

学校代码: 10246
学 号: 20110240002

復旦大學

博 士 学 位 论 文
(学术学位)

复杂网络中心性度量及其区分能力

**Centrality Measures and Their Discriminative Power
in Complex Networks.**

院 系: 计算机科学技术学院

专 业: 计算机软件与理论

姓 名: 包奇

指导教师: 章忠志 教授

完成日期: 2023 年 8 月 11 日

指导小组成员名单

阚海斌	教授	复旦大学
张军平	教授	复旦大学
陈翌佳	教授	上海交通大学
章忠志	教授	复旦大学

摘要

现实世界中的许多事物都可以用相互关联的实体组成的系统来描述。复杂网络作为一种描述现实系统的强大工具近年来在各个领域的研究中展现出了巨大的潜力和重要性。寻找关键点或关键边的问题作为网络科学领域的一个基本问题有着悠久的历史历史和广泛的应用。为了评估网络中节点和边的重要性，学者们提出了多种衡量点和边相对重要性的方法，即中心性度量。这些中心性度量各具特点，它们衡量网络中点和边重要性的方式也各不相同。因此，研究这些中心性度量的特性是十分必要的。为了更好的了解这些中心性度量的特征，学者们在这方面进行了大量的研究，并取得了一些重要的研究成果。其中，有的学者提出了中心性公理来定性的描述中心性度量的特征，还有的学者深入研究了中心性度量的敏感性并提出了相关的量化指标。然而，到目前为止，学者们对中心性度量的区分能力还没有进行深入研究。为了更好的理解中心性度量的区分能力，本文深入的研究了点和边中心性度量的区分能力，同时提出了区分能力更强的节点中心性度量和边中心性度量，具体内容如下：

基于自同构等价的概念提出了节点中心性度量区分能力的评价标准。本文深入的研究了节点中心性度量的区分能力并基于节点自同构等价的概念提出了节点中心性度量区分能力的量化评价指标。根据等价节点中心性值应当相等的原则，本文基于节点自同构等价的概念提出了相应的自同构等价公理，规定了节点中心性度量对于自同构等价的节点必须给出相等的中心性值。本文证明了几个常用的中心性度量均满足本文提出的自同构等价公理。本文再根据不等价节点的中心性值应当不同的原则，用节点中心性度量对于不等价的节点给出不同中心性值的概率定义了区分能力指数的概念。在满足自同构等价公理时，一个中心性度量的区分能力越强，它的区分能力指数就越大。同时，本文还提出了广义区分度的概念用以快速比较两种节点中心性度量的区分能力。本文在多个现实网络和模型网络上用上述方法深入的分析了几种常用的节点中心性度量的区分能力。此外，本文还把上述分析节点中心性度量区分能力的方法扩展到了有向加权网络。

基于置根生成森林提出了森林中心性并提出了相应的快速近似算法。为探究是否存在区分能力更为出色的节点中心性度量，本文基于置根生成森林的概念提出了森林中心性这一新的节点中心性度量。森林中心性基于以某个节点为根的所有置根生成森林中这个点所在连通分量的平均大小来定义一个节点的相对重要程度。相较于其它常用的节点中心性度量，森林中心性不仅可以作用于不连通的网络，它还有着更为出色的区分能力。此外，本文还基于 JL 引理以及快速求解器设计了一个用于计算森林中心性值的快速近似算法。该算法能够用几乎线性的时间和空间复杂度给出一个网络中所有节点的森林中心性的准确近似值。本文在多个现实网络和模型网络上比较了森林中心性和其他几个常用的节点中心性度量的区分能力。实验结果表明，相较于其他几个常用的节点中心性度量，森林中心性确实有着更为强大的区分能力。此外，本文在包含数百万节点的现实网络中验证了快速近似算法的运算效率和准确性。

提出了节点中心性度量区分能力的评价标准并提出了森林边中心性。本文还

深入的研究了边中心性度量的区分能力并基于自同构的概念提出了边中心性度量区分能力的量化评价指标。本文首先基于自同构的概念给出了边自同构等价的定义。再根据等价边的中心性值应当相等的原则，本文基于边的自同构等价提出了相应的自同构等价公理，规定了边中心性度量对于自同构等价的边须给出相等的边中心性值。本文证明了几个常用的边中心性度量均满足边中心性的自同构等价公理。本文接着根据不等价边的中心性值应当不同的原则，用边中心性度量对于不等价边给出不同中心性值的概率定义了边中心性的区分能力指数。在满足自同构等价公理时，一个边中心性度量的区分能力越强，它的区分能力指数就越大。同时，本文还提出了广义区分度的概念用以快速比较两种不同边中心性度量的区分能力。为得到区分能力更强的边中心性，本文还基于置根生成森林的概念提出了森林边中心性这一新的边中心性度量。本文在多个现实网络和模型网络上用上述方法深入的分析了几种常用的边中心性度量以及新提出的森林边中心性的区分能力。实验结果表明，森林边中心性明显拥有更加优秀的区分能力。

关键词：中心性度量，自同构，复杂网络，区分能力

中图分类号：TP393

Abstract

Many things in the real world can be described as systems composed of interconnected entities. Complex networks, as a powerful tool for describing real-world systems, have shown significant potential and importance in various fields of research in recent years. The problem of identifying key nodes or crucial edges, as a fundamental issue in network science, has a long history of study and wide-ranging applications. To assess the importance of nodes and edges in a network, scholars have proposed various methods for measuring the relative significance of nodes and edges, known as centrality measures. These centrality measures have distinct characteristics, and they vary in how they quantify the importance of nodes and edges in a network. Therefore, studying the properties of these centrality measures is essential. However, up to this point, scholars have not conducted an in-depth study of the discriminating power of centrality measures. To gain a better understanding of the discriminating power of centrality measures, we deeply study the discriminating power of node and edge centrality measures in this paper. We also propose node centrality measures and edge centrality measures with stronger discriminating power. The details are as follows:

We propose an evaluation criterion for discriminating power of node centrality measures based on the concept of automorphism equivalence. We conduct an in-depth study of the discriminating power of node centrality measures and introduce a quantitative evaluation index for discriminating power based on the concept of automorphism equivalence for nodes. According to the principle that equivalent nodes should have equal centrality values, we put forth the corresponding automorphism equivalence axioms based on the concept of node automorphism equivalence, stipulating that node centrality measures must assign equal centrality values to nodes that are automorphism equivalents. We prove that several commonly used centrality measures satisfy the automorphism equivalence axioms proposed in this paper. Furthermore, by considering that the centrality values of non-equivalent nodes should be distinct, we define the concept of a discriminating power index using the probability that a node centrality measure assigns different centrality values to non-equivalent nodes. When satisfying the automorphism equivalence axioms, a centrality measure exhibits stronger discriminating power with a higher discriminating power index. Additionally, we introduce the notion of generalized discrimination degree to efficiently compare the discriminating power of two node centrality measures. We thoroughly analyze the discriminating power of several commonly used node centrality measures using the proposed approach on multiple real-world and model networks. Moreover, we extend the method for analyzing the discriminating power of node centrality measures to directed weighted networks.

We propose Forest Centrality based on the concept of rooted spanning forests and propose a corresponding fast approximation algorithm. To explore the possibility of more effective node centrality measures with better discriminating power, we propose a new node centrality measure called Forest Centrality, based on the concept of rooted spanning forests. Forest Centrality defines the relative importance of a node

based on the average size of connected components containing that node in all rooted spanning forests rooted at that node. Compared to other commonly used node centrality measures, Forest Centrality not only works on disconnected networks but also exhibits superior discriminating power. Furthermore, we design a fast approximation algorithm for computing Forest Centrality values based on the Johnson-Lindenstrauss lemma and fast linear solvers. This algorithm provides accurate approximations of Forest Centrality values for all nodes in a network with nearly linear time and space complexity. We compare the discriminating power of Forest Centrality with several other commonly used node centrality measures on multiple real-world and model networks. Experimental results demonstrate that forest centrality indeed possesses stronger discriminating power. Additionally, we verify the efficiency and accuracy of the fast approximation algorithm in real networks with millions of nodes.

We propose an evaluation criterion for discriminating power of edge centrality measures and propose Forest Edge Centrality. We delve into the discriminating power of edge centrality measures and propose a quantitative evaluation index for the discriminating power of edge centrality measures based on the concept of automorphism. We first define edge automorphism equivalence based on the concept of automorphism. Then, following the principle that equivalent edges should have equal centrality values, we establish the corresponding automorphism equivalence axioms for edge centrality measures, stipulating that edge centrality measures must assign equal centrality values to edges that are automorphic equivalent. We prove that several commonly used edge centrality measures satisfy the automorphism equivalence axioms proposed in this paper. Subsequently, by considering that the centrality values of non-equivalent edges should be distinct, we define the concept of an edge centrality discriminating power index using the probability that an edge centrality measure assigns different centrality values to non-equivalent edges. When satisfying the automorphism equivalence axioms, an edge centrality measure exhibits stronger discriminating power with a higher discriminating power index. Moreover, we introduce the notion of generalized discrimination degree to efficiently compare the discriminating power of two different edge centrality measures. To obtain edge centrality measures with stronger discriminating power, we propose Forest Edge Centrality based on the concept of rooted spanning forests. We conduct a thorough analysis of several commonly used edge centrality measures as well as the newly proposed Forest Edge Centrality in terms of discriminating power on multiple real-world and model networks using the aforementioned approach. Experimental results show that Forest Edge Centrality indeed possesses superior discriminating power.

Keywords: Centrality Measures, Automorphism, Complex Networks, Discriminating Power

Chinese Library Classification: TP393

主要符号表

标量、向量、矩阵和集合

x	标量
\mathbf{x}	向量
x_i	向量元素
\mathbf{X}	矩阵
x_{ij}	矩阵元素
\mathcal{S}	集合
$f(\cdot)$	函数
$f: \mathcal{S}_1 \rightarrow \mathcal{S}_2$	映射
$P(\cdot)$	概率
\mathbf{x}^\top	向量转置
\mathbf{X}^\top	矩阵转置
\mathbf{X}^{-1}	矩阵的逆
\mathbf{X}^\dagger	Moore-Penrose 逆
$\text{Tr}(\mathbf{X})$	矩阵的迹
$ \mathbf{X} $	矩阵行列式
$\ \mathbf{x}\ $	向量的模
$\ \mathbf{x}\ $	向量的 范数
$\ \mathbf{X}\ _F$	矩阵的 Frobenius 范数
$ \mathcal{S} $	集合的元素个数

图相关

$G = (V, E)$	无向无权图
$G = (V, E, w)$	无向加权图
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	有向无权图
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$	有向无权图
d_u	节点的度
$R(u, v)$	等效电阻，电阻距离
$d(u, v)$	节点间最短路径长度
σ_{uv}	节点间最短路径个数
$K(G)$	基尔霍夫指数
$\mathcal{T}(G)$	生成树集合
$\mathcal{F}(G)$	置根生成森林集合

A	邻接矩阵
B	边-点关联矩阵
D	度对角阵
L	拉普拉斯矩阵

其它

\mathbb{R}	实数集合
\mathbb{R}^+	正实数集合
$\log_x(\cdot)$	以 x 为底的对数
\approx_ϵ	ϵ 近似
$O(\cdot)$	渐进行为记号

目录

摘要	IV
Abstract	VII
主要符号表	X
目录	XII
插图索引	XVII
表格索引	XVIII
1 绪论	1
1.1 研究背景和意义	1
1.2 研究历史和现状	3
1.2.1 中心性公理	3
1.2.2 中心性度量的敏感性研究	4
1.2.3 中心性度量的区分能力研究	4
1.3 本文的主要贡献	4
1.4 论文结构	6
2 预备知识	9
2.1 引言	9
2.2 图及其矩阵表示	9
2.3 生成树	9
2.4 电阻网络	10
3 点中心性区分能力的评价标准	13
3.1 引言	13
3.2 点中心性度量	13
3.2.1 度中心性	13
3.2.2 介数中心性	14
3.2.3 接近中心性	14
3.2.4 PageRank 中心性	14
3.2.5 特征向量中心性	15
3.2.6 信息中心性	15
3.3 节点之间的等价关系	16
3.3.1 结构等价	16

3.3.2	正则等价	17
3.3.3	自同构等价	17
3.3.4	小结	19
3.4	自同构等价公理	19
3.4.1	相关证明	19
3.4.2	公理	22
3.5	区分能力指数	23
3.6	实验	26
3.6.1	现实网络	26
3.6.2	模型网络	29
3.7	有向加权图	29
3.7.1	理论推广	29
3.7.2	应用	31
3.8	本章小结	32
4	森林中心性	35
4.1	引言	35
4.2	置根生成森林	35
4.3	森林中心性	36
4.3.1	定义	36
4.3.2	计算	37
4.4	相关性质	42
4.4.1	上下界	42
4.4.2	自同构等价公理	43
4.4.3	路径图	44
4.5	快速近似算法	48
4.5.1	误差分析	48
4.5.2	算法设计	56
4.6	实验	57
4.6.1	区分能力	57
4.6.2	近似算法	61
4.7	本章小结	64

5 边中心性度量及其区分能力研究	66
5.1 引言	66
5.2 边中心性度量	66
5.2.1 边介数中心性	66
5.2.2 生成边中心性	67
5.2.3 双调和距离中心性	67
5.3 边之间的等价关系	68
5.4 边中心性的自同构等价公理	70
5.4.1 公理	70
5.4.2 相关证明	71
5.5 区分能力指数	73
5.6 森林边中心性	76
5.6.1 定义	76
5.6.2 计算	79
5.7 相关性质	85
5.7.1 上下界	85
5.7.2 自同构等价公理	87
5.8 实验	88
5.8.1 示例网络	88
5.8.2 现实网络	89
5.8.3 模型网络	92
5.9 本章小结	93
6 总结与展望	95
6.1 全文总结	95
6.2 未来工作展望	96
参考文献	99
攻读学位期间的研究成果	107
致谢	109

插图索引

1.1	论文结构图	6
3.1	示例网络, 同一轨道中的点用相同颜色标出	17
3.2	一个包含 10 个节点的正则网络, 不同的轨道用不同的颜色标出	23
4.1	示例网络 G_0 以及它的全部 40 个置根生成森林.	36
4.2	示例网络 G_0 的置根生成森林集合 \mathcal{F}_{11} 、 \mathcal{F}_{22} 、 \mathcal{F}_{33} 以及 \mathcal{F}_{44}	38
4.3	映射 $f: \mathcal{S}_u \rightarrow \mathcal{F}$ 的构造方式	40
4.4	含 n 个节点的路径图 G_n^p	44
4.5	示例网络, 等价节点用相同颜色标出	57
5.1	一个包含 7 个节点和 9 条边的示例网络	69
5.2	示例网络 G_0 以及它的全部 40 个置根生成森林	77
5.3	集合 $\mathcal{F}_{34}(G_0)$ 中全部 9 个置根生成森林以及 $(3, 4)$ 在每个森林中的得分	79
5.4	映射 $f: \mathcal{S}_{uv}^1 \rightarrow \mathcal{F}_{uv}^*$ 的构造方式	81
5.5	映射 $g: \mathcal{S}_{uv}^0 \rightarrow \mathcal{F}_{vu}^*$ 的构造方式	83
5.6	一个包含 7 个节点和 9 条边的示例网络	88

表格索引

3.1	图 3.1 所示网络的全部 12 个自同构	18
3.2	图 3.2 所示网络的全部 4 个自同构	24
3.3	图 3.2 所示网络所有节点的中心性值	24
3.4	现实网络数据的统计信息	27
3.5	六个常用中心性度量在现实网络和模型网络中的 P_c 值	28
3.6	有向加权现实网络的统计信息以及 PR 和 EC 的 P_c 值 ($\times 10^{-2}$) . . .	32
4.1	图 4.5 所示网络所有节点的中心性值	58
4.2	现实网络数据的统计信息	59
4.3	中心性度量的广义区分度	60
4.4	现实网络的统计信息以及 Exact 和 Approx 的运行时间	62
4.5	算法 Approx 在不同 ϵ 取值下的平均相对误差	63
5.1	图 5.1 所示网络的唯一非平凡自同构	70
5.2	图 5.1 所示网络中所有边的中心性值.	74
5.3	图 5.6 所示网络的唯一非平凡自同构	89
5.4	图 5.6 所示网络中所有边的中心性值.	89
5.5	现实网络数据的统计信息	90
5.6	边中心性度量在现实网络上的广义区分度	91
5.7	边中心性度量在三种模型网络上的广义区分度	92

第一章 绪论

1.1 研究背景和意义

现实世界中的许多事物都可以用相互关联的实体组成的系统来描述。例如，计算机网络可以用大量相互连接的计算机实体组成的复杂系统来表示 [1]；人类社会可以用大量有着各自社交关系的人类实体组成的社交系统来表示 [2]；多细胞生物有时也可以被描述成大量有着相互作用关系的细胞实体组成的复杂系统 [3]。学者们对这些系统进行了多个方面的研究。有人研究了这些系统中实体的性质，如计算机的工作原理、人类的行为方式、细胞的特征等；有人则会研究实体间互动的方式，如互联网上使用的通信协议、人际关系的形成过程 [2]、细胞间的相互作用 [3] 等。然而，实体之间的连接模式往往会对这些系统的运行产生更为重要的影响。这也是网络科学研究的重点。系统的连接模式可以用网络的形式来表示，其中实体用节点来表示，实体之间的关系用边来表示。学者们经过大量研究后发现，网络结构有时可以对系统的行为产生重大的影响。例如，计算机网络的结构可以影响数据在网络上的传输路径和传输效率；社交网络的结构会影响人们形成观点和收集新闻的方式。因此，研究网络的结构对于人们理解对应系统的运行方式至关重要。网络科学的本质是将现实世界的复杂系统简化为仅保留连接模式等基本信息的抽象结构后进行相应的研究。尽管网络中的节点和边可以携带附加信息，如节点的名称、边的权重、节点的附加属性等，来保留现实系统中更多的信息，但即使如此，这种方法在将现实世界的系统简化为相应的网络表示的过程中通常会丢失大量的信息。这可能给后续的研究造成不小的阻碍，但有时也能很好的排除无用信息的干扰，从而对后续的研究工作产生积极的作用。多年来，各个领域的科学家们开发了许多工具来分析、建模和理解网络，如中心性度量 [4-9]、最短路径算法 [10-14]、图嵌入 [15-25] 等。由于这些工具应用于复杂网络这个抽象结构，所以理论上这些工具可以用于研究几乎任何能用网络表示的现实系统。因此，复杂网络是用于描述现实系统的一种通用而强大的方式。

复杂网络 [26-29, 6, 30-31] 作为一种描述现实世界复杂关系的数学模型，近年来在各个领域的研究中展现出了巨大的潜力和重要性。复杂网络在社交网络、生物网络、信息网络和社交网络等领域的广泛应用已经深刻地改变了人们对事物之间相互联系的认识。复杂网络不仅有助于理解现实世界中的复杂性，还为解决

实际问题提供了有力的工具和方法。一方面, 复杂网络模型揭示了现实世界中复杂系统的结构和性质。通过对复杂网络的拓扑结构进行分析, 可以发现网络中的重要节点、社团结构以及信息传播的规律。这些发现对于社会科学、生物学、物理学等学科有重要的意义, 且为人们理解人类社会、生物系统、信息传播等现象提供了新的视角。另一方面, 复杂网络模型也可以成为解决实际问题的具体工具。通过对复杂网络的中心性等关键指标进行研究, 可以识别出网络中的重要节点和关键链接, 从而为优化网络结构、提高网络效率提供指导。总之, 复杂网络模型对于深入理解现实世界的复杂系统、解决问题具有重要的意义。

寻找关键点或关键边的问题作为社交网络研究 [32]、人工智能 [7-9] 以及网络结构分析 [4-5] 等领域的一个基本问题有着悠久的历史 and 广泛的应用 [6]。例如, 在控制论领域, 节点中心性已被用于一致性协议 [33]、同步同步校准 [34] 和上下文分析 [35]。因此, 为了更好的评估网络中点和边的重要性, 学者们提出了多种衡量点和边相对重要性的方法 [4], 即中心性度量。这些中心性度量各具特点, 它们衡量网络中点和边重要性的方式也各不相同。因此, 在不同的应用场景中, 研究者们往往需要不同的中心性度量来解决问题。例如, 介数中心性能很好的解释 Medici 家族的崛起过程 [36-38]; Katz 中心性有助于人们理解社交乘数的概念 [39]; 扩散中心性有助于人们了解社交网络中信息传播的过程 [40-41]; 研究者们可以用特征向量中心性来判断一个社交网络是否可以很好的汇总意见 [42]; 度中心性有助于人们理解社会规范中的系统性偏差 [43]。

网络科学的主要工作是研究网络的结构、行为以及动力学, 而中心性度量则是评估网络中节点的重要性或影响力的重要手段。从上文中的几个例子中可以看出, 找到合适的中心性度量对于能否很好的解决网络科学中的具体问题非常重要。要想选到合适的中心性度量, 就必须深入理解这些中心性度量的相关特性。中心性度量数量繁多, 包括度中心性、接近度中心性、中介中心性和特征向量中心性等。每种中心性度量都有其独特的计算方式和解释方式, 适用于不同类型的网络和研究问题。研究者们需要深入理解这些中心性度量的特性, 以帮助他们选择适合特定问题的中心性度量, 并对研究结果进行合理的解释。

因此, 研究中心性度量的特性是十分必要的。通过研究中心性度量的特性, 研究者们可以了解不同中心性度量的优势和局限性, 从而更好的选择适合具体问题的中心性度量。此外, 深入理解中心性度量的特性还可以帮助研究者们发现网络中节点和边的重要特征, 进而深入理解网络结构和节点之间的相互关系。在这方面, 本文对中心性度量的区分能力这一特性进行了系统性的研究。区分能

力是评估中心性度量的一个重要指标，它衡量了中心性度量在区分节点重要性方面的能力。本文提出了一些相关的重要理论，并成功将中心性度量的区分能力量化表示出来。本文通过对不同中心性度量的比较和分析，揭示了这些中心性度量在不同网络中的适用性和效果差异。

本文的研究对深入理解中心性度量的区分能力具有重要的意义。它为研究者们提供了一个理论框架和实证分析方法，帮助研究者们评估和选择合适的中心性度量。通过选取区分能力更好的中心性度量，研究者们可以更准确地揭示网络中节点的重要性，为网络科学的相关领域提供更好的理论支持和实践指导。

1.2 研究历史和现状

为了深入理解各种中心性度量的特性，学者们进行了大量的研究，并取得了一些重要的研究成果。本节将介绍一些与本文研究内容密切相关的研究成果。

1.2.1 中心性公理

为了深入了解度中心性、接近中心性和衰减中心性等一些中心性度量的特征，Manuj Garg 于 2009 年率先提出了用公理化的方式刻画中心性度量特征的研究方法 [44]。这种方法通过明确的公理化表达，使得学者们能够更加准确地描述中心性度量的特征。随后，越来越多的中心性公理被提出 [45-49]，例如匿名性公理 [45]、得分单调性公理 [49]、局部性公理等 [50]。这些中心性公理对人们深入理解中心性度量的特性而言是不可或缺的。例如，一个满足匿名性公理的中心性度量对节点的评判结果不会随着节点标号的改变而改变 [45]；一个满足得分单调性公理的中心性度量在节点新连上一条边时会赋予这个节点更高的重要性得分 [49]。这些中心性公理不仅使学者们能够深入理解中心性度量的相关特性，还激发了他们提出更多新的中心性度量 [51, 50, 52-55]。通过对中心性公理的研究，学者们可以设计出满足不同要求的中心性度量，从而更好地应用于不同的领域和问题。然而，中心性公理也存在着局限性。对中心性公理而言，一个中心性度量要么满足公理，要么不满足公理。这使得学者们对中心性度量特性的研究停留在定性分析的阶段，无法进一步对中心性度量的特性进行定量化的研究。虽然公理化的方法为中心性度量的研究提供了重要的框架，但人们仍然需要更多的研究来进一步探索中心性度量的定量特征以及其在实际应用中的效果。

1.2.2 中心性度量的敏感性研究

与此同时，学者们还分别从实验以及理论的角度对点中心性度量的敏感性做了一系列系统性的研究。就实验方面而言，作者们在 [56] 中实验性地分析了各种中心性对图的某些修改的敏感性。在 [57] 中，学者们通过实证研究揭示了网络拓扑结构对中心性度量敏感性的影响。此外，研究者们还在 [58] 中通过节点采样生成图的方法进行了中心性度量敏感性的分析。同时，学者们还从理论角度对中心性度量的敏感性进行了探索。在 [59] 中，学者们对 PageRank 中心性的敏感性进行了理论分析。在 [60] 和 [61] 中，作者们评估了度中心性对网络中随机错误的敏感性。在 [62] 中，学者们对加权网络上的中心性度量进行了敏感性分析。最近，有学者提出了一种评价中心性度量敏感性的量化指标，使得人们可以衡量中心性度量在网络中随机边被删除时的敏感性，并应用该量化指标比较了六种常用中心性度量的敏感性 [63]。这一系列对中心性度量敏感性的研究使得人们对中心性度量的敏感性有了更为深刻的认识。通过这些实验结果和理论研究成果的相互支持，人们可以更好的理解不同中心性度量在敏感性方面的特点，这也为网络科学的进一步发展提供更加坚实的基础。然而，中心性度量的特征是多样化的，所以仅仅从敏感性这一角度去研究中心性的特性是远远不够的。中心性度量的选择应该根据具体的应用场景和研究问题来进行，在考虑到中心性度量的敏感性的同时还需要考虑其他因素，如计算效率、中心性度量的可解释性等。若想全面了解中心性度量的特征，学者们还需进一步研究中心性度量在其它方面的特性。

1.2.3 中心性度量的区分能力研究

此外，尽管有多篇文献 [64-65, 53, 66] 提到过中心性度量区分能力的概念，但这些研究仅仅局限于在某些简单的示例网络中比较不同中心性度量区分能力。它们并没有对中心性度量的区分能力进行系统性的研究，导致了人们至今对中心性度量区分能力的概念缺乏清晰的认识。

1.3 本文的主要贡献

一、基于自同构等价的概念提出了节点中心性度量区分能力的评价标准。本文深入的研究了节点中心性度量的区分能力并基于节点自同构等价的概念提出了节点中心性度量区分能力的量化评价指标。根据等价节点中心性值应当相等的原则，本文基于节点自同构等价的概念提出了相应的自同构等价公理，规定了

节点中心性度量对于自同构等价的节点必须给出相等的中心性值。本文证明了几个常用的中心性度量均满足本文提出的自同构等价公理。本文再根据不等价节点的中心性值应当不同的原则，用节点中心性度量对于不等价的节点给出不同中心性值的概率定义了区分能力指数的概念。在满足自同构等价公理时，一个中心性度量的区分能力越强，它的区分能力指数就越大。同时，本文还提出了广义区分度的概念用以快速比较两种节点中心性度量的区分能力。本文在多个现实网络和模型网络上用上述方法深入的分析了几种常用的节点中心性度量的区分能力。此外，本文还把上述分析节点中心性度量区分能力的方法扩展到了有向加权网络。

二、基于置根生成森林提出了森林中心性并提出了相应的快速近似算法。为探究是否存在区分能力更为出色的节点中心性度量，本文基于置根生成森林的概念提出了森林中心性这一新的节点中心性度量。森林中心性基于以某个节点为根的所有置根生成森林中这个点所在连通分量的平均大小来定义一个节点的相对重要程度。相较于其它常用的节点中心性度量，森林中心性不仅可以作用于不连通的网络，它还有着更为出色的区分能力。此外，本文还基于 JL 引理以及快速求解器设计了一个用于计算森林中心性值的快速近似算法。该算法能够用几乎线性的时间和空间复杂度给出一个网络中所有节点的森林中心性的准确近似值。本文在多个现实网络和模型网络上比较了森林中心性和其他几个常用的节点中心性度量的区分能力。实验结果表明，相较于其他几个常用的节点中心性度量，森林中心性确实有着更为强大的区分能力。此外，本文在包含数百万节点的现实网络中验证了快速近似算法的运算效率和准确性。

三、提出了节点中心性度量区分能力的评价标准并提出了森林边中心性。本文还深入的研究了边中心性度量的区分能力并基于自同构的概念提出了边中心性度量区分能力的量化评价指标。本文首先基于自同构的概念给出了边自同构等价的定义。再根据等价边的中心性值应当相等的原则，本文基于边的自同构等价提出了相应的自同构等价公理，规定了边中心性度量对于自同构等价的边须给出相等的边中心性值。本文证明了几个常用的边中心性度量均满足边中心性的自同构等价公理。本文接着根据不等价边的中心性值应当不同的原则，用边中心性度量对于不等价边给出不同中心性值的概率定义了边中心性的区分能力指数。在满足自同构等价公理时，一个边中心性度量的区分能力越强，它的区分能力指数就越大。同时，本文还提出了广义区分度的概念用以快速比较两种不同边中心性度量的区分能力。为得到区分能力更强的边中心性，本文还基于置根生成

森林的概念提出了森林边中心性这一新的边中心性度量。本文在多个现实网络和模型网络上用上述方法深入的分析了几种常用的边中心性度量以及新提出的森林边中心性的区分能力。实验结果表明，森林边中心性明显拥有更加优秀的区分能力。

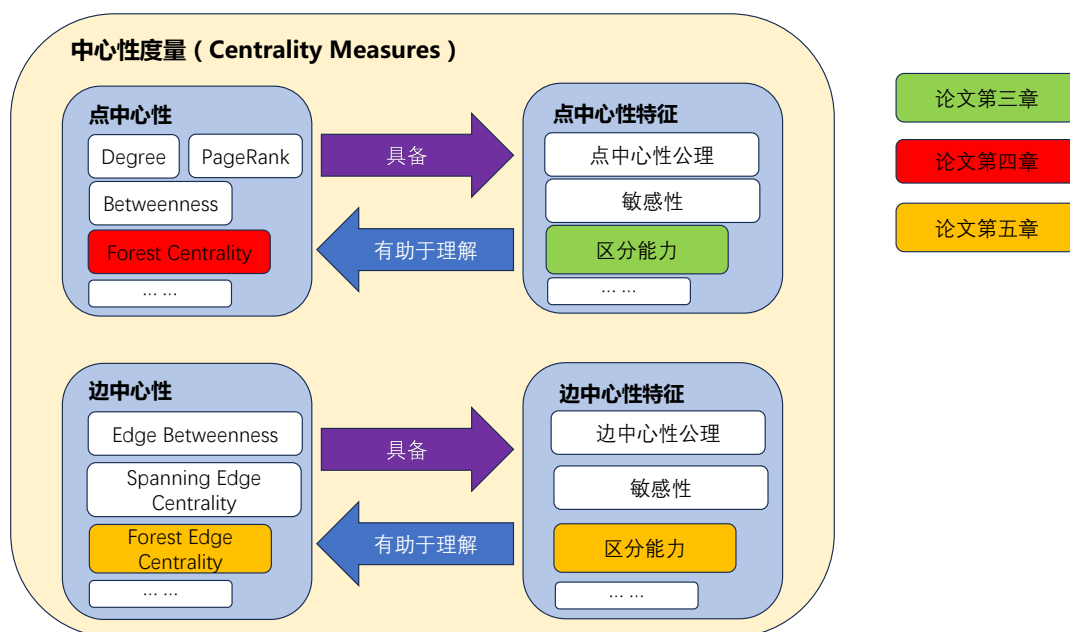


图 1.1: 论文结构图

1.4 论文结构

本文的结构详见图 1.1。本文后续的章节安排如下：

第二章主要介绍一些基础知识以及一些重要的概念。内容包括图、生成树、电阻网络、基尔霍夫系数等重要概念。

第三章重点研究了节点中心性度量区分能力的评价标准问题，给出了节点中心性度量的自同构等价公理，同时提出了区分能力指数的概念。本章指出，满足自同构等价公理的节点中心性度量的区分能力指数越大，意味着它的区分能力越强。本章还提出了一种快速比较两种不同的节点中心性度量区分能力指数的方法。本章最后在多个现实网络和模型网络上，用上述方法分析了几种主流点中心性度量的区分能力。此外，本章还把上述分析节点中心性度量区分能力的方法扩展到了有向加权网络。

第五章基于置根生成森林的概念提出了一种新的点中心性度量，即森林中心性。本章分析了森林中心性的相关特性，并比较了它和其它几个主流点中心性度量的区分能力。结果表明，森林中心性的区分能力优于其它几个主流点中心性度量。此外，本章还基于 JL 引理以及快速求解器设计了一个用于计算森林中心性值的快速近似算法。

第四章主要研究边中心性度量区分能力的评价标准问题，给出了边中心性度量的自同构等价公理，同时提出了边中心性度量区分能力指数的概念。本章认为，在满足自同构等价公理的前提下，一个边中心性度量的区分能力指数越高，它的区分能力就越强。同时，本章还提出了一种快速比较两种边中心性度量区分能力指数的方法。最后，本章在多个现实网络和模型网络上，用上述方法分析了几种主流边中心性度量的区分能力。

第五章基于置根生成森林的概念提出了一种新的点中心性度量，即森林中心性。本章分析了森林中心性的相关特性，并比较了它和其它几个主流点中心性度量的区分能力。结果表明，森林中心性的区分能力优于其它几个主流点中心性度量。此外，本章还提出了一种快速计算森林中心性近似值的算法。

第六章基于置根生成森林的概念提出了一种新的边中心性度量，即森林边中心性。本章分析了森林边中心性的一些特性，并比较了它和其它几个主流边中心性度量的区分能力。结果表明，森林边中心性的区分能力优于其它几个主流边中心性度量。

第七章总结了全文的研究内容，同时展望未来的研究方向。

第二章 预备知识

2.1 引言

本章介绍文中使用到的一些符号、概念和定义，包括图及其矩阵表示、生成树、电阻网络等。

2.2 图及其矩阵表示

令 $G = (V, E)$ 为包含 n 个点和 m 条边的无权无向图，其中 $V = \{1, 2, \dots, n\}$ 表示点集， $E = \{(u_1, v_1), (u_2, v_2), \dots, (u_m, v_m)\}$ 表示边集。若 G 是无环连通的，则 G 可以被称作树。图 G 的邻接矩阵 \mathbf{A} 是一个 $n \times n$ 的矩阵，其中矩阵 \mathbf{A} 的第 i 行第 j 列元素 a_{ij} 表示节点 i 和 j 之间的邻接关系。若 $(i, j) \in E$ ，则 $a_{ij} = a_{ji} = 1$ ，否则 $a_{ij} = a_{ji} = 0$ 。图 G 的度对角矩阵是 $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_n)$ ，其中 $d_i = \sum_{j=1}^n a_{ij}$ 表示节点 i 的度。图 G 的拉普拉斯矩阵定义为 $\mathbf{L} = \mathbf{D} - \mathbf{A}$ ，而拉普拉斯矩阵 \mathbf{L} 的 Moore-Penrose 逆，简称伪逆，可以表示为 $\mathbf{L}^\dagger = (\mathbf{L} + \mathbf{J})^{-1} - \mathbf{J}$ ，其中 $\mathbf{J} = \frac{1}{n} \mathbf{I} \mathbf{I}^\top$ ， \mathbf{I} 是 n 维全 1 向量。

若给图 G 的每条无向边任意分配一个方向，即给每条边指定一个头和一个尾，就能得到图 G 的一个 $m \times n$ 的带符号的关联矩阵 \mathbf{B} 。令 b_{ij} 表示关联矩阵 \mathbf{B} 第 i 行第 j 列的元素。若节点 j 是第 i 条无向边的头，则有 $b_{ij} = 1$ ；若节点 j 是第 i 条无向边的尾，则 $b_{ij} = -1$ ；否则 $b_{ij} = 0$ 。上文中提到的拉普拉斯矩阵 \mathbf{L} 也可以用关联矩阵 \mathbf{B} 来表示，即 $\mathbf{L} = \mathbf{B}^\top \mathbf{B}$ 。根据这个表达形式可以观察到，拉普拉斯矩阵 \mathbf{L} 是一个半正定矩阵，且若图 G 是连通的，则它的拉普拉斯矩阵 \mathbf{L} 的特征值中仅有一个特征值为 0，其余特征值均为正数。

2.3 生成树

若无向图 $G = (V, E)$ 的一个子图 $T = (V_T, E_T)$ 满足 $V_T = V$ 以及 $E_T \subseteq E$ ，且子图 T 是无环连通的，则子图 T 是图 G 的一个生成树 (spanning tree)。换言之，图的生成树就是包含原图所有顶点的无环连通子图。一般来说，一个图往往有多个生成树，但非连通的图没有任何生成树。若无向图 G 本身是一棵树，则此时 G 有唯一的一棵生成树，即图 G 本身。本文用 $\mathcal{T}(G)$ 来表示图 G 的所有生成树构成的集合，用 $|\mathcal{T}(G)|$ 表示图 G 的生成树个数。

一个图的生成树的数量可能十分庞大，例如，一个包含 n 个节点的完全图的生成树的数量可达 n^{n-2} 。因此，用逐一枚举的方法计算一个图的生成树个数是不切实际的。在大多数情况下，一个图的生成树个数往往通过基尔霍夫矩阵树定理（Kirchhoff matrix-tree theorem）来计算 [67]。

定理 2.1 (基尔霍夫矩阵树定理) 给定一个包含 n 个节点的无权无向图 $G = (V, E)$ ，其拉普拉斯矩阵为 \mathbf{L} 。令 \mathbf{L}_{-u} 表示由矩阵 \mathbf{L} 删去第 u 行和第 u 列后得到的子矩阵，则图 G 的生成树的数量，记作 $|\mathcal{T}(G)|$ ，可以表示为

$$|\mathcal{T}(G)| = |\mathbf{L}_{-u}|,$$

其中 $|\mathbf{L}_{-u}|$ 表示子矩阵 \mathbf{L}_{-u} 的行列式值。

2.4 电阻网络

图 $G = (V, E)$ 也可以用来构造一个电阻网络（electrical network） $G_R = (V, E, r)$ 。图 G 中的每条无权无向边都对应电阻网络中一个阻值为 1 的电阻，图 G 中的节点对应电路中的连接点。本文用 $R(u, v)$ 表示电阻网络 G_R 中节点 u 和节点 v 之间的等效电阻（effective resistance），即节点 u 和节点 v 之间的电势差。令 $\mathbf{i} \in \mathbb{R}^m$ 表示每条边上的电流，令 $\mathbf{v} \in \mathbb{R}^n$ 表示每个节点上的电压，再令 $\mathbf{c} \in \mathbb{R}^n$ 表示电阻网络中每个节点上净流入的电流，则根据基尔霍夫定律 $\mathbf{B}^\top \mathbf{i} = \mathbf{c}$ 以及欧姆定律 $\mathbf{i} = \mathbf{B}\mathbf{v}$ 可以推导出 $\mathbf{L}\mathbf{v} = \mathbf{c}$ 。由此可知，任意两个节点对之间的等效电阻可以用图 G 的拉普拉斯矩阵 \mathbf{L} 的伪逆 \mathbf{L}^\dagger 的一个二次型来表示 [68]。

定理 2.2 (等效电阻定理) 给定一个无权无向图 $G = (V, E)$ ，其拉普拉斯矩阵为 \mathbf{L} ，则图 G 中节点 u 和节点 v 之间的等效电阻，记作 $R(u, v)$ ，可以表示为

$$R(u, v) = (\mathbf{e}_u - \mathbf{e}_v)^\top \mathbf{L}^\dagger (\mathbf{e}_u - \mathbf{e}_v),$$

其中 \mathbf{L}^\dagger 为矩阵 \mathbf{L} 的伪逆。

文献 [68] 证明了等效电阻是一种距离函数，即等效电阻同时满足非负性、同一性、对称性以及三角不等式。因此，等效电阻有时又被称作电阻距离。在不引起混淆的情况下，本文有时会交替使用这两个等价的概念。

介绍完等效电阻的相关概念后，本节接下来介绍基尔霍夫系数（Kirchhoff index）的概念 [68]。一个图的基尔霍夫系数等于这个图中所有节点对之间等效电阻之和，其具体表述如下。

定义 2.3 (基尔霍夫系数) 令 $G = (V, E)$ 表示一个无权无向图，其拉普拉斯矩阵记为 \mathbf{L} ，则图 G 的基尔霍夫系数，记作 $K(G)$ ，定义为

$$K(G) = \sum_{i < j} R(i, j),$$

其中 $R(i, j)$ 表示节点 i 和节点 j 之间的等效电阻。

基尔霍夫系数的概念十分重要，它在许多应用场景中都有着举足轻重的作用 [69-70]。

第三章 点中心性区分能力的评价标准

3.1 引言

本章主要研究节点中心性度量在无权无向简单图上的区分能力。本章首先在 3.2 节中介绍六中常用的节点中心性度量，随后在 3.3 节中确定节点等价的标准。之后，本章将分别讨论中心性度量在节点等价时和不等价时的情况，其中节点等价时的情况在 3.4 节中讨论，节点不等价时的情况在 3.5 节中讨论。在建立了中心性度量区分能力的理论分析框架后，3.6 节将在现实网络以及模型网络中测试六个常用的中心性度量的区分能力。在 3.7 节中，本章的研究成果将被推广至有向加权图。3.8 节将总结本章的研究成果。

3.2 点中心性度量

过去多年，学者们提出了许多中心性度量。在这一节中，我们主要回顾六种常用的点中心性度量，分别是：度中心性（Degree Centrality, DC）、介数中心性（Betweenness Centrality, BC）、接近中心性（Closeness Centrality, CC）、PageRank 中心性（PR）、特征向量中心性（Eigenvector Centrality, EC）以及信息中心性（Information Centrality, IC）。

3.2.1 度中心性

度中心性是最早的中心性度量之一。节点 u 的度中心性指的是与 u 点直接相连的节点的数量，其具体定义如下 [71]。

定义 3.1 (度中心性) 在无权无向图 $G = (V, E)$ 中，节点 u 的度中心性定义为：

$$f_{\text{DC}}(u) = d_u = \mathbf{e}_u^\top \mathbf{A} \mathbf{1},$$

其中 \mathbf{e}_u 是第 u 维为 1 其余全为 0 的单位向量。

度中心性的计算成本很低。在一个包含 m 条边的图中，计算所有节点的度中心性的时间复杂度仅为 $O(m)$ 。然而，度中心性也有其局限性，它仅仅表示了节点的邻居个数，并未考虑到更多的全局信息。

3.2.2 介数中心性

一个节点的介数中心性可以理解为这个节点出现在其它节点对的最短路径上的概率，其具体定义如下 [72]。

定义 3.2 (介数中心性) 在无权无向图 $G = (V, E)$ 中，节点 u 的介数中心性的定义如下：

$$f_{\text{BC}}(u) = \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}},$$

其中 σ_{st} 表示节点 s 和 t 之间最短路径的个数，而 $\sigma_{st}(u)$ 表示节点 s 和 t 之间经过节点 u 的最短路径的个数。

介数中心性考虑了最短路径这个全局信息，然而它的计算成本偏高，在一个包含 n 个节点和 m 条边的简单图中，计算所有节点的介数中心性的时间复杂度为 $O(nm)$ [73]。

3.2.3 接近中心性

一个节点的接近中心性指的是从这个节点出发到图中其它各节点的距离之和的倒数，其具体定义如下 [74-75]。

定义 3.3 (接近中心性) 在无权无向图 $G = (V, E)$ 中，节点 u 的接近中心性的定义如下：

$$f_{\text{CC}}(u) = \frac{1}{\sum_{v \neq u} d(u, v)},$$

其中 $d(u, v)$ 是节点 u 与节点 v 之间最短路径的长度。

接近中心性考虑了距离这个全局信息，然而它有两个明显的局限性：计算复杂度偏高以及不能应用于不连通图。在一个包含 n 个节点和 m 条边的图中，计算所有节点的接近中心性的时间复杂度是 $O(nm)$ 。

3.2.4 PageRank 中心性

PageRank 作为一种知名的中心性度量曾经在 Google 搜索引擎以及其它商业场景中有着广泛的应用。PageRank 中心性的具体定义如下 [76]。

定义 3.4 (PageRank 中心性) 令 $G = (V, E)$ 为一个无权无向的简单图。令 $\alpha \in (0, 1)$ 为随机跳跃因子，且用 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 表示一个 n 维列向量。 \mathbf{x} 的所有分量均大于等于 0（即 $\mathbf{x} \geq \mathbf{0}$ ），且满足 $\mathbf{x}^\top \mathbf{I} = 1$ 以及

$$(\mathbf{I} - \alpha \mathbf{A} \mathbf{D}^{-1}) \mathbf{x} = \frac{1 - \alpha}{n} \mathbf{I},$$

其中 \mathbf{I} 为 n 阶单位矩阵。则在图 G 中，节点 u 的 *PageRank* 中心性定义为 $f_{\text{PR}}^\alpha(u) = x_u$ 。

PageRank 中心性提供了节点的全局信息，而且它的计算复杂度很低。在一个包含 m 个节点的图中，计算所有节点的 *PageRank* 中心性的时间复杂度为 $O(m)$ [76]。这两个特性也是 *PageRank* 得以广泛应用的重要原因。

3.2.5 特征向量中心性

学者 Bonacich 在 1971 年提出了用邻接矩阵的最大特征值所对应的特征向量来评估网络中节点的重要性。特征向量中心性认为一个节点的重要性取决于与它相邻节点的重要性，其具体定义如下 [77]。为了便于后续的证明工作，本文仅考虑经过归一化之后的特征向量。

定义 3.5 (特征向量中心性) 在无权无向图 $G = (V, E)$ 中，令 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 为一 n 维列向量，满足 $\mathbf{x}^\top \mathbf{I} = 1$ 以及

$$\mathbf{A} \mathbf{x} = \lambda_{\max} \mathbf{x},$$

其中 λ_{\max} 是邻接矩阵 \mathbf{A} 的最大特征值。则图 G 中节点 u 的特征向量中心性定义为 $f_{\text{EC}}(u) = x_u$ 。

特征向量中心性可以用迭代法快速的计算出来。在一个包含 m 个节点的网络中，计算出所有节点的特征向量中心性所需的时间复杂度是 $O(rm)$ ，其中 r 为迭代次数 [77]。此外，特征向量中心性也可以应用于神经网络 [78]。

3.2.6 信息中心性

信息中心性的思想最早由学者 Stephenson 和 Zelen 提出 [79]。一个节点的信息中心性等于该节点与其它各点之间电阻距离之和的倒数，其数学定义如下 [80]。

定义 3.6 (信息中心性) 在无权无向图 $G = (V, E)$ 中, 节点 u 的信息中心性定义为:

$$f_{IC}(u) = \frac{n}{\sum_{v \neq u} R(u, v)} = \frac{n}{n\mathbf{L}_{uu}^\dagger + \text{Tr}(\mathbf{L}^\dagger)},$$

其中 $R(u, v)$ 表示节点 u 和 v 之间的电阻距离, $\text{Tr}(\mathbf{L}^\dagger)$ 表示矩阵 \mathbf{L}^\dagger 的迹, \mathbf{L}_{uu}^\dagger 表示矩阵 \mathbf{L}^\dagger 第 u 行第 u 列的元素。

信息中心性考虑到了节点之间的所有路径, 然而它的计算成本很高。在一个包含 n 个节点的网络中, 计算出所有节点的信息中心性所需的时间复杂度为 $O(n^3)$, 所需的空间复杂度为 $O(n^2)$ 。

3.3 节点之间的等价关系

在研究中心性度量的区分能力之前, 我们首先要对图中节点之间的等价关系有一个清晰的定义。否则, 人们只需定义一个随机中心性, 即对图中的每个节点赋予一个随机的实数作为它的中心性得分, 就可以使中心性度量的区分能力最大化, 而这显然背离了这项研究的初衷。本节中, 我们将讨论三种常用的节点之间的等价关系, 分别是结构等价 (structural equivalence)、正则等价 (regular equivalence) 以及自同构等价 (automorphic equivalence)。

3.3.1 结构等价

Lorrain 等人在 1971 年提出了结构等价的概念 [81]。他们认为若两个节点的邻居集合完全一样, 那么这两个节点就是等价的。结构等价概念的具体表述如下。

定义 3.7 (结构等价) 在无权无向图 $G = (V, E)$ 中, 令 $N_u = \{v \mid (u, v) \in E\}$ 为节点 u 的邻居集合。若节点 u 和节点 v 满足 $N_u = N_v$, 则节点 u 与节点 v 结构等价。

结构等价的优势是计算成本很低, 判断两个节点是否结构等价只需求出它们的邻居集合然后排序依次比较即可。然而, 在许多应用场景中, 结构等价的定义却往往与常识相违背。例如, 在包含四个节点的路径图 $G_4^p = (\{1, 2, 3, 4\}, \{(1, 2), (2, 3), (3, 4)\})$ 中, 节点 1 的邻居集合是 $N_1 = \{2\}$, 节点 4 的邻居集合是 $N_4 = \{3\}$ 。显然, $N_1 \neq N_4$, 所以根据定义 3.7 可知节点 1 与节点 4 不是结构等价的。但我们根据常识认为: 路径图的两端应该是等价的。因此, 在这个场景中, 结构等价的定义是与常识相违背的。违反常识是结构等价概念的一个重要缺陷, 也是结构等价未能得到广泛应用的重要原因。

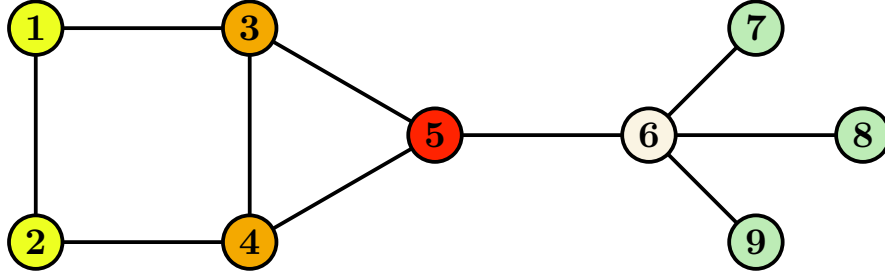


图 3.1: 示例网络, 同一轨道中的点用相同颜色标出

3.3.2 正则等价

除了上文提到的结构等价之外, 学者们还提出了其它节点等价的概念, 而正则等价正是其中之一。正则等价的概念认为若两个节点是等价的, 那么它们的邻居节点也应该是对应等价的, 其具体表述如下 [82]。

定义 3.8 (正则等价) 在无权无向图 $G = (V, E)$ 中, 令 $lab(u)$ 为节点 u 的等价类标签, 令 N_u 为节点 u 的邻居集合。若存在双射 $f: N_u \rightarrow N_v$ 使得对于任意的 $u' \in N_u$ 均成立 $lab(u') = lab(f(u'))$, 则节点 u 与节点 v 正则等价且 $lab(u) = lab(v)$ 。

正则等价的概念看似合理, 但它有一个致命的缺陷, 即等价划分不唯一。例如, 在包含四个节点的路径图 $G_4^p = \{\{1, 2, 3, 4\}, \{(2, 3), (2, 3), (3, 4)\}\}$ 中存在两种符合定义 3.8 的标签方案, 第一种是 $lab(1) = lab(4) = 1$ 、 $lab(2) = lab(3) = 2$; 第二种是 $lab(1) = 1$ 、 $lab(2) = 2$ 、 $lab(3) = 3$ 、 $lab(4) = 4$ 。在第一种方案中, 节点 1 和节点 4 是等价的, 而在第二种方案中节点 1 和节点 4 不等价。我们发现, 两个节点在一种标签方案中是等价的, 在另一种方案中可能是不等价的, 这也导致了正则等价在实际应用场景中难有作为。

3.3.3 自同构等价

在介绍自同构等价的概念之前, 我们先介绍自同构的定义。图的自同构可以看作是一个能够使节点之间邻接关系不变的节点的排列, 其具体定义如下 [83-84]。

定义 3.9 (自同构) 在无权无向简单图 $G = (V, E)$ 中, 图 G 的一个自同构是一个双射函数 $f: V \rightarrow V$, 满足对于任意的 $(u, v) \in E$, 都有 $(f(u), f(v)) \in E$ 。

表 3.1: 图 3.1 所示网络的全部 12 个自同构

u	1	2	3	4	5	6	7	8	9
$f_1(u)$	1	2	3	4	5	6	7	8	9
$f_2(u)$	1	2	3	4	5	6	7	9	8
$f_3(u)$	1	2	3	4	5	6	8	7	9
$f_4(u)$	1	2	3	4	5	6	8	9	7
$f_5(u)$	1	2	3	4	5	6	9	7	8
$f_6(u)$	1	2	3	4	5	6	9	8	7
$f_7(u)$	2	1	4	3	5	6	7	8	9
$f_8(u)$	2	1	4	3	5	6	7	9	8
$f_9(u)$	2	1	4	3	5	6	8	7	9
$f_{10}(u)$	2	1	4	3	5	6	8	9	7
$f_{11}(u)$	2	1	4	3	5	6	9	7	8
$f_{12}(u)$	2	1	4	3	5	6	9	8	7

若一个自同构中，所有节点均映射到它自身，则我们称这个自同构为平凡自同构。除此之外的自同构均是非平凡自同构。值得注意的是，任意自同构 f 的逆映射 f^{-1} 都是图 G 的自同构；若 f 和 g 都是图 G 的自同构，则它们的复合 $f \circ g$ 也是图 G 的自同构。

在一个包含 n 个节点的图 $G = (V, E)$ 中， f 是图 G 的一个自同构，我们令 \mathbf{P} 为 f 对应的排列矩阵。在这个 $n \times n$ 矩阵 \mathbf{P} 中，若 $j = f(i)$ ，则 $p_{ij} = 1$ ，否则 $p_{ij} = 0$ 。例如，图 G 的平凡自同构所对应的排列矩阵是 n 阶单位阵 \mathbf{I} 。显然，排列矩阵 \mathbf{P} 是一个正交矩阵，有 $\mathbf{P}\mathbf{P}^\top = \mathbf{P}^\top\mathbf{P} = \mathbf{I}$ 。根据定义 3.9 可以推断出： $\mathbf{P}\mathbf{A}\mathbf{P}^\top = \mathbf{P}^\top\mathbf{A}\mathbf{P} = \mathbf{A}$ 。

基于自同构的概念，我们接下来给出自同构等价的定义 [85]。

定义 3.10 (自同构等价) 在无权无向简单图 $G = (V, E)$ 中，节点 u 和节点 v 是自同构等价的，记为 $u \cong v$ ，当且仅当存在图 G 的一个自同构 f 满足 $f(u) = v$ 。

令 $\text{Aut}(G)$ 为包含图 G 所有自同构的集合。有研究指出，自同构集合 $\text{Aut}(G)$ 以及定义在自同构之间的复合运算构成一个群 [83-84]。基于定义在节点集合 V 上的自同构等价关系，我们可以得到自同构划分 $\mathcal{P} = \{V_1, V_2, \dots, V_k\}$ 。自同构划分中的子集称之为 $\text{Aut}(G)$ 的轨道。显然，在同一个轨道中的节点是相互自同构等价的。

为便于读者理解，我们给出一个例子。如图 3.1 所示，该网络有 9 个节点和 10 条边。它总共有 12 个自同构，如表 3.1 所示。它的自同构划分是

$\mathcal{P} = \{\{1, 2\}, \{3, 4\}, \{5\}, \{6\}, \{7, 8, 9\}\}$, 同一个轨道中的节点在图 3.1 中用相同的颜色标出。根据自同构等价的定义, 我们可以得出 $1 \cong 2$ 、 $3 \cong 4$ 以及 $7 \cong 8 \cong 9$ 。

3.3.4 小结

在这一节中, 我们介绍了三种定义节点之间等价关系的方法, 即结构等价、正则等价以及自同构等价。前两种等价概念都有明显的缺陷: 结构等价虽易于应用却往往与常识相悖; 正则等价的划分结果不唯一难以应用。自同构等价的判定结果基本与常识一致且划分结果唯一。因此, 本文将选用自同构等价作为判定节点是否等价的标准和依据。

3.4 自同构等价公理

本节中, 我们将讨论两个节点在自同构等价时的中心性值。本节将考察 6 个常用的中心性度量, 分别是: 度中心性、介数中心性、接近中心性、PageRank 中心性、特征向量中心性以及信息中心性。我们将证明, 在任意网络中, 上述中心性度量都能保证在节点等价时给出相同的中心性值。在本节最后, 我们将给出中心性度量的自同构等价公理。

3.4.1 相关证明

I 度中心性

定理 3.11 在任意无权无向简单图 $G = (V, E)$ 中, 若节点 u 和节点 v 满足 $u \cong v$, 则它们的度中心性值 $f_{DC}(u)$ 和 $f_{DC}(v)$ 相等, 即 $f_{DC}(u) = f_{DC}(v)$ 。

证明: 因为节点 u 和节点 v 是自同构等价的, 所以必然存在图 G 的一个自同构 f 使得 $f(u) = v$ 。令 \mathbf{P} 为自同构 f 所对应的排列矩阵, 则根据度中心性的相关定义, 即定义 3.1, 有

$$\begin{aligned} f_{DC}(u) &= \mathbf{e}_u^\top \mathbf{A} \mathbf{1} = \mathbf{e}_u^\top (\mathbf{P} \mathbf{A} \mathbf{P}^\top) \mathbf{1} \\ &= (\mathbf{e}_u^\top \mathbf{P}) \mathbf{A} (\mathbf{P}^\top \mathbf{1}) = \mathbf{e}_v^\top \mathbf{A} \mathbf{1} = f_{DC}(v), \end{aligned}$$

得证。 □

II 介数中心性

定理 3.12 在任意无权无向简单图 $G = (V, E)$ 中, 若节点 u 和节点 v 满足 $u \cong v$, 则它们的介数中心性值 $f_{BC}(u)$ 和 $f_{BC}(v)$ 相等, 即 $f_{BC}(u) = f_{BC}(v)$ 。

证明: 由 $u \cong v$ 可知, 存在图 G 的一个自同构 f 使得 $f(u) = v$ 。令 $\mathbf{p}_{st} = (s, v_1, v_2, \dots, t)$ 表示图 G 中从节点 s 到节点 t 的一条最短路径。由于自同构 f 是一个双射, 所以每个 \mathbf{p}_{st} 都对应一个唯一的从节点 $f(s)$ 至节点 $f(t)$ 的最短路径 $\mathbf{p}_{f(s)f(t)} = (f(s), f(v_1), f(v_2), \dots, f(t))$, 由此可知 $\sigma_{st} = \sigma_{f(s)f(t)}$ 。同理, 也可以证明 $\sigma_{st}(u) = \sigma_{f(s)f(t)}(f(u)) = \sigma_{f(s)f(t)}(v)$ 。根据介数中心性的相关定义, 即定义 3.2, 可知

$$f_{BC}(u) = \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}} = \sum_{s \neq u \neq t} \frac{\sigma_{f(s)f(t)}(v)}{\sigma_{f(s)f(t)}}.$$

又因为自同构 f 是一个双射, 我们有

$$\sum_{s \neq u \neq t} \frac{\sigma_{f(s)f(t)}(v)}{\sigma_{f(s)f(t)}} = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} = f_{BC}(v),$$

得证。 □

III 接近中心性

定理 3.13 在任意无权无向简单图 $G = (V, E)$ 中, 若节点 u 和节点 v 是自同构等价的, 即 $u \cong v$, 则它们的接近中心性值 $f_{CC}(u)$ 和 $f_{CC}(v)$ 相等, 即 $f_{CC}(u) = f_{CC}(v)$ 。

证明: 由 $u \cong v$ 可知, 存在图 G 的一个自同构 f 使得 $f(u) = v$ 。根据定理 3.12 中关于 $\sigma_{st} = \sigma_{f(s)f(t)}$ 的相关证明, 我们可以类似的得出: 节点 u 和节点 i 之间最短路径的长度等于节点 $f(u) = v$ 和节点 $f(i)$ 之间最短路径的长度。因此, 根据接近中心性的相关定义, 即定义 3.3, 可知

$$f_{CC}(u) = \frac{1}{\sum_{i \neq u} d(u, i)} = \frac{1}{\sum_{i \neq u} d(v, f(i))}.$$

又因为自同构 f 是一个双射, 所以

$$\frac{1}{\sum_{i \neq u} d(v, f(i))} = \frac{1}{\sum_{i \neq v} d(v, i)} = f_{CC}(v),$$

得证。 □

IV PageRank 中心性

定理 3.14 在任意无权无向简单图 $G = (V, E)$ 中, 若节点 u 和节点 v 是自同构等价的, 即 $u \cong v$, 则它们的 PageRank 中心性值 $f_{\text{PR}}(u)$ 和 $f_{\text{PR}}(v)$ 相等, 即 $f_{\text{PR}}(u) = f_{\text{PR}}(v)$ 。

证明: 因为节点 u 和节点 v 是自同构等价的, 所以必然存在图 G 的一个自同构 f 使得 $f(u) = v$ 。令 \mathbf{P} 为自同构 f 所对应的排列矩阵, 且令 $\mathbf{x} = (f_{\text{PR}}^\alpha(1), f_{\text{PR}}^\alpha(2), \dots, f_{\text{PR}}^\alpha(n))^\top$ 为图 G 的归一化 PageRank 向量。根据排列矩阵 \mathbf{P} 的特性, 我们有 $\mathbf{Px} \geq \mathbf{0}$ 、 $\mathbf{I}^\top (\mathbf{Px}) = \mathbf{I}^\top \mathbf{x} = 1$ 以及

$$\begin{aligned} & (\mathbf{I} - \alpha \mathbf{AD}^{-1}) (\mathbf{Px}) \\ &= (\mathbf{PP}^\top) (\mathbf{I} - \alpha \mathbf{AD}^{-1}) (\mathbf{Px}) \\ &= \mathbf{P} (\mathbf{P}^\top \mathbf{P} - \alpha \mathbf{P}^\top \mathbf{AD}^{-1} \mathbf{P}) \mathbf{x} \\ &= \mathbf{P} (\mathbf{I} - \alpha \mathbf{AD}^{-1}) \mathbf{x} \\ &= \frac{1-\alpha}{n} \mathbf{PI} = \frac{1-\alpha}{n} \mathbf{I}. \end{aligned}$$

因此, 向量 \mathbf{Px} 也是图 G 的归一化 PageRank 向量。又因为一个给定网络的归一化 PageRank 向量是唯一的 [76], 所以我们有 $\mathbf{Px} = \mathbf{x}$ 。由此可得

$$f_{\text{PR}}^\alpha(u) = \mathbf{e}_u^\top \mathbf{x} = \mathbf{e}_u^\top \mathbf{Px} = \mathbf{e}_v^\top \mathbf{x} = f_{\text{PR}}^\alpha(v),$$

得证。 □

V 特征向量中心性

定理 3.15 在任意无权无向简单图 $G = (V, E)$ 中, 若节点 u 和节点 v 是自同构等价的, 即 $u \cong v$, 则它们的特征向量中心性值 $f_{\text{EC}}(u)$ 和 $f_{\text{EC}}(v)$ 相等, 即 $f_{\text{EC}}(u) = f_{\text{EC}}(v)$ 。

证明: 因为节点 u 和节点 v 是自同构等价的, 所以必然存在图 G 的一个自同构 f 使得 $f(u) = v$ 。令 \mathbf{P} 为自同构 f 所对应的排列矩阵, 令 \mathbf{A} 为图 G 的邻接矩阵, 且令 \mathbf{x} 为矩阵 \mathbf{A} 的最大特征值 λ_{\max} 所对应的归一化特征向量。根据排列矩阵 \mathbf{P} 的特性, 我们有 $\mathbf{I}^\top (\mathbf{Px}) = \mathbf{I}^\top \mathbf{x} = 1$ 以及

$$\mathbf{A} (\mathbf{Px}) = (\mathbf{PP}^\top) \mathbf{A} (\mathbf{Px}) = \mathbf{P} (\mathbf{P}^\top \mathbf{AP}) \mathbf{x}$$

$$= \mathbf{P}\mathbf{A}\mathbf{x} = \lambda_{\max}(\mathbf{P}\mathbf{x}),$$

由此可得向量 $\mathbf{P}\mathbf{x}$ 也是图 G 邻接矩阵 \mathbf{A} 的最大特征值 λ_{\max} 所对应的归一化特征向量。又因为邻接矩阵 \mathbf{A} 的最大特征值 λ_{\max} 所对应的归一化特征向量是唯一的 [77], 我们有 $\mathbf{P}\mathbf{x} = \mathbf{x}$ 。因此

$$f_{\text{EC}}(u) = \mathbf{e}_u^\top \mathbf{x} = \mathbf{e}_u^\top \mathbf{P}\mathbf{x} = \mathbf{e}_v^\top \mathbf{x} = f_{\text{EC}}(v),$$

得证。 \square

VI 信息中心性

定理 3.16 在任意无权无向简单图 $G = (V, E)$ 中, 若节点 u 和节点 v 是自同构等价的, 即 $u \cong v$, 则它们的信息中心性值 $f_{\text{IC}}(u)$ 和 $f_{\text{IC}}(v)$ 相等, 即 $f_{\text{IC}}(u) = f_{\text{IC}}(v)$ 。

证明: 因为节点 u 和节点 v 是自同构等价的, 所以必然存在图 G 的一个自同构 f 使得 $f(u) = v$ 。令 \mathbf{P} 为自同构 f 所对应的排列矩阵。根据自同构的定义, 即定义 3.9 可得, $\mathbf{P}\mathbf{A}\mathbf{P}^\top = \mathbf{P}^\top \mathbf{A}\mathbf{P} = \mathbf{A}$ 。再根据定理 3.11 可知, $\mathbf{P}\mathbf{D}\mathbf{P}^\top = \mathbf{P}^\top \mathbf{D}\mathbf{P} = \mathbf{D}$ 。又因为 $\mathbf{L} = \mathbf{D} - \mathbf{A}$, 所以 $\mathbf{P}\mathbf{L}\mathbf{P}^\top = \mathbf{P}^\top \mathbf{L}\mathbf{P} = \mathbf{L}$, 由此可得

$$\begin{aligned} \mathbf{L}_{vv}^\dagger &= \mathbf{e}_v^\top ((\mathbf{L} + \mathbf{J})^{-1} - \mathbf{J}) \mathbf{e}_v \\ &= \mathbf{e}_u^\top \mathbf{P} ((\mathbf{L} + \mathbf{J})^{-1} - \mathbf{J}) \mathbf{P}^\top \mathbf{e}_u \\ &= \mathbf{e}_u^\top (\mathbf{P}(\mathbf{L} + \mathbf{J})^{-1} \mathbf{P}^\top - \mathbf{P}\mathbf{J}\mathbf{P}^\top) \mathbf{e}_u \\ &= \mathbf{e}_u^\top ((\mathbf{P}\mathbf{L}\mathbf{P}^\top + \mathbf{P}\mathbf{J}\mathbf{P}^\top)^{-1} - \mathbf{J}) \mathbf{e}_u \\ &= \mathbf{e}_u^\top ((\mathbf{L} + \mathbf{J})^{-1} - \mathbf{J}) \mathbf{e}_u = \mathbf{L}_{uu}^\dagger. \end{aligned}$$

根据信息中心性的定义, 即定义 3.5, 可以得出

$$f_{\text{IC}}(u) = \frac{n}{n\mathbf{L}_{uu}^\dagger + \text{Tr}(\mathbf{L}^\dagger)} = \frac{n}{n\mathbf{L}_{vv}^\dagger + \text{Tr}(\mathbf{L}^\dagger)} = f_{\text{IC}}(v),$$

得证。 \square

3.4.2 公理

上一小节中, 我们证明了六个常用的中心性度量对于自同构等价的节点都能保证给出相同的中心性值。由此, 我们可以推测所有的中心性度量都应该在节点自同构等价时给出相同的中心性值。在这一节中, 我们给出中心性度量的自同构等价公理, 其具体表述如下。

公理 3.17 (自同构等价公理) 在无权无向简单图 $G = (V, E)$ 中, 任意中心性度量 c 都应满足以下特性: 对于任意的自同构等价的两个节点 u 和 v , 记为 $u \cong v$, 均成立 $f_c(u) = f_c(v)$, 其中 $f_c(u)$ 表示节点 u 基于中心性度量 c 的中心性值。

自同构等价公理, 即公理 3.17, 是研究中心性度量区分能力的前提。它规定了中心性度量必须在两个节点自同构等价时给出相同的中心性值。若一个中心性度量, 例如给每个节点赋予一个随机的中心性值, 不满足自同构等价公理, 那么它的区分能力就无从谈起。

3.5 区分能力指数

上一节讨论了中心性度量在节点等价时的情况, 给出了中心性度量的自同构等价公理, 即公理 3.17, 规定了中心性度量在节点自同构等价时必须给出相等的中心性值。本节将进一步讨论中心性度量在节点不等价时的情况。

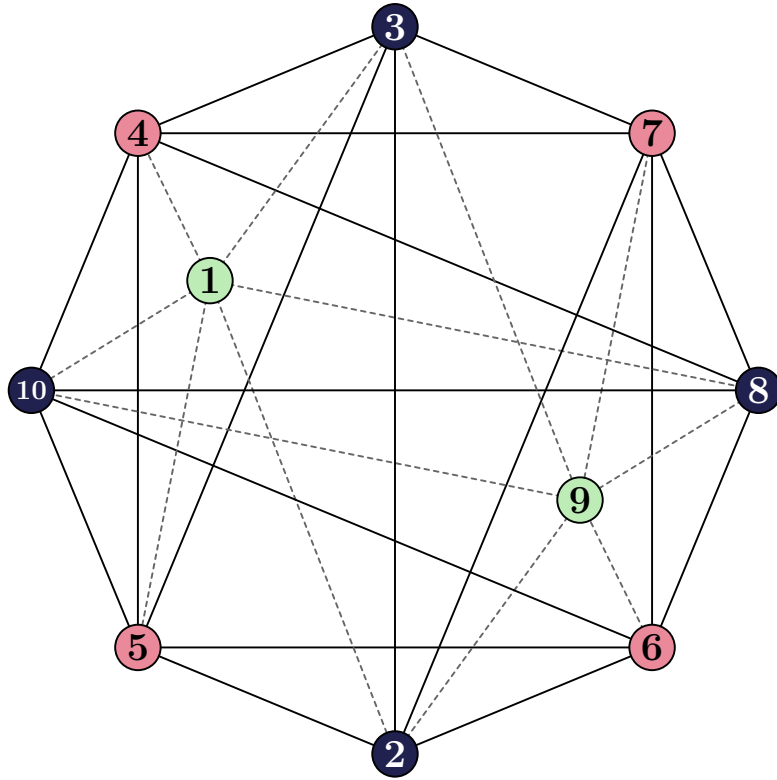


图 3.2: 一个包含 10 个节点的正则网络, 不同的轨道用不同的颜色标出

本质上, 中心性度量的区分能力可以理解为区分不等价节点的能力。换言之, 一个区分能力强的中心性度量能在节点等价时保证给出相等的中心性值, 同时

表 3.2: 图 3.2 所示网络的全部 4 个自同构

u	1	2	3	4	5	6	7	8	9	10
$f_1(u)$	1	2	3	4	5	6	7	8	9	10
$f_2(u)$	1	8	10	5	4	7	6	2	9	3
$f_3(u)$	9	3	2	6	7	4	5	10	1	8
$f_4(u)$	9	10	8	7	6	5	4	3	1	2

表 3.3: 图 3.2 所示网络所有节点的中心性值

u	1,9	2,3,8,10	4,5,6,7
$f_{DC}(u)$	6	6	6
$f_{BC}(u)$	3	3	3
$f_{CC}(u)$	0.8333	0.8333	0.8333
$f_{PR}^{0.85}(u)$	0.1	0.1	0.1
$f_{EC}(u)$	0.1	0.1	0.1
$f_{IC}(u)$	3.5008	3.5035	3.5031

能在节点不等价时尽可能的给出不同的值。实际上,许多常用的中心性度量在节点不等价时往往也会给出相同的中心性值。关于这一点,本节将举例来说明。图 3.2 给出了一个包含 10 个节点的正则网络,其中每个节点的度都等于 6。这个正则网络共有 4 个自同构,如表 3.2 所示。根据自同构等价的定义可以推断出 $1 \cong 9$ 、 $4 \cong 5 \cong 6 \cong 7$ 以及 $2 \cong 3 \cong 8 \cong 10$ 。本节在此列出这个正则网络中所有节点基于 6 个常用中心性度量的中心性值,如表 3.3 所示。从表 3.3 的数据可以看出,所有 6 个常用的中心性度量都能够在节点自同构等价时给出相同的中心性值,这与之前证明的结论是一致的。此外,对于不等价的节点,如节点 1 和节点 2,只有信息中心性给出了不同的中心性值,而度中心性、介数中心性、接近中心性、PageRank 中心性以及特征向量中心性均给出了相同的中心性值。

上述分析指出,所有中心性度量都能在节点等价时给出相同的中心性值,同时只有部分的中心性度量在节点不等价时能够给出不同的中心性值。由此,本文提出了区分能力指数的概念。简而言之,一个中心性度量的区分能力指数是指它在节点不等价时给出不同中心性值的概率,其具体定义如下。

定义 3.18 (区分能力指数) 在无权无向简单图 $G = (V, E)$ 中,若中心性度量 c 满足自同构等价公理,即公理 3.17,则它的区分能力指数,记为 D_c ,定义为中心性度量 c 对于从图 G 的节点集合 V 中随机选出的一对不等价节点给出不同中心

性值的概率，即

$$D_c = P(f_c(u) \neq f_c(v) \mid u \not\cong v),$$

其中 $u \not\cong v$ 表示节点 u 和 v 不是自同构等价的。

根据区分能力指数的定义，即定义 3.18，可知，一个中心性度量的区分能力指数值，即 D_c 值，越高，它的区分能力就越强。然而，直接计算定义 3.18 中的 D_c 值是不现实的。因为计算 D_c 需要找出图 G 的所有轨道，而这个问题到目前为止还不存在一个多项式时间内的解法。为解决区分能力指数难以直接计算的问题，本节接下来将给出一个简单的技巧以实现在不直接计算 D_c 值的情况下比较两个不同中心性度量的区分能力指数。为此，本节先介绍广义区分度的概念，其具体定义如下。

定义 3.19 (广义区分度) 在无权无向简单图 $G = (V, E)$ 中，若中心性度量 c 满足自同构等价公理，即公理 3.17，则它的广义区分度，记为 P_c ，定义为中心性度量 c 对于从图 G 的节点集合 V 中随机选出的一对节点给出不同中心性值的概率，即

$$P_c = P(f_c(u) \neq f_c(v)).$$

接下来，我们将说明区分能力指数与广义区分度之间的关联。

定理 3.20 在无权无向简单图 $G = (V, E)$ 中，令 c_1 和 c_2 为满足自同构等价公理的两个不同的中心性度量，则 $D_{c_1} < D_{c_2}$ 当且仅当 $P_{c_1} < P_{c_2}$ 。

证明： 令 c 为一个满足自同构等价公理的中心性度量，则根据全概率公式，我们有

$$\begin{aligned} P_c &= P(f_c(u) \neq f_c(v) \mid u \cong v) \cdot P(u \cong v) \\ &\quad + P(f_c(u) \neq f_c(v) \mid u \not\cong v) \cdot P(u \not\cong v) \\ &= 0 \cdot P(u \cong v) + D_c \cdot P(u \not\cong v) \\ &= D_c (1 - P(u \cong v)). \end{aligned}$$

由于在一个给定的图中， $P(u \cong v)$ 可视为一个常数，所以 P_c 值等于 D_c 值乘以一个常数。换言之， P_c 值的大小仅由 D_c 值的大小决定，得证。 \square

定理 3.20 表明, 在比较不同中心性度量的区分能力大小时, 广义区分度, 即 P_c , 与区分能力指数, 即 D_c , 效果一致。然而, 相较于区分能力指数, 广义区分度更易于计算。在已知所有节点的中心性值的情况下, 我们可以用快速排序和二分查找算法在 $O(n \log n)$ 的时间内计算出 P_c 的值 [86]。

3.6 实验

本节将在现实网络以及模型网络中用上一节提出的方法来比较六个常用的中心性度量, 即度中心性 (DC)、介数中心性 (BC)、接近中心性 (CC)、PageRank 中心性 (PR)、特征向量中心性 (EC) 以及信息中心性 (IC), 的区分能力。鉴于这六个中心性度量均满足自同构等价公理, 即公理 3.17, 本节实验中可以通过比较它们的广义区分度, 即 P_c 值, 来比较它们的区分能力。在一个包含 n 个节点的网络 $G = (V, E)$ 中, 中心性度量 c 的 P_c 值以如下方式计算。

$$P_c = \frac{|\{(u, v) \mid f_c(u) \neq f_c(v), u, v \in V, u \neq v\}|}{n(n-1)}.$$

在所有的实验中, PageRank 中心性的随机跳跃因子 α 设置为 0.85。此外, 若两个不同节点的中心性值的差距在 10^{-9} 以内, 则视为相等。

3.6.1 现实网络

本节首先在 30 个不同的现实网络中比较六个常用中心性度量的区分能力。这些网络都来自 Koblenz Network Collection [87] 以及 Network Repository [88]。本节实验中的现实网络均为无权无向简单图。本节在每个网络的最大连通分支 (Largest Connected Component, LCC) 上运行实验, 因为接近中心性、特征向量中心性以及信息中心性在不连通图上没有明确的定义。所有现实网络的相关信息如表 3.4 所示, 其中 n 和 m 分别表示网络包含的节点数和边数; n' 和 m' 分别表示网络的最大连通分支包含的节点数和边数。

表 3.5 列出了这六个中心性度量在这些现实网络中的 P_c 值, 每个现实网络中最高的 P_c 值用黄色标出。从表 3.5 所示的实验结果可以看出: ① 度中心性在所有现实网络中的 P_c 值与该网络的最大 P_c 值都有着显著的差距, 这表明度中心性的区分能力普遍较差; ② 介数中心性的 P_c 值有时与最高值比较接近, 如在网络 USFCA72 以及 Bcspwr10 中, 有时则低于度中心性, 如在网络 NipsEgo 和 Reality 中, 这表明介数中心性的区分能力不稳定且在多数时候较差; ③ 接近中心性的 P_c 值在几乎所有的现实网络 (网络 Karate 除外) 中与最高值相差不大, 这表明

表 3.4: 现实网络数据的统计信息

Network	n	m	n'	m'
Karate	34	78	34	78
Dolphins	62	159	62	159
Celegans	453	2,025	453	2,025
Diseasome	516	1,188	516	1,188
GridMouse	1,450	1,636	791	1,098
CrimeMoreno	829	1,473	829	1,473
WikiVote	889	2,914	889	2,914
Reed98	962	18,812	962	18,812
EmailUniv	1,133	5,451	1,133	5,451
GridPlant	1,717	3,098	1,272	2,726
Yeast	1,458	1,948	1,458	1,948
Hamster	2,426	16,630	2,000	16,097
USFCA72	2,682	65,252	2,672	65,244
NipsEgo	2,888	2,981	2,888	2,981
GridWorm	3,507	6,531	3,343	6,437
GrQc	4,158	13,422	4,158	13,422
USgrid	4,941	6,594	4,941	6,594
Erdos992	5,094	7,515	4,991	7,428
Advogato	5,167	39,432	5,054	39,374
Bcspwr10	5,300	8,271	5,300	8,271
Reality	6,809	7,680	6,809	7,680
PagesGovernment	7,057	89,429	7,057	89,429
WikiElec	7,115	100,753	7,066	100,727
Dmela	7,393	25,569	7,393	25,569
HepPh	11,204	117,619	11,204	117,619
Anybeat	12,645	49,132	12,645	49,132
PagesCompany	14,113	52,126	14,113	52,126
AstroPh	17,903	196,972	17,903	196,972
CondMat	21,363	91,286	21,363	91,286
Gplus	23,628	39,194	23,613	39,182

表 3.5: 六个常用中心性度量在现实网络和模型网络中的 P_c 值

Network	不同中心性度量的 P_c 值 ($\times 10^{-2}$)					
	P_{DC}	P_{BC}	P_{CC}	P_{PR}	P_{EC}	P_{IC}
Karate	84.13547	87.87879	94.83066	97.68271	97.68271	97.68271
Dolphins	91.16869	98.09625	98.67795	99.89424	99.89424	99.89424
Celegans	89.00545	93.47907	99.49794	99.94725	99.93358	99.95605
Diseasome	87.78882	52.36773	99.73658	99.78701	99.76594	99.78475
GridMouse	67.84778	70.26997	99.36757	99.38805	97.46451	99.38325
CrimeMoreno	83.59848	93.10706	99.90880	99.96970	99.97028	99.97028
WikiVote	89.67181	93.65214	99.93818	99.98505	99.98556	99.98556
Reed98	98.75887	99.65451	99.89335	99.99762	99.99784	99.99784
EmailUniv	94.20018	96.95748	99.94012	99.99470	99.99485	99.99485
GridPlant	79.58313	79.34672	99.75889	99.77807	93.01830	99.77770
Yeast	69.88786	71.18673	99.78384	99.80116	99.80069	99.80069
Hamster	96.16198	92.34762	99.93692	99.96508	99.96498	99.96518
USFCA72	98.98914	99.77220	99.96393	99.99944	99.99992	99.99992
NipsEgo	6.59288	1.44921	82.88608	82.88608	82.88608	82.88608
GridWorm	67.30131	68.92715	98.91129	98.93478	98.93598	98.93603
GrQc	88.97215	69.68902	99.95909	99.96818	99.76556	99.96876
USgrid	76.98386	91.24768	99.99145	99.99356	10.96066	99.99426
Erdos992	48.54567	41.45339	99.81120	99.82120	99.82166	99.82166
Advogato	92.45751	91.88240	99.96094	99.97316	99.97454	99.97457
Bcspwr10	74.70909	99.43660	99.99722	99.99819	55.56521	99.99919
Reality	14.62505	12.12196	98.54250	98.54259	98.54262	98.54262
PagesGovernment	97.59780	99.10966	99.99083	99.99804	99.99855	99.99888
WikiElec	88.09263	87.77373	99.97009	99.97233	99.98116	99.98116
Dmela	87.05891	92.45899	99.98207	99.98874	99.99039	99.99039
HepPh	93.70279	75.13847	99.97706	99.97926	99.84894	99.98062
Anybeat	72.24760	63.94228	96.94333	96.95285	96.95613	96.95623
PagesCompany	91.27264	94.56037	99.99492	99.99551	99.98403	99.99762
AstroPh	96.58229	81.83606	99.99136	99.99191	99.99209	99.99421
CondMat	92.82196	73.30385	99.99477	99.99396	99.99573	99.99690
Gplus	48.83511	32.75618	98.87605	98.87107	98.87817	98.87817
ER	85.11403	99.44611	99.93232	99.99919	99.99938	99.99938
BA	69.79423	99.97225	99.93265	99.99833	99.99897	99.99897

接近中心性的区分能力较好，且总体上比较稳定；④ PageRank 中心性的 P_c 值在所有现实网络中与最高值几乎相等，这表明 PageRank 中心性有着非常优秀的区分能力，且非常稳定；⑤ 特征向量中心性的 P_c 值在绝大多数现实网络中与最高值几乎相等，然而在网络 USgrid 以及 Bcspwr10 中却很低，甚至低于度中心性的 P_c 值，这表明特征向量中心性的区分能力总体上很优秀但不太稳定；⑥ 信息中心性的 P_c 值在所有现实网络中与最高值几乎相等，这表明信息中心性的区分能力非常优秀且很稳定。

3.6.2 模型网络

本节接着在两种模型网络，即 Erdős-Rényi (ER) 模型网络 [89] 以及 Barabási-Albert (BA) 模型网络 [90]，上比较这六个中心性度量的区分能力。对于 ER 模型网络，我们共生成了 50 个网络，每个网络包含 1000 个节点。每个 ER 网络中任意节点对连接的概率为 0.004。由于生成出的 ER 网络可能是不连通的，本节在每个 ER 网络的最大连通分支上进行实验。这些中心性度量在 ER 网络上的平均 P_c 值如表 3.5 所示。对于 BA 网络，我们共生成了 50 个网络，每个网络包含 1000 个节点。每个 BA 网络中节点的平均度为 4。这些中心性度量在 BA 网络上的平均 P_c 值如表 3.5 所示。此外，由于这些 P_c 值的方差很小，即小于 10^{-6} ，可以忽略不计，因此表 3.5 中不再列出它们的方差值。表 3.5 中模型网络的实验结果表明：① 度中心性的 P_c 值与其余中心性度量有明显差距，这表明度中心性的区分能力明显差于其余的中心性度量；② 介数中心性以及接近中心性的 P_c 值低于最高值但差距不大且显著高于度中心性的 P_c 值，这表明这两个中心性度量的区分能力较好；③ PageRank 中心性、特征向量中心性以及信息中心性的 P_c 值与最高值几乎一样，这表明这三个中心性度量有着十分优秀的区分能力。

3.7 有向加权图

至此，我们已经研究了无向无权图上中心性度量的区分能力。本节中，我们将继续研究有向加权图上中心性度量的区分能力。

3.7.1 理论推广

本节首先介绍一些关于有向加权图的基本概念。令 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$ 表示一个包含 n 个节点和 m 条有向边的有向加权图，其中 $\mathcal{V} = \{1, 2, \dots, n\}$ 是节点集， $\mathcal{E} = \{\langle u_1, v_1 \rangle, \langle u_2, v_2 \rangle, \dots, \langle u_m, v_m \rangle\}$ 表示有向边集， $\langle u, v \rangle$ 表示从节点 u 到节点

v 的一条有向边, $\varepsilon: \mathcal{E} \rightarrow \mathbb{R}_+$ 表示权重函数。我们用 ε_{uv} 表示有向边 $\langle u, v \rangle$ 的权重。图 \mathcal{G} 的邻接矩阵 \mathbf{A} 的定义如下: 若 $\langle i, j \rangle \in \mathcal{E}$, 则 $a_{ij} = \varepsilon_{ij}$, 否则 $a_{ij} = 0$ 。节点 u 的入度记为 $d_u^- = \sum_{\langle v, u \rangle \in \mathcal{E}} \varepsilon_{vu}$, 它的出度记为 $d_u^+ = \sum_{\langle u, v \rangle \in \mathcal{E}} \varepsilon_{uv}$ 。有向加权图 \mathcal{G} 的入度对角阵以及出度对角阵分别是 $\mathbf{D}_- = \text{diag}(d_1^-, d_2^-, \dots, d_n^-)$ 和 $\mathbf{D}_+ = \text{diag}(d_1^+, d_2^+, \dots, d_n^+)$ 。

有研究指出, 无权无向图中自同构的定义, 即定义 3.9, 可以推广到有向加权图, 其具体表述如下 [91-92]。

定义 3.21 (有向加权图上的自同构) 在有向加权图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$ 中, 双射 $f: \mathcal{V} \rightarrow \mathcal{V}$ 是图 \mathcal{G} 的一个自同构, 当且仅当对于任意的 $\langle u, v \rangle \in \mathcal{E}$ 均成立 $\langle f(u), f(v) \rangle \in \mathcal{E}$ 以及 $\varepsilon_{uv} = \varepsilon_{f(u)f(v)}$ 。

基于定义 3.21, 我们可以给出有向加权图上自同构等价的定义, 其具体表述如下。

定义 3.22 (有向加权图上的自同构等价) 在有向加权图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$ 中, 节点 u 和节点 v 是自同构等价的, 记为 $u \cong v$, 当且仅当存在图 \mathcal{G} 的一个自同构 f (基于定义 3.21) 使得 $f(u) = v$ 。

在有了有向加权图上自同构等价的定义, 即定义 3.22, 之后, 我们根据“等价节点必然有相同的中心性值”的思想, 可以将无权无向图中自同构等价公理, 即公理 3.17, 推广至有向加权图。

公理 3.23 (有向加权图上的自同构等价公理) 在有向加权图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$ 中, 任意中心性度量 c 都应满足以下特性: 对于任意的自同构等价 (基于定义 3.22) 的两个节点 u 和 v , 记为 $u \cong v$, 均成立 $f_c(u) = f_c(v)$, 其中 $f_c(u)$ 表示节点 u 基于中心性度量 c 的中心性值。

同样的, 我们也可以将无权无向图中区分能力指数以及广义区分度的概念推广至有向加权图, 其具体表述如下。

定义 3.24 (有向加权图上的区分能力指数) 在有向加权图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$ 中, 若中心性度量 c 满足自同构等价公理, 即公理 3.23, 则它的区分能力指数, 记为 D_c ,

定义为中心性度量 c 对于从图 \mathcal{G} 的节点集合 \mathcal{V} 中随机选出的一对不等价节点给出不同中心性值的概率，即

$$D_c = P(f_c(u) \neq f_c(v) \mid u \not\sim v),$$

其中 $u \not\sim v$ 表示节点 u 和 v 不等价（基于定义 3.22）。

定义 3.25 (有向加权图上的广义区分度) 在有向加权图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$ 中，若中心性度量 c 满足自同构等价公理，即公理 3.23，则它的广义区分度，记为 P_c ，定义为中心性度量 c 对于从图 \mathcal{G} 的节点集合 \mathcal{V} 中随机选出的一对节点给出不同中心性值的概率，即

$$P_c = P(f_c(u) \neq f_c(v)).$$

显然，对于有向加权图上的两个不同的中心性度量 c_1 和 c_2 ，有 $D_{c_1} < D_{c_2}$ 当且仅当 $P_{c_1} < P_{c_2}$ 。由于证明过程与定理 3.20 的证明过程完全一致，我们在此不再重复。

至此，我们成功的将无向无权图上研究中心性度量区分能力的全部理论推广至有向加权图。

3.7.2 应用

在这一小节中，我们将研究有向加权图上的两个常用的中心性度量，即 PageRank 中心性以及特征向量中心性。我们首先给出它们的数学定义。

定义 3.26 (有向加权图上的 PageRank 中心性) 在有向加权图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$ 中，令 $\alpha \in (0, 1)$ 为随机跳跃因子，且用 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 表示一个 n 维列向量。 \mathbf{x} 的所有分量均大于等于 0（即 $\mathbf{x} \geq \mathbf{0}$ ），且满足 $\mathbf{x}^\top \mathbf{I} = 1$ 以及

$$(\mathbf{I} - \alpha \mathbf{A}^\top \mathbf{D}_+^{-1}) \mathbf{x} = \frac{1 - \alpha}{n} \mathbf{I}.$$

则在图 \mathcal{G} 中，节点 u 的 PageRank 中心性定义为 $f_{\text{PR}}^\alpha(u) = x_u$ 。

定义 3.27 (有向加权图上的特征向量中心性) 在有向加权图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \varepsilon)$ 中，令 $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ 为一 n 维列向量，满足 $\mathbf{x}^\top \mathbf{I} = 1$ 以及

$$\mathbf{A}\mathbf{x} = \lambda_{\max}\mathbf{x},$$

表 3.6: 有向加权现实网络的统计信息以及 PR 和 EC 的 P_c 值 ($\times 10^{-2}$)

Network	n	m	n'	m'	P_{PR}	P_{EC}
emailDnc	1,892	5,518	520	3,518	99.2404	99.3738
wikiVote	7,115	103,689	1,300	39,456	99.9999	99.9988
socEpinions1	75,879	508,837	32,223	443,506	99.9817	99.9970
egoTwitter	81,306	1,768,135	68,413	1,685,152	99.9737	95.8989
emailEuAll	265,214	418,956	34,203	151,132	99.7983	99.8660

其中 λ_{\max} 是邻接矩阵 \mathbf{A} 的最大特征值。则图 \mathcal{G} 中节点 u 的特征向量中心性定义为 $f_{EC}(u) = x_u$ 。

显然，有向加权图上的 PageRank 中心性以及特征向量中心性满足自同构等价公理，即公理 3.23。由于证明过程与定理 3.14 和定理 3.15 的证明过程基本一致，因此不再重复。

接下来，我们通过计算这两个中心性度量在有向加权的现实网络上的广义区分度，即 P_c 值，来测试它们的区分能力。我们共挑选了 5 个现实网络所示，其中网络 emailDnc 来自 Network Repository¹，其余的现实网络来自 SNAP²。这些网络的相关统计信息如表 3.6 所示，其中 n 和 m 分别表示这些网络包含的节点和有向边的数量， n' 和 m' 分别表示这些网络的最大强连通分支包含的节点和有向边的数量。我们在这些现实网络的最大强连通分支上进行实验，实验结果如表 3.6 所示。从实验结果来看，PageRank 中心性与特征向量中心性的广义区分度值基本相当，这说明这两个中心性度量的区分能力基本相同。然而 PageRank 中心性的稳定性要由于特征向量中心性，因为在网络 egoTwitter 上，PageRank 中心性的 P_c 值要明显高于特征向量中心性且在其余的四个网络中这两个中心性的表现并没有明显差距。

3.8 本章小结

在这一章中，我们对无权无向简单图上的中心性度量的区分能力做了系统性的研究。我们首先确立了节点之间等价关系的判断标准。在综合考量了三种常用的节点等价关系定义，即结构等价、正则等价以及自同构等价，之后，我们根据自同构等价具有唯一性且符合常识的特点，决定选用自同构等价作为判断两个

¹<http://networkrepository.com>

²<https://snap.stanford.edu>

节点是否等价的依据。随后，我们讨论了六种常用的中心性度量，即度中心性、介数中心性、接近中心性、PageRank 中心性、特征向量中心性以及信息中心性，在节点等价时的情况，发现这些常用的中心性度量都能够在节点等价时给出相等的中心性值。因此，我们给出了自同构等价公理，即公理 3.17，并规定任何具备区分能力的中心性度量都应满足这个公理。我们接着讨论了中心性度量在节点不等价时的情况，发现并非所有的中心性度量都能在节点不等价时给出不同的中心性值。根据这个现象，我们提出了区分能力指数的概念，即定义 3.18，为量化研究中心性度量的区分能力提供了理论依据。由于直接计算区分能力指数非常困难，我们提出了用广义区分度，即 P_c 值，来间接比较不同中心性度量的区分能力。随后，我们在 30 个现实网络以及两种模型网络上测试了六种中心性度量的区分能力，并大致判断出它们区分能力的强弱，即信息中心性、特征向量中心性、PageRank 中心性和接近中心性的区分能力较强；度中心性和介数中心性的区分能力较弱。最后，我们把无权无向图上的研究结论推广至有向加权图并对有向加权图上的两个常用的中心性度量，即 PageRank 中心性和特征向量中心性，的区分能力做了研究。

第四章 森林中心性

4.1 引言

本章将提出一种具备较强区分能力的节点中心性度量，即森林中心性（Forest Centrality, FC）。4.2节将介绍图的置根生成森林的相关概念。森林中心性的定义将在4.3节中给出。4.4节将研究森林中心性的几个重要的特性。此外，4.5节将提出一种近似计算森林中心性的快速算法。本章的实验部分以及总结部分将分别在4.6和4.7节中呈现。

4.2 置根生成森林

置根生成森林在图数据挖掘中有着广泛的应用 [93-94, 66, 95]。对于一个无权无向简单图 $G = (V, E)$ 而言，它的置根生成森林 $F = (V_F, E_F, r_F)$ 是图 G 的一个无环子图，其中 V_F 表示森林 F 的节点集合且有 $V_F = V$ ； E_F 表示森林 F 的边集合且有 $E_F \subseteq E$ ； $r_F : V_F \rightarrow \{0, 1\}$ 为置根函数，若 V_F 中的一个节点 u 是根节点则 $r_F(u) = 1$ 否则 $r_F(u) = 0$ 。由于置根生成森林 F 是无环的，因此 F 中的每个连通分支都是一颗树，且每个连通分支有且仅有一个节点是根节点。本章用 $\mathcal{F} = \mathcal{F}(G)$ 表示图 G 的所有置根生成森林构成的集合，用 $\mathcal{F}_{ij} = \mathcal{F}_{ij}(G)$ 表示所有节点 i 和 j 在同一个连通分支且这个连通分支的根节点是 i 的置根生成森林构成的集合。此外， $\mathcal{F}_{ii} = \mathcal{F}_{ii}(G)$ 可用于表示所有节点 i 是根节点的置根生成森林构成的集合。

为便于读者理解置根生成森林的概念，本章在此给出一个示例。如图4.1所示，示例网络 $G_0 = (\{1, 2, 3, 4\}, \{(1, 2), (1, 3), (1, 4), (3, 4)\})$ 共有 $|\mathcal{F}| = |\mathcal{F}(G_0)| = 40$ 个置根生成森林，其中节点1是根节点的置根生成森林共有 $|\mathcal{F}_{11}| = |\mathcal{F}_{11}(G_0)| = 16$ 个，这些置根生成森林在图4.1中用黄色底色标出。

本节接着介绍森林矩阵的概念。图 $G = (V, E)$ 的森林矩阵 [96-97]，记为 \mathbf{W} ，定义为 $\mathbf{W} = (\mathbf{I} + \mathbf{L})^{-1} = (w_{ij})_{n \times n}$ ，其中 \mathbf{I} 为 n 阶单位矩阵； \mathbf{L} 为图 G 的拉普拉斯矩阵。有研究指出， $|\mathcal{F}| = |\mathbf{I} + \mathbf{L}|$ 且 $w_{ij} = |\mathcal{F}_{ij}|/|\mathcal{F}|$ [98-99]。为便于读者理解，本节给出了示例网络 G_0 的森林矩阵，即

$$\mathbf{W} = \frac{1}{40} \begin{pmatrix} 16 & 8 & 8 & 8 \\ 8 & 24 & 4 & 4 \\ 8 & 4 & 19 & 9 \\ 8 & 4 & 9 & 19 \end{pmatrix},$$

其中 $w_{11} = 16/40 = |\mathcal{F}_{11}|/|\mathcal{F}|$ 。显然，森林矩阵 \mathbf{W} 中所有元素的值都在 0 到 1 之间，即 $w_{ij} \in [0, 1]$ 。

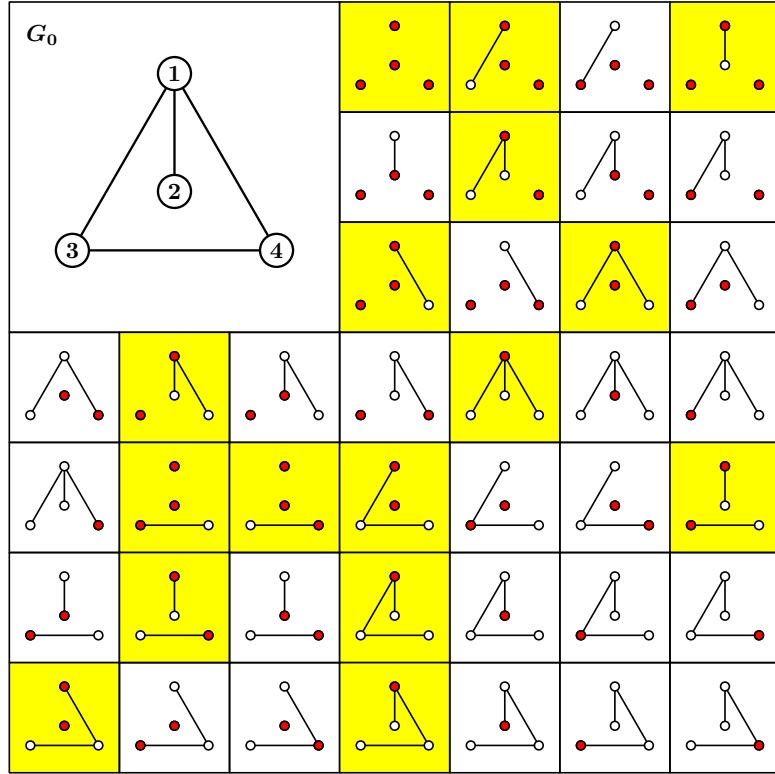


图 4.1: 示例网络 G_0 以及它的全部 40 个置根生成森林.

4.3 森林中心性

基于图的置根生成森林的概念，本节将提出一个新的节点中心性度量，即森林中心性（Forest Centrality, FC）。

4.3.1 定义

在无权无向简单图 $G = (V, E)$ 中，令 $F = (V_F, E_F, r_F)$ 为图 G 的一个置根生成森林，其中 u 节点是根节点，即 $r_F(u) = 1$ 。显然，置根生成森林 F 可以分为这样两个部分：一个是与节点 u 连通的部分；另一个是与节点 u 不连通的部分。与节点 u 不连通那个部分和节点 u 的重要性之间没有明显的关联，因此本节对于这部分不予考虑。本节中重点讨论的是置根生成森林 F 中与节点 u 连通的部分，即森林 F 中包含节点 u 的连通分支。直觉上，节点 u 所在的那个连通分支越庞大越能够显明出节点 u 处于一个重要的位置。因此，根据这个认识，本节提出了森林中心性的概念，其具体表述如下。

定义 4.1 (森林中心性) 在无权无向简单图 $G = (V, E)$ 中, 节点 u 的森林中心性定义为图 G 的所有 u 点是根的置根生成森林中节点 u 所在的连通分支的平均大小, 即

$$f_{\text{FC}} = \frac{1}{|\mathcal{F}_{uu}|} \sum_{F \in \mathcal{F}_{uu}} |F|_u, \quad (4.1)$$

其中 $|F|_u$ 表示置根生成森林 F 中节点 u 所在的连通分支包含的节点数。

为便于读者理解森林中心性的定义, 即定义 4.1, 本节以 4.2 节中的示例网络 G_0 为例作一个说明。网络 G_0 的四个置根生成森林集合 \mathcal{F}_{11} 、 \mathcal{F}_{22} 、 \mathcal{F}_{33} 以及 \mathcal{F}_{44} 中的所有元素如图 4.2 所示, 其中 $|F|_u$ 的数值标在每个置根生成森林图片的右上角。因此, 根据森林中心性的定义, 即定义 4.1, 可以计算出示例网络 G_0 中每个节点的森林中心性值, 即

$$\begin{aligned} f_{\text{FC}}(1) &= \frac{1 \times 3 + 2 \times 5 + 3 \times 5 + 4 \times 3}{3 + 5 + 5 + 3} = 2.5 \\ f_{\text{FC}}(2) &= \frac{1 \times 16 + 2 \times 3 + 3 \times 2 + 4 \times 3}{16 + 3 + 2 + 3} \approx 1.67 \\ f_{\text{FC}}(3) &= \frac{1 \times 8 + 2 \times 4 + 3 \times 4 + 4 \times 3}{8 + 4 + 4 + 3} \approx 2.11 \\ f_{\text{FC}}(4) &= \frac{1 \times 8 + 2 \times 4 + 3 \times 4 + 4 \times 3}{8 + 4 + 4 + 3} \approx 2.11. \end{aligned}$$

在示例网络 G_0 中, 节点 1 与网络中其余三个节点都是直接相连的, 占据着最重要的位置。它的森林中心性值是 $f_{\text{FC}}(1) = 2.5$, 在 G_0 的四个节点中是最大的; 节点 2 仅仅与节点 1 直接相连, 和网络中其它两个节点的距离比较远, 处于边缘的位置。它的森林中心性值是 $f_{\text{FC}}(2) \approx 1.67$, 在 G_0 的四个节点中是最小的; 节点 3 和节点 4 相互对称, 与除节点 2 之外的节点都是直接相连的。这两个节点的重要程度介于节点 1 和节点 2 之间。它们的森林中心性值是 $f_{\text{FC}}(3) = f_{\text{FC}}(4) \approx 2.11$, 大于 $f_{\text{FC}}(2)$ 小于 $f_{\text{FC}}(1)$ 。从示例网络 G_0 上的情况来看, 森林中心性的定义, 即定义 4.1, 有一定的合理性。

4.3.2 计算

由于一个图的置根生成森林的数量十分庞大, 本节无法通过直接枚举所有置根生成森林的方式来计算森林中心性的值。因此, 为高效率的计算森林中心性的值, 本节接下来将证明森林中心性的值可以用森林矩阵 \mathbf{W} 的对应对角线元素值来表示。

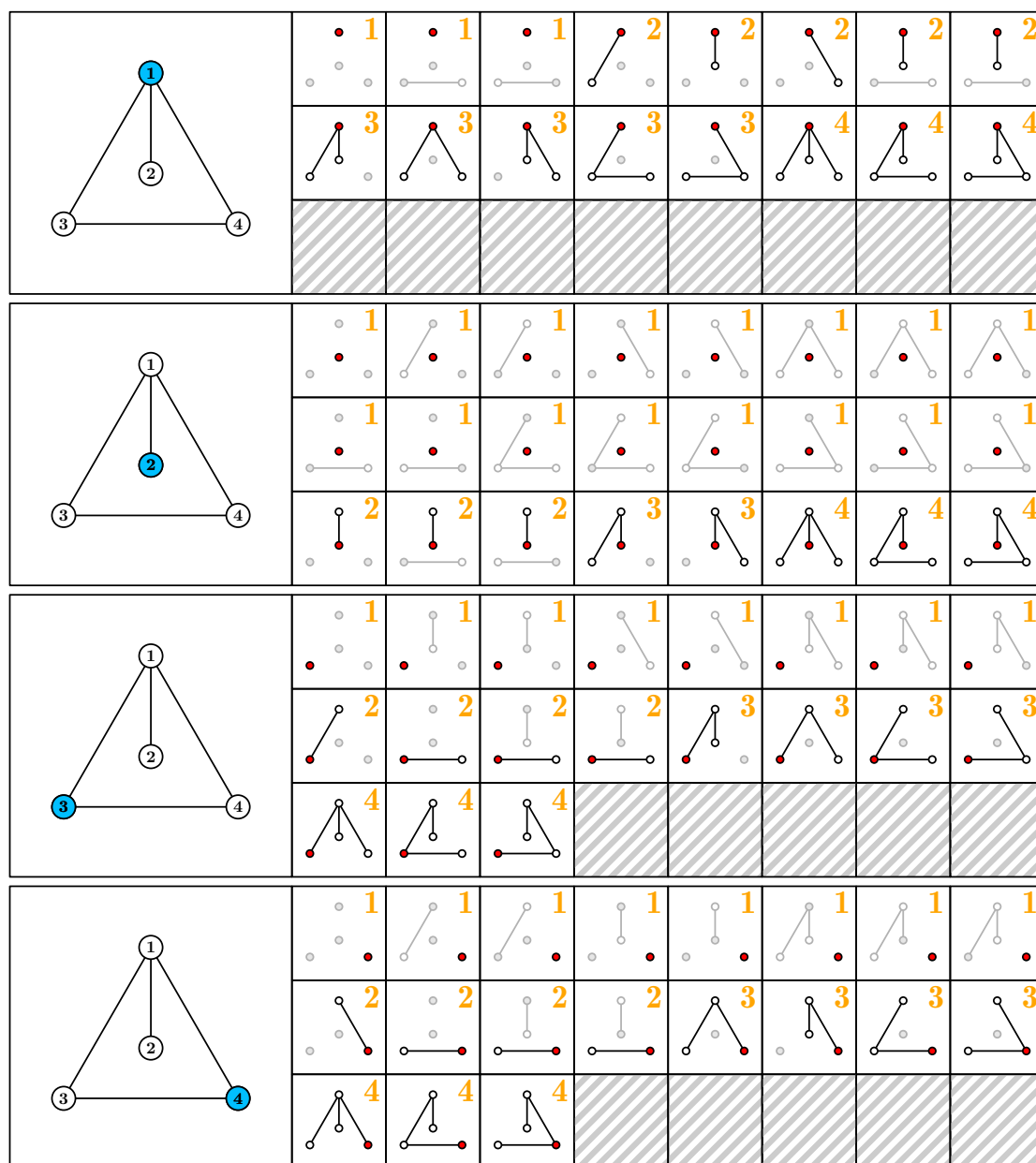


图 4.2: 示例网络 G_0 的置根生成森林集合 \mathcal{F}_{11} 、 \mathcal{F}_{22} 、 \mathcal{F}_{33} 以及 \mathcal{F}_{44} .

已知 $|\mathcal{F}| = |\mathbf{I} + \mathbf{L}|$, w_{uu} 是矩阵 $\mathbf{W} = (\mathbf{I} + \mathbf{L})^{-1}$ 中的元素并且矩阵的行列式和矩阵的逆都可以在 $O(n^3)$ 时间内求解出来, 则 $|\mathcal{F}|$ 和 w_{uu} 的值都能在 $O(n^3)$ 时间内求出。所以, 根据 $w_{uu} = |\mathcal{F}_{uu}|/|\mathcal{F}|$ 可知, $|\mathcal{F}_{uu}| = w_{uu}|\mathcal{F}|$ 的值也可以在 $O(n^3)$ 时间内算出。因此, 只需在多项式时间内求出 (4.1) 式的后半部分, 即 $\sum_{F \in \mathcal{F}_{uu}} |F|_u$, 的值, 就能够在多项式时间内计算出森林中心性的值。

本节先把计算 $\sum_{F \in \mathcal{F}_{uu}} |F|_u$ 的问题转换为求解一个集合大小的问题。

引理 4.2 在无权无向图 $G = (V, E)$ 中, 令集合 \mathcal{S}_u 定义为

$$\mathcal{S}_u = \{(F, v^*) \mid F \in \mathcal{F}_{uu}, v^* \in T(F, u)\},$$

其中 $T(F, u)$ 表示置根生成森林 F 中节点 u 所在的那个连通分支。则

$$\sum_{F \in \mathcal{F}_{uu}} |F|_u = |\mathcal{S}_u|$$

成立。

证明: 已知 $|F|_u$ 表示置根生成森林 F 中节点 u 所在连通分支包含节点的个数, 因此,

$$\begin{aligned} |F|_u &= |\{v^* \mid v^* \in T(F, u)\}| \\ &= |\{(F, v^*) \mid v^* \in T(F, u)\}|. \end{aligned}$$

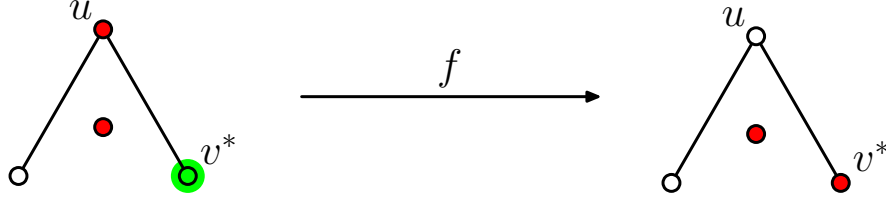
所以,

$$\begin{aligned} \sum_{F \in \mathcal{F}_{uu}} |F|_u &= \sum_{F \in \mathcal{F}_{uu}} |\{(F, v^*) \mid v^* \in T(F, u)\}| \\ &= |\{(F, v^*) \mid F \in \mathcal{F}_{uu}, v^* \in T(F, u)\}| = |\mathcal{S}_u|, \end{aligned}$$

得证。 □

本节接下来将证明集合 \mathcal{S}_u 的与集合 \mathcal{F} 大小相等。

引理 4.3 在无权无向图 $G = (V, E)$ 中, 等式 $|\mathcal{S}_u| = |\mathcal{F}|$ 对于任意图 G 中的节点 u 成立。

图 4.3: 映射 $f : \mathcal{S}_u \rightarrow \mathcal{F}$ 的构造方式

证明: 本节通过构造映射 $f : \mathcal{S}_u \rightarrow \mathcal{F}$ 的方式来证明集合 \mathcal{S}_u 的与集合 \mathcal{F} 大小相等。令 (F_0, v_0^*) 为集合 \mathcal{S}_u 中任意一个元素，则映射 f 可以理解为对置根生成森林 F_0 做一个操作，即令节点 v^* 为 F_0 中节点 u 所在连通分支的根节点。这样就能得到新的置根生成森林 $f(F_0, v_0^*) \in \mathcal{F}$ 。图 4.3 揭示了映射 f 的构造过程。

首先证明映射 $f : \mathcal{S}_u \rightarrow \mathcal{F}$ 是一个单射。令 $(F_0, v_0^*) \neq (F_1, v_1^*)$ 为集合 \mathcal{S}_u 中任意两个不相等的元素。接下来，本节将讨论以下三种情况。

- ① 若 $F_0 = F_1$ ，则由 $(F_0, v_0^*) \neq (F_1, v_1^*)$ 可得 $v_0^* \neq v_1^*$ 。根据映射 f 的构造方式可知：节点 v_0^* 和 v_1^* 分别是置根生成森林 $f(F_0, v_0^*)$ 和 $f(F_1, v_1^*)$ 中的根节点。因为 (F_0, v_0^*) 和 (F_1, v_1^*) 都是集合 \mathcal{S}_u 中的元素，所以根据集合 \mathcal{S}_u 的定义可知，节点 v_0^* 和 v_1^* 同属于森林 $F_0 = F_1$ 中节点 u 所在的连通分支，即 $v_0^*, v_1^* \in T(F_0, u)$ 。由于映射 f 所对应的操作并不改变置根生成森林的边集，因此有 $v_0^*, v_1^* \in T(f(F_0, v_0^*), u)$ 以及 $v_0^*, v_1^* \in T(f(F_1, v_1^*), u)$ 。又因为置根生成森林中每个连通分支中有且仅有一个根节点，所以在置根生成森林 $f(F_0, v_0^*)$ 中节点 v_1^* 不是根节点。由于节点 v_1^* 在置根生成森林 $f(F_1, v_1^*)$ 中是根节点，因此 $f(F_0, v_0^*) \neq f(F_1, v_1^*)$ 。
- ② 若 $F_0 \neq F_1$ 且 F_0 和 F_1 的边集合不相等，则由映射 f 所对应的操作并不改变置根生成森林的边集可知，置根生成森林 $f(F_0, v_0^*)$ 和 $f(F_1, v_1^*)$ 的边集合也不相等。因此 $f(F_0, v_0^*) \neq f(F_1, v_1^*)$ 。
- ③ 若 $F_0 \neq F_1$ 且 F_0 和 F_1 的边集合相等，则必然存在一个节点 v 使得 $r_{F_0}(v) \neq r_{F_1}(v)$ ，即节点 v 在其中一个置根生成森林是根节点同时在另一个森林中不是根节点。不妨设 $r_{F_0}(v) = 1$ ，则 $r_{F_1}(v) = 0$ 。假设 v 属于森林 F_0 中节点 u 所在的连通分支，即 $v \in T(F_0, u)$ 。由 $(F_0, v_0^*) \in \mathcal{S}_u$ 可知 $F_0 \in \mathcal{F}_{uu}$ ，所以 $r_{F_0}(u) = 1$ 。又因为 $r_{F_0}(v) = 1$ 且 $v \in T(F_0, u)$ ，所以有 $u = v$ 。再根据 $(F_1, v_1^*) \in \mathcal{S}_u$ 可知 $r_{F_1}(u) = 1$ 。因此有 $r_{F_1}(v) = 1$ ，与 $r_{F_1}(v) = 0$ 矛盾。所

以, 节点 v 必定不属于森林 F_0 中节点 u 所在的连通分支, 即 $v \notin T(F_0, u)$ 。又因为 F_0 和 F_1 的边集合相等, 所以 $v \notin T(F_1, u)$ 。由于映射 f 所对应的操作仅仅改变节点 u 所在的连通分支, 因此节点 v 在森林 $f(F_0, v_0^*)$ 中是根节点, 在森林 $f(F_0, v_0^*)$ 中不是根节点。所以, $f(F_0, v_0^*) \neq f(F_1, v_1^*)$ 。

由 ① ② ③ 可知, 对于任意 $(F_0, v_0^*), (F_1, v_1^*) \in \mathcal{S}_u$, 若 $(F_0, v_0^*) \neq (F_1, v_1^*)$, 则 $f(F_0, v_0^*) \neq f(F_1, v_1^*)$ 必定成立。所以, 映射 $f: \mathcal{S}_u \rightarrow \mathcal{F}$ 是一个单射。

接下来证明映射 $f: \mathcal{S}_u \rightarrow \mathcal{F}$ 是一个满射。令 F' 集合 \mathcal{F} 中任意一个元素, v_0^* 为置根生成森林 F' 中 u 点所在连通分支, 即 $T(F', u)$, 的根节点。只要把 $T(F', u)$ 的根节点重新设置成节点 u , 就能得到一个新的置根生成森林 F_0 。根据映射 f 的构造方式可知, $f(F_0, v_0^*) = F'$ 。又因为 $r_{F_0}(u) = 1$ 且 $v_0^* \in T(F_0, u)$, 所以有 $(F_0, v_0^*) \in \mathcal{S}_u$ 。因此, 对于任意的 $F' \in \mathcal{F}$, 都存在 $(F_0, v_0^*) \in \mathcal{S}_u$ 使得 $f(F_0, v_0^*) = F'$ 。所以, 映射 $f: \mathcal{S}_u \rightarrow \mathcal{F}$ 是一个满射。

综上所述, 映射 f 既是单射又是满射。所以, 映射 $f: \mathcal{S}_u \rightarrow \mathcal{F}$ 是一个双射。因此, $|\mathcal{S}_u| = |\mathcal{F}|$, 得证。 \square

基于引理 4.3 给出的结论, 森林中心性的值就可以用森林矩阵 \mathbf{W} 的对应对角线元素值来表示。

定理 4.4 在无权无向图 $G = (V, E)$ 中, 节点 u 的森林中心性值等于图 G 的森林矩阵的第 u 个对角线元素值的倒数, 即

$$f_{\text{FC}}(u) = \frac{1}{w_{uu}},$$

其中 $f_{\text{FC}}(u)$ 表示节点 u 的森林中心性值, w_{uu} 表示图 G 的森林矩阵 \mathbf{W} 中第 u 行第 u 列的元素。

证明: 由引理 4.2 以及引理 4.3 的结论可知,

$$f_{\text{FC}}(u) = \frac{1}{|\mathcal{F}_{uu}|} \sum_{F \in \mathcal{F}_{uu}} |F|_u = \frac{|\mathcal{S}_u|}{|\mathcal{F}_{uu}|} = \frac{|\mathcal{F}|}{|\mathcal{F}_{uu}|}.$$

又因为 $w_{uu} = |\mathcal{F}_{uu}|/|\mathcal{F}|$, 所以

$$f_{\text{FC}}(u) = \frac{|\mathcal{F}|}{|\mathcal{F}_{uu}|} = \frac{1}{w_{uu}},$$

由此得证。 \square

定理 4.4 的结论指出, 只要求出森林矩阵的所有对角线元素, 就可以计算出图中所有节点的森林中心性值。因此, 在一个包含 n 个节点的无权无向图 G 中, 仅需花费 $O(n^3)$ 的时间就可以求出所有节点的森林中心性值。

4.4 相关性质

本节将从三个方面研究森林中心性的性质, 即森林中心性值的上下界, 森林中心性是否满足自同构等价公理以及森林中心性在路径图上的表现。

4.4.1 上下界

本节首先研究森林中心性值的上下界问题。

有学者曾经研究过森林矩阵对角线元素的上下界问题 [99], 结论如下。

引理 4.5 在无权无向图 $G = (V, E)$ 中, 令 \mathbf{W} 为图 G 的森林矩阵, w_{uu} 为矩阵 \mathbf{W} 第 u 行第 u 列的元素, 则对于任意的 $u \in V$, 都有

$$w_{uu} \in \left[\frac{1}{d_u + 1}, \frac{2}{d_u + 2} \right],$$

其中 d_u 表示节点 u 的度。

由于森林中心性值可以用森林矩阵的对角线元素来表示。因此, 本节可以通过森林矩阵对角线元素的上下界推算出森林中心性值的上下界。

定理 4.6 在无权无向图 $G = (V, E)$ 中, 对于任意的节点 $u \in V$, 均成立

$$f_{\text{FC}}(u) \in \left[1 + \frac{d_u}{2}, d_u + 1 \right],$$

其中 $f_{\text{FC}}(u)$ 表示节点 u 的森林中心性值, d_u 表示节点 u 的度。

证明: 由引理 4.5 可得

$$\frac{1}{d_u + 1} \leq w_{uu} \leq \frac{2}{d_u + 2}.$$

再根据定理 4.4 可知

$$f_{\text{FC}}(u) = \frac{1}{w_{uu}}.$$

因此，有

$$1 + \frac{d_u}{2} \leq w_{uu} \leq d_u + 1,$$

得证。 \square

本节通过定理 4.6 给出了森林中心性的上下界，即一个节点 u 的森林中心性的值必定大于等于 $1 + d_u/2$ 小于等于 $d_u + 1$ 。

4.4.2 自同构等价公理

本节接着证明森林中心性满足自同构等价公理，即公理 3.17。

定理 4.7 在任意无权无向简单图 $G = (V, E)$ 中，若节点 u 和节点 v 自同构等价，记为 $u \cong v$ ，则它们的森林中心性值 $f_{\text{FC}}(u)$ 和 $f_{\text{FC}}(v)$ 相等，即 $f_{\text{FC}}(u) = f_{\text{FC}}(v)$ 。

证明： 因为节点 u 和节点 v 是自同构等价的，所以必然存在图 G 的一个自同构 f 使得 $f(u) = v$ 。令 \mathbf{P} 为自同构 f 所对应的排列矩阵。根据定理 3.16 的证明过程可知， $\mathbf{P}\mathbf{L}\mathbf{P}^\top = \mathbf{P}^\top\mathbf{L}\mathbf{P} = \mathbf{L}$ 。所以，

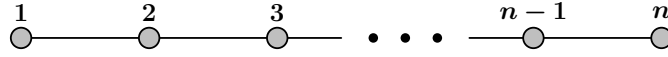
$$\begin{aligned} w_{vv} &= \mathbf{e}_v^\top (\mathbf{I} + \mathbf{L})^{-1} \mathbf{e}_v \\ &= \mathbf{e}_u^\top \mathbf{P}(\mathbf{I} + \mathbf{L})^{-1} \mathbf{P}^\top \mathbf{e}_u \\ &= \mathbf{e}_u^\top (\mathbf{P}\mathbf{P}^\top + \mathbf{P}\mathbf{L}\mathbf{P}^\top)^{-1} \mathbf{e}_u \\ &= \mathbf{e}_u^\top (\mathbf{I} + \mathbf{L})^{-1} \mathbf{e}_u = w_{uu}. \end{aligned}$$

再根据森林中心性的定义，即定义 4.1，以及定理 4.4 的结论可知，

$$f_{\text{FC}}(u) = \frac{1}{w_{uu}} = \frac{1}{w_{vv}} = f_{\text{FC}}(v),$$

得证。 \square

定理 4.7 指出，森林中心性必定会给自同构等价的节点赋予相同的中心性值，即森林中心性满足自同构等价公理。这个结论为本章接下来量化研究森林中心性的区分能力奠定了重要的理论基础。

图 4.4: 含 n 个节点的路径图 G_n^p

4.4.3 路径图

本节最后分析森林中心性在路径图上的表现。

本节用 $G_n^p = (V, E)$ 表示包含 n 个节点的路径图。如图 4.4 所示，路径图 G_n^p 中节点的标号从左到右依次是 1 至 n 。

引理 4.8 令 $G_n^p = (V, E)$ 为一个包含 n 个节点的路径图，则有

$$N_n = \left(\frac{5 - 3\sqrt{5}}{10} \right) \left(\frac{3 - \sqrt{5}}{2} \right)^{n-1} + \left(\frac{5 + 3\sqrt{5}}{10} \right) \left(\frac{3 + \sqrt{5}}{2} \right)^{n-1},$$

其中 N_n 表示路径图 G_n^p 的置根生成森林的个数。

证明: 令 N_n ($n = 0, 1, 2, \dots$) 分别为路径图 G_n^p ($n = 0, 1, 2, \dots$) 的置根生成森林的个数，且令 $N_0 = 1$ 。

若已知路径图 $G_0^p, G_1^p, \dots, G_n^p$ 的置根生成森林的个数是 N_0, N_1, \dots, N_n ，则如果在路径图 G_{n+1}^p 中，标号为 $n+1$ 的节点所在的连通分支包含了 x 个点，那么对于前 $n+1-x$ 个节点有 N_{n+1-x} 个置根生成森林，对于最后的 x 个点而言，它们只能构成一条链，但这条链中每个节点都可以是根节点，所以最后 x 个节点有 x 个置根生成森林。由于两边的情况是可以任意组合的，所以这样的置根生成森林的个数等于 xN_{n+1-x} 。

由此可以得出一个递推关系，即

$$N_{n+1} = N_n + 2N_{n-1} + \dots + nN_1 + (n+1)N_0. \quad (4.2)$$

根据 (4.2) 式给出的递推关系可知，

$$N_n = N_{n-1} + 2N_{n-2} + \dots + nN_0. \quad (4.3)$$

用 (4.2) 式减去 (4.3) 式得

$$N_{n+1} - N_n = \sum_{i=0}^n N_i$$

令 $S_n = \sum_{i=0}^n N_i$, 则有

$$\begin{aligned} (S_{n+1} - S_n) - (S_n - S_{n-1}) &= S_n \\ \Rightarrow S_{n+1} &= 3S_n - S_{n-1} \end{aligned}$$

列出关于 S_n 的递推关系的特征方程, 即

$$x^2 - 3x + 1 = 0$$

解得

$$x_1 = \frac{3 - \sqrt{5}}{2}, x_2 = \frac{3 + \sqrt{5}}{2}$$

则该递推关系的通解为

$$S_n = \left(\frac{3 - \sqrt{5}}{2} \right)^n c_1 + \left(\frac{3 + \sqrt{5}}{2} \right)^n c_2$$

由于 $N_0 = N_1 = 1$, 所以其初始条件是 $S_0 = 1$ 和 $S_1 = 2$, 因此

$$\begin{cases} c_1 + c_2 = 1 \\ \frac{3 - \sqrt{5}}{2} \cdot c_1 + \frac{3 + \sqrt{5}}{2} \cdot c_2 = 2 \end{cases}$$

解得

$$c_1 = \frac{5 - \sqrt{5}}{10}, c_2 = \frac{5 + \sqrt{5}}{10}$$

所以, S_n 的通项公式是

$$S_n = \left(\frac{5 - \sqrt{5}}{10} \right) \left(\frac{3 - \sqrt{5}}{2} \right)^n + \left(\frac{5 + \sqrt{5}}{10} \right) \left(\frac{3 + \sqrt{5}}{2} \right)^n$$

再用 S_n 减去 S_{n-1} 得

$$N_n = \left(\frac{5 - 3\sqrt{5}}{10} \right) \left(\frac{3 - \sqrt{5}}{2} \right)^{n-1} + \left(\frac{5 + 3\sqrt{5}}{10} \right) \left(\frac{3 + \sqrt{5}}{2} \right)^{n-1},$$

由此得证。 □

令 $f_{ii} = |\mathcal{F}_{ii}|$ 表示路径图 G_n^p 的节点 i 是根节点的置根生成森林的个数, 则根据引理 4.8 的结论, 本节可以接着推导出 f_{ii} 的表达式。

引理 4.9 令 $G_n^p = (V, E)$ 为一个包含 n 个节点的路径图, 则有

$$f_{ii} = \frac{1}{5} \left(\left(\frac{3+\sqrt{5}}{2} \right)^n + \left(\frac{3-\sqrt{5}}{2} \right)^n \right) + \frac{1}{5} \left(\left(\frac{3+\sqrt{5}}{2} \right)^{n-2i+1} + \left(\frac{2}{3+\sqrt{5}} \right)^{n-2i+1} \right),$$

其中 f_{ii} 表示路径图 G_n^p 的节点 i 是根节点的置根生成森林的个数。

证明: 设置根生成森林中节点 i 是根节点, 且节点 i 左侧有 x 个节点在以节点 i 为根的连通分支中, 节点 i 右侧有 y 个节点在以节点 i 为根的连通分支中, 则这样的置根生成森林的个数是 $N_{i-x}N_{n-i-y}$ 。因此

$$\begin{aligned} f_{ii} &= \sum_{x=0}^{i-1} \sum_{y=0}^{n-i} N_{i-x} N_{n-i-y} \\ &= (N_0 + N_1 + \cdots + N_{i-1}) (N_0 + N_1 + \cdots + N_{n-i}) \\ &= S_{i-1} S_{n-i}. \end{aligned}$$

再将 S_n 的通项公式代入得

$$f_{ii} = \frac{1}{5} \left(\left(\frac{3+\sqrt{5}}{2} \right)^n + \left(\frac{3-\sqrt{5}}{2} \right)^n \right) + \frac{1}{5} \left(\left(\frac{3+\sqrt{5}}{2} \right)^{n-2i+1} + \left(\frac{2}{3+\sqrt{5}} \right)^{n-2i+1} \right),$$

由此得证。 \square

根据 f_{ii} 的表达式, 本节可以比较路径图中节点森林中心性值的大小。

定理 4.10 在含 n ($n \geq 3$) 个节点的路径图 G_n^p 中, 若 n 为奇数, 则有

$$f_{FC}\left(\frac{n+1}{2}\right) > f_{FC}\left(\frac{n+3}{2}\right) = f_{FC}\left(\frac{n-1}{2}\right) > \cdots > f_{FC}(n) = f_{FC}(1),$$

若 n 为偶数, 则有

$$f_{FC}\left(\frac{n}{2}\right) = f_{FC}\left(\frac{n+2}{2}\right) > \cdots > f_{FC}(n) = f_{FC}(1),$$

其中 $f_{FC}(u)$ 表示节点 u 的森林中心性值。

证明: 为便于证明, 首先令

$$c_0 = \frac{1}{5} \left(\left(\frac{3+\sqrt{5}}{2} \right)^n + \left(\frac{3-\sqrt{5}}{2} \right)^n \right),$$

$$c_1 = \frac{3 + \sqrt{5}}{2},$$

则根据引理 4.9 的结论, 有

$$f_{ii} = c_0 + \frac{1}{5} (c_1^{n-2i+1} + c_1^{2i-n-1}).$$

由于

$$\begin{aligned} f_{n+1-i, n+1-i} &= c_0 + \frac{1}{5} (c_1^{n-2(n+1-i)+1} + c_1^{2(n+1-i)-n-1}) \\ &= c_0 + \frac{1}{5} (c_1^{2i-n-1} + c_1^{n-2i+1}) \\ &= f_{ii}, \end{aligned}$$

因此, 根据定理 4.4 的结论可以推导出

$$f_{\text{FC}}(i) = f_{\text{FC}}(n+1-i).$$

对 f_{ii} 的表达式求偏导, 得

$$\frac{\partial f_{ii}}{\partial i} = -\frac{2}{5} (c_1^{n-2i+1} \ln c_1 - c_1^{2i-n-1} \ln c_1). \quad (4.4)$$

令 (4.4) 式的值等于 0, 得 $i = n/2$ 。又因为 $i < n/2$ 时 (4.4) 式的值小于 0 且 $i > n/2$ 时 (4.4) 式的值大于 0, 因此 f_{ii} 在 $i \in [1, n/2)$ 时单调下降, 在 $i \in (n/2, n]$ 时单调上升, 在 $i = n/2$ 时取得最小值。

由定理 4.4 的结论可知

$$f_{\text{FC}}(i) = \frac{|\mathcal{F}|}{|\mathcal{F}_{uu}|} = \frac{N_n}{f_{ii}}.$$

因此, 节点 i 的森林中心性值仅由 f_{ii} 决定, 且 f_{ii} 的值越小, 节点 i 的森林中心性值就越大。所以, 当 n 为奇数时, 有

$$f_{\text{FC}}(\frac{n+1}{2}) > f_{\text{FC}}(\frac{n+3}{2}) = f_{\text{FC}}(\frac{n-1}{2}) > \cdots > f_{\text{FC}}(n) = f_{\text{FC}}(1),$$

当 n 为偶数时, 有

$$f_{\text{FC}}(\frac{n}{2}) = f_{\text{FC}}(\frac{n+2}{2}) > \cdots > f_{\text{FC}}(n) = f_{\text{FC}}(1),$$

得证。 □

定理 4.10 的结论指出，路径图中间节点的森林中心性值较大而两边节点的森林中心性值较小，且路径图对称节点的森林中心性值是相等的。因此，森林中心性赋予路径图中节点的中心性值是合理的，这也部分的说明了森林中心性的有效性。

4.5 快速近似算法

本节将提出一种估计节点森林中心性值的快速近似算法。本节先给出一些误差分析，随后在这些分析结果的基础上设计出一个能够快速计算节点森林中心性近似值的算法。

4.5.1 误差分析

从定理 4.4 的结论可以看出，只要计算出森林矩阵 \mathbf{W} 的所有对角线元素值就能得到图中所有节点的森林中心性值。本节首先把森林矩阵 \mathbf{W} 所有对角线元素值的计算问题转化为求解向量模长的问题。

引理 4.11 令 \mathbf{W} 为无权无向图 $G = (V, E)$ 的森林矩阵， w_{ii} 为矩阵 \mathbf{W} 第 i 行第 i 列的元素，则

$$w_{ii} = \|\mathbf{B}\mathbf{W}\mathbf{e}_i\|^2 + \|\mathbf{W}\mathbf{e}_i\|^2,$$

其中 \mathbf{B} 为图 G 的关联矩阵。

证明： 由 $\mathbf{W} = (\mathbf{I} + \mathbf{L})^{-1}$ 得

$$\begin{aligned} w_{ii} &= \mathbf{e}_i^\top \mathbf{W} \mathbf{e}_i \\ &= \mathbf{e}_i^\top (\mathbf{I} + \mathbf{L})^{-1} \mathbf{e}_i \\ &= \mathbf{e}_i^\top (\mathbf{I} + \mathbf{L})^{-1} (\mathbf{I} + \mathbf{L}) (\mathbf{I} + \mathbf{L})^{-1} \mathbf{e}_i. \end{aligned}$$

又因为 $\mathbf{L} = \mathbf{B}^\top \mathbf{B}$ ，所以

$$\begin{aligned} w_{ii} &= \mathbf{e}_i^\top (\mathbf{I} + \mathbf{L})^{-1} (\mathbf{I} + \mathbf{L}) (\mathbf{I} + \mathbf{L})^{-1} \mathbf{e}_i \\ &= \mathbf{e}_i^\top (\mathbf{I} + \mathbf{L})^{-1} (\mathbf{I} + \mathbf{B}^\top \mathbf{B}) (\mathbf{I} + \mathbf{L})^{-1} \mathbf{e}_i \\ &= \mathbf{e}_i^\top \mathbf{W} (\mathbf{I} + \mathbf{B}^\top \mathbf{B}) \mathbf{W} \mathbf{e}_i \\ &= \mathbf{e}_i^\top \mathbf{W}^2 \mathbf{e}_i + \mathbf{e}_i^\top \mathbf{W} \mathbf{B}^\top \mathbf{B} \mathbf{W} \mathbf{e}_i. \end{aligned}$$

由于森林矩阵 \mathbf{W} 是对称阵，因此有 $\mathbf{W}^\top = \mathbf{W}$ 。所以，

$$\begin{aligned} w_{ii} &= \mathbf{e}_i^\top \mathbf{W}^2 \mathbf{e}_i + \mathbf{e}_i^\top \mathbf{W} \mathbf{B}^\top \mathbf{B} \mathbf{W} \mathbf{e}_i \\ &= \mathbf{e}_i^\top \mathbf{W}^\top \mathbf{W} \mathbf{e}_i + \mathbf{e}_i^\top \mathbf{W}^\top \mathbf{B}^\top \mathbf{B} \mathbf{W} \mathbf{e}_i \\ &= (\mathbf{W} \mathbf{e}_i)^\top (\mathbf{W} \mathbf{e}_i) + (\mathbf{B} \mathbf{W} \mathbf{e}_i)^\top (\mathbf{B} \mathbf{W} \mathbf{e}_i) \\ &= \|\mathbf{B} \mathbf{W} \mathbf{e}_i\|^2 + \|\mathbf{W} \mathbf{e}_i\|^2, \end{aligned}$$

得证。 \square

本节主要通过 JL 引理 (Johnson–Lindenstrauss Lemma) 以及快速线性方程组求解器来实现森林中心性的快速计算。JL 引理的表述如下。

引理 4.12 (JL 引理 [100-101]) 给定一正实数 $\epsilon > 0$ 以及 x 个 d 维列向量 $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_x \in \mathbb{R}^d$ ，令 $\mathbf{Q}_{k \times d}$ 为一个 k 行 d 列的随机矩阵且矩阵 \mathbf{Q} 中的元素有 $1/2$ 概率是 $1/\sqrt{k}$ ， $1/2$ 概率是 $-1/\sqrt{k}$ ，则当 $k \geq 24 \log x / \epsilon^2$ 时，对于任意的 $\mathbf{v}_i, \mathbf{v}_j (1 \leq i, j \leq x)$ ，不等式

$$(1 - \epsilon) \|\mathbf{v}_i - \mathbf{v}_j\|^2 \leq \|\mathbf{Q} \mathbf{v}_i - \mathbf{Q} \mathbf{v}_j\|^2 \leq (1 + \epsilon) \|\mathbf{v}_i - \mathbf{v}_j\|^2$$

有 $1 - 1/x$ 的概率成立。

本节接着给出快速线性方程组求解器的相关引理，这个求解器能够快速的给出线性方程组的近似解。

引理 4.13 (快速线性方程组求解器 [102]) 令 \mathbf{T} 为一个 $n \times n$ 的含 m 个非零元素的对角线元素占优的对阵正定稀疏矩阵， \mathbf{b} 为一个 n 维列向量，正实数 δ 为精确度参数，则存在一个快速线性方程组求解器 $\mathbf{y}' = \text{LapSolve}(\mathbf{T}, \mathbf{b}, \delta)$ 使得其输出向量 \mathbf{y} 满足

$$\|\mathbf{y}' - \mathbf{y}\| \leq \delta \|\mathbf{y}\|_{\mathbf{T}},$$

其中 \mathbf{y} 是方程组 $\mathbf{T} \mathbf{x} = \mathbf{b}$ 的解， $\|\mathbf{y}\|_{\mathbf{T}} = \sqrt{\mathbf{y}^\top \mathbf{T} \mathbf{y}}$ 。求解器运行一次的时间复杂度为 $O(m \log^3 n \log(1/\delta))$ 。

基于 JL 引理，即引理 4.12，本节可以用 $\|\mathbf{Q} \mathbf{B} \mathbf{W} \mathbf{e}_i\|^2$ 和 $\|\mathbf{P} \mathbf{W} \mathbf{e}_i\|^2$ 的值分别近似的估计 $\|\mathbf{B} \mathbf{W} \mathbf{e}_i\|^2$ 以及 $\|\mathbf{W} \mathbf{e}_i\|^2$ 的大小，其中 \mathbf{Q} 是一个 $k \times m$ 的随机矩阵， \mathbf{P} 是一个 $k \times n$ 的随机矩阵，矩阵 \mathbf{P} 和 \mathbf{Q} 中的元素有 $1/2$ 概率是 $1/\sqrt{k}$ ，有 $1/2$ 概率

是 $-1/\sqrt{k}$ 。引理 4.12 指出, k 越大, 计算误差越小。然而, 仅仅做到这种程度显然是不够的。

由于求出森林矩阵 \mathbf{W} 需要 $O(n^3)$ 的时间复杂度, 因此在近似算法中不能直接计算森林矩阵, 而引理 4.13 中提到的快速线性方程组求解器可以很好的解决这个问题。因为无向图的森林矩阵 \mathbf{W} 是对称阵, 所以矩阵 \mathbf{QBW} 的第 i 行 $(\mathbf{QBW})_{[i]}$ 可以看作是线性方程组 $(\mathbf{I} + \mathbf{L})\mathbf{x} = (\mathbf{QB})_{[i]}$ 的解, 又因为矩阵 $\mathbf{I} + \mathbf{L}$ 是正定阵且它的每个对角线元素值的绝对值都大于同行或同列所有非对角线元素绝对值之和。因此, 本节可以用快速线性方程组求解器来近似的计算 $(\mathbf{QBW})_{[i]}$ 。所以, 矩阵 \mathbf{QBW} 的每一行都可以用快速线性方程组求解器高效率的近似计算出来。同理, 矩阵 \mathbf{PW} 的每一行也可以用求解器计算。

本节接下来评估这种近似算法的误差。为便于证明, 本节令 $\|\mathbf{M}\|_F$ 为矩阵 \mathbf{M} 的 Frobenius 范数, 定义为 $\|\mathbf{M}\|_F = \sqrt{\text{Tr}(\mathbf{M}^\top \mathbf{M})}$ 。

引理 4.14 令 \mathbf{W} 为含 n 个节点和 m 条边的无权无向图 $G = (V, E)$ 的森林矩阵, \mathbf{B} 为图 G 的关联矩阵, \mathbf{L} 为图 G 的拉普拉斯矩阵。令 $\epsilon \in (0, 1/2)$ 为误差参数。令 \mathbf{Q} 和 \mathbf{P} 分别为 $k \times m$ 以及 $k \times n$ 的随机矩阵。矩阵 \mathbf{P} 和 \mathbf{Q} 中的元素有 $1/2$ 概率是 $1/\sqrt{k}$, 有 $1/2$ 概率是 $-1/\sqrt{k}$, 且

$$k \geq \frac{24 \log(n+1)}{(\epsilon/32)^2}.$$

令 $\hat{\mathbf{X}} = \mathbf{QB}$, $\mathbf{X} = \mathbf{BW}$, $\mathbf{X}' = \mathbf{QBW}$ 。令 $\tilde{\mathbf{X}}$ 为一个 k 行 n 列的矩阵, 其中 $\tilde{\mathbf{X}}$ 的第 i 行 $\tilde{\mathbf{X}}_{[i]} = \text{LapSolve}(\mathbf{I} + \mathbf{L}, \hat{\mathbf{X}}_{[i]}, \delta)$ 。令 $\hat{\mathbf{Y}} = \mathbf{P}$, $\mathbf{Y} = \mathbf{W}$, $\mathbf{Y}' = \mathbf{PW}$ 。令 $\tilde{\mathbf{Y}}$ 为一个 k 行 n 列的矩阵, 其中 $\tilde{\mathbf{Y}}$ 的第 i 行 $\tilde{\mathbf{Y}}_{[i]} = \text{LapSolve}(\mathbf{I} + \mathbf{L}, \hat{\mathbf{Y}}_{[i]}, \delta)$, 且

$$\delta \leq \frac{\epsilon \sqrt{1 - \epsilon/32}}{96\sqrt{n}(n+1)\sqrt{n+1}\sqrt{1 + \epsilon/32}}.$$

则对于任意的 $u \in V$, 都有

$$(1 - \frac{\epsilon}{4})w_{uu} \leq \tilde{w}_{uu} \leq (1 + \frac{\epsilon}{4})w_{uu},$$

其中 $\tilde{w}_{uu} = \|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2$ 。

证明： 由 JL 引理，即引理 4.12，可知

$$\begin{aligned}(1 - \epsilon/32) \|\mathbf{X}\mathbf{e}_u\|^2 &\leq \|\mathbf{X}'\mathbf{e}_u\|^2 \leq (1 + \epsilon/32) \|\mathbf{X}\mathbf{e}_u\|^2, \\ (1 - \epsilon/32) \|\mathbf{Y}\mathbf{e}_u\|^2 &\leq \|\mathbf{Y}'\mathbf{e}_u\|^2 \leq (1 + \epsilon/32) \|\mathbf{Y}\mathbf{e}_u\|^2.\end{aligned}$$

再由引理 4.13 的结论可得

$$\begin{aligned}\|\mathbf{X}'_{[i]} - \tilde{\mathbf{X}}_{[i]}\|_{(\mathbf{L}+\mathbf{I})} &\leq \delta \|\mathbf{X}'_{[i]}\|_{(\mathbf{L}+\mathbf{I})}, \\ \|\mathbf{Y}'_{[i]} - \tilde{\mathbf{Y}}_{[i]}\|_{(\mathbf{L}+\mathbf{I})} &\leq \delta \|\mathbf{Y}'_{[i]}\|_{(\mathbf{L}+\mathbf{I})}.\end{aligned}$$

基于上述推论，可以给出 $|\|\tilde{\mathbf{X}}\mathbf{e}_u\| - \|\mathbf{X}'\mathbf{e}_u\||$ 的上界，即

$$\begin{aligned}|\|\tilde{\mathbf{X}}\mathbf{e}_u\| - \|\mathbf{X}'\mathbf{e}_u\|| &\leq \|\tilde{\mathbf{X}}\mathbf{e}_u - \mathbf{X}'\mathbf{e}_u\| \leq \|\tilde{\mathbf{X}} - \mathbf{X}'\|_F \\ &= \sqrt{\sum_{i=1}^k \|\tilde{\mathbf{X}}_{[i]} - \mathbf{X}'_{[i]}\|^2} \leq \sqrt{\sum_{i=1}^k \|\tilde{\mathbf{X}}_{[i]} - \mathbf{X}'_{[i]}\|_{(\mathbf{L}+\mathbf{I})}^2} \\ &\leq \sqrt{\sum_{i=1}^k \delta^2 \|\mathbf{X}'_{[i]}\|_{(\mathbf{L}+\mathbf{I})}^2} \leq \delta \sqrt{n+1} \|\mathbf{X}'\|_F \\ &\leq \delta \sqrt{n+1} \sqrt{\sum_{u=1}^n (1 + \epsilon/32) \|\mathbf{X}\mathbf{e}_u\|^2} \\ &\leq \delta \sqrt{n+1} \sqrt{1 + \epsilon/32} \sqrt{\sum_{u=1}^n \mathbf{e}_u^\top (\mathbf{L} + \mathbf{I})^{-1} \mathbf{e}_u} \\ &\leq \delta \sqrt{n} \sqrt{n+1} \sqrt{1 + \epsilon/32}.\end{aligned}$$

同理，也可以求得 $|\|\tilde{\mathbf{Y}}\mathbf{e}_u\| - \|\mathbf{Y}'\mathbf{e}_u\||$ 的上界，即

$$\begin{aligned}|\|\tilde{\mathbf{Y}}\mathbf{e}_u\| - \|\mathbf{Y}'\mathbf{e}_u\|| &\leq \|\tilde{\mathbf{Y}}\mathbf{e}_u - \mathbf{Y}'\mathbf{e}_u\| \leq \|\tilde{\mathbf{Y}} - \mathbf{Y}'\|_F \\ &= \sqrt{\sum_{i=1}^k \|\tilde{\mathbf{Y}}_{[i]} - \mathbf{Y}'_{[i]}\|^2} \leq \sqrt{\sum_{i=1}^k \|\tilde{\mathbf{Y}}_{[i]} - \mathbf{Y}'_{[i]}\|_{(\mathbf{L}+\mathbf{I})}^2} \\ &\leq \sqrt{\sum_{i=1}^k \delta^2 \|\mathbf{Y}'_{[i]}\|_{(\mathbf{L}+\mathbf{I})}^2} \leq \delta \sqrt{n+1} \|\mathbf{Y}'\|_F \\ &\leq \delta \sqrt{n+1} \sqrt{\sum_{u=1}^n (1 + \epsilon/32) \|\mathbf{Y}\mathbf{e}_u\|^2}\end{aligned}$$

$$\begin{aligned}
&\leq \delta \sqrt{n+1} \sqrt{1+\epsilon/32} \sqrt{\sum_{u=1}^n \mathbf{e}_u^\top (\mathbf{L} + \mathbf{I})^{-1} \mathbf{e}_u} \\
&\leq \delta \sqrt{n} \sqrt{n+1} \sqrt{1+\epsilon/32}.
\end{aligned}$$

此外，有

$$\begin{aligned}
&\|\mathbf{Y}'\mathbf{e}_u\|^2 \geq (1 - \epsilon/32) \|\mathbf{Y}\mathbf{e}_u\|^2 \\
&= (1 - \epsilon/32) \mathbf{e}_u^\top (\mathbf{L} + \mathbf{I})^{-2} \mathbf{e}_u \\
&= (1 - \epsilon/32) \sum_{k=1}^n w_{uk} w_{ku} \\
&= (1 - \epsilon/32) \sum_{k=1}^n w_{uk}^2 \\
&\geq (1 - \epsilon/32) w_{uu}^2 \\
&\geq \frac{4(1 - \epsilon/32)}{(n+1)^2}.
\end{aligned}$$

接下来，本节讨论以下两种情况。

① 若 $\|\mathbf{X}'\mathbf{e}_u\|^2 \geq \|\mathbf{Y}'\mathbf{e}_u\|^2$ ，则有

$$\begin{aligned}
&\frac{||\tilde{\mathbf{X}}\mathbf{e}_u|| - \|\mathbf{X}'\mathbf{e}_u\|}{\|\mathbf{X}'\mathbf{e}_u\|} \\
&\leq \frac{||\tilde{\mathbf{X}}\mathbf{e}_u|| - \|\mathbf{X}'\mathbf{e}_u\|}{\|\mathbf{Y}'\mathbf{e}_u\|} \\
&\leq \frac{\delta \sqrt{n} (n+1) \sqrt{n+1} \sqrt{1+\epsilon/32}}{2\sqrt{1-\epsilon/32}} \\
&\leq \frac{\epsilon}{192}.
\end{aligned}$$

所以，

$$\begin{aligned}
&||\tilde{\mathbf{X}}\mathbf{e}_u\|^2 - \|\mathbf{X}'\mathbf{e}_u\|^2| \\
&= ||\tilde{\mathbf{X}}\mathbf{e}_u\| - \|\mathbf{X}'\mathbf{e}_u\| \cdot ||\tilde{\mathbf{X}}\mathbf{e}_u\| + \|\mathbf{X}'\mathbf{e}_u\| \\
&\leq \frac{\epsilon}{192} \left(2 + \frac{\epsilon}{192}\right) \|\mathbf{X}'\mathbf{e}_u\|^2 \\
&\leq \frac{\epsilon}{64} \|\mathbf{X}'\mathbf{e}_u\|^2 \\
&\leq \frac{\epsilon}{32} \|\mathbf{X}'\mathbf{e}_u\|^2.
\end{aligned}$$

同理可得

$$||\tilde{\mathbf{Y}}\mathbf{e}_u||^2 - \|\mathbf{Y}'\mathbf{e}_u\|^2 \leq \frac{\epsilon}{32} \|\mathbf{Y}'\mathbf{e}_u\|^2.$$

因此, 有

$$\begin{aligned} (1 - \epsilon/32) (\|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2) &\leq \|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2, \\ \|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2 &\leq (1 + \epsilon/32) (\|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2). \end{aligned}$$

这是第一种情况。

② 若 $\|\mathbf{X}'\mathbf{e}_u\|^2 \leq \|\mathbf{Y}'\mathbf{e}_u\|^2$, 则有

$$\|\mathbf{Y}'\mathbf{e}_u\|^2 \leq \|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2 \leq 2\|\mathbf{Y}'\mathbf{e}_u\|^2.$$

所以

$$\begin{aligned} &\frac{||\tilde{\mathbf{X}}\mathbf{e}_u|| - \|\mathbf{X}'\mathbf{e}_u\|}{\|\mathbf{X}'\mathbf{e}_u\|} \\ &\leq \frac{||\tilde{\mathbf{X}}\mathbf{e}_u|| - \|\mathbf{X}'\mathbf{e}_u\|}{2\|\mathbf{Y}'\mathbf{e}_u\|} \\ &\leq \frac{\delta\sqrt{n}(n+1)\sqrt{n+1}\sqrt{1+\epsilon/32}}{4\sqrt{1-\epsilon/32}} \\ &\leq \frac{\epsilon}{384}. \end{aligned}$$

所以,

$$\begin{aligned} &||\tilde{\mathbf{X}}\mathbf{e}_u\|^2 - \|\mathbf{X}'\mathbf{e}_u\|^2 \\ &= ||\tilde{\mathbf{X}}\mathbf{e}_u\| - \|\mathbf{X}'\mathbf{e}_u\| \cdot ||\tilde{\mathbf{X}}\mathbf{e}_u\| + \|\mathbf{X}'\mathbf{e}_u\| \\ &\leq \frac{\epsilon}{384} \left(2 + \frac{\epsilon}{384}\right) \|\mathbf{X}'\mathbf{e}_u\|^2 \\ &\leq \frac{\epsilon}{64} \|\mathbf{X}'\mathbf{e}_u\|^2 \\ &\leq \frac{\epsilon}{64} \|\mathbf{Y}'\mathbf{e}_u\|^2 \\ &\leq \frac{\epsilon}{64} (\|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2). \end{aligned}$$

同理可证

$$||\tilde{\mathbf{Y}}\mathbf{e}_u\|^2 - \|\mathbf{Y}'\mathbf{e}_u\|^2 \leq \frac{\epsilon}{64} (\|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2).$$

因此,

$$\begin{aligned} & |(\|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2) - (\|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2)| \\ & \leq ||\|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 - \|\mathbf{X}'\mathbf{e}_u\|^2| + ||\|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2 - \|\mathbf{Y}'\mathbf{e}_u\|^2| \\ & \leq \frac{\epsilon}{32}(\|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2). \end{aligned}$$

这是第二种情况。

由①②可知, 对于任意的 $u \in V$ 都有

$$\begin{aligned} (1 - \epsilon/32)(\|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2) & \leq \|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2, \\ \|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2 & \leq (1 + \epsilon/32)(\|\mathbf{X}'\mathbf{e}_u\|^2 + \|\mathbf{Y}'\mathbf{e}_u\|^2). \end{aligned}$$

又因为

$$\begin{aligned} (1 - \epsilon/32)\|\mathbf{X}\mathbf{e}_u\|^2 & \leq \|\mathbf{X}'\mathbf{e}_u\|^2 \leq (1 + \epsilon/32)\|\mathbf{X}\mathbf{e}_u\|^2, \\ (1 - \epsilon/32)\|\mathbf{Y}\mathbf{e}_u\|^2 & \leq \|\mathbf{Y}'\mathbf{e}_u\|^2 \leq (1 + \epsilon/32)\|\mathbf{Y}\mathbf{e}_u\|^2, \end{aligned}$$

以及 $w_{uu} = \|\mathbf{X}\mathbf{e}_u\|^2 + \|\mathbf{Y}\mathbf{e}_u\|^2$, 所以

$$\begin{aligned} \|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2 & \geq (1 - \epsilon/32)^2(\|\mathbf{X}\mathbf{e}_u\|^2 + \|\mathbf{Y}\mathbf{e}_u\|^2) \geq (1 - \epsilon/4)w_{uu}, \\ \|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2 & \leq (1 + \epsilon/32)^2(\|\mathbf{X}\mathbf{e}_u\|^2 + \|\mathbf{Y}\mathbf{e}_u\|^2) \leq (1 + \epsilon/4)w_{uu}, \end{aligned}$$

即

$$(1 - \frac{\epsilon}{4})w_{uu} \leq \tilde{w}_{uu} \leq (1 + \frac{\epsilon}{4})w_{uu},$$

由此得证。 □

根据引理 4.14 的结论, 本节最后分析森林中心性的近似误差。

定理 4.15 在无权无向图 $G = (V, E)$ 中, 令 \tilde{w}_{uu} 为森林矩阵 \mathbf{W} 对角线元素 w_{uu} 的估计值, 且

$$\left(1 - \frac{\epsilon}{4}\right)w_{uu} \leq \tilde{w}_{uu} \leq \left(1 + \frac{\epsilon}{4}\right)w_{uu},$$

其中 $\epsilon \in (0, 1/2)$ 。则对任意的 $u \in V$ 均成立

$$(1 - \epsilon)f_{\text{FC}}(u) \leq \tilde{f}_{\text{FC}}(u) \leq (1 + \epsilon)f_{\text{FC}}(u),$$

其中 $f_{\text{FC}}(u)$ 表示节点 u 的森林中心性值, $\tilde{f}_{\text{FC}}(u) = 1/\tilde{w}_{uu}$ 。

证明： 因为

$$\tilde{w}_{uu} \leq \left(1 + \frac{\epsilon}{4}\right) w_{uu},$$

所以

$$\begin{aligned} \frac{1}{\tilde{w}_{uu}} &\geq \frac{1}{1 + \epsilon/4} \cdot \frac{1}{w_{uu}} \\ &= \frac{4}{4 + \epsilon} \cdot \frac{1}{w_{uu}} = \left(1 - \frac{\epsilon}{4 + \epsilon}\right) \frac{1}{w_{uu}} \\ &\geq (1 - \epsilon) \frac{1}{w_{uu}}. \end{aligned}$$

由于

$$\tilde{w}_{uu} \geq \left(1 - \frac{\epsilon}{4}\right) w_{uu},$$

因此

$$\begin{aligned} \frac{1}{\tilde{w}_{uu}} &\leq \frac{1}{1 - \epsilon/4} \cdot \frac{1}{w_{uu}} \\ &= \frac{4}{4 - \epsilon} \cdot \frac{1}{w_{uu}} = \left(1 + \frac{\epsilon}{4 - \epsilon}\right) \frac{1}{w_{uu}} \\ &\leq (1 + \epsilon) \frac{1}{w_{uu}}. \end{aligned}$$

综上所述，

$$(1 - \epsilon) \frac{1}{w_{uu}} \leq \frac{1}{\tilde{w}_{uu}} \leq (1 + \epsilon) \frac{1}{w_{uu}}.$$

又因为 $f_{\text{FC}}(u) = 1/w_{uu}$ 且 $\tilde{f}_{\text{FC}}(u) = 1/\tilde{w}_{uu}$ ，所以

$$(1 - \epsilon)f_{\text{FC}}(u) \leq \tilde{f}_{\text{FC}}(u) \leq (1 + \epsilon)f_{\text{FC}}(u),$$

得证。 □

引理 4.14 和定理 4.15 的结论表明，只要正确设定参数 ϵ 以及 δ 的值，就能够将森林中心性的近似误差控制在理想的范围内。这些结论为后面的算法设计工作奠定了十分重要的理论基础。

4.5.2 算法设计

完成误差分析工作后，本节提出一个用于计算所有节点森林中心性值的快速近似算法。这个算法的主要思想是用 JL 引理实现降维以及用求解器避免矩阵的求逆运算。该算法的具体表述如 Algorithm 1 所示。

Algorithm 1: Approx(G, ϵ)

Input : G : a graph
 ϵ : the error parameter
Output : $\{u, \tilde{f}_{\text{FC}}(u)\}$, $u \in V$: every vertex u and its approximate forest centrality $f_{\text{FC}}(u)$

- 1 $\delta = \frac{\epsilon \sqrt{1-\epsilon/32}}{96\sqrt{n(n+1)}\sqrt{n+1}\sqrt{1+\epsilon/32}}$
- 2 Initialize matrices $\mathbf{Q}_{k \times m}$ and $\mathbf{P}_{k \times n}$ with entries being $\frac{1}{\sqrt{k}}$ or $-\frac{1}{\sqrt{k}}$, where
 $k = \left\lceil 24 \cdot \frac{\log(n+1)}{(\epsilon/32)^2} \right\rceil$
- 3 $\hat{\mathbf{X}} = \mathbf{Q}\mathbf{B}$, $\hat{\mathbf{Y}} = \mathbf{P}$
- 4 **for** $i = 1$ **to** k **do**
- 5 $\tilde{\mathbf{X}}_{[i:]} = \text{LapSolve}(\mathbf{L} + \mathbf{I}, \hat{\mathbf{X}}_{[i:]}, \delta)$
- 6 $\tilde{\mathbf{Y}}_{[i:]} = \text{LapSolve}(\mathbf{L} + \mathbf{I}, \hat{\mathbf{Y}}_{[i:]}, \delta)$
- 7 **for** $u = 1$ **to** n **do**
- 8 $\tilde{w}_{uu} = \|\tilde{\mathbf{X}}\mathbf{e}_u\|^2 + \|\tilde{\mathbf{Y}}\mathbf{e}_u\|^2$
- 9 $\tilde{f}_{\text{FC}}(u) = \frac{1}{\tilde{w}_{uu}}$
- 10 **return** $\{u, \tilde{f}_{\text{FC}}(u)\}$

根据引理 4.14 和定理 4.15 的结论，算法 Approx 输出的森林中心性的近似值 $\tilde{f}_{\text{FC}}(u)$ 与森林中心性的精确值之间的相对误差不超过 ϵ 。因此，只需将参数 ϵ 设定为一个较小的值就可以获得很好的近似效果。

本节接着分析算法 Approx 的时间复杂度。算法前两行对应的操作是设定参数 k 和 δ 的值以及生成随机矩阵 \mathbf{Q} 和 \mathbf{P} ，时间复杂度为 $O(km)$ 。算法第三行是求稠密矩阵 \mathbf{Q} 与稀疏矩阵 \mathbf{B} 的积。由于 \mathbf{Q} 是一个 k 行 m 列的矩阵，且稀疏矩阵 \mathbf{B} 有 $O(m)$ 个非零元素。因此，这个矩阵乘法的时间复杂度是 $O(km)$ 。算法的 4 到 6 行对应的操作是运行 $2k$ 次求解器。因为求解器运行一次的时间复杂度是 $O(m \log^3 n \log(1/\delta))$ ，所以这部分的时间复杂度是 $O(km \log^3 n \log(1/\delta))$ 。算法的 4 至 6 行是求解 $2n$ 个 k 维向量的模的平方，时间复杂度是 $O(kn)$ 。综上所述，算法 Approx 耗时最多的部分是该算法 4 到 6 行的部分。又因为 $O(k) = O(\epsilon^{-2} \log n)$ 且 $O(\log(1/\delta)) = O(\log n \log(1/\epsilon))$ ，所以算法 Approx 的时间复杂度是 $O(\epsilon^{-2} m \log^5 n \log(1/\epsilon))$ 。

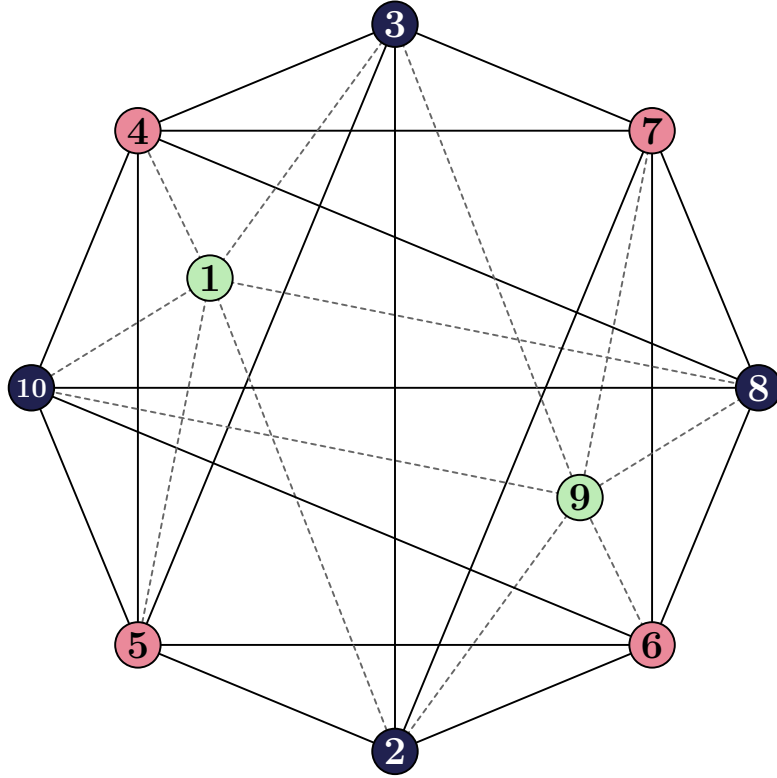


图 4.5: 示例网络，等价节点用相同颜色标出

4.6 实验

本章的实验分两个部分进行。第一部分实验主要测试森林中心性（Forest Centrality, FC）的区分能力；第二部分实验则着重于测试 4.5 节中提出的快速近似算法的性能。

4.6.1 区分能力

本节分别在示例网络、现实网络以及模型网络上测试森林中心性的区分能力。因为本章在定理 4.7 中已经证明了森林中心性是满足自同构等价公理的，所以本节可以通过计算它的广义区分度，即 P_{FC} 值，的方式来评估它的区分能力。

I 示例网络

本节首先在一个小型的示例网络上测试森林中心性的区分能力。如图 4.5 所示，该示例网络共包含 10 个节点。网络中所有节点的中心性值如表 4.1 所示。表 4.1 的第一列中的 DC、BC、CC、PR、EC、IC 和 FC 分别对应度中心性、介数中心性、接近中心性、PageRank 中心性、特征向量中心性、信息中心性以及森

表 4.1: 图 4.5 所示网络所有节点的中心性值

u	1,9	2,3,8,10	4,5,6,7
$f_{DC}(u)$	6	6	6
$f_{BC}(u)$	3	3	3
$f_{CC}(u)$	0.8333	0.8333	0.8333
$f_{PR}^{0.85}(u)$	0.1	0.1	0.1
$f_{EC}(u)$	0.1	0.1	0.1
$f_{IC}(u)$	3.5008	3.5035	3.5031
$f_{FC}(u)$	4.4961	4.4982	4.4979

林中心性。示例网络上的实验结果表明，森林中心性不仅能给等价的节点赋予相同的中心性值，而且可以给不等价赋予不同的中心性值以区分这些节点。在这个实验中，森林中心性的区分能力与信息中心性相当，且明显优于其它几个中心性度量。

II 现实网络

本节接着在现实网络上评估森林中心性的区分能力。在现实网络以及模型网络上的实验中，本节主要通过计算森林中心性以及其它五个常用中心性度量，即介数中心性（BC）、接近中心性（CC）、PageRank 中心性（PR）、特征向量中心性（EC）以及信息中心性（IC），的广义区分度来评估森林中心性的区分能力。在一个包含 n 个节点的网络 $G = (V, E)$ 中，中心性度量 c 的 P_c 值的计算方式如 (4.5) 式所示。

$$P_c = \frac{|\{(u, v) \mid f_c(u) \neq f_c(v), u, v \in V, u \neq v\}|}{n(n-1)}. \quad (4.5)$$

在现实网络以及模型网络上的所有实验中，PageRank 中心性的随机跳跃因子 α 设置为 0.85。此外，若两个不同节点的中心性值的差距在 10^{-9} 以内，则视为相等。

本节在 30 个不同的现实网络中比较森林中心性与五个常用中心性度量的区分能力。这些现实网络都来自 Koblenz Network Collection [87] 以及 Network Repository [88]。实验中的现实网络均为无权无向简单图。实验在每个网络的最大连通分支（Largest Connected Component, LCC）上进行，因为接近中心性、特征向量中心性以及信息中心性在不连通图上没有明确的定义。所有现实网络的相关信息如表 4.2 所示，其中 n 和 m 分别表示网络包含的节点数和边数； n' 和 m' 分别表示网络的最大连通分支包含的节点数和边数。

表 4.2: 现实网络数据的统计信息

Network	n	m	n'	m'
Karate	34	78	34	78
Dolphins	62	159	62	159
Celegans	453	2,025	453	2,025
Diseasome	516	1,188	516	1,188
GridMouse	1,450	1,636	791	1,098
CrimeMoreno	829	1,473	829	1,473
WikiVote	889	2,914	889	2,914
Reed98	962	18,812	962	18,812
EmailUniv	1,133	5,451	1,133	5,451
GridPlant	1,717	3,098	1,272	2,726
Yeast	1,458	1,948	1,458	1,948
Hamster	2,426	16,630	2,000	16,097
USFCA72	2,682	65,252	2,672	65,244
NipsEgo	2,888	2,981	2,888	2,981
GridWorm	3,507	6,531	3,343	6,437
GrQc	4,158	13,422	4,158	13,422
USgrid	4,941	6,594	4,941	6,594
Erdos992	5,094	7,515	4,991	7,428
Advogato	5,167	39,432	5,054	39,374
Bcspwr10	5,300	8,271	5,300	8,271
Reality	6,809	7,680	6,809	7,680
PagesGovernment	7,057	89,429	7,057	89,429
WikiElec	7,115	100,753	7,066	100,727
Dmela	7,393	25,569	7,393	25,569
HepPh	11,204	117,619	11,204	117,619
Anybeat	12,645	49,132	12,645	49,132
PagesCompany	14,113	52,126	14,113	52,126
AstroPh	17,903	196,972	17,903	196,972
CondMat	21,363	91,286	21,363	91,286
Gplus	23,628	39,194	23,613	39,182

表 4.3: 中心性度量的广义区分度

Network	不同中心性度量的 P_c 值 ($\times 10^{-2}$)					
	P_{BC}	P_{CC}	P_{PR}	P_{EC}	P_{IC}	P_{FC}
Karate	87.87879	94.83066	97.68271	97.68271	97.68271	97.68271
Dolphins	98.09625	98.67795	99.89424	99.89424	99.89424	99.89424
Celegans	93.47907	99.49794	99.94725	99.93358	99.95605	99.95116
Diseasome	52.36773	99.73658	99.78701	99.76594	99.78475	99.78701
GridMouse	70.26997	99.36757	99.38805	97.46451	99.38325	99.38805
CrimeMoreno	93.10706	99.90880	99.96970	99.97028	99.97028	99.97028
WikiVote	93.65214	99.93818	99.98505	99.98556	99.98556	99.98556
Reed98	99.65451	99.89335	99.99762	99.99784	99.99784	99.99784
EmailUniv	96.95748	99.94012	99.99470	99.99485	99.99485	99.99485
GridPlant	79.34672	99.75889	99.77807	93.01830	99.77770	99.77819
Yeast	71.18673	99.78384	99.80116	99.80069	99.80069	99.80172
Hamster	92.34762	99.93692	99.96508	99.96498	99.96518	99.96518
USFCA72	99.77220	99.96393	99.99944	99.99992	99.99992	99.99989
NipsEgo	1.44921	82.88608	82.88608	82.88608	82.88608	82.88608
GridWorm	68.92715	98.91129	98.93478	98.93598	98.93603	98.93617
GrQc	69.68902	99.95909	99.96818	99.76556	99.96876	99.96877
USgrid	91.24768	99.99145	99.99356	10.96066	99.99426	99.99442
Erdos992	41.45339	99.81120	99.82120	99.82166	99.82166	99.82166
Advogato	91.88240	99.96094	99.97316	99.97454	99.97457	99.97457
Bcspwr10	99.43660	99.99722	99.99819	55.56521	99.99919	99.99919
Reality	12.12196	98.54250	98.54259	98.54262	98.54262	98.54262
PagesGovernment	99.10966	99.99083	99.99804	99.99855	99.99888	99.99888
WikiElec	87.77373	99.97009	99.97233	99.98116	99.98116	99.98116
Dmela	92.45899	99.98207	99.98874	99.99039	99.99039	99.99038
HepPh	75.13847	99.97706	99.97926	99.84894	99.98062	99.98062
Anybeat	63.94228	96.94333	96.95285	96.95613	96.95623	96.95631
PagesCompany	94.56037	99.99492	99.99551	99.98403	99.99762	99.99762
AstroPh	81.83606	99.99136	99.99191	99.99209	99.99421	99.99422
CondMat	73.30385	99.99477	99.99396	99.99573	99.99690	99.99690
Gplus	32.75618	98.87605	98.87107	98.87817	98.87817	98.87817
ER	99.44611	99.93232	99.99919	99.99938	99.99938	99.99938
BA	99.97225	99.93265	99.99833	99.99897	99.99897	99.99897

表 4.3 列出了森林中心性以及其它五个中心性度量在这些现实网络中的广义区分度，每个现实网络中最高的 P_c 值用黄色标出。表 4.3 所示的实验结果显示，森林中心性在其中的 27 个现实网络中取得了最高的 P_c 值，并且在其余三个现实网络，即 *Celegans*、*USFCA72* 以及 *Dmela*，中的广义区分度值也与最高值相差无几。现实网络上的实验结果表明，森林中心性的区分能力非常优秀，在不同网络中的表现也十分稳定，并未出现大的波动。

III 模型网络

本节接着在两种模型网络，即 Erdős-Rényi (ER) 模型网络 [89] 以及 Barabási-Albert (BA) 模型网络 [90]，上比较森林中心性与其它五个中心性度量的区分能力。

本节共生成了 50 个 ER 模型网络，每个模型网络包含 1000 个节点。每个 ER 网络中任意节点对连接的概率为 0.004。由于生成出的 ER 网络可能是不连通的，本节在每个 ER 网络的最大连通分支上进行实验。这些中心性度量在这些 ER 网络最大连通分支上的平均 P_c 值如表 4.3 所示。本节还生成了 50 个 BA 模型网络，每个网络包含 1000 个节点。每个 BA 网络中节点的平均度为 4。这些中心性度量在这些 BA 网络上的平均 P_c 值如表 4.3 所示。此外，由于这些 P_c 值的方差很小，即小于 10^{-6} ，可以忽略不计，因此表 4.3 中不再列出它们的方差值。

表 4.3 中模型网络的实验结果显示，森林中心性在这两种模型网络上都取得了最高的平均 P_c 值。模型网络上的实验结果再一次表明了森林中心性拥有十分优秀的区分能力且表现稳定。

4.6.2 近似算法

本章第二部分的实验内容是评估 4.5 节中提出的快速近似算法 Approx 的性能。这部分实验主要考察快速近似算法的运行时间以及计算结果的相对误差。本节将在 22 个现实网络上比较 Exact 与 Approx 的运行时间以及相对误差，其中 Exact 表示用矩阵求逆的方式直接计算森林矩阵 \mathbf{W} 从而算出所有节点森林中心性值的算法。

本节的所有实验在一个 Linux 系统上运行。计算机 CPU 的型号是 4 核 4.2GHz 英特尔 i7-7700K，内存大小为 32GB。本节实验的所有代码都用 Julia 语言编写。快速求解器 LapSolve 的代码参考自 [102]。

表 4.4: 现实网络的统计信息以及 Exact 和 Approx 的运行时间

Network	n	m	Time of Exact	Time of Approx with various ϵ (s)					
				0.10	0.15	0.20	0.25	0.30	
PowerGrid	4941	6594	1.6	4.5	3.0	2.4	2.2	2.1	
GrQc	5241	14484	1.9	5.4	3.5	2.8	2.5	2.3	
Reactome	6229	146160	2.3	25.3	17.0	13.2	12.0	11.4	
RouteViews	6474	12572	2.6	4.6	3.2	2.6	2.4	2.2	
HighEnergy	7610	15751	5.6	6.9	4.2	3.2	2.7	2.5	
HepTh	9875	25973	11.7	9.2	5.3	3.9	3.3	2.9	
PrettyGoodPrivacy	10680	24316	14.8	12.7	6.7	4.6	3.8	3.2	
HepPh	12006	118489	24.3	31.6	18.2	13.5	11.3	10.2	
AstroPh	18771	198050	79.2	69.3	38.1	29.7	24.8	23.0	
Internet	22963	48436	153.8	22.2	12.0	8.0	6.3	5.3	
CAIDA	26475	53381	230.8	28.5	12.9	9.5	7.1	6.3	
EnronEmail	36692	183831	706.4	77.6	43.0	32.5	27.6	24.8	
CondensedMatter	39577	175692	847.7	86.5	47.2	35.3	27.3	24.4	
Brightkite	58228	214078	2595.6	120.1	65.2	49.4	40.5	36.2	
Livemocha	104103	2193083	—	6394.4	6004.7	5902.7	5768.4	5723.1	
WordNet	146005	656999	—	713.2	526.7	463.4	436.6	420.1	
Gowalla	196591	950327	—	1546.1	1231.7	1135.4	1094.3	1063.1	
DBLP	317080	1049866	—	2222.5	1766.2	1599.5	1537.3	1504.7	
Amazon	334863	925872	—	1976.1	1499.7	1315.9	1246.2	1203.9	
Pennsylvania	1088092	1541898	—	7759.1	6227.1	5692.2	5462.2	5321.1	
Youtube	1134890	2987624	—	19632.6	18010.0	17422.1	17176.3	17147.8	
Texas	1379917	1921660	—	11842.8	9780.4	9086.7	8768.5	8616.8	

表 4.5: 算法 Approx 在不同 ϵ 取值下的平均相对误差

Network	σ with various $\epsilon (\times 10^{-2})$				
	0.10	0.15	0.20	0.25	0.30
PowerGrid	1.131	1.736	2.282	2.866	3.460
GrQc	1.105	1.724	2.218	2.760	3.313
Reactome	0.937	1.422	1.847	2.285	2.795
RouteViews	1.114	1.669	2.233	2.779	3.352
HighEnergy	1.096	1.641	2.236	2.755	3.333
HepTh	1.062	1.610	2.169	2.681	3.269
PrettyGoodPrivacy	1.052	1.575	2.110	2.683	3.159
HepPh	1.005	1.494	2.028	2.534	3.019
AstroPh	0.974	1.452	1.935	2.412	2.934
Internet	1.040	1.559	2.114	2.623	3.133
CAIDA	1.038	1.546	2.070	2.590	3.080
EnronEmail	0.983	1.484	1.970	2.442	2.946
CondensedMatter	1.004	1.500	2.003	2.501	2.999
Brightkite	0.953	1.436	1.908	2.387	2.875

实验中所有的现实网络都来自 Koblenz Network Collection [87] 以及 Network Repository [88]。这些现实网络均为无权无向简单图。表 4.4 的前三列给出了这些现实网络的统计信息，其中第一列是每个现实网络的名称，第二列是网络中的节点数，第三列是网络中的边数。最小的现实网络有 4941 个节点，最大的现实网络的节点数接近 140 万。表 4.4 中的现实网络以节点数升序排列，其中许多网络是不连通的。

表 4.4 的后 6 列给出了 Exact 以及 Approx 在不同规模的现实网络上的运行时间。由于 Exact 的时间复杂度和空间复杂度都很高，因此 Exact 无法在最后 7 个大型网络上正常运行，而快速近似算法 Approx 却能够在这些大型网络上正常运行。本节还测试了 Approx 在不同 ϵ 取值下的运行时间。结果显示，精度参数 ϵ 的取值越小，快速近似算法 Approx 的运行时间就越长。在网络节点数小于 20000 时，Approx 的运行时间略高于 Exact。然而，在节点数超过 20000 的大型网络上，快速近似算法 Approx 的运行时间明显少于 Exact。表 4.4 所示的实验结果表明，快速近似算法 Approx 运算速度很快，效率很高。

本节接着考察快速近似算法 Approx 给出的近似值与 Exact 给出的精确值之间的相对误差。相对误差 σ 的计算方式为

$$\sigma = \frac{1}{n} \sum_{u=1}^n \frac{|f_{\text{FC}}(u) - \tilde{f}_{\text{FC}}(u)|}{f_{\text{FC}}(u)},$$

其中 $f_{\text{FC}}(u)$ 表示节点 u 森林中心性的精确值, $\tilde{f}_{\text{FC}}(u)$ 表示由 **Approx** 给出的节点 u 森林中心性的近似值。表 4.5 列出了快速近似算法 **Approx** 在不同精度参数 ϵ 取值下的平均相对误差。值得注意的是, 表 4.5 中列出的最大平均相对误差也仅仅是 3.46%。由此可见, 快速近似算法 **Approx** 给出的近似值与矩阵求逆得出的精确值相差无几。

本章第二部分实验的结果表明, 用于计算森林中心性值的快速近似算法 **Approx** 的计算速度很快且运算精度很高。

4.7 本章小结

本章基于置根生成森林的概念提出了一个新的节点中心性度量, 即森林中心性。本章证明了森林中心性的值等于森林矩阵对应对角线元素的倒数, 使得森林中心性的值能在 $O(n^3)$ 的时间内算出。本章接着根据森林矩阵对角线元素的上下界推导出了森林中心性取值的上下界。本章随后证明了森林中心性满足节点中心性的自同构等价公理。此外, 本章还研究了森林中心性在路径图中的表现, 发现路径图中各个节点的森林中心性值是合理的, 部分的说明了森林中心性的有效性。为快速计算森林中心性的值, 本章提出了快速近似算法 **Approx** 且证明了该算法能够在 $O(m \log^5 n \log(1/\epsilon))$ 的时间内给出相对误差不超过 ϵ 的森林中心性的近似值。本章最后分别测试了森林中心性的区分能力以及算法 **Approx** 的性能。实验结果表明, 森林中心性的区分能力十分优秀且用于快速计算森林中心性值的近似算法 **Approx** 是高效精准的。

第五章 边中心性度量及其区分能力研究

5.1 引言

本章主要研究无权无向简单图上的边中心性度量及其区分能力。本章首先对边中心性度量的其区分能力展开深入研究，提出相应的量化评价指标。随后，本章将提出一个区分能力更强的边中心性度量，即森林边中心性（Forest Edge Centrality）。本章的具体章节安排如下。首先，本章在 5.2 节中将介绍三中常用的边中心性度量，在 5.3 节中会给出无向边等价的判断标准。随后，本章将分别讨论边中心性度量在无向边等价时和不等价时的情况，其中边等价时的情况在 5.4 节中讨论，边不等价时的情况在 5.5 节中讨论。之后，本章将提出一个区分能力更强的边中心性度量，即森林边中心性。其中，本章将在 5.6 节中介绍森林边中心性的定义并证明森林边中心性的值可以用森林矩阵中的元素来表示，在 4.3 节中研究森林中心性的一些重要特性。本章的实验部分以及总结部分将分别在 4.4 节和 4.5 节中呈现。

5.2 边中心性度量

为判断边在网络中的相对重要性，学者们提出了很多边中心性度量。在这一节中，本节主要回顾三种常用的边中心性度量，分别是：边介数中心性（Edge Betweenness, EB）、生成边中心性（Spanning Edge Centrality, SEC）以及双调和距离中心性（Biharmonic Distance-related Centrality, BDRC）。

5.2.1 边介数中心性

边介数中心性在社团挖掘 [103]、图可视化 [104] 等领域有着广泛的应用。与介数中心性类似，一条边的边介数中心性指的是通过这条边的最短路径的个数，其具体定义如下 [73]。

定义 5.1 (边界数中心性) 在一个无权无向简单图 $G = (V, E)$ 中，边 $(u, v) \in E$ 的边介数中心性定义为

$$f_{EB}(u, v) = \sum_{s, t \in V} \frac{\sigma_{st}(u, v)}{\sigma_{st}},$$

其中 σ_{st} 表示节点 s 与 t 之间最短路径的数量， $\sigma_{st}(u, v)$ 表示节点 s 与 t 之间经过边 (u, v) 的最短路径的数量。

边介数中心性的计算成本与介数中心性一样。在一个包含 n 个节点和 m 条边的简单图中，计算所有边的边介数中心性的时间复杂度为 $O(nm)$ [73]。

5.2.2 生成边中心性

另一个被广泛使用的边中心性度量是生成边中心性。对于图中的一条边而言，它的生成边中心性指的是这条边出现在一个从这个图的所有生成树中随机选出的一个生成树中的概率，其具体定义如下 [105]。

定义 5.2 (生成边中心性) 在一个无权无向简单图 $G = (V, E)$ 中，边 $(u, v) \in E$ 的生成边中心性定义为

$$f_{\text{SEC}}(u, v) = \frac{|\mathcal{T}_{(u,v)}|}{|\mathcal{T}|},$$

其中 \mathcal{T} 表示包含图 G 所有生成树的集合，而 $\mathcal{T}_{(u,v)}$ 表示包含图 G 的所有含边 (u, v) 的生成树的集合。

有研究指出，边 (u, v) 的生成边中心性等价于节点 u 和 v 之间的电阻距离 $R(u, v)$ [106-107]，即

引理 5.3 在无权无向图 $G = (V, E)$ 中，有

$$f_{\text{SEC}}(u, v) = R(u, v) = (\mathbf{e}_u - \mathbf{e}_v)^\top \mathbf{L}^\dagger (\mathbf{e}_u - \mathbf{e}_v),$$

其中 $\mathbf{L}^\dagger = (\mathbf{L} + \mathbf{J})^{-1} - \mathbf{J}$ 为图 G 的拉普拉斯矩阵的伪逆。

由引理 5.3 可知，在一个包含 n 个节点的网络中，计算出所有边的生成边中心性所需的时间复杂度为 $O(n^3)$ 。

5.2.3 双调和距离中心性

在许多现实场景中，一条边的重要性往往与基尔霍夫指数有着密切的关联 [69-70]。因此，有学者提出了双调和距离中心性。对于一条边 (u, v) 而言，它的双调和距离中心性等于通过改变这条边的电阻值 r_{uv} 所导致的基尔霍夫指数的改变量，其具体定义如下 [65]。

定义 5.4 (双调和距离中心性) 在无权无向图 $G = (V, E)$ 中, 边 (u, v) 的双调和中心性定义为

$$f_{\text{BDRC}}(u, v) = \frac{\partial K(G)}{\partial r_{uv}} = (\mathbf{e}_u - \mathbf{e}_v)^\top \mathbf{L}^{2\dagger} (\mathbf{e}_u - \mathbf{e}_v),$$

其中 $K(G)$ 表示图 G 的基尔霍夫指数, $\mathbf{L}^{2\dagger} = (\mathbf{L}^\dagger)^2$ 为图 G 的拉普拉斯矩阵的伪逆的平方。

由此可知, 在一个包含 n 个节点的网络中, 计算出所有边的双调和距离中心性所需的时间复杂度为 $O(n^3)$ 。

5.3 边之间的等价关系

在研究边中心性度量的区分能力时, 首先要弄清楚两条边在什么样的情况下是等价的。由于之前的研究并未涉及到边的等价问题, 因此, 本章首先要做的工作就是定义边之间的等价关系。在研究点中心性度量的区分能力时, 上一章使用了自同构等价 [85] 作为判别节点是否等价的标准。同样的, 本章也可以用图的自同构 [83-84] 来给出边自同构等价的定义, 其具体表述如下。

定义 5.5 (边自同构等价) 在无权无向简单图 $G = (V, E)$ 中, 令 (u_i, v_i) 和 (u_j, v_j) 为图 G 的边集 E 中的两条无向边。若存在图 G 的一个自同构 f 满足 $f(u_i) = u_j, f(v_i) = v_j$ 或者 $f(u_i) = v_j, f(v_i) = u_j$, 则 (u_i, v_i) 和 (u_j, v_j) 自同构等价, 记为 $(u_i, v_i) \cong (u_j, v_j)$ 。

接下来, 本节将说明由边自同构等价, 即定义 5.5, 所定义的关系是等价关系。

定理 5.6 在无权无向简单图 $G = (V, E)$ 中, 由定义 5.5 所给出的边集合 E 上的关系 \cong 是一个等价关系。

证明: 首先证明图 $G = (V, E)$ 中定义在边集合 E 上的关系 \cong 是自反的。对于任意的无向边 $(u, v) \in E$, 均存在图 G 的平凡自同构 $f: V \rightarrow V$ 使得 $f(u) = u$ 且 $f(v) = v$ 。由定义 5.5 可知, $(u, v) \cong (u, v)$ 。因此, 关系 \cong 是自反的。

接着证明关系 \cong 是对称的。由定义 5.5 可知, 对于任意自同构等价的两条边 (u_1, v_1) 和 (u_2, v_2) , 记为 $(u_1, v_1) \cong (u_2, v_2)$, 而言, 存在图 G 的一个自

同构 f 使得 $f(u_1) = u_2, f(v_1) = v_2$ 或者 $f(u_1) = v_2, f(v_1) = u_2$ 。在此不妨设 $f(u_1) = u_2, f(v_1) = v_2$ 。由于自同构 f 是一个双射，所以必然存在 f 的逆映射 $f^{-1}: V \rightarrow V$ 使得 $f(u_2) = u_1, f(v_2) = v_1$ 。由于 f 的逆映射 f^{-1} 也是图 G 的自同构，所以，由定义 5.5 可知， $(u_2, v_2) \cong (u_1, v_1)$ 。因此，关系 \cong 是对称的。

最后证明关系 \cong 是传递的。令 (u_1, v_1) 、 (u_2, v_2) 和 (u_3, v_3) 为图 G 的边集合 E 中满足 $(u_1, v_1) \cong (u_2, v_2)$ 以及 $(u_2, v_2) \cong (u_3, v_3)$ 的三条无向边。根据边自同构等价的定义，即定义 5.5，不妨设存在图 G 的两个自同构 f_1 和 f_2 ，使得 $f_1(u_1) = u_2, f_1(v_1) = v_2$ 以及 $f_2(u_2) = u_3, f_2(v_2) = v_3$ 。令 $f_3 = f_1 \circ f_2$ ，则有 $f_3(u_1) = u_3$ 以及 $f_3(v_1) = v_3$ 。因为任意两个自同构的复合仍是自同构，所以，由定义 5.5 可知， $(u_1, v_1) \cong (u_3, v_3)$ 。因此，关系 \cong 是传递的。

因此，定义在无权无向图 $G = (V, E)$ 的边集合 E 上的关系 \cong 是自反的、对称的以及传递的。所以，关系 \cong 是一个等价关系，得证。 \square

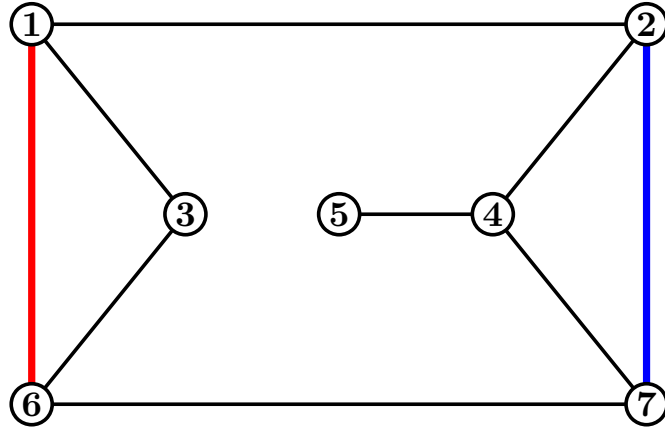


图 5.1: 一个包含 7 个节点和 9 条边的示例网络

为了便于读者理解边自同构等价的概念，本节在此给出一个示例。如图 5.1 所示，该网络共包含 7 个节点和 9 条无向边。除平凡自同构外，该示例网络只有一个非平凡自同构，如表 5.1 所示。根据边自同构等价的定义，即定义 5.5，可以推断出除自反等价关系外的三组自同构等价的无向边，分别是： $(1, 2) \cong (6, 7)$ 、 $(1, 3) \cong (6, 3)$ 以及 $(4, 2) \cong (4, 7)$ 。

合理的等价定义是研究中心性度量区分能力的前提。与节点的自同构等价的定义一样，边自同构等价的判定结果唯一且判定结果基本与常识相符。本节给出了边自同构等价的定义，即定义 5.5，作为判断边是否等价的标准。本节的结论为进一步探究边中心性度量的区分能力打下了坚实的理论基础。

表 5.1: 图 5.1所示网络的唯一非平凡自同构

u	(u, v)	$f(u)$	$(f(u), f(v))$
1	(1,2) (1,3) (1,6)	6	(6,7) (6,3) (6,1)
2	(2,1) (2,4) (2,7)	7	(7,6) (7,4) (7,2)
3	(3,1) (3,6)	3	(3,6) (3,1)
4	(4,2) (4,5) (4,7)	4	(4,7) (4,5) (4,2)
5	(5,4)	5	(5,4)
6	(6,1) (6,3) (6,7)	1	(1,6) (1,3) (1,2)
7	(7,2) (7,4) (7,6)	2	(2,7) (2,4) (2,1)

5.4 边中心性的自同构等价公理

本文在研究点中心性度量的区分能力时有这样一个发现：给等价的节点赋予相同的中心性值是一个点中心性度量具备区分能力的前提。上一章基于这个认识归纳出了自同构等价公理，并规定任何具备区分能力的点中心性度量必须满足这个公理。本章在研究边中心性度量的区分能力时也可以采用类似的手段。本节首先给出边中心性的自同构等价公理，随后将证明三种常用的边中心性度量，即边介数中心性、生成边中心性以及双调和距离中心性，均满足边中心性的自同构等价公理。

5.4.1 公理

本节先仿照点中心性的自同构等价公理，即公理 3.17，给出边中心性的自同构等价公理，其具体表述如下。

公理 5.7 (边中心性的自同构等价公理) 在无权无向简单图 $G = (V, E)$ 中，任意边中心性度量 c 都应满足以下特性：对于任意的自同构等价的两条无向边 (u_1, v_1) 和 (u_2, v_2) ，记为 $(u_1, v_1) \cong (u_2, v_2)$ ，均成立 $f_c(u_1, v_1) = f_c(u_2, v_2)$ ，其中 $f_c(u, v)$ 表示无向边 (u, v) 基于边中心性度量 c 的中心性值。

边中心性的自同构等价公理，即公理 5.7，是研究边中心性度量区分能力的前提。它规定了边中心性度量必须在两条边自同构等价时给出相同的中心性值。若一个边中心性度量，例如给每条无向边赋予一个随机的中心性值，不满足边中心性的自同构等价公理，则它的区分能力就无从谈起。

5.4.2 相关证明

接下来将证明三种常用的边中心性度量，即边介数中心性（EB）、生成边中心性（SEC）以及双调和距离中心性（BDRC），都满足边中心性的自同构等价公理，即公理 5.7。

I 边介数中心性

定理 5.8 在无权无向简单图 $G = (V, E)$ 中，若无向边 (u_i, v_i) 和 (u_j, v_j) 是自同构等价的，记为 $(u_i, v_i) \cong (u_j, v_j)$ ，则它们的边介数中心性值 $f_{EB}(u_i, v_i)$ 和 $f_{EB}(u_j, v_j)$ 相等，即 $f_{EB}(u_i, v_i) = f_{EB}(u_j, v_j)$ 。

证明： 已知 $(u_i, v_i) \cong (u_j, v_j)$ 。根据边自同构等价的定义，即定义 5.5，不妨设存在图 G 的自同构 f 使得 $f(u_i) = u_j$ 且 $f(v_i) = v_j$ 。令 $\mathbf{p}_{st} = (s, v_1, v_2, \dots, t)$ 表示图 G 中从节点 s 到节点 t 的一条最短路径。由于自同构 f 是一个双射，所以每个 \mathbf{p}_{st} 都对应一个唯一的从节点 $f(s)$ 至节点 $f(t)$ 的最短路径 $\mathbf{p}_{f(s)f(t)} = (f(s), f(v_1), f(v_2), \dots, f(t))$ ，由此可知 $\sigma_{st} = \sigma_{f(s)f(t)}$ 。同理，也可以证明 $\sigma_{st}(u) = \sigma_{f(s)f(t)}(f(u)) = \sigma_{f(s)f(t)}(v)$ 。由边介数中心性的相关定义，即定义 5.1，可知

$$\begin{aligned} f_{EB}(u_i, v_i) &= \sum_{s, t \in V} \frac{\sigma_{st}(u_i, v_i)}{\sigma_{st}} \\ &= \sum_{s, t \in V} \frac{\sigma_{f(s)f(t)}(f(u_i), f(v_i))}{\sigma_{f(s)f(t)}} \\ &= \sum_{s, t \in V} \frac{\sigma_{f(s)f(t)}(u_j, v_j)}{\sigma_{f(s)f(t)}}. \end{aligned}$$

因为自同构 f 是一个双射，所以有

$$\sum_{s, t \in V} \frac{\sigma_{f(s)f(t)}(u_j, v_j)}{\sigma_{f(s)f(t)}} = \sum_{s, t \in V} \frac{\sigma_{st}(u_j, v_j)}{\sigma_{st}}.$$

因此，

$$\begin{aligned} f_{EB}(u_i, v_i) &= \sum_{s, t \in V} \frac{\sigma_{f(s)f(t)}(u_j, v_j)}{\sigma_{f(s)f(t)}} \\ &= \sum_{s, t \in V} \frac{\sigma_{st}(u_j, v_j)}{\sigma_{st}} \\ &= f_{EB}(u_j, v_j), \end{aligned}$$

由此得证。 \square

II 生成边中心性

定理 5.9 在无权无向简单图 $G = (V, E)$ 中, 若无向边 (u_i, v_i) 和 (u_j, v_j) 是自同构等价的, 记为 $(u_i, v_i) \cong (u_j, v_j)$, 则它们的生成边中心性值 $f_{\text{SEC}}(u_i, v_i)$ 和 $f_{\text{SEC}}(u_j, v_j)$ 相等, 即 $f_{\text{SEC}}(u_i, v_i) = f_{\text{SEC}}(u_j, v_j)$ 。

证明: 已知 $(u_i, v_i) \cong (u_j, v_j)$ 。根据边自同构等价的定义, 即定义 5.5, 不妨设存在图 G 的自同构 f 使得 $f(u_i) = u_j$ 且 $f(v_i) = v_j$ 。令 \mathbf{P} 为自同构 f 所对应的排列矩阵, 则根据排列矩阵的定义可知 $\mathbf{e}_{u_i}^\top \mathbf{P} = \mathbf{e}_{u_j}^\top$ 以及 $\mathbf{e}_{v_i}^\top \mathbf{P} = \mathbf{e}_{v_j}^\top$ 。

由生成边中心性度量的定义, 即定义 5.2, 以及引理 5.3 的结论可知

$$f_{\text{SEC}}(u, v) = \frac{|\mathcal{T}_{(u,v)}|}{|\mathcal{T}|} = (\mathbf{e}_u - \mathbf{e}_v)^\top \mathbf{L}^\dagger (\mathbf{e}_u - \mathbf{e}_v).$$

因此,

$$\begin{aligned} & f_{\text{SEC}}(u_j, v_j) \\ &= (\mathbf{e}_{u_j} - \mathbf{e}_{v_j})^\top \mathbf{L}^\dagger (\mathbf{e}_{u_j} - \mathbf{e}_{v_j}) \\ &= (\mathbf{e}_{u_j} - \mathbf{e}_{v_j})^\top ((\mathbf{L} + \mathbf{J})^{-1} - \mathbf{J}) (\mathbf{e}_{u_j} - \mathbf{e}_{v_j}) \\ &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top \mathbf{P} ((\mathbf{L} + \mathbf{J})^{-1} - \mathbf{J}) \mathbf{P}^\top (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\ &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top ((\mathbf{P}\mathbf{L}\mathbf{P}^\top + \mathbf{P}\mathbf{J}\mathbf{P}^\top)^{-1} - \mathbf{P}\mathbf{J}\mathbf{P}^\top) (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\ &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top ((\mathbf{L} + \mathbf{J})^{-1} - \mathbf{J}) (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\ &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top \mathbf{L}^\dagger (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\ &= f_{\text{SEC}}(u_i, v_i), \end{aligned}$$

得证。 \square

III 双调和距离中心性

定理 5.10 在无权无向简单图 $G = (V, E)$ 中, 若无向边 (u_i, v_i) 和 (u_j, v_j) 是自同构等价的, 记为 $(u_i, v_i) \cong (u_j, v_j)$, 则它们的双调和距离中心性值 $f_{\text{BDRC}}(u_i, v_i)$ 和 $f_{\text{BDRC}}(u_j, v_j)$ 相等, 即 $f_{\text{BDRC}}(u_i, v_i) = f_{\text{BDRC}}(u_j, v_j)$ 。

证明： 已知 $(u_i, v_i) \cong (u_j, v_j)$ 。根据边自同构等价的定义，即定义 5.5，不妨设存在图 G 的自同构 f 使得 $f(u_i) = u_j$ 且 $f(v_i) = v_j$ 。令 \mathbf{P} 为自同构 f 所对应的排列矩阵，则根据排列矩阵的定义可知 $\mathbf{e}_{u_i}^\top \mathbf{P} = \mathbf{e}_{u_j}^\top$ 以及 $\mathbf{e}_{v_i}^\top \mathbf{P} = \mathbf{e}_{v_j}^\top$ 。由定理 5.9 的证明过程可知 $\mathbf{P} \mathbf{L}^\dagger \mathbf{P}^\top = \mathbf{L}^\dagger$ 。

因此，根据双调和距离中心性度量的相关定义，即定义 5.4，可得

$$\begin{aligned}
 & f_{\text{BDRC}}(u_j, v_j) \\
 &= (\mathbf{e}_{u_j} - \mathbf{e}_{v_j})^\top \mathbf{L}^{2\dagger} (\mathbf{e}_{u_j} - \mathbf{e}_{v_j}) \\
 &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top \mathbf{P} \mathbf{L}^{2\dagger} \mathbf{P}^\top (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\
 &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top \mathbf{P} \mathbf{L}^\dagger (\mathbf{P}^\top \mathbf{P}) \mathbf{L}^\dagger \mathbf{P}^\top (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\
 &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top (\mathbf{P} \mathbf{L}^\dagger \mathbf{P}^\top) (\mathbf{P} \mathbf{L}^\dagger \mathbf{P}^\top) (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\
 &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top \mathbf{L}^{2\dagger} (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\
 &= f_{\text{BDRC}}(u_i, v_i),
 \end{aligned}$$

由此得证。 □

至此，本节成功的证明了三种常用的边中心性度量，即边介数中心性、生成边中心性以及双调和距离中心性，均满足边中心性的自同构等价公理，即公理 5.7。本节中的几个重要的证明为进一步研究这三种常用的边中心性度量的区分能力打下了重要基础，同时也是研究这几种边中心性度量的区分能力的非常关键的第一步。

5.5 区分能力指数

上一节讨论了边中心性度量在无向边等价时的情况，给出了边中心性度量的自同构等价公理，即公理 5.7，规定了边中心性度量在无向边自同构等价时必须给出相等的中心性值。本节将进一步讨论边中心性度量在无向边不等价时的情况。

边中心性度量的区分能力本质上可以理解为它区分不等价无向边的能力。换言之，一个区分能力强的边中心性度量能在无向边等价时保证给出相等的中心性值，同时也能在边不等价时尽可能的给出不同的中心性值。实际上，许多常用的边中心性度量在无向边不等价时往往也会给出相同的中心性值。关于这一点，本节将举例来说明。图 5.1 给出了一个包含 7 个节点和 9 条边的小型网络。

表 5.2: 图 5.1 所示网络中所有边的中心性值.

边中心性度量	图 5.1 所示网络中边的中心性值					
	(1,2) (6,7)	(3,1) (3,6)	(4,2) (4,7)	(1,6)	(2,7)	(4,5)
EB	12	6	10	4	4	12
SEC	0.7	0.63	0.63	0.53	0.53	1
BDRC	0.47	0.25	0.39	0.15	0.15	0.86

这个网络除了平凡自同构之外仅有 1 个非平凡自同构, 如表 5.1 所示。根据边自同构等价的定义可以推断出 $(1, 2) \cong (6, 7)$ 、 $(1, 3) \cong (6, 3)$ 以及 $(4, 2) \cong (4, 7)$ 。本节在此列出这个小型网络中所有无向边基于 3 个常用边中心性度量的中心性值, 如表 5.2 所示, 其中表格第一列中的 EB、SEC、BDRC 分别表示边介数中心性、生成边中心性以及双调和距离中心性。从表 5.2 所给出的结果可以看出, 所有这 3 个常用的边中心性度量都能够在边自同构等价时给出相同的中心性值, 这与 5.4.2 节中证明的结论是一致的。此外, 对于一些不等价的无向边, 这些边中心性度量有时能给出不同的中心性值, 如 $(1, 2)$ 和 $(3, 1)$, 有时则会给出相同的中心性值, 如 $(1, 6)$ 和 $(2, 7)$ 。

上述分析指出, 所有边中心性度量都能在无向边等价时给出相同的中心性值, 同时边中心性度量在无向边不等价时有时能给出不同的中心性值予以区分, 有时则不能区分。由此, 本章提出了边中心性的区分能力指数的概念。简而言之, 一个边中心性的区分能力指数指的是它在无向边不等价时给出不同中心性值的概率。为便于表述, 在本节以及下一节中将使用 ϵ_i 表示无向边 (u_i, v_i) 。

定义 5.11 (边中心性的区分能力指数) 在无权无向简单图 $G = (V, E)$ 中, 若边中心性度量 c 满足边中心性的自同构等价公理, 即公理 5.7, 则它的区分能力指数, 记为 D_c , 定义为边中心性度量 c 对于从图 G 的无向边集合 E 中随机选出的一对不等价无向边给出不同中心性值的概率, 即

$$D_c = P(f_c(\epsilon_i) \neq f_c(\epsilon_j) \mid \epsilon_i \not\cong \epsilon_j),$$

其中 $\epsilon_i \not\cong \epsilon_j$ 表示无向边 ϵ_i 和 ϵ_j 不等价。

根据边中心性度量的区分能力指数的定义, 即定义 5.11, 可知, 一个边中心性度量的区分能力指数值, 即 D_c 值, 越高, 它区分不等价边的能力就越强。然而, 直接计算定义 5.11 中的区分能力指数, 即 D_c 值, 是不现实的。因为计算

D_c 需要找出图 G 的自同构群 $[\text{Aut}(G), \circ]$ 的所有生成元, 而这个问题到目前为止还找不到一个多项式时间内的解法。为解决边中心性度量的区分能力指数难以直接计算的问题, 本节接下来将给出一个简单的技巧以实现在不直接计算能力指数, 即 D_c 值, 的情况下比较两个不同边中心性度量的区分能力指数。为此, 本节先介绍边中心性的广义区分度的概念, 其具体表述如下。

定义 5.12 (边中心性的广义区分度) 在无权无向简单图 $G = (V, E)$ 中, 若边中心性度量 c 满足边中心性的自同构等价公理, 即公理 5.7, 则它的广义区分度, 记为 P_c , 定义为边中心性度量 c 对于从图 G 的无相边集合 E 中随机选出的一对无向边给出不同中心性值的概率, 即

$$P_c = P(f_c(\epsilon_i) \neq f_c(\epsilon_j)).$$

接下来, 本节将说明边中心性的区分能力指数与它的广义区分度之间的关联。

定理 5.13 在无权无向简单图 $G = (V, E)$ 中, 令 c_1 和 c_2 为满足边中心性自同构等价公理, 即公理 5.7, 的两个不同的边中心性度量, 则 $D_{c_1} < D_{c_2}$ 当且仅当 $P_{c_1} < P_{c_2}$ 。

证明: 令 c 为一个满足边中心性自同构等价公理的边中心性度量, 则根据公理 5.7 的表述可知

$$P(f_c(\epsilon_i) \neq f_c(\epsilon_j) \mid \epsilon_i \cong \epsilon_j) = 0.$$

因此, 由全概率公式可得

$$\begin{aligned} P_c &= P(f_c(\epsilon_i) \neq f_c(\epsilon_j) \mid \epsilon_i \cong \epsilon_j) \cdot P(\epsilon_i \cong \epsilon_j) \\ &\quad + P(f_c(\epsilon_i) \neq f_c(\epsilon_j) \mid \epsilon_i \not\cong \epsilon_j) \cdot P(\epsilon_i \not\cong \epsilon_j) \\ &= 0 \cdot P(\epsilon_i \cong \epsilon_j) + D_c \cdot P(\epsilon_i \not\cong \epsilon_j) \\ &= D_c (1 - P(\epsilon_i \cong \epsilon_j)). \end{aligned}$$

由于在一个给定的图中, $P(\epsilon_i \cong \epsilon_j)$ 可视为一个常数, 所以 P_c 值等于 D_c 值乘以一个常数。换言之, P_c 值的大小仅由 D_c 值的大小决定, 得证。□

定理 5.13 表明, 在比较不同边中心性度量的区分能力指数的大小时, 边中心性的广义区分度, 即 P_c , 与它的区分能力指数, 即 D_c , 效果一致。然而, 相较

于边中心性的区分能力指数而言, 它的广义区分度更易于计算。在已知所有 m 条无向边的中心性值的情况下, 用快速排序和二分查找算法能够在 $O(m \log m)$ 的时间内计算出广义区分度的值 [86]。

5.6 森林边中心性

本节主要介绍森林边中心性 (Forest Edge Centrality, FEC) 的定义。此外, 本节还将证明一条无向边的森林边中心性值可以用森林矩阵中的相关元素来表示。

5.6.1 定义

为便于读者理解, 本节在介绍森林边中心性的定义之前先通过一个例子简单回顾置根生成森林以及森林矩阵的定义。

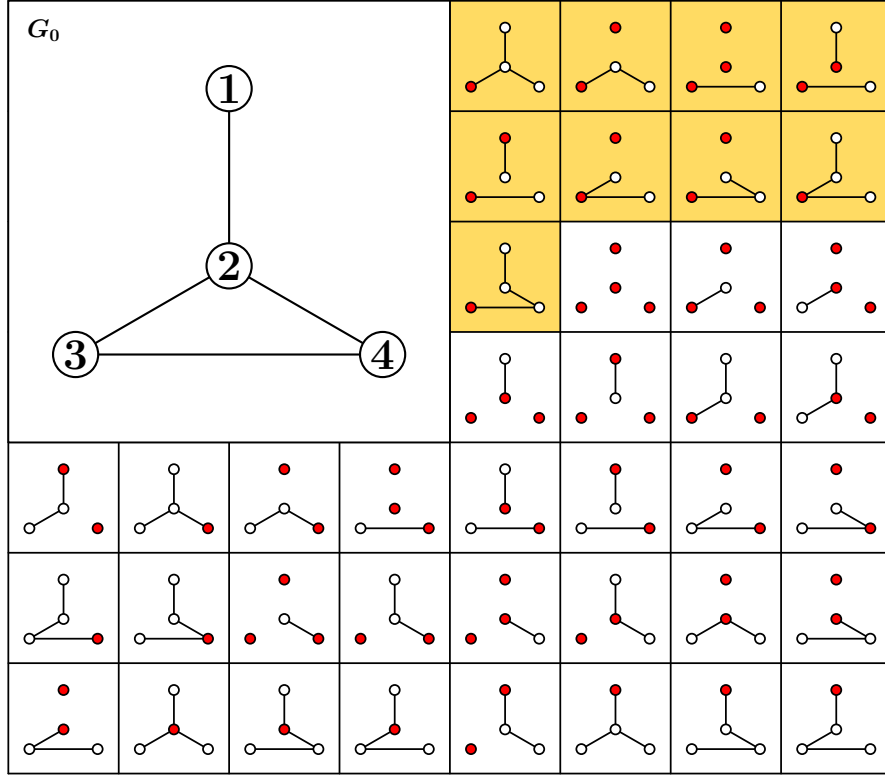
对于一个无权无向简单图 $G = (V, E)$ 而言, 它的置根生成森林 $F = (V_F, E_F, r_F)$ 是图 G 的一个无环子图, 其中 V_F 表示森林 F 的节点集合且有 $V_F = V$; E_F 表示森林 F 的边集合且有 $E_F \subseteq E$; $r_F : V_F \rightarrow \{0, 1\}$ 为置根函数, 若 V_F 中的一个节点 u 是根节点则 $r_F(u) = 1$ 否则 $r_F(u) = 0$ 。由于置根生成森林 F 是无环的, 因此 F 中的每个连通分支都是一棵树, 且每个连通分支有且仅有一个节点是根节点。本章用 $\mathcal{F} = \mathcal{F}(G)$ 表示图 G 的所有置根生成森林构成的集合, 用 $\mathcal{F}_{ij} = \mathcal{F}_{ij}(G)$ 表示所有节点 i 和 j 在同一个连通分支且这个连通分支的根节点是 i 的置根生成森林构成的集合。如图 5.2 所示, 示例网络 $G_0 = (\{1, 2, 3, 4\}, \{(1, 2), (2, 3), (2, 4), (3, 4)\})$ 共有 $|\mathcal{F}| = |\mathcal{F}(G_0)| = 40$ 个置根生成森林, 其中节点 3 是根节点且节点 4 和 3 在同一个连通分支的置根生成森林共有 $|\mathcal{F}_{34}| = |\mathcal{F}_{34}(G_0)| = 9$ 个, 这些置根生成森林在图 5.2 中用黄色底色标出。

图 $G = (V, E)$ 的森林矩阵 [96-97], 记为 \mathbf{W} , 定义为 $\mathbf{W} = (\mathbf{I} + \mathbf{L})^{-1} = (w_{ij})_{n \times n}$, 其中 \mathbf{I} 为 n 阶单位矩阵; \mathbf{L} 为图 G 的拉普拉斯矩阵。有研究指出, $|\mathcal{F}| = |\mathbf{I} + \mathbf{L}|$ 且 $w_{ij} = |\mathcal{F}_{ij}|/|\mathcal{F}|$ [98-99]。本节在此给出示例网络 G_0 的森林矩阵, 即

$$\mathbf{W} = \frac{1}{40} \begin{pmatrix} 24 & 8 & 4 & 4 \\ 8 & 16 & 8 & 8 \\ 4 & 8 & 19 & 9 \\ 4 & 8 & 9 & 19 \end{pmatrix},$$

其中矩阵第 3 行第 4 列的元素是 $w_{34} = 9/40 = |\mathcal{F}_{11}|/|\mathcal{F}|$ 。

森林边中心性的核心思想是在置根生成森林中考察一条边的重要程度。本节利用置根生成森林 F 考察无向边 (u, v) 时, 将出现以下三种情况。

图 5.2: 示例网络 G_0 以及它的全部 40 个置根生成森林

- ① 无向边 (u, v) 在置根生成森林 F 中且节点 u 和节点 v 连通。由于置根生成森林是无环的，因此无向边 (u, v) 是节点 u 和节点 v 连接的唯一路径。若把无向边 (u, v) 、节点 u 和节点 v 这三者看作是一个整体，则根据上一章中节点的森林中心性的定义可以发现，如果无向边 (u, v) 处于重要的位置，那么无向边 (u, v) 所在的连通分支包含的节点数大概率比较多，反之，无向边 (u, v) 处于比较边缘的位置，那么无向边 (u, v) 所在的连通分支包含的节点数大概率比较少。因此，在这种情况下，无向边 (u, v) 、节点 u 和节点 v 这三者所在的连通分支的大小将直接反映出无向边 (u, v) 的重要程度。
- ② 无向边 (u, v) 不在置根生成森林 F 中且节点 u 和节点 v 连通。因为无向边 (u, v) 最基础的功能是把节点 u 和节点 v 这两个节点直接连接起来，因此若节点 u 和节点 v 在没有无向边 (u, v) 的情况下仍然能够连通，则说明无向边 (u, v) 的作用在某种程度上可以被其它无向边替代。因此，此类情况出现的越多，无向边 (u, v) 越不重要。

- ③ 无向边 (u, v) 不在置根生成森林 F 中且节点 u 和节点 v 不连通。这种情况说明无向边 (u, v) 连通节点 u 和节点 v 的功能不能被其它无向边替代。因此, 此类情况出现的越多, 无向边 (u, v) 越重要。

值得注意的是, 前两种情况下无向边 (u, v) 的两个端点在置根生成森林 F 中都是连通的, 并且在集合 \mathcal{F}_{uv} 中的每个置根生成森林中节点 u 和节点 v 也都是连通的。因此, 本节可以用置根生成森林集合 \mathcal{F}_{uv} 来考察无向边 (u, v) 的重要性。

本节首先根据 ① 和 ② 中的分析来定义出无向边 (u, v) 在置根生成森林 F 中的得分, 其具体表述如下。

定义 5.14 在无权无向图 $G = (V, E)$ 中, 令 F 为集合 $\mathcal{F}_{uv} = \mathcal{F}_{uv}(G)$ 中的一个置根生成森林, 则无向边 (u, v) 在置根生成森林 F 中的得分, 记为 $s(F, u, v)$, 定义为

$$s(F, u, v) = \begin{cases} 0, & (u, v) \notin F; \\ |F|_u, & (u, v) \in F, \end{cases}$$

其中 $|F|_u$ 表示置根生成森林 F 中节点 u 所在连通分支包含的节点数量。

由于集合 \mathcal{F}_{uv} 中的每一个置根生成森林中节点 u 都是根节点且节点 u 和节点 v 处于同一个连通分支中, 因此若 $F \in \mathcal{F}_{uv}$, 则森林 F 中 u 点和 v 点必定是连通的。根据定义 5.14 的描述, 如果无向边 (u, v) 不在置根生成森林 F 中并且节点 u 和节点 v 仍然能够保持连通, 那么无向边 (u, v) 在这种情况下的得分为 0 分; 若节点 u 和节点 v 通过无向边 (u, v) 保持连通, 则 (u, v) 此时的得分为置根生成森林 F 中无向边 (u, v) 所在连通分支包含的节点数量。定义 5.14 制定的得分规则与上文中 ① 和 ② 两种情况的分析是吻合的。

为便于读者理解定义 5.14 中的得分规则, 本节以图 5.2 所示的示例网络中无向边 $(3, 4)$ 的情况为例来说明。如图 5.3 所示, 示例网络 G_0 的置根生成森林集合 $\mathcal{F}_{34} = \mathcal{F}_{34}(G_0)$ 共有 9 个森林。集合 \mathcal{F}_{34} 中的每个置根生成森林中节点 3 和节点 4 都是连通的且节点 3 都是根节点。在图 5.2 所展示的前两个置根生成森林都不包含无向边 $(3, 4)$, 所以 $(3, 4)$ 在这两个置根生成森林中的得分为 0。在接下来的三个置根生成森林中, 无向边 $(3, 4)$ 所在的连通分支仅仅包含节点 3 和节点 4 这两个节点, 所以 $(3, 4)$ 在这三个置根生成森林中的得分为 2。在图 5.2 所展示的第六个以及第七个置根生成森林中, 无向边 $(3, 4)$ 所在的连通分支包含节点 2、

	F									
	$s(F, 3, 4)$	0	0	2	2	2	3	3	4	4

图 5.3: 集合 $\mathcal{F}_{34}(G_0)$ 中全部 9 个置根生成森林以及 $(3, 4)$ 在每个森林中的得分

节点 3 以及节点 4 这三个节点，因此 $(3, 4)$ 在这两个置根生成森林中的得分为 3。在最后两个置根生成森林中，无向边 $(3, 4)$ 所在的连通分支包含示例网络 G_0 的全部 4 个节点，因此 $(3, 4)$ 在最后两个置根生成森林中的得分为 4。

基于定义 5.14 表述的得分规则，本节用无向边 (u, v) 在集合 \mathcal{F}_{uv} 的所有置根生成森林中的平均得分作为无向边 (u, v) 的森林边中心性 (Forest Edge Centrality, FEC)。森林边中心性的具体定义如下。

定义 5.15 (森林边中心性) 在无权无向图 $G = (V, E)$ 中，无向边 (u, v) 的森林边中心性等于 (u, v) 在集合 \mathcal{F}_{uv} 的所有置根生成森林中的平均得分，即

$$f_{\text{FEC}}(u, v) = \frac{1}{|\mathcal{F}_{uv}|} \sum_{F \in \mathcal{F}_{uv}} s(F, u, v), \quad (5.1)$$

其中 $s(F, u, v)$ 表示定义 5.14 规定的无向边 (u, v) 在置根生成森林 F 中的得分。

根据定义 5.15 的表述，示例网络 G_0 中无向边 $(3, 4)$ 的森林中心性值为 $(0 + 0 + 2 + 2 + 2 + 3 + 3 + 4 + 4)/9 = 20/9$ 。

5.6.2 计算

由于一个图的置根生成森林的数量十分庞大，本节无法通过直接枚举所有置根生成森林的方式来计算森林边中心性的值。因此，为高效率的计算森林边中心性的值，本节接下来将证明森林边中心性的值可以用森林矩阵 \mathbf{W} 中的元素来表示。

已知 $|\mathcal{F}|$ 等于矩阵 $\mathbf{I} + \mathbf{L}$ 行列式的值， w_{uv} 是矩阵 $\mathbf{W} = (\mathbf{I} + \mathbf{L})^{-1}$ 中的元素并且矩阵的行列式和矩阵的逆都可以在 $O(n^3)$ 时间内求解出来，则 $|\mathcal{F}|$ 和 w_{uv} 的值都能在 $O(n^3)$ 时间内求出。所以，根据 $w_{uv} = |\mathcal{F}_{uv}|/|\mathcal{F}|$ 可知， $|\mathcal{F}_{uv}| = w_{uv}|\mathcal{F}|$ 的值也可以在 $O(n^3)$ 时间内求出。因此，只需在多项式时间内求出 (5.1) 式的后半部分，即 $\sum_{F \in \mathcal{F}_{uv}} s(F, u, v)$ ，的值，就能够在多项式时间内计算出森林边中心性的值。

本节先把计算 $\sum_{F \in \mathcal{F}_{uv}} s(F, u, v)$ 的问题转换为求解一个集合大小的问题。

引理 5.16 在无权无向图 $G = (V, E)$ 中, 集合 \mathcal{S}_{uv} 定义为

$$\mathcal{S}_{uv} = \{(F, v^*) \mid F \in \mathcal{F}_{uv}, (u, v) \in F, v^* \in T(F, u)\}, \quad (5.2)$$

其中 $T(F, u)$ 表示置根生成森林 F 中节点 u 所在的连通分支, 则有

$$\sum_{F \in \mathcal{F}_{uv}} s(F, u, v) = |\mathcal{S}_{uv}|,$$

其中 $s(F, u, v)$ 是由定义 5.14 给出的无向边 (u, v) 在置根生成森林 F 中的得分。

证明: 由定义 5.14 可知, 只有无向边 (u, v) 出现在置根生成森林 F 中时, 得分 $s(F, u, v)$ 才不为 0。同时当 $(u, v) \in F$ 时, 得分 $s(F, u, v)$ 等于置根生成森林 F 中节点 u 所在的连通分支包含的节点数。因此, 有

$$s(F, u, v) = |\{v^* \mid (u, v) \in F, v^* \in T(F, u)\}|.$$

所以,

$$\begin{aligned} & \sum_{F \in \mathcal{F}_{uv}} s(F, u, v) \\ &= \sum_{F \in \mathcal{F}_{uv}} |\{v^* \mid (u, v) \in F, v^* \in T(F, u)\}| \\ &= |\{(F, v^*) \mid F \in \mathcal{F}_{uv}, (u, v) \in F, v^* \in T(F, u)\}| \\ &= |\mathcal{S}_{uv}|, \end{aligned}$$

得证。 □

引理 5.16 成功的把计算 $\sum_{F \in \mathcal{F}_{uv}} s(F, u, v)$ 的问题转换为求解集合 \mathcal{S}_{uv} 大小的问题。本节接着把集合 \mathcal{S}_{uv} 划分为两个不同的部分随后分别求出它们的大小。

引理 5.17 在无权无向图 $G = (V, E)$ 中, 集合 \mathcal{S}_{uv} 的定义如 (5.2) 式所示, 集合 \mathcal{S}_{uv}^0 和集合 \mathcal{S}_{uv}^1 分别定义为

$$\begin{aligned} \mathcal{S}_{uv}^0 &= \{(F, v^*) \mid (F, v^*) \in \mathcal{S}_{uv}, (u, v) \notin \mathbf{p}(F, u, v^*)\}, \\ \mathcal{S}_{uv}^1 &= \{(F, v^*) \mid (F, v^*) \in \mathcal{S}_{uv}, (u, v) \in \mathbf{p}(F, u, v^*)\}, \end{aligned}$$

其中 $\mathbf{p}(F, u, v^*)$ 表示置根生成森林 F 中节点 u 到节点 v^* 的路径, 则 \mathcal{S}_{uv}^0 和 \mathcal{S}_{uv}^1 构成集合 \mathcal{S}_{uv} 的一个划分。

证明: 由于置根生成森林 F 是无环的, 因此森林 F 中任意两个不同的节点之间要么没有路径, 要么仅有唯一的一条路径。根据集合 \mathcal{S}_{uv} 的定义可知, 若 $(F, v^*) \in \mathcal{S}_{uv}$, 则必有 $v^* \in T(F, u)$, 即节点 v^* 和节点 u 处于同一个连通分支。换言之, 此时节点 v^* 和节点 u 是连通的。又因为置根生成森林 F 是无环的, 所以节点 v^* 和节点 u 之间仅有唯一的一条路径 $\mathbf{p}(F, u, v^*)$ 。由于无向边 (u, v) 要么在路径 $\mathbf{p}(F, u, v^*)$ 中, 要么不在路径 $\mathbf{p}(F, u, v^*)$, 因此集合 \mathcal{S}_{uv} 中的元素要么属于 \mathcal{S}_{uv}^0 , 要么属于 \mathcal{S}_{uv}^1 。又因为路径 $\mathbf{p}(F, u, v^*)$ 是唯一的, 所以 $(u, v) \notin \mathbf{p}(F, u, v^*)$ 与 $(u, v) \in \mathbf{p}(F, u, v^*)$ 不可能同时成立, 所以集合 \mathcal{S}_{uv}^0 和集合 \mathcal{S}_{uv}^1 的交集为空集。综上所述, \mathcal{S}_{uv}^0 和 \mathcal{S}_{uv}^1 构成集合 \mathcal{S}_{uv} 的一个划分, 得证。 \square

有了引理 5.17 的结论, 本节仅需分别求出集合 \mathcal{S}_{uv}^0 和集合 \mathcal{S}_{uv}^1 的大小就能求得集合 \mathcal{S}_{uv} 的大小。

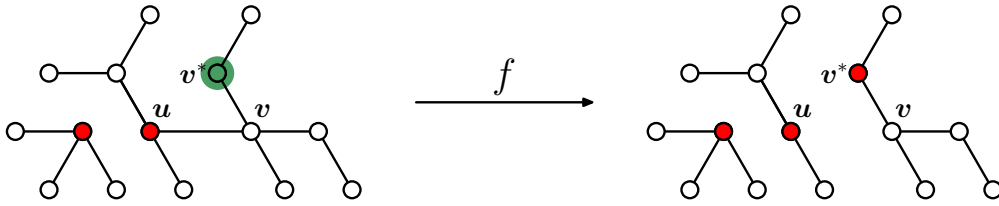


图 5.4: 映射 $f: \mathcal{S}_{uv}^1 \rightarrow \mathcal{F}_{uv}^*$ 的构造方式

引理 5.18 在无权无向图 $G = (V, E)$ 中, 令 \mathcal{F}_{uv}^* 为所有节点 u 是根节点且节点 u 和节点 v 不在同一连通分支的置根生成森林构成的集合, 即 $\mathcal{F}_{uv}^* = \mathcal{F}_{uu} \setminus \mathcal{F}_{uv}$, 则 $|\mathcal{S}_{uv}^1| = |\mathcal{F}_{uv}^*|$ 对于所有 $(u, v) \in E$ 成立。

证明: 本节通过构造映射 $f: \mathcal{S}_{uv}^1 \rightarrow \mathcal{F}_{uv}^*$ 的方式来证明集合 \mathcal{S}_{uv}^1 的与集合 \mathcal{F}_{uv}^* 大小相等。令 (F_0, v_0^*) 为集合 \mathcal{S}_{uv}^1 中任意一个元素, 则映射 f 可以理解为对置根生成森林 F_0 做一个操作, 即删去 F_0 中的无向边 (u, v) 且令节点 v^* 为 F_0 中节点 v^* 所在连通分支的根节点。这样就能得到新的置根生成森林 $f(F_0, v_0^*) \in \mathcal{F}_{uv}^*$ 。图 5.4 揭示了映射 f 的构造过程。

首先证明映射 $f: \mathcal{S}_{uv}^1 \rightarrow \mathcal{F}_{uv}^*$ 是一个单射。令 $(F_0, v_0^*) \neq (F_1, v_1^*)$ 为集合 \mathcal{S}_{uv}^1 中任意两个不相等的元素。接下来, 本节将讨论以下三种情况。

- ① 若 $F_0 = F_1$, 则由 $(F_0, v_0^*) \neq (F_1, v_1^*)$ 可知, 必有 $v_0^* \neq v_1^*$ 。根据映射 f 的构造方式可知: 节点 v_0^* 和 v_1^* 分别是置根生成森林 $f(F_0, v_0^*)$ 以及 $f(F_1, v_1^*)$ 中的根节点。因为 (F_0, v_0^*) 和 (F_1, v_1^*) 都是集合 \mathcal{S}_{uv}^1 中的元素, 所以根据集合 \mathcal{S}_{uv}^1 的定义可知, 节点 v_0^* 和 v_1^* 同属于森林 $F_0 = F_1$ 中节点 u 所在的连通分支, 即 $v_0^*, v_1^* \in T(F_0, u)$ 。由于映射 f 所对应的操作除删去无向边 (u, v) 外不改变置根生成森林的边集且节点 v_0^* 和节点 v_1^* 到节点 u 的路径中均包含无向边 (u, v) , 因此有 $v_0^*, v_1^* \in T(f(F_0, v_0^*), v_0^*)$ 以及 $v_0^*, v_1^* \in T(f(F_1, v_1^*), v_0^*)$ 。又因为置根生成森林中每个连通分支中有且仅有一个根节点, 所以在置根生成森林 $f(F_0, v_0^*)$ 中节点 v_1^* 不是根节点。又因为节点 v_1^* 在置根生成森林 $f(F_1, v_1^*)$ 中是根节点, 所以 $f(F_0, v_0^*) \neq f(F_1, v_1^*)$ 。
- ② 若 $F_0 \neq F_1$ 且 F_0 和 F_1 的边集合不相等, 则由映射 f 所对应的操作除删去无向边 (u, v) 外不改变置根生成森林的边集可知, 置根生成森林 $f(F_0, v_0^*)$ 和 $f(F_1, v_1^*)$ 的边集合也不相等。因此 $f(F_0, v_0^*) \neq f(F_1, v_1^*)$ 。
- ③ 若 $F_0 \neq F_1$ 且 F_0 和 F_1 的边集合相等, 则必然存在一个节点 v' 使得 $r_{F_0}(v') \neq r_{F_1}(v')$, 即节点 v' 在其中一个置根生成森林是根节点同时在另一个森林中不是根节点。不妨设 $r_{F_0}(v') = 1$, 则 $r_{F_1}(v') = 0$ 。假设 v' 属于森林 F_0 中节点 u 所在的连通分支, 即 $v' \in T(F_0, u)$ 。由 $(F_0, v_0^*) \in \mathcal{S}_{uv}^1$ 可知 $F_0 \in \mathcal{F}_{uv}$, 所以 $r_{F_0}(u) = 1$ 。又因为 $r_{F_0}(v') = 1$ 且 $v' \in T(F_0, u)$, 所以有 $u = v'$ 。再根据 $(F_1, v_1^*) \in \mathcal{S}_{uv}$ 可知 $r_{F_1}(u) = 1$ 。因此有 $r_{F_1}(v') = 1$, 与 $r_{F_1}(v') = 0$ 矛盾。所以, 节点 v' 必定不属于森林 F_0 中节点 u 所在的连通分支, 即 $v' \notin T(F_0, u)$ 。又因为 F_0 和 F_1 的边集合相等, 所以 $v' \notin T(F_1, u)$ 。由于映射 f 所对应的操作仅仅改变节点 u 所在的连通分支, 因此节点 v' 在森林 $f(F_0, v_0^*)$ 中是根节点, 在森林 $f(F_1, v_1^*)$ 中不是根节点。所以, $f(F_0, v_0^*) \neq f(F_1, v_1^*)$ 。

由 ① ② ③ 可知, 对于任意 $(F_0, v_0^*), (F_1, v_1^*) \in \mathcal{S}_{uv}^1$, 若 $(F_0, v_0^*) \neq (F_1, v_1^*)$, 则 $f(F_0, v_0^*) \neq f(F_1, v_1^*)$ 必定成立。所以, 映射 $f: \mathcal{S}_u \rightarrow \mathcal{F}$ 是一个单射。

接下来证明映射 $f: \mathcal{S}_{uv}^1 \rightarrow \mathcal{F}_{uv}^*$ 是一个满射。令 F' 集合 \mathcal{F}_{uv}^* 中任意一个元素, v_0^* 为置根生成森林 F' 中 v 点所在连通分支, 即 $T(F', v)$, 的根节点。只要将无向边 (u, v) 加入置根生成森林 F' 并且把 $T(F', u)$ 的根节点重新设置成节点 u , 就能得到一个新的置根生成森林 F_0 。根据映射 f 的构造方式可知, $f(F_0, v_0^*) = F'$ 。

又因为 $F' \in \mathcal{F}_{uv}^*$ 且 $v_0^* \in T(F', v)$, 所以有 $(F_0, v_0^*) \in \mathcal{S}_{uv}^1$ 。因此, 对于任意的 $F' \in \mathcal{F}_{uv}^*$, 都存在 $(F_0, v_0^*) \in \mathcal{S}_{uv}^1$ 使得 $f(F_0, v_0^*) = F'$ 。所以, 映射 $f: \mathcal{S}_{uv}^1 \rightarrow \mathcal{F}_{uv}^*$ 是一个满射。

综上所述, 映射 f 既是单射又是满射。所以, 映射 $f: \mathcal{S}_{uv}^1 \rightarrow \mathcal{F}_{uv}^*$ 是一个双射。因此, $|\mathcal{S}_{uv}^1| = |\mathcal{F}_{uv}^*|$, 得证。 \square

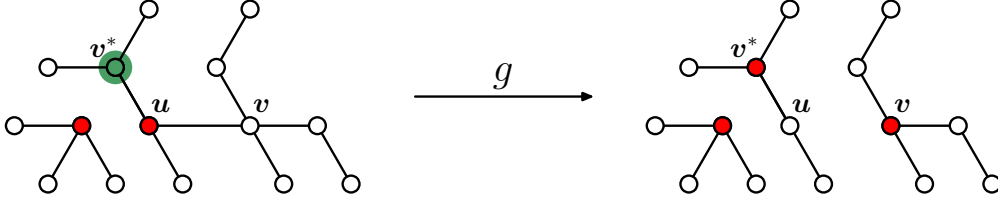


图 5.5: 映射 $g: \mathcal{S}_{uv}^0 \rightarrow \mathcal{F}_{vu}^*$ 的构造方式

引理 5.19 在无权无向图 $G = (V, E)$ 中, 令 \mathcal{F}_{vu}^* 为所有节点 v 是根节点且节点 u 和节点 v 不在同一连通分支的置根生成森林构成的集合, 即 $\mathcal{F}_{vu}^* = \mathcal{F}_{vv} \setminus \mathcal{F}_{vu}$, 则 $|\mathcal{S}_{uv}^0| = |\mathcal{F}_{vu}^*|$ 对于所有 $(u, v) \in E$ 成立。

证明: 本节通过构造映射 $g: \mathcal{S}_{uv}^0 \rightarrow \mathcal{F}_{vu}^*$ 的方式来证明集合 \mathcal{S}_{uv}^0 的与集合 \mathcal{F}_{vu}^* 大小相等。令 (F_0, v_0^*) 为集合 \mathcal{S}_{uv}^0 中任意一个元素, 则映射 g 可以理解为对置根生成森林 F_0 做一个操作, 即删去 F_0 中的无向边 (u, v) 且令节点 v^* 和节点 v 分别为 F_0 中节点 v^* 和节点 v 所在连通分支的根节点。这样就能得到新的置根生成森林 $g(F_0, v_0^*) \in \mathcal{F}_{vu}^*$ 。图 5.5 揭示了映射 g 的构造过程。

首先证明映射 $g: \mathcal{S}_{uv}^0 \rightarrow \mathcal{F}_{vu}^*$ 是一个单射。令 $(F_0, v_0^*) \neq (F_1, v_1^*)$ 为集合 \mathcal{S}_{uv}^0 中任意两个不相等的元素。接下来, 本节将讨论以下三种情况。

- ① 若 $F_0 = F_1$, 则由 $(F_0, v_0^*) \neq (F_1, v_1^*)$ 可知, 必有 $v_0^* \neq v_1^*$ 。根据映射 g 的构造方式可知: 节点 v_0^* 和 v_1^* 分别是置根生成森林 $g(F_0, v_0^*)$ 以及 $g(F_1, v_1^*)$ 中的根节点。因为 (F_0, v_0^*) 和 (F_1, v_1^*) 都是集合 \mathcal{S}_{uv}^0 中的元素, 所以根据集合 \mathcal{S}_{uv}^0 的定义可知, 节点 v_0^* 和 v_1^* 同属于森林 $F_0 = F_1$ 中节点 u 所在的连通分支, 即 $v_0^*, v_1^* \in T(F_0, u)$ 。由于映射 g 所对应的操作除删去无向边 (u, v) 外不改变置根生成森林的边集且节点 v_0^* 和节点 v_1^* 到节点 u 的路径中均不包含无向边 (u, v) , 因此有 $v_0^*, v_1^* \in T(g(F_0, v_0^*), u)$ 以及 $v_0^*, v_1^* \in T(g(F_1, v_1^*), u)$ 。又因为置根生成森林中每个连通分支中有且仅有一个根节点, 所以在置根生成森林 $g(F_0, v_0^*)$ 中节点 v_1^* 不是根节点。又因为节点 v_1^* 在置根生成森林 $g(F_1, v_1^*)$ 中是根节点, 所以 $g(F_0, v_0^*) \neq g(F_1, v_1^*)$ 。

- ② 若 $F_0 \neq F_1$ 且 F_0 和 F_1 的边集合不相等, 则由映射 g 所对应的操作删除去无向边 (u, v) 外不改变置根生成森林的边集可知, 置根生成森林 $g(F_0, v_0^*)$ 和 $g(F_1, v_1^*)$ 的边集合也不相等。因此 $g(F_0, v_0^*) \neq g(F_1, v_1^*)$ 。
- ③ 若 $F_0 \neq F_1$ 且 F_0 和 F_1 的边集合相等, 则必然存在一个节点 v' 使得 $r_{F_0}(v') \neq r_{F_1}(v')$, 即节点 v' 在其中一个置根生成森林是根节点同时在另一个森林中不是根节点。不妨设 $r_{F_0}(v') = 1$, 则 $r_{F_1}(v') = 0$ 。假设 v' 属于森林 F_0 中节点 u 所在的连通分支, 即 $v' \in T(F_0, u)$ 。由 $(F_0, v_0^*) \in \mathcal{S}_{uv}^0$ 可知 $F_0 \in \mathcal{F}_{uv}$, 所以 $r_{F_0}(u) = 1$ 。又因为 $r_{F_0}(v') = 1$ 且 $v' \in T(F_0, u)$, 所以有 $u = v'$ 。再根据 $(F_1, v_1^*) \in \mathcal{S}_{uv}$ 可知 $r_{F_1}(u) = 1$ 。因此有 $r_{F_1}(v') = 1$, 与 $r_{F_1}(v') = 0$ 矛盾。所以, 节点 v' 必定不属于森林 F_0 中节点 u 所在的连通分支, 即 $v' \notin T(F_0, u)$ 。又因为 F_0 和 F_1 的边集合相等, 所以 $v' \notin T(F_1, u)$ 。由于映射 g 所对应的操作仅仅改变节点 u 所在的连通分支, 因此节点 v' 在森林 $g(F_0, v_0^*)$ 中是根节点, 在森林 $g(F_1, v_1^*)$ 中不是根节点。所以, $g(F_0, v_0^*) \neq g(F_1, v_1^*)$ 。

由 ① ② ③ 可知, 对于任意 $(F_0, v_0^*), (F_1, v_1^*) \in \mathcal{S}_{uv}^0$, 若 $(F_0, v_0^*) \neq (F_1, v_1^*)$, 则 $g(F_0, v_0^*) \neq g(F_1, v_1^*)$ 必定成立。所以, 映射 $g: \mathcal{S}_u \rightarrow \mathcal{F}$ 是一个单射。

接下来证明映射 $g: \mathcal{S}_{uv}^0 \rightarrow \mathcal{F}_{vu}^*$ 是一个满射。令 F' 集合 \mathcal{F}_{vu}^* 中任意一个元素, v_0^* 为置根生成森林 F' 中 u 点所在连通分支, 即 $T(F', u)$, 的根节点。只要将无向边 (u, v) 加入置根生成森林 F' 并且把 $T(F', u)$ 的根节点重新设置成节点 u , 就能得到一个新的置根生成森林 F_0 。根据映射 g 的构造方式可知, $g(F_0, v_0^*) = F'$ 。又因为 $F' \in \mathcal{F}_{vu}^*$ 且 $v_0^* \in T(F', u)$, 所以有 $(F_0, v_0^*) \in \mathcal{S}_{uv}^0$ 。因此, 对于任意的 $F' \in \mathcal{F}_{vu}^*$, 都存在 $(F_0, v_0^*) \in \mathcal{S}_{uv}^0$ 使得 $g(F_0, v_0^*) = F'$ 。所以, 映射 $g: \mathcal{S}_{uv}^0 \rightarrow \mathcal{F}_{vu}^*$ 是一个满射。

综上所述, 映射 g 既是单射又是满射。所以, 映射 $g: \mathcal{S}_{uv}^0 \rightarrow \mathcal{F}_{vu}^*$ 是一个双射。因此, $|\mathcal{S}_{uv}^0| = |\mathcal{F}_{vu}^*|$, 得证。□

本节最后将证明森林边中心性的值可以用森林矩阵中的相关元素表示。

定理 5.20 在无权无向图 $G = (V, E)$ 中, 令 \mathbf{W} 为图 G 的森林矩阵, 且令 $f_{\text{FEC}}(u, v)$ 为无向边 (u, v) 的森林边中心性值, 则等式

$$f_{\text{FEC}}(u, v) = \frac{w_{uu} + w_{vv} - 2w_{uv}}{w_{uv}}$$

对于任意的 $(u, v) \in E$ 成立。

证明： 根据森林边中心性的定义，即定义 5.15，以及引理 5.16、引理 5.17、引理 5.18 和引理 5.19 的结论可知

$$\begin{aligned}
 f_{\text{FEC}}(u, v) &= \frac{1}{|\mathcal{F}_{uv}|} \sum_{F \in \mathcal{F}_{uv}} s(F, u, v) \\
 &= \frac{|\mathcal{S}_{uv}|}{|\mathcal{F}_{uv}|} = \frac{|\mathcal{S}_{uv}^0| + |\mathcal{S}_{uv}^1|}{|\mathcal{F}_{uv}|} = \frac{|\mathcal{F}_{uv}^*| + |\mathcal{F}_{vu}^*|}{|\mathcal{F}_{uv}|} \\
 &= \frac{|\mathcal{F}_{uu}| - |\mathcal{F}_{uv}| + |\mathcal{F}_{vv}| - |\mathcal{F}_{vu}|}{|\mathcal{F}_{uv}|}.
 \end{aligned}$$

又因为 $w_{uv} = |\mathcal{F}_{uv}| / |\mathcal{F}|$ ，所以

$$\begin{aligned}
 f_{\text{FEC}}(u, v) &= \frac{|\mathcal{F}_{uu}|}{|\mathcal{F}_{uv}|} - \frac{|\mathcal{F}_{uv}|}{|\mathcal{F}_{uv}|} + \frac{|\mathcal{F}_{vv}|}{|\mathcal{F}_{uv}|} - \frac{|\mathcal{F}_{vu}|}{|\mathcal{F}_{uv}|} \\
 &= \frac{|\mathcal{F}_{uu}|}{|\mathcal{F}|} \cdot \frac{|\mathcal{F}|}{|\mathcal{F}_{uv}|} - 1 + \frac{|\mathcal{F}_{vv}|}{|\mathcal{F}|} \cdot \frac{|\mathcal{F}|}{|\mathcal{F}_{uv}|} - 1 \\
 &= \frac{w_{uu} + w_{vv} - 2w_{uv}}{w_{uv}},
 \end{aligned}$$

得证。 □

定理 5.20 的结论指出，只要求出森林矩阵的所有元素，就可以计算出图中每条无向边的森林边中心性值。因此，在一个包含 n 个节点和 m 条无向边的无权无向图 G 中，仅需花费 $O(n^3)$ 的时间就可以求出所有无向边的森林边中心性值。

5.7 相关性质

本节将给出森林边中心性取值的上下界并证明森林边中心性满足边中心性的自同构等价公理。

5.7.1 上下界

本节首先研究森林边中心性取值的上下界。

引理 5.21 在无权无向图 $G = (V, E)$ 中，令 $\mathbf{L} = (l_{ij})_{n \times n}$ 为图 G 的拉普拉斯矩阵，且令 $\mathbf{W} = (w_{ij})_{n \times n}$ 为图 G 的森林矩阵，则对于任意的 $(u, v) \in E$ ，都有

$$\frac{w_{uu}}{w_{uv}} \leq d_v + 1,$$

其中 d_v 表示节点 v 的度。

证明: 根据森林矩阵的定义可知, $\mathbf{W}(\mathbf{I} + \mathbf{L}) = \mathbf{I}$ 。因此, 对于任意的 $(u, v) \in E$, 都有

$$\sum_{i=1}^n w_{ui} l_{iv} + w_{uv} = 0,$$

即

$$\sum_{i \neq u, v} w_{ui} l_{iv} + w_{uu} l_{uv} + w_{uv} (l_{vv} + 1) = 0.$$

又因为 $l_{vv} = d_v$, $l_{uv} = -1$, 以及拉普拉斯矩阵非对角线元素均小于等于 0, 所以

$$w_{uu} = \sum_{i \neq u, v} w_{ui} l_{iv} + w_{uv} (d_v + 1) \leq w_{uv} (d_v + 1),$$

得证。 □

根据引理 5.21 的结论, 本节可以结合森林边中心性的定义顺势推导出森林边中心性取值的上下界。

定理 5.22 在无权无向图 $G = (V, E)$ 中, 令 $f_{\text{FEC}}(u, v)$ 表示无向边 (u, v) 的森林边中心性值, 则对于任意的 $(u, v) \in E$ 均成立

$$2 \leq f_{\text{FEC}}(u, v) \leq d_u + d_v,$$

其中 d_u 和 d_v 分别表示节点 u 和 v 的度。

证明: 有研究指出 [108], 对于森林矩阵 \mathbf{W} 的任意非对角线元素 w_{ij} ($i \neq j$) 都有

$$w_{ii} \geq 2w_{ij}.$$

又因为森林矩阵 \mathbf{W} 是对称阵, 因此

$$\begin{aligned} f_{\text{FEC}}(u, v) &= \frac{w_{uu} + w_{vv} - 2w_{uv}}{w_{uv}} \\ &= \frac{w_{uu}}{w_{uv}} + \frac{w_{vv}}{w_{vu}} - 2 \\ &\geq 2 + 2 - 2 = 2. \end{aligned}$$

再根据引理 5.21 的结论可知

$$\begin{aligned}
 f_{\text{FC}}(u, v) &= \frac{w_{uu} + w_{vv} - 2w_{uv}}{w_{uv}} \\
 &= \frac{w_{uu}}{w_{uv}} + \frac{w_{vv}}{w_{vu}} - 2 \\
 &\leq (d_v + 1) + (d_u + 1) - 2 \\
 &\leq d_u + d_v,
 \end{aligned}$$

得证。 \square

根据定理 5.22 给出的上下界可以看出，一条无向边的森林边中心性的值至少为 2 至多为这条边两个端点的度之和。

5.7.2 自同构等价公理

本节接着证明森林边中心性满足边中心性的自同构等价公理，即公理 5.7。

定理 5.23 在无权无向简单图 $G = (V, E)$ 中，若无向边 (u_i, v_i) 和 (u_j, v_j) 是自同构等价的，记为 $(u_i, v_i) \cong (u_j, v_j)$ ，则它们的森林边中心性值 $f_{\text{FEC}}(u_i, v_i)$ 和 $f_{\text{FEC}}(u_j, v_j)$ 相等，即 $f_{\text{FEC}}(u_i, v_i) = f_{\text{FEC}}(u_j, v_j)$ 。

证明： 已知 $(u_i, v_i) \cong (u_j, v_j)$ 。根据边自同构等价的定义，即定义 5.5，不妨设存在图 G 的自同构 f 使得 $f(u_i) = u_j$ 且 $f(v_i) = v_j$ 。令 \mathbf{P} 为自同构 f 所对应的排列矩阵，则根据排列矩阵的定义可知 $\mathbf{e}_{u_i}^\top \mathbf{P} = \mathbf{e}_{u_j}^\top$ 以及 $\mathbf{e}_{v_i}^\top \mathbf{P} = \mathbf{e}_{v_j}^\top$ 。因此，有

$$\begin{aligned}
 &w_{u_j u_j} + w_{v_j v_j} - 2w_{u_j v_j} \\
 &= (\mathbf{e}_{u_j} - \mathbf{e}_{v_j})^\top (\mathbf{I} + \mathbf{L})^{-1} (\mathbf{e}_{u_j} - \mathbf{e}_{v_j}) \\
 &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top \mathbf{P} (\mathbf{I} + \mathbf{L})^{-1} \mathbf{P}^\top (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\
 &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top (\mathbf{P} \mathbf{P}^\top + \mathbf{P} \mathbf{L} \mathbf{P}^\top)^{-1} (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\
 &= (\mathbf{e}_{u_i} - \mathbf{e}_{v_i})^\top (\mathbf{I} + \mathbf{L})^{-1} (\mathbf{e}_{u_i} - \mathbf{e}_{v_i}) \\
 &= w_{u_i u_i} + w_{v_i v_i} - 2w_{u_i v_i}.
 \end{aligned}$$

此外，

$$\begin{aligned}
 w_{u_j v_j} &= \mathbf{e}_{u_j}^\top (\mathbf{I} + \mathbf{L})^{-1} \mathbf{e}_{v_j} \\
 &= \mathbf{e}_{u_i}^\top \mathbf{P} (\mathbf{I} + \mathbf{L})^{-1} \mathbf{P}^\top \mathbf{e}_{v_i}
 \end{aligned}$$

$$\begin{aligned}
&= \mathbf{e}_{u_i}^\top (\mathbf{I} + \mathbf{L})^{-1} \mathbf{e}_{v_i} \\
&= w_{u_i v_i}.
\end{aligned}$$

根据森林边中心性的定义，即定义 5.15，可知

$$\begin{aligned}
f_{\text{FEC}}(u_j, v_j) &= \frac{w_{u_j u_j} + w_{v_j v_j} - 2w_{u_j v_j}}{w_{u_j v_j}} \\
&= \frac{w_{u_i u_i} + w_{v_i v_i} - 2w_{u_i v_i}}{w_{u_i v_i}} \\
&= f_{\text{FEC}}(u_i, v_i),
\end{aligned}$$

得证。 □

定理 5.23 指出，森林边中心性必定会给自同构等价的无向边赋予相同的中心性值，即森林边中心性满足边中心性的自同构等价公理。这个结论为本章接下来量化研究森林边中心性的区分能力奠定了重要的理论基础。

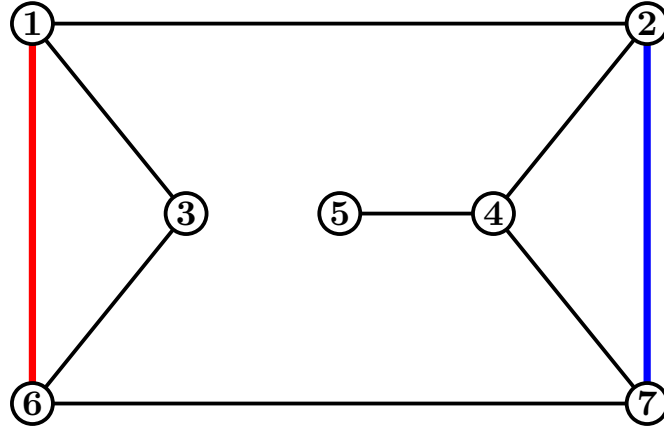


图 5.6: 一个包含 7 个节点和 9 条边的示例网络

5.8 实验

本节将分别在示例网络、现实网络以及模型网络上比较森林边中心性（Forest Edge Centrality, FEC）与其它三种常用边中心性度量，即边介数中心性（EB）、生成边中心性（SEC）和双调和距离中心性（BDRC），的区分能力。

5.8.1 示例网络

本节首先在一个小型的示例网络上测试森林边中心性的区分能力。如图 5.6 所示，示例网络包含 7 个节点和 9 条无向边。该网络仅有一个非平凡自同构，如

表 5.3: 图 5.6 所示网络的唯一非平凡自同构

u	(u, v)	$f(u)$	$(f(u), f(v))$
1	(1,2) (1,3) (1,6)	6	(6,7) (6,3) (6,1)
2	(2,1) (2,4) (2,7)	7	(7,6) (7,4) (7,2)
3	(3,1) (3,6)	3	(3,6) (3,1)
4	(4,2) (4,5) (4,7)	4	(4,7) (4,5) (4,2)
5	(5,4)	5	(5,4)
6	(6,1) (6,3) (6,7)	1	(1,6) (1,3) (1,2)
7	(7,2) (7,4) (7,6)	2	(2,7) (2,4) (2,1)

表 5.3，则根据边自同构等价的定义，即定义 5.5，可以推断出除自反等价关系外的三组自同构等价的无向边，分别是：(1,2) \cong (6,7)、(1,3) \cong (6,3) 以及 (4,2) \cong (4,7)。

表 5.4: 图 5.6 所示网络中所有边的中心性值.

边中心性度量	图 5.6 所示网络中边的中心性值					
	(1,2) (6,7)	(3,1) (3,6)	(4,2) (4,7)	(1,6)	(2,7)	(4,5)
EB	12	6	10	4	4	12
SEC	0.7	0.63	0.63	0.53	0.53	1
BDRC	0.47	0.25	0.39	0.15	0.15	0.86
FEC	3.61	2.69	3.05	2.67	2.86	3.22

表 5.4 列出了该示例网络中所有无向边的边中心性值，其中表格第一列中的 EB、SEC、BDRC 以及 FEC 分别表示边介数中心性、生成边中心性、双调和距离中心性以及森林边中心性。从表 5.4 所示的结果可以发现，(1,6) 和 (2,7) 两条无向边并不等价，但其它三种边中心性度量，即边介数中心性、生成边中心性和双调和距离中心性都给出了相同的中心性值，唯有森林边中心性给出了不同的中心性值。因此，示例网络上的结果表明，森林边中心性拥有比其它三种边中心性度量更好的区分能力。

5.8.2 现实网络

本节接着在 20 个不同的现实网络中比较森林边中心性与其他三个常用边中心性度量的区分能力。这些现实网络都来自 SNAP¹，且所有现实网络均为无权无向的简单图。由于生成边中心性以及双调和距离中心性在不连通的网络上

¹<https://snap.stanford.edu>

没有明确的定义, 因此, 本节中所有的实验将在每个现实网络的最大连通分支 (Largest Connected Component, LCC) 上运行。所有现实网络数据的相关统计信息如表 5.5 所示, 其中 n 和 m 分别表示网络包含的节点数和边数; n' 和 m' 分别表示网络的最大连通分支包含的节点数和边数。

表 5.5: 现实网络数据的统计信息

Network	n	m	n'	m'
Karate	34	78	34	78
Dolphins	62	159	62	159
Lesmis	77	254	77	254
Adjnoun	112	425	112	425
Celegansneural	297	2148	297	2148
Chicago	1467	1298	823	822
HamsterFull	2426	16631	2000	16098
GrQc	5241	14484	4158	13422
PowerGrid	4941	6594	4941	6594
HighEnergy	7610	15751	5835	13815
Reactome	6229	146160	5973	145778
RouteViews	6474	12572	6474	12572
HepTh	9875	25973	8638	24806
PrettyGoodPrivacy	10680	24316	10680	24316
HepPh	12006	118489	11204	117619
AstroPh	18771	198050	17903	196972
Internet	22963	48436	22963	48436
CAIDA	26475	53381	26475	53381
EnronEmail	36692	183831	33696	180811
CondensedMatter	39577	175692	36458	171735

由于森林边中心性以及其它三个边中心性度量都满足边中心性的自同构等价公理, 即公理 5.7。因此, 根据定理 5.13 的结论, 本节可以通过比较它们边中心性的广义区分度, 即 P_c 值, 来比较它们的区分能力。在一个包含 m 条无向边的网络 $G = (V, E)$ 中, 边中心性度量 c 的广义区分度, 即 P_c 值, 以如下方式计算。

$$P_c = \frac{|\{(\epsilon_i, \epsilon_j) \mid f_c(\epsilon_i) \neq f_c(\epsilon_j), \epsilon_i, \epsilon_j \in E, \epsilon_i \neq \epsilon_j\}|}{m(m-1)}.$$

在本节剩余的实验中, 若两个不同无向边的中心性值的差距在 10^{-9} 以内, 则视为相等。

表 5.6 列出了这三个边中心性度量在这 20 个无权无向的现实网络中的广义区分度, 即 P_c 值, 表格第一行中的 EB、BDRC、SEC 和 FEC 分别表示边介数

表 5.6: 边中心性度量在现实网络上的广义区分度

Network	P_c for different metrics ($\times 10^{-2}$)			
	P_{EB}	P_{BDRC}	P_{SEC}	P_{FEC}
Karate	99.1009	99.1342	98.0686	99.1342
Dolphins	99.7134	99.7134	99.6497	99.9841
Lesmis	97.9117	98.8360	98.5248	99.1628
Adjnoun	99.9501	99.9501	99.9367	99.9989
Celegansneural	99.9954	99.9954	99.9953	99.9978
Chicago	25.2158	25.2158	0.0000	95.0544
HamsterFull	99.8155	99.9763	99.9702	99.9910
GrQc	96.7400	99.3732	99.0569	99.7526
PowerGrid	96.4240	96.4399	93.9758	99.9964
HighEnergy	99.0778	99.3576	98.8529	99.9877
Reactome	99.4433	99.8644	99.8656	99.8703
RouteViews	96.3630	96.3630	96.1153	99.8909
HepTh	99.1677	99.5526	99.3418	99.9671
PrettyGoodPrivacy	96.7787	96.8726	94.7064	99.9920
HepPh	99.0290	99.9417	99.9358	99.9638
AstroPh	99.7930	99.9903	99.9885	99.9969
Internet	97.3593	97.3593	97.2322	99.9607
CAIDA	96.5176	96.5176	96.3283	99.9625
EnronEmail	99.6371	99.7035	99.6713	99.9940
CondensedMatter	99.5623	99.9305	99.8952	99.9992

中心性、双调和距离中心性、生成边中心性以及森林边中心性，每个现实网络中最高的 P_c 值用粗体字标出。从表 3.5 所示的实验结果可以看出：①边介数中心性在 7 个现实网络中取得了最高的 P_c 值，同时它在其它现实网络中的广义区分度也与最高值相差无几，仅仅在 GrQc 这个网络中的广义区分度与最高值有明显差距，因此边介数中心性的区分能力很强且基本上比较稳定；②双调和距离中心性在几乎所有的现实网络中都取得了最高的 P_c 值，且在唯一未取得最高值的网络 Reactome 中的 P_c 值也与最高值相差不大，总的来说，双调和距离中心性的区分能力很强且非常稳定；③生成边中心性仅仅在 Reactome 这个网络中取得了最高的 P_c 值，并且，它在多个现实网络，如 Chicago、PowerGrid 以及 PrettyGoodPrivacy，中的表现并不理想。值得注意的是，生成边中心性表现不好的网络往往比较稀疏，例如：Chicago 这个网络的 LCC 的节点平均度是 1.998；网络 PowerGrid 的 LCC 的节点平均度是 2.669；网络 PrettyGoodPrivacy 的 LCC 的节点平均度是 4.554。同时，生成边中心性表现非常优秀的网络往往比较稠密，

如网络 Reactome 的 LCC 的节点平均度高达 48.812 远高于那三个生成边中心性表现差的现实网络。因此,生成边中心性在稠密网络中区分能力很强,但在稀疏网络中区分能力很差。此外,从上述实验结果可以看出,森林边中心性在每个现实网络中的广义区分度都是最大的。此外,由于网络 Chicago 的最大连通分支是一个树,所以网络中每一条边在生成树中的概率都是 1,因此生成边中心性在网络 Chicago 上的广义区分度是 0。总之,现实网络上的实验结果表明,本章提出的新的边中心性度量在区分能力方面相较于其它几个常用的边中心性度量有着明显的优势。

5.8.3 模型网络

本节接着在三种模型网络,即 Erdős-Rényi (ER) 模型网络 [89]、Barbási-Albert (BA) 模型网络 [90] 以及 Watts-Strogatz (WS) 模型网络 [109], 上比较森林边中心性和其它三个常用边中心性度量的区分能力。

表 5.7: 边中心性度量在三种模型网络上的广义区分度

Network	P_c for different metrics ($\times 10^{-2}$)			
	P_{EB}	P_{BDRC}	P_{SEC}	P_{FEC}
ER[p=0.0004]	94.3842	94.3842	88.5171	99.1342
ER[p=0.0008]	99.0878	99.0875	98.7007	99.3867
ER[p=0.0012]	99.9814	99.9803	99.9793	99.9913
ER[p=0.0016]	99.9985	99.9957	99.9987	99.9987
BA[m=1]	53.0480	53.0480	0.0000	95.0384
BA[m=2]	99.9961	99.9932	99.9704	99.9976
BA[m=3]	99.9973	99.9963	99.9985	99.9987
BA[m=4]	99.9968	99.9976	99.9993	99.9993
WS[K=2,p=0.15]	92.4537	92.4537	84.4173	97.3648
WS[K=4,p=0.15]	96.2944	96.2873	96.0239	99.4133
WS[K=6,p=0.15]	99.9211	99.9324	99.9318	99.9938
WS[K=8,p=0.15]	99.9925	99.9946	99.9976	99.9976

这三种模型网络的生成方式分别如下: 对于 ER 模型网络, 每组实验共生成 100 个 ER 模型网络, 每个生成的 ER 网络含 10000 个节点, 节点间的连接概率 p 依次取 $p = 0.0004$ 、 $p = 0.0008$ 、 $p = 0.0012$ 以及 $p = 0.0016$ 。由于生成出的 ER 网络可能是不连通的, 因此将提取每个 ER 网络的最大连通分支进行实验。每组实验记录这 100 个 ER 网络的平均 P_c 值; 对于 BA 模型网络, 每组实验共生成 100 个 BA 模型网络, 每个生成的 BA 网络含 10000 个节点, 每次添加的边数 m

依次取 $m = 1$ 、 $m = 2$ 、 $m = 3$ 和 $m = 4$ 。每组实验记录这 100 个 BA 网络的平均 P_c 值；对于 WS 模型网络，每组实验共生成 100 个 WS 模型网络，每个生成的 WS 网络含 10000 个节点，网络的初始连接数 K 依次取 $K = 2$ 、 $K = 4$ 、 $K = 6$ 和 $K = 8$ ，WS 模型的边置换概率 p 设置为 $p = 0.15$ 。每组实验记录这 100 个 WS 网络的平均 P_c 值。

模型网络的实验结果如表 5.7 所示，表格第一行中的 EB、BDRC、SEC 和 FEC 分别表示边介数中心性、双调和距离中心性、生成边中心性以及森林边中心性。从表 5.7 中的实验结果可以看出：①边介数中心性和双调和距离中心性的表现相差不大，它们在所有的模型网络中的 P_c 值都十分接近最高值，由此看来，这两个边中心性度量的区分能力非常优秀且很稳定。②生成边中心性的区分能力在稀疏图中很弱，在稠密图中很强。这个现象与现实网络实验中观察到的结果基本一致。此外，从表 5.7 所列出的实验结果可以看出，森林边中心性在每组试验中的广义区分度都是最大的。模型网络上的实验结果进一步说明了森林边中心性在区分能力方面相较于其它几个常用的边中心性度量有着明显的优势。

5.9 本章小结

本章基于置根生成森林的概念提出了一个新的边中心性度量，即森林边中心性。本章首先通过评估无向边在置根生成森林中的重要程度给出了森林边中心性的定义。随后，本章又证明了任意一条无向边的森林边中心性值都可以用森林矩阵中的对应元素来表示，大大降低了森林边中心性的计算成本。本章接着分析了森林边中心性取值的上下界并且证明了森林边中心性满足边中心性度量的自同构等价公理。最后，本章分别在示例网络、现实网络以及模型网络上测试了森林边中心性以及其它三种常用边中心性度量的区分能力。实验结果表明，本章提出的森林边中心性在区分能力方面相较于其它几个常用的边中心性度量有着明显的优势。

第六章 总结与展望

6.1 全文总结

现实世界中的许多事物都可以用相互关联的实体组成的系统来描述。复杂网络作为一种描述现实系统的强大工具近年来在各个领域的研究中展现出了巨大的潜力和重要性。寻找关键点或关键边的问题作为网络科学领域的一个基本问题有着悠久的历史历史和广泛的应用。为了评估网络中节点和边的重要性，学者们提出了多种衡量点和边相对重要性的方法，即中心性度量。这些中心性度量各具特点，它们衡量网络中点和边重要性的方式也各不相同。因此，研究这些中心性度量的特性是十分必要的。为了更好的了解这些中心性度量的特征，学者们在这方面进行了大量的研究，并取得了一些重要的研究成果。其中，有的学者提出了中心性公理来定性的描述中心性度量的特征，还有的学者深入研究了中心性度量的敏感性并提出了相关的量化指标。然而，到目前为止，学者们对中心性度量的区分能力还没有进行深入研究。为了更好的理解中心性度量的区分能力，本文深入的研究了点和边中心性度量的区分能力，同时提出了区分能力更强的节点中心性度量和边中心性度量，具体内容如下：

基于自同构等价的概念提出了节点中心性度量区分能力的评价标准。本文深入的研究了节点中心性度量的区分能力并基于节点自同构等价的概念提出了节点中心性度量区分能力的量化评价指标。根据等价节点中心性值应当相等的原则，本文基于节点自同构等价的概念提出了相应的自同构等价公理，规定了节点中心性度量对于自同构等价的节点必须给出相等的中心性值。本文证明了几个常用的中心性度量均满足本文提出的自同构等价公理。本文再根据不等价节点的中心性值应当不同的原则，用节点中心性度量对于不等价的节点给出不同中心性值的概率定义了区分能力指数的概念。在满足自同构等价公理时，一个中心性度量的区分能力越强，它的区分能力指数就越大。同时，本文还提出了广义区分度的概念用以快速比较两种节点中心性度量的区分能力。本文在多个现实网络和模型网络上用上述方法深入的分析了几种常用的节点心性度量的区分能力。此外，本文还把上述分析节点中心性度量区分能力的方法扩展到了有向加权网络。

基于置根生成森林提出了森林中心性并提出了相应的快速近似算法。为探究是否存在区分能力更为出色的节点中心性度量，本文基于置根生成森林的概念

提出了森林中心性这一新的节点中心性度量。森林中心性基于以某个节点为根的所有置根生成森林中这个点所在连通分量的平均大小来定义一个节点的相对重要程度。相较于其它常用的节点中心性度量,森林中心性不仅可以作用于不连通的网络,它还有着更为出色的区分能力。此外,本文还基于 JL 引理以及快速求解器设计了一个用于计算森林中心性值的快速近似算法。该算法能够用几乎线性的时间和空间复杂度给出一个网络中所有节点的森林中心性的准确近似值。本文在多个现实网络和模型网络上比较了森林中心性和其他几个常用的节点中心性度量的区分能力。实验结果表明,相较于其他几个常用的节点中心性度量,森林中心性确实有着更为强大的区分能力。此外,本文在包含数百万节点的现实网络中验证了快速近似算法的运算效率和准确性。

提出了节点中心性度量区分能力的评价标准并提出了森林边中心性。本文还深入的研究了边中心性度量的区分能力并基于自同构的概念提出了边中心性度量区分能力的量化评价指标。本文首先基于自同构的概念给出了边自同构等价的定义。再根据等价边的中心性值应当相等的原则,本文基于边的自同构等价提出了相应的自同构等价公理,规定了边中心性度量对于自同构等价的边须给出相等的边中心性值。本文证明了几个常用的边中心性度量均满足边中心性的自同构等价公理。本文接着根据不等价边的中心性值应当不同的原则,用边中心性度量对于不等价边给出不同中心性值的概率定义了边中心性的区分能力指数。在满足自同构等价公理时,一个边中心性度量的区分能力越强,它的区分能力指数就越大。同时,本文还提出了广义区分度的概念用以快速比较两种不同边中心性度量的区分能力。为得到区分能力更强的边中心性,本文还基于置根生成森林的概念提出了森林边中心性这一新的边中心性度量。本文在多个现实网络和模型网络上用上述方法深入的分析了几种常用的边中心性度量以及新提出的森林边中心性的区分能力。实验结果表明,森林边中心性明显拥有更加优秀的区分能力。

6.2 未来工作展望

本文系统的研究了点和边中心性度量的区分能力,并分别提出了区分能力更强的点中心性度量以及边中心性度量。本节主要展望一些未来可能的研究方向。

系统性的研究中心性度量的其它特征: 本文系统性的研究了中心性度量的区分能力并提出了相应的量化描述指标,即区分能力指数。然而,中心性度量还有许多其他重要的特征没有得到系统性的研究,如中心性度量的单调性、不同中心

性度量之间的相关性等。实际上,到目前为止,学者们对中心性公理的研究还远远没有接近尾声。近几年中,不断有新的中心性公理被提出。因此,若学者们能对中心性度量的其它特征展开系统性的研究,将很有可能取得不错的进展。

中心性的更多实际应用:相较于机器学习而言,与中心性度量相关的研究工作目前并不多见。其中一个重要的因素就是中心性度量的应用相较于机器学习而言还不够广泛。幸运的是,近几年中中心性度量在深度学习等研究领域得到了进一步的运用。如果能将节点或者边的中心性度量用到更多的地方去,那么中心性度量成为网络科学领域的热门研究方向将指日可待。

提出新的中心性度量:本文基于置根生成森林提出了一个新的节点中心性度量和一个新的边中心性度量。实际上,近几年中不断有学者提出新的中心性度量以更好的解决实际问题。然而,系统性的研究中心性度量的更多的其它特征以及发现中心性的更多实际应用才是治本之策。只要做好前文中提到的这两点,那么,学者们自然能够提出更多新的更好的中心性度量。

提出计算现有中心性度量的快速算法:本文提出了一种用于计算森林中心性的快速近似算法。目前,仍然存在许多特性良好但计算复杂度高的中心性度量。因此,提出计算现有的这些中心性度量的快速算法将是一项有趣而富有挑战性的工作。

参考文献

- [1] Eyinagho M O, Falaki S O. Appropriate buffer sizes for internet nodal devices: A networks' topologies-based solution[J]. International Journal of Communication Systems, 2020, 33(8): e4359.
- [2] Akhmadieva R S, Guryanova T Y, Kurakin A V, et al. Student attitude to intercultural communication and intercultural interaction in social networks[J]. Contemporary Educational Technology, 2020, 11(1): 21-29.
- [3] Nagpal S, Baksi K D, Kuntal B K, et al. Netconfer: a web application for comparative analysis of multiple biological networks[J]. BMC Biology, 2020, 18: 1-12.
- [4] Koschützki D, Lehmann K A, Peeters L, et al. Centrality indices[M]//Network Analysis. Springer, 2005: 16-61.
- [5] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167-256.
- [6] Newman M E J. Networks: An introduction[M]. Oxford University Press, 2010.
- [7] Grando F, Granville L Z, Lamb L C. Machine learning in network centrality measures: Tutorial and outlook[J]. ACM Computing Surveys, 2018, 51(5): 102.
- [8] Skibski O, Michalak T P, Rahwan T. Axiomatic characterization of game-theoretic centrality[J]. Journal of Artificial Intelligence Research, 2018, 62: 33-68.
- [9] Skibski O, Rahwan T, Michalak T P, et al. Attachment centrality: Measure for connectivity in networks[J]. Artificial Intelligence, 2019, 274: 151-179.
- [10] Dreyfus S E. An appraisal of some shortest-path algorithms[J]. Operations research, 1969, 17(3): 395-412.
- [11] Golden B. Shortest-path algorithms: A comparison[J]. Operations Research, 1976, 24(6): 1164-1168.
- [12] Gallo G, Pallottino S. Shortest path algorithms[J]. Operations Research, 1988, 13(1): 1-79.
- [13] Fu L, Sun D, Rilett L R. Heuristic shortest path algorithms for transportation applications: State of the art[J]. Computers and Operations Research, 2006, 33 (11): 3324-3343.
- [14] Magzhan K, Jani H M. A review and evaluations of shortest path algorithms[J]. International Journal of Scientific and Technology Research, 2013, 2(6): 99-104.

- [15] Yan S, Xu D, Zhang B, et al. Graph embedding and extensions: A general framework for dimensionality reduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 29(1): 40-51.
- [16] Luo D, Nie F, Huang H, et al. Cauchy graph embedding[C]//Proceedings of the 28th International Conference on Machine Learning. 2011: 553-560.
- [17] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 701-710.
- [18] Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.
- [19] Ou M, Cui P, Pei J, et al. Asymmetric transitivity preserving graph embedding[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1105-1114.
- [20] Ribeiro L F, Saverese P H, Figueiredo D R. struc2vec: Learning node representations from structural identity[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 385-394.
- [21] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.
- [22] Cai H, Zheng V W, Chang K C C. A comprehensive survey of graph embedding: Problems, techniques, and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616-1637.
- [23] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey[J]. Knowledge-Based Systems, 2018, 151: 78-94.
- [24] Xu M. Understanding graph embedding methods and their applications[J]. SIAM Review, 2021, 63(4): 825-853.
- [25] Wang X, Bo D, Shi C, et al. A survey on heterogeneous graph embedding: methods, techniques, applications and sources[J]. IEEE Transactions on Big Data, 2022, 9(2): 415-436.
- [26] Strogatz S H. Exploring complex networks[J]. Nature, 2001, 410(6825): 268-276.
- [27] Albert R, Barabási A L. Statistical mechanics of complex networks[J]. Reviews of Modern Physics, 2002, 74(1): 47.
- [28] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics[J]. Physics Reports, 2006, 424(4): 175-308.

- [29] Börner K, Sanyal S, Vespignani A. Network science[J]. Annual Review of Information Science and Technology, 2007, 41(1): 537-607.
- [30] Newman M. Networks[M]. Oxford University Press, 2018.
- [31] Chen G, Wang X, Li X. Fundamentals of complex networks: models, structures and dynamics[M]. John Wiley & Sons, 2014.
- [32] Doostmohammadian M, Rabiee H R, Khan U A. Centrality-based epidemic control in complex social networks.[J]. Social Network Analysis and Mining, 2020, 10(1): 32.
- [33] Park M J, Lee S H, Kwon O M, et al. Betweenness centrality-based consensus protocol for second-order multiagent systems with sampled-data[J]. IEEE Transactions on Cybernetics, 2017, 47(8): 2067-2078.
- [34] Park M, Lee S H, Kwon O M, et al. Closeness-centrality-based synchronization criteria for complex dynamical networks with interval time-varying coupling delays[J]. IEEE Transactions on Cybernetics, 2018, 48(7): 2192-2202.
- [35] Cantero S V A B, Gonçalves D N, dos Santos Scabini L F, et al. Importance of vertices in complex networks applied to texture analysis[J]. IEEE Transactions on Cybernetics, 2020, 50(2): 777-786.
- [36] Padgett J F, Ansell C K. Robust action and the rise of the medici, 1400-1434[J]. American Journal of Sociology, 1993, 98(6): 1259-1319.
- [37] Jackson M O. Social and economic networks[M]. Princeton University Press, 2008.
- [38] Jackson M O. The human network: How your social position determines your power, beliefs, and behaviors[M]. Vintage, 2019.
- [39] Ballester C, Calvó-Armengol A, Zenou Y. Who's who in networks. wanted: The key player[J]. Econometrica, 2006, 74(5): 1403-1417.
- [40] Banerjee A, Chandrasekhar A G, Duflo E, et al. The diffusion of microfinance [J]. Science, 2013, 341(6144): 1236498.
- [41] Banerjee A, Chandrasekhar A G, Duflo E, et al. Using gossips to spread information: Theory and evidence from two randomized controlled trials[J]. The Review of Economic Studies, 2019, 86(6): 2453-2490.
- [42] Golub B, Jackson M O. Naive learning in social networks and the wisdom of crowds[J]. American Economic Journal, 2010, 2(1): 112-149.
- [43] Jackson M O. The friendship paradox and systematic biases in perceptions and social norms[J]. Journal of Political Economy, 2019, 127(2): 777-818.
- [44] Garg M. Axiomatic foundations of centrality in networks[J]. SSRN Electronic Journal, 2009.

- [45] Boldi P, Vigna S. Axioms for centrality[J]. Internet Mathematics, 2014, 10(3-4): 222-262.
- [46] Kitti M. Axioms for centrality scoring with principal eigenvectors[J]. Social Choice and Welfare, 2016, 46: 639-653.
- [47] Boldi P, Luongo A, Vigna S. Rank monotonicity in centrality measures[J]. Network Science, 2017, 5(4): 529-550.
- [48] Wąs T, Skibski O. Axiomatization of the pagerank centrality[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 3898-3904.
- [49] Bloch F, Jackson M O, Tebaldi P. Centrality measures in networks[J]. Social Choice and Welfare, 2023: 1-41.
- [50] Sosnowska J, Skibski O. Attachment centrality for weighted graphs.[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017: 416-422.
- [51] Schoch D, Brandes U. Re-conceptualizing centrality in social networks[J]. European Journal of Applied Mathematics, 2016, 27(6): 971-985.
- [52] de Andrade R L, Rêgo L C. p -means centrality[J]. Communications in Nonlinear Science and Numerical Simulation, 2019, 68: 41-55.
- [53] Wąs T, Rahwan T, Skibski O. Random walk decay centrality[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 33. 2019: 2197-2204.
- [54] Sadler E. Ordinal centrality[J]. Journal of Political Economy, 2022, 130(4): 926-955.
- [55] Skibski O. Closeness centrality via the condorcet principle[J]. Social Networks, 2023, 74: 13-18.
- [56] Borgatti S P, Carley K M, Krackhardt D. On the robustness of centrality measures under conditions of imperfect data[J]. Soc. Netw., 2006, 28(2): 124-136.
- [57] Frantz T L, Cataldo M, Carley K M. Robustness of centrality measures under uncertainty: Examining the role of network topology[J]. Comput. Math. Organ. Theory, 2009, 15(4): 303.
- [58] Costenbader E, Valente T W. The stability of centrality measures when networks are sampled[J]. Soc. Netw., 2003, 25(4): 283-307.
- [59] Ghoshal G, Barabási A L. Ranking stability and super-stable nodes in complex networks[J]. Nat. Commun., 2011, 2(1): 1-7.
- [60] Platig J, Ott E, Girvan M. Robustness of network measures to link errors[J]. Phys. Rev. E., 2013, 88(6): 062812.

- [61] Tsugawa S, Ohsaki H. Analysis of the robustness of degree centrality against random errors in graphs[M]//Complex Networks VI. Springer, 2015: 25-36.
- [62] Segarra S, Ribeiro A. Stability and continuity of centrality measures in weighted graphs[J]. IEEE Tran. Signal Process., 2015, 64(3): 543-555.
- [63] Murai S, Yoshida Y. Sensitivity analysis of centralities on unweighted networks [C]//Proceedings of the 2019 World Wide Web Conference. 2019: 1332-1342.
- [64] Li H, Zhang Z. Kirchhoff index as a measure of edge centrality in weighted networks: Nearly linear time algorithms[C]//Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM, 2018: 2377-2396.
- [65] Yi Y, Shan L, Li H, et al. Biharmonic distance related centrality for edges in weighted networks[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 3620-3626.
- [66] Jin Y, Bao Q, Zhang Z. Forest distance closeness centrality in disconnected graphs[C]//2019 IEEE International Conference on Data Mining. IEEE, 2019: 339-348.
- [67] Kirchhoff G. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird[J]. Annalen der Physik, 1847, 148(12): 497-508.
- [68] Klein D J, Randić M. Resistance distance[J]. Journal of Mathematical Chemistry, 1993, 12(1): 81-95.
- [69] Tizghadam A, Leon-Garcia A. Autonomic traffic engineering for network robustness[J]. IEEE Journal on Selected Areas in Communications, 2009, 28(1): 39-50.
- [70] Patterson S, Bamieh B. Consensus and coherence in fractal networks[J]. IEEE Transactions on Control of Network Systems, 2014, 1(4): 338-348.
- [71] Shaw M E. Group structure and the behavior of individuals in small groups[J]. J. Psychol., 1954, 38(1): 139-149.
- [72] Freeman L C. Centrality in social networks conceptual clarification[J]. Social Networks, 1978, 1(3): 215-239.
- [73] Brandes U. A faster algorithm for betweenness centrality[J]. J. Math. Sociol., 2001, 25(2): 163-177.
- [74] Bavelas A. A mathematical model for group structures[J]. Applied Anthropology, 1948, 7(3): 16-30.
- [75] Sabidussi G. The centrality index of a graph[J]. Psychometrika, 1966, 31(4): 581-603.
- [76] Gleich D F. Pagerank beyond the web[J]. SIAM Rev., 2015, 57(3): 321-363.

- [77] Bonacich P. Factoring and weighting approaches to status scores and clique identification[J]. J. Math. Sociol., 1972, 2(1): 113-120.
- [78] Fletcher J M, Wennekers T. From structure to activity: using centrality measures to predict neuronal activity[J]. Int. J. Neural Syst., 2018, 28(02): 1750013.
- [79] Stephenson K, Zelen M. Rethinking centrality: Methods and examples[J]. Soc. Netw., 1989, 11(1): 1-37.
- [80] Brandes U, Fleischer D. Centrality measures based on current flow[C]//Annual symposium on theoretical aspects of computer science. Springer, 2005: 533-544.
- [81] Lorrain F, White H C. Structural equivalence of individuals in social networks [J]. J. Math. Sociol., 1971, 1(1): 49-80.
- [82] Hanneman R A, Riddle M. Concepts and measures for basic network analysis[J]. The SAGE handbook of social network analysis, 2011: 340-369.
- [83] Biggs N. Algebraic graph theory: Vol. 67[M]. Cambridge University Press, 1993.
- [84] Godsil C, Royle G F. Algebraic graph theory[M]. Springer Science & Business Media, 2001.
- [85] Borgatti S P, Everett M G. Notions of position in social network analysis[J]. Sociol. Methodol., 1992: 1-35.
- [86] Cormen T H, Leiserson C E, Rivest R L, et al. Introduction to algorithms[M]. MIT Press, 2009.
- [87] Kunegis J. Konect: the koblenz network collection[C]//Proceedings of the 22th World Wide Web Conference. ACM, 2013: 1343-1350.
- [88] Rossi R, Ahmed N. The network data repository with interactive graph analytics and visualization[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. AAAI, 2015: 4292-4293.
- [89] Erdős P, Rényi A. On random graphs i[J]. Publ Math Debrecen, 1959, 6: 290-297.
- [90] Barabási A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.
- [91] Balasubramanian K. Computer generation of automorphism groups of weighted graphs[J]. J. Chem. Inf. Comput. Sci., 1994, 34(5): 1146-1150.
- [92] Grech M, Imrich W, Krystek A D, et al. Direct product of automorphism groups of digraphs[J]. Ars Math. Contemp., 2019, 17(1): 89-101.
- [93] Fouss F, Pirotte A, Renders J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. IEEE Trans. Knowl. Data Eng., 2007, 19(3): 355-369.

- [94] Senelle M, Garcia-Diez S, Mantrach A, et al. The sum-over-forests density index: identifying dense regions in a graph[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2013, 36(6): 1268-1274.
- [95] van der Grinten A, Angriman E, Predari M, et al. New approximation algorithms for forest closeness centrality—for individual vertices and vertex groups [C]//Proceedings of the 2021 SIAM International Conference on Data Mining. SIAM, 2021: 136-144.
- [96] Chebotarev P. Spanning forests and the golden ratio[J]. Discrete Appl. Math., 2008, 156(5): 813-821.
- [97] Golender V, Drboglav V, Rosenblit A. Graph potentials method and its application for chemical information processing[J]. J. Chem. Inform. Comput. Sci., 1981, 21(4): 196-204.
- [98] Merris R. Doubly stochastic graph matrices[J]. Publikacije Elektrotehničkog Fakulteta. Serija Matematika, 1997(8): 64-71.
- [99] Zhang X D. Vertex degrees and doubly stochastic graph matrices[J]. J. Graph Theory, 2011, 66(2): 104-114.
- [100] Johnson W B, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space[J]. Contemporary Mathematics, 1984, 26: 189-206.
- [101] Achlioptas D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins[J]. Journal of Computer and System Sciences, 2003, 66(4): 671-687.
- [102] Kyng R, Sachdeva S. Approximate gaussian elimination for laplacians-fast, sparse, and simple[C]//Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science. IEEE, 2016: 573-582.
- [103] Arasteh M, Alizadeh S. A fast divisive community detection algorithm based on edge degree betweenness centrality[J]. Applied Intelligence, 2019, 49(2): 689-702.
- [104] Lin C C, Huang W, Liu W Y, et al. A novel centrality-based method for visual analytics of small-world networks[J]. Journal of Visualization, 2019, 22(5): 973-990.
- [105] Teixeira A S, Monteiro P T, Carriço J A, et al. Spanning edge betweenness[C]//Workshop on mining and learning with graphs: Vol. 24. 2013: 27-31.
- [106] Hayashi T, Akiba T, Yoshida Y. Efficient algorithms for spanning tree centrality [C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016: 3733-3739.
- [107] Doyle P G, Snell J L. Random walks and electric networks[M]. American Mathematical Society, 1984.
- [108] Merris R. Doubly stochastic graph matrices, ii[J]. Linear and Multilinear Algebra, 1998, 45(2-3): 275-285.
- [109] Watts D J, Strogatz S H. Collective dynamics of ‘small-world’ networks[J]. Nature, 1998, 393(6684): 440-442.

攻读学位期间的研究成果

已发表论文

1. **Qi Bao**, Zhongzhi Zhang. Discriminating power of centrality measures in complex networks. IEEE Transactions on Cybernetics 52 (11), 12583-12593. (中科院一区, CCF B 类, IF=11.8, 对应本文第三章)
2. **Qi Bao**, Wanyue Xu, Zhongzhi Zhang. Benchmark for discriminating power of edge centrality metrics. The Computer Journal 65 (12), 3141-3155. (CCF B 类, IF=1.4, 对应本文第五章)
3. Yujia Jin, **Qi Bao**, Zhongzhi Zhang. Forest distance closeness centrality in disconnected graphs. IEEE International Conference on Data Mining 2019. 2019 : 339 –348. (部分对应本文第四章)
4. Wanyue Xu, **Qi Bao**, Zhongzhi Zhang. Fast evaluation for relevant quantities of opinion dynamics. Proceedings of the Web Conference 2021. 2021 : 2037 – 2045.
5. Liwang Zhu, **Qi Bao**, Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. Advances in Neural Information Processing Systems, 2021, 34 : 2072 –2084.
6. Yucheng Wang, **Qi Bao**, Zhongzhi Zhang. Combinatorial properties of farey graphs. Theoretical Computer Science 796, 70-89
7. Liwang Zhu, **Qi Bao**, Zhongzhi Zhang. Measures and optimization for robustness and vulnerability in disconnected networks. IEEE Transactions on Information Forensics and Security 18, 3350-3362.

拟发表论文

1. **Qi Bao**, Zhongzhi Zhang. Role Similarity Metric Based on Spanning Rooted Forest (已投 IEEE Transactions on Network Science and Engineering, 目前处于大修阶段, 部分对应本文第四章)

致谢

在此，我衷心感谢那些曾经帮助过我的人。

首先，我要感谢章忠志教授对我的学术指导。他的严谨治学态度和卓越的学术造诣是我学习的榜样。在整个研究过程中，他耐心地解答我的问题，给予我宝贵的建议和意见。他总是鼓励我尝试新的研究方向，推动我克服困难和挑战。在他的指导下，我逐渐成长为一名独立思考和解决问题的研究者。

其次，我要感谢章忠志教授对我的个人关怀和鼓励。他不仅关心我的学术进展，还关心我的生活和工作状态。他经常与我交流，了解我的困惑和挑战，并给予我支持和鼓励。在我面临困难和挫折时，他总是给我勇气和力量，让我坚持不懈地追求自己的目标。

此外，我要感谢实验室的老师 and 同学们。他们与我一起工作、学习和交流，互相支持和帮助。在实验室的氛围中，我不断学习和成长，获得了宝贵的经验和知识。没有他们的合作和支持，我无法顺利完成我的研究工作。

最后，我要特别感谢我的家人。他们一直以来对我的支持和理解是我最大的动力和鼓励。他们在我学习和研究的道路上默默支持着我，给了我无尽的勇气和动力。

再次向章忠志教授表示最衷心的感谢！感谢章忠志教授对我的信任和培养。

复旦大学

学位论文独创性声明

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。论文中除特别标注的内容外，不包含任何其他个人或机构已经发表或撰写过的研究成果。对本研究做出重要贡献的个人和集体，均已在论文中作了明确的声明并表示了谢意。本声明的法律结果由本人承担。

作者签名：_____ 日期：_____

复旦大学

学位论文使用授权声明

本人完全了解复旦大学有关收藏和利用博士、硕士学位论文的规定，即：学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。涉密学位论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：_____