关重要. 更精确地, 扩散空间中的量化失真限制了扩散算子的近似的误差.

2.4.2.1　粗粒化的随机游动的构造

考虑节点集 S 的任意分区 $\{S_k\}_{1\le k\le N}$. Lafon 和 Lee 的目标是将点聚集在每个集合中, 来粗化状态空间 S 和随机游动的时间演化. 为此, 将每个集合 S_k 作为 N 个节点的图 $\hat{G}$ 相应的节点, 其权重函数定义为

$$\hat{w}(S_k,S_l) = \sum_{x\in S_k, y\in S_l} \mu(x)p_t(x,y), \tag{2.64}$$

其中求和包含所有的 $x\in S_k$ 和 $y\in S_l$ 间的转移概率, 如图 2.3 所示.

由可逆性条件 (2.52), 可证明这个图是对称的, 即 $\hat{w}(S_k,S_l)=\hat{w}(S_l,S_k)$. 令

$$\hat{\mu}(S_k) = \sum_{x\in S_k} \mu(x), \tag{2.65}$$

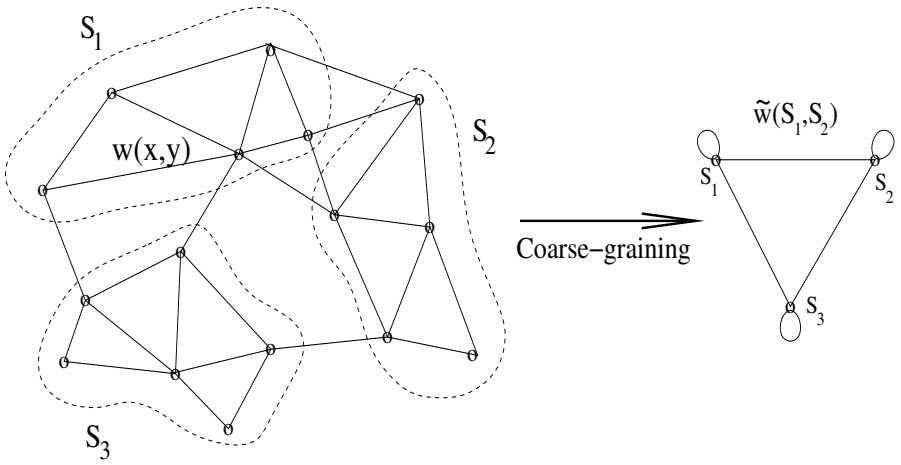


图 2.3: 图的粗粒化的例子: 给定图中节点集合的分区 $S = S_1 \bigcup S_2 \bigcup S_3$, 通过将所有节点聚集到子集 S_k 后形成超节点 (meta-node) 来定义粗粒化的图 $\hat{G}$[107]. 通过适当将 $x \in S_k$ 和 $y \in S_l$ 间的转移概率平均化, 可计算出新的权重为 $\hat{w}(S_k, S_l)$ 和转移概率为 $\hat{p}(S_k, S_l)$ 的马氏链, $k, l = 1, 2, 3$.

则可在这个图上定义可逆马氏链, 具有平稳分布 $\hat{\mu} \in \mathbb{R}^N$ 和转移概率

$$\hat{p}(S_k, S_l) = \frac{\hat{w}(S_k, S_l)}{\sum_{m=1}^{N} \hat{w}(S_k, S_l)} = \sum_{x \in S_k, y \in S_l} \frac{\mu(x)}{\hat{\mu}(S_k)} p_t(x, y), \tag{2.66}$$

故 $\hat{P} = \{\hat{p}(S_k, S_l)\}$ 为粗粒化的图上的 $N \times N$ 的转移矩阵. 更一般地, 对于 $0 \le j \le n-1$, 通过在一个分区中将所有节点求和来定义粗粒化形式的 ψ_j

$$\hat{\psi}_j(S_k) = \sum_{x \in S_k} \psi_j(x), \tag{2.67}$$

正如 (2.55), 定义粗粒化形式的 φ_j 依照对偶条件

$$\hat{\varphi}_j(S_k) = \frac{\hat{\psi}_j(S_k)}{\hat{\mu}(S_k)}, \tag{2.68}$$

这等价于取 φ_j 在 S_k 上的加权平均

$$\hat{\psi}_j(S_k) = \sum_{x \in S_k} \frac{\mu(x)}{\hat{\mu}(S_k)} \varphi_j(x). \tag{2.69}$$

粗粒化的 $\hat{p}(S_k, S_l)$ 包含了关于图 $\hat{G}$ 中新节点连接程度的数据的全部信息. 上述向量构成 $\hat{P}$ 的左右特征向量的近似的程度依赖于分区 $\{S_k\}$ 的特殊选取, 这在下述中更精确地研究.

2.4.2.2　近似误差

类似于 (2.56) 和 (2.57), 定义粗粒化的广义测度 $\hat{\psi}_j$ 上的范数

$$\|\hat{\psi}_j\|_{1/\hat{\mu}}^2 = \sum_{k=1}^{N} \frac{\hat{\psi}_j^2(S_k)}{\hat{\mu}(S_k)}, \tag{2.70}$$

和粗粒化测试函数 $\hat{\varphi}_j$ 上的范数

$$\|\hat{\varphi}_j\|_{\hat{\mu}}^2 = \sum_{k=1}^{N} \hat{\varphi}_j^2(S_k)\hat{\mu}(S_k). \tag{2.71}$$

现在引入每个社团 S_k 的几何质心的定义如下

定义 2.15 (几何质心) 设 $1 \le k \le N$, S 的子集 S_k 的几何质心 $c(S_k)$ 定义为加权和

$$c(S_k) = \sum_{x \in S_k} \frac{\mu(x)}{\hat{\mu}(S_k)} \Phi_t(x). \tag{2.72}$$

下述结论表明对于很小的 j, $\hat{\psi}_j$ 和 $\hat{\varphi}_j$ 是 $\hat{P}$ 的关于特征值 λ_j^t 的近似左, 右特征向量.

定理 2.16 对于 $1 \le k \le N$, 有如下关系成立

$$\hat{\psi}_j^T \hat{P} = \lambda_j^t \hat{\psi}_j^T + e_j, \quad \hat{P}\hat{\varphi}_j = \lambda_j^t \hat{\varphi}_j + f_j, \tag{2.73}$$

其中 e_j 和 f_j 满足

$$\|e_j\|_{1/\hat{\mu}}^2 \le 2\mathcal{D}, \quad \|f_j\|_{\hat{\mu}}^2 \le 2\mathcal{D}, \tag{2.74}$$

这里

$$\mathcal{D} = \sum_{k=1}^{N} \sum_{x \in S_k} \mu(x) \|\Phi_t(x) - c(S_k)\|^2. \tag{2.75}$$

这意味着如果 $|\lambda_j|^t \gg \sqrt{\mathcal{D}}$, 则 $\hat{\psi}_j$ 和 $\hat{\varphi}_j$ 是 $\hat{P}$ 的关于特征值 λ_j^t 的近似左, 右特征向量. 定理的证明参见 [107]. 这个结论也指出为了使近似质量最优, 需要在扩散空间中极小化如下失真度

$$\mathcal{D} = \sum_{k=1}^{N} \sum_{x \in S_k} \mu(x) \|\Phi_t(x) - c(S_k)\|^2$$

$$\approx \mathbb{E}_k\{\mathbb{E}_{X|k}\{\|\Phi_t(X)-c(S_k)\|^2|X\in S_k\}\}, \tag{2.76}$$

也可写成关于成对距离的加权和的形式

$$\begin{aligned}\mathcal{D} &= \frac{1}{2}\sum_{k=1}^{N}\hat{\mu}(S_k)\sum_{x,y\in S_k}\frac{\mu(x)}{\hat{\mu}(S_k)}\frac{\mu(y)}{\hat{\mu}(S_k)}\|\Phi_t(x)-\Phi_t(y)\|^2\\ &\approx \frac{1}{2}\mathbb{E}_k\{\mathbb{E}_{X,Y|k}\{\|\Phi_t(X)-\Phi_t(Y)\|^2|X,Y\in S_k\}\}.\end{aligned} \tag{2.77}$$

2.4.2.3 极小化失真度的算法

最后, Lafon 和 Lee 将极小化问题的算法方面与 k-means 算法联系起来. (2.75) 中 $\mathcal{D}$ 的形式在信息论中是经典的, 其极小化等价于求解基于样本点集 $\Phi_t(S)$ 的质量分布的量化 N 个编码的扩散空间问题, 这个最优化问题通常由保证收敛到局部极小值的 k-means 算法[86]来解决:

(1) 在扩散空间中随机初始化分划 $\{S_k^{(0)}\}_{1\le k\le N}$.

(2) 对于 $n>0$, 更新分划如下

$$S_k^{(n)}=\{x: k=\arg\min_{1\le l\le N}\|\Phi_t(x)-c(S_l^{(n-1)})\|^2\},\quad 1\le k\le N, \tag{2.78}$$

其中 $c(S_l^{(n-1)})$ 是 $S_l^{(n-1)}$ 的几何质心.

(3) 重复 (2) 直到收敛.

这个方法的缺点是每个质心 $\{c(S_k)\}$ 可能不属于集合 $\{\Phi_t(S)\}$, 这在应用中将导致这样的组合没有意义的问题, 例如基因数据的情况. 为得到属于原始集合 S 的社团代表 $\{c_k\}$, 从而引入如下的扩散中心的定义

定义 2.17 (扩散中心) *S 的一个子集 S_k 的扩散中心 $u(S_k)$ 定义为*

$$u(S_k)=\arg\min_{x\in S}\|\Phi_t(x)-c(S_k)\|^2. \tag{2.79}$$

这个概念没有定义唯一的扩散中心, 但是对于极小化失真度的目的是充分的. 注意到 $\{u(S_k)\}$ 是质心思想在图上的推广. 现在, 如果 $\{S_k\}$ 是 k-means 算法的输出, 则可以将 S_k 中每个点标记为代表性的中心 $u(S_k)$. 在这个意义上, 图 $\hat{G}$ 是 G 的二次抽样形式, 对给定的 N 保持图的谱性质. 上述分析为扩散空间的 k-means 算法提供了严格的证明, 此外还使得本节的内容与谱图分割 (通常仅考虑图的 Laplace 矩阵的第一个非平凡特征向量) 和特征映射 (用谱坐标来实现数据参数化) 联系起来.

第三章　基于最优预测的确定性分区方法

本章的内容将介绍利用最优预测的框架[31–35]来解决复杂网络的分区问题, 特别是寻找网络的一个最优分区. 为此可以定义网络之间的距离, 但不是像扩散距离 (2.53) 那样的网络上的距离. 所提出的策略与已有的计算机科学中发展出来的图形分割方法不同, 例如 Meila 和 Shi 的 MNCut 算法[132]以及 Lafon 和 Lee 的扩散映射的方法[107], 也和物理学家关于网络分析所提出的方法不同. 在 [43] 中对于近来主要相关的工作作了回顾.

下面将在网络的框架中发展最优预测理论, 并展示如何利用这个框架来最优约化网络的维度. 特别是将展现如何利用最优预测来将网络分划成社团, 并且将所提出的策略与近来网络分区中用到的其它降维方法进行比较. 同时, 也指出当网络良分区时, 即在马氏链呈现谱间隙 (spectral gap), 且靠近 1 的特征值对应的特征向量近似分片常数的意义下, 本章中的方法渐近等价于基于谱分割的方法. 最后根据此策略提出一个算法来分割网络, 并用几个具有代表性的算例加以阐明. 所有推导的细节见附录 B.

本章内容主要参考 [60].

3.1　基于最优预测的方法的框架

3.1.1　网络与马氏链

设 $G = (S, E)$ 为一个具有 n 个节点和 m 条边的网络 (或有限加权有向图), 其中 S 为节点集合, $E = \{e(x, y)\}_{x,y\in S}$ 为权重矩阵, $e(x, y)$ 为连接节点 x 和 y 的边上的权重. 权重矩阵的一个简单的例子是邻接矩阵 $e(x, y) = 1$ 或 0, 依赖于 x 和 y 是否连接.

考虑网络 G 上的离散随机游动或扩散过程. 在每个时间步, 一个游动者位于某一个节点并且移动到一个从其邻居中随机均匀选取的一个节点. 访问节点的序列构成了一个马氏链, 其状态就是网络中的节点. 在每一步, 从节点 x 到节点 y 的转移概率为

$$p(x,y)=\frac{e(x,y)}{d(x)},\quad d(x)=\sum_{z\in S}e(x,z), \tag{3.1}$$

其中 $d(x)$ 为节点 x 的度[36, 120]. 这就定于了转移矩阵 $P=\{p(x,y)\}_{x,y\in S}$. 如果取矩阵 P 的 t 次幂, 即让这个随机游动随时间 t 演化, 则此过程由 $P^t=\{p_t(x,y)\}_{x,y\in S}$ 驱动, 其中 $p_t(x,y)$ 表示随机游动从节点 x 经过 t 个时间步移动到节点 y 的概率, 并且为表述一致, 有 $p_1(x,y)=p(x,y)$. 如果网络不是有向的, 即 $e(x,y)=e(y,x)$, 则由 [36] 知, 当起始于节点 x 的随机游动时到达节点 y 的时间步 t 趋于无穷时的概率仅依赖于 y 的度, 而不再依赖于初始节点 x, 即

$$\lim_{t\to+\infty}p_t(x,y)=\mu(y),\quad \forall x,y\in S, \tag{3.2}$$

其中 μ 是马氏链的平稳分布, 它具有形式

$$\mu(x)=\frac{d(x)}{\sum_{z\in S}d(z)},\quad x\in S, \tag{3.3}$$

满足平稳分布的定义

$$\sum_{x\in S}\mu(x)p_t(x,y)=\mu(y),\quad \forall t\geq 1, \tag{3.4}$$

进一步, μ 还满足细致平衡条件

$$\mu(x)p_t(x,y)=\mu(y)p_t(y,x),\quad \forall t\geq 1, \tag{3.5}$$

从而马氏链是可逆的. 为简便, 在以后的章节中, 均假设网络是无向的, 尽管多数结果可以简单地推广到有向网络.

一个基本想法是从随机游动者在网络上移动来推断网络的性质, 这也是下面的内容将采取的思想. 本文将利用下述的关于马氏链的基本事实. 假设 S 上随机游动者的初始分布为 $\mu_0(x)$. 在后来的时刻 $t\in\mathbb{N}$, 它们的概率分布

为 $\mu_t(x) = \sum_{y\in S}\mu_0(y)p_t(y,x)$. 为计算 P^t, 可利用谱表示的方法. 设 $\{\varphi_j\}_{j=0}^{n-1}$ 和 $\{\psi_j\}_{j=0}^{n-1}$ 分别为 P 的右和左特征向量

$$P\varphi_j = \lambda_j\varphi_j, \quad \psi_j^T P = \lambda_j\psi_j^T, j = 0,1,\ldots,n-1. \tag{3.6}$$

在可逆的情况下, 所有特征值均为实数且位于区间 $[-1,1]$, 记为 $1 = \lambda_0 \geq |\lambda_1| \geq \ldots \geq |\lambda_{n-1}|$. 显然有 $\varphi_0 = e = (1,\ldots,1)^T$ 和 $\psi_0 = \mu$. 此外, 还有如下结论.

命题 3.1 *设 $P = \{p(x,y)\}_{x,y\in S}$ 为随机矩阵, 具有平稳分布 μ, $\{\varphi_j\}_{j=0}^{n-1}$ 和 $\{\psi_j\}_{j=0}^{n-1}$ 分别为 P 的右和左特征向量, 则有下述关系式成立*[①]:

(a) $\psi_j(x) = \varphi_j(x)\mu(x), \quad j = 0,\ldots,n-1, \quad x \in S;$

(b) $\psi_i^T\varphi_j = \delta_{ij}, \quad i,j = 0,\ldots,n-1.$

证明 事实上, 对于 (a), 可将 (3.6) 写成分量形式, 有

$$\sum_{y\in S} p(x,y)\varphi_j(y) = \lambda_j\varphi_j(x), \tag{3.7}$$

两边同乘以 $\mu(x)$, 得到

$$\sum_{y\in S}\mu(x)p(x,y)\varphi_j(y) = \sum_{y\in S}\mu(y)p(y,x) = \lambda_j\mu(x)\varphi_j(x), \tag{3.8}$$

令 $\tilde{\psi}_j(x) = \mu(x)\varphi_j(x)$, $\forall x \in S, j = 0,\ldots,n-1$, 则上式化为

$$\sum_{y\in S} p(y,x)\tilde{\psi}_j(y) = \lambda_j\tilde{\psi}_j(x), \tag{3.9}$$

即 $\tilde{\psi}_j^T P = \lambda_j\tilde{\psi}_j^T$, 由 (3.6) 知, $\psi_j = \tilde{\psi}_j$, $j = 0,\ldots,n-1$, 这就得到了 (a) 中结论.

对于 (b), 由于随机矩阵 P 在 $L_\mu^2(n)$ 中是对称的, 由 Hermite 矩阵的性质[②], 其右特征向量 $(\varphi_0\cdots\varphi_{n-1})$ 构成 μ 正交矩阵, 满足

$$(\varphi_i,\varphi_j)_\mu = \sum_{x\in S}\mu(x)\varphi_i(x)\varphi_j(x) = \delta_{ij}, \tag{3.10}$$

再由 (a) 知, 结论 (b) 成立. □

[①] 在第二章的命题 2.5 中也给出相同结论.

[②] 若 A 是实对称矩阵, 则存在正交矩阵 U 和实对角矩阵 $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$, 使得 $U^TAU = \Lambda$, 即 $AU = U\Lambda$. 详见 [79].

因此, 由 $P^t(\varphi_0\cdots\varphi_{n-1})=(\varphi_0\cdots\varphi_{n-1})\Lambda^t$ 和命题 3.1 得到 P^t 的谱分解

$$P^t=(\varphi_0\cdots\varphi_{n-1})\Lambda^t(\psi_0\cdots\psi_{n-1})^T, \tag{3.11}$$

写成分量形式, 则有

$$p_t(x,y)=\sum_{j=0}^{n-1}\lambda_j^t\varphi_j(x)\psi_j(y)=\sum_{j=0}^{n-1}\lambda_j^t\varphi_j(x)\varphi_j(y)\mu(y). \tag{3.12}$$

3.1.2 最优预测

本章要解决的主要问题是: 给定远小于 n 的 N, 如何寻找一个具有 N 的节点的网络上的马氏链来最好地表达原始马氏链的动力性质? E 等人从最优预测[31–35]的观点解决了这个问题. 这是一个从变化角度实现模型约化的框架, 想法是在某个目标函数极小化的意义下寻找约化模型来最佳逼近原始模型. E 等人将沿着这种思路来构造一个关于复杂网络最优分区的框架. 模型的空间就是网络上马氏链的空间 S, 由它们相应的随机矩阵表示. 约化模型将会是 S 的具有 N 个元素的分区上的可聚团的 (lumpable) 马氏链, 这些马氏链自然地嵌入 S 上马氏链自身的空间.

现引入 S 上马氏链 (随机矩阵) 的状态空间关于平稳分布 μ 的度量. 若 $p(x,y)$ 是具有平稳分布 μ 的随机矩阵, 定义它的 μ 范数如下

$$\|P\|_\mu^2=\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}|p(x,y)|^2, \tag{3.13}$$

则对于这个范数的含义, 有如下结论

引理 3.2 *设 $P=\{p(x,y)\}_{x,y\in S}$ 为随机矩阵, 具有平稳分布 μ, 则由 (3.13) 定义的 μ 范数等价于 $L_\mu^2(n)$ 空间中的 Hilbert-Schmidt 范数, 即 $\|P\|_\mu^2=\|P\|_{\mathrm{HS}}^2$.*

证明 首先证明: 对于前向自伴算子 $F:L_\mu^2(n)\to L_\mu^2(n)$, 满足

$$Ff(x)=\sum_{y\in S}K(x,y)\mu(y)f(y),\quad \forall f\in L_\mu^2(n), \tag{3.14}$$

其中 $K(x,y)$ 为核函数, 则算子 F 的 Hilbert-Schmidt 范数为

$$\|F\|_{\mathrm{HS}}^2=\sum_{x,y\in S}K^2(x,y)\mu(x)\mu(y). \tag{3.15}$$

事实上, 设 $\{\phi_j\}_{j=0}^{n-1}$ 是 $L_\mu^2(n)$ 的一组 μ 正交基, 则

$$K(x,y)=\sum_{i,j=0}^{n-1}c_{ij}\phi_i(x)\phi_j(y). \tag{3.16}$$

一方面, 由于 F 自伴, 记 $\mathrm{tr}(\cdot)$ 为算子的迹, 则有

$$\|F\|_{\mathrm{HS}}^2=\mathrm{tr}(F^*F)=\sum_{j=0}^{n-1}(\phi_j,F^*F\phi_j)_\mu=\sum_{j=0}^{n-1}(F\phi_j,F\phi_j)_\mu, \tag{3.17}$$

将 (3.14) 的形式代入上式, 得到

$$\begin{aligned}
\|F\|_{\mathrm{HS}}^2 &= \sum_{j=0}^{n-1}\left(\sum_{y\in S}K(x,y)\phi_j(y)\mu(y)\right)\cdot\left(\sum_{z\in S}K(x,z)\phi_j(z)\mu(z)\right)\cdot\mu(x)\\
&= \sum_{j=0}^{n-1}\sum_{x,y,z\in S}K(x,y)K(x,z)\phi_j(y)\phi_j(z)\mu(x)\mu(y)\mu(z)\\
&= \sum_{j=0}^{n-1}\sum_{x,y,z\in S}\sum_{s,t,u,v=0}^{n-1}c_{st}\phi_s(x)\phi_t(y)c_{uv}\phi_u(x)\phi_v(z)\phi_j(y)\phi_j(z)\mu(z)\mu(y)\mu(z)\\
&= \sum_{j=0}^{n-1}\sum_{x,y,z\in S}\sum_{s,t,u,v=0}^{n-1}c_{st}c_{uv}\delta_{su}\delta_{tj}\delta_{vj}\mu(z)\mu(y)\mu(z)\\
&= \sum_{s,t,u,v=0}^{n-1}c_{st}c_{uv}\delta_{su}\delta_{tv}=\sum_{s,t=0}^{n-1}c_{st}^2.
\end{aligned} \tag{3.18}$$

另一方面, 注意到 (3.15) 的右端为

$$\begin{aligned}
\sum_{x,y\in S}K^2(x,y)\mu(x)\mu(y) &= \sum_{x,y\in S}\mu(x)\mu(y)\sum_{s,t,u,v=0}^{n-1}c_{st}\phi_s(x)\phi_t(y)c_{uv}\phi_u(x)\phi_v(y)\\
&= \sum_{s,t,u,v=0}^{n-1}c_{st}c_{uv}\delta_{su}\delta_{tv}=\sum_{s,t=0}^{n-1}c_{st}^2.
\end{aligned} \tag{3.19}$$

从而证明了上述关于 $L_\mu^2(n)$ 空间自伴算子 F 的结论 (3.15).

其次, 注意到随机矩阵 P 是 $L_\mu^2(n)$ 空间自伴算子, 满足

$$Pf(x)=\sum_{y\in S}p(x,y)f(y)=\sum_{y\in S}\frac{p(x,y)}{\mu(y)}f(y)\mu(y),\quad \forall f\in L_\mu^2(n), \tag{3.20}$$

则由 (3.14) 知, $K(x,y)=p(x,y)/\mu(y)$, 再将其代入 (3.15), 得到 (3.13). □

将谱分解 (3.12) 代入 (3.13), 并由命题 3.1, 可以得到

$$\|p\|_\mu^2 = \sum_{x,y\in S} \mu(x)\mu(y) \sum_{i,j=0}^{n-1} \lambda_i\lambda_j\varphi_i(x)\varphi_i(y)\varphi_j(x)\varphi_j(y) = \sum_{i=0}^{n-1} \lambda_i^2. \tag{3.21}$$

从而范数 (3.13) 为 P 的特征值平方和. 若随机矩阵 P_1 和 P_2 具有共同的平稳分布 μ, 则范数 (3.13) 可以定义这两个矩阵之间的距离 $\|P_1 - P_2\|_\mu^2$, 选取 $P_1 = P$ 为原始马氏链的随机矩阵, P_2 是某一类随机矩阵, 则可以通过极小化 $\|P_1 - P_2\|_\mu^2$ 来在这一类中寻找最佳逼近原始马氏链的随机矩阵.

接下来, 选取 S 的一个分区 $S = \bigcup_{k=1}^N S_k$, 且 $S_k \bigcap S_l = \emptyset$ 若 $k \neq l$. 设 $\hat{P} = \{\hat{p}(S_k, S_l)\}_{k,l=1}^N$ 为状态空间 $\mathbb{S} = \{S_1, \ldots, S_N\}$ 上的随机矩阵, 这个矩阵可以被自然地提升到原始状态空间 S 上的随机矩阵空间中的 $\tilde{P} = \{\tilde{p}(x, y)\}_{x,y\in S}$

$$\tilde{p}(x, y) = \sum_{k,l=1}^N \mathbf{1}_{S_k}(x)\hat{p}(S_k, S_l)\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)}, \tag{3.22}$$

其中 $\mathbf{1}_{S_k}(x) = 1$ 若 $x \in S_k$, 否则 $\mathbf{1}_{S_k}(x) = 0$, 并且有

$$\hat{\mu}(S_k) = \sum_{x\in S_k} \mu(x). \tag{3.23}$$

上述 (3.22) 表明从 S_k 中的任何状态跳出的概率是相同的, 并且按照平稳分布进入 S_l. 这与粗粒化初始动力性质到新状态空间 $\mathbb{S} = \{S_1, \ldots, S_N\}$, 并忽略集合 S_k 中动力性质细节的思想是一致的.

命题 3.3 *如果 $\hat{P}$ 是 $\mathbb{S}$ 上的随机矩阵, 具有平稳分布 $\hat{\mu}$, 则由 (3.22) 定义的 $\tilde{P}$ 为 S 上的随机矩阵, 具有平稳分布 μ. 若进一步, 有 $\hat{P}$ 满足关于 $\hat{\mu}$ 的细致平衡条件, 则 $\tilde{P}$ 满足关于 μ 的细致平衡条件.*

证明 由已知, 对于 $\forall k, l = 1, \ldots, N$, 有

$$\hat{p}(S_k, S_l) \geq 0, \quad \sum_{l=1}^N \hat{p}(S_k, S_l) = 1, \quad \sum_{k=1}^N \hat{\mu}(S_k)\hat{p}(S_k, S_l) = \hat{\mu}(S_l), \tag{3.24}$$

则由 (3.22) 知, 显然 $\tilde{p}(x, y) \geq 0$, $x, y \in S$, 并可得出

$$\sum_{y\in S} \tilde{p}(x, y) = \sum_{y\in S} \sum_{k,l=1}^N \mathbf{1}_{S_k}(x)\hat{p}(S_k, S_l)\frac{\mathbf{1}_{S_l}(y)\mu(y)}{\hat{\mu}(S_l)}$$

$$= \sum_{k,l=1}^{N} \mathbf{1}_{S_k}(x)\hat{p}(S_k,S_l)\frac{\sum_{y\in S_l)}\mu(y)}{\hat{\mu}(S_l)}$$
$$= \sum_{k=1}^{N}\mathbf{1}_{S_k}(x)\sum_{l=1}^{N}\hat{p}(S_k,S_l)=1, \tag{3.25}$$

$$\begin{aligned}\sum_{x\in S}\mu(x)\tilde{p}(x,y) &= \sum_{k,l=1}^{N}\left(\sum_{x\in S}\mu(x)\mathbf{1}_{S_k}(x)\right)\hat{p}(S_k,S_l)\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)}\\ &= \mu(y)\sum_{l=1}^{N}\mathbf{1}_{S_l}(y)\frac{1}{\hat{\mu}(S_l)}\sum_{k=1}^{N}\hat{\mu}(S_k)\hat{p}(S_k,S_l)\\ &= \mu(y)\sum_{l=1}^{N}\mathbf{1}_{S_l}(y)=\mu(y),\end{aligned} \tag{3.26}$$

从而 $\tilde{P}$ 为 S 上的随机矩阵, 且具有平稳分布 μ. 进一步, 如果 $\hat{P}$ 满足关于 $\hat{\mu}$ 的细致平衡条件, 即 $\hat{\mu}(S_k)\hat{p}(S_k,S_l)=\hat{\mu}(S_l)\hat{p}(S_l,S_k)$, 于是有

$$\begin{aligned}\mu(x)\tilde{p}(x,y) &= \sum_{k,l=1}^{N}\mu(x)\mu(y)\mathbf{1}_{S_k}\mathbf{1}_{S_l}\frac{\hat{p}(S_k,S_l)}{\hat{\mu}(S_l)}\\ &= \sum_{k,l=1}^{N}\mu(x)\mu(y)\mathbf{1}_{S_k}\mathbf{1}_{S_l}\frac{\hat{p}(S_l,S_k)}{\hat{\mu}(S_k)}=\mu(y)\tilde{p}(y,x),\end{aligned} \tag{3.27}$$

则 $\tilde{P}$ 满足关于 μ 的细致平衡条件. 于是结论得证. □

现考虑: 给定分区 $\mathbb{S}$ 和某个 $t\geq 1$, 如何计算 P^t 的最佳逼近 $\tilde{P}$? 如果最优是在度量 (3.13) 的意义下, 则最佳的 $\tilde{P}$ 是对于所有的形如 (3.22) 的 $\tilde{P}$ 中使

$$E(\tilde{P})=\|P^t-\tilde{P}\|_\mu^2 \tag{3.28}$$

极小化的极小值点. 附录 B.1 中的直接计算表明 $E(\tilde{P})$ 的极小值点 $\tilde{P}$ 中的 $\hat{P}$ 有如下形式

$$\hat{p}^*(S_k,S_l)=\frac{1}{\hat{\mu}(S_k)}\sum_{x\in S_k,y\in S_l}\mu(x)p_t(x,y)=\sum_{j=0}^{n-1}\lambda_j^t\hat{\varphi}_j(S_k)\hat{\varphi}_j(S_l)\hat{\mu}(S_l), \tag{3.29}$$

其中第二个等号用到了 (3.12), 且

$$\hat{\varphi}_j(S_k)=\frac{\sum_{x\in S_k}\mu(x)\varphi_j(x)}{\sum_{x\in S_k}\mu(x)}. \tag{3.30}$$

命题 3.4 *方程 (3.29) 中的 $\hat{P}^*$ 是随机矩阵, $\hat{\mu}$ 是 $\mathbb{S}$ 上转移矩阵为 $\hat{P}^*$ 的马氏链的平稳分布, 且 $\hat{P}^*$ 满足关于 $\hat{\mu}$ 的细致平衡条件.*

证明 由 (3.29), 有 $\hat{p}^*(S_k, S_l) \geq 0$, $\forall k, l = 1, \ldots, N$, 且

$$\sum_{l=1}^{N} \hat{p}^*(S_k, S_l) = \frac{1}{\hat{\mu}(S_k)} \sum_{l=1}^{N} \sum_{x \in S_k} \mu(x) \sum_{y \in S_l} p_t(x, y) = \frac{1}{\hat{\mu}(S_k)} \hat{\mu}(S_k) = 1, \tag{3.31}$$

故 $\hat{P}^*$ 是随机矩阵. 由 (3.23) 知, $\hat{\mu}(S_k) \geq 0$, $\forall k = 1, \ldots, N$, 且 $\sum_{k=1}^{N} \mu(S_k) = 1$, 此外,

$$\sum_{k=1}^{N} \hat{\mu}(S_k) \hat{p}^*(S_k, S_l) = \sum_{k=1}^{N} \sum_{x \in S_k, y \in S_l} \mu(x) p_t(x, y) = \sum_{y \in S_l} \sum_{x \in S} \mu(x) p_t(x, y) = \hat{\mu}(S_l), \tag{3.32}$$

从而得到 $\hat{\mu}$ 为 $\mathbb{S}$ 上的转移矩阵为 $\hat{P}^*$ 的马氏链的平稳分布. 进一步, 有

$$\begin{aligned} \hat{\mu}(S_k) \hat{p}^*(S_k, S_l) &= \sum_{x \in S_k, y \in S_l} \mu(x) p_t(x, y) \\ &= \sum_{y \in S_l, x \in S_k} \mu(y) p_t(y, x) = \hat{\mu}(S_l) \hat{p}^*(S_l, S_k), \end{aligned} \tag{3.33}$$

则 $\hat{\mu}$ 还满足细致平衡条件. 结论得证. □

从而, 由命题 3.3, 知

$$\tilde{p}^*(x, y) = \sum_{k,l=1}^{N} \mathbf{1}_{S_k}(x) \hat{p}^*(S_k, S_l) \mathbf{1}_{S_l}(y) \frac{\mu(y)}{\hat{\mu}(S_l)} \tag{3.34}$$

是 S 上的随机矩阵, 其平稳分布为 μ, 那么矩阵 $\tilde{P}^*$ 是在形如 (3.22) 的类中最佳逼近原始 P^t 的随机矩阵. 此外, 由附录 B.2, 还可以得到 $E(\tilde{P}^*) \equiv E^*$, 其中

$$\begin{aligned} E^* &= \sum_{x,y \in S} \frac{\mu(x)}{\mu(y)} |p_t(x, y)|^2 - \sum_{k,l=1}^{N} \frac{\hat{\mu}(S_k)}{\hat{\mu}(S_l)} |\hat{p}^*(S_k, S_l)|^2 \\ &\equiv \|P^t\|_\mu^2 - \|\hat{P}^*\|_{\hat{\mu}}^2. \end{aligned} \tag{3.35}$$

注意到 (3.22) 是中的矩阵 $\tilde{P}$ 的秩 $\mathrm{rank}(\tilde{P}) \leq N$, 从而将 (3.29) 代入 (3.22) 所得到的残量$E^* \geq 0$, 不小于对所有秩 N 矩阵 $\tilde{P}$ 来极小化 (3.28) 所得到的残量. 直

接的计算[③]可以表明对所有秩 N 矩阵 $\tilde{P}$, (3.28) 的极小值点 $\tilde{P}^{**}=\{\tilde{p}^{**}(x,y)\}_{x,y\in S}$ 为

$$\tilde{p}^{**}(x,y)=\sum_{j=0}^{N-1}\lambda_j^t\varphi_j(x)\varphi_j(y)\mu(y). \tag{3.37}$$

注意到在一般情况下, 由于 (3.37) 中的某些元素的可能为负, 故 $\tilde{P}^{**}$ 不是随机矩阵, 然而 $E^{**}=\|P^t-\tilde{P}^{**}\|_\mu^2$ 给出了残量 E^* 的一个下界.

3.2　聚团性 (lumpability) 与最优分区

3.2.1　马氏链关于分区的聚团性

定义 3.5 随机矩阵为 P 的马氏链关于分区 $\mathbb{S}=\{S_1,\dots,S_N\}$ 是可聚团的 (lumpable), 当且仅当原始链上的随机游动在这些集合上也是马氏的.

定理 3.6 假设由 (3.29) 定义的 $\hat{P}^*$ 是非奇异的, 则 $E^*\geq E^{**}$, 其中 $E^*=E^{**}$ 当且仅当马氏链关于 $\mathbb{S}$ 是可聚团的. 此时, $\tilde{p}^{**}(x,y)=\tilde{p}^*(x,y)$, $x,y\in S$.

这个定理是 [132] 中结果的简单推论 (命题 2.12). 在 [132] 中, Meila 和 Shi 证明了矩阵随机矩阵 P 的马氏链在集合 $\mathbb{S}=\{S_1,\dots,S_N\}$ 上是可聚团的, 当且仅当满足下面两个条件之一:

(a) 对于每个 S_k, $\sum_{y\in S_l}p(x,y)$ 与 $x\in S_k$ 独立, 且矩阵 $\hat{p}^*(S_k,S_l)=\sum_{y\in S_l}p(x,y)$, $x\in S_k$, 是非奇异的.

(b) 前 N 个特征向量 $\varphi_j(x)$, $j=0,\dots,N-1$, 关于分划 $\mathbb{S}=\{S_1,\dots,S_N\}$ 是分片常数的.

[③](低秩逼近定理) 设 A 为 n 阶秩为 p 的实对称矩阵, 则满足

$$\|A-A_q\|_{HS}=\min\left\{\|A-B\|_{HS}:B\text{ 为 }n\times n\text{ 阶秩为 }q\text{ 的矩阵},1\leq q\leq p\right\}$$

的 n 阶秩为 q 矩阵 A_q 有如下形式

$$A_q=U\mathrm{diag}(\lambda_1,\dots,\lambda_q,0,\dots,0)U^T,$$

其中 $\lambda_1\geq\cdots\geq\lambda_n$ 为 A 的特征值, $U=(\varphi_1\dots\varphi_n)$ 为 A 的右特征向量构成的正交矩阵. 详见 [79].

在这种情况下, 易知对于 $i=0,\dots,N-1$, $\varphi_i(x)$ 必须具有形式

$$\varphi_i(x)=\sum_{k=1}^{N} c_{ik}\mathbf{1}_{S_k}(x), \tag{3.38}$$

其中 $c_{0k}=1$, 且对于 $i=1,\dots,N-1$, 系数 c_{ik} 满足

$$\sum_{k=1}^{N} c_{ik}\hat{\mu}(S_k)=0 \quad \sum_{k=1}^{N} c_{ik}c_{jk}\hat{\mu}(S_k)=\delta_{ij}. \tag{3.39}$$

特征向量的 μ 正交性条件蕴含着, 对于 $i=N,\dots,n-1$ 和 $\forall k=1,\dots,N$, 有

$$\sum_{x\in S_k}\varphi_i(x)\mu(x)=0. \tag{3.40}$$

因此, 得到

$$\hat{\varphi}_i(S_k)=\begin{cases} c_{ik}, & i=0,\dots,N-1,\\ 0 & i=N,\dots,n-1. \end{cases} \tag{3.41}$$

再由 (3.29) 和 (3.22), 得到

$$\tilde{p}^{*}(x,y)=\sum_{k,l=1}^{N}\mathbf{1}_{S_k}(x)\sum_{i=0}^{N-1}\lambda_i^t c_{ik}c_{il}\mathbf{1}_{S_l}(y)\mu(y)\equiv\tilde{p}^{**}(x,y). \tag{3.42}$$

在一般的情况中, 马氏链关于分区 $\mathbb{S}=\{s_1,\dots,S_N\}$ 是非可聚团的, 并且 (3.37) 也将不再是随机矩阵, 因此, 它不是一个可接受的近似. 但是 (3.34) 仍然是原始马氏链随机矩阵的最优逼近, 且 E^*-E^{**} 给出了这个逼近关于聚团性的质量的的度量. 稍后, 可以看到定理 3.6 的条件对于良分区的网络是近似满足的.

3.2.2 最优分区

本章要解决的下一个问题是: 给定 N, 最好的分区 $\{S_1,\dots,S_N\}$ 是什么? 为了回答这个问题, E 等人将 (3.35) 看作是 $\{S_1,\dots,S_N\}$ 的函数, 即 $E^*\equiv E(S_1,\dots,S_N)$, 并且计算

$$\min_{\{S_1,\dots,S_N\}} E(S_1,\dots,S_N)=-\max_{\{S_1,\dots,S_N\}}\sum_{k,l=1}^{N}\frac{\hat{\mu}(S_k)}{\hat{\mu}(S_l)}|\hat{p}^{*}(S_k,S_l)|^2. \tag{3.43}$$

作为定理 3.6 的一个直接推广, 有如下结论

定理 3.7 记 $\{S_1^*,\ldots,S_N^*\}$ 为极小化 (3.43) 的分区, 并设 $E^{**}=\|P^t-\tilde{P}^{**}\|_\mu^2$, 其中 $\tilde{P}^{**}$ 由 (3.37) 给出, 则有 $E(S_1^*,\ldots,S_N^*)\geq E^{**}$, 并且 $E(S_1^*,\ldots,S_N^*)=E^{**}$ 当且仅当马氏链关于分区 $\mathbb{S}^*=\{S_1^*,\ldots,S_N^*\}$ 是可聚团的.

换句话说, 如果马氏链关于一个具有 N 个集合的分区是可聚团的, 则极小化问题 (3.43) 可识别这些集合. 下面, E 等人提出了一个 k-means 算法的变形来求解这个极小化问题. 但做这件事情之前, 先将这里提出的准则与先前提出的分割网络的其它准则进行比较.

3.2.2.1　与其它分区策略的比较

在 [107] 中, Lafon 和 Lee 提出了一个网络分区的统一框架. 基本思想是引入下述网络节点之间的扩散距离 (这可以与网络之间的距离 (3.28) 形成对照)

$$D_t^2(x,y)=\sum_{z\in S}\frac{(p_t(x,z)-p_t(y,z))^2}{\mu(z)}=\sum_{j=0}^{n-1}\lambda_j^{2t}(\varphi_j(x)-\varphi_j(y))^2. \tag{3.44}$$

基于这个扩散距离, Lafon 和 Lee 指出通过极小化下述的失真度来分割网络

$$\min_{\{S_1,\ldots,S_N\}}\sum_{k=1}^{N}\sum_{x\in S_k}\mu(x)\sum_{j=0}^{n-1}\lambda_j^{2t}(\varphi_j(x)-\hat{\varphi}_j(S_k))^2, \tag{3.45}$$

其中 $\hat{\varphi}_j(S_k)$ 如 (3.30) 所定义. 这个目标, 或更精确地说, 向量

$$(\lambda_1^{2t}\hat{\varphi}_1(S_k),\ldots,\lambda_{n-1}^{2t}\hat{\varphi}_{n-1}(S_k)) \tag{3.46}$$

在 [107] 中称为几何质心. 通过展开 (3.45) 并利用 (3.12) 和 (3.29), 易知 (3.45) 可重新表示为

$$\min_{\{S_1,\ldots,S_N\}}\left(\sum_{x\in S}p_t(x,x)-\sum_{k=1}^{N}\hat{p}^*(S_k,S_k)\right), \tag{3.47}$$

或等价地, 有

$$\max_{\{S_1,\ldots,S_N\}}\sum_{k=1}^{N}\hat{p}^*(S_k,S_k)=\max_{\{S_1,\ldots,S_N\}}\sum_{k=1}^{N}\frac{\sum_{x,y\in S_k}\mu(x)p_t(x,y)}{\sum_{x\in S_k}\mu(x)}. \tag{3.48}$$

这个准则也在 [132] 中提出的 MNCut 算法, 以及在 [47] 中引入并在 [49, 50] 中得到进一步发展的几乎不变集合体中采用过.

为了看清 (3.48) 与 (3.43) 之间的差异, E 等人注意到当马氏链关于 $\{S_1,\dots,S_N\}$ 是可聚团的, 则 (3.48) 的极小值点可能会是分区 $\{S'_1,\dots,S'_N\}$, 它与 $\{S_1,\dots,S_N\}$ 不同, 从而与 (3.43) 的极小值点不同.

下面介绍一个简单的例子来举例说明这个观点. 假设 $S=\{1,\dots,2n\}$, $p(x,y)=\frac{1}{2}$ 如果 $x=2,\dots,2n-1$ 且 $y=x\pm1$, $p(1,2)=p(2n,2n-1)=1$, $p(x,y)=0$ 对所有其它情形 (即每个节点仅与直线上它的两个直接邻居相连). 这个链关于两个状态的链 $S_1=\{1,3,\dots,2n-1\}$ 和 $S_2=\{2,3\dots,2n\}$ 是可聚团的, 其中

$$\hat{p}^*(S_1,S_1)=\hat{p}^*(S_2,S_2)=0,$$
$$\hat{p}^*(S_1,S_2)=\hat{p}^*(S_2,S_1)=1.$$

根据这个选取, 实际上残量 $E(S_1,S_2)=E^{**}$, 这与定理 3.7 是一致的. 另一方面, (3.48) 导致 $S'_1=\{1,2,\dots,n\}$ 和 $S'_2=\{n+1,n+2,\dots,2n\}$, 并且马氏链关于这个分区不是可聚团的. 因此基于 (3.43) 和 (3.48) 的分区算法实际上是不同的. 可以猜想 (3.48) 可能在数据分割中更为有用 (即如果网络上的动力学是不相关的), 而如果对于网络的动力学性质更感兴趣, 则 (3.43) 较之更为可取.

最后, 令人感兴趣的是, 虽然 (3.43) 和 (3.48) 是不同的, 但是对于下面将要描述的良分区网络, 它们是等价的.

3.2.2.2 良分区网络的情况

由定义知, 称一个网络为良分区的, 如果与其相关联的马氏链具有一个谱间隙, 即 (3.6) 中定义的 P 的特征值满足

$$\lambda_j=1-\eta_j\delta,\qquad k=1,\dots,N-1,$$
$$|\lambda_j|<\lambda_\star,\qquad k=N,\dots,n-1,$$

其中 $0<\delta\ll1$, $\eta_j>0$, $\lambda_\star\in(0,1)$ 关于 δ 是 $O(1)$ 的. 在这种情况下可以证明, 存在 S 的一个具有 N 个集合的分区 $\{S_1,\dots,S_N\}$, 使得 P 的前 N 个特征向量在这些

集合上是近似分片常数的

$$\varphi_j(x) = \sum_{k=1}^{N} c_{jk} \mathbf{1}_{S_k}(x) + o(1), \quad k = 1, \ldots, N-1, \tag{3.49}$$

其中 $c_{0k} = 1$, 且系数 c_{jk} 对于 $j = 1, \ldots, N-1$ 满足 (3.39). 这表明马氏链关于分区 $\{S_1, \ldots, S_N\}$ 是近似可聚团的, 并且由定理 3.7, 关于这些集合上的残量 $E(S_1, \ldots, S_N)$, 当 $\delta \to 0$ 时, 趋于 E^{**}. 事实上, 在这种情况下, 有

$$\|\tilde{P}^* - P^{t(\delta)}\|_\mu^2 \to 0, \quad \delta \to 0, \tag{3.50}$$

其中 $t(\delta) = 1/\delta$, $\tilde{P}^*$ 由 (3.34) 给出. 这是离散时间形式的 Khasminskii 平均定理[101].

在 $p_t(x, y)$ 中取 $t(\delta) = 1/\delta$, 则相似的计算表明, $\{S_1, \ldots, S_N\}$ 也是依照 (3.48) 的最优分区. 因此对于良分区的网络, (3.43) 和 (3.48) 是渐近等价的.

3.3　算法的构造

在实践中, 能够简易地处理极小化问题 (3.43) 是极其重要的. 在 [107] 中, 极小化问题 (3.45) 可利用 k-means 算法[86]求解. 这里将指出这个算法的一个变形也可以用来处理 (3.43).

算法 3.8 (变形 k-means 算法)

(1) 随机初始化分区 $\{S_k^{(0)}\}_{k=1}^N$.

(2) 对于 $n \geq 0$, 计算 $\hat{p}_{kl}^{(n)}$ 如下

$$\hat{p}_{kl}^{(n)} = \frac{1}{\hat{\mu}(S_k^{(n)})} \sum_{x \in S_k^{(n)}, y \in S_l^{(n)}} \mu(x) p_t(x, y). \tag{3.51}$$

(3) 对于 $n \geq 0$, 更新分区 $\{S_k^{(n+1)}\}$ 如下

$$S_k^{(n+1)} = \left\{ x : k = \arg\min_l \bar{E}(x, S_l^{(n)}) \right\}, \quad k = 1, \cdots, N, \tag{3.52}$$

其中

$$\bar{E}(x, S_k) = \sum_{l=1}^{N} \sum_{y \in S_l} \mu(x)\mu(y) \left| \frac{p_t(x, y)}{\mu(y)} - \frac{\hat{p}^*(S_k, S_l)}{\hat{\mu}(S_l)} \right|^2. \tag{3.53}$$

(4) 重复步骤 (2) 到 (3) 直至目标函数不再下降.

这个算法像所有的 k-means 型算法一样, 尽管不收敛于全局极小值, 但是具有收敛速度很快的优点. 实际上, 这里的情形比通常情形稍坏一些, 因为 $\bar{E}(x, S_k)$ 隐式地依赖于先前的分区, 这是由于 (3.53) 中 S_l 的出现. 因此目标函数 $E(S_1, \ldots, S_N)$ 不能保证每一次迭代都下降. 这个问题可以通过若目标函数上升或保持常数则终止迭代来解决, 然后用一些不同的初始分区 $\{S_k^{(0)}\}_{k=1}^N$ 作为初值来重复这个计算过程, 并选取最佳的结果, 就像使用 k-means 算法时通常所做的那样.

下面考虑算法的计算量. 由附录 B.3 可知, 上述变形 k-means 算法中每次迭代的花费, 即计算 (3.52) 和 (3.53) 的花费, 为 $O(N(n+m))$, 其中 n 是网络中的节点数, m 是边数, N 是社团数目, 这个量从而为变形 k-means 算法的花费提供了一个下界. 若估计它的实际花费, 还要估计需要用多少随机初始分区来确定实际的极小值, 和在达到一个局部极小值之前算法需要做多少次迭代. 这些数字很难解析地计算出来. 在规模增长的 ad hoc 网络的实验中, 可以发现这些数字看起来仅是微弱地依赖于网络的规模: 通常, 500 个随机初始分区已经足够, 并且对于每一个, 算法于几十步内收敛, 甚至对于大型的网络 (即 n 取值在 128 和 1280 之间). 如果这些结果是一般性的, 这将使变形 k-means 算法成为在 [43] 中比较的诸多方法中花费最少的算法之一, 然而这一点仍需要进一步的研究.

3.4 数值实验

3.4.1 空手道俱乐部网络

作为第一个测试, E 等人采用空手道俱乐部这个众所周知的例子. 这个网络是 Wayne Zachary 在观察一个美国大学的空手道俱乐部成员之间的社交活动后所构建的[210], 具体描述见 1.5.2. E 等人将算法应用于 Zachary 原始网络, 并选取随机分区作为变形 k-means 算法的初始条件. 图 3.1 显示了变形 k-means 算法所得到的分区结果, 对应于 $S_1 = \{1 : 8, 10 : 14, 17, 18, 20, 22\}$ 和 $S_2 = \{9, 15, 16, 19, 21, 23 : 34\}$. 这与 Zachary 观察到的分裂后的小俱乐部的实际结构 $S_1 = \{1 : 8, 11 : 14, 17, 18, 20, 22\}$ 和 $S_2 = \{9, 10, 15, 16, 19, 21, 23 : 34\}$ 非常相似, 仅有一个节点 10 被误分区.

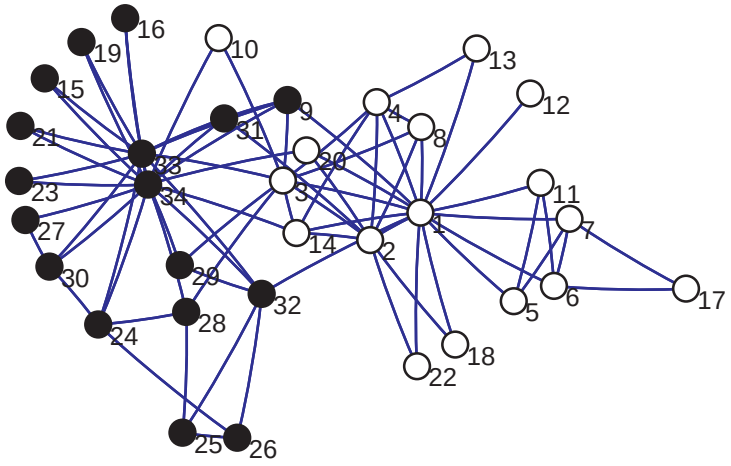


图 3.1: 利用变形 k-means 方法确定的 Zachary 空手道俱乐部网络[210]的两个社团. 节点 1 和节点 33 分别表示管理者和主教练. 选取随机分区作为变形 k-means 算法的初始条件, 算法得到的分区为 $S_1 = \{1 : 8, 10 : 14, 17, 18, 20, 22\}$ 和 $S_2 = \{9, 15, 16, 19, 21, 23 : 34\}$, 这和 Zachary 实际的观察非常相似: 仅有一个节点 10 被误分区.

3.4.2　128 个节点的 ad hoc 网络

作为第二个测试, E 等人采用 ad hoc 网络的例子, 这类网络具有已知的社团结构, 构造如 1.5.1 所述. 通常定义 z_{out} 为某个节点与属于其它社团节点之间连接的平均数, z_{out} 越大, 社团就变得越模糊 (diffusive).

首先, E 等人在假设社团数目 $N = 4$ 已知的情况下用变形 k-means 算法去分割网络. 考虑 0 和 8 之间的一些 z_{out} 的值, 并计算变形 k-means 算法得到的节点识别的正确率 f (为计算这个分数 f, E 等人采用了 [144] 中提出的评判准则, 在 [43] 中也用它来进行比较研究). 为了使结果较少地依赖于特殊的网络选取, 对于 z_{out} 的每个值, E 等人选取网络的 100 个实现, 计算这 100 个实现中的 f 的均值和标准差. 在每种情形下, 他们取 100, 300, 500, 1000 个随机初始分区来应用变形 k-means 算法, 并取最佳结果, 即具有最小残量 E^* 的结果. 最终关于 f 的均值的结果如图 3.2 所示. 这表明方法在识别正确的社团方面表现得很好, 一直到 $z_{\text{out}} = 7.5$ (其中 $z_{\text{out}}/\langle d\rangle = 0.4688$, 500 次试验的 $f = 0.93$), 仅对于 $z_{\text{out}} = 8$ 的情况才出现恶化 (其中 $z_{\text{out}}/\langle d\rangle = 0.5$, 500 次试验的 $f = 0.7367$). 即使在这种情况下, 变形

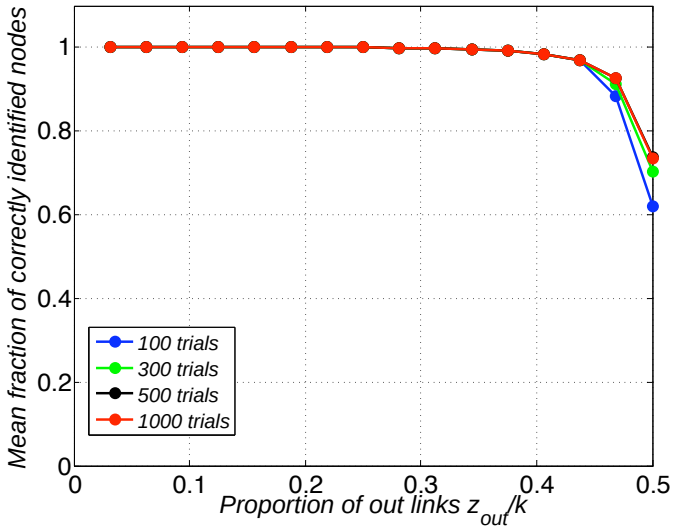


图 3.2: 节点识别的正确率随 $z_{\rm out}$ 的变化. 四条曲线分别表示 k-means 算法取 100, 300, 500, 1000 次随机初始分区所得到的结果[60]. 如图所示, 结果随初始条件数量的增加而改进, 但最终饱和 (500 次试验的曲线几乎无法与其上面的 1000 次试验的曲线区分开). 结果表明变形 k-means 算法与 [43] 中提及的诸多算法相比起来是最好的算法之一.

k-means 算法与 [43] 中陈述的方法相比较仍然非常具有竞争力, 其中仅有两种方法在最后的数据点 $z_{\rm out}/\langle d\rangle = 0.5$ 上优于变形 k-means 算法. 下面将更详细地讨论变形 k-means 算法关于精度方面的性能.

3.4.3 算法的精度

图 3.2 中展示的结果表明变形 k-means 算法可以正确地识别多于 90% 的 f 的节点, 直到 $z_{\rm out} = 7.5$. 方法当 $z_{\rm out} = 8$ 时恶化, 然而令人感兴趣的是去研究此时发生了什么. 需要解决的第一个问题是, 变形 k-means 算法是否能够确定目标函数 (3.43) 的极小值 E^*, 因为它依赖于采用的随机初始分区的数目. 图 3.2 中的结果表明若初始分区数目大于 500, 那么的确是这种情况. 图 3.3 中的结果, 对于 $z_{\rm out} = 8$ 的 ad hoc 网络的 100 次实现, 取 100, 300, 500, 1000 个随机初始分划, 通过展示变形 k-means 算法所得到的残量 E^*, 来确定这个发现. 图 3.3 也解释了对于分割诸如 $z_{\rm out} = 8$ 时的 ad hoc 网络这样的具有扩散社团结构的网络的固有的困难. 实际上, 可以看出从已知社团结构计算出的残量 E^* 通常大于变形 k-means 算法确定

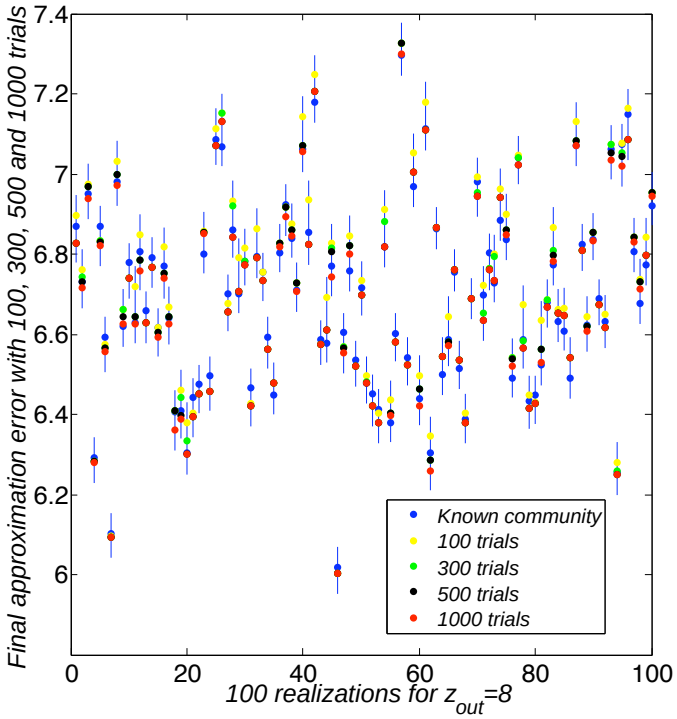


图 3.3: 对于 ad hoc 网络的 $z_{\text{out}} = 8$ 的 100 次独立实现, 利用变形 k-means 算法分别取 100, 300, 500, 1000 次随机初始分区所得到的残量 E^*[60]. 图中也展示了利用网络的已知分区计算出的残量 E^*. 可以看出由变形 k-means 算法确定的实际残量 E^* 通常小于利用已知社团计算的残量. 这反映了当 $z_{\text{out}} = 8$ 时 ad hoc 网络中社团结构分散的性质. 图中置入垂直线为了可视化不同实现所确定的不同的点.

的残量. 这意味着, 至少关于聚团性, 当 $z_{\text{out}} = 8$ 时, 社团结构通常变得十分模糊 (diffusive), 使得其它社团实际上较好, 这就是为什么在这个情况下节点识别的正确率 f 变得较小.

3.4.4　确定社团数目 N

到目前为止, E 等人都假设社团数目 N 是给定的. 然而, 在许多应用中, 这个数目是预先未知的, 并需要通过分割方法本身来确定. 结合变形 k-means 算法, 一种处理方式是取一些不同的 N 值来应用算法, 然后比较结果. 例如, 设 $z_{\text{out}} = 6$, 并对于 $N = 2, 3 \ldots, 8$ 个社团来应用变形 k-means 算法, 可发现如下结果. 当 $N = 2$ 时, 变形 k-means 算法确定出一个包含 32 个节点的社团, 这是一个正确的社团,

和一个包含96个节点的社团, 这是余下 3 个正确社团的合并. 当 $N = 3$ 时, 变形 k-means 算法正确识别出两个包含 32 个节点的社团, 并将其它两个分区在一起. 当 $N = 4$ 时, 变形 k-means 算法得到正确的分区. 当 $N = 5$ 时, 变形 k-means 算法也得到正确的分区, 除了其中一个包含 32 个节点的社团分裂成两个. 当 $N = 6, 7, 8$ 时出现相似的情形, 分别有两个, 三个和四个包含 32 个节点的正确的社团分裂成两个.

如何预先确定这些结果中哪个 N 值才是真实值? 最自然的方式是考虑每种情形中的相对残量 $E^* - E^{**}$. 在 $z_{\text{out}} = 6$ 的 ad hoc 网络的例子中, E 等人发现 $E^* - E^{**}$ 从 $N = 2$ 到 $N = 4$ 是缓慢增长的, 然后当 $N > 4$ 时增长变得更快. 这看起来巩固了 $N = 4$ 使最优选择的结论 (由于使用更多的社团使结果更差).

与传统的 k-means 算法族[86]一样, 这里的变形 k-means 算法是建立在对一个社团数已知的给定的目标函数的最优化的基础之上的. 这个目标函数 (3.28) 随社团数目增加而减少, 无法用它来实现自动模型选择. 这激发了作者产生构造有效性指标 (validity index) 函数[17, 18, 44, 45, 59, 73, 74, 148, 149, 160, 171, 189, 207, 208, 211]来衡量社团结构的质量的想法, 最优社团数目可通过选择指标的最小或最大值来确定. 在网络分析的框架中, 广泛使用的著名的模量 (modularity) 函数[144]也扮演了类似的角色. 在第五章中, 作者将构造了一个新的有效性指标, 它包含每个分区的社团内部紧密程度 (compactness) 与社团间分离程度 (separation), 来度量网络社团结构的优良性. 所构造的算法不仅可以有效得到网络的社团结构, 而且不用任何关于社团结构的先验信息就可以自动确定出社团的数目[118]. 这将在第五章进行详细介绍.

3.5 小结

总之, E 等人发展了基于 Hilbert-Schmidt 度量粗粒化可逆马氏链的理论框架, 提出了一个基于最优预测理论的复杂网络分区的新方法. 这个方法切合了有关网络上动力学的情形, 并给出了服从原始网络上最优动力学的粗粒化网络. 正如 E 等人所指出的那样, 变形 k-means 算法可以应用于复杂网络分区的问题中, 效果良好, 并且相对于 [43] 中所提及的那些现有的方法, 无论在精度方面还是在计算量方面都是一个令人吸引的选择.

第四章　基于最优预测的概率性分区方法

本章将第三章介绍的工作[60]扩展到概率性的框架中[114]. 此时网络中每个节点以某一概率从属于某一社团, 而不是将节点分配到确定的社团中. 本章提出一个概率分布空间的自由能函数, 当温度为 $-\infty$ 时, 该自由能函数退化成第三章提出的目标函数(3.28). 对于这个概率性的框架, 本章也构造了相应的网络分区的算法. 这种扩展是十分自然和有价值的, 特别是对于那些没有显著社团结构的网络. 这种想法也类似于数据挖掘中的模糊分区 (fuzzy clustering)[89]以及 fuzzy c-means 算法[19, 58]. 后面的阐述中将表明模糊分区通常包含更多详细的信息.

本章提出了三种算法: 基于极小化自由能得到的 Euler-Lagrange 方程组之间的交替迭代法, 关于自由能的带投影算子的梯度下降法, 以及于关于自由能的指数变换的最速下降法. 将这些算法实施于四个算例: Zachary 空手道俱乐部网络, Gauss 混合模型生成的样本网络, 1280 个节点的 ad hoc 网络和 Mueller 势生成的样本网络. 数值结果表明交替迭代算法 AIP 通常具有更高的效率和精度. 但是作为非线性问题的迭代算法, 其收敛性不能保证. 这种情形下, 梯度下降的方法提供了一个合理的可选择的方法.

本章将阐述的内容如下. 4.1 首先简要回顾第三章中的框架, 然后介绍相应的网络概率性分区的形式. 4.2 提出三种算法, 别是交替迭代法 AIP, 带投影算子的梯度下降法 SDP, CGP 和指数变换的最速下降法 ETSD. 4.3 将算法 AIP 和 ETSD 应用于之前提到的四个算例, 并比较数值结果和算法的性能. 带投影算子的最速下降法 SDP 和共轭梯度法 CGP 的数值结果详见 [116]. 所有推导的细节参见附录 C.

本章内容主要参考 [114].

4.1 网络概率性分割的框架

首先简要回顾一下第三章[60]中提出的网络最优分区的框架. 设 $G=(S,E)$ 为一个具有 n 个节点和 m 条边的网络, 其中 S 是节点集合, $E=\{e(x,y)\}_{x,y\in S}$ 是权重矩阵, $e(x,y)$ 是连接节点 x 和 y 的边的权重. 权重矩阵的一个简单的例子就是邻接矩阵 $e(x,y)=0$ 或 1, 取决于 x 与 y 是否连接. 于是可以通过随机矩阵 $p=\{p(x,y)\}_{x,y\in S}$ 将这个网络与一个离散马氏链联系起来

$$p(x,y)=\frac{e(x,y)}{d(x)},\qquad d(x)=\sum_{z\in S}e(x,z), \tag{4.1}$$

其中 $d(x)$ 是节点 x 的度[36, 120], 这对应于网络上各向同性的随机游动, 该马氏链具有平稳分布

$$\mu(x)=\frac{d(x)}{\sum_{z\in S}d(z)}, \tag{4.2}$$

并且满足细致平衡条件

$$\mu(x)p(x,y)=\mu(y)p(y,x). \tag{4.3}$$

给定 S 的一个分区 $S=\bigcup_{k=1}^{N}S_k$, 且 $S_k\bigcap S_l=\emptyset$ 若 $k\neq l$. 设 $\hat{p}_{kl}$ 是状态空间 $\mathbb{S}=\{S_1,\ldots,S_N\}$ 中从 S_k 到 S_l 的粗粒化的转移概率, 满足

$$\hat{p}_{kl}\geq 0,\quad \sum_{l=1}^{N}\hat{p}_{kl}=1, \tag{4.4}$$

这个矩阵可以通过下述表达式自然地提升到原始状态空间上的随机矩阵空间去

$$\tilde{p}(x,y)=\sum_{k,l=1}^{N}\mathbf{1}_{S_k}(x)\hat{p}_{kl}\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}_l}, \tag{4.5}$$

其中 $\mathbf{1}_{S_k}(x)=1$ 若 $x\in S_k$, 否则 $\mathbf{1}_{S_k}(x)=0$, 并且有

$$\hat{\mu}_k=\sum_{x\in S_k}\mu(x). \tag{4.6}$$

这种压缩和提升随机矩阵的规模同时保留其随机矩阵性质的思想在 [75, 76, 107] 等文献中也曾提到过.

在第三章[60]中 E 等人介绍了随机矩阵空间中的一种度量 (Hilbert-Schmidt 范数). 设两个随机矩阵 $p_1 = (p_1(x,y))$ 和 $p_2 = (p_2(x,y))$, 定义

$$\|p_1 - p_2\|_\mu^2 = \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} |p_1(x,y) - p_2(x,y)|^2 \tag{4.7}$$

则最优分区可通过极小化 $\|\tilde{p} - p\|_\mu$ 得到.

在上述的公式中, 每个节点在分区后仅属于一个社团. 这在很多情形下非常具有局限性, 比如网络中那些共享一个或几个社团的节点, 或者说在图形表示中位于不同社团之间边缘的节点. 在社会网络中, 将人们按照他们相互之间的社会联系分成不同的社团时, 其中会有一些人以非零的概率属于不同的社团, 他们起到了中间媒介的作用. 在分子动力学中, 将轨道分成服从于不同亚稳态的不同区域时, 处于过渡带的节点停留在中间, 并起到了瓶颈的作用. 这就促使了最优分区理论向概率性的框架扩展和推广.

本章的主要思想是在 (4.5) 中, 用一般性的概率函数 $\rho_k(x)$ 代替示性函数 $\mathbf{1}_{S_k}(x)$, 这里 $\rho_k(x)$ 是节点 x 属于第 k 个社团 S_k 的概率, 从而自然地要求

$$\rho_k(x) \geq 0, \quad \sum_{k=1}^{N} \rho_k(x) = 1, \quad x \in S. \tag{4.8}$$

同理, 可以定义其诱导的马氏链的转移概率矩阵

$$\tilde{p}(x,y) = \sum_{k,l=1}^{N} \rho_k(x) \hat{p}_{kl} \rho_l(y) \frac{\mu(y)}{\hat{\mu}_l}, \qquad x,y \in S, \tag{4.9}$$

其中

$$\hat{\mu}_k = \sum_{z\in S} \rho_k(z)\mu(z). \tag{4.10}$$

这里提升随机矩阵大小的思想类似于确定性分区的情形, 它表明了这样的观点, 即节点 x 通过由社团 S_k 到社团 S_l 的不同的渠道, 并以它们相应的从属概率转移到节点 y, 最终停留在稳态. 不难验证

命题 4.1 *如果 $\hat{p}$ 是 $\mathbb{S}$ 上的随机矩阵, 具有平稳分布 $\hat{\mu}$, 则由 (4.9) 定义的 $\tilde{p}$ 为 S 上的随机矩阵, 具有平稳分布 μ. 若进一步, 有 $\hat{p}$ 满足关于 $\hat{\mu}$ 的细致平衡条件, 则 $\tilde{p}$ 满足关于 μ 的细致平衡条件.*

给定社团数目 N, 可以通过考虑如下的极小化问题

$$\begin{aligned}\min_{\rho_k(x),\ \hat{p}_{kl}} J &= \|p-\tilde{p}\|_\mu^2 = \sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}|p(x,y)-\tilde{p}(x,y)|^2 \\ &= \sum_{x,y\in S}\mu(x)\mu(y)\left(\frac{p(x,y)}{\mu(y)}-\sum_{k,l=1}^{N}\rho_k(x)\rho_l(y)\frac{\hat{p}_{kl}}{\hat{\mu}_l}\right)^2 \end{aligned} \tag{4.11}$$

服从于约束条件 (4.4) 和 (4.8), 并且从网络动力学的角度最优地约化马氏链.

极小化问题 (4.11) 可以理解为如下问题当温度 T 取 0 时的形式

$$\min_{\rho_k(x),\ \hat{p}_{kl}}\left(J+T\sum_{x\in S}\sum_{k=1}^{N}\rho_k(x)\ln\rho_k(x)\right), \tag{4.12}$$

其中非负参数 T 起到温度的作用. 当 $T=+\infty$ 时, 由于泛函 (4.12) 中表示熵的项为非正数, 故 $\rho_k(x)$ 为常数; 当 $T=-\infty$ 时, 表示熵的项变成了一个强约束, 即 $\rho_k(x)$ 或者是 0 或者是 1, 这退化成确定性分区的情形; 当 $T=0$ 时, 又得到了概率性分区的目标函数 (4.11).

为了极小化 (4.11) 中的目标函数 J, 可以定义

$$\hat{p}^*_{kl}=\frac{1}{\hat{\mu}_k}\sum_{x,y\in S}\mu(x)\rho_k(x)p(x,y)\rho_l(y), \tag{4.13}$$

这是对于确定性分区的情形进行推广所得到的. 容易验证

命题 4.2 *$\hat{p}^*$ 是 $\mathbb{S}$ 上的随机矩阵, 平稳分布为 $\hat{\mu}$, 且满足关于 $\hat{\mu}$ 的细致平衡条件.*

带有约束条件 (4.4) 和 (4.8) 的 J 的最优化问题, 相当于求 (4.11) 的稳定点, 本章用三种形式来构造这个极小化问题的算法. 为描述方便, 记 $\rho=\{\rho_k(x)\}_{k=1,\dots,N,x\in S}$ 为 $N\times n$ 的矩阵, 定义 $\hat{\mu}$ 为 $N\times N$ 的矩阵, 其元素为

$$\hat{\mu}_{kl}=\sum_{z\in S}\mu(z)\rho_k(z)\rho_l(z)=(\rho\cdot I_\mu\cdot\rho^T)_{kl}. \tag{4.14}$$

对角矩阵 I_μ 和 $I_{\hat{\mu}}$ 分别为 $n\times n$ 和 $N\times N$ 的矩阵, 其元素为

$$I_\mu(x,y)=\mu(x)\delta(x,y),\quad x,y\in S, \tag{4.15a}$$

$$(I_{\hat{\mu}})_{kl}=\hat{\mu}_k\delta_{kl},\quad k,l=1,\dots,N, \tag{4.15b}$$

其中 $\delta(x,y)$ 和 δ_{kl} 均为 Kronecker delta 符号.

4.2　算法的构造

4.2.1　基于 Euler-Lagrange 方程组的交替迭代法

引理 4.3 带有约束条件 $\sum_{k=1}^{N}\rho_k(x)=1$ 的极小化问题 (4.11) 的 Euler-Lagrange 组如下所述

$$\left(I_{\hat{\mu}}^{-1}\cdot\hat{\hat{\mu}}\right)\cdot\hat{p}\cdot\left(I_{\hat{\mu}}^{-1}\cdot\hat{\hat{\mu}}\right)=\hat{p}^{*}, \tag{4.16a}$$

$$\rho=I_{\hat{\mu}}\hat{p}^{-1}\hat{\hat{\mu}}^{-1}\rho p^{T}, \tag{4.16b}$$

关于引理 4.3 的证明见附录 C.1, 它们给出了极小值点应该满足的必要条件.

命题 4.4 如果 $\hat{\hat{\mu}}$ 可逆, 则 (4.16a) 中的 $\hat{p}$ 是一个随机矩阵. 进一步, 如果 $\hat{p}^{*}$ 满足关于 $\hat{\mu}$ 的细致平衡条件, 则 $\hat{p}$ 也满足关于 $\hat{\mu}$ 的细致平衡条件.

证明 注意到

$$\hat{\hat{\mu}}\cdot\mathbf{1}_{N\times1}=\hat{\mu},\qquad I_{\hat{\mu}}\cdot\mathbf{1}_{N\times1}=\hat{\mu}, \tag{4.17}$$

其中 $\mathbf{1}_{N\times1}=(1,\ldots,1)^{T}$ 表示 N 阶全1向量. 如果 $\hat{\hat{\mu}}$ 可逆, 通过直接计算可得

$$\hat{p}\cdot\mathbf{1}_{N\times1}=\hat{\hat{\mu}}^{-1}I_{\hat{\mu}}\hat{p}^{*}\hat{\hat{\mu}}^{-1}I_{\hat{\mu}}\cdot\mathbf{1}_{N\times1}=\mathbf{1}_{N\times1} \tag{4.18}$$

这表明 $\hat{p}$ 是为随机矩阵. 进一步, 当 $\hat{p}^{*}$ 满足关于 $\hat{\mu}$ 的细致平衡条件 $I_{\hat{\mu}}\hat{p}^{*}=\hat{p}^{*T}I_{\hat{\mu}}$ 时, 得到

$$\hat{p}\cdot I_{\hat{\mu}}^{-1}=\hat{\hat{\mu}}^{-1}I_{\hat{\mu}}\hat{p}^{*}\hat{\hat{\mu}}^{-1}=\hat{\hat{\mu}}^{-1}\hat{p}^{*T}I_{\hat{\mu}}\hat{\hat{\mu}}^{-1}=I_{\hat{\mu}}^{-1}\cdot\hat{p}^{T}, \tag{4.19}$$

从而 $\hat{p}$ 满足关于 $\hat{\mu}$ 的细致平衡条件. □

由 Euler-Lagrange 方程组 (4.16) 立即得到的一个策略是在关于 $\hat{p}$ 和 ρ 的方程之间交替迭代. 为了保证算法的可实现性, 即 $\hat{p}$ 和 ρ 的非负性和归一化条件, 需要在每次迭代后加入一个投影步, 即将最优性条件 (4.16) 变为

$$\hat{p}=\mathcal{P}\left(\hat{\hat{\mu}}^{-1}\cdot I_{\hat{\mu}}\cdot\hat{p}^{*}\cdot\hat{\hat{\mu}}^{-1}\cdot I_{\hat{\mu}}\right), \tag{4.20a}$$

$$\rho=\mathcal{P}\left(I_{\hat{\mu}}\hat{p}^{-1}\hat{\hat{\mu}}^{-1}\rho p^{T}\right). \tag{4.20b}$$

这里 $\mathcal{P}$ 是一个投影算子, 它可将一个实向量映射成一个具有非负归一化分量的向量. 这就得到了如下的算法.

算法 4.5 (带投影算子的交替迭代法, Alternating Iteration algorithm with Projections — AIP)

(1) 设置初始状态 $\rho^{(0)}$ 为算法 3.8 中变形 k-means 得到的关于网络中每个节点确定性分区的示性矩阵, $n = 0$.

(2) 执行下面的简单迭代, 直到 $\|\rho^{(n+1)} - \rho^{(n)}\| \leq E_{\text{tol}}$:

$$\hat{p}^{(n+1)} = \mathcal{P}[(\hat{\hat{\mu}}^{-1} \cdot I_{\hat{\mu}} \cdot \hat{p}^* \cdot \hat{\hat{\mu}}^{-1} \cdot I_{\hat{\mu}})^{(n)}], \tag{4.21a}$$

$$\rho^{(n+1)} = \mathcal{P}[(I_{\hat{\mu}} \hat{p}^{-1} \hat{\hat{\mu}}^{-1} \rho p^T)^{(n)}]. \tag{4.21b}$$

这里 E_{tol} 是给定的精度.

(3) 最终的 $\rho^{(n)}$ 给出每个节点的概率性分区.

本章的计算中用到了投影算子 $\mathcal{P}$ 的两种选择, 这两种投影算子的选择对于数值结果影响不大. 设 $\boldsymbol{u} = (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n$, 且 $u_i < 0$ 当 $i \in \Lambda$.

(i) 选择*1:* 投影到边界的直接投影. 当 $i \in \Lambda$, 令 $\mathcal{P}u_i = 0$; 否则, 令 $\mathcal{P}u_i = u_i / \sum_{j \notin \Lambda} u_j$.

(ii) 选择*2:* 反复投影. 首先将 $\boldsymbol{u}$ 投影到超平面 $\sum_{i=1}^n u_i = 1$, 然后检查投影过的 $\boldsymbol{u}$ 的每个分量. 如果 $u_{i_0} < 0$, 令 $\mathcal{P}u_{i_0} = 0$, 并向一个约化的超平面 $\sum_{i \neq i_0} u_i = 1$ 再次投影. 对于更低维的超平面重复此投影过程, 直到没有负的分量为止.

作者发现算法的收敛速率取决于网络的结构. 对于一个有良好分区的社团结构的复杂网络, 收敛通常很快. 但是对于一个分区较为模糊网络, 收敛就会很慢. 现在来估计每次迭代中的计算量. 注意到在 $\hat{p}$ 的迭代步中, 所有的矩阵都是 $N \times N$ 阶.

(a) $\hat{p}$ 迭代步的计算量. 易知计算 $\hat{\mu}$ 花费 $O(Nn)$, 而计算 $\hat{\hat{\mu}}$ 花费 $O(N^2 n)$. 计算 $\hat{p}^*$ 花费 $O(N^2 m)$, 其中 m 表示边数, 其在真实网络中通常假定是 $O(n)$. 计算 $\hat{\hat{\mu}}^{-1}$ 的花费是 $O(N^3)$. 因此, 计算 $\hat{p}$ 的步骤中的总花费是 $O(N^2(m+n))$.

(b) ρ 迭代步的计算量. ρp^T 花费 $O(Nm)$, $I_{\hat{\mu}} \hat{p}^{-1} \hat{\hat{\mu}}^{-1}$ 花费 $O(N^3)$. 所以计算 ρ 的总花费是 $O(N^2 n + Nm)$.

4.2.2　带投影算子的梯度下降方法

另一种显而易见的选择是用最速下降法来极小化目标函数. 易知 (4.11) 的梯度流 (gradient flow) 为

$$\frac{d\hat{p}}{dt} = -\frac{\partial J}{\partial \hat{p}}(\hat{p}, \rho), \tag{4.22a}$$

$$\frac{d\rho}{dt} = -\frac{\partial J}{\partial \rho}(\hat{p}, \rho). \tag{4.22b}$$

引理 4.6　*目标函数 J 关于 $\hat{p}$ 和 ρ 的偏导数分别为*

$$\frac{\partial J}{\partial \hat{p}} = 2\left(\hat{\mu}\hat{p}I_{\hat{\mu}}^{-1}\hat{\mu}I_{\hat{\mu}}^{-1} - I_{\hat{\mu}}\hat{p}^* I_{\hat{\mu}}^{-1}\right), \tag{4.23a}$$

$$\begin{aligned}\frac{\partial J}{\partial \rho} =& 2\Bigg[\left(\hat{p}I_{\hat{\mu}}^{-1}\hat{\mu}I_{\hat{\mu}}^{-1}\hat{p}^T + I_{\hat{\mu}}^{-1}\hat{p}^T\hat{\mu}\hat{p}I_{\hat{\mu}}^{-1}\right)\cdot \rho I_\mu - \left(\hat{p}I_{\hat{\mu}}^{-1} + I_{\hat{\mu}}^{-1}\hat{p}^T\right)\cdot \rho p^T I_\mu \\ & - \left(\left(I_{\hat{\mu}}^{-2}\hat{p}^T\right) * \left(\hat{\mu}I_{\hat{\mu}}^{-1}\hat{p}^T\hat{\mu}\right)\right)\cdot \mathbf{1}_{N\times 1}\mu^T + \left(\left(I_{\hat{\mu}}^{-2}\hat{p}^T\right) * \left(\hat{p}^*\right)^T\right)\cdot \hat{\mu}\mu^T\Bigg],\end{aligned} \tag{4.23b}$$

引理 4.6 的证明见附录 C.2. 方程组 (4.23) 给出了目标函数 (4.11) 的偏导数, 这是构造梯度方法的关键. 于是可以利用最速下降法[79, 86]求解约束优化问题 (4.11). 为了保证 $\hat{p}$ 和 ρ 的非负性和归一性, 依然需要在每一次更新后加入投影步, 于是得到如下算法

算法 4.7 (*带投影算子的最速下降法*, Steepest Descent method with Projections —SDP)

(1) *设置初始状态 $\rho^{(0)}$ 为算法 3.8 中变形 k-means 得到的关于网络中每个节点确定性分区的示性矩阵*, $n = 0$.

(2) *执行下面的简单迭代, 直到* $\|\rho^{(n+1)} - \rho^{(n)}\| \le E_{\text{tol}}$:

$$\hat{p}^{(n+1)} = \mathcal{P}\left(\hat{p}^{(n)} - \alpha\frac{\partial J}{\partial \hat{p}}(\hat{p}^{(n)}, \rho^{(n)})\right), \tag{4.24a}$$

$$\rho^{(n+1)} = \mathcal{P}\left(\rho^{(n)} - \alpha\frac{\partial J}{\partial \rho}(\hat{p}^{(n)}, \rho^{(n)})\right). \tag{4.24b}$$

其中 $\alpha > 0$ 为时间步长, E_{tol} 为给定的精度.

(3) 最终的 $\rho^{(n)}$ 给出每个节点的概率性分区.

时间步长 α 通常选取为始于一个合理的初值, 然后随迭代步数 n 减少到 0, 满足 $0 \leq \alpha(n) \leq 1$, 并且有

$$\lim_{n\to\infty} \alpha(n) = 0, \quad \sum_{n=1}^{\infty} \alpha(n) = \infty. \tag{4.25}$$

关于这种情形的一个典型的例子是 $\alpha(n) = \alpha_0/n$, 其中 α_0 为正常数[124]. 另一种更为简易的方法是将时间步长 α 固定为一个正常数[123, 125], 这也是本章采用的方法, 因为初始分化已经足够好, 使得当 α 变小时, 目标函数 (4.11) 衰减得十分缓慢, 而较大的 α 值将引起爆炸.

下面来估计每次迭代中的计算量.

(a) $\hat{p}$ 迭代步的计算量. 同 AIP 一样, 易知 $\hat{\mu}$ 的计算量为 $O(N^2 n)$, $\hat{p}^*$ 的花费为 $O(N^2 m)$, 故计算 $\hat{p}$ 的花费为 $O(N^2(m+n))$.

(b) ρ 迭代步的计算量. 由于 (4.23b) 包含 $\hat{p}^*$, 故计算 ρ的花费也为 $O(N^2(m+n))$.

另一种选择是利用共轭梯度法的简单形式来极小化目标函数, 这种技巧在机器学习中经常使用[79, 157], 它可以看作是上述最速下降法加上一个非零的动量项, 使得下降的效率大量提升. 这就引入了下面的算法.

算法 4.8 (带投影算子的共轭梯度法, Conjugate Gradient method with Projections — CGP)

(1) 设置初始状态 $\rho^{(0)}$ 为算法 3.8 中变形k-means得到的关于网络中每个节点确定性分区的示性矩阵, $n = 0$.

(2) 执行下面的简单迭代, 直到 $\|\rho^{(n+1)} - \rho^{(n)}\| \leq E_{\text{tol}}$:

$$\hat{p}^{(n+1)} = \mathcal{P}\left(\hat{p}^{(n)} - \alpha \frac{\partial J}{\partial \hat{p}}(\hat{p}^{(n)}, \rho^{(n)}) + \beta(\hat{p}^{(n)} - \hat{p}^{(n-1)})\right), \tag{4.26a}$$

$$\rho^{(n+1)} = \mathcal{P}\left(\rho^{(n)} - \alpha \frac{\partial J}{\partial \rho}(\hat{p}^{(n)}, \rho^{(n)}) + \beta(\rho^{(n)} - \rho^{(n-1)})\right), \tag{4.26b}$$

其中 $\alpha, \beta > 0$ 为时间步长, E_{tol} 为给定的精度.

(3) 最终的 $\rho^{(n)}$ 给出每个节点的概率性分区.

4.2.3　指数变换的最速下降法

如果不考虑加入投影步, 则另一种策略是利用如下形式的简单变形

$$\hat{p}_{kl}=\frac{e^{Y_{kl}}}{\sum_{m=1}^{N}e^{Y_{km}}},\quad \rho_k(x)=\frac{e^{Z_k(x)}}{\sum_{m=1}^{N}e^{Z_m(x)}}, \tag{4.27}$$

其中 $\{Y_{kl}\},\{Z_k(x)\}\in\mathbb{R}$ 分别为 $\hat{p}_{kl}$ 和 $\rho_k(x)$ 的广义坐标.

引理 4.9 定义矩阵

$$M_1=\hat{p}I_{\hat{\mu}}^{-1}\hat{\hat{\mu}}I_{\hat{\mu}}^{-1}\hat{p}^T\rho-\hat{p}I_{\hat{\mu}}^{-1}\rho p^T, \tag{4.28a}$$

$$M_2=I_{\hat{\mu}}^{-1}\hat{p}^T\hat{\hat{\mu}}\hat{p}I_{\hat{\mu}}^{-1}\rho-I_{\hat{\mu}}^{-1}\hat{p}^T\rho p^T. \tag{4.28b}$$

则极小化问题 (4.11) 的目标函数 J 关于广义坐标 (4.27) 的偏导数为

$$\begin{aligned}\frac{\partial J}{\partial Y}=&2\Bigg[\left(\hat{\hat{\mu}}\hat{p}I_{\hat{\mu}}^{-1}\hat{\hat{\mu}}I_{\hat{\mu}}^{-1}\right)*\hat{p}-(\hat{p}^*)^T*\hat{p}\\&-\operatorname{diag}\left(\hat{\hat{\mu}}\hat{p}I_{\hat{\mu}}^{-1}\hat{\hat{\mu}}I_{\hat{\mu}}^{-1}\cdot\hat{p}^T\right)\cdot\hat{p}+\operatorname{diag}\left((\hat{p}^*)^T\cdot\hat{p}^T\right)\cdot\hat{p}\Bigg],\end{aligned} \tag{4.29a}$$

$$\begin{aligned}\frac{\partial J}{\partial Z}=&2\Bigg[(M_1+M_2)*\rho-\rho\cdot\operatorname{diag}\left(\rho^T\cdot(M_1+M_2)\right)\\&-\operatorname{diag}(I_{\hat{\mu}}^{-2}\hat{\hat{\mu}}I_{\hat{\mu}}^{-1}\hat{p}^T\hat{\hat{\mu}}\hat{p})\cdot\rho+\operatorname{diag}(\hat{p}^*\hat{p}I_{\hat{\mu}}^{-1})\cdot\rho\\&+\rho\cdot\operatorname{diag}_{vm}\left(\mathbf{1}_{1\times N}\cdot\left((\hat{\hat{\mu}}\hat{p}I_{\hat{\mu}}^{-1}\hat{\hat{\mu}}I_{\hat{\mu}}^{-2})*\hat{p}\right)\cdot\rho\right)\\&-\rho\cdot\operatorname{diag}_{vm}\left(\mathbf{1}_{1\times N}\cdot\left((\hat{p}^*)^T*\hat{p}\right)\cdot I_{\hat{\mu}}^{-1}\rho\right)\Bigg]I_{\mu}.\end{aligned} \tag{4.29b}$$

其中 $*$ 表示矩阵对应元素相乘的运算, $\operatorname{diag}(A)$ 是矩阵 A 的对角线部分, $\operatorname{diag}_{vm}(u)$ 是由向量 u 的分量形成的对角矩阵.

引理 4.9 的证明见附录 C.3. 这就得到了如下形式的指数变换的最速下降法.

算法 4.10 (指数变换的最速下降法, Exponentially Transformed Steepest Descent method — ETSD)

(1) 由算法 3.8 中变形 k-means 得到 $\hat{p}^*$ 和 ρ, $n=0$.

(2) 设置矩阵初始状态 $Y_{kl}^{(0)} = \ln \hat{p}_{kl}^*$, 为简便, 取 $Z_k^{(0)}(x) = 0$ 若 $\rho_k(x) = 1$, $Z_k^{(0)}(x) = -5$ 若 $\rho_k(x) = 0$ ($\exp(-5) \approx 0.0067$).

(3) 执行下面的简单迭代, 更新 Y 和 Z, 直到 $|J^{(n+1)} - J^{(n)}| \leq E_{\text{tol}}$:

$$Y^{(n+1)} = Y^{(n)} - \alpha \frac{\partial J}{\partial Y}(Y^{(n)}, Z^{(n)}), \tag{4.30a}$$

$$Z^{(n+1)} = Z^{(n)} - \alpha \frac{\partial J}{\partial Z}(Y^{(n)}, Z^{(n)}), \tag{4.30b}$$

其中 α 是 Y 和 Z 的步长.

(4) 最终的 $\rho^{(n)}$ 给出每个节点的概率性分区.

这里在初始化步骤中取 $Z_k^{(0)}(x) = -5$ 当 $\rho_k(x) = 0$ 是很多合理选择之一, 这并不影响最终结果. 于是可以估计每次迭代的计算量如下.

(a) Y 迭代步的计算量. 类似于 AIP 算法, 计算 $\hat{\hat{\mu}}$ 花费 $O(N^2 n)$, 计算 $\hat{p}^*$ 花费 $O(N^2 m)$, 其它项由这两项可推出, 从而计算 $\partial J/\partial Y$ 花费 $O(N^2(m+n))$. 所以计算 Y 的总花费是 $O(N^2(m+n))$ 个乘法和 $O(N^2)$ 个指数运算.

(b) Z 迭代步的计算量. 由于 $\hat{p}^*$ 包含在方程中, 故计算 $\partial J/\partial Z$ 也花费 $O(N^2(m+n))$. 所以计算 Z 的总花费是 $O(N^2(m+n))$ 个乘法和 $O(Nn)$ 个指数运算.

注意到除了指数运算, ETSD 算法中每个迭代步的计算量与 AIP 算法同阶. 为了展现 AIP 算法的优越性能, 这里仅用 AIP 与 ETSD 来进行数值试验, 带有两种不同投影算子的算法 SDP 和 CGP 的实验结果详见 [116].

4.3 数值实验

本节将对以下四个算例测试本章提出的算法, 分别为: Zachary 空手道俱乐部网络, Gauss 混合模型生成的样本网络, 1280 个节点的 ad hoc 网络以及 Mueller 势生成的样本网络. 这里将比较上述 AIP 和 ETSD 的收敛速率和数值结果.

4.3.1　空手道俱乐部网络

这个网络由 Zachary 观察了美国一所大学的一个空手道俱乐部成员中的社会联系后构造的[210], 具体介绍见 1.5.2. 在空手道俱乐部网络中有 34 个节点, 如图 4.2 和 4.3 所示. 在 Zachary 最初的分区中, 每个节点仅属于一个分裂后的子俱乐部. 这里在图中将它标示为黑色或白色来表示其属性. 根据概率性分区的观点, 每个节点的在分区后的从属性不再是一个示性函数, 而是一个离散的概率分布. 下述记号中, 联合概率 ρ_K 和 ρ_W 分别表示每个节点属于黑色或白色社团的概率.

收敛速率. 图 4.1 给出了两种方法 AIP 和 ETSD 的收敛过程. 令 $E_{\mathrm{tol}} = 10^{-6}$, AIP 选用的控制为 $\|\rho^{(n+1)} - \rho^{(n)}\|$, ETSD 选用的控制为 $|J^{(n+1)} - J^{(n)}|$. 这里在计算中简单地选取 $\alpha = 20$, 这是由于数值上更大的 α 将导致爆炸. 对于 AIP 算法, 迭代步数为 47, $J_{\min} = 4.039030$, 这比算法 3.8 中的变形 k-means 得到的结果 $J_{\min} = 4.179811$ 要小. 对于 ETSD, 迭代步数需要 631, $J_{\min} = 4.039674$. 为了改进 ETSD 的精度, 可以选用更小的 E_{tol}, 其结果如表 4.1 所示. 作者发现即使取 $E_{\mathrm{tol}} = 10^{-9}$ 并经过 1944 步迭代, $J_{\min}$ 仍然没有 AIP 的结果好. 对于这种现象, 作者做了如下解释: 首先回想 Euler-Lagrange 方程组 (4.16a) 和 (4.16b) 的直接迭代给出负的分量, 这意味着当带有凸区域 $\sum_{k=1}^{N} \rho_k(x) = 1$ 的约束时, 最终的 ρ 可能会有零分量, 即 $\rho_{k_0}(x_0) = 0$. 这些零分量在 AIP 的投影步可达到. 但是在 ETSD 中做了指数变换, 这意味着相应的分量 $Z_{k_0}(x_0) = -\infty$. 为了达到这个极限, 就应该有足够长的迭代步数. 在实际的计算中, 最速下降法使分量 $Z_{k_0}(x_0)$ 趋于负数, 但是在终止准则 $|J^{(n+1)} - J^{(n)}| < E_{\mathrm{tol}}$ 的控制下进行一些步后它将会停止. 这种停止会引入对于 $\rho_k(x)$ 的显著误差. 为了得到更高的精度, 应该设定 E_{tol} 尽可能的小并迭代尽可能多的步数, 但是这可能会导致数值效率问题.

联合概率 ρ. 表 4.2 中列出了最终的分区结果, 其中 ρ_K 和 ρ_W 分别表示属于黑色或白色社团的概率, 如图 4.2 所示. 比较 AIP 和 ETSD 的 ρ_K 和 ρ_W, 作者发现几乎所有的误差都小于 10^{-2}, 但是联合概率, 或者说, 模糊分区概率 ρ, 与变形 k-means 算法得到的 0-1 分布非常不同.

现在来比较 AIP 得到的联合概率 ρ_K, ρ_W 和 Zachary 得到的最初分区结果. 在 [210] 中, Zachary 给出分区 $S_W = \{1 : 8, 11 : 14, 17, 18, 20, 22\}$ 和 $S_K =$

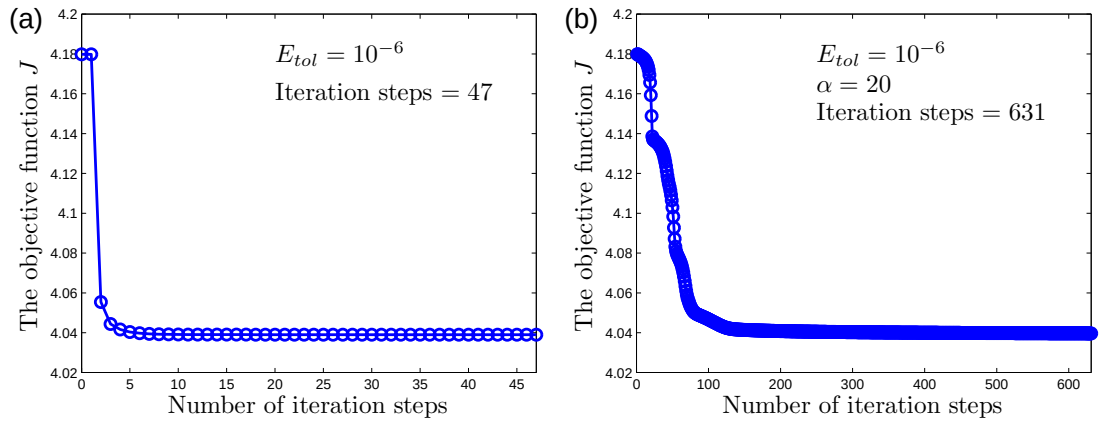


图 4.1: 目标函数 J 的收敛过程, (a) 和 (b) 分别表示 AIP 和 ETSD 的结果. 对于 AIP, 当 $E_{\rm tol} = 10^{-6}$ 时只需要迭代 47 步; 而对于 ETSD, 当 $E_{\rm tol} = 10^{-6}$ 和 $\alpha = 20$ 时则需要迭代 631 步.

表 4.1: 对于 ETSD 算法取不同的精度时的目标函数极小值. 这里 $\alpha = 20.0$.

$E_{\rm tol}$	Iteration Steps	$J_{\rm min}$
10^{-5}	183	4.040980
10^{-6}	631	4.039674
10^{-7}	1861	4.039190
10^{-8}	1901	4.039187
10^{-9}	1944	4.039188

$\{9, 10, 15, 16, 19, 21, 23 : 34\}$. 如果根据多数决定原则 (majority rule) 将节点分区, 即若 $\rho_K(x) > \rho_W(x)$ 则令 $x \in S_K$, 否则令 $x \in S_W$, 这样就得到与 Zachary 一样的分区, 如图 4.2 所示. 但实际上所得结果拥有了更多的信息, 从表 4.2 中可以发现对那些位于白色社团边缘的节点 $\{5 : 7, 11 : 13, 17 : 18, 22\}$ 有 $\rho_W = 1$, 对那些位于黑色社团边缘的节点 $\{15 : 16, 19, 21, 23 : 27, 30, 33\}$ 有 $\rho_K = 1$ (除了位于黑色社团中心的节点 33). 其它节点均以非零的概率属于黑色或白色社团, 且符合图 4.2 的直观表示. 节点 $\{3, 9, 10, 14, 20, 31\}$ 具有更模糊的概率, 它们起到了黑色和白色社团之间的过渡点的作用. 特别地, 节点 3 是最模糊的点. 用图 4.3 可以更明显地将数据 ρ 可视化, 其做法如下: 假设可视化工具中不同颜色的向量表示为

表 4.2: 网络中每个节点的联合概率, 其中 ρ_K 和 ρ_W 分别表示属于图 4.2 中黑色或白色社团的概率.

Nodes		1	2	3	4	5	6	7	8	9	10	11	12
AIP	ρ_K	0.0427	0.0821	0.4314	0.0015	0	0	0	0.0111	0.6619	0.7430	0	0
	ρ_W	0.9573	0.9179	0.5686	0.9985	1.0000	1.0000	1.0000	0.9889	0.3381	0.2570	1.0000	1.0000
ETSD	ρ_K	0.0485	0.0898	0.4412	0.0046	0.0010	0.0007	0.0007	0.0087	0.6718	0.7564	0.0010	0.0027
	ρ_W	0.9515	0.9102	0.5588	0.9954	0.9990	0.9993	0.9993	0.9913	0.3282	0.2436	0.9990	0.9973

Nodes		13	14	15	16	17	18	19	20	21	22	23	24
AIP	ρ_K	0	0.2262	1.0000	1.0000	0	0	1.0000	0.3012	1.0000	0	1.0000	1.0000
	ρ_W	1.0000	0.7738	0	0	1.0000	1.0000	0	0.6988	0	1.0000	0	0
ETSD	ρ_K	0.0014	0.2359	0.9984	0.9984	0.0012	0.0019	0.9984	0.3114	0.9984	0.0019	0.9984	0.9993
	ρ_W	0.9986	0.7641	0.0016	0.0016	0.9988	0.9981	0.0016	0.6886	0.0016	0.9981	0.0016	0.0007

Nodes		25	26	27	28	29	30	31	32	33	34
AIP	ρ_K	1.0000	1.0000	1.0000	0.9496	0.8344	1.0000	0.7210	0.8956	1.0000	0.9475
	ρ_W	0	0	0	0.0504	0.1656	0	0.2790	0.1044	0	0.0525
ETSD	ρ_K	0.9987	0.9988	0.9984	0.9570	0.8473	0.9992	0.7305	0.9026	0.9982	0.9550
	ρ_W	0.0013	0.0012	0.0016	0.0430	0.1527	0.0008	0.2695	0.0974	0.0018	0.0450

$\boldsymbol{v}_k, k = 1, \ldots, N$, 则节点 x 的颜色向量由加权平均给出

$$\boldsymbol{v}(x) = \sum_{k=1}^{N} \rho_k(x)\boldsymbol{v}_k, \quad x \in S. \tag{4.31}$$

这里可视化工具中黑色和白色的向量表示分别为 $\boldsymbol{v}_K$ 和 $\boldsymbol{v}_W$, 于是节点 x 的颜色向量为 $\rho_K(x)\boldsymbol{v}_K + \rho_W(x)\boldsymbol{v}_W$. 这样可将不同社团之间的转移更清楚地表示出来.

从这个结果中可以自然地推测位于中间的成员与两个俱乐部都有紧密联系, 他们可能很难决定当俱乐部分裂成两个后去参加那一个组, 虽然得到这样的结论, 但是作者没有额外的数据来证实这个结论.

4.3.2 Gauss 混合模型生成的样本网络

本节的第二个算例是 Gauss 混合模型生成的样本网络. 这个模型与 Penrose

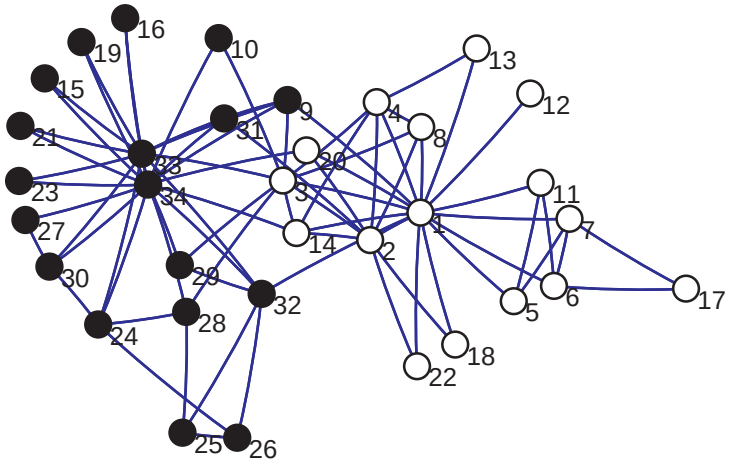


图 4.2: 利用多数决定原则得到的分区结果, 即如果 $\rho_K(x) > \rho_W(x)$ 则令 $x \in S_K$, 否则令 $x \in S_W$. 这个结果与 Zachary 给出的结果相同.

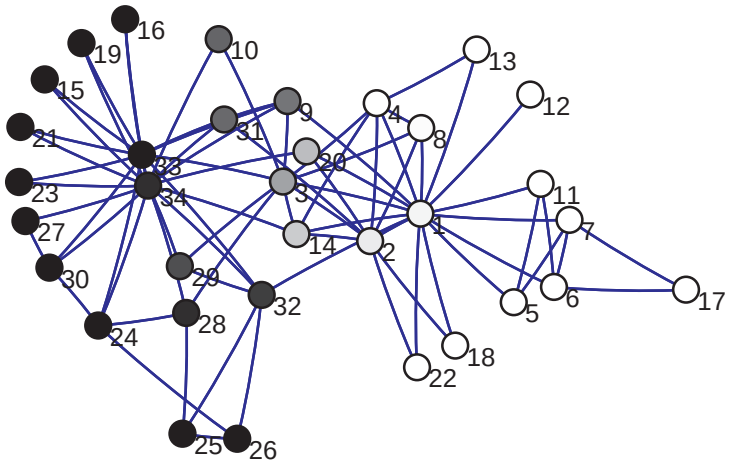
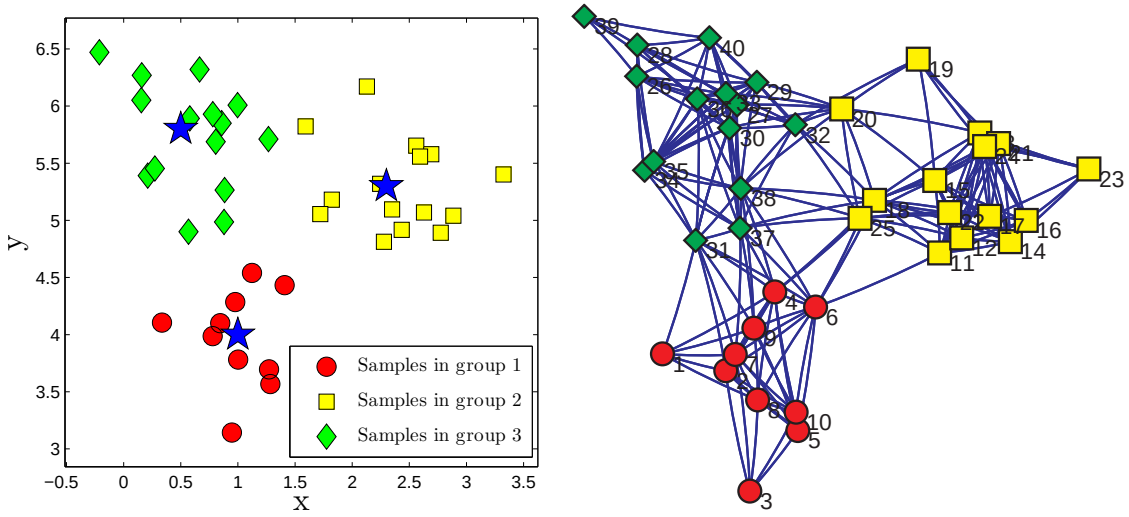


图 4.3: 空手道俱乐部网络每个节点的权重 ρ_K 和 ρ_W 的可视化. 每个节点的颜色向量为 $\rho_K \boldsymbol{v}_K + \rho_W \boldsymbol{v}_W$, 其中 $\boldsymbol{v}_K$ 和 $\boldsymbol{v}_W$ 分别表示黑色和白色的向量. 颜色越深意味着 ρ_K 的值越大, 过渡点或中立点被清楚地表示出来.

提出的随机几何图的概念[153]有关, 只是这里选取 Gauss 混合模型, 而不再是 [153] 中的均匀分布. 网络的构造过程和意义如 1.5.1 中所描述.



(a)　　(b)

图 4.4: (a) 由 3-Gauss 混合模型生成的 40 个样本点. 其中星形符号表示每个 Gauss 分量的中心, 圆形, 方形和菱形符号分别表示三个不同分量中的样本点. (b) 由 (a) 中样本点根据参数 $dist = 1.0$ 生成的网络.

首先, 选取 $n = 40$ 和 $K = N = 3$, 根据均值

$$\boldsymbol{\mu}_1 = (1.0, 4.0)^T, \boldsymbol{\mu}_2 = (2.3, 5.3)^T, \boldsymbol{\mu}_3 = (0.5, 5.8)^T, \tag{4.32a}$$

和协方差矩阵

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = 0.15\boldsymbol{I} = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{4.32b}$$

产生样本点. 为简便, 这里选取节点 1 : 10 在第 1 组, 节点 11 : 25 在第 2 组, 节点 26 : 40 在第 3 组. 于是近似地有 $q_1 = 10/40$, $q_2 = q_3 = 15/40$. 阀值则取为 $dist = 1.0$. 样本点如图 4.4(a) 所示, 相应的网络如图 4.4(b) 所示.

为了评价上述算法的结果, 首先对任意的 $\boldsymbol{x}$ 定义一个先验的模糊分区概率 $\rho_k^{\text{priori}}(\boldsymbol{x})$

$$\rho_k^{\text{priori}}(\boldsymbol{x}) = \frac{q_k p_k(\boldsymbol{x})}{\sum_{l=1}^{N} q_l p_l(\boldsymbol{x})}, \tag{4.33}$$

其中 $p_k(\boldsymbol{x})$ 是均值为 $\boldsymbol{\mu}_k$ 和协方差为 Σ_k 的 Gauss 概率密度函数. 注意到这个先验

表 4.3: 关于 3-Gauss 混合模型生成的40个节点的样本网络, AIP 和 ETSD 算法的迭代步数, 目标函数极小值 $J_{\min}$ 以及与传统 fuzzy c-means 算法和先验概率相比的 ρ 的平均和最大 L^∞ 误差.

	Iterstep	$J_{\min}$	$E_\rho^{m①}$	$E_\rho^{\infty②}$	$\bar{E}_\rho^{m③}$	$\bar{E}_\rho^{\infty④}$
AIP	27	1.1554	0.0810	0.2143	0.0628	0.3984
ETSD	859	1.1557	0.0821	0.2130	0.0628	0.4015

概率是独立于网络的拓扑结构的, 它可以被认为是一个合理的参考值而不是一个精确对象.

将本章算法的结果与 fuzzy c-means 算法[19, 58]的结果进行比较是有指导意义, 这是由于在这个情形下度量是已知的. 下面可应用 fuzzy c-means 算法来将样本点分区. 传统 fuzzy c-means 算法的主要思想是极小化目标函数

$$J_{\mathrm{FCM}} = \sum_{k=1}^{N}\sum_{i=1}^{n} \rho_k(\boldsymbol{x}_i)^b \|\boldsymbol{x}_i - \boldsymbol{m}_k\|^2, \quad b \geq 1, \tag{4.34}$$

其中 $\boldsymbol{x}_i$ 是样本点, $\boldsymbol{m}_k$ 是社团中心, 在计算中通常取 $b = 2$. $\rho_k(\boldsymbol{x}_i)$ 表示 $\boldsymbol{x}_i$ 属于第 k 个社团的概率, 它满足条件

$$0 \leq \rho_k(\boldsymbol{x}_i), \quad \sum_{k=1}^{N} \rho_k(\boldsymbol{x}_i) = 1, \quad i = 1, 2, \cdots, n. \tag{4.35}$$

于是可以对这个目标函数推导关于 $\boldsymbol{m}$ 和 ρ 的 Euler-Lagrange 方程组, 然后迭代直到找到稳定点. 更多细节参见 [19, 58].

在表 4.3 中, 对于 AIP 和 ETSD, 比较了迭代步数, 目标函数极小值 $J_{\min}$ 以及 ρ 和传统 fuzzy c-means 算法, 先验概率之间的平均和最大 L^∞ 误差. 表 4.4 列出了中间节点的联合概率. 比较 AIP 和 ETSD 可知, AIP 是更有效的. 两种算法的 ρ 之间的最大偏差小于 0.03. 将本章的算法与传统 fuzzy c-means 算法比较, ρ 的平均偏

①平均 L^∞ 误差: $\frac{1}{n}\sum_{i=1}^{n} \|\rho(\boldsymbol{x}_i) - \rho^{\mathrm{FCM}}(\boldsymbol{x}_i)\|_\infty$.

②最大 L^∞ 误差: $\max_i \|\rho(\boldsymbol{x}_i) - \rho^{\mathrm{FCM}}(\boldsymbol{x}_i)\|_\infty$.

③平均 L^∞ 误差: $\frac{1}{n}\sum_{i=1}^{n} \|\rho(\boldsymbol{x}_i) - \rho^{\mathrm{priori}}(\boldsymbol{x}_i)\|_\infty$.

④最大 L^∞ 误差: $\max_i \|\rho(\boldsymbol{x}_i) - \rho^{\mathrm{priori}}(\boldsymbol{x}_i)\|_\infty$.

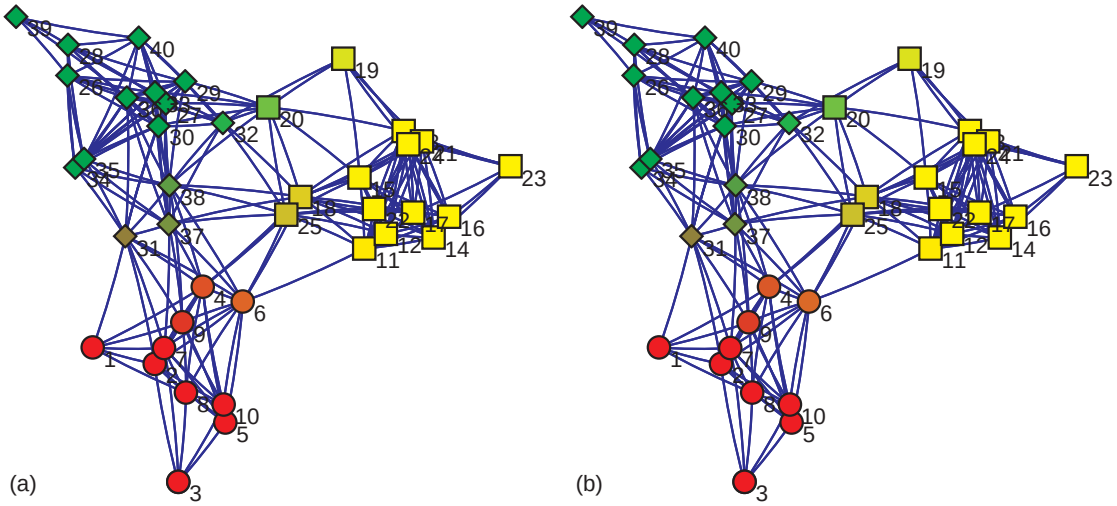


图 4.5: 权重 $\{\rho_k(x)\}$ 的可视化. 每个节点的颜色向量为加权平均 $\rho_R\boldsymbol{v}_R+\rho_Y\boldsymbol{v}_Y+\rho_G\boldsymbol{v}_G$, 其中 $\boldsymbol{v}_R,\boldsymbol{v}_Y,\boldsymbol{v}_G$ 分别表示红色, 黄色和绿色的向量. (a)和(b)分别给出了 AIP 和 ETSD 的结果. 节点 $\{4,6,9,11,18:20,25,31:32,37:38\}$ 具有明显的过渡颜色, 它们在网络中起到了过渡点的作用.

差小于 0.083, 但最大偏差约为 0.22; 与先验概率比较, ρ 的平均偏差小于 0.063, 但最大偏差约为 0.40. 可以看出具有较大偏差的节点全都位于过渡区域, 最大偏差发生在节点 20. 通过上述比较可知本章的方法达到了合理的结果并符合网络拓扑结构的直观.

权重 $\{\rho_k(x)\}$ 如图 4.5 所示. 设可视化工具中的红色, 黄色和绿色的向量表示分别为 $\boldsymbol{v}_R,\boldsymbol{v}_Y$ 和 $\boldsymbol{v}_G$, 则由 (4.31) 知, 节点 x 的颜色向量由加权平均 $\rho_R(x)\boldsymbol{v}_R+\rho_Y(x)\boldsymbol{v}_Y+\rho_G(x)\boldsymbol{v}_G$ 给出. 这就更清楚地表示了不同社团之间的过渡. 特别地, 节点 $\{4,6,9,11,18:20,25,31:32,37:38\}$ 表现出了显著的过渡行为. 如果根据多数决定原则进行分区, 即根据节点的最大权重分区, 除了节点 31, AIP 和 ETSD 给出相同的结果. 从表 4.4 中可见节点 31 以几乎相等的概率从属于绿色或红色的社团.

接下来, 选取 $n=400$ 和 $K=N=3$, 其中节点 $1:100$ 在第 1 组, 节点 $101:250$ 在第 2 组, 节点 $251:400$ 在第 3 组. 这表明近似地有 $q_1=100/400$,

表 4.4: 3-Gauss 混合模型生成的样本网络的具有中间权重的节点属于不同社团的概率. ρ_R, ρ_Y 和 ρ_G 分别表示属于红色, 黄色或绿色社团的权重, 其它节点具有 0 或 1 的权重. 节点 $\{1:3,5,7,8,10\}$ 有 $\rho_R=1$, 节点 $\{12:14,16,17,21:24\}$ 有 $\rho_Y=1$, 节点 $\{26,28,29,33,35,36,39,40\}$ 有 $\rho_G=1$. 两种算法中 $E_{\text{tol}}=10^{-6}$, ETSD 的步长为 $\alpha=26.0$.

Nodes		4	6	9	11	15	18	19	20
AIP	ρ_G	0.0944	0.0987	0.0757	0	0.0160	0.1509	0.1811	0.5965
	ρ_R	0.8247	0.7392	0.9243	0.0417	0	0.1275	0	0
	ρ_Y	0.0809	0.1621	0	0.9583	0.9840	0.7216	0.8189	0.4035
ETSD	ρ_G	0.1124	0.1169	0.0916	0.0026	0.0054	0.1646	0.1764	0.5977
	ρ_R	0.7985	0.7122	0.9051	0.0301	0.0015	0.1069	0.0019	0.0019
	ρ_Y	0.0891	0.1709	0.0033	0.9673	0.9931	0.7285	0.8217	0.4004
Nodes		25	27	30	31	32	34	37	38
AIP	ρ_G	0.2222	0.9980	0.9980	0.4994	0.7941	0.9981	0.6152	0.6943
	ρ_R	0.1805	0.0020	0.0020	0.5006	0.0084	0.0019	0.3098	0.2114
	ρ_Y	0.5973	0	0	0	0.1975	0	0.0750	0.0943
ETSD	ρ_G	0.2386	0.9977	0.9977	0.5147	0.8022	0.9981	0.6351	0.7138
	ρ_R	0.1563	0.0011	0.0011	0.4833	0.0032	0.0012	0.2845	0.1873
	ρ_Y	0.6051	0.0012	0.0012	0.0020	0.1945	0.0007	0.0804	0.0989

$q_2=q_3=150/400$. 其它的模型参数为

$$\boldsymbol{\mu}_1=(1.0,4.0)^T,\boldsymbol{\mu}_2=(2.5,5.5)^T,\boldsymbol{\mu}_3=(0.5,6.0)^T, \tag{4.36a}$$

$$\Sigma_1=\Sigma_2=\Sigma_3=0.15\boldsymbol{I}, \tag{4.36b}$$

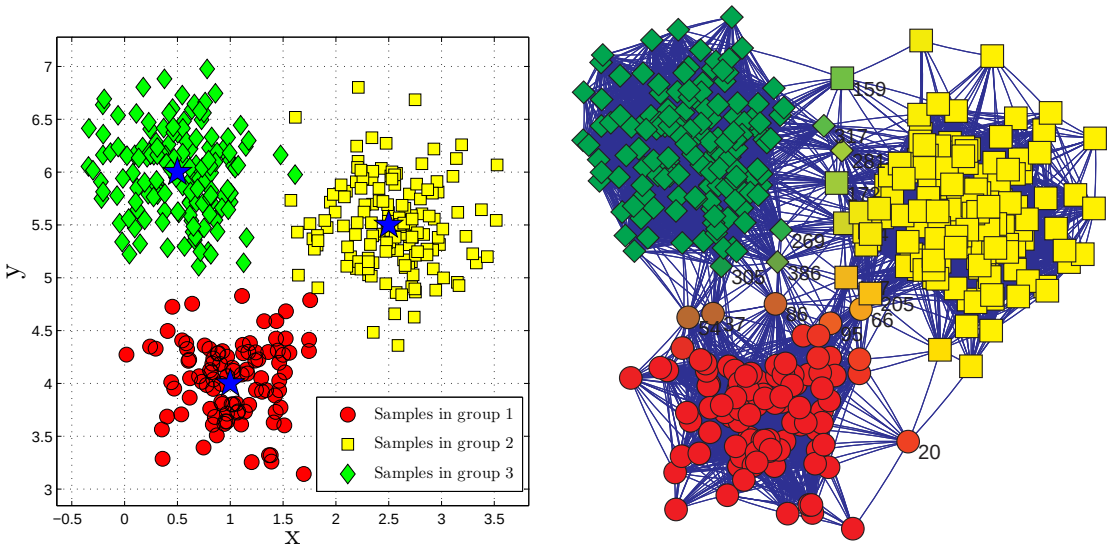
这里阀值取为 $dist=0.8$. 然后生成网络并用上述的算法进行分区. 数值结果如表 4.5 和图 4.6 所示. AIP 和 ETSD 根据多数决定原则得到的确定性分区结果相同.

4.3.3 1280 个节点的 ad hoc 网络

本节的第三个算例是 1280 个节点的 ad hoc 网络, 这类网络具有已知的社团结构, 构造如 1.5.1 所述. 但这里仅考虑大型网络: 假设选取节点数 $n=1280$, 将它们

表 4.5: 关于 3-Gauss 混合模型生成的400个节点的样本网络, AIP 和 ETSD 算法的迭代步数, 目标函数极小值 $J_{\min}$ 以及与传统 fuzzy c-means 算法和先验概率相比的 ρ 的平均和最大 L^∞ 误差.

	IterStep	$J_{\min}$	E_ρ^m	E_ρ^∞	$\bar{E}_\rho^m$	$\bar{E}_\rho^\infty$
AIP	16	1.7942	0.1037	0.3837	0.0116	0.2243
ETSD	104	1.7962	0.1014	0.4045	0.0126	0.3193



(a)　　(b)

图 4.6: 3-Gauss 混合模型生成的 400 个节点的网络由 AIP 算法得到的权重 $\{\rho_k(x)\}$ 的可视化. 每个节点的颜色向量由加权平均 $\rho_R \boldsymbol{v}_R + \rho_Y \boldsymbol{v}_Y + \rho_G \boldsymbol{v}_G$ 给出. 节点 $\{20, 37, 54, 66, 86, 95, 104, 147, 159, 172, 205, 269, 281, 305, 317, 386\}$ 具有比其它节点更混合的权重, 如图中过渡颜色所示.

分成 $N = 4$ 个社团, 每个社团有 320 个节点. 并且节点的平均度为 $\langle d \rangle = 160$. 换句话说, p_{in} 和 p_{out} 满足如下关系

$$319 p_{\text{in}} + 960 p_{\text{out}} = 160. \tag{4.37}$$

这里记 $S_1 = \{1 : 320\}, S_2 = \{321 : 640\}, S_3 = \{641 : 960\}, S_4 = \{961 : 1280\}$. 为了对一个社团结构较为模糊的网络进行测试, 故取定 $z_{\text{out}} = 960 p_{\text{out}} = 80$. 数值结

表 4.6: 关于 1280 个节点的 $z_{\text{out}} = 80$ 的 ad hoc 网络的数值结果. 两种方法的精度均为 $E_{\text{tol}} = 10^{-6}$. 最后两列显示了 ρ 和 (4.38) 中所定义的边比例 $\tilde{\rho}$ 的偏差.

	IterStep	$J_{\min}$	E_ρ^m	E_ρ^∞
AIP	907	6.603824	0.182283	0.269223
ETSD	494	6.604187	0.182256	0.266623

果如表 4.6 和图 4.7 所示. 在表 4.6 中, 作者将 $\rho_k(x)$ 与一个有趣的量 $\tilde{\rho}_k(x)$ 做比较, $\tilde{\rho}_k(x)$ 定义如下

$$\tilde{\rho}_k(x) = \frac{E_k(x)}{d(x)}, \quad k = 1, \ldots, 4, \quad x \in S, \tag{4.38}$$

其中 $E_k(x)$ 是社团 S_k 中与节点 x 连接的节点数, 从而有 $\sum_{k=1}^4 E_k(x) = d(x)$. 根据这个定义, $\tilde{\rho}_k(x)$ 表示第 k 个社团中与节点 x 连接的边所占的比例. 注意到这个量和模糊分区概率不同, 但这是一个可以用来作比较的有趣的量. 可以发现这两个量的偏差约为 0.2. 同时, 也能够注意到在这个算例中 ETSD 的迭代步数比 AIP 少, 尽管它的最终精度较差.

图 4.7 中画出了 ρ_k 和 $\tilde{\rho}_k$ $(k = 1, 2, 3, 4)$ 的概率分布函数. 可以发现 ρ_k 和 $\tilde{\rho}_k$ 的概率分布函数曲线的形状几乎相同. 注意到所有的 ρ_k 的函数在 0.7 处有一个较底的峰值, 这对应于这个社团中的节点; 并在 0.1 处达到较高的峰值, 这对应于这个社团外的节点. $\tilde{\rho}_k$ 的情形与之类似, 只是在 0.5 处达到较低峰值, 在 0.5/3 处达到较高峰值. 这里 0.5 精确地对应于参数选择 $z_{\text{out}}/d = 0.5$. 如果根据多数决定原则分割网络, 可以得到四个社团的准确分区. 这也证实了算法的精确性, 但是本章的新算法对每个节点给出了更详细的信息.

4.3.4 Mueller 势生成的样本网络

本节中将考虑将 Langevin 轨道点与阀值准则结合起来而形成的网络, 类似于 4.3.2. 结果表明了社团结构问题和转移状态理论之间的有趣的联系.

考虑 Langevin 动力学

$$d\mathbf{X}_t = -\nabla V(\mathbf{X}_t)dt + \sqrt{\varepsilon}d\mathbf{W}_t, \tag{4.39}$$

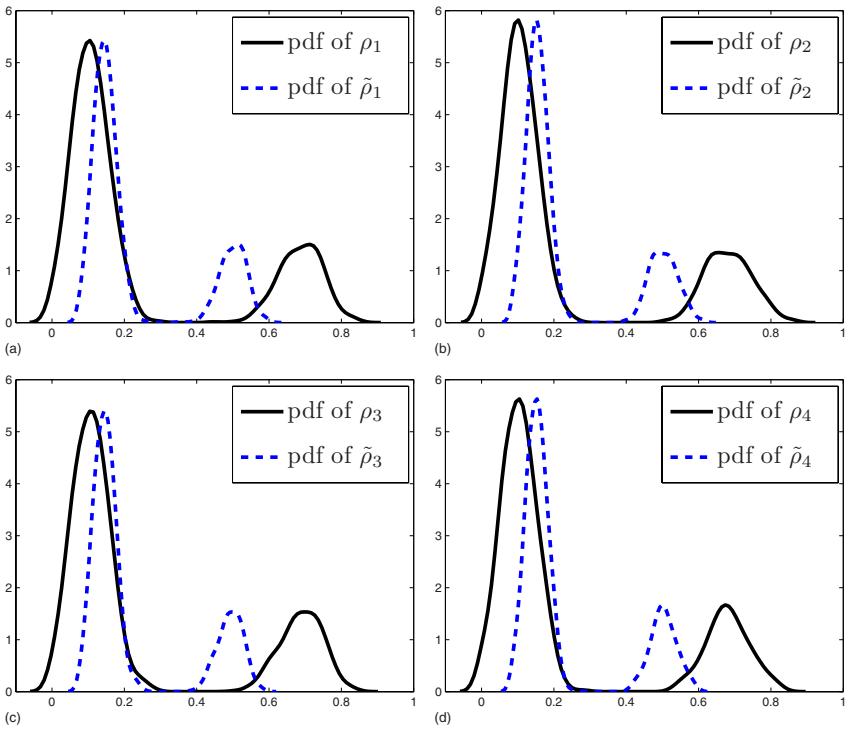


图 4.7: 1280 个节点的 ad hoc 网络的 ρ_k 和 $\tilde{\rho}_k$ $(k = 1, 2, 3, 4)$ 的概率分布函数. 实线和虚线分别表示 ρ_k 和 $\tilde{\rho}_k$ 的概率分布. 在每个图中, 较低的峰值对应于这个社团内部的节点, 较高的峰值对应于这个社团外部的节点.

这里选取 Mueller 势 $V(x, y)$ 具有如下形式

$$V(x,y) = \sum_{i=1}^{4} A_i \exp(a_i(x - x_i)^2 + b_i(x - x_i)(y - y_i) + c_i(y - y_i)^2) \tag{4.40}$$

其中参数为

$$\begin{aligned} A &= (-200, -100, -170, 15), \\ a &= (-1, -1, -6.5, 0.7), \\ b &= (0, 0, 11, 0.6), \\ c &= (-10, -10, -6.5, 0.7), \\ x &= (1, 0, -0.5, -1), \\ y &= (0, 0.5, 1.5, 1). \end{aligned}$$

如图 4.8 和图 4.9 所示, 它有三个局部极小值点, 分别标记为 A, B 和 C; 两个鞍点,

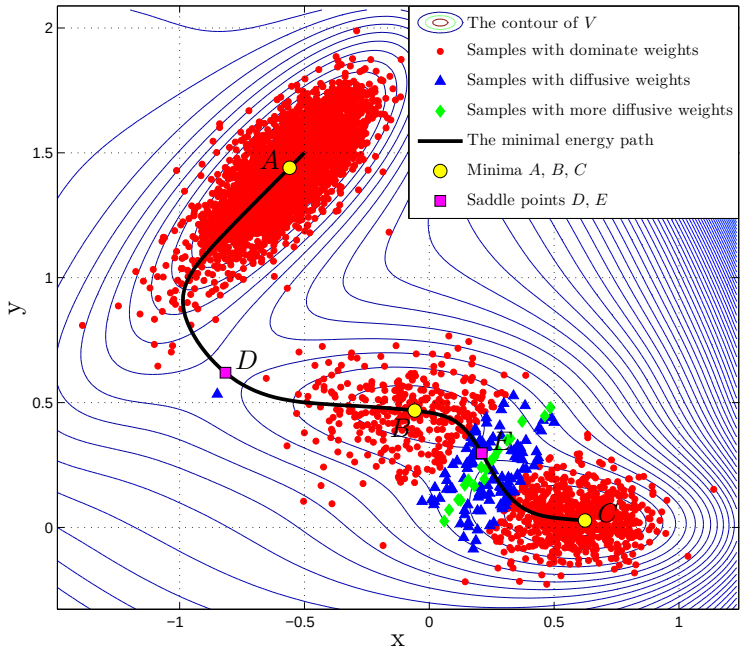


图 4.8: 网络用 AIP 得到的权重 $\{\rho_k\}$ 的可视化. 红色圆形, 蓝色三角形和绿色菱形分别表示最大权重位于区间 $[0.9, 1], [0.6, 0.9)$ 和 $[0.5, 0.6)$ 的点. 邻近于鞍点 D 的三角形节点具有权重 $(\rho_A, \rho_B, \rho_C) = (0.7211, 0.2789, 0)$, 它起到了社团 A 和 B 之间转移节点的作用. 但是社团 B 和 C 之间的转移比较扩散. 为清晰可视化故没有画出网络拓扑.

分别标记为 D 和 E. 由弦方法[61, 62]得到的从 A 到 C 得最小能量路径也在这两个图中绘出. 作为反应路径中的瓶颈的鞍点 D 和 E 起到了不同能量盆地之间的转移状态的作用.

利用如下的 Euler-Maruyama 格式[105]可以得到样本点

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \delta t \nabla V(\mathbf{X}_n) + \sqrt{\varepsilon}\delta \mathbf{W}_n, \tag{4.41}$$

其中 $\delta \mathbf{W}_n$ 为标准 Gauss 随机变量 $G(0, \delta t \boldsymbol{I})$. 这里选取 $\delta t = 0.1, \sqrt{\varepsilon} = 0.23$ 并每 10 步取一个样本点, 直到 5000 个样本点取到为止. 然后类似于 Gauss 混合模型, 根据阀值 $dist = 0.28$ 生成网络. 由于三个极小值点 A, B 和 C, 故社团结果数目的选取为 $N = 3$.

将 AIP 应用于这个网络, 算法得到目标函数值 $J_{\min} = 2.8602$, 迭代步数为 39. 图 4.8 中简明地显示了概率性分区的权重 ρ_k $(k = A, B, C)$. 最大权重位于区间

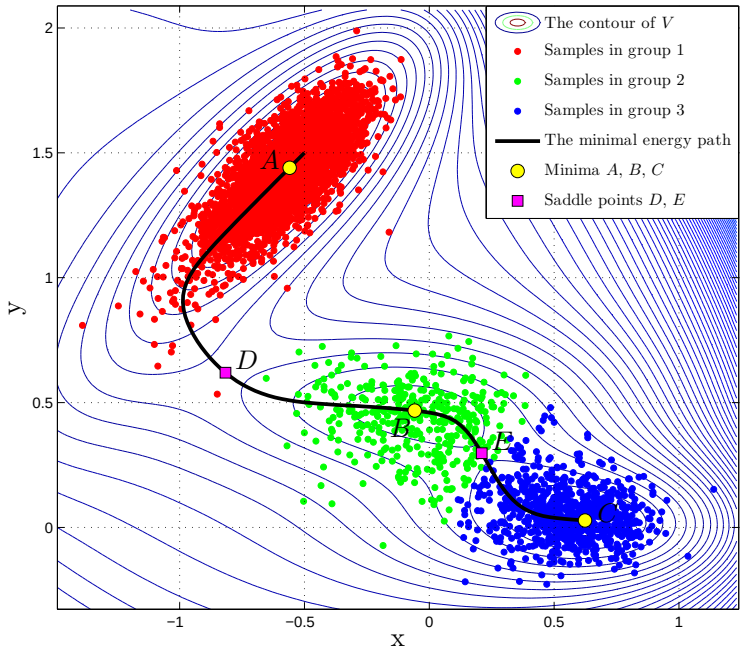


图 4.9: 根据多数决定原则分割网络. 红色, 绿色和蓝色的点分别表示属于社团 A, B 和 C 的节点. 社团结构也反映出了能量地形结构.

$[0.9, 1]$, $[0.6, 0.9)$ 和 $[0.5, 0.6)$的点分别用红色圆型, 蓝色三角形和绿色菱形作为记号. 注意到所有具有扩散权重的点聚集在鞍点 D 和 E 周围. 特别地, 唯一的邻近鞍点 D 的蓝色三角形点具有权重 $(\rho_A, \rho_B, \rho_C) = (0.7211, 0.2789, 0)$. 其它蓝色三角形和绿色菱形点由于远离社团 A 故均具有权重 $\rho_A = 0$. 除去概率性的信息, 在图 4.8 中没有给出由分区结果确定的点. 这些结果正确地表明了转移状态理论和社团结果的概率性框架之间的联系. 另外, 这个方法关于势的有限温度噪声扰动是稳定的. 根据多数决定原则分割网络则给出图 4.9, 这完美地对应于不同的能量盆地.

4.3.5　社团个数的确定

到目前为止, 本章都假定社团个数 N 是给定的. 然而在很多应用中, 这个数目是未知的并需要被确定的. 假设对于一个固定的网络, 存在一个最优的分区数目 N_0, 那么自然希望当人工选取更大的社团数目 $N > N_0$ 时, 每个节点的概率性分区的权重会趋于一个共同的极限. 也就是说, $\rho(x)$ 相应于不存在的社团的分量将会是

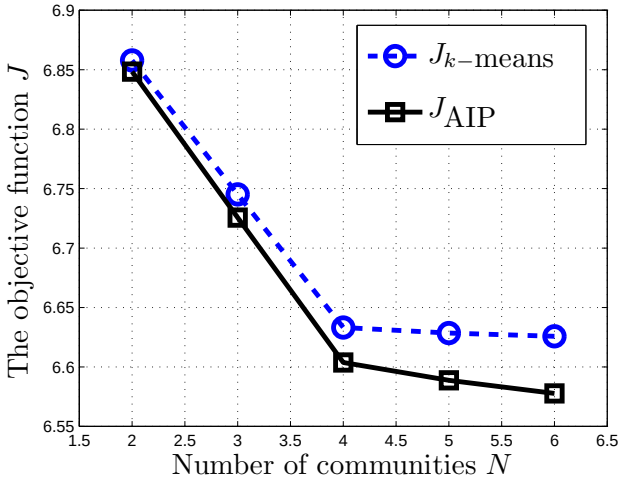


图 4.10: 目标函数 J 的极小值相应于社团数目的变化. 带圆圈的虚线表示变形 k-means 算法的结果, 带方块的实线表示 AIP 的结果. 可见随着社团数目的增加, 目标函数的极小值减小, 而且由 AIP 得到的最终的目标函数极小值比用变形 k-means 得到的小.

零. 然而这对于本章的模型是不成立的. 假设当 $N = N_0$ 时已经得到 ρ 和 $\hat{p}$, 现选取更大的 N, 并做如下扩充: 将 $\rho_k(x)$ 相应于新增加的社团的的值扩充为零, $\hat{p}$ 增加一个 $N - N_0$ 维单位矩阵. 由扩充的 ρ, $\hat{p}$ 以及社团数 N, 并忽略 $\hat{\mu}$ 的奇异性 (相应于不存在的社团有 $\hat{\mu}_k = 0$), 则目标函数 J 将和 $N = N_0$ 时相等. 这可以从 (4.9) 和 (4.10) 中看出. 变形 k-means 和 AIP 算法的极小化结果如图 4.10 所示. 当指定的社团数目增加时, 目标函数 J 是递减的, 这与第三章[60]中的变形 k-means 算法的情形类似. 事实上, 即使对于欧式空间中的点, 也不能简单地用 fuzzy c-means 算法去做模型选择.

虽然本章中的几个算法均可以成功地实施, 但是需要已知社团数目作为模型参数, 并且初值分区是由变形 k-means 算法[60]得到的确定性分区. 这无疑增加了大量的计算, 因为每个初始分区需要 500-1000 次变形 k-means 迭代来避免陷入局部极小值. 为了克服这两个弱点, 作者将在第六章 6.2 中提出了一个有效方法来实现概率性分区的自动模型选择[117]. 作者将模量的概念[144]自然地推广到概率性的形式, 即模糊模量 (fuzzy modularity), 来量化概率性分区的质量, 所构造的算法不仅可以确定每个节点属于不同社团的概率, 而且可以自动确定最优的社团个数而不

需要任何关于社团结构的先验信息. 此外, 模糊分区的概率 ρ 的初值可以被随机选取. 这些将在后面的第六章 6.2 中进行详细介绍.

4.4 小结

本章提出了关于网络分区的一个概率性框架, 它既可以看做是统计中的 fuzzy c-means 算法[19, 58]向网络分区问题的自然扩展, 也可以看做是第三章[60]中网络最优分区的确定性框架的推广. 本章构造了三种算法: 带投影的交替迭代算法 AIP, 带投影的梯度下降法 SDP, CGP 和指数变换的最速下降法 ETSD, 并成功地将它们应用于四个有代表性的算例. 数值结果表明它们可以得到相似的结果, 但是 AIP 算法具有更高的效率和精度[114, 116].

这里描述的概率性框架为网络分区问题的研究提供了一种更成熟的方法. 更重要的是, 它比传统的网络分区方法更具有预测性能. 可以想象, 例如, 将本文中的算法应用于美国参议员的投票记录, 那么可以预测谁将最有可能转变立场.

第五章　基于有效性指标的确定性分区的自动模型选择

在第三章[60]中提出了一种基于最优预测理论[31–35]的网路社团结构检测方法, 其基本思想是将网络与随机游动 Markov 动力学[120]联系起来, 然后引入马氏链空间中的一种度量, 即前向算子的 Hilbert-Schmidt 范数, 并且在这个度量下最优约化马氏链. 最终的极小化问题由传统 k-means 算法[86]的一个变形来求解. 这个方法与图像分割中的 MNCut 算法[132, 182]和数据挖掘中的扩散映射方法[107]具有一些相似之处.

在传统的聚类方法中, 标准的 k-means 算法族是建立在对一个类数已知的给定的目标函数的最优化的基础之上的[86]. 然而, 人们有时需要确定网络最优分区的社团的数目, 并遇到了 k-means 的目标函数通常随社团数目增加而减少的的困难. 这激发了人们产生构造有效性指标函数[17, 18, 44, 45, 59, 73, 74, 148, 149, 160, 171, 189, 207, 208, 211]来衡量分区结果的质量的想法, 最优分区数目可通过选择指标的最小或最大值来确定. 根据同样的思想, 本章构造了一个新的有效性指标, 它包含每个分区的社团内部紧密程度 (compactness) 与社团间分离程度 (separation), 来度量网络社团结构的优良性. 然后利用模拟退火的策略[83, 103, 133]来得到这个函数的极小值. 这种结合了之前的变形 k-means 的模拟退火方法具有高度有效性和精确性, 这是由于迭代的过程可以加速极小化有效性指标的趋势. 这个方法不仅可以有效得到网络的社团结构, 而且不用任何关于社团结构的先验信息就可以自动确定出社团的数目. 此外, 本章提出的有效性指标与网路社团结构的模量 (modularity)[144]相比具有一定的竞争力.

本章构造算法: 极小化有效性指标并结合变形 k-means 迭代的模拟退火方法 (SAVI), 来实现网络分区的自动模型选择. 算法测试于三个人工网络, 包括 ad hoc 网络, Gauss 混合模型生成的样本网络与 LFR 基准网络. 数值结果表明算法以合

理的计算量有效地实现, 并可以得到准确的分区结果. 此外, 算法成功地应用于三个真实世界中的网络, 包括空手道俱乐部网络, 宽吻海豚网络和美国足球队网络, 这巩固了算法的有效性. 需要指出的是, 尽管有效性指标不是一个崭新的概念, 但是将其应用于马氏链约化和复杂网络社团结构检测则是新颖的工作. 这也可以看作是之前的变形 k-means 算法在马氏链集合背景下的自然扩展.

本章内容组织如下. 在 5.1 中简单介绍基于最优预测理论的网络分区的框架. 在 5.2 中, 回顾一些著名的有效性指标之后, 将详细描述所提出的关于网路分区的有效性指标函数. 算法 SAVI 以及相应的策略将在 5.3 中叙述. 在 5.4 中, 将 SAVI 应用于之前提到的典型算例, 并分析分区的结果和算法的性能.

本章内容主要参考 [118].

5.1 基于最优预测的网络分区

在第三章[60]中, 介绍了一种基于最优预测[31–35]的约化随机游动 Markov 动力学的新方法. 设 $G=(S,E)$ 为具有 n 个节点和 m 条边的网络, 其中 S 为节点集合, $E=\{e(x,y)\}_{x,y\in S}$ 为权重矩阵且 $e(x,y)$ 为连接节点 x 和 y 的边上的权重. 于是可以将这个网络与离散时间的马氏链联系起来, 其随机矩阵为 $p=(p(x,y))$ 如下

$$p(x,y)=\frac{e(x,y)}{d(x)},\qquad d(x)=\sum_{z\in S}e(x,z), \tag{5.1}$$

其中 $d(x)$ 为节点 x 的度[36, 120]. 这个马氏链具有平稳分布

$$\mu(x)=\frac{d(x)}{\sum_{z\in S}d(z)}, \tag{5.2}$$

并满足细致平衡条件 $\mu(x)p(x,y)=\mu(y)p(y,x)$.

在 [60] 中的基本思想是引入随机矩阵 $p(x,y)$ 的 μ 范数, 即 Hilbert 空间 $L^2_\mu(n)$ 中与 p 相关的前向算子的 Hilbert-Schmidt 范数

$$\|p\|^2_\mu=\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}|p(x,y)|^2 \tag{5.3}$$

并通过极小化距离 $\|\tilde{p}-p\|_\mu$ 来寻找约化的马氏链 $\tilde{p}$.

给定状态空间按 S 的分区 $S=\bigcup_{k=1}^{N}S_k$, 且 $S_k\bigcap S_l=\emptyset$ 若 $k\neq l$. 设 $\hat{p}_{kl}$ 为状态空间 $\mathbb{S}=\{S_1,\ldots,S_N\}$ 上的从 S_k 到 S_l 的粗粒化的转移概率. 这个矩阵可以通过下述表达式自然地提升到原始状态空间 S 中的随机矩阵空间去

$$\tilde{p}(x,y)=\sum_{k,l=1}^{N}\mathbf{1}_{S_k}(x)\hat{p}_{kl}\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}_l}, \tag{5.4}$$

其中 $\mathbf{1}_{S_k}(x)=1$ 如果 $x\in S_k$, 否则 $\mathbf{1}_{S_k}(x)=0$, 并且

$$\mu_k=\sum_{z\in S_k}\mu(z). \tag{5.5}$$

基于这个形式, 对于任何固定的分区, 则可以找出最优的 $\hat{p}_{kl}$; 根据这个最优形式的 $\hat{p}_{kl}$, 则进一步寻找最佳分区 $\{S_1,\ldots,S_N\}$. 这是通过对于给定社团个数 N 进而极小化最优预测误差

$$\begin{aligned}\min_{\{S_k\},\hat{p}_{kl}}J &= \left\|\tilde{p}-p\right\|_{\mu}^{2}=\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}\Big[\tilde{p}(x,y)-p(x,y)\Big]^2\\ &= \sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}p^2(x,y)-\sum_{k,l=1}^{N}\frac{\hat{\mu}_k}{\hat{\mu}_l}\hat{p}_{kl}^2\end{aligned} \tag{5.6}$$

来实现的. 由直接计算可知, 当已知分区时, 则 (5.6) 的极小值点唯一, 并具有如下形式

$$\hat{p}_{kl}=\frac{1}{\hat{\mu}_k}\sum_{x\in S_k,y\in S_l}\mu(x)p(x,y). \tag{5.7}$$

可以验证 (5.7) 是一个随机矩阵, 并且 (5.5) 中的 $\hat{\mu}$ 是 $\mathbb{S}$ 上的转移矩阵为 (5.7) 的马氏链的平稳分布; 进一步, (5.7) 还满足关于 $\hat{\mu}$ 的细致平衡条件. 一个类似于 k-means 算法的方法 (变形 k-means) 用来极小化组合最优化问题 (5.6). 给定初始分区 $\{S_k^{(0)}\}_{k=1}^{N}$, 对于第 t 步, 利用

$$S_k^{(t+1)}=\left\{x:k=\arg\min_{l}\bar{E}(x,S_l^{(t)})\right\} \tag{5.8}$$

来更新状态, 其中

$$\bar{E}(x,S_k)=\sum_{l=1}^{N}\sum_{y\in S_l}\mu(x)\mu(y)\left(\frac{p(x,y)}{\mu(y)}-\frac{\hat{p}_{kl}}{\hat{\mu}_l}\right)^2. \tag{5.9}$$

这就是第三章[60]中构造变形 k-means 算法的理论基础. 此外, 这个框架已成功地推广到第四章[114, 116]的概率性分区形式.

5.2 有效性指标准则

有效性指标 (validity index) 是度量分区结果反映数据集 S 的结构的好坏程度的量. 结构最重要的指示器是分区数目, 而大多数基本分区 (聚类) 算法都假设它是一个用户定义的参数. 然而, 分区数目恰是一个关于数据结构复杂性的参数. 换句话说, 分区算法通过设置分区数目的不同初始值来运行, 并比较结果以确定合适的分区数目. 为此, 文献中介绍了各种各样的有效性指标函数. 接下来将简单地介绍一些现有的关于确定性分区和概率性分区的有效性指标的定义, 但这里并不是列出一个完整的清单. 在这里提出的新的有效性指标函数中, 则主要效仿了 Xie-Beni 定义[208], 这是归因于它的简单形式及其有效性, 这从 5.4 的数值结果中将得到证明. Xie-Beni 指标与其之前提出的其它指标相比更为精确, 并且后来的许多模式识别中的有效性指标都是从它而衍生出来的[148, 189, 207, 211].

5.2.1 确定性分区的有效性指标

5.2.1.1 Dunn 指标

一个已有的良好的确定性分区的有效性指标是分离指标 V_D[59], 它可以鉴定内部紧密且之间分离的分区, 定义如下

$$V_D = \min_{1\le k\le N}\left\{\min_{k+1\le l\le N-1}\left\{\frac{\mathrm{dis}(S_k,S_l)}{\max_{1\le m\le N}\{\mathrm{dia}(S_m)\}}\right\}\right\}, \tag{5.10}$$

其中

$$\mathrm{dia}(S_k) = \max_{x,y\in S_k}\|x-y\|, \tag{5.11}$$

$$\mathrm{dis}(S_k,S_l) = \min_{x\in S_k,y\in S_l}\|x-y\|. \tag{5.12}$$

这里 $\|\cdot\|$ 是 $\mathbb{R}^n$ 上的内积诱导出的任意距离. S 的内部紧密且之间分离的分区可通过求解 $\max_{2\le N\le n}\left\{\max_{\{S_1,\cdots,S_N\}} V_D\right\}$ 而得到, 其中 $\{S_1,\cdots,S_N\}$ 表示对于固定的 N 的最优分区. 在 [59] 中证明了如果 $V_D>1$, 则 S 的确定性分区包含 N 个紧密且分离的社团. 进一步, 如果 $V_D>1$, 则 S 至多存在一个分区. 这个有效性指标的直接实现在计算上存在主要缺陷, 这是由于当 N 和 n 增加时, 计算 V_D 的计算花费非常昂贵.

5.2.1.2 Davies-Bouldin 指标

另一个度量紧密且分离的社团的有效性指标是由 Davies 和 Bouldin 介绍的[45]. 这个指标是一个分区内部分散程度与分区之间分离程度的比值的函数. 第 k 个社团内部的分散程度计算如下

$$SC_k = \frac{1}{|S_k|} \sum_{x \in S_k} \|x - m_k\| \tag{5.13}$$

其中 $|S_k|$ 是 S_k 中数据点的个数, m_k 是社团的中心. 中心间的距离为

$$d_{kl} = \|m_k - m_l\|, \tag{5.14}$$

于是 Davies-Bouldin 指标 V_{DB} 定义为

$$V_{DB} = \frac{1}{N} \sum_{k=1}^{N} \max_{l:l \neq k} \left\{ \frac{SC_k + SC_l}{d_{kl}} \right\}. \tag{5.15}$$

这里的目标是极小化 V_{DB} 以得到合适的分区. 它与 V_D 的不同在于通过利用每个分区的平均误差来考虑平均情况.

5.2.2 概率性分区的有效性指标

在欧式空间中实现概率性分区的一个常用的方法是 fuzzy c-means 算法[19, 58]. 传统 fuzzy c-means 算法的主要思想是极小化目标函数

$$J(\rho, m) = \sum_{k=1}^{N} \sum_{x \in S} \rho_k^b(x) \|x - m_k\|^2, \quad b \geq 1, \tag{5.16}$$

其中 x 是样本点, m_k 是中心, 计算中常选取 $b = 2$. $\rho_k(x)$ 表示 x 属于第 k 社团的概率, 满足条件

$$\rho_k(x) \geq 0, \quad \sum_{k=1}^{N} \rho_k(x) = 1, \quad x \in S. \tag{5.17}$$

文献中针对于概率性分区问题而提出的有效性指标有许多形式[73, 74, 148, 149, 160, 189, 207, 208, 211], 但是这与完整清单相差甚远, 因为不同的构造根据不同的动机所提出. 这里仅仅列出其中的几个具有简单几何直观的指标.

5.2.2.1 分割系数 (partition coefficient)

作为概率性分区的有效性指标, Bezdek 设计了分割系数 (partition coefficient) V_{PC} 来度量分区之间的重叠总量[17]

$$V_{PC} = \frac{1}{n}\sum_{x\in S}\sum_{k=1}^{N}\rho_k^2(x). \tag{5.18}$$

这个形式中, V_{PC} 与模糊分区对之间的全部平均重叠量成反比. 特别地, 如果 $V_{PC}=1$, 则不存在 $\rho_k(x)$ 共享任何一对模糊分区. 求解 $\max_N\left\{\max_{\{S_1,\cdots,S_N\}} V_{PC}\right\}$ 可以得到 S 有效的分区.

修正的分割系数 (modified partition coefficient) 由 Dave 提出[44], 通过定义

$$V_{MPC} = 1 - \frac{N}{N-1}(1-V_{PC}). \tag{5.19}$$

来减少单调趋势, 注意到这个修正的分割系数与非模糊性指标[171]等价.

5.2.2.2 分割熵 (partition entropy)

分割熵 (partition entropy)也可度量模糊性 (fuzziness) 的总量[18], 其定义为

$$V_{PE} = -\frac{1}{n}\sum_{x\in S}\sum_{k=1}^{N}\rho_k(x)\log_2\rho_k(x). \tag{5.20}$$

一般而言, 可以通过求解 $\min_{2\le N\le n-1} V_{PE}$ 实现数据集 S 得最佳分区来寻找最优的社团数目.

上述的有效性指标仅利用了模糊分区概率 ρ, 这或许缺乏与数据几何结构的联系. 下面的指标同时考虑了模糊分区概率与数据结构.

5.2.2.3 Fukuyama-Sugeno 指标

Fukuyama 和 Sugeno 提出了一个有效性指标 V_{FS}[73], 其定义为

$$\begin{aligned} V_{FS} &= \sum_{x\in S}\sum_{k=1}^{N}\rho_k^2(x)\|x-m_k\|^2 - \sum_{x\in S}\sum_{k=1}^{N}\rho_k^2(x)\|m_k-\bar{m}\|^2 \\ &= J(\rho,m)+K(\rho,m), \end{aligned} \tag{5.21}$$

其中 $\bar{m}=\sum_{k=1}^{N} m_k/N$, 这里的 $J(\rho,m)$ 是传统 fuzzy c-means 的目标函数取 $b=2$ 的情形, 它度量分区内部紧密程度, 而 $K(\rho,m)$ 度量分区间的分离程度. 从而这里的目标是通过极小化 V_{FS} 来寻找模糊分区.

5.2.2.4　Xie-Beni 指标

另一个著名的有效性指标称为 Xie-Beni 指标[208] , 它可以显式地写成

$$V_{XB}=\frac{\sum_{x\in S}\sum_{k=1}^{N}\rho_k^2(x)\|x-m_k\|^2}{n\min_{k\neq l}\|m_k-m_l\|^2}=\frac{J(\rho,m)}{nK(m)}. \tag{5.22}$$

更重要地, 极小化 V_{XB} 对应于极小化取 $b=2$ 时的传统 fuzzy c-means 的目标函数 $J(\rho,m)$. 另外的因子 $K(m)$ 为分区间分离程度度量. 分区之间越分离, $K(m)$ 值越大, 故 V_{XB} 值越小. 更多的关于 Xie-Beni 指标的分析参见 [149]. 此后大量的各种形式的有效性指标函数层出不穷, 于此不再赘述[74, 148, 160, 189, 207, 211].

5.2.3　网络分区的有效性指标

作者在 5.1 的形式中, 采用同时考虑社团内部紧密程度 (compactness) 与社团间分离程度 (separation) 的思想, 并构造如下的对于网络分区的有效性指标

$$V_{\text{net}}=J(\hat{p})\cdot K(\hat{p}), \tag{5.23}$$

其中 $J(\hat{p})$ 为 (5.6) 中的目标函数, 它反映社团内部的紧密程度; 而因子

$$K(\hat{p})=\frac{N-\sum_{k=1}^{N}\hat{p}_{kk}}{\sum_{k=1}^{N}\hat{p}_{kk}}\cdot\frac{1}{N-1}=\frac{\sum_{k\neq l}\hat{p}_{kl}}{\sum_{k=1}^{N}\hat{p}_{kk}}\cdot\frac{1}{N-1}=\frac{\frac{1}{N(N-1)}\sum_{k\neq l}\hat{p}_{kl}}{\frac{1}{N}\sum_{k=1}^{N}\hat{p}_{kk}} \tag{5.24}$$

起到了类似 (5.22) 中 $K(m)$ 那样的度量社团之间分离程度的作用. 这里 $\frac{1}{N(N-1)}\sum_{k\neq l}\hat{p}_{kl}$ 表示从一个社团转移到另一个社团的平均转移概率, 而 $\frac{1}{N}\sum_{k=1}^{N}\hat{p}_{kk}$ 表示从一个社团转移到其自身的平均转移概率. 一个理想的分区要求应是空间 $\mathbb{S}=\{S_1,\ldots,S_N\}$ 中的一个比较稳定的状态, 具有较小的 $\frac{1}{N(N-1)}\sum_{k\neq l}\hat{p}_{kl}$ 和较大的 $\frac{1}{N}\sum_{k=1}^{N}\hat{p}_{kk}$. 因此, 最优分区可通过求解

$$\min_{N}\left\{\min_{\{S_1,\cdots,S_N\}}V_{\text{net}}\right\} \tag{5.25}$$

而得到. 根据 (5.6), 于是将

$$V_{\text{net}} = \frac{1}{N-1}\Big[\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}p^2(x,y) - \sum_{k,l=1}^{N}\frac{\hat{\mu}_k}{\hat{\mu}_l}\hat{p}_{kl}^2\Big]\cdot\frac{N-\sum_{k=1}^{N}\hat{p}_{kk}}{\sum_{k=1}^{N}\hat{p}_{kk}} \tag{5.26}$$

作为紧随框架 [60] 的对于网络分区的一个新的有效性指标.

作为一个模型选择的框架, 在上述的设置中通常会遇到如何调整参数来控制社团内部紧密程度 $J(\hat{p})$ 和社团之间分离程度 $K(\hat{p})$ 之间竞争力的问题, 这可简单包含于当前情形中. 现在引入定义

$$V_{\text{net}}^{\lambda} = J(\hat{p})\cdot K(\hat{p})^{\lambda}, \quad \lambda\in\mathbb{R} \tag{5.27}$$

其中 λ 是正则化参数. 当 $\lambda=0$ 时, V_{net}^{λ} 退化成 $J(\hat{p})$. 在下述的所有数值实验中, 将仅考虑情形 $\lambda=1$, 它已然给出了满意的结果.

5.3 算法的构造

模拟退火算法最初是由模拟退火固体的物理过程而产生的[133]. 这个过程描述如下: 首先, 一个固体晶体由高温而受热, 然后慢慢冷却直到系统在任何时间近似热力学平衡. 在平衡状态, 可能存在许多构造, 每一个对应于一个特定的能量级. 接受从当前构造变为一个新的构造的机会与两个状态之间的能量差有关. 模拟退火策略广泛地应用于最优化问题[103].

设 $E = V_{\text{net}}$, $E^{(t)}$ 和 $E^{(t+1)}$ 分别表示当前能量和新能量. $E^{(t+1)}$ 总被接受如果它满足 $E^{(t+1)} < E^{(t)}$, 但是如果 $E^{(t+1)} > E^{(t)}$ 则新能量级仅以概率 $\exp(-\frac{1}{T}\triangle E^{(t)})$ 被接受, 其中 $\triangle E^{(t)} = E^{(t+1)} - E^{(t)}$ 为能量差, T 为当前温度. 较高的能量状态有可能被接受这一事实使得避免了被困于局部极小值. 于是温度逐渐降低, 重复退火过程直到没有更新的改进出现或者达到某个终止准则.

在给定温度, 新状态 $\{S_k^{(t+1)}\}_{k=1}^{N}$ 以概率 $\exp(-\frac{1}{T}\triangle E^{(t)})$ 被接受, 其中能量用来衡量一个分划. 初始状态由变形 k-means 算法[60]取随机初始 N 个社团计算后而产生, 其中 N 为区间 $[N_{\min}, N_{\max}]$ 中的整数, 在下面的计算中选取 $N_{\min}=2, N_{\max}=n/3$. 初始温度设置为一个较高的温度 $T_{\max}$, 则下一个提议的状态是由将变形 k-means 应用于根据下述两个建议产生的初始状态而得到的, 这个

新的状态被接受如果满足接受条件. 这个过程将在给定的温度上重复 R 次. 设置冷却速率 $0 < \alpha < 1$ 来降低当前温度直到温度下界 $T_{\min}$ 被达到. 整个的结合变形 k-means 迭代算法的模拟退火来极小化有效性指标的过程概括如下.

算法 5.1 (Simulated annealing algorithm to minimize the validity index — SAVI[①])

(1) 设置参数 $T_{\max}$, $T_{\min}$, α 和 R. 在区间 $[N_{\min}, N_{\max}]$ 中随机选取 N, 并随机初始化分区 $\{S_k^{(0)}\}_{k=1}^{N}$. 设当前温度 $T = T_{\max}$.

(2) 根据 (5.7) 和 (5.8), 用变形 k-means 计算相应的 $\hat{p}_{kl}^{(0)}$ 和 $\{S_k^{(0)}\}_{k=1}^{N}$. 利用定义 (5.26) 计算能量 E^*.

(3) 对于 $t = 0, 1, \ldots, R$, 做如下迭代:

(3.1) 根据下述提议产生新的分区 $\{S_k^{(t)}\}_{k=1}^{N'}$ 作为初始分区, 并设 $N = N'$.

(3.2) 根据 (5.7) 和 (5.8), 用变形 k-means 更新相应的粗粒化转移概率 $\hat{p}_{kl}^{(t)}$ 和分区 $\{S_k^{(t+1)}\}_{k=1}^{N}$. 根据 (5.26) 更新新能量 $E^{(t+1)}$.

(3.3) 根据标准 Metropolis 准则接受新的分区, 即以概率 $\min\{1, \exp(-\frac{1}{T}\Delta E^{(t)})\}$ 接受; 令 $t = t + 1$.

(3.4) 更新最优状态. 如果 $E^{(t)} < E^*$, 则令 $E^* = E^{(t)}$, 并记录当前分区.

(4) 降温 $T = \alpha \cdot T$. 如果 $T < T_{\min}$, 执行 (5); 否则重复 (3).

(5) 输出最优分区 $\{S_k\}_{k=1}^{N}$ 和最小能量 E^*.

对于步骤 (3.1) 中产生一个新分区的集合的过程的提议由两个操作组成, 分别是删除一个当前社团和分裂一个当前社团. 在每次变形 k-means 迭代时, 随机选择两个操作中的一个操作, 并且社团强度

$$M_k = \hat{p}_{kk}, \quad k = 1, \cdots, N \tag{5.28}$$

被用来选择一个社团, 它反映第 k 个社团呆在自身而不趋向于转移到其它社团中的概率. 显然, 如果社团强度越大, 则它的社团结构性越强. 这两个操作描述如下.

(a) 删除一个社团. 具有最小社团强度 M_d 的社团被确定, 其中 $d = \arg\min_k M_k$. 将其从当前的分区中删除, 并将其中包含的节点合并到社团 S_k, 其中 $k = \arg\max_m \hat{p}_{dm}$.

[①] 算法的 matlab 程序下载链接为: http://dsec.pku.edu.cn/~tieli/software/SAVI.zip.

(b) **分裂一个社团**. 具有最大社团强度 M_s 的社团被选择, 并将其随机分裂成两个相等大小的新社团. 如果节点数 n_s 是奇数, 两个子社团的大小则分别为 $(n_s+1)/2$ 和 $(n_s-1)/2$.

注意到另外也可以通过对于所有可能的 N 利用变形 k-means 来得到有效性指标 (5.26) 的全局极小值. 但是这将花费太大, 因为对于每个固定的 N, 变形 k-means 需要重复 $O(10^2)$ 次来得到一个可以信赖的良好分区. 然而上述的模拟退火过程可以避免这种重复以及一个接一个地搜寻最优社团数目. 接下来的实验表明整个算法的数值性能是非常有效且成功的.

5.4 数值实验

在本节中, 将算法 SAVI 测试于具有已知社团结构的人工生成的网络, 包括 128 个节点的 ad hoc 网络, Gauss 混合模型生成的样本网络和 LFR 基准网络. 随后, 算法成功地应用于真实世界中的网络, 包括空手道俱乐部网络, 神奇湾宽吻海豚网络和美国足球队网络.

5.4.1 人工生成的网络

5.4.1.1 128 个节点的 ad hoc 网络

首先将本章的方法用于 128 个节点的 ad hoc 网络, 这类网络具有已知的社团结构, 构造如 1.5.1 所述. 通常定义 z_{out} 为某个节点与属于其它社团节点之间连接的平均数, z_{out} 越大, 社团就变得越模糊 (diffuse).

现将 $z_{\text{out}} = 96p_{\text{out}}$ 从 0.5 变化到 8, 并观察节点识别的正确率. 通过设置参数 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 20$, 用 SAVI 进行分区. 节点识别的正确率随 z_{out} 的变化如图 5.1 所示. 与 [144] 中的两种方法进行比较, 可以看出 SAVI 的性能显著优于先前的两种方法, 特别是当 z_{out} 很大时的较为模糊的情形.

为了测试于一个良分区的网络, 选取 $z_{\text{out}} = 5$. 当没有应用退火策略时, 即仅应用变形 k-means 算法, 有效性指标 V_{net} 和目标函数 J 随社团数目 N 的变化如图 5.2(a) 和表 5.1 所示. 可以看出最优社团结构在 $N = 4$ 处达到, 相应的有效性指标

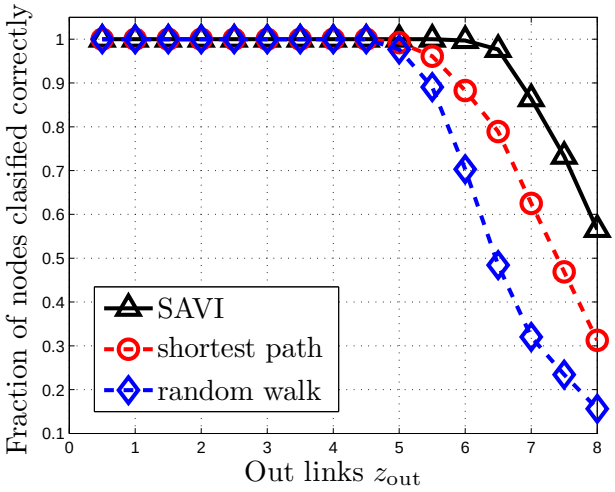
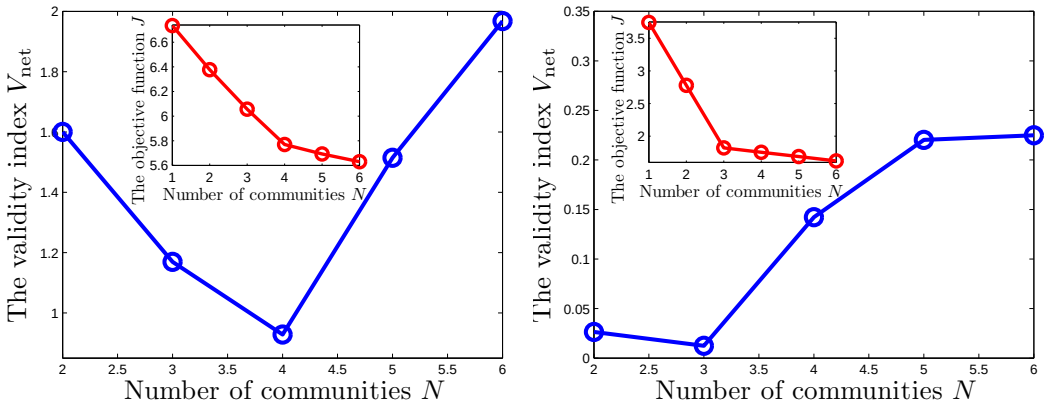


图 5.1: 由 SAVI 和 [144] 中方法所得到的节点识别的正确率随 z_{out} 的变化. 从图中可见 SAVI 的性能优于最短路径方法和随机游动方法[144].



(a)　　(b)

图 5.2: 由变形 k means 得到的有效性指标 V_{net} 和目标函数 J 随社团数目 N 的变化. 其中每个图的全局极小值点恰为 SAVI 得到的最优社团数目. (a) $z_{\text{out}} = 5$ 的 ad hoc 网络. (b) 400 个节点的 Gauss 混合模型生成的样本网络.

的值为 $V_{\text{net}} = 0.9281$. 本章的 SAVI 算法可以在不知道社团数目作为先验参数时得到期望的结果.

表 5.1: 对于图 5.2 中的 $z_{\text{out}} = 5$ 的 ad hoc 网络和一个 400 个节点的 Gauss 混合模型生成的网络, 其有效性指标 V_{net} 和目标函数 J 的值随社团数目 N 的变化.

	N	2	3	4	5	6
Ad hoc network	V_{net}	1.6001	1.1694	**0.9281**	1.5143	1.9677
($z_{\text{out}} = 5$)	J	6.3766	6.0560	5.7696	5.6931	5.6293
Gaussian mixture	V_{net}	0.0264	**0.0124**	0.1422	0.2203	0.2249
network	J	2.7795	1.8218	1.7550	1.6891	1.6223

5.4.1.2 Gauss 混合模型生成的样本网络

本节的第二个算例是 Gauss 混合模型生成的样本网络. 这个模型与 Penrose 提出的随机几何图的概念[153]有关, 只是这里选取 Gauss 混合模型, 而不再是 [153] 中的均匀分布. 网络的构造过程和意义如 1.5.1 中所描述.

选取 $n = 400$ 和 $K = 3$, 然后根据如下的均值和协方差矩阵生成样本点

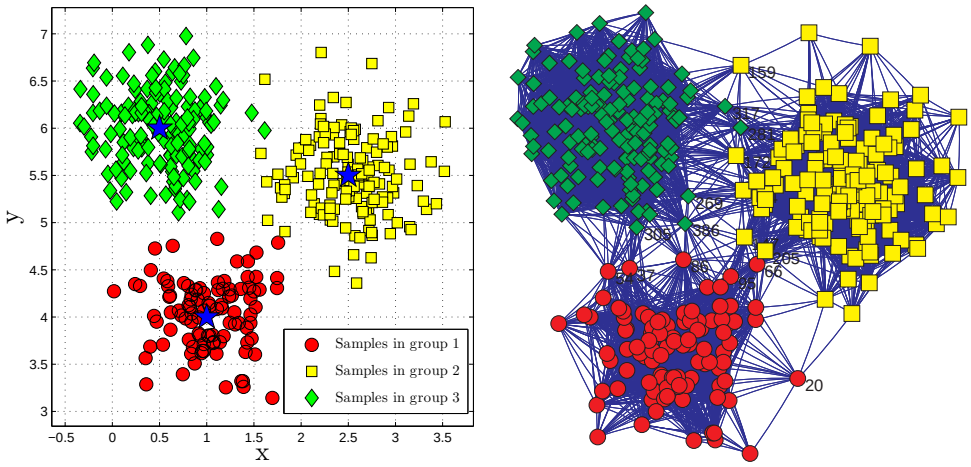
$$\boldsymbol{\mu}_1 = (1.0, 4.0)^T, \boldsymbol{\mu}_2 = (2.5, 5.5)^T, \boldsymbol{\mu}_3 = (0.5, 6.0)^T, \tag{5.29a}$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = 0.15\boldsymbol{I} = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{5.29b}$$

这里选择节点 1 : 100 在第 1 组, 节点 101 : 250 在第 2 组, 节点 251 : 400 在第 3 组. 根据这个选择, 近似地有 $q_1 = 100/400$, $q_2 = q_3 = 150/400$. 这个实验中取阀值 $dist = 0.8$. 样本点及其相应的网络如图 5.3 所示. 由变形 k-means 计算出的有效性指标 V_{net} 与目标函数 J 随社团数目 N 的变化如图 5.2(b) 和表 5.1 所示. 可以发现最优社团结构在 $N = 3$ 达到, 相应的有效性指标值为 $V_{\text{net}} = 0.0124$. 取 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 20$, 应用 SAVI, 也得到 $N = 3$ 和 $V_{\text{net}} = 0.0124$. 分区结果如图 5.4 所示. 只有节点 $\{66, 159, 281\}$ 与欧式空间中生成的初始样本组不一致. 本章的 SAVI 算法得到了合理的分区结果, 这是符合网络拓扑可视化的直观的.

5.4.1.3 LFR 基准网络

LFR 基准网络[109, 111]是为探检测社团结构的一个现实主义的网络, 如 1.5.2 中所述: 它同时要求节点度和社团规模的非均匀性. 节点度服从指数为 γ 的幂律分



(a)　　(b)

图 5.3: (a) 由 3-Gauss 混合模型生成的 400 个样本点. 星形符号表示每个 Gauss 分量的中心; 圆形, 方形和菱形符号分别表示三个不同分量中的样本点. (b) 由 (a) 中样本点根据参数 $dist = 0.8$ 生成的网络.

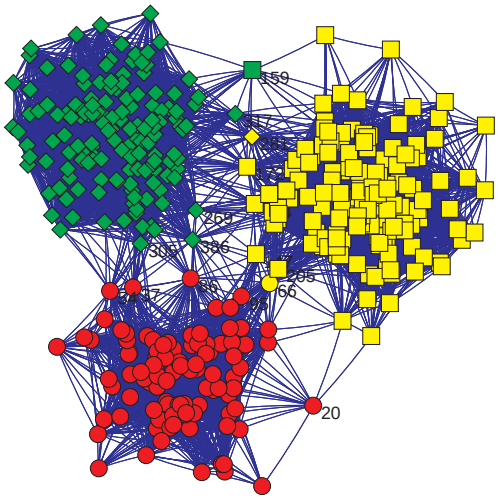


图 5.4: 由 SAVI 方法得到的 400 个节点的 Gauss 混合网络的社团结构. 只有节点 $\{66, 159, 281\}$ 与欧式空间中生成的初始样本组不一致.

布, 而社团规模服从指数为 β 的幂律分布. 混合参数 μ 作为独立参数, 它表示一个节点关于它所在社团的外面的度与全部度之间的比率. 为比较固定模块结构与算

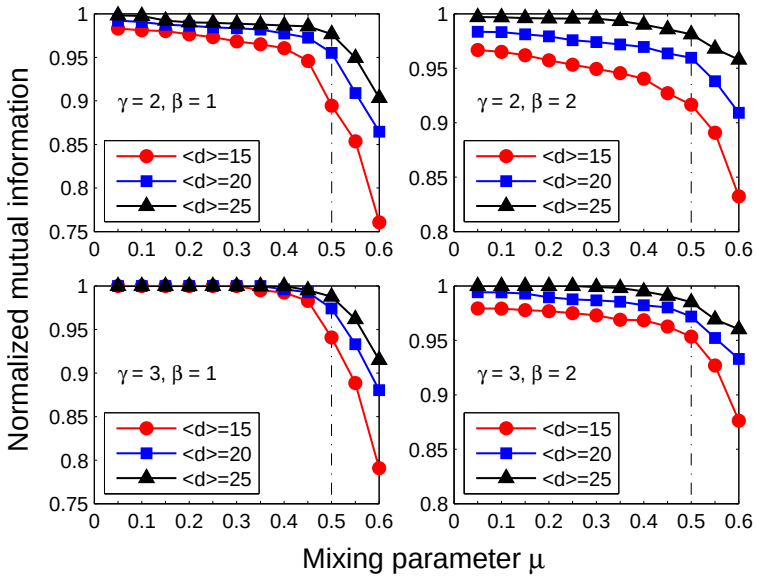


图 5.5: 将 SAVI 算法测试于 LFR 基准网络[109, 111]. 节点数为 $n = 500$. 结果明显地依赖于基准网络的所有参数, 从指数 γ 和 β 到平均度 $\langle d\rangle$. 由垂直虚线表示的阀值 $\mu_c = 0.5$ 标记出一个边界; 超出这个边界则社团不再是强意义下(即每个节点在自己从属的社团中比在其它社团中具有更多的邻居)所定义的. 每个点对应于超过 20 次的图实现的平均值. 全部结果表明 SAVI 算法对于检测社团结构给出很好的精度. 对于归一化互信息, 当 $\mu \leq \mu_c$ 时所得的结果都大于 0.9, 并且对于社团结构较为模糊的情形也是非常具有竞争力的.

法得到的结构, 这里不再使用 ad hoc 算例中的节点识别的正确率, 而是采用归一化互信息 (1.8), 其定义在 1.5.1 中也有过详细介绍[43, 109, 111].

在图 5.5 中, 展示了将 SAVI 算法应用于 $n = 500$ 的基准网络的结果. 算法的参数设置为 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 20$. 图中所示的四个子图分别对应于取值为四对指数 $(\gamma, \beta) = (2,1), (2,2), (3,1), (3,2)$ 的结果. 为了探索网络结构的万象, 这里选择指数范围的极端的组合. 每条曲线展示了归一化互信息随混合参数 μ 的变化. 可以看出当平均度 $\langle d\rangle$ 较大时 SAVI 算法的性能更好, 但是当混合参数较大时性能变差. 图中垂直虚线表示的阀值 $\mu_c = 0.5$ 标记出一个边界; 超出这个边界则社团不再是强意义下 (即每个节点在自己从属的社团中比在其它社团中具有更多的邻居) 所定义的. 在图 5.6 中, 对于 $\langle d\rangle = 20$ 且其它参数取值

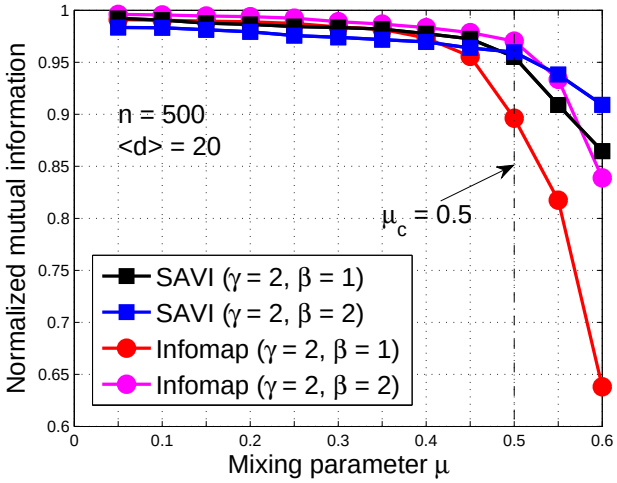


图 5.6: 将 SAVI 算法与 Informap 算法[170]测试于 LFR 基准网络[109, 111]并进行比较. 节点数为 $n = 500$, 平均度为$\langle d\rangle = 20$. 结果表明 SAVI 算法与 Informap 算法相比非常具有竞争力. 当 μ 很小时, 两种方法都给出归一化互信息接近于 1 的很好的精度. 对于社团较为模糊的情形 $\mu > \mu_c = 0.5$, SAVI 算法的性能优于 Infomap 算法.

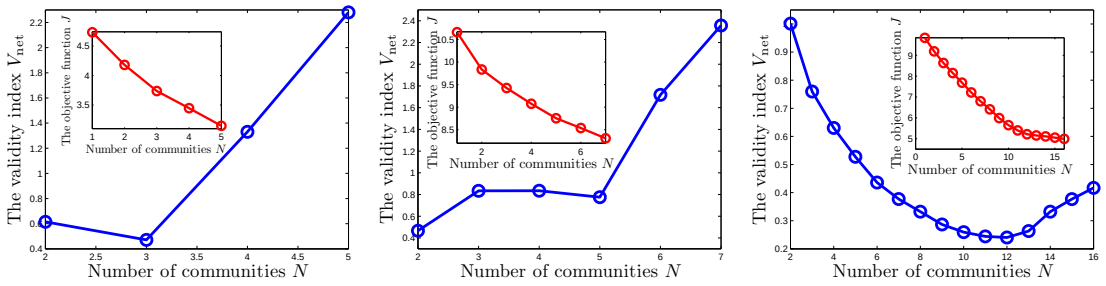
为$(\gamma, \beta) = (2, 1), (2, 2)$ 的情形, 比较了 SAVI 算法与 Infomap 算法[170]. 结果表明 SAVI 与 Infomap 相比非常具有竞争力, 特别是对于当混合参数 μ 很大时的较为扩散的情形. 这些结果都支持了 SAVI 算法的有效性.

5.4.2　真实世界中的网络

5.4.2.1　空手道俱乐部网络

这个网络是由 Wayne Zachary 在观察一所美国大学空手道俱乐部成员之间的社交而构建的[210], 具体介绍见 1.5.2.

当仅应用变形 k means 时, 有效性指标 V_{net} 和目标函数 J 随社团数目 N 的变化如图 5.7(a) 和表 5.2 所示. 取 $N = 2$ 和 $N = 3$ 所得到的社团结构如图 5.8 所示. 从中可以发现最优社团结构于 $N = 3$ 达到, 相应的有效性指标为 $V_{\text{net}} = 0.4711$. 注意到这个结果与 Zachary 观察到的原始分区不同, 但是从网络拓扑及其最终分区 (图 5.8(b)) 来看是合理的, 正如 [144] 中最大化模量得到 4 个社团一样. 如果选取参数 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 50$, 执行 SAVI, 得到与 $N = 3$ 的变形



(a) (b) (c)

图 5.7: 由变形 k means 得到的有效性指标 $V_{\rm net}$ 和目标函数 J 随社团数目 N 的变化. 其中每个图的全局极小值点恰为 SAVI 得到的最优社团数目. (a) 空手道俱乐部网络. (b) 宽吻海豚网络. (c) 美国足球队网路.

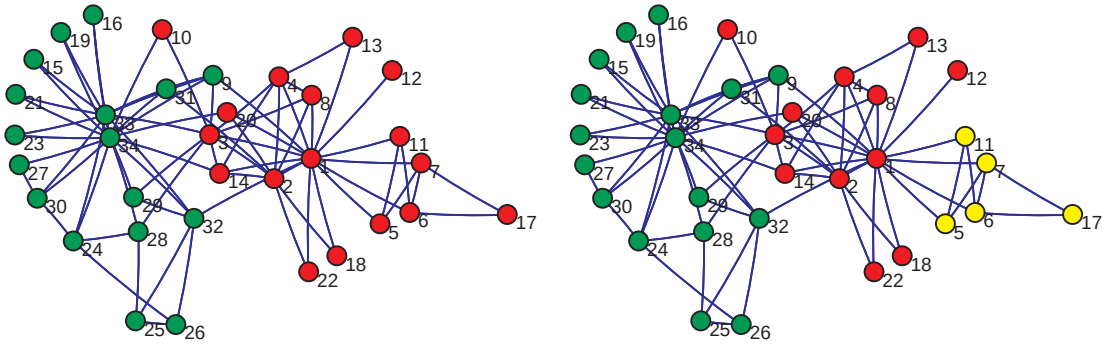
表 5.2: 对于图 5.7 中的空手道俱乐部网络, 宽吻海豚网络和美国足球队网路, 其有效性指标 $V_{\rm net}$ 和目标函数 J 的值随社团数目 N 的变化.

The karate club network			The Dolphins network			The football team network		
N	$V_{\rm net}$	J	N	$V_{\rm net}$	J	N	$V_{\rm net}$	J
2	0.6147	4.1798	2	**0.4667**	9.8349	10	0.2594	5.6511
3	**0.4711**	3.7372	3	0.8344	9.4243	11	0.2444	5.3985
4	1.3308	3.4463	4	0.8349	9.0751	12	**0.2403**	5.2169
5	2.2806	3.1472	5	0.7757	8.7538	13	0.2634	5.1557

k means 相同的分区. 这个现象与空手道俱乐部网络的社团结构较为模糊的性质紧密相关.

5.4.2.2 宽吻海豚网络

宽吻海豚网络由生活在新西兰道尔福峡湾 (神奇湾) 的一个组织中的 62 只宽吻海豚之间的频繁联系所构成的网络[121, 122]. 这个网络是由 Lusseau 对这些海豚七年的现场研究而构造的, 通过对于统计上的重要且频繁的联系的观察而建立每对海豚之间的边[121]. 关于这个网络的详细介绍参见 1.5.2.



(a)　　(b)

图 5.8: 由变形 k means 算法[60]得到的空手道俱乐部网络的社团结构. (a) 给定 $N = 2$ 所得到的分区. (b) 给定 $N = 3$ 所得到的分区, 这与 SAVI 算法得到的分区相同.

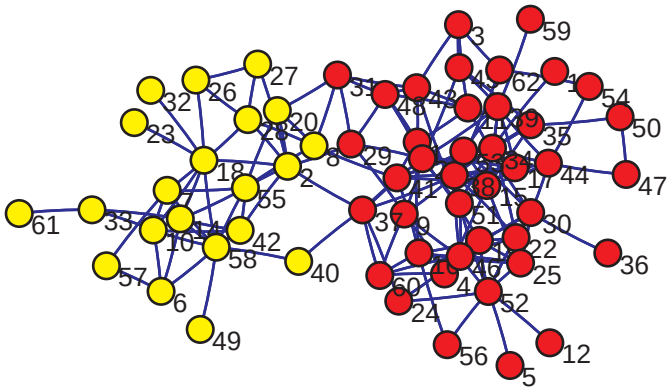


图 5.9: 图中红色和黄色的节点对应于 SAVI 所得到宽吻海豚网络的分划. SAVI 算法所得的社团结构与这个海豚组织的一个已知分割一致[121, 122].

当仅应用变形 k means 算法[60]时, 有效性指标 V_{net} 和目标函数 J 随社团数目 N 的变化如图 5.7(b) 和表 5.2 所示. 可以看出全局最优社团结构于 $N = 2$ 处达到. 当设置参数 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 20$ 从而执行 SAVI 算法时, 也得到 $N = 2$ 和相应的 $V_{\text{net}} = 0.4667$. 分区结果如图 5.9 所示. 根据 SAVI 的结果, 网络看起来分裂成两个较大的社团, 分别由红色和黄色表示, 这对应于这个海豚组织的一个已知的分割[122]. 这表明有效性指标可以有效反映内在社团结构特征.

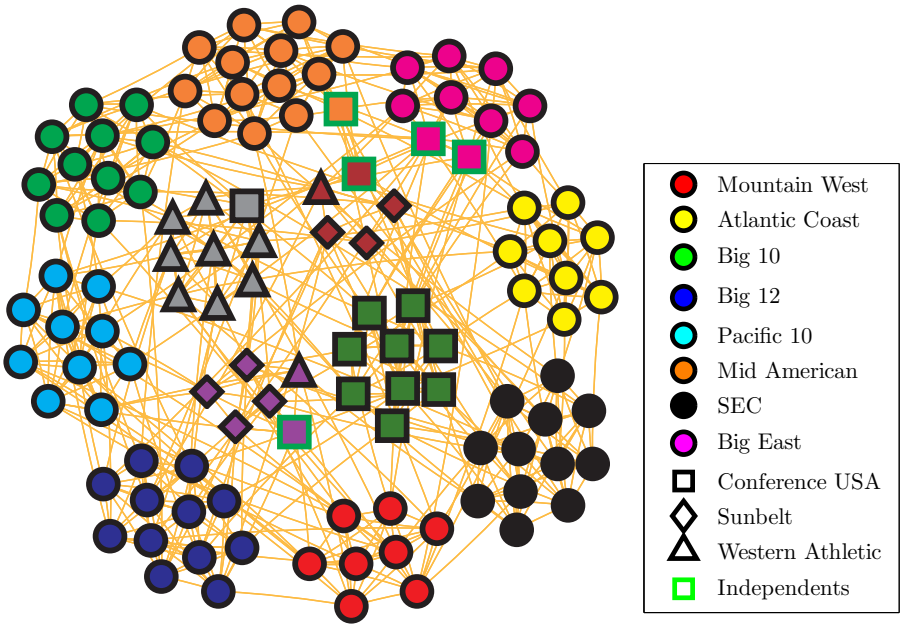


图 5.10: 由 SAVI 算法得到的美国足球队网络的社团结构. 网络中节点表示球队, 边表示球队之间的比赛. 12 个真实联盟由右边图例中列出的不同符号表示. SAVI 算法确定出网络中几乎所有的社团, 并用不同的颜色来表示.

5.4.2.3 美国足球队网络

这个网络表示美国大学生足球联联赛 2000年 第一季度的比赛日程[77]. 网络中的节点表示 115 个由学校名字命名的足球队, 连接两个节点的边表示他们之间的规则季度赛. 这个网络包含了一个已知的社团结构: 这些足球队被分成一些联盟, 每个联盟包含 8 到 12 个足球队, 同一个联盟中的球队之间的比赛比不同联盟球队之间的比赛要频繁. 关于这个网络的详细介绍参见 1.5.2.

当仅应用变形 k means 算法[60]时, 有效性指标 V_{net} 和目标函数 J 随社团数目 N 的变化如图 5.7(c) 和表 5.2 所示. 全局最优社团结构于 $N = 12$ 处达到. 当通过设置参数 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 20$ 来运行 SAVI 算法时, 也得到 $N = 12$ 和相应的 $V_{\text{net}} = 0.2403$. 分划结果如图 5.9 所示. 分划结果如图 5.10 所示. 根据 SAVI 算法所得到的结果, 方法高度准确地确定出了社团结构. 几乎所有的足球队都被正确分区在与他们所属联盟一直的社团中. Independents 联盟中的球队

(绿边方形) 看起来不属于任何一个社团, 但是他们趋向于与他们最为紧密联系的联盟分区在一起. Sunbelt 联盟 (菱形) 分裂成两个社团, 每一个社团各与 Western Athletic 联盟 (三角形) 中连接松弛的一个球队分区在一起. Conference USA 联盟 (黑边方形) 中仅有一个球队 Texas Christian, 与 Western Athletic 联盟的绝大多数球队分区在一起. 所有其它的社团 (彩色圆形) 与已知的社团结构一致, 这表明 SAVI 算法性能极佳.

5.5　小结

本章中作者提出了一个新的有效性指标函数来衡量网络社团结构的优良程度, 它包含了社团内部紧密程度和社团之间分离程度两个因素. 所构造的算法 SAVI, 即结合变形 k-means 迭代的极小化有效性指标的模拟退火, 成功地应用于几个具有代表性的网络. 关于具有已知社团结构的人工生成网络的实验展示了非常满意的结果, 即 SAVI 算法可以高效率和高精度地确定网络的社团. 它可以用随机初始分区, 经过冷却过程, 最终得到正确的社团结构. 在没有任何关于社团结构的先验信息的情况下, 最优社团数目可以被自动确定. 所提出有效性指标与针对于网路社团结构所提出的模量 (modularity) 函数[144]相比具有竞争力. 此外, 对于空手道俱乐部网络, 宽吻海豚网络以及美国足球队网络这三个真实世界网络的成功应用巩固了 SAVI 算法的有效性.

第六章　基于模量和模糊模量的自动模型选择

本章将继续在随机游动的框架下, 利用模量 (modularity) 以及新提出模糊模量 (fuzzy modularity) 这一扩展形式, 分别来实现复杂网络社团结构的确定性分区和概率性分区的自动模型选择. 其理论框架, 算法的构造以及数值试验将分别在 6.1 和 6.2 中详细阐述.

本章的内容主要参考 [119] 和 [117].

6.1　基于模量的确定性分区的自动模型选择

在 Zhou 的工作[213] 中提出了每对节点的相异性指标, 它可以度量网络节点之间的接近程度和表示两个节点属于同一个社团的程度, 其基本思想是将网络与随机游动 Markov 动力学[120]联系起来. 这就促使作者用这个度量下的 k-means 方法[86]求解网络分区问题. 另一个工作是 Lafon 和 Lee 的扩散映射[107], 它也紧随随机游动 Markov 动力学, 但它在此框架下引入了节点空间的扩散距离并定义了几何形心. 这个相似度反映了扩散过程中节点的连接程度. 最终的在这个距离下的极小化问题也可以用 k-means 方法求解. 这在第二章 2.4 中有过介绍.

如第五章[118]所指出的那样, 传统的 k-means 算法族是建立在对一个分区数目已知的给定的目标函数的最优化的基础之上的[86]. 故人们在需要确定网络最优分区的社团的数目时遇到了 k-means 的目标函数随社团数目增加而减少的的困难. 为克服这个弱点, 除了使用第五章[118]所提出的有效性指标, 还可以选择广泛使用的模量 (modularity) 的概念[144]作为网络分区的有效度量, 越大的模量值表明越强的社团结构. 然后将模拟退火方法[83, 103, 133]用来寻找模量的最大值. 冷却过程结合了基于上述网络中两种度量的 k-means 迭代而实现. 本节中结合 k-means 的模拟退火与第五章中算法 5.1 的过程不同, 它在给定温度时的每次试只执行 k-means

的一步迭代而不是整个的为收敛为止的 k-means; 本节的算法也与单纯退火过程[43]不同, 因为迭代的过程加速了极大化模量的趋势. 本节的算法可以找到已有方法更优的模量值[39, 57, 140–144]. 算法的另一个有点是不仅可以确定出社团结构以及社团数目, 并且可以给出每个社团的中心节点. 社团的中心节点可以传达这样的信息: 它在同一组的成员中具有多重要的地位; 这是因为人们有时关注社团中的通讯特点, 并假定在结构中心和社团进程的影响之间存在某种关系[72].

本节构造网络分区的算法: 结合基于相异性指标的 k-means 的模拟退火 (SADI) 和结合基于扩散距离的 k-means 的模拟退火 (SADD). 算法测试于两个人工生成的网络, 包括 128 个节点的 ad hoc 网络和 Gauss 混合模型生成的样本网络. 两种方法都可以用合理的计算量有效实现并得到精确的分区结果. 此外, 算法还成功地应用于一些真实世界中的网络, 包括空手道俱乐部网络, 宽吻海豚网络, 政治书籍网络, 小说《悲惨世界》的人物关系网络以及美国足球队网络.

本节余下部分内容如下. 在 6.1.1 中简单介绍关于网络节点相似程度的两种度量, 包括相异性指标和扩散距离. 回顾 6.1.2 中模量的定义及其意义之后, 在 6.1.3 中提出所构造的算法和相应的策略. 在 6.1.4 中, 将方法应用于上述的具有代表性的算例, 并比较数值结果和算法的性能.

本节内容主要参考 [119].

6.1.1 网络中节点之间接近程度的度量

6.1.1.1 相异性指标与其相应的中心

在 [213, 214] 中 Zhou 定义了节点对之间的相异性指标, 它可以度量网络中节点的接近程度. 设 $G = (S, E)$ 为 n 个节点和 m 条边的网络, 其中 S 为节点集合, $E = \{e(x, y)\}_{x,y\in S}$ 为权重矩阵, 且 $e(x, y)$ 为连接节点 x 和 y 的边上的权重. 于是可以通过如下的随机矩阵 $p = (p(x, y))$ 将这个网络与离散时间的马氏链联系起来

$$p(x, y) = \frac{e(x, y)}{d(x)}, \qquad d(x) = \sum_{z\in S} e(x, z), \tag{6.1}$$

其中 $d(x)$ 为节点 x 的度[36, 120]. 由前面几章的内容知, 这个马氏链具有如下形式的平稳分布 μ

$$\mu(x) = \frac{d(x)}{\sum_{z\in S} d(z)}, \tag{6.2}$$

并满足细致平衡条件 $\mu(x)p(x,y) = \mu(y)p(y,x)$.

假设随机游动者位于节点 x, 则平均首达时 $t(x,y)$ 为它首次到达节点 y 之前需要经过的平均步数, 有如下形式

$$t(x,y) = p(x,y) + \sum_{j=1}^{+\infty}(j+1)\cdot \sum_{z_1,\cdots,z_j\neq y} p(x,z_1)p(z_1,z_2)\cdots p(z_j,y). \tag{6.3}$$

易知 $t(x,y)$ 是下述线性方程的解

$$[I - B(y)]\begin{pmatrix} t(1,y) \\ \vdots \\ t(n,y) \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tag{6.4}$$

其中 $B(y)$ 是将矩阵 p 的第 y 列替换成一列零而形成的矩阵[215]. 节点 x 和 y 关于网络的将来状态中的差异可以被定量地度量. 相异性指标由下面的表达式定义

$$\Lambda(x,y) = \frac{1}{n-2}\left[\sum_{z\in S, z\neq x,y}\Big(t(x,z) - t(y,z)\Big)^2\right]^{\frac{1}{2}}. \tag{6.5}$$

选取 S 的分划 $S = \bigcup_{k=1}^{N} S_k$, 且 $S_k \bigcap S_l = \emptyset$ 若 $k \neq l$. 如果两个节点 x 和 y 属于相同的社团, 则平均首达时 $t(x,z)$ 将与 $t(y,z)$ 非常相似, 因此网络的两个将来状态将非常相似. 因此, 如果 x 和 y 属于相同的社团, 则 $\Lambda(x,y)$ 将会很小; 若它们属于不同的社团, 则 $\Lambda(x,y)$ 将会很大. 社团 S_k 的中心 $m^I(S_k)$ 可定义如下

$$m^I(S_k) = \arg\min_{x\in S_k} \frac{1}{|S_k|}\sum_{y\in S_k, y\neq x}\Lambda(x,y), \quad k = 1,\ldots,N, \tag{6.6}$$

其中 $|S_k|$ 是社团 S_k 中节点数目. 这是一个很直观的想法, 即选择与之所属社团内其它节点的平均相异性指标最小的节点作为这个社团的中心.

6.1.1.2 扩散距离与扩散中心

在 [107] 中 Lafon 和 Lee 的主要思想是定义关于反映给定网络中节点连接程度的显式度量的坐标系统, 其构造也是基于网络上的随机游动. 节点 x 和 y 之间的扩散距离 $D(x,y)$ 定义为加权 L^2 距离

$$D^2(x,y)=\sum_{z\in S}\frac{(p(x,z)-p(y,z))^2}{\mu(z)}, \tag{6.7}$$

其中权重 $1/\mu(x)$ 惩罚低密度区域多于高密度区域的差异. 这个图中点的接近程度的概率反映了在扩散过程中关于数据点连接程度的集合的内在几何性. 两点间的扩散距离将较小如果它们由图中许多路径相连接. 转移矩阵 P 具有特征值 $\lambda_0\geq|\lambda_1|\geq\cdots\geq|\lambda_{n-1}|\geq 0$, 以及右, 左特征向量

$$P\varphi_j=\lambda_j\varphi_j,\quad \psi_j^TP=\lambda_j\psi_j^T,\quad j=0,1,\cdots,n-1. \tag{6.8}$$

注意到 $\psi_0=\mu$ 且 $\varphi_0\equiv 1$. 此外还有 $\psi_j(x)=\varphi_j(x)\mu(x)$. 令 q 为使得 $|\lambda_j|>\delta|\lambda_1|$ 的最大的指标 j, 并且如果引入扩散映射

$$\Phi: x\longmapsto\begin{pmatrix}\lambda_1\varphi_1(x)\\ \lambda_2\varphi_2(x)\\ \vdots\\ \lambda_q\varphi_q(x)\end{pmatrix}, \tag{6.9}$$

则精度为 δ 的扩散距离 $D(x,y)$ 可用前 q 个非平凡特征向量和特征值近似地表示

$$D^2(x,y)\simeq\sum_{j=1}^{q}\lambda_j^2\Big(\varphi_j(x)-\varphi_j(y)\Big)^2=\|\Phi(x)-\Phi(y)\|^2. \tag{6.10}$$

社团 S_k 的几何形心 $c(S_k)$ 定义为

$$c(S_k)=\sum_{x\in S_k}\frac{\mu(x)}{\hat{\mu}(S_k)}\Phi(x),\quad k=1,\ldots,N, \tag{6.11}$$

其中 $\hat{\mu}(S_k)=\sum_{x\in S_k}\mu(x)$. 这里的 $c(S_k)$ 可能并不属于集合$\{\Phi(x)\}_{x\in S}$. 为了得到属于原始集合 S 的代表社团的中心节点, 引入如下的扩散中心 $m^D(S_k)$

$$m^D(S_k)=\arg\min_{x\in S_k}\|\Phi(x)-c(S_k)\|^2,\quad k=1,\ldots,N. \tag{6.12}$$

6.1.2　模量的定义

近年来, 由 Newman 提出的模量 (modularity) 的概念[144]作为社团结构优良性的度量被广泛地使用. 将他指出将网络分成社团的好的分区不是仅仅要求连接社团之间的边数较少, 而是要求连接社团之间的边数少于期望的值. 这些考虑得到了模量 Q 的定义如下

$$Q = \text{社团内部的边数} - \text{这些边数的期望}. \tag{6.13}$$

这是一个关于将网络分成社团的特定分区的函数, 其值越大表示社团结构越强.

模量的定义包括一个真实网络与等价的随机模型网络的社团内部边数比较, 这个等价的随机模型网络中的边是不考虑社团结构而置入的[141]. 空模型 (null model) 具有与原始网络相同的 n 个节点. 每对节点 x 和 y 之间置入一条边的概率 $p^E(x,y)$ 是指定的. 更精确地说, $p^E(x,y)$是 x 和 y 之间边数的期望值, 这个定义允许一对节点之间存在多于一条边的可能性, 这种情况发生在某些类型的网络中. 对于给定的分划 $\{S_k\}_{k=1}^N$, 模量可以写成

$$Q = \frac{1}{2m_e}\sum_{k=1}^{N}\sum_{x,y\in S_k}\Big(e(x,y) - p^E(x,y)\Big), \quad p^E(x,y) = \frac{d(x)d(y)}{2m_e}, \tag{6.14}$$

其中 m_e 是边上权重的和 $\sum_{x,y\in S} e(x,y)/2$, 如果网络是无权图则 $m_e = m$. 这个模型与物理中广泛研究的构造模型紧密相关[141]. 一些已有的方法[39, 57, 140–144]通过对可能的分割来最优化模量的方式, 以寻找将网络分成社团的好的分区, 这在实践中被证明是非常有效的.

尽管模量最大化是流行的方法, 但是当将其应用于未知社团结构的真实世界网络中时, 对于输出结果的质量和重要性依然有很多未知. 模量并不是社团结构的一个精确的指示器, 这基于下面两个原因: 第一, 这个量具有一个著名的分辨率极限现象, 使它对于大网络的应用出现问题, 例如小模块仍然无法发现[71]. 模量最优化可能对于识别小于一个尺度的模块出现失败, 这个尺度依赖于网络边上权重之和 m_e 和模块间的互相连接的度, 甚至在模块不含糊地定义的情形也是如此. 如果模块内部的连接数与 $\sqrt{2m_e}$ 同阶或更小, 则这个模块隐藏其良定义子结构的概率是最高的. 因此, 事先不可能去知道通过模量最优化得到的一个模块是否真正的

是一个单个模块或者是一个由较小模块构成的社团. 从而这个结果引入了使用模量去检测社团结构中的一些警告. 第二, 模块的地形是“像玻璃的”, 即大量分区具有与最大模量值非常接近的模量值, 但是结构是不同的, 故最大化的含义也有待讨论[80]. 模量展示极端退化: 它极具特色地容许指数量级的不同的高分解, 并缺乏清晰的全局极大值. 退化的解可从根本上与许多但非全部的分区性质不一致, 例如最大模块的组成和模块规模的分布. 这些结果意味着任何模量最优化过程的输出都应进行谨慎的解释.

6.1.3 算法的构造

正如 5.3 中所介绍的, 模拟退火算法最初是由模拟退火固体的物理过程而产生的[133], 并广泛地应用于最优化问题[103].

设 $E=-Q$, $E^{(n)}$ 和 $E^{(n+1)}$ 分别表示当前能量和新能量. $E^{(n+1)}$ 总被接受如果它满足 $E^{(n+1)}<E^{(n)}$, 但是如果 $E^{(n+1)}>E^{(n)}$ 则新能量级仅以概率 $\exp(-\frac{1}{T}\triangle E^{(n)})$ 被接受, 其中 $\triangle E^{(n)}=E^{(n+1)}-E^{(n)}$ 为能量差, T 为当前温度. 较差的解基于解的质量的变化而被接受, 这使搜索避免了被困于局部极小值. 于是温度逐渐降低, 重复退火过程直到没有更新的改进出现或者达到某个终止准则. 初始状态为随机生成的 N 个社团, 其中 N 为区间 $[N_{\min},N_{\max}]$ 中的整数, 在下面的计算中选取 $N_{\min}=2, N_{\max}=n/3$. 初始温度设置为一个较高的温度 $T_{\max}$. 当前状态的下一个状态由随机选择下述的提议之一而产生, 然后计算这个新状态的能量. 这个新的状态被接受如果满足接受条件. 这个过程将在给定的温度上重复 R 次. 设置冷却速率 $0<\alpha<1$ 来降低当前温度直到温度下界 $T_{\min}$ 被达到. 整个的结合基于相异性指标的 k-means 迭代算法的模拟退火过程概括如下.

算法 6.1 (Simulated Annealing with the Dissimilarity-Index-based k-means algorithm — SADI)

(1) 设置参数 $T_{\max}$, $T_{\min}$, α 和 R. 在区间 $[N_{\min},N_{\max}]$ 中随机选取 N, 并随机初始化分区 $\{S_k^{(0)}\}_{k=1}^N$. 设当前温度 $T=T_{\max}$.

(2) 根据 (6.6) 计算中心 $\{m^I(S_k^{(0)})\}_{k=1}^N$, 然后根据 (6.14) 计算初始能量 $E^{(0)}$; 令 $n^*=0$.

(3) 对于 $n = 0, 1, \ldots, R$, 做如下迭代:

(3.1) 根据下述提议产生一组新的中心 $\{m^I(S_k^{(n)})\}_{k=1}^{N'}$, 并令 $N = N'$.

(3.2) 分别根据

$$S_k^{(n+1)} = \left\{x : k = \arg\min_l \Lambda(x, m^I(S_l^{(n)}))\right\}, \quad k = 1, \ldots, N, \tag{6.15}$$

和 (6.6) 来更新分区 $\{S_k^{(n+1)}\}_{k=1}^N$ 和中心 $\{m^I(S_k^{(n+1)})\}_{k=1}^N$, 然后根据 (6.14) 计算新能量 $E^{(n+1)}$.

(3.3) 接受或拒绝新状态. 如果 $E^{(n+1)} < E^{(n)}$, 或者 $E^{(n+1)} > E^{(n)}$ 且 $u \sim \mathcal{U}[0,1]$, $u < \exp\{-\frac{1}{T}\triangle E^{(n)}\}$, 则接受新状态, 令 $n = n + 1$; 否则拒绝.

(3.4) 更新最优状态, 即如果 $E^{(n)} < E^{(n^*)}$, 则令 $n^* = n$.

(4) 降温 $T = \alpha \cdot T$. 如果 $T < T_{\min}$, 执行 (5); 否则令 $n = n^*$, 并重复 (3).

(5) 输出整个过程的最优分区 $\{S_k^{(n^*)}\}_{k=1}^N$ 和最小能量 $E^{(n^*)}$.

通过将上述算法 6.1 中的 (6.6) 和 (6.15) 分别替换为 (6.12) 和

$$S_k^{(n+1)} = \left\{x : k = \arg\min_l \|\Phi(x) - c(S_l^{(n)})\|^2\right\}, \quad k = 1, \ldots, N, \tag{6.16}$$

则得到结合基于扩散距离的 k-means 迭代算法的模拟退火算法 (Simulated Annealing with Diffusion-Distance-based k-means algorithm — SADD).

对于步骤 (3.1) 中产生一组新中心的过程的提议由三个函数组成, 分别是保持一个当前社团, 删除一个当前社团和分裂一个当前社团. 在每次迭代时, 随机选择三个函数中的一个, 并且社团强度[158]

$$M(S_k) = \sum_{x \in S_k} (d^{\text{in}}(x) - d^{\text{out}}(x)), \quad k = 1, \ldots, N, \tag{6.17}$$

被用来选择一个社团, 其中 $d^{\text{in}}(x) = \sum_{z \in S_k} e(x, z)$ 且 $d^{\text{out}}(x) = \sum_{z \notin S_k} e(x, z)$. 这三个函数描述如下.

(a) 保持一个社团. 保持这组中心.

(b) 删除一个社团. 由 (6.17) 具有最小社团强度的社团 S_d 被选择, 其中心将被从中心集合中删除.

(c) **分裂一个社团**. 具有最小平均社团强度的社团

$$S_s = \arg\min_{S_l} \frac{M(S_l)}{|S_l|} \tag{6.18}$$

被选择. 对于 SADI, 新的中心由

$$m^I(S_{N+1}) = \arg \min_{x \in S_s, x \neq m(S_s)} \Lambda(x, m^I(S_s)), \tag{6.19}$$

得到; 对于 SADD, 当前的几何形心 $c(S_s)$ 由下面两个新的几何形心代替

$$\begin{aligned} c(S_{N+1}) &= c(S_s) - |c(S_s) - m(S_s)|, \\ c(S_s) &= c(S_s) + |c(S_s) - m(S_s)|. \end{aligned} \tag{6.20}$$

算法的迭代步数依赖于初始和终止温度, 冷却速率和在给定温度下的重复次数, 约为 $R \log_\alpha \frac{T_{\min}}{T_{\max}}$. 对于 SADI, 由于真正需要知道的是平均首达时的差, 即 $t(x,z) - t(y,z)$, 故可以以计算量 $O(n^3)$ 计算所有不同的差. 对于 SADD, 每次迭代计算 $\{c(S_k)\}_{k=1}^N$ 和 $\{S_k\}_{k=1}^N$ 的花费均为 $O(qNn)$.

本节的算法的优点是它们克服了传统聚类方法的弱点, 即最优预测误差随分区数目增加而减少, 例如由梯度方法构造的 k-means[127]和 fuzzy c-means 算法[19, 58]. 本节方法的冷却过程可以有效且自动地确定社团数目 N 而不再是将它固定作为已知模型参数, 并且初始分区 $\{S_k^{(0)}\}_{k=1}^N$ 可以随机选取. 这个问题也可通过另一种方式求解, 即对于所有可能的 N, 分别遍历两种 k-means 算法. 但是这将花费巨大, 因为对于每个固定的 N, k-means 过程需运行 1000 到 5000 次试验来避免陷入局部极小值. 另一方面, 与对于所有可能的 N 遍历两种 k-means 算法相比, 本节的算法有时会得到具有更大模量值的更优分区 (如图 6.4 所示). 因此这里的模拟退火策略可以避免无效的重复并得到高效率与高精度.

6.1.4 数值试验

本节中, 将算法测试于具有已知社团结构的人工生成的网络, 包括 128 个节点的 ad hoc 网络和 Gauss 混合模型生成的样本网络. 随后, 算法成功地应用于真实世界中的网络, 包括空手道俱乐部网络, 神奇湾宽吻海豚网络, 政治书籍网络, 《悲惨世界》人物关系网络和美国足球队网络.

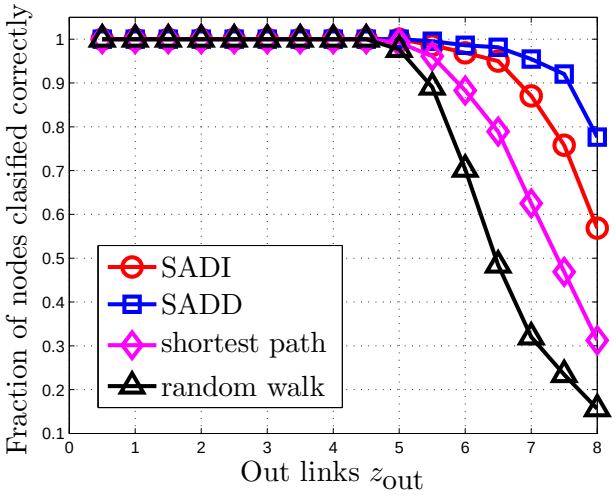


图 6.1: 由本节中的算法和 [144] 中方法所得到的 ad hoc 网络的节点识别的正确率随 z_{out} 的变化. 从图中可见 SADI 和 SADD 的性能优于最短路径方法和随机游动方法[144].

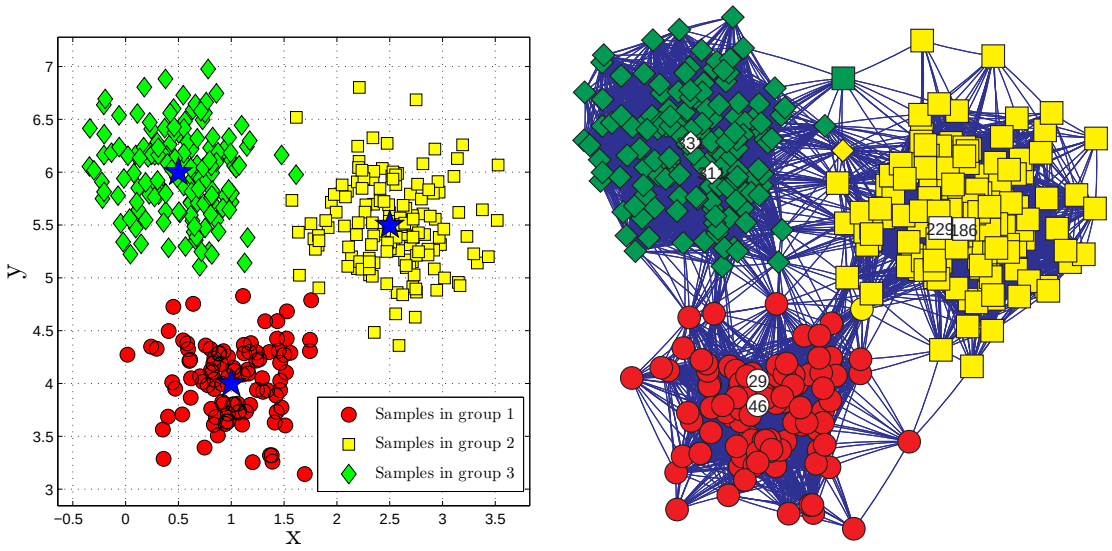
6.1.4.1　人工生成的网络

128 个节点的 ad hoc 网络. 本节的第一个算例是 128 个节点的 ad hoc 网络. 这类网络具有已知的社团结构, 构造如 1.5.1 所述. 通常定义 z_{out} 为某个节点与属于其它社团节点之间连接的平均数, z_{out} 越大, 社团就变得越模糊 (diffuse). 这里将 $z_{\text{out}} = 96p_{\text{out}}$ 从 0.5 变化到 8, 并观察节点识别的正确率. 在这个模型计算中参数设置为 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 20$. 分区结果的节点识别正确率如图 6.1 所示. 与 [144] 中的两种方法进行比较, 可以看出 SADI 和 SADD 的性能显著优于先前的两种方法, 特别是当 z_{out} 很大时的社团结果较为模糊的情形.

Gauss 混合模型生成的样本网络. 本节的第二个算例是 Gauss 混合模型生成的样本网络. 这个模型与 Penrose 提出的随机几何图的概念[153]有关, 只是这里选取 Gauss 混合模型, 而不再是 [153] 中的均匀分布. 网络的构造过程和意义如 1.5.1 中所描述. 选取 $n = 400$ 和 $K = 3$, 然后根据如下的均值和协方差矩阵生成样本点

$$\boldsymbol{\mu}_1 = (1.0, 4.0)^T, \boldsymbol{\mu}_2 = (2.5, 5.5)^T, \boldsymbol{\mu}_3 = (0.5, 6.0)^T, \tag{6.21a}$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = 0.15\boldsymbol{I} = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{6.21b}$$



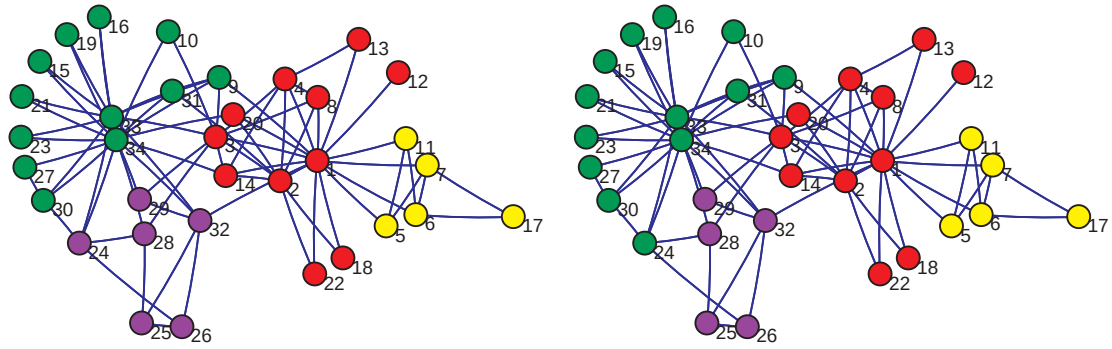
(a) (b)

图 6.2: (a) 由 3-Gauss 混合模型生成的 400 个样本点. 星形符号表示每个 Gauss 分量的中心; 圆形, 方形和菱形符号分别表示三个不同分量中的样本点. (b) 算法关于由 (a) 中样本点根据 $dist = 0.8$ 生成的网络的分区结果. 不同的社团由不同的颜色表示. 中心 $m^I = \{46, 186, 331\}$ 和 $m^D = \{29, 229, 311\}$ 由白色表示.

这里选择节点 1 : 100 在第 1 组, 节点 101 : 250 在第 2 组, 节点 251 : 400 在第 3 组. 根据这个选择, 近似地有 $q_1 = 100/400$, $q_2 = q_3 = 150/400$. 这个实验中取阀值 $dist = 0.8$. 样本点如图 6.2(a) 所示. 这里参数取 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 15$ 来实现所提出的算法. 分区结果如图 6.2(a) 所示. 两种算法 SADI 和 SADD 均得到同样的 $N = 3$ 和 $Q = 0.6241$, 中心分别为 $m^I = \{46, 186, 331\}$ 和 $m^D = \{29, 229, 311\}$. 对于 SADI 和 SADD, 所得到的中心与均值 $\boldsymbol{\mu}$ 之间的平均 L^2 误差

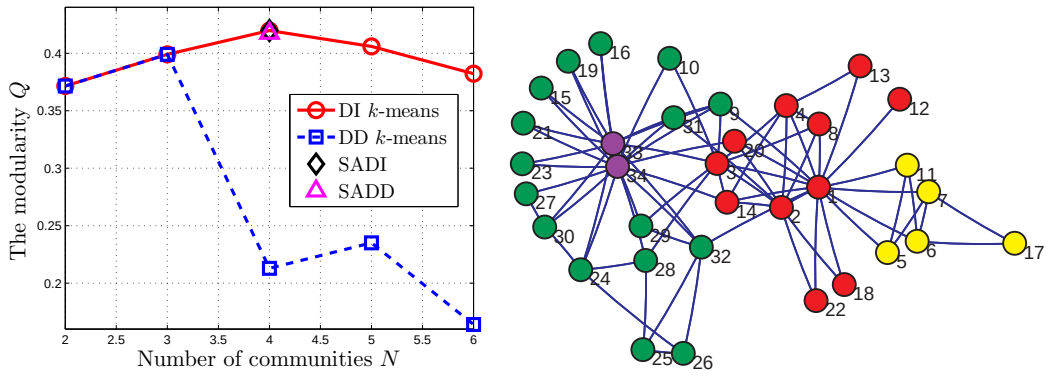
$$\frac{1}{N}\sum_{k=1}^{N} \|\boldsymbol{x}_{m(S_k)} - \boldsymbol{\mu}_k\|_2 \tag{6.22}$$

分别为 0.0804 和 0.2211. 结果是非常合理的, 这表明本节所提出的算法可以成功计算较大型的网络.



(a)　　(b)

图 6.3: 用本节中的方法得到的空手道俱乐部网络的社团结构. 两种方法产生相同的分区除了节点 24. (a) SADI 的分区结果. (b) SADD 的分区结果.



(a)　　(b)

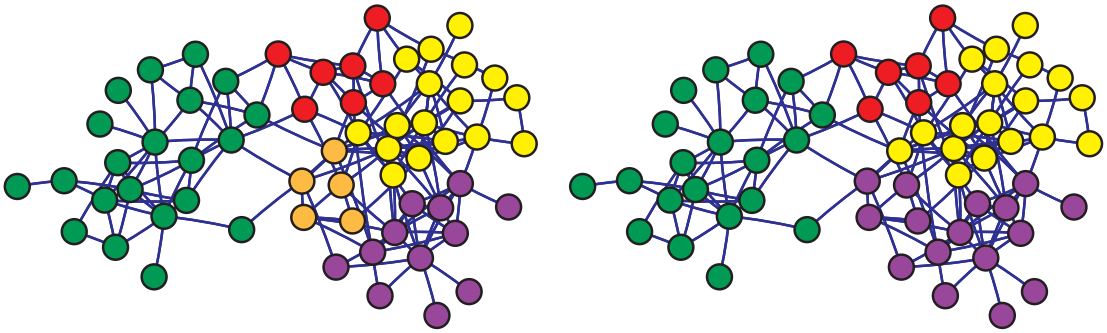
图 6.4: (a) 由基于相异性指标的 k-means, 基于扩散距离的 k-means, SADI 和 SADD 所得到的最大化的模量值. 图中清楚地显示出 SADI 达到最大模量值 $Q = 0.4198$, 这与其相应的 k-means 当 $N = 4$ 时的结果相同. 而 SADD 可以达到比相应的 k-means 当 $N = 4$ 时更大的模量值 $Q = 0.4174$. (b) 基于扩散距离的 k-means 算法当 $N = 4$ 时得到的社团结构. 基于相异性指标的 k-means 算法得到的结果与图 6.3(a) 相同.

6.1.4.2　真实世界中的网络

空手道俱乐部网络. 这个网络是由 Wayne Zachary 在观察一所美国大学空手道俱乐部成员之间的社交而构建的[210], 具体介绍见 1.5.2. 本节提出的方法得到的

表 6.1: 两种算法对于空手道俱乐部网络, 宽吻海豚网络以及政治书籍网络得到的数值结果.

	karate club network		dolphins network		political books network	
	N	Q	N	Q	N	Q
SADI	4	0.4198	5	0.5176	4	0.5260
SADD	4	0.4174	4	0.5235	4	0.5266

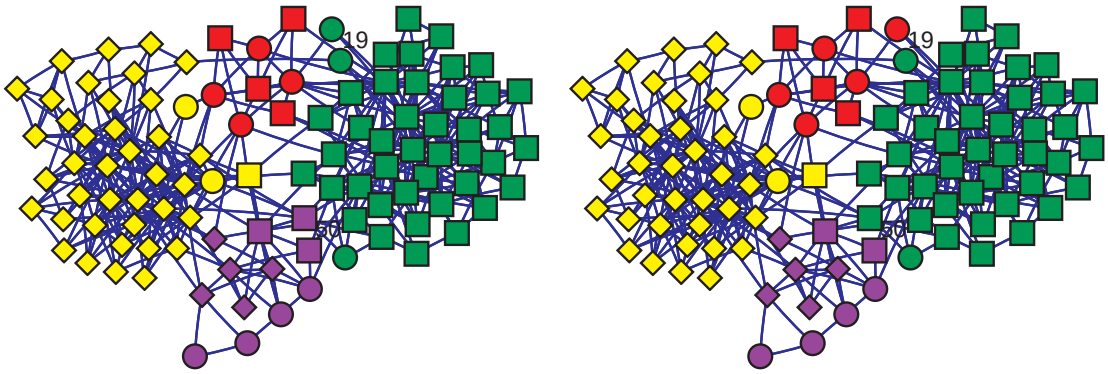


(a) (b)

图 6.5: 用本节中的方法得到的宽吻海豚网络的社团结构. (a) SADI 的分区结果. (b) SADD 的分区结果.

结果如表 6.1 和图 6.3 所示. 两种方法均得到 $N = 4$, 但 SADI 达到更高的 Q 值, 这胜过了多数现有方法[39, 57, 140–144]. 另一方面, 模拟退火的方法可以得到比对所有可能的 N 遍历两种 k-means 更优的分区结果, 这如图 6.4 所示, SADI 达到最大模量值 $Q = 0.4198$, 这与其相应的 k-means 当 $N = 4$ 时的结果相同. 而 SADD 可以达到比其相应的 k-means 当 $N = 4$ 时更大的模量值 $Q = 0.4174$.

宽吻海豚网络. 宽吻海豚网络由生活在新西兰道尔福峡湾 (神奇湾) 的一个组织中的 62 只宽吻海豚之间的频繁联系所构成的网络[121, 122]. 关于这个网络的详细介绍参见 1.5.2. 本节所提出的算法得到的结果如表 6.1 和图 6.5 所示. 根据结果, 网络看起来分裂成绿色部分和较大部分这两个大的社团, 其中较大部分继续分裂成几个更小的社团, 分别用不同颜色表示. 分裂成两个社团对应于海豚组织的一个已知分区[122]. 而较大的那部分中的几个子社团也对应于海豚的真实的分区, 黄色部分几乎全部由雌性组成, 而其它部分几乎全部都是雄性.



(a)　　(b)

图 6.6: 用本节中的方法得到的政治书籍网络网络的社团结构. 两种方法产生几乎相同的分区除了节点 19 和 50. (a) SADI 的分区结果. (b) SADD 的分区结果.

政治书籍网络. 这个网络是由 V. Krebs 编制的关于美国政治的书籍的网络, 详细介绍参见 1.5.2. 网络中的节点表示从在线书商 Amazon.com 上购买的最近的 105 本关于美国政治的书籍, 连接书籍对的边表示这两本书频繁地由相同顾客购买. 书籍的分类是按照它们所陈述的明显的政治立场, 自由党或者保守党, 除了一小部分书籍是明确的两党派或中立者, 或者没有明确的从属关系[142]. 如图 6.6 所示, 节点的给出是根据为它们属于保守的 (方形) 还是自由的 (菱形), 除此之外还有小部分书籍是中立的 (圆形). 计算结果如表 6.1 和图 6.6 所示. 可以发现四个社团, 分别用不同的颜色表示. 看起来这些社团中的其中一个几乎全部由属于自由的书籍组成, 一个乎全部由属于保守的书籍组成. 多数属于中立的书籍落入余下的两个社团中. 因此这些书似乎形成了与政治观点密切相关的联合购买的社团, 本节的算法能够从原始数据网络中提取有意义的结果.

小说《悲惨世界》人物关系网络. 这是维克多·雨果的关于法国恢复后的犯罪与救赎的长篇巨著《悲惨世界》中的主要人物之间的相互关系的网络, 它是由 Knuth 根据戏剧的场次中出现的人物列表而构造的[106]. 网络中的节点代表人物, 两个节点之间的边代表与相关人物共同出现在一场或多场中, 详细介绍参见 1.5.2. 由算法 SADD 得到结果的最优社团结构具有模量 $Q = 0.5654$, 并给出如图 6.7 所示的 6 个社团, 这达到了比 [144] 中方法更大的模量值. 社团清楚地反映了书中次

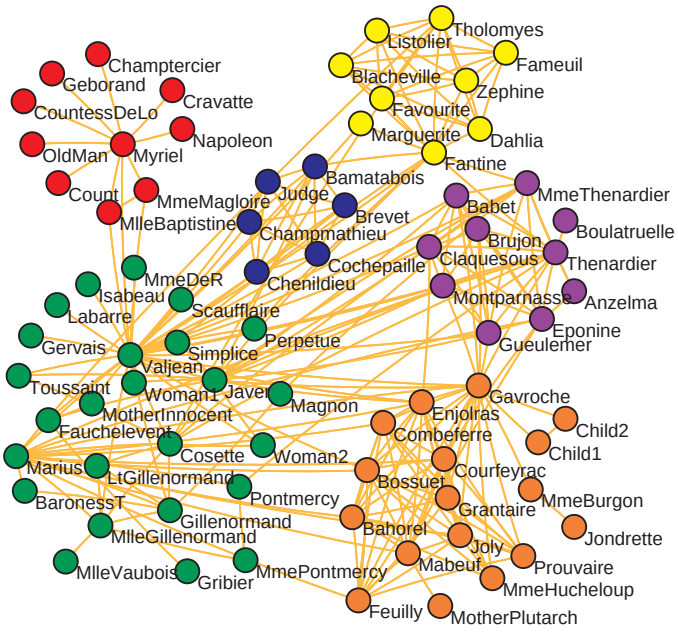


图 6.7: 雨果的小说《悲惨世界》主要人物之间联系的网络的社团结构. 利用 SADD 方法得到的最大模量 $Q = 0.5654$, 对应于不同颜色表示的 6 个社团.

要情节的结构: 主角 Jean Valjean 和他的复仇者, 警务人员 Javert 均是网络的重要成员, 并形成由他们的拥护者组成的社团的中心. 其它集中在 Marius, Cosette, Fantine 和主教 Myrial 的次要情节也在图中表现出来.

美国足球队网络. 本节研究的最后一个网络是美国大学生足球联联赛 2000年第一季度的比赛日程[77]. 关于这个网络的详细介绍参见 1.5.2. 利用算法 SADI 得到的网络的最优社团结构具有很强的模量 $Q = 0.6032$, 并给出如图 6.8 所示的 11 个社团, 这个分区结果优于多数现有方法[77, 213]. 根据所得到的结果, 算法 SADI 高度准确地确定出了社团结构, 几乎所有的足球队都被正确地与他们所属联盟中的其它球队分在一个社团中. Independents 联盟中的球队 (绿边方形) 看起来不属于任何一个社团, 但是他们趋向于与他们最为紧密联系的联盟分区在一起. Sunbelt 联盟 (菱形) 分裂成两个社团, 其中一个社团与 Western Athletic 联盟 (三角形) 中连接松弛的一个球队分区在一起, 另一个社团与 Mountain West 分区在一起. Conference USA 联盟 (黑边方形) 中仅有一个球队 Texas Christian, 与 Western Athletic 联盟的绝大多数球队分区在一起. 所有其它的社团 (彩色圆形) 与已知的

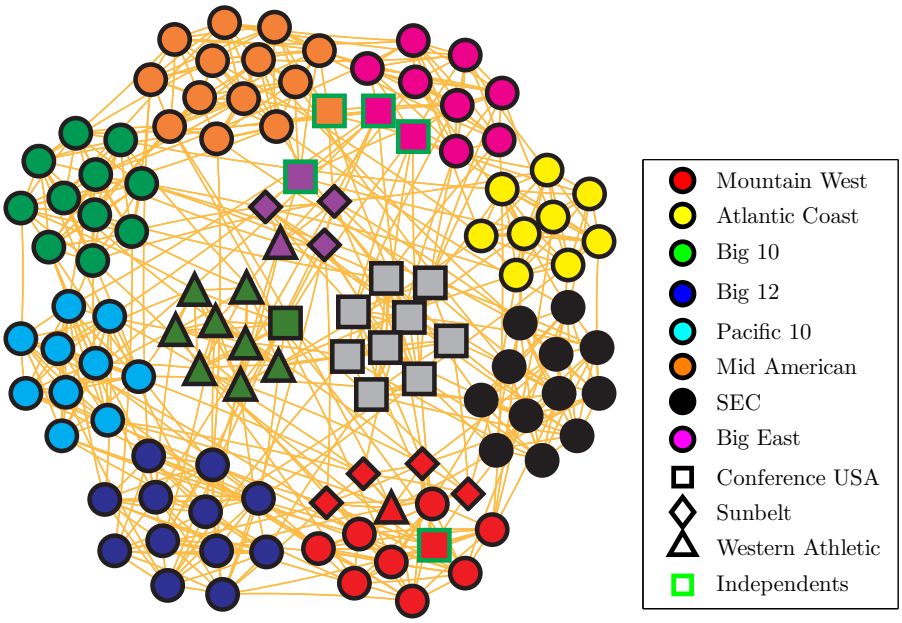


图 6.8: 美国足球队网络的社团结构. 利用 SADI 方法得到的最大模量 $Q = 0.6032$, 对应于不同颜色表示的 11 个社团.

社团结构一致.

6.1.5　小结

本节中提出了实现网络确定性分区的自动模型选择方法 SADI 和 SADD, 并成功地应用于几个具有代表性的网络. 数值结果表明它们产生相似的结果, 但是 SADD 在多数情况下具有更好的效率和精度. 两种算法均得到比多数现有方法更优的模量[39, 57, 140–144]. 这里再一次指出本节所提出的算法不仅可以确定社团结构, 还可以确定每个社团的中心节点. 最优社团数目在降温过程中可以被有效自动地确定, 而不需要任何关于社团结构的先验信息.

6.2　基于模糊模量的概率性分区的自动模型选择

在第四章[114]中提出了关于网络分区的概率性框架, 它既可以看做是统计中的 fuzzy c-means 算法[19, 58]向网络分区问题的自然扩展, 也可以看做是第三章[60]中网络最优分区的确定性框架的推广. 在传统的网络分区问题中, 每个节点在分区后

仅处于一个社团. 然而这通常是过于限制的, 因为在许多扩散网络中, 位于社团之间边界处得节点共享多余一个社团的公共性, 并起到了过渡的作用. 这就激发了概率性分区的思想来为讨论网络结构提供一种更加成熟的方式. 每个节点在概率性分区后以某一概率从属于某一社团, 而不是将节点分配到特定的社团中. 第四章[114]中的方法可成功实施, 但是需要已知社团数目作为模型参数, 和由变形 k-means 算法[60]得到的确定性分区作为特定的初始化.

为了克服这两个弱点, 在本节中提出了一个有效方法来实现概率性分区的自动模型选择. 本节将著名的衡量网络社团优良性的模量的概念[144]自然地推广到概率性的形式, 即模糊模量 (fuzzy modularity), 来量化概率性分区的质量; 并且协助实现自动模型选择, 而不再是要求用户固定社团数目. 同样地, 模拟退火方法[83, 103, 133]被用来寻找模糊模量的最大值. 冷却过程结合了基于概率性分区算法的 AIP 交替迭代[114]而实现. 这种方法避免无效重复, 具有高效率和高精度, 迭代过程加速了极大化模糊模量的趋势, 并可以得到比对所有可能的 N 遍历使用 AIP 更大的模糊模量值, 如图 6.12 所示. 总之, 这个方法不仅可以确定每个节点属于不同社团的概率, 而且可以自动确定最优的社团个数而不需要任何关于社团结构的先验信息. 此外, 模糊分区概率 ρ 的初始值可以被随机选取. 模糊社团结构包含了更多详细信息并且与之前做网络分区的方法相比具有更多预测性的功能.

本节构造算法: 极大化模糊模量的结合基于第四章[114]中 Euler-Lagrange 方程组的交替迭代 AIP 的模拟退火方法 (SAFM). 算法测试于两个个人工网络, 包括 128 个节点的 ad hoc 网络和 LFR 基准网络及其具有重叠社团的网络. 数值结果表明算法以合理的计算量有效地实现并可以达到准确的分区结果. 此外, 算法成功地应用于四个真实世界中的网络, 包括空手道俱乐部网络, 宽吻海豚网络, 政治书籍网络与圣达菲研究所科学家合作网络, 巩固了算法的有效性.

本节余下部分内容如下. 在 6.2.1 中简单回顾第四章[114]中的网络分区的概率性框架. 在 6.2.2 中将扩展模量的定义, 提出模糊模量函数. 在 6.2.3 中提出算法 SAFM 和相应的策略. 在 6.2.4 中, 将方法应用于上述的有代表性的算例, 并分析数值结果和算法的性能.

本节内容主要参考 [117].

6.2.1　网络的概率性分区

第四章[114]的基本思想是引入关于 (6.1) 中的随机矩阵 $p(x,y)$ 的一种度量

$$\|p\|_\mu^2 = \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} |p(x,y)|^2, \tag{6.23}$$

并通过极小化距离 $\|\tilde{p}-p\|_\mu$ 来寻找约化的马氏链 $\tilde{p}$.

给定 S 的分区 $S=\bigcup_{k=1}^N S_k$, 且 $S_k \bigcap S_l = \emptyset$ 若 $k\neq l$. 设 $\hat{p}_{kl}$ 是状态空间 $\mathbb{S}=\{S_1,\ldots,S_N\}$ 中从 S_k 到 S_l 的粗粒化的转移概率, 满足

$$\hat{p}_{kl}\geq 0, \quad \sum_{l=1}^N \hat{p}_{kl} = 1. \tag{6.24}$$

设 $\rho_k(x)$ 为节点 x 属于第 k 个社团 S_k 的概率, 满足条件

$$\rho_k(x)\geq 0 \quad \text{and} \quad \sum_{k=1}^N \rho_k(x) = 1, \quad \text{for all} \quad x\in S. \tag{6.25}$$

这个矩阵可以通过下述表达式自然地提升到原始状态空间 S 中的随机矩阵空间去

$$\tilde{p}(x,y) = \sum_{k,l=1}^N \rho_k(x)\hat{p}_{kl}\rho_l(y)\frac{\mu(y)}{\hat{\mu}_l}, \qquad x,y\in S, \tag{6.26}$$

其中

$$\hat{\mu}_k = \sum_{z\in S} \rho_k(z)\mu(z). \tag{6.27}$$

这种将随机矩阵提升的思想表达了这样的现象: 节点 x 通过由社团 S_k 到社团 S_l 的不同的渠道, 并以它们相应的从属概率转移到节点 y, 最终停留在平衡状态. 不难验证如果 $\hat{p}$ 是 $\mathbb{S}$ 上的随机矩阵, 具有平稳分布 $\hat{\mu}$, 则由 (4.9) 定义的 $\tilde{p}$ 为 S 上的随机矩阵, 具有平稳分布 μ. 若进一步, 有$\hat{p}$ 满足关于 $\hat{\mu}$ 的细致平衡条件, 则 $\tilde{p}$ 满足关于 μ 的细致平衡条件 (命题 4.1).

给定社团数目 N, 最优地约化 Markov 随机游动动力学, 通过考虑如下的极小化问题

$$\min_{\rho_k(x),\ \hat{p}_{kl}} J \quad = \quad \|p-\tilde{p}\|_\mu^2 = \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} |p(x,y)-\tilde{p}(x,y)|^2$$

$$= \sum_{x,y\in S} \mu(x)\mu(y)\left(\frac{p(x,y)}{\mu(y)} - \sum_{k,l=1}^{N} \rho_k(x)\rho_l(y)\frac{\hat{p}_{kl}}{\hat{\mu}_l}\right)^2 \tag{6.28}$$

服从于约束条件 (6.25) 和 (6.24). 为了极小化 (6.28) 中的目标函数 J, 定义

$$\hat{p}^*_{kl} = \frac{1}{\hat{\mu}_k}\sum_{x,y\in S} \mu(x)\rho_k(x)p(x,y)\rho_l(y), \tag{6.29}$$

则 $\hat{p}^*$ 是 $\mathbb{S}$ 上的随机矩阵, 平稳分布为 $\hat{\mu}$, 且满足关于 $\hat{\mu}$ 的细致平衡条件.

带有约束条件 $\sum_{k=1}^{N}\rho_k(x) = 1$ 的 J 的最优化问题, 相当于求 (6.28) 的稳定点, 其 Euler-Lagrange 组如下所述

$$\left(I_{\hat{\mu}}^{-1}\cdot\hat{\hat{\mu}}\right)\cdot\hat{p}\cdot\left(I_{\hat{\mu}}^{-1}\cdot\hat{\hat{\mu}}\right) = \hat{p}^*, \tag{6.30a}$$

$$\rho = I_{\hat{\mu}}\hat{p}^{-1}\hat{\hat{\mu}}^{-1}\rho p^T, \tag{6.30b}$$

其中 $\rho = \{\rho_k(x)\}_{k=1,\ldots,N,x\in S}$ 为 $N\times n$ 的矩阵, 定义 $\hat{\hat{\mu}}$ 为 $N\times N$ 的矩阵, 其元素为

$$\hat{\hat{\mu}}_{kl} = \sum_{z\in S}\mu(z)\rho_k(z)\rho_l(z) = (\rho\cdot I_\mu\cdot\rho^T)_{kl}. \tag{6.31}$$

对角矩阵 I_μ 和 $I_{\hat{\mu}}$ 分别为 $n\times n$ 和 $N\times N$ 的矩阵, 其元素为

$$I_\mu(x,y) = \mu(x)\delta(x,y), \quad x,y\in S, \tag{6.32a}$$

$$(I_{\hat{\mu}})_{kl} = \hat{\mu}_k\delta_{kl}, \quad k,l = 1,\ldots,N, \tag{6.32b}$$

其中 $\delta(x,y)$ 和 δ_{kl} 均为 Kronecker delta 符号.

由 Euler-Lagrange 方程组 (6.30) 立即得到的一个策略是在关于 $\hat{p}$ 和 ρ 的方程之间交替迭代. 为了保证算法的可实现性, 即 $\hat{p}$ 和 ρ 的非负性和归一化条件, 需要在每次迭代后加入一个投影步, 即将最优性条件 (6.30) 变为

$$\hat{p} = \mathcal{P}\left(\hat{\hat{\mu}}^{-1}\cdot I_{\hat{\mu}}\cdot\hat{p}^*\cdot\hat{\hat{\mu}}^{-1}\cdot I_{\hat{\mu}}\right), \tag{6.33a}$$

$$\rho = \mathcal{P}\left(I_{\hat{\mu}}\hat{p}^{-1}\hat{\hat{\mu}}^{-1}\rho p^T\right). \tag{6.33b}$$

这里 $\mathcal{P}$ 是一个投影算子, 它可将一个实向量映射成一个有非负归一化分量的向量.

6.2.2　模糊模量的定义

为了衡量网络不同的模糊分区的优良性, 本节将模量 (6.14) 扩展到一个概率性的形式中. 根据上述框架, 连接一对节点的边不再属于某个特定的社团, 而是以非零的概率属于不同的社团. 这得归结为相应的节点属于这些社团的概率. 这就激发了模糊模量的定义

$$Q_f = \text{社团内部的边的概率的和} - \text{这些边的概率的期望的和}. \tag{6.34}$$

它可以看作是量化概率性分区的质量的一个标准, 从而协助自动模型选择过程, 而不再要求社团数目已知.

对于给定的模糊分区 $\{\rho_k(x)\}_{k=1}^N$, 根据多数决定原则 (majority rule) 将节点分区, 即如果对于给定的节点 x, 有 $k = \arg\max_l \rho_l(x)$, 则令 $x \in S_k$. 于是模糊模量 Q_f 也可以基于空模型[141]来给出

$$Q_f = \frac{1}{2m}\sum_{k=1}^{N}\sum_{x,y\in S_k}\Big(\frac{\rho_k(x)+\rho_k(y)}{2}e(x,y) - p_f^E(x,y)\Big), \tag{6.35}$$

其中 $p_f^E(x,y)$ 为边 $e(x,y)$ 属于 S_k 的概率的期望值, 具有形式

$$p_f^E(x,y) = \frac{d_f(x)d_f(y)}{2m}, \quad x,y \in S_k \tag{6.36}$$

这里 $d_f(x)$ 是概率性框架下, 节点 x 在社团 S_k 中的广义度, 形式如下

$$d_f(x) = \sum_{z\in S_k}\frac{\rho_k(x)+\rho_k(z)}{2}e(x,z) + \sum_{z\notin S_k}\frac{\rho_k(x)+(1-\rho_k(z))}{2}e(x,z). \tag{6.37}$$

这个扩展形式 (6.35) 可以看作是传统模量 (6.14) 的推广. 对于固定社团数目 N 的一个理想的分区要求在 $\{\rho_k(x)\}_{k=1}^N$ 中的较为稳定的状态. 因此, 最优概率性分区可以通过求解如下问题得到

$$\max_N\left\{\max_{\{\rho_k(x)\}_{k=1}^N} Q_f\right\}. \tag{6.38}$$

需要指出在得到节点关于社团的模糊分区 ρ 之后, 模糊模量仅利用了 ρ 的最大分量属于相同社团的节点对来定量化. 这看起来扔掉了模糊分区 ρ 中的其它信息, 不仅是最大分量的其它信息. 然而, 数值实验结果表明 (6.35) 的实际应用产生与使用原始模量不同的分区, 如图 6.12 所示. 而根据下述的算法, ρ 的余下信息在当产生重叠社团时予以考虑, 如图 6.11 所示.

6.2.3 算法的构造

正如 6.1.3 中所介绍的, 可以采用模拟退火方法[83, 103, 133]用来寻找模量的最大值.

设 $E = -Q$, $E^{(n)}$ 和 $E^{(n+1)}$ 分别表示当前能量和新能量. $E^{(n+1)}$ 总被接受如果它满足 $E^{(n+1)} < E^{(n)}$, 但是如果 $E^{(n+1)} > E^{(n)}$ 则新能量级仅以概率 $\exp(-\frac{1}{T}\triangle E^{(n)})$ 被接受, 其中 $\triangle E^{(n)} = E^{(n+1)} - E^{(n)}$ 为能量差, T 为当前温度. 初始状态 $\{\rho_k^{(0)}(x)\}_{k=1}^N$ 随机生成, 其中 N 为区间 $[N_{\min}, N_{\max}]$ 中的整数, 在下面的计算中选取 $N_{\min} = 2, N_{\max} = n/3$. 初始温度设置为一个较高的温度 $T_{\max}$. 当前状态的下一个状态由随机选择下述的提议之一而产生, 然后计算这个新状态的能量. 这个新的状态被接受如果满足接受条件. 这个过程将在给定的温度上重复 R 次. 设置冷却速率 $0 < \alpha < 1$ 来降低当前温度直到温度下界 $T_{\min}$ 被达到. 整个的极大化模糊模量的结合 AIP 迭代算法的模拟退火过程概括如下.

算法 6.2 (Simulated Annealing to maximize the Fuzzy Modularity associating with AIP — SAFM)

(1) 设置参数 $T_{\max}$, $T_{\min}$, α 和 R. 在区间 $[N_{\min}, N_{\max}]$ 中随机选取 N, 并随机初始化模糊分区 $\{\rho_k^{(0)}\}_{k=1}^N$. 设当前温度 $T = T_{\max}$.

(2) 根据 (6.33a) 计算 $\hat{p}^{(0)}$, 然后根据 (6.35) 计算初始能量 $E^{(0)}$; 令 $n^* = 0$.

(3) 对于 $n = 0, 1, \ldots, R$, 做如下迭代:

  (3.1) 根据下述提议产生一组新的模糊分区 $\{\rho_k^{(n)}\}_{k=1}^{N'}$, 并令 $N = N'$.

  (3.2) 分别根据 (6.33a), (6.33b) 和 (6.35) 来更新 $\hat{p}^{(n+1)}$, $\rho^{(n+1)}$ 和新能量 $E^{(n+1)}$.

  (3.3) 接受或拒绝新状态. 如果 $E^{(n+1)} < E^{(n)}$, 或者 $E^{(n+1)} > E^{(n)}$ 且 $u \sim \mathcal{U}[0,1]$, $u < \exp\{-\frac{1}{T}\triangle E^{(n)}\}$, 则接受新状态, 令 $n = n + 1$; 否则拒绝.

  (3.4) 更新最优状态, 即如果 $E^{(n)} < E^{(n^*)}$, 则令 $n^* = n$.

(4) 降温 $T = \alpha \cdot T$. 如果 $T < T_{\min}$, 执行 (5); 否则令 $n = n^*$, 并重复 (3).

(5) 输出整个过程的最优模糊分区 $\{\rho_k^{(n^*)}\}_{k=1}^N$ 和最小能量 $E^{(n^*)}$. 多数决定原则

$$S_k = \{x : k = \arg\max_l \rho_l^{(n^*)}(x)\} \tag{6.39}$$

给出确定性分区, 而

$$S_k = \{x : \rho_k^{(n^*)}(x) > \eta\} \tag{6.40}$$

给出重叠的社团结构, 其中 $\eta > 0$ 为阀值.

在下面的计算中选择投影算子 $\mathcal{P}$ 为投影到边界的直接投影. 设 $\boldsymbol{u} = (u_1, u_2, \ldots, u_N) \in \mathbb{R}^N$, 且 $\Lambda = \{i : u_i \geq 0\}$. 当 $i \notin \Lambda$ 时, 令$\mathcal{P}u_i = 0$; 否则令 $\mathcal{P}u_i = u_i / \sum_{j\in\Lambda} u_j$.

对于步骤 (3.1) 中产生一组新的模糊分区过程的提议由三个函数组成, 分别是保持一个当前社团, 删除一个当前社团和分裂一个当前社团. 在每次迭代时, 随机选择三个函数中的一个, 并且社团规模

$$M_k = \sum_{x\in S} \rho_k(x), \quad k = 1, \ldots, N, \tag{6.41}$$

被用来选择一个社团, 显然如果社团的规模越大则它存在性越强. 这三个函数描述如下.

(a) 保持一个社团. 保持这当前的社团结构.

(b) 删除一个社团. 具有最小社团规模 M_d 的社团被选择, 于是从当前模糊分区矩阵 ρ 中删除第 d 行, 并将这一行加在第 k 行上, 即 $\rho_k = \rho_k + \rho_d$, 其中 $k = \arg\max_m \hat{p}^*_{dm}$, $\hat{p}^*$ 如 (6.29) 所定义.

(c) 分裂一个社团. 具有最大社团规模 M_s 的社团被选择, 于是将其分成两个新社团. 取 $r(x) \sim^{i.i.d.} \mathcal{U}[0,1]$, 则 $\rho_{N+1}(x) = r(x) \cdot \rho_s(x), \rho_s(x) = (1 - r(x)) \cdot \rho_s(x), \forall x \in S$.

算法的迭代步数依赖于初始和终止温度以及冷却速率. 每次迭代中, 计算 $\hat{p}$ 的总花费为 $O(N^2(m+n))$, 计算 ρ 的花费为 $O(N^2 n + Nm)$. 全局最大化问题 (6.38) 也可以通过对所有可能的 N 遍历使用 AIP 迭代算法, 但这将花费巨大. 因为对于每个固定的参数 N, AIP 初始化时的变形 k-means 需要执行大约 2000 次试验以避免陷入局部极小值. 然而 SAFM 可以避免无效重复并具有高效性和高精度. 此外, 它还可以得到比对所有可能的 N 遍历使用 AIP 算法更大的 Q_f 值, 从而实现更优的分区 (如图 6.12 所示). SAFM 的另一个优点是它克服了第四章[114]中方法的弱

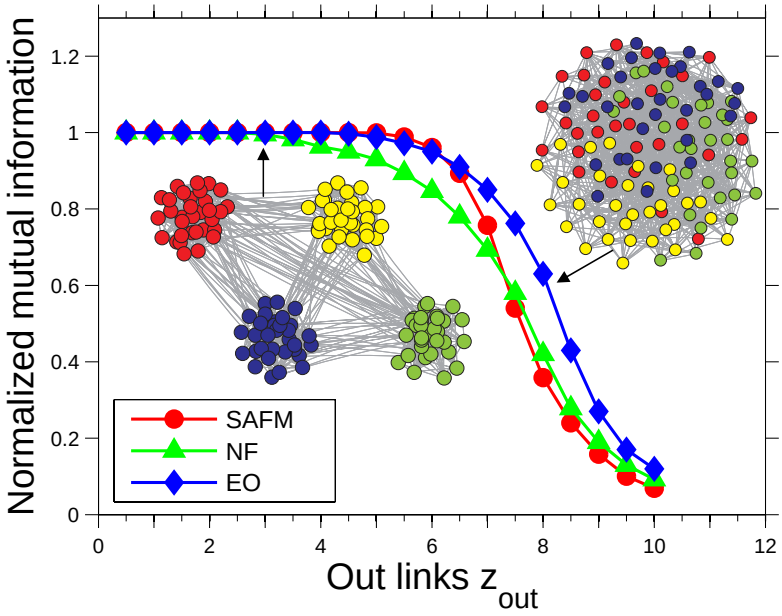


图 6.9: 将 SAFM 算法与 Newman 快速算法 (NF)[140] 和极值最优化算法 (EO)[57]共同测试于 128 个节点的 ad hoc 网络[43, 77, 144]并进行比较. Ad hoc 网络具有四个社团: 对于较低的 z_{out}, 社团可以轻松地识别; 而对于较高的 z_{out}, 社团边的更加复杂.

点. 冷却过程可以自动地确定社团数目 N 而不再是固定它作为已知的模型参数, 并且初始的模糊分区 $\{\rho_k^{(0)}\}$ 可以随机选取, 而不再选用由变形 k-means 得到的关于每个节点确定性分区的示性矩阵.

6.2.4 数值试验

本节中, 将算法测试于具有已知社团结构的人工生成的网络, 包括 128 个节点的 ad hoc 网络和 LFR 基准网络. 随后, 算法成功地应用于真实世界中的网络, 包括空手道俱乐部网络, 神奇湾宽吻海豚网络, 美国政治书籍网络和圣非研究所科学家合作网络.

6.2.4.1 人工生成的网络

128个节点的 ad hoc 网络. 本节的第一个算例是 128 个节点的 ad hoc 网络, 这类网络具有已知的社团结构, 构造如 1.5.1 所述. 通常定义 z_{out} 为某个节点与属

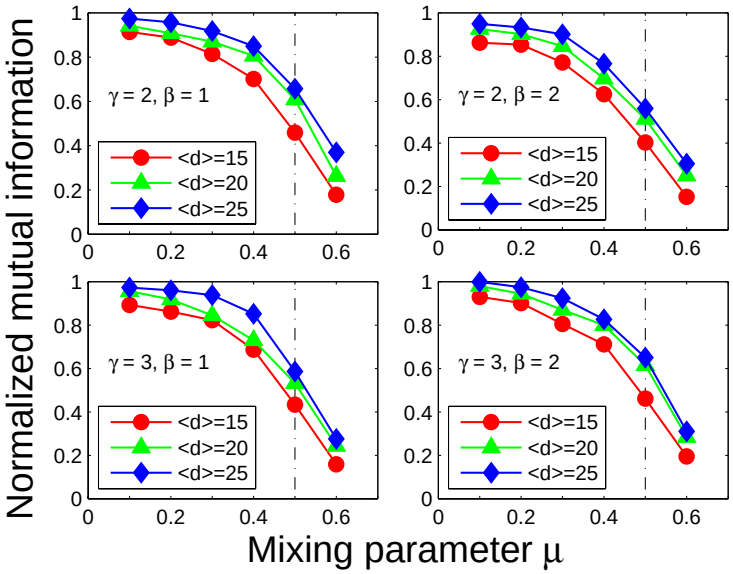


图 6.10: 将 SAFM 算法测试于 LFR 基准网络[111]. 节点数为 $n = 500$. 结果明显地依赖于基准网络的所有参数, 从指数 γ 和 β 到平均度 $\langle d\rangle$. 由垂直虚线表示的阀值 $\mu_c = 0.5$ 标记出一个边界; 超出这个边界则社团不再是强意义下 (即每个节点在自己从属的社团中比在其它社团中具有更多的邻居) 所定义的. 每个点对应于超过 20 次的图实现的平均值.

于其它社团节点之间连接的平均数, z_{out} 越大, 社团就变得越模糊 (diffuse). 在这个模型计算中参数设置为 $T_{\max} = 3$, $T_{\min} = 10^{-2}$, $\alpha = 0.9$ 和 $R = 50$. 如果在应用 SAFM 算法之后根据多数决定原则分割网络, 即根据节点的最大权重分区, 则得到了一个确定性的分区. 为比较固定模块结构与算法得到的结构, 这里采用归一化互信息[43, 109, 111], 如 (1.8) 所定义. 将 $z_{\text{out}} = 96p_{\text{out}}$ 从 0.5 变到 10, 并观察归一化互信息. 正如图 6.9 所示, 不同方法求得的归一化互信息随 z_{out} 的增长而呈现相似方式的变化, 同时社团结构也变得更加模糊. 正常的分区总被找到直到 $z_{\text{out}} = 6$ 为止, 之后方法开始失效. SAFM 看起来与 Newman 快速算法[140] 和极值最优化算法[57]相比具有竞争力, 特别是对于 z_{out} 更高的社团结构更为模糊的情形. 这也证实了 SAFM 算法的精确性, 但 SAFM 算法给出了关于每个节点更多详细的信息.

LFR 基准网络. LFR 基准网络[108, 109, 111]是为研究社团结构而构造的一个现实的基准网络, 它同时要求节点度和社团规模的非均匀性. 节点度服从指数为 γ 的幂律分布, 而社团规模服从指数为 β 的幂律分布. 混合参数 μ 作为独立参数, 它表

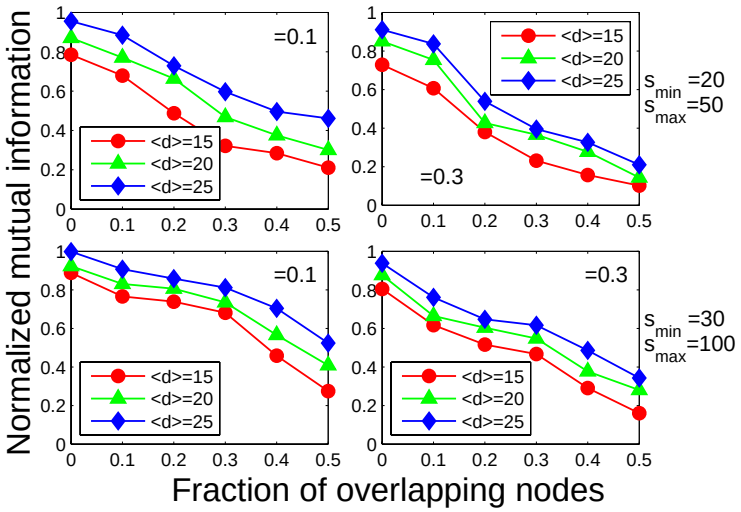


图 6.11: 将 SAFM 算法测试于无向无权但具有重叠社团的 LFR 基准网络[108]. 图中展现了已知的重叠分区和重新获得的分区之间的针对重叠社团的广义形式的归一化互信息[110]随重叠节点比率的变化. 络具有 $n = 500$ 个节点, 其它参数为 $\gamma = 2, \beta = 1$ 和 $d_{\max} = 50$. 每个点对应于超过 20 次的图实现的平均值.

示一个节点关于它所在社团的外面的度与全部度之间的比率[111]. LFR 基准网络可进一步地推广到具有重叠社团的情形[108], 相应的对于重叠社团的广义化的归一化互信息被提出并用来实现测试算法的目的[110]. 详细介绍如 1.5.1 所述.

在图 6.10 中, 展现了将 SAFM 算法实施于 $n = 500$ 的标准 LFR 基准网络[111]的结果. 计算中的模型参数设为 $T_{\max} = 3.0$, $T_{\min} = 0.01$, $\alpha = 0.9$ 和 $R = 20$. 四个子图分别对应于取值为四对指数 $(\gamma, \beta) = (2, 1), (2, 2), (3, 1), (3, 2)$ 的结果. 为了探索网络结构的万象, 选择指数范围的极端的组合. 每条曲线表现了归一化互信息随混合参数 μ 的变化. 可以看出平均度 $\langle d \rangle$ 越大则算法的性能越好, 但是当混合参数变大时算法性能变差. 图中垂直虚线表示的阀值 $\mu_c = 0.5$ 标记出一个边界; 超出这个边界则社团不再是强意义下 (即每个节点在自己从属的社团中比在其它社团中具有更多的邻居) 所定义的. 总而言之, 可以推断 SAFM 方法给出好的结果.

在图 6.11 中, 展现了将 SAFM 算法实施于具有重叠社团的 LFR 基准网络[108]的结果. 网络具有 $n = 500$ 个节点, 其它参数为 $\gamma = 2, \beta = 1$ 和 $d_{\max} = 50$. 在这个情形下, 固定混合参数 μ, 而变化社团之间重叠节点的比率. 通过设置与图 6.10

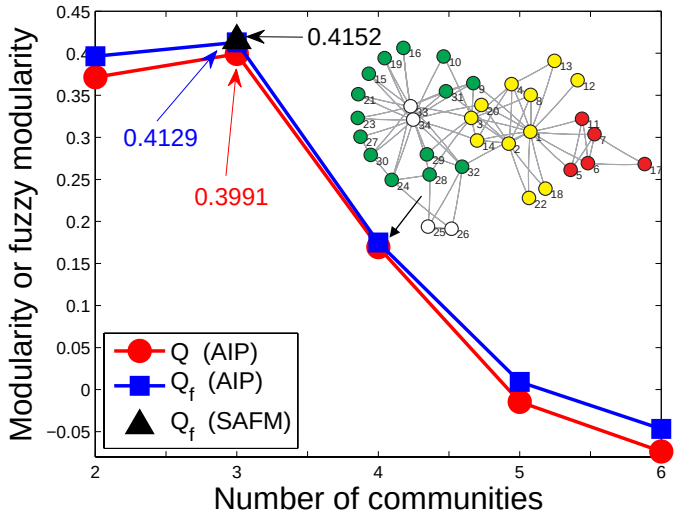


图 6.12: 利用算法 AIP 所得到的原始模量和模糊模量的值. 图中清楚地表明 SAFM 可以找到一个比对所有可能的 N 遍历 AIP 算法[114]更大的模糊模量值 $Q_f = 0.4152$. 插入图表示 $N = 4$ 时的 AIP 算法得到的社团结构, 当 $N \geq 4$ 时分区结果变得更为复杂.

的计算相同的参数, 观察对于不同的平均度 $\langle d \rangle$, 人工分区和重新获得的分区之间的针对重叠社团的广义形式的归一化互信息[110]的变化. 这里选择 $\eta = 0.3$ 来产生重叠社团. 也可注意到平均度 $\langle d \rangle$ 越大则算法的性能越好, 而当重叠节点比率变大时算法性能变差. 在上面两个子图中社团尺寸的范围是在 $s_{\min} = 20$ 和 $s_{\max} = 50$ 之间, 而下面两个子图中社团尺寸的范围是在 $s_{\min} = 30$ 和 $s_{\max} = 100$ 之间. 通过比较上排和下排的图, 可发现当社团平均尺寸越大时算法性能越好.

6.2.4.2　真实世界中的网络

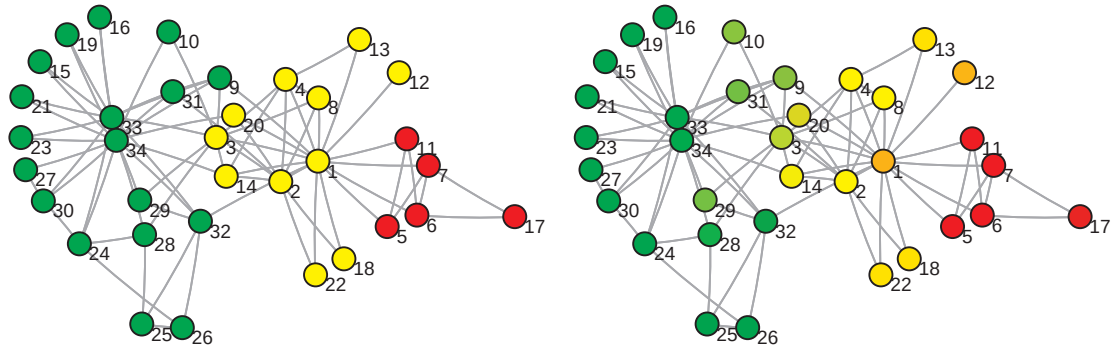
空手道俱乐部网络. 这个网络是由 Wayne Zachary 在观察一所美国大学空手道俱乐部成员之间的社交而构建的[210], 具体介绍见 1.5.2. 通过设定参数 $T_{\max} = 3.0$, $T_{\min} = 10^{-5}$, $\alpha = 0.9$ 和 $R = 50$, 执行 SAFM 算法. 图 6.12 中详细详细阐明了结合迭代的模拟退火方法可以得到一个比对所有可能的 N 遍历 AIP 算法[114]的更大的模糊模量 $Q_f = 0.4152$. SAFM 算法得到的数值结果如表 6.2 所示. 图 6.13(a) 展示了通过多数决定原则产生的 3 个社团, 分别用不同颜色表示. 然而实际上算法获得了更多详细信息. 从表 6.2 中, 发现对于节点 $\{5, 6, 7, 11\}$ 有

表 6.2: 空手打俱乐部网络中每个节点属于不同社团的联合概率. ρ_R, ρ_Y 和 ρ_G 分别表示属于图 6.13 中红色, 黄色和绿色社团的概率.

Nodes	1	2	3	4	5	6	7	8	9	10	11	12
ρ_R	0.3322	0	0	0	1.0000	1.0000	1.0000	0	0.0293	0	1.0000	0.3165
ρ_Y	0.6678	1.0000	0.6841	1.0000	0	0	0	1.0000	0.4598	0.4920	0	0.6835
ρ_G	0	0	0.3159	0	0	0	0	0	0.5109	0.5080	0	0
Nodes	13	14	15	16	17	18	19	20	21	22	23	24
ρ_R	0.0780	0	0	0	0.9741	0.0780	0	0.0587	0	0.0780	0	0.0105
ρ_Y	0.9220	0.9482	0	0	0	0.9220	0	0.7695	0	0.9220	0	0
ρ_G	0	0.0518	1.0000	1.0000	0.0259	0	1.0000	0.1718	1.0000	0	1.0000	0.9895
Nodes	25	26	27	28	29	30	31	32	33	34		
ρ_R	0.0340	0.0322	0	0	0	0.0009	0	0.0753	0	0		
ρ_Y	0	0	0	0.1615	0.4170	0	0.4074	0.0653	0.0289	0.0801		
ρ_G	0.9660	0.9678	1.0000	0.8385	0.5830	0.9991	0.5926	0.8594	0.9711	0.9199		

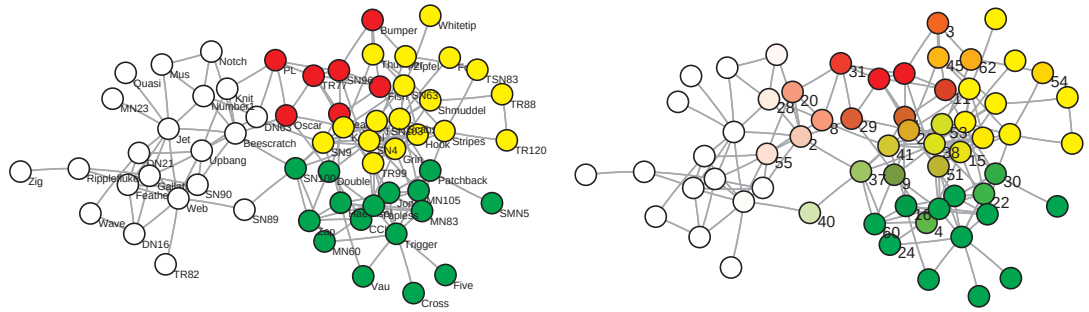
$\rho_R = 1$, 对于节点 $\{2, 4, 8\}$ 有 $\rho_Y = 1$, 对于节点 $\{15, 16, 19, 21, 23, 27\}$ 有 $\rho_G = 1$, 它们都位于相应颜色社团的边界. 其它节点均以非零概率属于三个社团, 特别是节点 $\{1, 3, 9, 10, 12, 20, 29, 31\}$ 具有更加模糊的权重, 并起到其相应社团之间转移节点的作用. 由权重平均可视化给出的模糊社团结构如图 6.13(b) 所示, 节点的颜色由社团颜色的加权平均 (4.31) 给出, 这里假设 $\boldsymbol{v}_R$, $\boldsymbol{v}_Y$ 和 $\boldsymbol{v}_G$ 分别表示可视化工具中红色, 黄色和绿色的向量, 则节点 x 的颜色向量为 $\rho_R(x)\boldsymbol{v}_R + \rho_Y(x)\boldsymbol{v}_Y + \rho_G(x)\boldsymbol{v}_G$. 这样可将不同社团之间的转移更清楚地表示出来. 可以推测位于中心地带的成员与社团联系更为紧密.

宽吻海豚网络. 宽吻海豚网络由生活在新西兰道尔福峡湾 (神奇湾) 的一个组织中的 62 只宽吻海豚之间的频繁联系所构成的网络[121, 122]. 关于这个网络的详细介绍参见 1.5.2. 分区结果如表 6.3 和图 6.14(a) 所示. 根据图 6.14(a), 网络看起来分裂成白色部分和较大部分这两个大的社团, 其中较大部分继续分裂成几个更小的社团, 分别用不同颜色表示. 分裂成两个社团对应于根据海豚年龄所形成的一个已知分区[122]. 而较大的那部分中的几个子社团也对应于海豚的真实的分区, 黄色



(a)　　(b)

图 6.13: (a) 由 SAFM 算法得到的经多数决定原则处理之后的空手道俱乐部网络社团结构, 三个社团由不同颜色表示. (b) 由权重 $\{\rho_k(x)\}$ 作可视化的模糊社团结构, 对应的模糊模量为 $Q_f = 0.4152$.



(a)　　(b)

图 6.14: (a) 由 SAFM 算法得到的经多数决定原则处理之后的宽吻海豚网络社团结构, 相应的 4 个社团由不同颜色表示. (b) 由权重 $\{\rho_k(x)\}$ 作可视化的模糊社团结构, 对应的模糊模量为 $Q_f = 0.5050$.

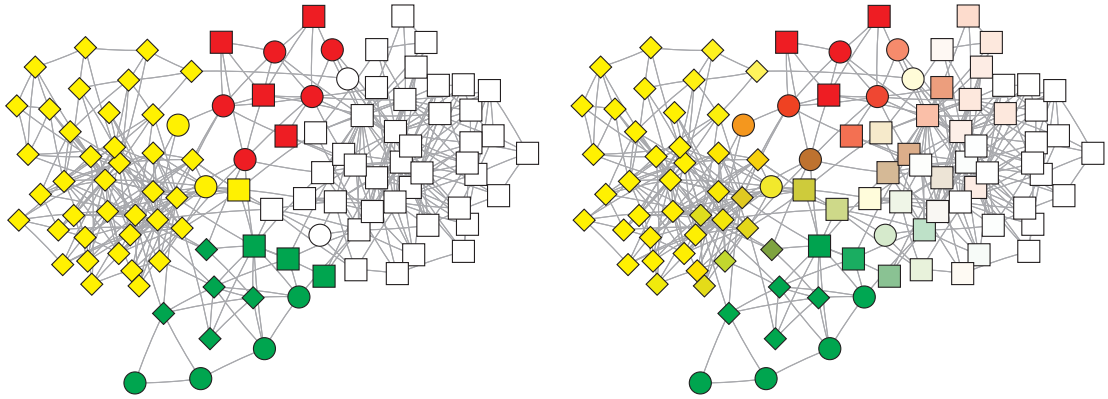
部分几乎全部由雌性组成, 而其它部分几乎全部都是雄性[122]. 表 6.3 中列出了具有中间权重的节点的联合概率. 对于其它节点, 即时它们不具有 0-1 权重, 也会具有一个权重强度大于 0.95 的主导权重. 看起来节点 $\{2, 8, 20, 29, 31, 37, 40\}$ 起到了两个大社团之间的主要连接的作用, 而表中的其它节点也起着所检测到的 4 个社团之间的过渡作用. 权重 $\{\rho_k(x)\}$ 的可视化如图 6.14(b) 所示. 总之, SAFM 算法得到的模糊分区再一次反映了这个社会网络中海豚之间频繁联系的强弱程度.

表 6.3: 宽吻海豚网络中具有中间权重的节点属于不同社团的联合概率. ρ_R, ρ_Y, ρ_G 和 ρ_W 分别表示属于图 6.14 中红色, 黄色, 绿色和白色社团的概率. 对于网络中的其它节点, 即使它们不具有 0-1 权重, 也会具有一个权重强度大于 0.95 的主导权重.

Nodes	1	2	3	4	8	9	11	15	16	20	21	22	24	28
ρ_R	0.7389	0.2364	0.7398	0.0216	0.4818	0.2539	0.8918	0.0209	0.1124	0.4829	0.3320	0	0	0.0652
ρ_Y	0.1092	0	0.2602	0.3188	0.0163	0.1229	0	0.8763	0	0	0.5252	0.2589	0	0.0412
ρ_G	0.1519	0.0277	0	0.6596	0	0.6232	0.1082	0.1028	0.8876	0	0.1427	0.7411	0.9445	0
ρ_W	0	0.7359	0	0	0.5019	0	0	0	0	0.5171	0	0	0.0555	0.8936
Nodes	29	30	31	37	38	40	41	45	51	53	54	55	60	62
ρ_R	0.7688	0.0648	0.9053	0.0540	0	0	0.1309	0.3057	0.2103	0	0.1360	0.1341	0.0406	0.3663
ρ_Y	0.0027	0.1602	0	0.3176	0.8352	0.1972	0.6239	0.6943	0.4928	0.7828	0.8640	0.0027	0.0218	0.6337
ρ_G	0.1046	0.7750	0	0.4156	0.1648	0.1836	0.1945	0	0.2969	0.2172	0	0	0.9125	0
ρ_W	0.1239	0	0.0947	0.2128	0	0.6192	0.0507	0	0	0	0	0.8632	0.0251	0

美国政治书籍网络. 这个网络是由 V. Krebs 编制的关于美国政治的书籍的网络. 关于这个网络的详细介绍参见 1.5.2. 如图 6.15 所示, 节点的给出的根据为它们属于保守的 (方形) 还是自由的 (菱形), 除此之外还有小部分书籍是中立的 (圆形). 计算结果如图 6.15 所示. 算法发现四个社团, 分别用不同的颜色表示. 看起来这些社团中的其中一个几乎全部由属于自由的书籍组成, 一个乎全部由属于保守的书籍组成. 多数属于中立的书籍落入余下的两个社团中. 因此这些书似乎形成了与政治观点密切相关的联合购买的社团, SAFM 算法能够从原始数据网络中提取有意义的结果. 此外, 模糊社团结构指出了每本书在政治观点上的倾向性.

圣达菲研究所科学家合作网络. 最后一个例子是是美国新墨西哥州圣达菲 (Santa Fe) 的一个交叉学科研究中心, 圣达菲研究所中的科学家之间的合作网络[77]. 在这个网络中的 271 节点代表在 1999-2000 年居住在圣达菲研究所的科学家以及他们的合作者. 如果在同样的时间段内, 两个科学家之间合作过一篇或者更多的论文, 则他们之间就画上一条带权重的边. 关于这个网络的详细介绍参见 1.5.2. 图 6.16(a) 中给出了 SAFM 算法应用于合作网络的最大分量, 并根据多数决定原则所得到的结果. 这个网络由 118 个科学家组成, 分成 6 个社团由不同颜色表示. 图中位于顶部的社团 (红色) 表示利用基于智能体模型来研究经济和交通流量



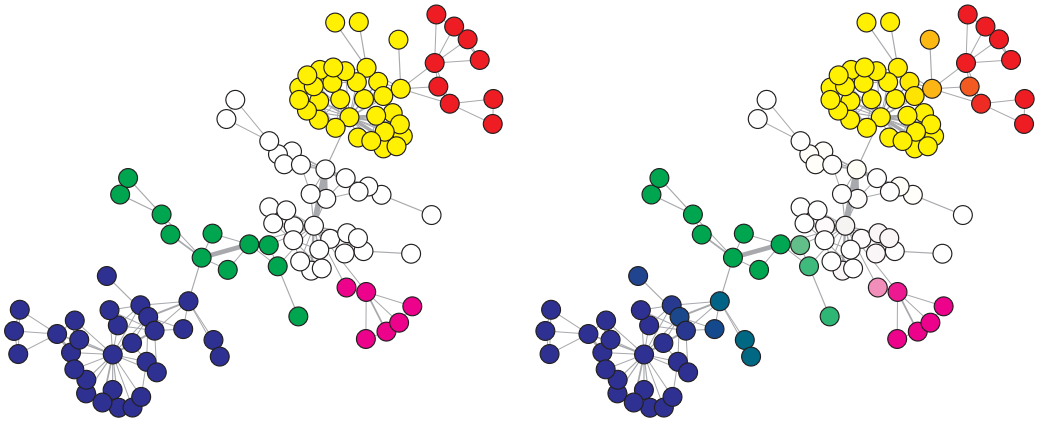
(a)　　(b)

图 6.15: (a) 由 SAFM 算法得到的经多数决定原则处理之后的美国政治书籍网络社团结构, 相应的 4 个社团由不同颜色表示. (b) 由权重 $\{\rho_k(x)\}$ 作可视化的模糊社团结构, 对应的模糊模量为 $Q_f = 0.5184$.

问题的科学家社团. 下面一个社团 (黄色) 表示研究生态学中数学模型的科学家社团, 形成了一个相当凝聚的结构. 最大的社团 (白色, 紫红色, 绿色) 是主要研究统计物理的科学家社团, 并看起来继续分成 3 个更小的社团. 在这个情形下, 每个子社团看起来围绕在一个主导成员的研究兴趣周围. 最后的位于图中底部的社团 (蓝色) 是主要研究 RNA 结构的科学家社团. 权重 $\{\rho_k(x)\}$ 的可视化如图 6.16(b) 所示, 它清楚地显示出圣达菲研究所科学家之间的合作程度, 并指出他们以不同的概率倾向于加入哪一个研究领域, 致力于交叉学科领域的成员也可以被找到并衡量.

6.2.5　小结

本节中作者提出了模糊模量函数来衡量网络概率性分区的优良性, 并构造了相应的算法 SAFM, 且算法成功地应用于几个具有代表性的网络. 数值实验展现了非常满意的结果, 即 SAFM 算法可以高效率且高精度地确定每个节点属于不同社团的概率. 结合迭代法的模拟退火过程的实施避免了无效重复, 并可以获得比对于所有 N 遍历迭代算法更大的模糊模量值. 所提出的 SAFM 方法成功地克服了第四章[114]中的弱点, 此时社团数目 N 可以被自动确定而不再是将它固定为已知的模型参数, 并且初始模糊分区 ρ 可以随机选取, 而不再是取变形 k-means 算法[60]所



(a) (b)

图 6.16: (a) 由 SAFM 算法得到的经多数决定原则处理之后的圣达菲研究所科学家合作网络社团结构, 相应的 6 个社团由不同颜色表示. (b) 由权重 $\{\rho_k(x)\}$ 作可视化的模糊社团结构, 对应的模糊模量为 $Q_f = 0.7075$.

得到的确定性分区的示性矩阵. 根据多数决定原则和阀值操作分别产生确定性分区和重叠社团结构. 总之, 模糊社团结构包含了更多详细信息并且较之先前的网络分区方法具有更多预测性的功能.

第七章　总结与展望

通过博士阶段的研究工作, 我打开了探视更广阔自然世界的缝隙. 我相信, 在自然科学的很多领域, 因为科学计算的发展和进步, 数学将发挥超越传统工具的限制而起到更加现实的影响; 具体到复杂网络社团结构的研究, 本文中的工作只是事情开始的第一步而已, 后续可做的事情还有很多. 我期望, 通过亲自动手使得对复杂网络社团结构的研究的理解更加深入, 更加形象, 更加具体.

本章内容组织如下: 在 7.1 中对本论文的研究工作进行了简要的总结, 包括本论文创新点的展现以及方法不足之处的描述. 在 7.2 中将本论文提出的一些算法与文献中的其它方法进行比较, 并分析本文方法的优势和劣势. 在 7.3 中对未来将要继续研究的内容进行了展望和讨论.

7.1　本文研究的总结

在本论文中, 作者主要研究一类基于随机游动的动力学方法. 本论文的基石是最近由 E, Li 和 Vanden-Eijnden 发展的基于 Hilbert-Schmidt 度量粗粒化可逆马氏链的理论框架 (*Proc. Natl. Acad. Sci. USA* **105** (2008), 7907–7912), 作者进一步发展并完善了由此理论所建立的复杂网络社团结构的确定性分区方法, 所得到的主要创新成果如下:

(a) 提出了复杂网络社团结构的一个概率性框架, 其中每个节点以某一概率从属于某一个社团. 这可以看作是统计中的 fuzzy c-means 算法向网络分区问题的自然扩展, 也可以看做是之前的网络最优分区的确定性框架的推广. 提出的算法成功地应用于几个具有代表性的算例. 概率性框架为网络分区问题的研究提供更详细的信息. 更重要的是, 它比传统的网络确定性分区方法更具有预测性能.

(b) 设计了一个基于有效性指标 (validity index) 的方法来实现确定性分区的自动模型选择. 提出的有效性指标函数可以为社团结构的优良程度提供一种度

量, 它是由两个因素的乘积所定义的, 分别是每个分区的社团内部紧密程度 (compactness) 与社团间分离程度 (separation). 数值试验表明算法在降温过程中可以有效找出社团结构, 并且无需任何关于社团结构的先验信息就可以自动确定社团的个数. 算法的 matlab 程序可以从网上免费下载使用, 下载链接为:

http://dsec.pku.edu.cn/~tieli/software/SAVI.zip.

(c) 分别利用结合了两种 k-means 迭代的模拟退火方法来最大化模量 (modularity), 以实现确定性分区的自动模型选择. 这两种 k-means 分别基于相异性指标和扩散距离. 算法可以得到较之许多已有方法更大的模量的值, 从这个意义上来说胜过了许多已有的方法. 算法不仅可以确定社团的个数以及社团结构, 还可以给出每个社团的中心节点.

(d) 构造了实现网络概率性分区的自动模型选择的方法. 提出了模糊模量 (fuzzy modularity) 函数, 它可以看作是传统模量的一个推广, 并为网络模糊社团结构的优良性提供了度量. 算法可以有效确定每个节点属于不同社团的概率, 并且初始的模糊分区可以随机选取, 社团的个数也可以自动确定而不再是将其固定为已知的模型参数.

对于基于最优预测的确定性分区算法: 算法 3.8 (变形 k-means), 仅仅用一个类似于 k-means 的过程来极小化目标函数 (3.28), 这是由于网络中的节点并未嵌入到一个度量空间中去, 从而用 (3.53) 来表示某个节点到某个社团的“距离”, 这导致了算法的迭代不能保证目标函数每一次都下降, 于是可以通过如果目标函数上升或保持常数则终止迭代的方式来解决这个问题. 这个算法的构思新颖. 第三章的数值结果也表明变形 k-means 算法的有效性良好. 但是在空手道俱乐部这个算例中, 节点 10 被误分区, 而这在多数的社团结构检测的算法中是不会发生的. 但是算法的计算量很少, 仅为每次迭代 $O(N(n+m))$, 这又与多数算法相比十分优越, 这将在下面具体分析. 这个算法的一个不足是无法自动确定社团的个数, 社团个数 N 必须作为已知的模型参数. 在第五章中解决了这个问题, 即实现了确定性分区的自动模型选择. 所提出的有效性指标函数可与模量函数相媲美, 但是也存在着不足. 例如宽吻海豚网络的例子中, 用最大化模量的方法可以得到四个社团, 而用极小化

有效性指标的方法仅得到两个社团, 即是一个前者的初级分区. 这样, 一些小规模的或者更高级分划中社团无法检测出来. 这是否可以通过设置 (5.27) 中的参数 λ 来调节分离程度和紧密程度的比例关系, 进而实现不同规模的社团检测有待于将来的验证. 由于算法中的迭代使用了变形 k-means, 故对于空手道俱乐部这个算例中, 节点 10 仍然被误分区.

对于基于最优预测的概率性分区算法, 例如算法 4.5 (AIP), 第四章中的数值实验表明其计算结构非常令人满意, 尤其是对于空手道俱乐部网络的算例, 由概率性分区结果根据多数决定原则所得到的确定性分区时, 节点 10 被正确分区. 但是 AIP 的初始区是用变形 k-means 所得到的确定性分区, 即一个示性矩阵, 对于固定的 N, 变形 k-means 大约执行 1000 次试验以避免陷入局部极小值, 从而增加了算法总的计算量; 并且和变形 k-means 的情形一样, 社团个数 N 必须作为已知的模型参数, 这是算法的局限所在. 在第六章的 6.2 中解决了这个问题, 即提出了传统模量的一个扩展形式, 模糊模量, 进而实现了确定性分区的自动模型选择. 算法 6.2 成功地克服了 AIP 的两个缺陷, 此时初始的模糊分区 $\{\rho_k^{(0)}\}$ 可以随机选取, 并且社团数目 N 可以自动被确定出来. 从数值结果中可见模糊模量具有一定的有效性, 但是由定义 (6.35) 知, 模糊模量仅利用了 ρ 的最大分量属于相同社团的节点对来定量化, 而扔掉了 ρ 中的其它信息, 不仅是最大分量的其它信息. ρ 的余下信息在当产生重叠社团时予以考虑. 如何构造出更合理的衡量概率性分区优良程度的函数还有待于进一步讨论和发展.

7.2 与其它方法的比较

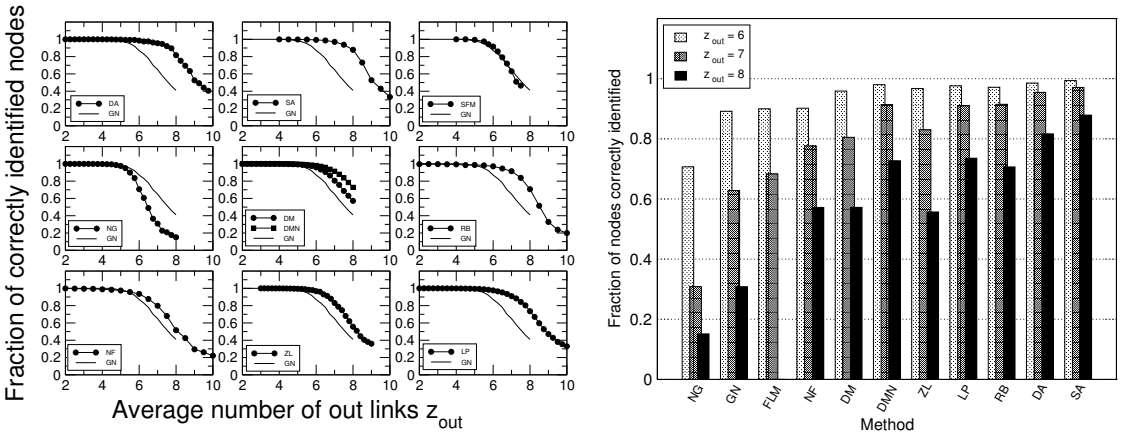
在 Danon 等人的分析[43]中, 比较了一些著名算法的计算量和算法应用于 ad hoc 网络的结果. 这些算法设计者的名字, 相关工作的出处, 表征算法的符号以及算法的计算复杂度如表 7.1 所示. 将这些算法应用于 ad hoc 网络中来考察算法精度的结果如图 7.1(a) 所示, 这里考察的是由不同方法得到的节点识别正确率随 z_{out} 的变化. 在三个特殊值 $z_{\text{out}} = 6, 7, 8$ 处的节点识别正确率如图 7.1(b) 所示, 注意到对于 FLM 算法 $z_{\text{out}} = 8$ 的数据无效. 可见多数方法当 z_{out} 的值上升到 6 时都可以很好地确定出正确的社团结构. 当 $z_{\text{out}} = 8$ 时一些方法开始“动摇”但仍可以正确

表 7.1: Danon 等人的比较分析[43]中所涉及到的算法的列表. 表中的四列分别显示了算法设计者的名字, 相关工作的出处, 表征算法的符号 (将在图 7.1 中使用) 以及算法的计算复杂度. 其中 n 表示节点数, m 表示边数, $\langle d \rangle$ 表示平均度.

作者	参考文献	算法标记	阶数
Eckmann & Moses	[63]	EM	$O(m\langle d^2 \rangle)$
Zhou & Lipowsky	[215]	ZL	$O(n^3)$
Latapy & Pons	[155]	LP	$O(n^3)$
Clauset 等	[39]	NF	$O(n \log^2 n)$
Newman & Girvan	[144]	NG	$O(nm^2)$
Girvan & Newman	[77]	GN	$O(n^2 m)$
Duch & Arenas	[57]	DA	$O(n^2 \log n)$
Fortunato 等	[72]	FLM	$O(m^3 n)$
Radicchi 等	[158]	RCCLP	$O(m^4/n^2)$
Donetti & Muñoz	[53, 54]	DM/DMN	$O(n^3)$
Bagrow & Bollt	[10]	BB	$O(n^3)$
Capocci 等	[26]	CSCC	$O(n^2)$
Wu & Huberman	[205]	WH	$O(n+m)$
Palla 等	[151]	PK	$O(\exp(n))$
Reichardt & Bornholdt	[163]	PB	参数独立
Guimerà 等	[82, 83]	SA	参数独立

识别半数以上的节点. 当 $z_{\text{out}} = 8$ 时仅有四个方法仍然能够识别出正确的社团结构. 而对于本文中第三章中的变形 k-means 算法, 第五章中的 SAVI 算法, 第六章 6.1 中的 SADI 和 SADD 算法, 其应用于 ad hoc 网络的结果优于 GN 算法, 并与 [43] 中的其它算法相比具有竞争力. 此外, 考察一个算法的优良程度不仅要考察其精度, 还要考虑其效率, 即计算量. 对于一个具有 n 个节点和 m 条边的网络, 表 7.1 中算法中最快速的仅需 $O(m+n)$, 而最慢的则需 $O(\exp(n))$. 而对于第三章中的变形 k-means 算法, 每次迭代的花费为 $O(N(n+m))$. 在规模增长的 ad hoc 网络中, 通常 500 个随机初始分划已经足够, 并并且对于每一个初始分区算法于几十步内收敛. 如果这些结果是一般性的, 那么变形 k-means 算法在 [43] 中比较的诸多方法中花费最少的算法之一, 然而这一点需要进一步的研究.

在 Lancichinetti 和 Fortunato的分析[109]中, 比较了一些著名算法的计算量和算法应用于 ad hoc 网络的结果. 这些算法设计者的名字, 相关工作的出处, 表征算法的符号以及算法的计算复杂度如表 7.2 所示. 将这些算法分别应用于 ad hoc 网

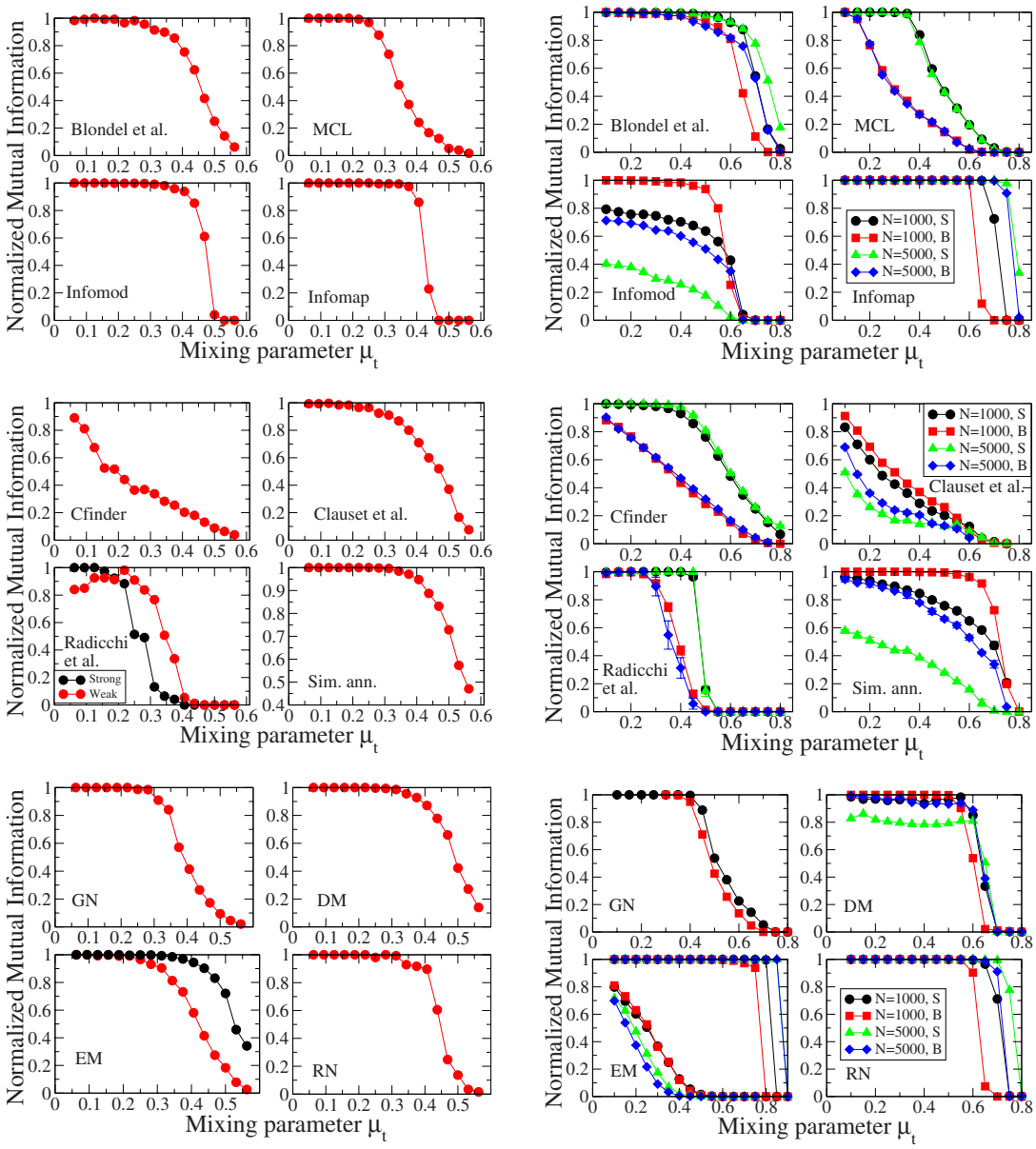


(a) (b)

图 7.1: (a) 利用 ad hoc 网络比较表 7.1 中算法的敏感度[43]. 这里考察的是由不同方法得到的节点识别正确率随 z_{out} 的变化. (b) 在三个特殊值 $z_{\text{out}} = 6, 7, 8$ 处的节点识别正确率[43]. 注意到对于 FLM 算法 $z_{\text{out}} = 8$ 的数据无效. 可见多数方法当 z_{out} 的值上升到 6 时都可以很好地确定出正确的社团结构. 当 $z_{\text{out}} = 8$ 时一些方法开始"动摇"但仍可以正确识别半数以上的节点. 当 $z_{\text{out}} = 8$ 时仅有四个方法仍然能够识别出正确的社团结构.

表 7.2: Lancichinetti 和 Fortunato的比较分析[109]中所涉及到的算法的列表. 表中的四列分别显示了算法设计者的名字, 相关工作的出处, 表征算法的符号 (将在图 7.2 中使用) 以及算法的计算复杂度.

作者	参考文献	算法标记	阶数
Girvan & Newman	[77, 144]	GN	$O(nm^2)$
Clauset 等	[39]	Clauset et al.	$O(n \log^2 n)$
Blondel 等	[20]	Blondel et al.	$O(m)$
Radicchi 等	[158]	Radicchi et al.	$O(m^4/n^2)$
Palla 等	[151]	Cfinder	$O(\exp(n))$
Van Dongen	[55]	MCL	$O(nk^2)$, $k < n$ 为参数
Donetti & Muñoz	[53, 54]	DM/DMN	$O(n^3)$
Ronhovde & Nussinov	[167]	RN	$O(m^\beta \log n)$, $\beta \sim 1.3$
Rosvall & Bergstrom	[170]	Infomap	$O(m)$
Rosvall & Bergstrom	[169]	Infomod	参数独立
Newman & Leicht	[145]	EM	参数独立
Guimerà 等	[82, 83]	Sim. Ann.	参数独立



(a) (b)

图 7.2: (a) 将表 7.2 中的算法测试于 ad hoc 网络[109]. (b) 将表 7.2 中的算法测试于无向无权的 LFR 网络[109]. 每个子图展现了算法所得到的分区与已知分区之间的归一化互信息随混合参数 μ_t 变化.

络和无向无权的 LFR 网络中来考察算法精度的结果如图 7.2 所示, 其中每个图展现了算法得到的分区与已知分区之间的归一化互信息随混合参数 μ_t 变化. 而对

于本文中第五章的 SAVI 算法, 也将其测试于 LFR 网络, 考察归一化互信息, 并与 Infomap 算法进行比较, 结果表明 SAVI 与 Infomap 相比非常具有竞争力, 特别是对于当混合参数 μ_t 很大时的较为扩散的情形. 对于第六章 6.2 中的 SAFM 算法, 将其分别应用于 ad hoc 网络和 LFR 网络, 考察归一化互信息, 结果表明 SAFM 算法效果良好, 与 Newman 快速算法[140] 和极值最优化算法[57]相比具有竞争力, 特别是对于 z_{out} 更高的更复杂的情形.

7.3 未来研究的展望

本文对无向的复杂网络的社团结构检测的讨论具有一般性, 可推广到具有类似结构的数据集中以及更复杂的网络中, 计算更实际的问题及更大的问题, 如 Internet, 生物网络, 交通网络以及社会学网络等等. 在后续的研究工作中, 将在以下几个方向深入下去.

统计解释. 本文关于复杂网络社团结构的研究基于动力学方法. 接下来希望从中寻找某种统计解释, 并在 Bayes 统计的框架下构造新的算法. 这部分可以参考的文献见第一章 1.4.4.

有向网络. 本文主要讨论的是无向网络中的社团结构. 无向网络的邻接矩阵为对称矩阵, 相应的转移矩阵满足关于其平稳分布的细致平衡条件, 因此 Markov 动力学具有许多良好性质, 有助于构造算法. 接下来将致力于把这些工作推广到有向网络中, 即推广到不一定满足细致平衡条件的更为一般的动力学中. 这个问题的相关文献见 [8, 84, 112, 145, 170].

层次结构. 本文给出的检测社团结构的算法均是初级分区, 即非层次结构. 文献中的多数算法也都是仅给出主要初级分区. 接下来将研究类似于分级聚类算法这样的能够识别出复杂网络中层次结构的算法. 这个问题的相关文献见 [37, 38, 172].

动态网络. 本文研究的是固定的复杂网络. 如何定义动态网络, 即随时间演化的网络中的社团结构, 并将其检测出来将是下一步的工作, 这个方向在实际中具有深远意义, 例如 Internet 网络等等. 这个问题的相关文献见 [9, 28, 29, 66, 91, 102, 115, 150].

实际应用. 将现有的算法应用到更大更实际的问题中去, 实现应用价值. 例如生物学网络, 包括细胞网络, 生态学网络, 蛋白质折叠, 神经网络等等; 社会学网络, 包括疾病传播网络, 科学文章引用网络等等; 以及一些经济学系统, 银行系统, 和流行的推荐系统等等. 这个问题的相关文献见 [26, 39, 70, 155].

总之, 我相信, 数学理论, 特别是复杂网络社团结构的动力学方法的相关理论, 将越来越多的渗透到物理学, 生物学, 计算机科学以及社会学等各个领域, 不但能够精确的刻画真实世界的现象, 而且一定也会给科学界带来崭新的思维.

附录 A　图论的基本要素

A.1　图中的基本定义

图 (graph) G 是一对集合 (V, E), 其中 V 是节点 (nodes or vertices) 集合, E 是 V^2 的子集, 这里 V^2 是 V 中元素的无序对构成的集合. E 中的元素称为边 (edges) 或连接 (links), 确定一条边得两个节点称为端点 (endpoints), 一条边邻接 (adjacent) 于它的每个端点. 若图的每条边都是节点的有序对则称为有向 (directed) 图, 此时有序对 (v, w) 是由 v 指向 w 的边, 或起点为 v 且终点为 w 的边. 如图 A.1 所示, 图表示为由线连接的点集. 在很多实际情况中, 图是加权的 (weighted), 即每条边上赋予一个实值. 图不包含环 (loops), 即连接节点与其自身的边; 也不包含多重边 (multiple edges), 即连接相同节点对有多条边. 具有环和多重边的图称为多重图 (multigraphs). 允许任意多个节点 (不仅限于两个) 之间存在边的广义的图称为超图 (hypergraphs).

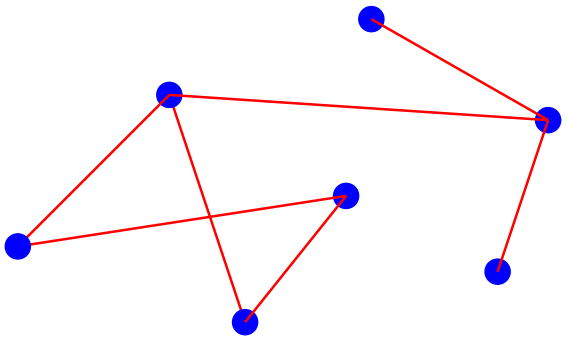


图 A.1: 由 7 个节点和 7 条边构成的简单图.

若 $V' = V$ 且 $E' = E$, 则图 $G' = (V', E')$ 称为图 $G = (V, E)$ 的子图 (subgraph). 若 G' 包含连接 V' 中节点并在 G 中的所有边, 则称子图 G' 是由 V' 诱导或生成的. 将节点集 V 分成两个子集 S 和 $V - S$ 的分划称为一个切割 (cut); G 中连接 S 中节点与 $V - S$ 中节点的边得数目称为切割规模 (cut size).

分别用 n 和 m 表示图中节点和边的数目, 称节点数目为图的阶 (order), 边的数目为图的规模 (size). 一个图的最大规模等于节点所有无序对的数目 $n(n-1)/2$. 若 $|V|=n$ 且 $|E|=m=n(n-1)/2$, 则称其为完全图 (complete graph, 或 clique), 记为 K_n. 若两个节点由边连接则称其为邻居 (neighbors) 或邻接的 (adjacent). 节点 v 的邻居的集合称为邻域 (neighborhood), 记为 $\Gamma(v)$. 节点 v 的邻居的数目称为节点的度 (degree), 记为 $d(v)$. 图中节点的度的序列 $d(v_1), d(v_2), \ldots, d(v_n)$ 称为度序列. 在有向图中, 节点 v 有两种类型的度: 入度 (in-degree), 即以 v 为起点的边的数目; 出度 (out-degree), 即以 v 为终点的边的数目. 加权图中的类似于度的量称为强度 (strength), 即邻接于节点的边上的权重之和. 图的一个重要的局部性质是传递性 (transitivity) 或聚集性 (clustering)[201], 它描述了一个节点邻居之间的凝聚程度①. 节点 v 的聚集系数 (clustering coefficient) $C(v)$ 定义为连接 v 的邻居对的边数与可能存在的边的总数 $d(v)(d(v)-1)/2$ 的比值, 它表示 v 的邻居对连接的概率. 由于 v 的所有邻居均与 v 相连接, 故连接 v 的邻居对的边与 v 形成三角形, 从而此定义通常以三角形数目的形式给出.

图 $P=(V(P), E(P))$ 称为路径 (path), 其中 $V(P)=\{x_0, x_1, \ldots, x_l\}$, $E(P)=\{x_0x_1, x_1x_2, \ldots, x_{l-1}x_l\}$. 节点 x_0 和 x_1 称为 P 的终端 (end-vertices), l 为其长度 (length). 若图的节点 (或边) 集合中任意两个元素非邻接, 则称其为独立的 (independent). 同理, 两条路径独立当它们仅共享终端. 节点和边均不不相同的封闭路径称为圈 (cycle). 长度为 l 的圈记为 c_l. 最小的非平凡圈为三角形 c_3.

若对于任意节点对, 至少存在一条由其中一个节点到另一个节点的路径, 则称图为连通的 (connected). 一般地, 连接两个节点存在长度不同的多条路径. 两个节点长度最短的路径称为最短路径 (shortest path, 或 geodesic), 这个最短的程度称为两个节点间的距离 (distance). 连通图中两个节点间距离的最大值称为直径 (diameter). 若两个节点之间不存在路径, 则图至少可分成两个连通子图, 每个最大连通子图称为连通分量 (connected component).

不存在圈的图称为森林 (forest), 连通的森林称为树 (tree). 从树的一个节点到另一个节点仅存在一条路径. 事实上, 如果相同的节点对间存在至少两条路径则将

①在一些学科中, 聚类性 (clustering) 通常被用来描述社团检测, 例如计算机科学, 并且在本文中也经常涉及. 为消除歧义, 本文特别使用聚集性以及聚集系数来指代一个节点邻域的局部性质.

形成一个圈, 而由定义知树是无圈图. 进一步, n 个节点的树具有 $n-1$ 条边. 如果树的任意一条边被移除, 则它分成了不连通的两部分; 如果加入一条新边, 则至少存在一个圈. 故树是给定阶时的最小连通图和最大无圈图. 每个连通图包含一个生成树 (spanning tree), 即共享图中所有节点的树. 在加权图中可定义最小 (最大) 生成树, 即边上权重之和最小 (最大) 的生成树. 最小和最大生成树在图最优化问题中经常使用.

若图 G 的节点集合 V 分成两个子集 V_1 和 V_2, 或类 (classes), 且每条边连接 V_1 中的一个节点和 V_2 中的一个节点, 则称 G 为二部的 (bipartite). 此定义可推广到 r-分割 (r-partition) 情形, 即节点类数为 r 且相同类中节点之间不存在边相连接, 这种图称为多部的 (multipartite).

A.2 图中的主要矩阵

一个 n 阶图的拓扑结构的全部信息都包含在邻接矩阵 (adjacency matrix) 中. 邻接矩阵 A 是一个 $n \times n$ 的矩阵, 当节点 i 和 j 有边连接时元素 a_{ij} 等于 1, 否则等于 0. 由于图中无环, 故邻接矩阵对角线元素均为 0. 对于无向图, 其邻接矩阵 A 是对称矩阵, 第 i 行或列的元素之和等于节点 i 的度. 对于加权图, 可以定义权重矩阵 (weighted matrix) W, 其元素 w_{ij} 表示连接节点 i 和 j 的边上的权重.

图 G 的邻接矩阵 A 的特征值的集合称为 G 的谱 (spectrum). 图矩阵的谱性质在图论研究中起着关键作用. 例如, 随机矩阵 (stochastic matrices) 决定了图中的扩散 (随机游动) 过程. 将 A 的第 i 行元素除以节点 i 的度得到右随机矩阵 (right stochastic matrix) R, R 的转置 T 为左随机矩阵 (left stochastic matrix). 随机矩阵的谱可估计随机游动的混合时间, 即达到过程平稳分布的时间, 是通过计算左随机矩阵的相应于最大特征值的特征向量得到的.

另一个重要的矩阵是 Laplace 矩阵 $L = D - A$, 其中 D 是 d_{ii} 等于节点 i 的度 $d(i)$ 的对角矩阵. 矩阵 L 通常称为非归一化 Laplace 矩阵. 文献中经常使用的归一化 Laplace 矩阵[36]有两种主要形式: $L_{\text{sym}} = D^{-1/2}LD^{-1/2}$ 和 $L_{\text{rw}} = D^{-1}L = I - D^{-1}A = I - T$. 矩阵 L_{sym} 是对称的; L_{rw} 非对称但与图上的随机游动紧密联系. 所有的 Laplace 矩阵均可直接推广到加权图的情形. Laplace 矩阵是最值得研究的

矩阵之一并且在诸多领域有广泛应用, 如图连通性[24], 同步性[14, 147], 扩散[36]和图形分割[156]. 由 Laplace 矩阵 (归一化或非归一化) 的构造可知, 每行元素之和为 0. 这意味着 L 总是具有至少一个 0 特征值, 相应的特征向量具有全部相等的分量, 如 $(1, 1, \ldots, 1)$. 相应于不同特征值的特征向量相互正交. 有趣的是, L 具有的 0 特征值的数目等于图中连通分量的数目, 故连通图的 Laplace 矩阵仅有一个 0 特征值而其它特征值均为正. Laplace 矩阵的特征向量经常用于谱方法 (见 1.4.1). 特别地, 相应于第二小特征值的特征向量, 称为 Fiedler 向量 (Fiedler vector)[67], 在图平分法中发挥着作用(见 1.4.1).

A.3 图中的主要模型

本节将介绍最流行的图模型来描述真实系统. 这些图是社团结构检测中的有用的空模型 (null model), 由于它们不具有社团结构, 故可以被用作分区算法的否定测试.

最古老的模型是由 Erdös 和 Rényi 提出的随机图 (random graph)[64]. 它具有两个参数: 节点数目 n 和连接概率 p. 每对节点以概率 p 相连接并独立于其它节点对. 图中边数的期望为 $pn(n-1)/2$, 且平均度得期望为 $\langle d\rangle = p(n-1)$. 随机图的节点的度分布是二项分布, 并且对于固定的 $\langle d\rangle$, 当 $n \to \infty$, $p \to 0$ 时, 节点分布收敛于 Poisson 分布. 因此所有节点有相同的度, 近似为 $\langle d\rangle$, 如图 A.2(a) 所示. 随机图最显著的性质是当 $n \to \infty$ 时变化 $\langle d\rangle$ 观察到得相变. 对于 $\langle d\rangle < 1$, 图分为若干连通分量, 每一个都是微观的, 占据了相对系统规模近乎消失的部分; 对于 $\langle d\rangle > 1$, 其中一个分量变得宏观, 占据了图中节点有限的部分.

具有 n 个节点的随机图的直径很小, 与 n 成对数增长. 这个小世界效应性质 (small world effect) 在许多真实图中普遍存在. 社会学家 Stanley Milgram 通过一系列著名的实验首次证明了社会网络具有长度很短的路径的特征[135]. 随机图节点聚集系数的期望为 p, 这是由于两个节点连接的概率等于它们是否为同一个节点的邻居的概率. 真实的图与相同大小的随机图相比具有更大的聚集系数. Watts 和 Strogatz 指出小世界性质和高聚集系数可在同一系统中共存[201]. 他们设计了一类由规则点阵填补得到的图, 具有高聚集系数, 和一个随机图, 具有小世界性质. 从

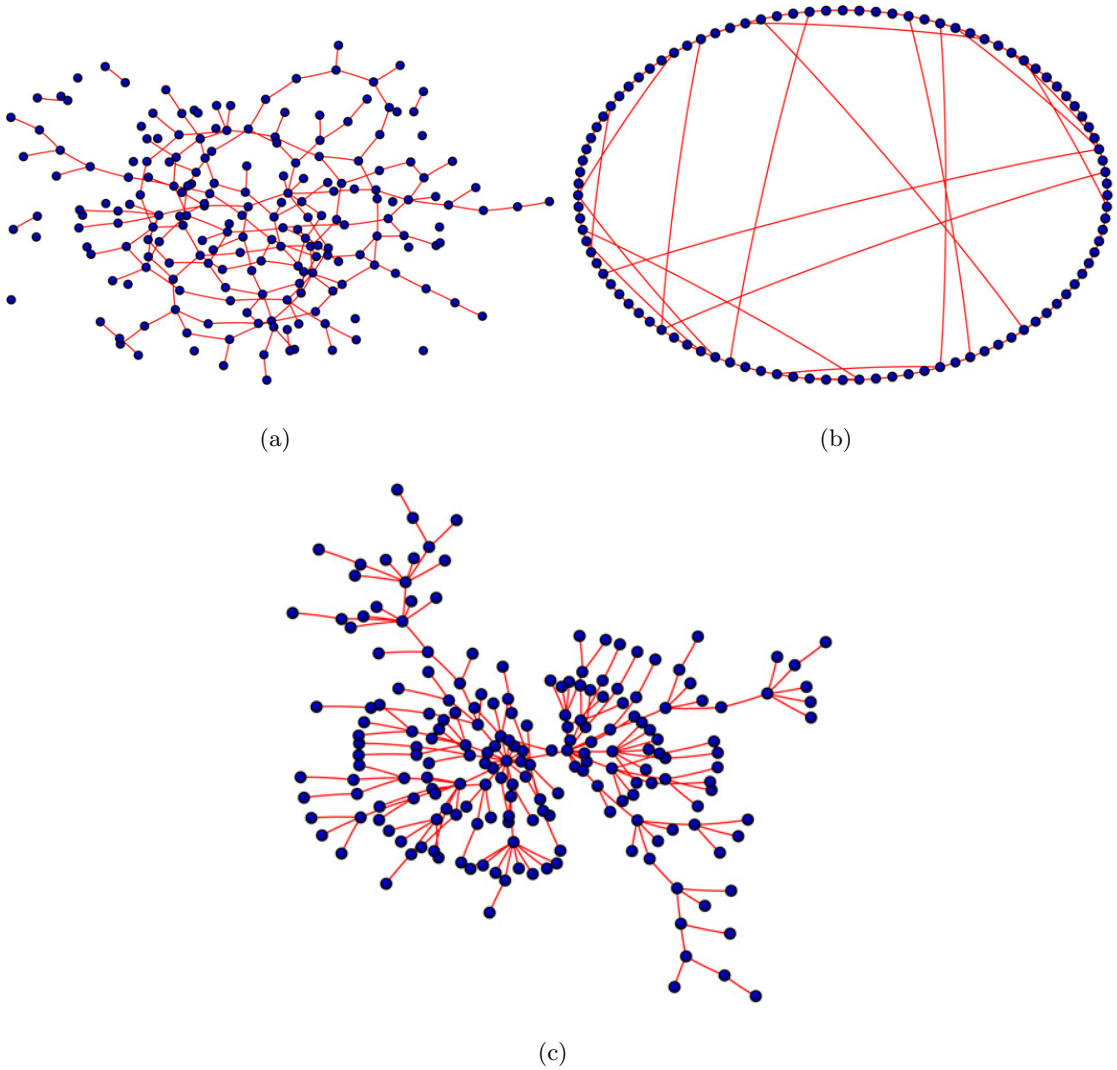


图 A.2: 复杂网络中的基本模型. (a) Erdös-Rényi 随机图, 其中节点数 $n = 100$, 连接概率 $p = 0.02$. (b) Watts-Strogatz 小世界图, 其中节点数 $n = 100$, 再连接概率 $p = 0.1$. (c) Barabási-Albert 无标度网络, 其中节点数 $n = 100$, 平均度为 2.

每个节点的度为 d 的环形点阵开始, 向不同的目标节点以概率 p 重新连上边, 如图 A.2(b) 所示. 这表明低的 p 值有足够的能力减少节点间最短路径的长度, 因为重新连接的边充当了连接原图中相距遥远的区域中节点的捷径的角色. 另一方面, 聚集系数仍保持较高的值, 因为少数几条重新连接的边没有明显的扰乱图的局部结构, 这仍类似于原来的环形点阵. 对于 $p = 1$, 所有的边均重新连接并导致结构为

Erdös 和 Rényi 提出的随机图.

Watts 和 Strogatz 的开创性工作触发了用图表示真实系统方面的大量兴趣. 最重要的发现之一为真实图的节点度分布是非常不均匀的[5], 具有很少和很多邻居的节点可以共存. 许多情况下这个分部的后部可由幂律分布很好的近似②, 故称无标度网络. 这种度的不均匀性是真实网络很多显著特征的形成原因[70]. 最著名的带有幂律度分布的图模型要数 Barabási 和 Albert 提出的模型[12]. 这个图由动力过程创建, 节点逐一地加在初始中心上. 新节点与已存在节点之间连接边的概率与存在节点的度成正比. 这样, 具有较高度得节点有更大的概率被新节点选为邻居; 如果被选中, 它们的度进一步地增加, 故它们在将来更有可能被选中. 当节点数目趋于无穷时, 这种滚雪球策略生成的图的度分布具有指数 3 的幂律尾部. 图 A.2(c) 中展示了 Barabási-Albert (BA) 图的一个例子. BA 图的聚集系数随图的规模衰减, 并且比真实网络低的多. 此外, 在真实网络中发现度分布的幂律衰减的指数范围通常在 2 和 3 之间, 而 BA 模型服从一个固定的值. 无论如何, 关于 BA 模型的很多提炼以及大量不同模型已在后来被介绍来考虑如何更接近真实系统的特征[4, 16, 22, 138, 143].

②幂律分布并不是描述复杂网络性质的必然方式, 仅仅由于度分布的尾部比较平缓, 即横跨度的数量级. 它们可能并非精确地服从幂律分布.

附录 B　基于最优预测的确定性分区算法中的推导

B.1　方程 (3.29) 的推导

由 $\tilde{P}$ 的定义

$$\tilde{p}(x,y)=\sum_{k,l=1}^{N}\mathbf{1}_{S_k}(x)\hat{p}(S_k,S_l)\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)}, \tag{B.1}$$

其中

$$\hat{\mu}(S_k)=\sum_{x\in S_k}\mu(x). \tag{B.2}$$

定义函数

$$J_{\mathrm{d}}=\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}|p_t(x,y)-\tilde{p}(x,y)|^2, \tag{B.3}$$

从而希望对于所有的 $\hat{p}(S_k,S_l)$ 寻求 J_{d} 的最小值. 注意到

$$\begin{aligned}
\frac{\partial J_{\mathrm{d}}}{\partial\hat{p}(S_k,S_l)} &= 2\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}\Big(p_t(x,y)-\tilde{p}(x,y)\Big)\frac{\partial\tilde{p}(x,y)}{\partial\hat{p}(S_k,S_l)}\\
&= 2\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}\Big(p_t(x,y)-\sum_{m,n=1}^{N}\mathbf{1}_{S_m}(x)\mathbf{1}_{S_n}(y)\hat{p}(S_m,S_n)\frac{\mu(y)}{\hat{\mu}(S_n)}\Big)\\
&\quad\cdot\mathbf{1}_{S_k}(x)\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)}\\
&= 2\sum_{x\in S_k,y\in S_l}\frac{\mu(x)}{\mu(y)}\Big(p_t(x,y)-\hat{p}(S_k,S_l)\frac{\mu(y)}{\hat{\mu}(S_l)}\Big)\frac{\mu(y)}{\hat{\mu}(S_l)}\\
&= \frac{2}{\hat{\mu}(S_l)}\Big(\sum_{x\in S_k,y\in S_l}\mu(x)p_t(x,y)-\hat{p}(S_k,S_l)\hat{\mu}(S_k)\Big),
\end{aligned} \tag{B.4}$$

并由最优化条件 $\partial J_{\mathrm{d}}/\partial\hat{p}(S_k,S_l)=0$, 从而得到

$$\hat{p}^*(S_k,S_l)=\frac{1}{\hat{\mu}(S_k)}\sum_{x\in S_k,y\in S_l}\mu(x)p_t(x,y). \tag{B.5}$$

再将 P^t 的谱分解

$$p_t(x,y)=\sum_{j=0}^{n-1}\lambda_j^t\varphi_j(x)\varphi_j(y)\mu(y) \tag{B.6}$$

代入 (B.5), 得到

$$\begin{aligned}\hat{p}^*(S_k,S_l) &= \frac{1}{\hat{\mu}(S_k)}\sum_{x\in S_k,y\in S_l}\mu(x)\sum_{j=0}^{n-1}\lambda_j^t\varphi_j(x)\varphi_j(y)\mu(y)\\ &= \sum_{j=0}^{n-1}\lambda_j^t\frac{1}{\hat{\mu}(S_k)}\sum_{x\in S_k}\mu(x)\varphi_j(x)\sum_{y\in S_l}\varphi_j(y)\mu(y)\\ &= \sum_{j=0}^{n-1}\lambda_j^t\left(\frac{\sum_{x\in S_k}\mu(x)\varphi_j(x)}{\hat{\mu}(S_k)}\right)\left(\frac{\sum_{y\in S_l}\mu(y)\varphi_j(y)}{\hat{\mu}(S_l)}\right)\hat{\mu}(S_l)\\ &= \sum_{j=0}^{n-1}\lambda_j^t\hat{\varphi}_j(S_k)\hat{\varphi}_j(S_l)\hat{\mu}(S_l),\end{aligned} \tag{B.7}$$

其中令

$$\hat{\varphi}_j(S_k)=\frac{\sum_{x\in S_k}\mu(x)\varphi_j(x)}{\sum_{x\in S_k}\mu(x)},\quad k=1,\dots,N. \tag{B.8}$$

于是方程 (3.29) 得证.

B.2 方程 (3.35) 的推导

由 $\hat{P}^*$ 的定义

$$\hat{p}^*(S_k,S_l)=\frac{1}{\hat{\mu}(S_k)}\sum_{x\in S_k,y\in S_l}\mu(x)p_t(x,y), \tag{B.9}$$

以及 $\tilde{P}^*$ 的定义

$$\tilde{p}^*(x,y)=\sum_{k,l=1}^{N}\mathbf{1}_{S_k}(x)\hat{p}^*(S_k,S_l)\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)}, \tag{B.10}$$

可以得到

$$\begin{aligned}E^* &\equiv E(\tilde{P}^*)=\|P^t-\tilde{P}^*\|_\mu^2=\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}|p_t(x,y)-\tilde{p}^*(x,y)|^2\\ &= \sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}|p_t(x,y)|^2-2\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}p_t(x,y)\tilde{p}^*(x,y)+\sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}|\tilde{p}^*(x,y)|^2\end{aligned}$$

$$\equiv \quad P_1 - 2P_2 + P_3, \tag{B.11}$$

其中

$$\begin{aligned}
P_2 &= \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} p_t(x,y) \sum_{k,l=1}^{N} \mathbf{1}_{S_k}(x)\hat{p}^*(S_k,S_l)\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)} \\
&= \sum_{k,l=1}^{N} \frac{1}{\hat{\mu}(S_l)}\hat{p}^*(S_k,S_l) \sum_{x\in S_k, y\in S_l} \mu(x)p_t(x,y) \\
&= \sum_{k,l=1}^{N} \frac{\hat{\mu}(S_k)}{\hat{\mu}(S_l)}|\hat{p}^*(S_k,S_l)|^2, \\
P_3 &= \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)} \sum_{k,l,m,n=1}^{N} \mathbf{1}_{S_k}(x)\hat{p}^*(S_k,S_l)\mathbf{1}_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)}\mathbf{1}_{S_m}(x)\hat{p}^*(S_m,S_n)\mathbf{1}_{S_n}(y)\frac{\mu(y)}{\hat{\mu}(S_n)} \\
&= \sum_{k,l=1}^{N} \frac{1}{\hat{\mu}(S_l)}\hat{p}^*(S_k,S_l) \sum_{x,y\in S} \mu(x)\mu(y)\mathbf{1}_{S_k}(x)\mathbf{1}_{S_l}(y) \sum_{m,n=1}^{N} \frac{1}{\hat{\mu}(S_n)}\hat{p}^*(S_m,S_n)\delta_{km}\delta_{ln} \\
&= \sum_{k,l=1}^{N} \frac{1}{\hat{\mu}(S_l)}\hat{p}^*(S_k,S_l)\frac{1}{\hat{\mu}(S_l)}\hat{p}^*(S_k,S_l) \sum_{x\in S_k}\mu(x) \sum_{y\in S_l}\mu(y) \\
&= \sum_{k,l=1}^{N} \frac{\hat{\mu}(S_k)}{\hat{\mu}(S_l)}|\hat{p}^*(S_k,S_l)|^2,
\end{aligned}$$

再将 P_2 和 P_3 代入 (B.11), 得到

$$E^* = \sum_{x,y\in S} \frac{\mu(x)}{\mu(y)}|p_t(x,y)|^2 - \sum_{k,l=1}^{N} \frac{\hat{\mu}(S_k)}{\hat{\mu}(S_l)}|\hat{p}^*(S_k,S_l)|^2 \equiv \|P^t\|_\mu^2 - \|\hat{P}^*\|_{\hat{\mu}}^2. \tag{B.12}$$

这就证明了 (3.35). 另一方面, 注意到

$$\begin{aligned}
P_1 &= \sum_{k,l=1}^{N} \sum_{x\in s_k, y\in S_l} \mu(x)\mu(y)|\frac{p_t(x,y)}{\mu(y)}|^2, \\
P_2 &= \sum_{k,l=1}^{N} \sum_{x\in s_k, y\in S_l} \mu(x)\mu(y)\frac{p_t(x,y)}{\mu(y)}\frac{\hat{p}^*(S_k,S_l)}{\hat{\mu}(S_l)}, \\
P_3 &= \sum_{k,l=1}^{N} \sum_{x\in s_k, y\in S_l} \mu(x)\mu(y)|\frac{\hat{p}^*(S_k,S_l)}{\hat{\mu}(S_l)}|^2,
\end{aligned}$$

从而 E^* 有另一种表达式

$$E^* = \sum_{k,l=1}^{N} \sum_{x\in S_k, y\in S_l} \mu(x)\mu(y)\left|\frac{p_t(x,y)}{\mu(y)} - \frac{\hat{p}^*(S_k,S_l)}{\hat{\mu}(S_l)}\right|^2. \tag{B.13}$$

B.3 算法 3.8 的计算量的估计

下面考虑算法 3.8 的计算量. 将 (3.53) 变换为如下形式

$$\begin{aligned}\bar{E}(x,S_k) &= \sum_{l=1}^{N}\sum_{y\in S_l}\mu(x)\mu(y)\left(\frac{p_t^2(x,y)}{\mu(y)}-2\frac{p_t(x,y)}{\mu(y)}\frac{\hat{p}^*(S_k,S_l)}{\hat{\mu}(S_l)}+\frac{\hat{p}^{*2}(S_k,S_l)}{\hat{\mu}^2(S_l)}\right)\\ &= \sum_{l=1}^{N}\sum_{y\in S_l}\frac{\mu(x)}{\mu(y)}p_t^2(x,y)+\mu(x)\sum_{l=1}^{N}\frac{\hat{p}^{*2}(S_k,S_l)}{\hat{\mu}(S_l)}\\ &\quad -2\mu(x)\sum_{l=1}^{N}\sum_{y\in S_l}p_t(x,y)\frac{\hat{p}^*(S_k,S_l)}{\hat{\mu}(S_l)}\\ &\equiv P_1+P_2-P_3. \end{aligned}\tag{B.14}$$

其中 n 是网络中的节点数, m 是边数, N 是社团数目. $\{\hat{\mu}(S_k)\}_{k=1}^N$ 的计算量为 $O(n)$. $\hat{p}^*(S_k,S_l)$ 为 $O(d_{kl})$, 其中 d_{kl} 为从第 k 个社团到第 l 个社团的连接的度, 故 $\{\hat{p}^*(S_k,S_l)\}_{k,l=1}^N$ 的计算量至多为 $O(m)$, 并且实际上远小于 m, 因为不需考虑社团内部的连接.

对于固定的 x 和 k, P_1 和 P_3 的计算量均为 $O(d(x))$. 注意到 P_2 中的求和部分, 对于所有的 k, 可以以 $O(N^2)$ 的计算量预计算, 故对于所有的 x 和 k, P_2 的计算量为 $O(Nn+N^2)$. 因此, 对于所有的 x 和 k, $\bar{E}(x,S_k)$ 的计算量为 $O(N(2m+n)+N^2+n+m)$. 考虑到对于现实中的网络, $N\ll n$, 且 m 为 $O(n)$ 的, 故得到每个迭代步的计算花费为 $O(N(m+n))$.

如果考虑平均迭代数 k_1, 以及试验次数 k_2, 于是得到算法最终的计算量为 $O(k_1k_2N(m+n))$.

附录 C　基于最优预测的概率性分区算法中的推导

C.1　引理 4.3 的证明

为了推导问题 (4.11) 的 Euler-Lagrange 方程组，首先对 J 关于 $\hat{p}_{kl}$ 求导，得到

$$\begin{aligned}\frac{\partial J}{\partial \hat{p}_{kl}} &= -2\sum_{x,y\in S}\mu(x)\mu(y)\left(\sum_{m,n=1}^{N}\rho_m(x)\rho_n(y)\Big(\frac{p(x,y)}{\mu(y)}-\frac{\hat{p}_{mn}}{\hat{\mu}_n}\Big)\right)\\ &\quad\cdot\left(\sum_{s,t=1}^{N}\rho_s(x)\rho_t(y)\frac{1}{\hat{\mu}_t}\delta_{ks}\delta_{lt}\right)=0.\end{aligned}\tag{C.1}$$

经过适当的计算得到

$$\begin{aligned}\sum_{x,y\in S}\sum_{m,n=1}^{N}\mu(x)\mu(y)\rho_m(x)\rho_n(y)\frac{\hat{p}_{mn}}{\hat{\mu}_n}\rho_k(x)\rho_l(y) &= \sum_{x,y\in S}\mu(x)p(x,y)\rho_k(x)\rho_l(y)\\ &= \hat{\mu}_k\hat{p}^{*}_{kl}.\end{aligned}\tag{C.2}$$

将上述结果用矩阵形式表示即可得到方程 (4.16a).

现在对于 J 在归一化条件 $\sum_{m=1}^{N}\rho_m(x)=1$ 下关于 $\rho_r(z)$ 求导. 定义带有 Lagrange 乘子 $\lambda(x)$ 的扩展的目标函数

$$\tilde{J}=J+\sum_{x\in S}\lambda(x)\Big(\sum_{m=1}^{N}\rho_m(x)-1\Big).\tag{C.3}$$

对 $\tilde{J}$ 关于 $\rho_r(z)$ 求导，得到

$$\begin{aligned}&\sum_{y\in S}\sum_{k,l=1}^{N}\sum_{n=1}^{N}\mu(y)\rho_k(z)\rho_l(y)\rho_n(y)\Big(\frac{p(z,y)}{\mu(y)}-\frac{\hat{p}_{kl}}{\hat{\mu}_l}\Big)\Big(\frac{p(z,y)}{\mu(y)}-\frac{\hat{p}_{rn}}{\hat{\mu}_n}\Big)\\ +&\sum_{y\in S}\sum_{k,l=1}^{N}\sum_{n=1}^{N}\mu(y)\rho_k(z)\rho_l(y)\rho_n(y)\Big(\frac{p(y,z)}{\mu(z)}-\frac{\hat{p}_{lk}}{\hat{\mu}_k}\Big)\Big(\frac{p(y,z)}{\mu(z)}-\frac{\hat{p}_{nr}}{\hat{\mu}_r}\Big)\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{x,y\in S}\sum_{k,l=1}^{N}\sum_{n=1}^{N}\mu(x)\mu(y)\rho_k(x)\rho_l(y)\rho_n(x)\rho_r(y)\Big(\frac{p(x,y)}{\mu(y)}-\frac{\hat{p}_{kl}}{\hat{\mu}_l}\Big)\frac{\hat{p}_{nr}}{\hat{\mu}_r^2}\\
&= -\frac{\lambda(z)}{2\mu(z)}.
\end{aligned}
\tag{C.4}
$$

将上述形式简单记为

$$
P_1+P_2+P_3=-\frac{\lambda(z)}{2\mu(z)}. \tag{C.5}
$$

并有

$$
P_3=\sum_{n=1}^{N}\hat{p}^*_{nr}\hat{\mu}_n\frac{\hat{p}_{nr}}{\hat{\mu}_r^2}-\sum_{k,l=1}^{N}\sum_{n=1}^{N}\hat{\hat{\mu}}_{nk}\hat{\hat{\mu}}_{lr}\frac{\hat{p}_{kl}}{\hat{\mu}_l}\frac{\hat{p}_{nr}}{\hat{\mu}_r^2}. \tag{C.6}
$$

由已经推导出的 (4.16a), 得到 $P_3=0$! 此外, 有

$$
\begin{aligned}
P_1 &= \mathbf{1}_{N\times 1}\cdot \mathrm{diag}_{mv}(p^2\cdot I_\mu^{-1}-p\cdot\rho^T\cdot I_{\hat{\mu}}^{-1}\cdot\hat{p}^T\cdot\rho)\\
&\quad -\hat{p}\cdot I_{\hat{\mu}}^{-1}\cdot\rho\cdot p^T+\hat{p}\cdot I_{\hat{\mu}}^{-1}\cdot\hat{\hat{\mu}}\cdot I_{\hat{\mu}}^{-1}\cdot\hat{p}^T\cdot\rho. \qquad (C.7)\\
P_2 &= \mathbf{1}_{N\times 1}\cdot \mathrm{diag}_{mv}(p^2\cdot I_\mu^{-1}-p\cdot\rho^T\cdot\hat{p}\cdot I_{\hat{\mu}}^{-1}\cdot\rho)\\
&\quad -I_{\hat{\mu}}^{-1}\cdot\hat{p}^T\cdot\rho\cdot p^T+I_{\hat{\mu}}^{-1}\cdot\hat{p}^T\cdot\hat{\hat{\mu}}\cdot\hat{p}\cdot I_{\hat{\mu}}^{-1}\cdot\rho. \qquad (C.8)
\end{aligned}
$$

这里符号 $\mathrm{diag}_{mv}(A)$ 是将矩阵对角线元素记成向量的算子. 由 $\hat{p}$ 满足的关于 $\hat{\mu}$ 的细致平衡条件 (4.19) 可得

$$
\begin{aligned}
P_1=P_2 &= \mathbf{1}_{N\times 1}\cdot \mathrm{diag}_{mv}(p^2\cdot I_\mu^{-1}-p\cdot\rho^T\cdot\hat{p}\cdot I_{\hat{\mu}}^{-1}\cdot\rho)\\
&\quad -\hat{p}\cdot I_{\hat{\mu}}^{-1}\cdot\rho\cdot p^T+\hat{p}\cdot I_{\hat{\mu}}^{-1}\cdot\hat{\hat{\mu}}\cdot\hat{p}\cdot I_{\hat{\mu}}^{-1}\cdot\rho.
\end{aligned}
\tag{C.9}
$$

通过适当的计算得到

$$
\rho=-\hat{\mu}\cdot\left[\mathrm{diag}_{mv}(p^2I_\mu^{-1}-p\rho^T\hat{p}I_{\hat{\mu}}^{-1}\rho)+\frac{1}{2}\mathrm{diag}_{mv}\Big(I_\lambda I_\mu^{-1}\Big)\right]+I_{\hat{\mu}}\hat{p}^{-1}\hat{\hat{\mu}}^{-1}\rho p^T. \tag{C.10}
$$

根据 ρ 的归一化条件, 得到 Lagrange 乘子

$$
\lambda(z)=\mu(z)\sum_{y\in S}\sum_{k,l=1}^{N}\rho_k(z)\rho_l(y)p(z,y)\frac{\hat{p}_{kl}}{\hat{\mu}_l}-\sum_{y\in S}p(z,y)p(y,z). \tag{C.11}
$$

将 (C.11) 代入 (C.10), 最终得到了关于 ρ 的表达式 (4.16b).

C.2　引理 4.6 的证明

首先, 将极小化问题 (4.11) 中的 J 关于 $\hat{p}_{kl}$ 求偏导数, 得到

$$\begin{aligned}
\frac{\partial J}{\partial \hat{p}_{kl}} &= 2\sum_{x,y\in S}\mu(x)\mu(y)\left(\sum_{m,n=1}^{K}\rho_m(x)\rho_n(y)\frac{\hat{p}_{mn}}{\hat{\mu}_n}-\frac{p(x,y)}{\mu(y)}\right)\sum_{s,t=1}^{N}\rho_s(x)\rho_t(y)\frac{\delta_{sk}\delta_{tl}}{\hat{\mu}_t} \\
&= 2\Bigg(\frac{1}{\hat{\mu}_l}\sum_{x,y\in S}\sum_{m,n=1}^{N}\mu(x)\mu(y)\rho_k(x)\rho_l(y)\rho_m(x)\rho_n(y)\frac{\hat{p}_{mn}}{\hat{\mu}_n} \\
&\quad -\frac{1}{\hat{\mu}_l}\sum_{x,y\in S}\mu(x)p(x,y)\rho_k(x)\rho_l(y)\Bigg) \\
&= 2\left(\frac{1}{\hat{\mu}_l}\sum_{m,n=1}^{N}\hat{\hat{\mu}}_{km}\frac{\hat{p}_{mn}}{\hat{\mu}_n}\hat{\hat{\mu}}_{nl}-\frac{\hat{\mu}_k}{\hat{\mu}_l}\hat{p}^*_{kl}\right)
\end{aligned} \tag{C.12}$$

将上述方程写成矩阵形式则得到 (4.23a).

接下来将对 J 关于 $\rho_r(z)$ 求偏导数, 得到

$$\begin{aligned}
\frac{\partial J}{\partial \rho_r(z)} &= 2\sum_{x,y\in S}\mu(x)\mu(y)\left(\sum_{m,n=1}^{K}\rho_m(x)\rho_n(y)\frac{\hat{p}_{mn}}{\hat{\mu}_n}-\frac{p(x,y)}{\mu(y)}\right) \\
&\quad \cdot\sum_{k,l=1}^{K}\Bigg[\delta_{kr}\delta(x,z)\rho_l(y)\frac{\hat{p}_{kl}}{\hat{\mu}_l}+\delta_{lr}\delta(y,z)\rho_k(x)\frac{\hat{p}_{kl}}{\hat{\mu}_l} \\
&\quad -\rho_k(x)\rho_l(y)\frac{\hat{p}_{kl}}{\hat{\mu}_l^2}\sum_{w\in S}\delta_{lr}\delta(w,z)\mu(w)\Bigg],
\end{aligned} \tag{C.13}$$

利用细致平衡条件 (4.3) 和 $\hat{p}^*$ 的定义 (4.13), 得到

$$\begin{aligned}
\frac{\partial J}{\partial \rho_r(z)} &= 2\Bigg[\sum_{y\in S}\mu(z)\mu(y)\left(\sum_{m,n=1}^{N}\rho_m(z)\rho_n(y)\frac{\hat{p}_{mn}}{\hat{\mu}_n}-\frac{p(z,y)}{\mu(y)}\right)\cdot\sum_{k=1}^{N}\rho_l(y)\frac{\hat{p}_{rl}}{\hat{\mu}_l} \\
&\quad +\sum_{x\in S}\mu(z)\mu(x)\left(\sum_{m,n=1}^{N}\rho_m(x)\rho_n(z)\frac{\hat{p}_{mn}}{\hat{\mu}_n}-\frac{p(x,z)}{\mu(z)}\right)\cdot\sum_{k=1}^{N}\rho_k(x)\frac{\hat{p}_{kr}}{\hat{\mu}_r} \\
&\quad -\sum_{x,y\in S}\mu(x)\mu(y)\left(\sum_{m,n=1}^{N}\rho_m(x)\rho_n(y)\frac{\hat{p}_{mn}}{\hat{\mu}_n}-\frac{p(x,y)}{\mu(y)}\right) \\
&\quad \cdot\sum_{k=1}^{N}\mu(z)\rho_k(x)\rho_r(y)\frac{\hat{p}_{kr}}{\hat{\mu}_r^2}\Bigg] \\
&= 2\left[\sum_{l,m,n=1}^{N}\hat{\hat{\mu}}_{ln}\rho_l(y)\frac{\hat{p}_{mn}}{\mu_n}\frac{\hat{p}_{rl}}{\hat{\mu}_l}-\sum_{y\in S}\sum_{l=1}^{N}p(z,y)\rho_l(y)\frac{\hat{p}_{rl}}{\hat{\mu}_l}\right.
\end{aligned}$$

$$+\sum_{k,m,n=1}^{N}\hat{\hat{\mu}}_{km}\rho_k(x)\frac{\hat{p}_{mn}}{\hat{\mu}_n}\frac{\hat{p}_{kr}}{\hat{\mu}_r}-\sum_{x\in S}\sum_{k=1}^{N}p(z,x)\rho_k(x)\frac{\hat{p}_{kr}}{\hat{\mu}_r}$$
$$-\sum_{k,m,n=1}^{N}\hat{\hat{\mu}}_{mk}\hat{\hat{\mu}}_{rn}\frac{\hat{p}_{mn}}{\hat{\mu}_n}\frac{\hat{p}_{kr}}{\hat{\mu}_r^2}+\sum_{k=1}^{N}\mu_k\hat{p}^*_{kr}\frac{\hat{p}_{kr}}{\hat{\mu}_r^2}\Bigg]\mu(z) \tag{C.14}$$

经适当的操作后最终得到 (4.23b).

C.3 引理 4.9 的证明

首先, 将 $\hat{p}_{kl}$ 的广义坐标 $\hat{p}_{kl}=\frac{e^{Y_{kl}}}{\sum_{m=1}^{N}e^{Y_{km}}}$ 代入 (4.11), 得到

$$J=\sum_{x,y\in S}\mu(x)\mu(y)\left(\frac{p(x,y)}{\mu(y)}-\sum_{k,l=1}^{N}\rho_k(x)\rho_l(y)\frac{e^{Y_{kl}}}{\sum_{m=1}^{N}e^{Y_{km}}}\frac{1}{\hat{\mu}_l}\right)^2,$$

将上述的 J 关于 Y_{kl} 求偏导数, 有

$$\begin{aligned}\frac{\partial J}{\partial Y_{kl}} &= 2\sum_{x,y\in S}\mu(x)\mu(y)\left(\sum_{s,t=1}^{N}\rho_s(x)\rho_t(y)\hat{p}_{st}\frac{1}{\hat{\mu}_t}-\frac{p(x,y)}{\mu(y)}\right)\\ &\quad\cdot\left(\sum_{m,n=1}^{N}\rho_m(x)\rho_n(y)\frac{1}{\hat{\mu}_n}\frac{e^{Y_{mn}}\delta_{mk}\delta_{nl}}{\sum_{r=1}^{N}e^{Y_{mr}}}\right.\\ &\quad\left.-\sum_{m,n=1}^{N}\rho_m(x)\rho_n(y)\frac{1}{\hat{\mu}_n}\frac{e^{Y_{mn}}}{(\sum_{r=1}^{N}e^{Y_{mr}})^2}\sum_{r=1}^{N}e^{mr}\delta_{mk}\delta_{rl}\right)\\ &= 2\sum_{x,y\in S}\mu(x)\mu(y)\left(\sum_{s,t=1}^{N}\rho_s(x)\rho_t(y)\hat{p}_{st}\frac{1}{\hat{\mu}_t}-\frac{p(x,y)}{\mu(y)}\right)\\ &\quad\cdot\left(\rho_k(x)\rho_l(y)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}-\sum_{n=1}^{N}\rho_k(x)\rho_n(y)\frac{1}{\hat{\mu}_n}\hat{p}_{kl}\hat{p}_{kn}\right)\\ &\equiv 2\left(P_1-P_2-P_3+P_4\right),\end{aligned}$$

将 $\hat{\hat{\mu}}$ 的定义 (4.14), $\hat{p}^*$ 的定义 (4.13) 以及细致平衡条件 (4.3) 代入上述 $P_i\,(1\le i\le 4)$ 中, 有

$$\begin{aligned}P_1 &= \sum_{x,y\in S}\sum_{s,t=1}^{N}\mu(x)\mu(y)\rho_s(x)\rho_t(y)\hat{p}_{st}\frac{1}{\hat{\mu}_t}\rho_k(x)\rho_l(y)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\\ &= \sum_{s,t=1}^{N}\hat{\hat{\mu}}_{ks}\hat{p}_{st}\frac{1}{\hat{\mu}_t}\hat{\hat{\mu}}_{tl}\frac{1}{\hat{\mu}_l}\cdot\hat{p}_{kl},\end{aligned}$$

$$
\begin{aligned}
P_2 &= \sum_{x,y\in S} \mu(x)p(x,y)\rho_k(x)\rho_l(y)\hat{p}_{kl}\frac{1}{\hat{\mu}_l} \\
&= (\hat{p}^*)^T_{kl}\cdot\hat{p}_{kl}, \\
P_3 &= \sum_{x,y\in S}\sum_{s,t,n=1}^{N} \mu(x)\mu(y)\rho_s(x)\rho_t(y)\hat{p}_{st}\frac{1}{\hat{\mu}_t}\rho_k(x)\rho_n(y)\frac{1}{\hat{\mu}_n}\hat{p}_{kl}\hat{p}_{kn} \\
&= \sum_{s,t,n=1}^{N} \hat{\hat{\mu}}_{ks}\hat{p}_{st}\frac{1}{\hat{\mu}_t}\hat{\hat{\mu}}_{tn}\frac{1}{\hat{\mu}_n}\hat{p}^T_{nk}\cdot\hat{p}_{kl} \\
P_4 &= \sum_{x,y\in S}\sum_{n=1}^{N} \mu(x)\mu(y)\rho_k(x)\rho_n(y)\frac{1}{\hat{\mu}_n}\hat{p}_{kl}\hat{p}_{kn} \\
&= \sum_{n=1}^{N}(\hat{p}^*)^T_{kn}\cdot\hat{p}^T_{nk}\cdot\hat{p}_{kl}.
\end{aligned}
$$

写成矩阵的形式, 有

$$
\begin{aligned}
\frac{\partial J}{\partial Y} &= 2\Bigg[\left(\hat{\hat{\mu}}\hat{p}I^{-1}_{\hat{\mu}}\hat{\hat{\mu}}I^{-1}_{\hat{\mu}}\right)*\hat{p}-(\hat{p}^*)^T*\hat{p} \\
&\qquad -\mathrm{diag}\left(\hat{\hat{\mu}}\hat{p}I^{-1}_{\hat{\mu}}\hat{\hat{\mu}}I^{-1}_{\hat{\mu}}\cdot\hat{p}^T\right)\cdot\hat{p}+\mathrm{diag}\left((\hat{p}^*)^T\cdot\hat{p}^T\right)\cdot\hat{p}\Bigg], \qquad \text{(C.15)}
\end{aligned}
$$

其中 $*$ 表示矩阵对应元素相乘的运算, $\mathrm{diag}(A)$ 是矩阵 A 的对角线部分. 这就得到了 (4.29a).

接下来, 将 $\rho_k(x)$ 的广义坐标 $\rho_k(x)=\frac{e^{Z_k(x)}}{\sum_{m=1}^{N}e^{Z_m(x)}}$ 代入 (4.11), 得到

$$
J=\sum_{x,y\in S}\mu(x)\mu(y)\left(\frac{p(x,y)}{\mu(y)}-\sum_{k,l=1}^{N}\hat{p}_{kl}\frac{e^{Z_k(x)}}{\sum_m e^{Z_m(x)}}\frac{e^{Z_l(y)}}{\sum_m e^{Z_m(y)}}\frac{1}{\sum_{z\in S}\mu(z)\frac{e^{Z_l(z)}}{\sum_m e^{Z_m(z)}}}\right)^2,
$$

将上述的 J 关于 Y_{kl} 求偏导数, 有

$$
\begin{aligned}
\frac{\partial J}{\partial Z_r(z)} &= 2\sum_{x,y\in S}\mu(x)\mu(y)\left(\sum_{k,l=1}^{N}\rho_k(x)\rho_l(y)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}-\frac{p(x,y)}{\mu(y)}\right) \\
&\cdot\left(\sum_{s,t=1}^{N}\rho_t(y)\frac{\hat{p}_{st}}{\hat{\mu}_t}\frac{e^{Z_s(x)}}{\sum_m e^{Z_m(x)}}\delta_{rs}\delta(z,x)-\sum_{s,t=1}^{N}\rho_t(y)\frac{\hat{p}_{st}}{\hat{\mu}_t}\frac{e^{Z_s(x)}}{(\sum_m e^{Z_m(x)})^2}\sum_{n=1}^{N}e^{Z_n(x)}\delta_{rn}\delta(z,x)\right. \\
&\left.+\sum_{s,t=1}^{N}\rho_s(x)\frac{\hat{p}_{st}}{\hat{\mu}_t}\frac{e^{Z_t(y)}}{\sum_m e^{Z_m(y)}}\delta_{rt}\delta(z,y)-\sum_{s,t=1}^{N}\rho_s(x)\frac{\hat{p}_{st}}{\hat{\mu}_t}\frac{e^{Z_t(y)}}{(\sum_m e^{Z_m(y)})^2}\sum_{n=1}^{N}e^{Z_n(y)}\delta_{rn}\delta(z,y)\right.
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{s,t=1}^{N} \rho_s(x)\rho_t(y)\frac{\hat{p}_{st}}{\hat{\mu}_t^2}\sum_{u\in S}\mu(u)\frac{e^{Z_t(u)}}{\sum_m e^{Z_m(u)}}\delta_{rt}\delta(z,u) \\
&+ \sum_{s,t=1}^{N} \rho_s(x)\rho_t(y)\frac{\hat{p}_{st}}{\hat{\mu}_t^2}\sum_{u\in S}\mu(u)\frac{e^{Z_t(u)}}{(\sum_m e^{Z_m(u)})^2}\sum_{n=1}^{N} e^{Z_n(u)}\delta_{rn}\delta(z,u)\Bigg) \\
&\qquad \equiv \; 2\left(P_1 - P_2 - P_3 + P_4 + P_5 - P_6 - P_7 + P_8 - P_9 + P_{10} + P_{11} - P_{12}\right),
\end{aligned}
$$

将 $\hat{\hat{\mu}}$ 的定义 (4.14), $\hat{p}^*$ 的定义 (4.13) 以及细致平衡条件 (4.3) 代入上述 P_i $(1 \le i \le 12)$ 中, 有

$$
\begin{aligned}
P_1 &= \sum_{y\in S}\sum_{k,l,t=1}^{N} \mu(z)\mu(y)\rho_k(x)\rho_l(y)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\rho_t(y)\hat{p}_{rt}\frac{1}{\hat{\mu}_t}\rho_r(z) \\
&= \sum_{k,l,t=1}^{N} \hat{p}_{rt}\frac{1}{\hat{\mu}_t}\hat{\hat{\mu}}_{tl}\frac{1}{\hat{\mu}_l}\hat{p}_{lk}^T\rho_k(z)\cdot\rho_r(z)\cdot\mu(z), \\
P_2 &= \sum_{y\in S}\sum_{t=1}^{N} \mu(z)p(z,y)\rho_t(y)\hat{p}_{rt}\frac{1}{\hat{\mu}_t}\rho_r(z) \\
&= \sum_{y\in S}\sum_{t=1}^{N} \hat{p}_{rt}\frac{1}{\hat{\mu}_t}\rho_t(y)p^T(y,z)\cdot\rho_r(z)\cdot\mu(z), \\
P_3 &= \sum_{y\in S}\sum_{k,l,s,t=1}^{N} \mu(z)\mu(y)\rho_k(z)\rho_l(y)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\rho_t(y)\rho_s(z)\rho_r(z)\hat{p}_{st}\frac{1}{\hat{\mu}_t} \\
&= \sum_{k,l,s,t=1}^{N} \rho_r(z)\cdot\left(\rho_k^T(z)\cdot\left(\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\hat{\hat{\mu}}_{lt}\frac{1}{\hat{\mu}_t}\hat{p}_{ts}^T\rho_s(z)\right)\right)\cdot\mu(z), \\
P_4 &= \sum_{y\in S}\sum_{s,t=1}^{N} \mu(z)p(z,y)\rho_t(y)\hat{p}_{st}\frac{1}{\hat{\mu}_t}\rho_s(z)\rho_r(z) \\
&= \sum_{y\in S}\sum_{s,t=1}^{N} \rho_z(r)\cdot\left(\rho_s^T(z)\cdot\left(\hat{p}_{st}\frac{1}{\hat{\mu}_t}\rho_t(y)p^T(y,z)\right)\right)\cdot\mu(z), \\
P_5 &= \sum_{x\in S}\sum_{k,l,s=1}^{N} \mu(x)\mu(z)\rho_k(z)\rho_l(z)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\rho_s(x)\rho_r(z)\hat{p}_{sr}\frac{1}{\hat{\mu}_r} \\
&= \sum_{k,l,s=1}^{N} \frac{1}{\hat{\mu}_r}\hat{p}_{rs}^T\hat{\hat{\mu}}_{sk}\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\rho_l(z)\cdot\rho_r(z)\cdot\mu(z), \\
P_6 &= \sum_{x\in S}\sum_{s=1}^{N} \mu(x)p(x,z)\rho_s(x)\rho_r(z)\hat{p}_{sr}\frac{1}{\hat{\mu}_r}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{x\in S}\sum_{s=1}^{N}\frac{1}{\hat{\mu}_r}\hat{p}_{rs}^{T}\rho_s(x)p^{T}(x,z)\cdot\rho_r(z)\cdot\mu(z),\\
P_7 &= \sum_{x\in S}\sum_{k,l,s,t=1}^{N}\mu(x)\mu(z)\rho_k(x)\rho_l(z)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\rho_s(x)\rho_t(z)\rho_r(z)\hat{p}_{st}\frac{1}{\hat{\mu}_t}\\
&= \sum_{k,l,s,t=1}^{N}\rho_r(z)\cdot\left(\rho_l^{T}(z)\cdot\left(\frac{1}{\hat{\mu}_l}\hat{p}_{lk}^{T}\hat{\hat{\mu}}_{ks}\hat{p}_{st}\frac{1}{\hat{\mu}_t}\rho_t(z)\right)\right)\cdot\mu(z),\\
P_8 &= \sum_{x\in S}\sum_{s,t=1}^{N}\mu(x)p(x,z)\rho_s(x)\rho_t(z)\rho_r(z)\hat{p}_{st}\frac{1}{\hat{\mu}_t}\\
&= \sum_{x\in S}\sum_{s,t=1}^{N}\rho_r(z)\cdot\left(\rho_t^{T}(z)\cdot\left(\frac{1}{\hat{\mu}_t}\hat{p}_{ts}^{T}\rho_s(x)p^{T}(x,z)\right)\right)\cdot\mu(z),\\
P_9 &= \sum_{x,y\in S}\sum_{k,l,s=1}^{N}\mu(x)\mu(y)\rho_k(x)\rho_l(y)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\rho_s(x)\rho_r(y)\hat{p}_{sr}\frac{1}{\hat{\mu}_r^2}\mu(z)\rho_r(z)\\
&= \sum_{k,l,s=1}^{N}\frac{1}{\hat{\mu}_r^2}\hat{\hat{\mu}}_{rl}\frac{1}{\hat{\mu}_l}\hat{p}_{lk}^{T}\hat{\hat{\mu}}_{ks}\cdot\hat{p}_{sr}\cdot\rho_r(z)\cdot\mu(z),\\
P_{10} &= \sum_{x,y\in S}\sum_{s=1}^{N}\mu(x)p(x,y)\rho_s(x)\rho_r(y)\hat{p}_{sr}\frac{1}{\hat{\mu}_r^2}\mu(z)\rho_r(z)\\
&= \sum_{s=1}^{N}\hat{p}_{rs}^{*}\hat{p}_{sr}\frac{1}{\hat{\mu}_r}\cdot\rho_r(z)\cdot\mu(z),\\
P_{11} &= \sum_{x,y\in S}\sum_{k,l,s,t=1}^{N}\mu(x)\mu(y)\rho_k(x)\rho_l(y)\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\rho_s(x)\rho_t(y)\hat{p}_{st}\frac{1}{\hat{\mu}_t^2}\mu(z)\rho_t(z)\rho_r(z)\\
&= \sum_{k,l,s,t=1}^{N}\rho_r(z)\cdot\left(\hat{\hat{\mu}}_{sk}\hat{p}_{kl}\frac{1}{\hat{\mu}_l}\hat{\hat{\mu}}_{lt}\frac{1}{\hat{\mu}_t^2}\cdot\hat{p}_{st}\cdot\rho_t(z)\right)\cdot\mu(z),\\
P_{12} &= \sum_{x,y\in S}\sum_{s,t=1}^{N}\mu(x)p(x,y)\rho_s(x)\rho_t(y)\hat{p}_{st}\frac{1}{\hat{\mu}_t^2}\mu(z)\rho_t(z)\rho_r(z)\\
&= \sum_{s,t=1}^{N}\rho_r(z)\cdot\left((\hat{p}^{*})_{st}^{T}\cdot\hat{p}_{st}\cdot\frac{1}{\hat{\mu}_t}\rho_t(z)\right)\cdot\mu(z),
\end{aligned}
$$

写成矩阵的形式, 并定义矩阵

$$
\begin{aligned}
M_1 &= \hat{p}I_{\hat{\mu}}^{-1}\hat{\hat{\mu}}I_{\hat{\mu}}^{-1}\hat{p}^{T}\rho-\hat{p}I_{\hat{\mu}}^{-1}\rho p^{T},\\
M_2 &= I_{\hat{\mu}}^{-1}\hat{p}^{T}\hat{\hat{\mu}}\hat{p}I_{\hat{\mu}}^{-1}\rho-I_{\hat{\mu}}^{-1}\hat{p}^{T}\rho p^{T},
\end{aligned}
$$

经简单运算, 得到

$$\begin{aligned}\frac{\partial J}{\partial Z} &= 2\Big[(M_1+M_2)*\rho-\rho\cdot\mathrm{diag}\Big(\rho^T\cdot(M_1+M_2)\Big)\\ &\quad -\mathrm{diag}(I_{\hat{\mu}}^{-2}\hat{\mu}I_{\hat{\mu}}^{-1}\hat{p}^T\hat{\mu}\hat{p})\cdot\rho+\mathrm{diag}(\hat{p}^*\hat{p}I_{\hat{\mu}}^{-1})\cdot\rho\\ &\quad +\rho\cdot\mathrm{diag}_{vm}\Big(\mathbf{1}_{1\times N}\cdot\big((\hat{\mu}\hat{p}I_{\hat{\mu}}^{-1}\hat{\mu}I_{\hat{\mu}}^{-2})*\hat{p}\big)\cdot\rho\Big)\\ &\quad -\rho\cdot\mathrm{diag}_{vm}\Big(\mathbf{1}_{1\times N}\cdot\big((\hat{p}^*)^T*\hat{p}\big)\cdot I_{\hat{\mu}}^{-1}\rho\Big)\Big]I_\mu,\end{aligned}\tag{C.16}$$

其中 $*$ 表示矩阵对应元素相乘的运算, $\mathrm{diag}(A)$ 是矩阵 A 的对角线部分, $\mathrm{diag}_{vm}(u)$ 是由向量 u 的分量形成的对角矩阵. 这就得到了 (4.29b).

参考文献

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

[2] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network flows: theory, algorithms and applications*. Prentice Hall, 1994.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[5] R. Albert, H. Jeong, and A.L. Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

[6] A. Arenas and A. Diaz-Guilera. Synchronization and modularity in complex networks. *The European Physical Journal: Special Topics*, 143(1):19–25, 2007.

[7] A. Arenas, A. Díaz-Guilera, and C.J. Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Physical Review Letters*, 96(11):114102, 2006.

[8] A. Arenas, J. Duch, A. Fernández, and S. Gómez. Size reduction of complex networks preserving modularity. *New Journal of Physics*, 9:176, 2007.

[9] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54. ACM, 2006.

[10] J.P. Bagrow and E.M. Bollt. Local method for detecting communities. *Physical Review E*, 72(4):046108, 2005.

[11] A.L. Barabási. *The new science of networks*. Cambridge: Perseus, 2002.

[12] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[13] A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.

[14] M. Barahona and L.M. Pecora. Synchronization in small-world systems. *Physical Review Letters*, 89(5):054101, 2002.

[15] E.R. Barnes. An algorithm for partitioning the nodes of a graph. *SIAM Journal on Algebraic and Discrete Methods*, 3:541–550, 1982.

[16] A. Barrat, M. Barthlemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge University Press, 2008.

[17] J.C. Bezdek. Cluster validity with fuzzy sets. *Cybernetics and Systems*, 3(3):58–73, 1973.

[18] J.C. Bezdek. Mathematical models for systematics and taxonomy. In *Proceedings of 8th International Conference on Numerical Taxonomy*, pages 143–166, 1975.

[19] J.C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.

[20] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

[21] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda. Detecting complex network modularity by dynamical clustering. *Physical Review E*, 75(4):045102, 2007.

[22] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[23] S. Boettcher and A.G. Percus. Optimization with extremal dynamics. *Physical Review Letters*, 86(23):5211–5214, 2001.

[24] B. Bollobas. *Modern graph theory*. Springer Verlag, 1998.

[25] K.P. Burnham and D.R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Verlag, 2002.

[26] A. Capocci, V.D.P. Servedio, G. Caldarelli, and F. Colaiori. Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications*, 352(2-4):669–676, 2005.

[27] D. Chakrabarti. Autopart: Parameter-free graph partitioning and outlier detection. *Lecture Notes in Computer Science*, 3202:112–124, 2004.

[28] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 554–560. ACM, New York, NY, USA, 2006.

[29] Y. Chi, X. Song, D. Zhou, K. Hino, and B.L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 153–162. ACM, New York, NY, USA, 2007.

[30] A.J. Chorin. Accurate evaluation of Wiener integrals. *Mathematics of Computation*, 27:1–15, 1973.

[31] A.J. Chorin. Conditional expectations and renormalization. *Multiscale Modeling and Simulation*, 1(1):105–118, 2003.

[32] A.J. Chorin, O.H. Hald, and R. Kupferman. Optimal prediction and the Mori-Zwanzig representation of irreversible processes. *Proceedings of the National Academy of Sciences of the United States of America*, 97(7):2968–2973, 2000.

[33] A.J. Chorin, A.P. Kast, and R. Kupferman. Optimal prediction of underresolved dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 95(8):4094–4098, 1998.

[34] A.J. Chorin, A.P. Kast, and R. Kupferman. Unresolved computation and optimal predictions. *Communications on Pure and Applied Mathematics*, 52(10):1231–1254, 1999.

[35] A.J. Chorin, R. Kupferman, and D. Levy. Optimal Prediction for Hamiltonian Partial Differential Equations. *Journal of Computational Physics*, 162(1):267–297, 2000.

[36] F.R.K. Chung. *Spectral graph theory*. American Mathematical Society, Providence, USA, 1997.

[37] A. Clauset, C. Moore, and M.E.J. Newman. Structural inference of hierarchies in networks. *Lecture Notes in Computer Science*, 4503:1–13, 2007.

[38] A. Clauset, C. Moore, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[39] A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

[40] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[41] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

[42] L. Danon, A. Díaz-Guilera, and A. Arenas. The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006:P11010, 2006.

[43] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005:P09008, 2005.

[44] R.N. Dave. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters*, 17(6):613–623, 1996.

[45] D.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):224–227, 1979.

[46] R. De Castro and J.W. Grossman. Famous trails to Paul Erdös. *The Mathematical Intelligencer*, 21(3):51–53, 1999.

[47] M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *SIAM Journal on Numerical Analysis*, 36(2):491–515, 1999.

[48] J.C. Delvenne, S.N. Yaliraki, and M. Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):12755, 2010.

[49] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte. Computation of essential molecular dynamics by subdivision techniques I: basic concept. *Lecture Notes in Computational Science and Engineering*, 4:98–115, 1999.

[50] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications*, 315(1-3):39–59, 2000.

[51] P.S. Dodds, R. Muhamad, and D.J. Watts. An experimental study of search in global social networks. *Science*, 301(5634):827, 2003.

[52] W.E. Donath and A.J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.

[53] L. Donetti and M.A. Muñoz. Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004:P10012, 2004.

[54] L. Donetti and M.A. Muñoz. Improved spectral algorithm for the detection of network communities. In *American Institute of Physics Conference Series*, volume 779, pages 104–107, 2005.

[55] S.V. Dongen. Graph clustering by flow simulation. *Computer Science Review*, 1(1):27–64, 2000.

[56] P. Doreian, V. Batagelj, and A. Ferligoj. *Generalized blockmodeling*. Cambridge University Press, New York, USA, 2005.

[57] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104, 2005.

[58] J.C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybernetics and Systems*, 3(3):32–57, 1973.

[59] J.C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems*, 4(1):95–104, 1974.

[60] W. E, T. Li, and E. Vanden-Eijnden. Optimal partition and effective dynamics of complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 105(23):7907–7912, 2008.

[61] W. E, W. Ren, and E. Vanden-Eijnden. String method for the study of rare events. *Physical Review B*, 66(5):052301, 2002.

[62] W. E, W. Ren, and E. Vanden-Eijnden. Finite temperature string method for the study of rare events. *Journal of Physical Chemistry B*, 109(14):6688–6693, 2005.

[63] J.P. Eckmann and E. Moses. Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5825, 2002.

[64] P. Erdös and A. Rényi. *On the evolution of random graphs*. Citeseer, 1960.

[65] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–140, 1736.

[66] D.J. Fenn, M.A. Porter, M. McDonald, S. Williams, N.F. Johnson, and N.S. Jones. Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007–2008 credit crisis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(3):033119, 2009.

[67] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

[68] G.W. Flake, S. Lawrence, C.L. Giles, and F.M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35(3):66–70, 2002.

[69] L.R. Ford and D.R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

[70] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[71] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41, 2007.

[72] S. Fortunato, V. Latora, and M. Marchiori. Method to find community structures based on information centrality. *Physical Review E*, 70(5):056104, 2004.

[73] Y. Fukuyama and M. Sugeno. A new method of choosing the number of clusters for the fuzzy *c*-means method. In *Proceedings of 5th Fuzzy Systems Symposium*, pages 247–250, 1989.

[74] I. Gath and A.B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–780, 1989.

[75] D. Gfeller and P. De Los Rios. Spectral coarse graining of complex networks. *Physical Review Letters*, 99(3):038701, 2007.

[76] D. Gfeller and P. De Los Rios. Spectral coarse graining and synchronization in oscillator networks. *Physical Review Letters*, 100(17):174104, 2008.

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

[78] A.V. Goldberg and R.E. Tarjan. A new approach to the maximum-flow problem. *Journal of the Association for Computing Machinery*, 35(4):921–940, 1988.

[79] G.H. Golub and C.F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996.

[80] B.H. Good, Y.A. De Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[81] M.S. Granovetter. The strength of weak ties. *The American journal of sociology*, 78(6):1360–1380, 1973.

[82] R. Guimerà and L.A.N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[83] R. Guimerà, M. Sales-Pardo, and L.A.N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.

[84] R. Guimerà, M. Sales-Pardo, and L.A.N. Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3):036102, 2007.

[85] D.J. Hartfiel and C.D. Meyer. On the structure of stochastic matrices with a subdominant eigenvalue near 1. *Linear Algebra and Its Applications*, 272(1-3):193–203, 1998.

[86] T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, Berlin, Germany,, 2001.

[87] M.B. Hastings. Community detection as an inference problem. *Physical Review E*, 74(3):035102, 2006.

[88] A. Hlaoui and W. Shengrui. A direct approach to graph clustering. *Neural Networks Computational Intelligence*, pages 158–163, 2004.

[89] J.M. Hofman and C.H. Wiggins. Bayesian approach to network modularity. *Physical Review Letters*, 100(25):258701, 2008.

[90] P. Holme, M. Huss, and H. Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19(4):532–538, 2003.

[91] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5249, 2004.

[92] Y. Hu, M. Li, P. Zhang, Y. Fan, and Z. Di. Community detection by signaling on complex networks. *Physical Review E*, 78(1):016115, 2008.

[93] B.D. Hughes. *Random walks and random environments*. Clarendon Press, 1995.

[94] I. Ispolatov, I. Mazo, and A. Yuryev. Finding mesoscopic communities in sparse networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006:P09014, 2006.

[95] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–4574, 2001.

[96] H. Jeong, S.P. Mason, A.L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[97] P.F. Jonsson, T. Cavanna, D. Zicha, and P.A. Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, 7(1):2, 2006.

[98] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, 2004.

[99] T. Kato. *Perturbation theory for linear operators*. Springer Verlag, 1995.

[100] B.W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.

[101] R.Z. Khasminskii. A limit theorem for the solutions of differential equations with random righthand sides. *Theory of Probability and Application*, 11:390–406, 1966.

[102] M.S. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. volume 2, pages 622–633. VLDB Endowment, 2009.

[103] S. Kirkpatrick, D.G. Jr., and M.P. Vecchi. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983.

[104] H. Kleinert. *Gauge fields in condensed matter*, volume 1. World Scientific Singapore, 1989.

[105] P.E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer, 1992.

[106] D.E. Knuth. *The Stanford GraphBase: a platform for combinatorial computing*. Addison-Wesley, Reading, MA, 1993.

[107] S. Lafon and A.B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–1403, 2006.

[108] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1):016118, 2009.

[109] A. Lancichinetti and S. Fortunato. Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5):056117, 2009.

[110] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping

and hierarchical community structure in complex networks. *New Journal of Physics*, 11:033015, 2009.

[111] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110, 2008.

[112] E.A. Leicht and M.E.J. Newman. Community structure in directed networks. *Physical Review Letters*, 100(11):118703, 2008.

[113] D. Li, I. Leyva, J.A. Almendral, I. Sendina-Nadal, J.M. Buldú, S. Havlin, and S. Boccaletti. Synchronization interfaces and overlapping communities in complex networks. *Physical Review Letters*, 101(16):168701, 2008.

[114] T. Li, J. Liu, and W. E. Probabilistic framework for network partition. *Physical Review E*, 80(2):026106, 2009.

[115] Y.R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B.L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In *Proceeding of the 17th International Conference on World Wide Web*, pages 685–694. ACM, New York, NY, USA, 2008.

[116] J. Liu. Detecting the fuzzy clusters of complex networks. *Pattern Recognition*, 43(4):1334–1345, 2010.

[117] J. Liu. Fuzzy modularity and fuzzy community structure in networks. *The European Physical Journal B: Condensed Matter and Complex Systems*, 77(4):547–557, 2010.

[118] J. Liu and T. Li. A validity index approach for network partitions. *Physica A: Statistical Mechanics and its Applications*, 390(20):3579–3591, 2011.

[119] J. Liu and T. Liu. Detecting community structure in complex networks using simulated annealing with k-means algorithms. *Physica A: Statistical Mechanics and its Applications*, 389(11):2300–2309, 2010.

[120] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.

[121] D. Lusseau. The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270:S186–S188, 2003.

[122] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, and S.M. Dawson. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[123] J. Ma, B. Gao, Y. Wang, and Q. Cheng. Conjugate and natural gradient rules for BYY harmony learning on Gaussian mixture with automated model selection. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(5):701–713, 2005.

[124] J. Ma and L. Wang. BYY harmony learning on finite mixture: adaptive gradient implementation and a floating RPCL mechanism. *Neural Processing Letters*, 24(1):19–40, 2006.

[125] J. Ma, T. Wang, and L. Xu. A gradient BYY harmony learning rule on Gaussian mixture with automated model selection. *Neurocomputing*, 56:481–487, 2004.

[126] D.J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK, 2003.

[127] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

[128] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[129] C.P. Massen and J.P.K. Doye. Identifying communities within energy landscapes. *Physical Review E*, 71(4):046101, 2005.

[130] W.D. McComb and A.G. Watt. Conditional averaging procedure for the elimination of the small-scale modes from incompressible fluid turbulence at high Reynolds numbers. *Physical Review Letters*, 65(26):3281–3284, 1990.

[131] A. Medus, G. Acuna, and CO Dorso. Detection of community structures in networks via global optimization. *Physica A: Statistical Mechanics and its Applications*, 358(2-4):593–604, 2005.

[132] M. Meila and J. Shi. A random walks view of spectral segmentation. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*, pages 92–97, Kaufmann, San Francisco, 2001. Citeseer.

[133] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[134] C.D. Meyer. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Review*, 31(2):240–272, 1989.

[135] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[136] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[137] B. Nadler, S. Lafon, R.R. Coifman, and I.G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

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

[139] M.E.J. Newman. Detecting community structure in networks. *The European Physical Journal B: Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

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

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

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

[143] M.E.J. Newman, A.L. Barabási, and D.J. Watts. *The structure and dynamics of networks*. Princeton, NJ: Princeton University Press, 2006.

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

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

[146] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14, pages 849–856, 2001.

[147] T. Nishikawa, A.E. Motter, Y.C. Lai, and F.C. Hoppensteadt. Heterogeneity in oscillator networks: Are smaller worlds easier to synchronize? *Physical Review Letters*, 91(1):014101, 2003.

[148] M.K. Pakhira, S. Bandyopadhyay, and U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3):487–501, 2004.

[149] N.R. Pal and J.C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, 3(3):370–379, 1995.

[150] G. Palla, A.L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[151] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[152] A. Papoulis, S.U. Pillai, and S. Unnikrishna. *Probability, random variables, and stochastic processes*. McGraw-Hill New York, 2002.

[153] M. Penrose. *Random geometric graphs*. Oxford University Press, USA, 2003.

[154] A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization: A universal concept in nonlinear sciences*. Cambridge University Press, 2003.

[155] P. Pons and M. Latapy. Computing communities in large networks using random walks. *Lecture Notes in Computer Science*, 3733:284–293, 2005.

[156] A. Pothen. Graph partitioning algorithms with applications to scientific computing. *ICASE LaRC Interdisciplinary Series in Science and Engineering*, 4:323–368, 1997.

[157] N. Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999.

[158] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[159] J.J. Ramasco and M. Mungan. Inversion method for content-based networks. *Physical Review E*, 77(3):036122, 2008.

[160] M. Ramze Rezaee, B.P.F. Lelieveldt, and J.H.C. Reiber. A new cluster validity index for the fuzzy *c*-mean. *Pattern Recognition Letters*, 19(3-4):237–246, 1998.

[161] M.J. Rattigan, M. Maier, and D. Jensen. Graph clustering with network structure indices. In *Proceedings of the 24th international conference on Machine learning*, pages 783–790. ACM, 2007.

[162] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[163] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters*, 93(21):218701, 2004.

[164] J. Reichardt and D.R. White. Role models for complex networks. *The European Physical Journal B: Condensed Matter and Complex Systems*, 60(2):217–224, 2007.

[165] T. Richardson, P.J. Mucha, and M.A. Porter. Spectral tripartitioning of networks. *Physical Review E*, 80(3):036111, 2009.

[166] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[167] P. Ronhovde and Z. Nussinov. Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, 80(1):016109, 2009.

[168] M. Rosvall, D. Axelsson, and C.T. Bergstrom. The map equation. *The European Physical Journal: Special Topics*, 178(1):13–23, 2009.

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

[170] M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–1123, 2008.

[171] M. Roubens. Pattern classification problems and fuzzy sets. *Fuzzy Sets and Systems*, 1(4):239–253, 1978.

[172] M. Sales-Pardo, R. Guimera, A.A. Moreira, and L.A.N. Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39):15224, 2007.

[173] A. Schenker, M. Last, H. Bunke, and A. Kandel. Graph representations for web document clustering. *Pattern Recognition and Image Analysis*, pages 935–942, 2003.

[174] W.H.A. Schilders, H.A. Van Der Vorst, and J. Rommes. *Model order reduction: theory, research aspects and applications*. Springer Verlag, 2008.

[175] P. Schuetz and A. Caflisch. Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Physical Review E*, 77(4):046112, 2008.

[176] P. Schuetz and A. Caflisch. Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. *Physical Review E*, 78(2):026112, 2008.

[177] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo. *Journal of Computational Physics*, 151(1):146–168, 1999.

[178] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[179] A. Scotti and C. Meneveau. Fractal model for coarse-grained nonlinear partial differential equations. *Physical Review Letters*, 78(5):867–870, 1997.

[180] E. Seneta. *Non-negative matrices and Markov chains*. Springer Verlag, 2006.

[181] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1):64–68, 2002.

[182] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[183] A. Sinclair. *Algorithms for random generation and counting: a Markov chain approach*. Springer, 1993.

[184] S.W. Son, H. Jeong, and J.D. Noh. Random field Ising model and community structure in complex networks. *The European Physical Journal B: Condensed Matter and Complex Systems*, 50(3):431–437, 2006.

[185] D.A. Spielmat and S.H. Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *IEEE Symposium on Foundations of Computer Science*, pages 96–105. IEEE Computer Society, 1996.

[186] G.W. Stewart. On the structure of nearly uncoupled Markov chains. *Mathematical Computer Performance and Reliability*, pages 287–302, 1984.

[187] S.H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[188] P.R. Suaris and G. Kedem. An algorithm for quadrisection and its application to standard cell placement. *IEEE Transactions on Circuits and Systems*, 35(3):294–303, 1988.

[189] H. Sun, S. Wang, and Q. Jiang. FCM-based model selection algorithms for determining the number of clusters. *Pattern Recognition*, 37(10):2027–2037, 2004.

[190] Y. Sun, B. Danila, K. Josić, and K.E. Bassler. Improved community structure detection using a modified fine-tuning strategy. *Europhysics Letters*, 86:28004, 2009.

[191] M. Szummer and T. Jaakkola. Partially labeled classification with Markov random walks. *Advances in neural information processing systems*, 2:945–952, 2002.

[192] J.R. Tyler, D.M. Wilkinson, and B.A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. In *International Conference on Communities and Technologies*, pages 81–96, Deventer, The Netherlands, 2003.

[193] A. Vazquez. Bayesian approach to clustering real value, categorical and network data: solution via variational methods. *Arxiv preprint arXiv:0805.2689*, 2008.

[194] J. Čopič, M.O. Jackson, and A. Kirman. Identifying community structures from network data. *URL http://www.hss.caltech.edu/jernej/netcommunity.pdf*, 2008.

[195] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[196] K. Wakita and T. Tsurumi. Finding community structure in mega-scale social networks:[extended abstract]. In *Proceedings of the 16th international conference on World Wide Web*, pages 1275–1276. ACM, 2007.

[197] C.S. Wallace and D.M. Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

[198] G. Wang, Y. Shen, and M. Ouyang. A vector partitioning approach to detecting community structure in complex networks. *Computers and Mathematics with Applications*, 55(12):2746–2752, 2008.

[199] X. Wang, X. Li, and G. Chen. *Complex networks theory and its application*. Tsinghua University Press, Beijing, 2006. (in Chinese).

[200] D.J. Watts. The "new" science of networks. *Annual review of sociology*, 30:243–270, 2004.

[201] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[202] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *International Conference on Computer Vision*, pages 975–982. IEEE Computer Society, 1999.

[203] D.M. Wilkinson and B.A. Huberman. A method for finding communities of related genes. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5241–5248, 2004.

[204] R.L. Winkler. *An introduction to Bayesian inference and decision*. Probabilistic Publishing, Gainesville, USA, 2003.

[205] F. Wu and B.A. Huberman. Finding communities in linear time: a physics approach. *The European Physical Journal B: Condensed Matter and Complex Systems*, 38(2):331–338, 2004.

[206] F.Y. Wu. The potts model. *Reviews of Modern Physics*, 54(1):235, 1982.

[207] K.L. Wu and M.S. Yang. A cluster validity index for fuzzy clustering. *Pattern Recognition Letters*, 26(9):1275–1291, 2005.

[208] X.L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 841–847, 1991.

[209] B. Yang and J. Liu. Discovering global network communities based on local centralities. *ACM Transactions on the Web*, 2(1):1–32, 2008.

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

[211] N. Zahid, M. Limouri, and A. Essaid. A new cluster-validity for fuzzy clustering. *Pattern Recognition*, 32(7):1089–1097, 1999.

[212] H. Zanghi, C. Ambroise, and V. Miele. Fast online graph clustering via Erdös-Rényi mixture. *Pattern Recognition*, 41(12):3592–3599, 2008.

[213] H. Zhou. Distance, dissimilarity index, and network community structure. *Physical Review E*, 67(6):061901, 2003.

[214] H. Zhou. Network landscape from a Brownian particle's perspective. *Physical Review E*, 67(4):041908, 2003.

[215] H. Zhou and R. Lipowsky. Network Brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. *Lecture Notes in Computer Science*, 3038:1062–1069, 2004.

[216] E. Ziv, M. Middendorf, and C.H. Wiggins. Information-theoretic approach to network modularity. *Physical Review E*, 71(4):046117, 2005.

致　谢

攻读博士研究生的五年时光值得我永生记忆。值此博士论文完成之际，我要真诚地感谢所有曾经给予我关怀、指导、支持和帮助的老师和同学：

首先，我要衷心地感谢我的导师鄂维南教授。鄂老师正直善良的品德，渊博的知识和杰出的成就是我终生学习的榜样。在我论文的选题和写作过程中，鄂老师都十分关心和支持，他严谨的治学态度、开阔的研究思路和敏锐的洞察力都将使我终身受益。感谢李铁军教授一直以来对我的关怀备至、悉心指导和帮助，论文的结构、研究的思路、论证的方法都倾注了李老师大量的心血和精力，他给予我的不仅有广博的专业知识，更有开拓创新的科研求知精神。他们的言传身教令我悟出许多做人做事做学问的真谛。

我要特别感谢张平文教授对我这些年来在思想、学业和生活等各个方面的培养、关心和帮助。张老师执着的科研热情、宽容的交流方式、严谨求实的治学态度深深地感染和激励着我。感谢张老师为我们组织的各类学术研讨会，给我们创造了非常好的科研环境和氛围。

在攻读学位和论文写作的过程中，有幸得到了许多老师的指导和帮助：感谢汤华中教授、李若教授、李治平教授、周铁教授、高立教授、胡俊副教授、吴金彪副教授。感谢曾给予我无私帮助和鼓励的马尽文教授和其他各位老师。他们精深的学术见解、不倦的教诲伴我度过了充实的博士研究生学习时光，这段时光将是我人生中最宝贵的财富。

感谢数学学院的各位领导，感谢田立青老师，刘雨龙老师和卢朓老师，感谢他们在我成长中对我的思想、生活和就业等方面的指导和关心。

感谢所有和我朝夕相处、攻读博士学位的同学们。在我学习的过程中，他们给予了我无私的关心和帮助。我们是同学，更是一生的朋友，同窗生涯结下的深厚友谊将让我永远珍惜和难忘。

感谢我的父母给予我温暖的关爱和鼓励，是他们的理解、支持与付出，让我

在感到疲惫之时，能够鼓起勇气并充满希望，得以顺利完成学业。

最后我要感谢论文评阅专家及答辩组专家，论文能够得到你们的审阅是我的荣幸，你们的宝贵意见和建议将会启发我不断努力进取和创新。感谢老师们的辛苦工作。

论文的完成凝聚我了几年的心血，它离不开各位老师和同学的指导和帮助。这些指导和帮助将永远铭记在我心中，伴我在今后的征途上不断前进，使我在面对生活时将不再艰难。

博士期间发表的学术论文

[1] Tiejun Li, **Jian Liu** and Weinan E, Probabilistic framework for network partition, *Physical Review E*, **80** (2009), 026106.

[2] **Jian Liu** and Na Wang, Detecting community structure of complex networks by simulated annealing with optimal prediction, Proceedings of *International Conference on Computational Intelligence and Software Engineering*, **2** (2009).

[3] **Jian Liu** and Na Wang, Detecting community structure of complex networks by affinity propagation, Proceedings of *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, **4** (2009), 13–19.

[4] **Jian Liu**, Detecting the fuzzy clusters of complex networks, *Pattern Recognition*, **43** (2010), 1334–1345.

[5] **Jian Liu** and Tingzhan Liu, Detecting community structure in complex networks using simulated annealing with k-means algorithms, *Physica A: Statistical Mechanics and its Applications*, **389** (2010), 2300–2309.

[6] **Jian Liu**, Fuzzy modularity and fuzzy community structure in networks, *European Physical Journal B: Condensed Matter and Complex Systems*, **77** (2010), 547–557.

[7] **Jian Liu** and Tingzhan Liu, Coarse-grained diffusion distance for community structure detection in complex networks, *Journal of Statistical Mechanics: Theory and Experiment*, (2010), P12030.

[8] **Jian Liu**, An extended validity index for identifying community structure in networks, *Lecture Notes in Computer Science*, **6064** (2010), 258–267.

[9] **Jian Liu**, Finding and evaluating fuzzy clusters in networks, *Lecture Notes in Computer Science*, **6146** (2010), 17–26.

[10] **Jian Liu**, Fuzzy algorithm based on diffusion maps for network partition, *Lecture Notes in Artificial Intelligence*, **6216** (2010), 163–172.

[11] **Jian Liu**, Comparing fuzzy algorithms on overlapping communities in networks, *Lecture Notes in Computer Science*, **6377** (2010), 269–276.

[12] **Jian Liu**, Comparative analysis for k-means algorithms in network community detection, *Lecture Notes in Computer Science*, **6382** (2010), 158–169.

[13] **Jian Liu** and Tiejun Li, A validity index approach for network partitions, *Physica A: Statistical Mechanics and its Applications*, **390** (2011), 3579–3591.

个人简历

基本情况

- 刘健, 女, 1984 年 4 月生于吉林省长春市.
- 电子邮箱: dugujian@pku.edu.cn

教育背景

- 2006.9 – 2011.7　北京大学数学科学学院, 科学与工程计算系, 理学博士
- 2002.9 – 2006.7　吉林大学数学学院, 信息与计算科学系, 理学学士

参与的科研项目

- 复杂网络的模型约化, 国家自然科学基金 *(10871010)*, 2009.1 – 2011.12
- 科技工作者职业发展空间和通道调查, 中国科学技术协会重点课题 *(2007DCYJ05)*, 2007.8 – 2008.8

所获荣誉与奖励

2011 年　北京大学 2011 年校优秀毕业生

2011 年　北京大学第十三届研究生“学术十杰”

2010 年　北京大学 2009 – 2010 学年世坤奖学金

2010 年　北京大学 2009 – 2010 学年学术创新奖

2009 年　北京大学 2008 – 2009 学年腾讯科技卓越特等奖学金

2009 年　北京大学 2008 – 2009 学年学习优秀奖

2007 年　北京大学 2006 – 2007 学年学习优秀奖

2006 年　吉林大学 2006 年校优秀毕业生

2006 年 吉林大学 2005 – 2006 学年校一等奖学金, 校优秀学生

2005 年 吉林大学 2005 年十佳大学生

2005 年 吉林大学大学生数学建模竞赛一等奖, 高教社杯全国大学生数学建模竞赛吉林省赛区一等奖

2005 年 吉林大学 2004 – 2005 学年校一等奖学金, 校优秀学生, 东荣奖学金一等奖, 惠普奖学金

2004 年 吉林大学 2003 – 2004 学年校一等奖学金, 校优秀学生, 中国石油奖学金优秀生奖

2003 年 吉林大学 2002 – 2003 学年校二等奖学金, 院优秀学生

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日