# Network Data Community Detection based on Spectral Clustering

**Abstract**：The majority of network data exhibits community structures, and discovering these structures contributes to uncovering potential information within communities. Spectral clustering algorithms are classical methods for solving community discovery, known for their accuracy, stability, efficiency, and ability to obtain global optimal solutions. However, traditional spectral clustering algorithms may be less effective for large and sparse network data. In this paper, we address this limitation by simulating the discovery of communities in sparse networks with degree heterogeneity. We utilize both synthetic and real datasets for community discovery and introduce modified spectral clustering algorithms, namely SCORE, SCORE+, SCOREq, and RSC. These algorithms outperform traditional spectral clustering algorithms like oPCA and nPCA, especially in sparse networks with node degree heterogeneity, providing better interpretations for the Degree-Corrected Block Model (DCBM). For the application of community discovery in a statistical co-authorship network that is both heterogeneous and sparse, we propose an effective analytical paradigm. We adopt a hierarchical community discovery approach, employing the SgnQ statistic as a hypothesis testing criterion for stopping divisions. The SCOREq algorithm is applied for community discovery at each layer. We use visual aids such as dendrograms, modularity, and intra/inter-community edge analysis to assist in determining the optimal number of communities (K), resulting in a hierarchical community discovery tree with four layers. Subsequently, we analyze the research directions of the top two centrality-ranked authors in each identified community, providing meaningful names to interpret the significance of different communities. Finally, through a comparative analysis with SCORE and RSC algorithms, we observe a certain degree of similarity in the results, indicating the robustness of the community partitioning outcomes.

Keywords: Community Discovery, Spectral Clustering, DCBM Model, Statistical Co-authorship Network

1. 1 Introduction

In recent years, with the advancement of massive data storage and processing capabilities, network data has become a research focus in various fields such as statistics, mathematics, information science, physics, and biology. Network data abstracts entities in the natural world as nodes and represents relationships between these entities as edges.

Many studies indicate that network data often exhibit community structures, where nodes within a community are densely connected, while connections between communities are relatively sparse. Discovering these communities simplifies the research process and aids researchers in uncovering useful potential information.

Community discovery is a typical clustering problem, aiming to map each node in a research network to at least one community. While past research has proposed various community discovery algorithms such as hierarchical clustering and modularity optimization, challenges persist in discovering communities in real-world networks.

Two main difficulties in real-world network community discovery are the large number of nodes with sparse and heterogeneous connections, making these networks complex and challenging for existing algorithms. Additionally, the absence of true community labels in network data complicates the accurate evaluation and comparison of community partitioning results.

This study focuses on network characterization methods based on statistical inference, aiming to fit generative models to network data for encoding community structures. Spectral clustering algorithms have proven to be effective for fitting random block models, demonstrating convergence properties in clustering error rates.

The paper analyzes classical spectral clustering algorithms, explores their extensions, and discusses various variations and modification methods. Through simulations and real data analysis, the study investigates the performance of these algorithms under different data characteristics, providing insights for the selection of community discovery algorithms in real-world networks.

Furthermore, the paper explores practical applications of community discovery, specifically in the context of discovering communities in a co-authorship network of statisticians. The network represents authors as nodes and collaboration relationships between authors as edges. Analyzing the community structure of this network helps understand collaboration patterns among statisticians, their characteristics, and research topics, contributing to the field of literature studies. Given the large scale, sparsity, and node degree heterogeneity of this network, the study considers the analysis results from earlier sections to inform algorithm selection and evaluates the rationality and interpretability of community discovery results from multiple perspectives.

1. 2 Basic Concepts

For large-scale network structural data, significant degree heterogeneity, sparse network structure, and susceptibility of crucial information to noise are common challenges. To explore these characteristics within the network, the following metrics are considered:

1) Degree Heterogeneity

The degree of a node characterizes the properties of an individual node within the network, while the network's degree characterizes the overall properties of the network. A high degree of a node indicates extensive connections within the network, reflecting its influence. In a network with a relatively even degree distribution, nodes exhibit homogeneity, and relationships among them are relatively flat. Conversely, in a network with an uneven degree distribution, there is a stronger heterogeneity among nodes [17]. In large-scale network data, node degrees often follow a power-law distribution, indicating pronounced heterogeneity.

Building upon the DCBM model, one can further define metrics for degree heterogeneity:

越小，度异质性越严重。当时，的下界为一个大于0的常数。在存在严重程度异质性的情况下，接近于0。

* + 1. 网络稀疏性

现实世界中的大型网络往往十分稀疏，稀疏网络是指连边很少的网络，其邻接矩阵中包含大量的0元素。刻画稀疏程度可以使用**平均度、网络密度**等指标。网络密度的定义如上文所示。

**平均度**为网络中所有节点的度的平均，平均度越高，则整体意义上网络中的边越密集，网络中节点间的联系越紧密。当网络节点的平均度小于时，传统谱聚类算法表现不理想[18]。

基于DCBM模型，还可以定义网络稀疏性性指标：

在该模型下，其取值范围为

* + 1. 信号与噪声[16]

信噪比是一种对于网络中弱信号程度的度量。在许多情况下，社区结构是微妙的，并被强噪声掩盖，其中信噪比（Signal-to-Noise Ratio, SNR）相对较低：

DCBM模型下，是邻接矩阵的第*K*大的特征值。如果依赖于*n*，则DCBM足以对弱信号情况进行建模，其中可能远小于，。具体解释为：

（1）视为稀疏性水平和噪声水平（即噪声矩阵*W*的谱范数）

（2）视为信号强度，因此是信噪比

* + 1. 综合指标

SCORE的聚类能力取决于度异质性与信噪比的合并指标：

1. 网络数据模型DCSBM与谱聚类算法

基于统计推断的网络刻画方法试图将生成模型拟合到网络数据中，对社区结构进行编码，具有很好的可解释性、表达能力、泛化能力及灵活性。其中最具代表性的模型是随机块模型(Stochastic Block Model, SBM)。下面将介绍该模型及其修正模型：

* 1. DCBM模型

首先定义一个无向无权网络，*V*为节点集合，*E*为边的集合，节点来自*K*个社区，即：，设为*N*的大小的邻接矩阵，

无向无权网络中有*K*个社区，则DCBM模型为：

即：

引入指示社区的向量，其余元素为0。则有

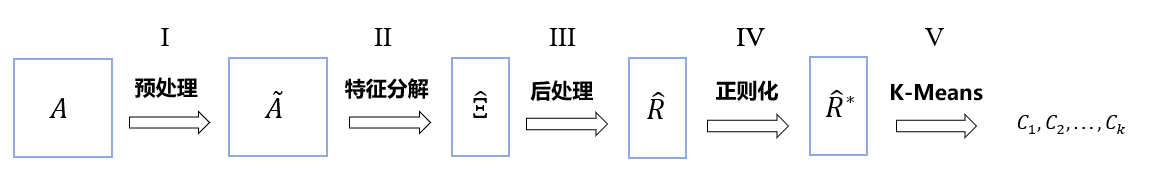
令，，则矩阵形式为

即：

其中*Ｐ*为描述社区结构的矩阵，假定满足对称、非奇异、非负（特指所有元素非负）、不可约。此外要求：*Ｐ*对角线元素为1，，*E*(*A*)代表*A*的期望，*W*被称为广义Wigner矩阵。可以这样理解这几个公式：元素对应非矩阵形式的节点*i*与*j*的连边概率，由于取值为0或1，自然，由于模型假定无自连接，因此矩阵形式需减去对角线元素构成的对角矩阵，最终。而*W*为中心伯努利分布，

即

* 1. 谱聚类算法

谱聚类一般分为如下5步骤：

* + 1. 预处理

对邻接矩阵*A*做“归一化”变换，得到变换后的矩阵。

* + 1. 特征分解（可以视作降低噪声）

对矩阵特征分解，获得前*K*个最大特征值对应特征向量，对应的特征值

记

注：SCORE+方法为解决弱信号，对于此步骤增加处理：

对于给定阈值*t* > 0，当时，设，否则。

* + 1. 后处理

对前*K*个特征向量做“归一化”变换：定义一个尺度不变映射，满足。

对于的每一行，应用这种映射，获得矩阵。

* + 1. 正则化

对于一个给定的阈值*T* (*T* > 0)，设是的正则化版本，其中，。

* + 1. 聚类

基于矩阵对*n*个节点做*K*-means聚类，得到*n*个节点的社区属性。

结合上文对各种修正处理的分析、讨论，选择和构建如下谱聚类算法进行分析讨论：

**表1 本文中使用谱聚类算法的变式**

|  |  |  |  |
| --- | --- | --- | --- |
| 算法名称 | 预处理 | 后处理 | 超参数取值 |
| OPCA  (传统谱聚类) | *A* | (无) |  |
| nPCA  （拉普拉斯变换谱聚类） |  | (无) | 还可尝试其他取值  如0.05 |
| SCORE  （度异质性修正谱聚类） | *A* | )是的第一项，最后删除第一项 |  |
| SCOREq | *A* |  |  |
| RSC[13] |  | )是的第一项 | 还可尝试其他取值 |
| SCORE+ |  | 加权：对于第*i*行， |  |

1. 算法的性能讨论

对于社区发现的效果评估并没有绝对的指标，常用的指标值与聚类的评估指标有相似之处，要求社区内部连接较为紧密，社区之间连接较少。评价指标有：错误率、模块度等等。

通过模拟实验能够在控制变量，研究不同条件下比较不同算法的性能。本文针对现实网络中存在节点度异质性、信号弱噪声大、网络规模大等问题，设计了4个实验，比较谱聚类算法oPCA、nPCA、SCORE、SCORE+、SCOREq、RSC的性能｡

**参数设置包括：**大小、社区数、实验重复次数(*n*, *K*, *rep*)、概率矩阵、度异质向量、标签向量

**实验步骤为：**

1. 生成主要信息，满足
2. 生成噪音矩阵

对角线为0的对称矩阵，上对角矩阵为以为参数的中心伯努利分布，即：

1. 生成的邻接矩阵

即

1. 求N的最大连通分量：为的大小。
2. 应用谱聚类算法进行社区发现，得到每个节点的社区标签。
3. 重复步骤(2)-(5)共*rep*次，得到节点分配错误率和方差。
   * 1. 实验1：比较无度异质性网络下算法的性能

* 大小､社区数､实验重复次数：(*n*, *K*, *rep*) = (1000, 2, 100)
* 概率矩阵
* 度异质向量：
* 标签向量：

在这种参数设定下，其网络结构的指标为：。注意，度异质性越弱，越大，当没有度异质性时，；噪声越强，SNR越小；当度异质性弱，信号强时，较大。

**表2 实验1算法表现结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **方法** | **SCORE** | **SCORE+** | **SCOREq2** | **RSC** |
| **均值(方差)** | 0.058(0.000) | 0.055(0.000) | 0.058 (0.000) | 0.055(0.000) |
| **方法** | **SCOREq1** | **oPCA** | **nPCA** |  |
| **均值(方差)** | 0.060(0.000) | 0.059(0.000) | 0.056(0.000) |  |

**结论1：**在无度异质性的DCBM模型生成的网络数据中SCORE､SCORE+､SCOREq2､RSC、SCOREq1、oPCA、nPCA均能很好的划分社区，节点分配错误率均小于0.06。

* + 1. 实验2：比较在具有度异质性时算法的性能
* 大小､社区数､实验重复次数：(*n*, *K*, *rep*) = (1500,3,100)
* 概率矩阵：
* 度异质向量：
* 标签向量：

在这种参数设定下，其网络结构的指标为：。

**表3 实验2算法表现结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **方法** | **SCORE** | **SCORE+** | **SCOREq2** | **RSC** |
| **均值(方差)** | 0.073 (0.000) | 0.044(0.000) | 0.068 (0.000) | 0.066(0.0003) |
| **方法** | **SCOREq1** | **oPCA** | **nPCA** |  |
| **均值(方差)** | 0.069815(0.000040) | 0.362450(0.000068) | 0.275(0.019) |  |

**结论2：**在有度异质性的DCBM模型生成的网络数据中SCORE､SCORE+､SCOREq2､RSC、SCOREq1均能很好的划分社区，节点分配错误率均小于0.08，其中SCORE+节点分配错误率最小为0.043，而oPCA、nPCA、表现欠佳，节点分配错误率均超过0.25。

* + 1. 实验3.1 比较在不同度异质性形式设定时算法的性能
* 大小､社区数､实验重复次数：(*n*, *K*, *rep*) = (1000,2,100)
* 概率矩阵：
* 度异质向量：比较三种形式的度异质性设定形式如下：

其中

注：第一种设定下(a)度异质性较弱，其他两种设定度异质性较强｡

* 标签向量：

**表4 实验3.1 算法表现结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **方法** | **SCORE** | **SCORE+** | **SCOREq2** | **RSC** |
| **a** | 0.015(0.000) | 0.012(0.000) | 0.015(0.000) | 0.013(0.000) |
| **b** | 0.074(0.000) | 0.067(0.000) | 0.074(0.000) | 0.067(0.000) |
| **c** | 0.119(0.000) | 0.117(0.000) | 0.118(0.000) | 0.117(0.000) |
| **方法** | **SCOREq1** | **oPCA** | **nPCA** |  |
| **a**  **b** | 0.015(0.000)  0.074(0.000) | 0.058(0.000)  0.250(0.000) | 0.013(0.000)  0.133(0.024) |  |
| **c** | 0.118(0.000) | 0.235(0.000) | 0.419(0.012) |  |

在参数设定(a)，(b)，(c)下，其网络结构的指标为：。

可以看到此种模型设定的度异质性无法仅用来度量，显然三种设定下均有，其异质性是由度异质性参数的中间值变化程度造成的。自然地，

**结论3：**在不同形式的度异质性的DCBM模型设定下，SCORE､SCORE+､SCOREq2､RSC、SCOREq1表现结果基本优于oPCA、nPCA。在度异质性程度较低时（a设定下）nPCA也能较好的划分出社区。

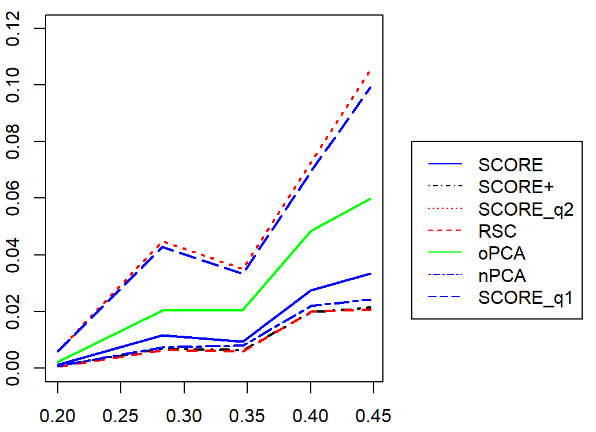
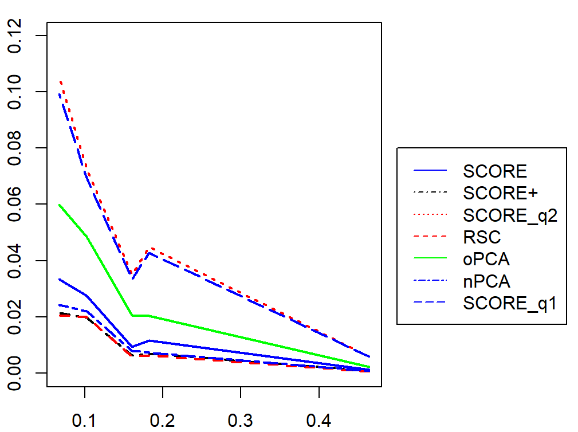
**结论4：**在DCBM模型设定下，随着度异质性的增加，节点分配错误率增加，即度异质性是影响算法表现结果的重要因素，且SCORE､SCORE+､SCOREq2､RSC、SCOREq1对度异质性有较强的抑制作用，而nPCA对度异质性只有一定的抑制作用，但这七种算法均不能完全消除度异质性的影响。

* + 1. 实验3.2 比较在不同度异质性形式设定时算法的性能
* 大小､社区数､实验重复次数：(*n*, *K*, *rep*) = (1000,2,100)
* 概率矩阵：
* 度异质向量：先生成,然后归一化：
* 标签向量：

在不同参数设定下()，其网络结构的指标为：

。

在此种模型设定下，由于度异质性参数的均由正态分布产生，其异质性的不同是由参数变化造成，因此能够很好地度量度异质性程度。

**图1 实验3.2 中不同算法的节点分配错误率**

图左：*y*轴为节点分配错误率，*x*轴为参数；图右：*y*轴为节点分配错误率，*x*轴为信噪比

**结论5：**在DCBM模型设定下，若度异质性参数生成模型形式相同，则指标是算法性能的一个预示，越大(对应越小)，节点分配错误率越低(如图1右)。

**结论6：**在此次实验中RSC、nPCA表现较好，SCOREq1、SCOREq2表现较差。在此次实验中，一个较为意外的结果是oPCA、nPCA在此种设定下表现较好。

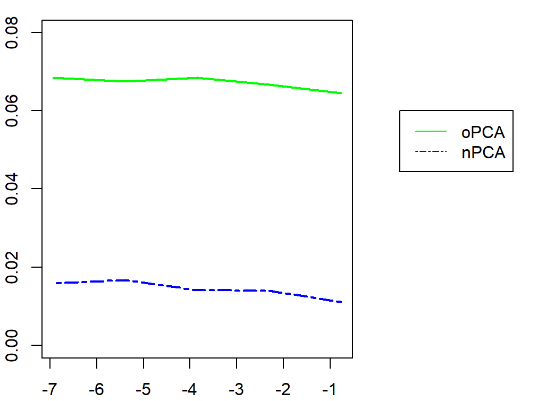
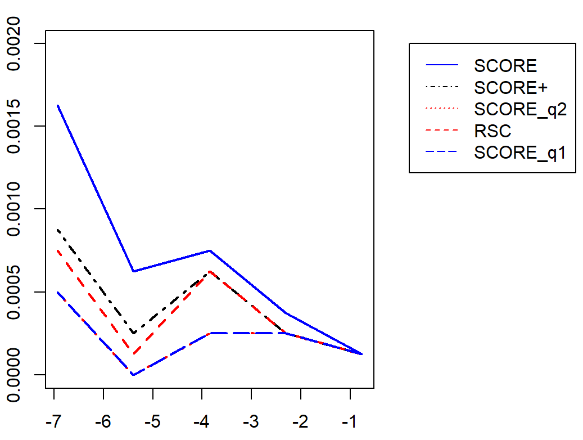
* + 1. 实验4 比较在较大样本量时算法的性能
* 大小、社区数、实验重复次数：(*n*, *K*, *rep*) = (4000,2,25)
* 概率矩阵：
* 度异质向量：首先生成

然后对值除以范数归一化，并令

* 标签向量：

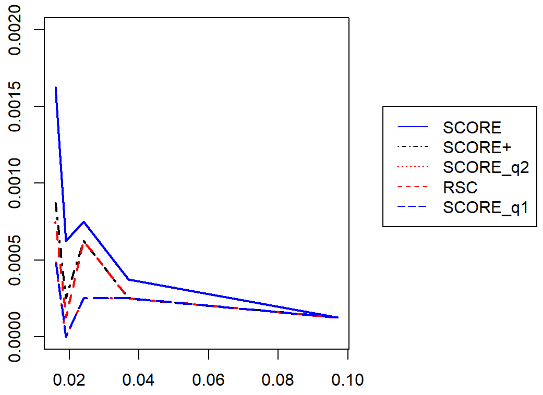
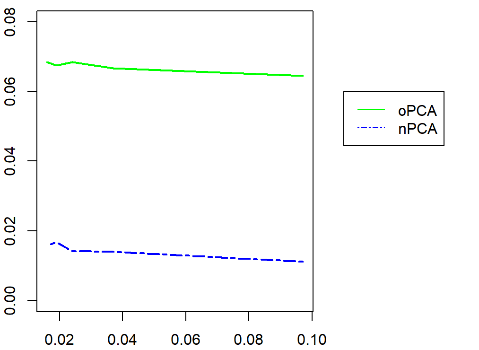
在不同参数设定下，其网络结构的指标为：。

在此种模型设定下，由于度异质性参数的均由线性函数产生，其异质性的不同是由参数变化造成,因此能够很好地度量度异质性程度。

**图2 实验4中不同算法的节点分配错误率**

上图：*y*轴为节点分配错误率，*x*轴为参数；下图：*y*轴为节点分配错误率，*x*轴为信噪比；

结论6：在此次实验中SCOREq1、SCOREq2表现较好，oPCA、nPCA表现较差。

 综上，在网络数据无度异质性时，SCORE，SCORE，SCOREq2，RSC，SCOREq1，oPCA，nPCA七种算法在进行社区发现时均表现出良好的性能，在具有度异质性时，SCORE､SCORE+､SCOREq2､RSC、SCOREq1五种算法总体而言表现性能优于nPCA、oPCA，在度异质性不高或某些特定度异质性生成形式下，nPCA也能抑制度异质性较好的划分社区。此外，根据以上模拟结果，在不同情形下，算法的表现也具有一定的差异，因此仍需结合具体数据特点选择合适的算法。

经过对比，可以得到结论：SCORE、SCORE+、SCOREq、RSC在面对异质性、稀疏性、噪声等问题时（Caltech和Simmons数据集已被证实具有较明显的噪声）表现比较好，其中SCORE+对具有明显噪声的数据表现较为优良，在后续的分析中可以结合数据的具体情况进一步分析。

1. 应用：统计学家合作者网络社区发现

在过去的几十年里，科学界的规模大幅增长。科学界的快速发展激发了许多有趣的大数据项目，其中之一是如何利用科学领域的大量出版物来描绘该领域的研究习惯、趋势和影响的全貌。这些研究有助于审查国家和全球与科学出版物相关的活动，对大学进行排名，并做出资助、晋升和奖励的决定。

研究统计学家合作者网络的社区结构就是一个有趣的问题，合作可能由许多因素驱动，例如，地理邻近性、学术谱系、文化联系等。为了更加精确地分析数据，考虑使用分层的社区发现方法，并尝试对不同社区进行解释。

* 1. 数据介绍与问题描述

本文中使用的MADStat数据集[[1]](#footnote-0)包含1975-2015年统计学、生物统计学、概率论、机器学习和相关领域的36个代表性期刊中的83331篇的基本信息，如作者、标题等，共涉及47331个作者。对该数据构建的统计学家合作者网络，进行社区发现。

36种期刊的选择遵循澳大利亚研究委员会(ARC)提供的175种2010年统计期刊排行榜，该榜单用于澳大利亚大学的绩效评估，是澳大利亚卓越研究项目的一部分。175种期刊被分为四类:A\*、A、B和C。具体选择情况为：对于9种A\*期刊，全部选择，其中两种AOP和PTRF是概率方向的期刊；其次，选取所有A类期刊，除应用概率或工程类主题较强的期刊（如应用概率进展、电子概率杂志、金融与随机学杂志、应用概率杂志、随机过程及其应用、概率论及其应用、技术计量学、排队系统、随机结构与算法）；B类大约有50种期刊，涵盖了广泛的主题，其中只选择方法论和理论方面的期刊，如澳大利亚和新西兰统计杂志，贝叶斯分析，加拿大统计杂志等。不选择任何C类期刊。

网络中的每个节点代表一个作者，本文主要关注长期活跃的研究人员子集，以及稳固的合作，所以认为如果两个作者合作发表了超过*m*=3篇论文，则为这两个作者的节点之间建立一条边。（取*m*=2也可能是一个合理的选择，但是得到的网络相对来说更密集、更大，共有10741个节点。由于需要逐个手动检查每个已识别的社区，因此选择*m*=2需要更多的时间和精力来解释结果，故暂不作考虑[19]）

本文选取这个构建的网络的最大连通分量进行分析，该连通分量中包含节4383个作者，6056条连边。此外，该统计学家合作者网络是无向的，所有边的权数为1，并且不计自连边，假设每个节点都只属于一个社区。

需要解决的问题是对核心网络进行社区发现：尝试分析网络中一共有几个社区，每个社区是怎样的结构，应当如何解释这种社区划分。下文将分步骤进行分析：

* 1. 数据描述统计

数据的统计描述有助于充分了解数据的特征，对数据产生初步的认识，并为算法选择提供支持和参考。

* + 1. 网络稀疏性指标——网络密度、平均度

网络密度反映了实际存在的连边数与可以存在的连边数的比值，反映了网络中各个节点连接的紧密程度，即稀疏性。经过计算，网络的密度为0.0006306262，即0.0631%，网络较稀疏。

网络的平均度为：2.763404，远小于log*|V|*，根据前文的分析，认为该网络稀疏。

* + 1. 节点度的分布、异质性

在该网络中，节点的度表示一个作者所合作的统计学家的数量。图3为关联网络节点度分布直方图，从图中可以看出网络具有严重的度异质性，节点度分布是严重左偏的，即大部分节点的度很小，存在少部分节点的度很大，最大为51。这也与社交网络的特点相符合，即存在少部分研究者的合作者众多，社交范围广，大部分统计学家的合作是处于较低水平的，而这些度较大的作者也是统计学家合作者网络中的重要节点。

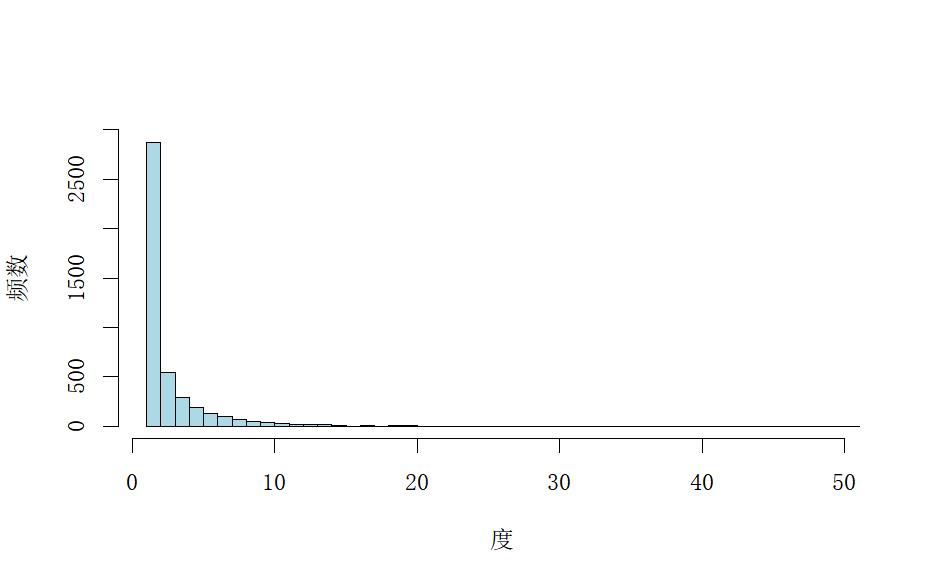
**图3 节点度的分布图**

表7展示了统计学家合作者网络中度最高的10个作者。排名第一的作者是Peter Hall，它的度为51，代表其合作者共有51个。

**表7 度排名前10的统计学家**

|  |  |  |
| --- | --- | --- |
| 排名 | 节点度 | 统计学家姓名 |
| 1 | 51 | Peter Hall |
| 2 | 46 | Raymond Carroll |
| 3 | 45 | Narayanaswamy Balakrishnan |
| 4 | 42 | Geert Molenberghs |
| 5 | 34 | Joseph Ibrahim |
| 6 | 32 | Jeremy Taylor |
| 7 | 27 | Holger Dette |
| 8 | 25 | Lixing Zhu |
| 9 | 24 | James S. Marron |
| 10 | 23 | David Dunson |

* + 1. 节点中心性——接近中心性

接近中心性可以反映一个节点与其他节点是否接近，接近中心性高的用户与许多其他研究者都有密切关联，是网络中的“中心”。表8展示了接近中心性最高的前10个作者。

**表8 接近中心性排名前10的作者相关信息**

|  |  |  |
| --- | --- | --- |
| 排名 | 接近中心性 | 统计学家姓名 |
| 1 | 0.1898860 | Peter Hall |
| 2 | 0.1850741 | Raymond Carroll |
| 3 | 0.1774448 | Yanyuan Ma |
| 4 | 0.1746513 | Matt P. Wand |
| 5 | 0.1726488 | Xihong Lin |
| 6 | 0.1716009 | Malay Ghosh |
| 7 | 0.1714397 | Jianqing Fan |
| 8 | 0.1713325 | Louise Ryan |
| 9 | 0.1707716 | Bingyi Jing |
| 10 | 0.1698121 | Hua Liang |

* + 1. 节点中心性——介数中心性

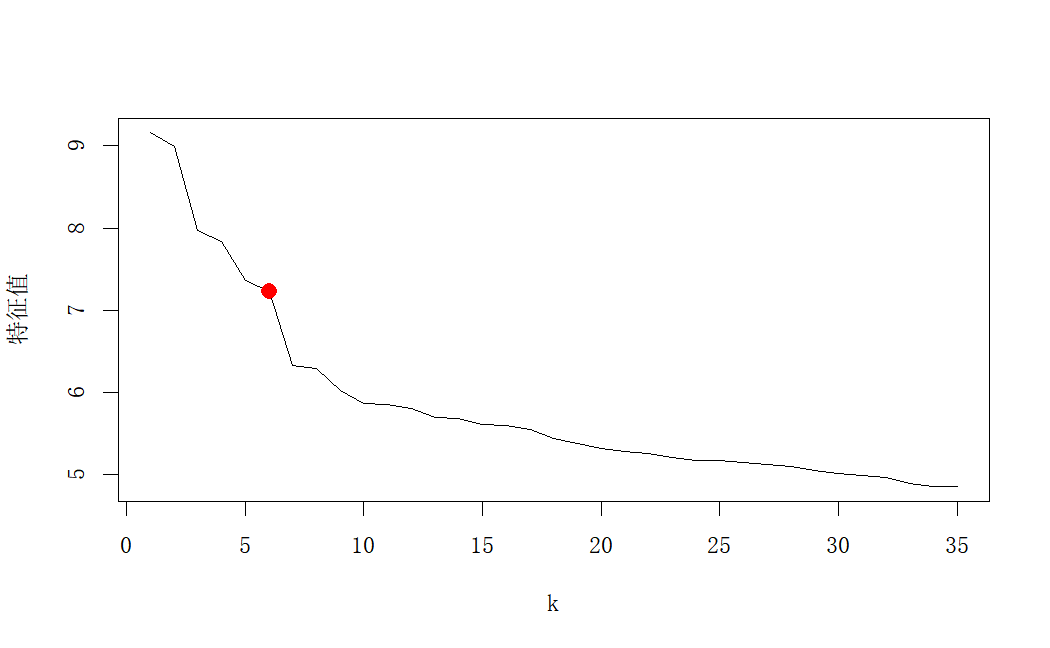
更进一步地，本文使用介数中心性指标来识别统计学家合作者网络中有影响力的作者。介数中心性可以反映一个节点在其他节点之间起"桥梁"作用的程度。如果某个作者的介数中心性较高，那么在该网络中它对交流合作的促进作用越强。表9展示了介数中心性最高的前10个作者。

**表9 介数中心性排名前10的作者相关信息**

|  |  |  |
| --- | --- | --- |
| 排名 | 介数中心性 | 统计学家姓名 |
| 1 | 0.24791518 | Peter Hall |
| 2 | 0.14638895 | Raymond Carroll |
| 3 | 0.09058327 | Qi Man Shao |
| 4 | 0.07665352 | Joseph Ibrahim |
| 5 | 0.07461228 | Malay Ghosh |
| 6 | 0.06308375 | Rahul Mukerjee |
| 7 | 0.05948326 | Narayanaswamy Balakrishnan |
| 8 | 0.05930359 | Geert Molenberghs |
| 9 | 0.05816338 | Ingrid Van Keilegom |
| 10 | 0.05633079 | Louise Ryan |

* + 1. 网络信噪比（SNR）

考虑到合作者网络节点社区标签未知、社区数也待定，所以结合上文对于强噪声、弱噪声网络的定义，可知：在进行社区发现之前，信噪比指标尚不能够得出。为了使得算法社区发现效果更好，考虑在进行社区选择时就使得该网络具备较强的信号。即令*/*取值较大。在图像上表示，可以视作是一个明显的拐点，取值较大，而骤然减小。

**图4 邻接矩阵的特征值碎石图**

综合上文对于网络的描述，可以得到统计学家合作者网络的特征是：节点异质性强、稀疏性显著。由于信噪比与社区发现结果有关，无法在划分社区前计算，所以从另一种思路考虑：为了避免强烈的噪声对社区发现效果的影响，在社区数*K*选择时倾向于选择特定的*K*值，以增强信噪比。例如，倾向于选择较小的*K*，以及满足*/*较大的*K*值。（显著大于但与较接近，如当*K*=6时）所以认为网络中的弱信号问题并不十分严峻。

所以，基于上述分析，对于统计学家合作者网络的社区发现有许多可选的算法，如SCORE及其变式SCOREq、RSC等，而由于网络信号较强，因此不选用SCORE+算法（该算法在信号弱的网络上效果更显著，事实上该算法在统计学家合作者网络中的社区划分结果也不十分理想）。

* 1. 统计学合作者网络社区发现

由于统计学中有不同的研究子领域，且不同子领域也由于地域等原因存在许多学术群体，所以合作者网络中应当存在许多具有价值的社区。下面将对由4383个节点构成的核心网络进行社区发现。

* + 1. 分层的社区发现思想

由于统计学合作者网络较大，且有不同的研究子领域，不同子领域内也由于地域等原因存在许多学术群体，所以合作者网络中应当存在具有层次的社区，即网络可能是不同级别的小社区的集合。下文考虑进行分层社区发现，以使得发现的结果更具有可解释性。

具体方法如下：

1. 将输入的统计学家合作者网络进行聚类，划分为个子网络，其中*，*是网络中社区的总数。聚类时，考虑运用了前文中性质优良的**SCORE、SCOREq2、RSC算法**。
2. 对于每个子网络做**SgnQ[16]**假设检验。原假设（*H0*）为子网络中只有一个社区，备择假设（*H1*）为子网络中有多个社区。

**SgnQ[16]（Signed-Quadrilateral）检验**定义如下：

*A*是子网络对应的邻接矩阵，令且

若原假设成立，在温和条件下，有。因此可以通过得到近似*p*值，其中是*N*(0,1)的分布函数。如果*p*值较小，则拒绝原假设，认为子网络中有多个社区。

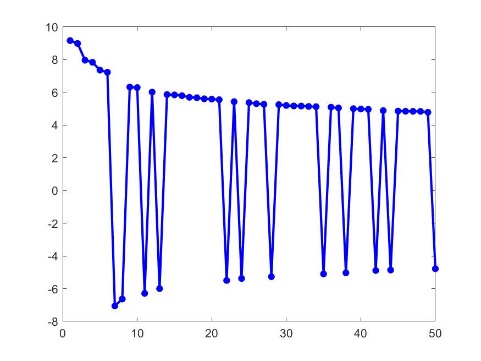
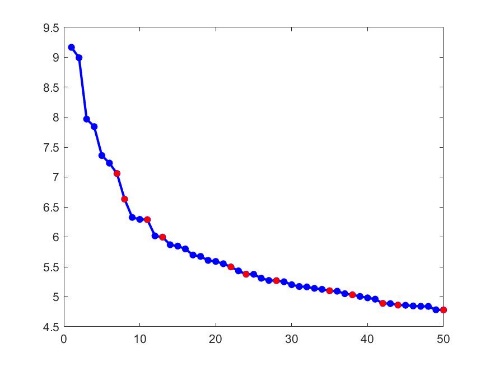
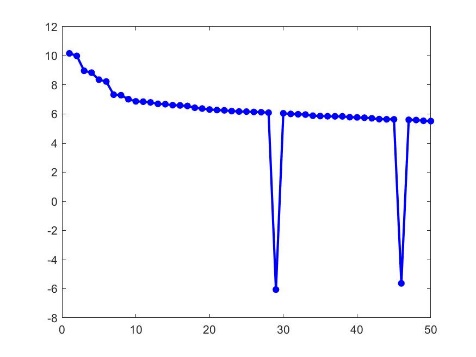
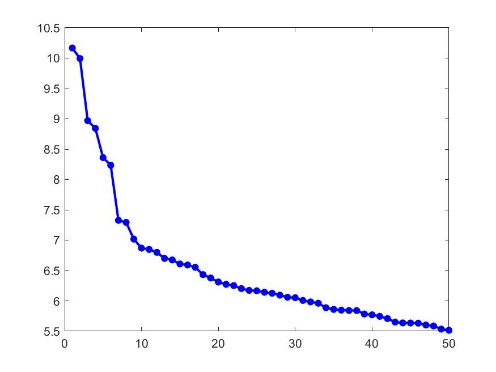
综合数据特征考虑，如果*p*>0.001，或子网络中节点数小于等于250，则认为当前的子网络只存在一个社区，不再对这个子网络进一步划分，该子网络中的节点构成一个单独的社区。否则，这个子网络将被进一步划分社区。

1. 当零假设在每个子网络中都被接受时，算法停止。
2. 算法的输出是一个层次树，树每个叶节点对应一个社区。
   * 1. *K*的选择准则

在社区发现时，由于使用的是谱聚类算法，所以都需要人为给定社区数*K*。如何选择一个较为合适的*K*是一个重要的问题。本文综合考虑了**特征值的碎石图、社区内部与社区间的连接情况、模块度**，三种因素，来选择第一层社区数*K*。第二层的初始*K*选择方法与第一层类似。

下面以第一层的社区数选择为例进行解释，在此使用的是SCOREq2算法。（SCORE、RSC对于第一层*K*的选择有相似结果）

**特征值的碎石图**是按照特征值大小排列的，以特征值大小为纵坐标、特征向量次序为横坐标生成的散点图。该碎石图有明显的拐点，一般取拐点前所有的因子及拐点后第一个因子作为主成分[12]。给定该网络的邻接矩阵，首先输出并观察该网络的碎石图，发现拐点是4、7、8和11。因此关注。但是第7、8和11个最大特征值是负的，所以由于合作者网络都是同质的个体，这些负特征值不太可能包含真实信号[15]。所以本文考虑通过进行特征分解来惩罚负特征值。特征分解结果如图5所示：

**图5 特征分解的碎石图**

左上：*A*的特征值；右上：*A*的特征值的绝对值；左下：*A*+I的特征值；右下：*A*+I的特征值的绝对值

由上述分析，初步认为*K*=6是较为理想的取值

**社区内部与社区间的连边情况**可以佐证使用碎石图选择的*K*值是可信的。对每个的分类结果进行研究，以*K*=6，7，8为例，分类结果示意图如图6：

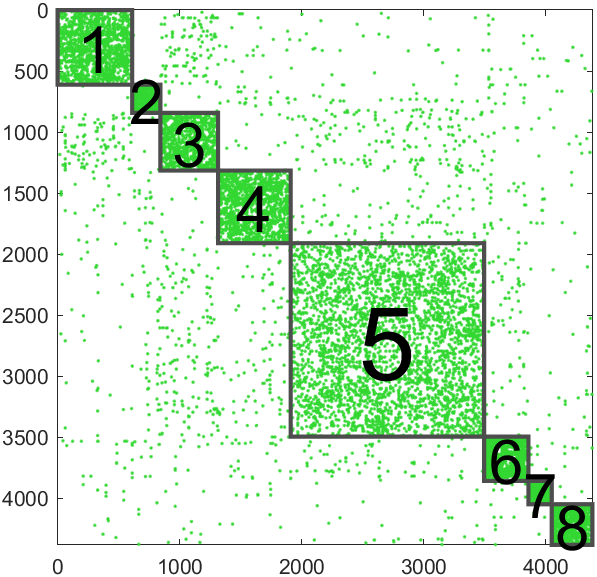
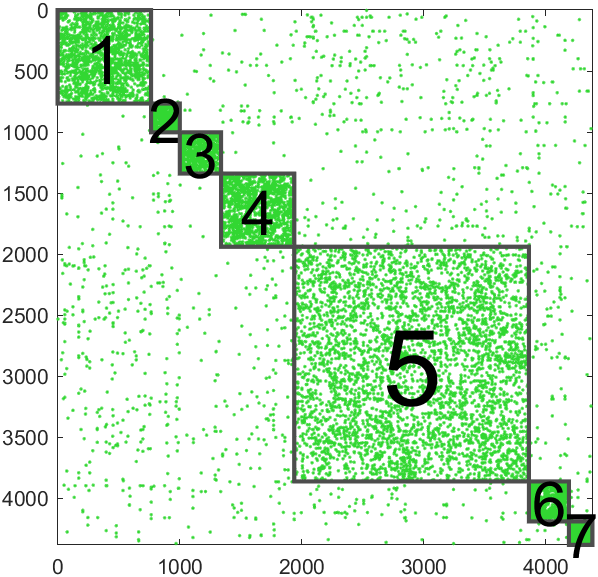
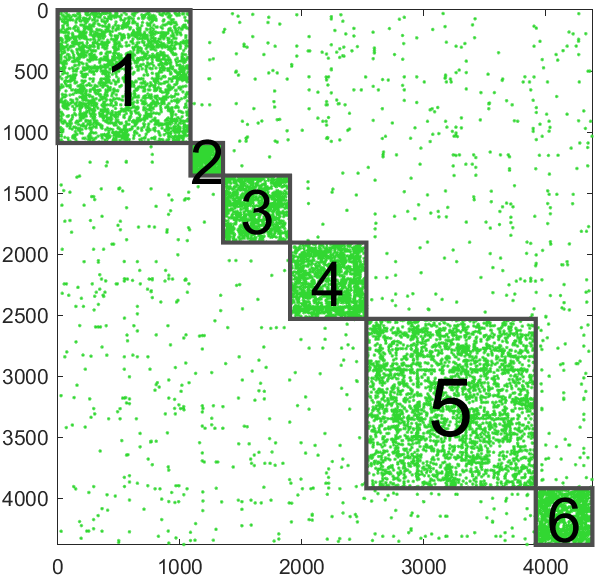
**图6 不同*K*值下的分类结果示意图（左至右：*K*=6，7，8）**

图6中每个散点代表一个作者，每个方块代表一个社区，方框内的绿色亮点代表社区内对应的两个节点存在连边，方框外的点代表该点与不同社区存在连边，即社区间有连边。基于本文的分析目的与“无混合成员”的假设，希望示意图中方框外的点尽可能的少，即社区间的连边尽可能地少。直观地，当*K*=7，8时，不同社区间连边相对多于*K*=6，并且当*K*>8时也有类似的结果，不再作展示。

结合**模块度**对*K*值的选择进行了进一步分析，结果如表10：

**表10 不同算法不同*K*值下聚类结果的模块度*Q***

|  |  |  |  |
| --- | --- | --- | --- |
| *K* | SCOREq | SCORE | RSC |
| 4 | 0.6300 | 0.6003 | 0.1611 |
| 5 | 0.6613 | 0.6358 | 0.6065 |
| 6 | 0.7123 | 0.6681 | 0.5905 |
| 7 | 0.6609 | 0.6068 | 0.6918 |
| 8 | 0.7031 | 0.6357 | 0.7078 |

模块度越大表明社区划分效果越好，*Q*值的范围在[-0.5,1)。研究表示当*Q*值在0.3~0.7之间时，说明聚类的效果很好。显然，表10中数据表明，当算法选择SCOREq，*K*=6时，模块度*Q*取到最大。

上述三种结果一致地支持*K*=6的社区数选择，所以我们对于第一层聚类的*K*值选择为*K*=6。第二层的社区数选择与此思想类似，不再赘述。

* 1. 社区发现结果

经过算法选择、社区数选择，发现SCOREq算法是较好的选择，最终选择使用SCOREq2算法对统计学家合作者网络进行社区发现。它产生了图中的社区树，这棵树在第一层有6个社区。

对于每个社区，观察其中具有较大度的节点的作者的研究方向，建议将这些社区命名解释如表11：

**表11 第一层社区划分与命名的详细信息**

|  |  |
| --- | --- |
| **社区名** | **描述** |
| C1 生物统计学（UNC） | 生存分析、纵向数据分析、UNC的生物统计学家及其密切合作者 |
| C2 生物统计学（欧洲） | 来自欧洲的生物统计学家及其密切合作者 |
| C3 数理统计 | 测试、计算统计学、概率论以及其它概率与统计学的经典理论 |
| C4 生物统计学（北美洲） | UM的生物统计学家及其密切合作者、其它北美洲生物统计学家 |
| C5 非参数统计 | 决策理论、非参数方法、高维统计、机器学习 |
| C6 半参数统计 | 半参数方法、生物统计学、贝叶斯、公共卫生 |

接下来分析树的其它层的具体表现。层次社区发现的停止规则是SgnQ统计量的*p*值大于0.001或社区大小小于等于250。但图中有一个例外：C3-1有275个节点，并且其*p*值约等于0，然而，在用SCOREq2将其进一步拆分为2个子社区后，一个子社区仅包含12个节点，另一个子社区的*p*值为0.24。因此，保持C3-1不变。

对于每个叶子社区（即，最终得到的，不再进一步划分的社区），使用两种常用的中心性度量：介数中心性、接近中心性，从而获得该社区的具有代表性的人，以为这个社区命名。给定一个叶子社区，使用介数中心度最大的两个节点来标记社区。附录一给出了划分得到的每个叶子社区的具体信息。层次树网络见图7。

研究结果证实，形成一个紧密联系的合作者社区有多种因素：相似的研究兴趣、学术谱系、友谊、同事关系、地理位置接近或密切的文化联系。以下是一些例子：

例1：类似的研究兴趣

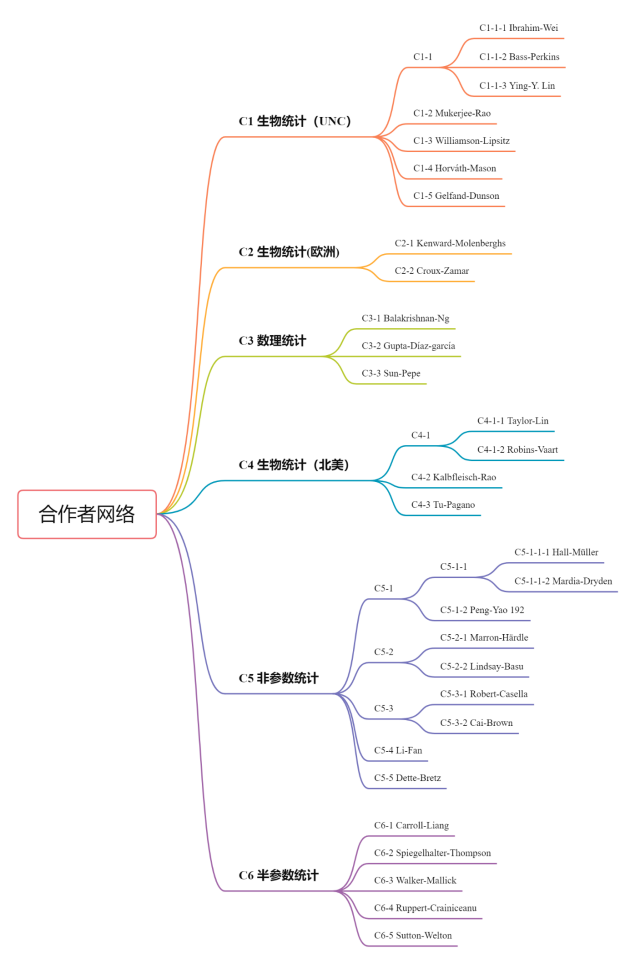
许多叶子社区是由具有相似研究兴趣的研究人员组成的，例如C6-1 Carroll-Liang（流行病统计学），C6-3 Walker-Mallick（贝叶斯），C5-1-1 Hall-Müller（非参数统计），C1-1-1 Ibrahim-Wei（遗传统计学）。

例2：地理和文化因素

在地理或文化上接近的人（例如同时、相邻研究所或同一地区或国家的研究人员）更有可能形成紧密的社区，例如C2-1 Kenward-Molenberghs（比利时的生物统计学家），C1-1-1 Ibrahim-Wei（北卡罗莱纳州研究三角区的统计员）。此外，与理论统计学家相比，地质和文化效应在生物统计学学家之间形成社区方面发挥着更重要的作用，一个可能的原因是生物统计学的合作研究更多地依赖于人力和数据共享。例如，为了遵守数据共享政策，一个人更容易与同一研究所、国家的人在一起进行协作。

例3：学术谱系

同一学术谱系下的师生之间往往有密切的学术合作。如，C5-2-1这个叶子社区有子社区，但是因为其社区尺寸小于250，没有进行进一步划分。它的一个子社区主要由三代师生组成，Jun Liu和Xiaotong Shen是Wing H Wong的学生，这三位作者和三人的学生共同构成了这个子社区。

**图7 社区发现的层次树**

* 1. 算法稳健性分析——基于不同算法的第一层划分结果

前文的分析分析结果已经支持SCORE、SCOREq、RSC是较优的社区发现算法，在前文的分析中使用的算法为SCOREq2。为了检验社区发现结果的稳健性，对三种算法进行比较。（其中范数q=1,2时结果类似便不再进行比较）为了简便考虑，本文仅对比进行第一层社区发现的结果。进行两组两两比较：SCOREq2-SCORE和SCOREq2-RSC，以检验社区发现结果的相似性。结果如表13、14所示，行代表使用SCOREq2或RSC方法划分出的第一层社区，列代表使用SCORE算法划分出的第一层社区。计算前者在后者社区中的节点比例，从而比较SCOREq2与RSC方法社区发现的异同。（小于10%的数字被省略）

在表13、14的最后一列，我们指出，如果一个社区和SCOREq2社区中的交集占据了该社区中60%以上的节点，则这个社区可以与SCOREq2中的社区对应（60%可更改为不同的阈值）。

**表13 应用SCORE算法与SCOREq2算法社区发现结果的比较（*K*=6）**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 社区序号 | | | SCOREq2 | | | | | | 主要部分  （>60%） |
| C1 | C2 | C3 | C4 | C5 | C6 |
| 1331 | 202 | 477 | 673 | 1436 | 264 |
| SCORE | C1 | 1090 |  |  |  | **60.37%** | 37.06% |  | C4 |
| C2 | 267 | 21.72% | **75.66%** |  |  |  |  | C2 |
| C3 | 549 |  |  | **84.15%** |  | 10.02% |  | C3 |
| C4 | 625 |  |  |  |  | 50.24% | 42.24% | - |
| C5 | 1388 | **86.17%** |  |  |  | 13.83% |  | C1 |
| C6 | 464 |  |  |  |  | **100.00%** |  | C5 |

对于SCORE算法，在6个社区中，有5个社区的大多数位于SCOREq2社区中，剩下的C4社区被较均匀地分在SCOREq2算法中的C5和C6社区中，说明两种算法的第一层社区划分这表明这两种算法社区发现结果整体上相似性较高。

**表14 SCOREq2算法与RSC算法社区发现结果的比较（*K*=6）**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 社区序号 | | | SCOREq2 | | | | | | 主要部分  （>60%） |
| C1 | C2 | C3 | C4 | C5 | C6 |
| 1331 | 202 | 477 | 673 | 1436 | 264 |
| RSC | C1 | 473 | **92.60%** |  |  |  |  |  | C1 |
| C2 | 148 |  |  | 14.86% | **71.62%** |  |  | C4 |
| C3 | 383 | 46.74% |  |  | 44.13% |  |  | - |
| C4 | 399 |  |  | **80.95%** |  |  |  | C3 |
| C5 | 2634 | 24.45% |  |  | 14.24% | 43.58% |  | - |
| C6 | 346 | 10.98% |  |  | 4.62% | 52.02% | 28.90% | - |

对于RSC算法，在6个社区中，有三个社区的节点大多数位于SCOREq2社区中，另外的三个社区分布在SCOREq2算法的2至3个社区之中。由此说明SCOREq2算法和RSC算法的社区发现结果较好，但相较SCORE差异稍大。

这两种算法社区发现的结果具有一定的相似性：例如SCORE算法中C6社区在SCOREq2算法中C5社区占比100%，RSC算法中C1社区在SCOREq2算法中C1社区占比92.60%。

总体上，可以认为使用SCOREq2算法对合作者网络进行社区发现的结果是较为稳健和合理的。

1. 结论与总结

大规模网络数据中，常常存在较大的节点度异质性，网络稀疏且存在噪声，这位社区发现问题带来了一定困难。本研究中，着重寻找解决此类问题的方法。其中谱聚类方法是一种性质优良的社区发现方法，且在拟合具有度异质性的网络模型是具有一致性。

本文首先对于具有度异质性的网络数据模型——度异质性修正的随机块模型DCBM进行分析，并对谱聚类算法进行了较为细致的梳理。分析了多种对于解决异质性、稀疏、噪声有助益的谱聚类算法修正方法，并利用模拟数据和真实有标签的数据集验证了其优良性质。发现相较于传统的谱聚类方法，修正后的谱聚类算法SCORE，SCOREq，RSC，SCORE+能较好地解决网络的异质性、稀疏性为社区发现带来的困扰，其中SCORE+算法对于解决噪声问题有出色表现。

接下来，对于真实数据集中的社区发现问题，本文基于统计学家合作者网络实现了一个有效的分析范式。首先对于网络的基本性质进行讨论，发现该网络中存在严重的节点度异质性且网络稀疏，从而考虑选择在度异质、稀疏网络中有较好表现的SCORE，SCOREq2，RSC算法。接下来，以SCOREq2算法为例较为详细地讨论了层次的社区发现思路，为社区发现的结果增加更多可解释性，并考虑使用碎石图、模块度、社区内外连边情况以辅助社区数*K*的选择，最终获得了具有4层的层次社区发现树，并通过与SCORE和RSC算法进行比较，说明了社区划分的结果是较为稳健的。

本文提供了一个对统计学合作者网络的社区发现层次树，并基于研究主题、地域、学术谱系提供了一定解释。本文的研究对社会科学和现实生活都有积极意义，该结果可以描述和可视化研究人员的研究概况，并辅助决策。本文的研究还为初级的研究人员选择研究主题、寻找参考资料和建立学术连结提供了一定指导。

参考文献

Lei, J. and A. Rinaldo. Consistency of spectral clustering in stochastic block models[J]. Annals of Statistics, 2015, 43 (1): 215–237.

Bickel P J, Chen A. A nonparametric view of network models and Newman–Girvan and other modularities[J]. Proceedings of the National Academy of Sciences, 2009, 106(50): 21068-21073.

1. Li Y, He K, Bindel D, et al. Uncovering the small community structure in large networks: A local spectral approach[C]. Proceedings of the 24th International Conference on World Wide Web. 2015: 658-668.

Holland P W, Laskey K B, Leinhardt S. Stochastic blockmodels: First steps[J]. Social Networks, 1983, 5(2): 109-137.

1. Abbe E, Sandon C. Proof of the achievability conjectures for the general stochastic block model[J]. Communications on Pure and Applied Mathematics, 2018, 71(7): 1334-1406.
2. 刘学艳. 面向复杂网络分析的随机块模型研究[D]. 吉林: 吉林大学, 2021.

Goldenberg A, Zheng A X, Fienberg S E, et al. A survey of statistical network models[J]. Foundations and Trends in Machine Learning, 2010, 2(2): 129-233.

1. Jin J. Fast community detection by SCORE[J]. The Annals of Statistics, 2015, 43(1): 57-89.
2. Levin K D, Roosta F, Tang M, et al. Limit theorems for out-of-sample extensions of the adjacency and Laplacian spectral embeddings[J]. Journal of Machine Learning Research, 2021, 22: 1-59.
3. Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416.

Gao, C., Z. Ma, A. Y. Zhang, H. H. Zhou. Community detection in degree-corrected block models[J]. Annals of Statistics, 2018, 46 (5): 2153–2185.

Jin J, Ke Z T, Luo S. Improvements on SCORE, especially for weak signals[J]. Sankhya A, 2021: 1-36.

Qin, T. K. Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel[J]. [Advances in Neural Information Processing Systems](https://www.researchgate.net/journal/Advances-in-Neural-Information-Processing-Systems-1049-5258" \t "_blank), 2013: 3120–3128.

Ke, Z. T., M. Wang. A new SVD approach to optimal topic estimation[J]. Annals of Statistics, 2017: 1-58.

Ji, P., J. Jin, Z. T. Ke, W. Li. Co-citation and co-authorship networks of statisticians(with discussion)[J]. [Journal of Business & Economic Statistics](https://www.tandfonline.com/toc/ubes20/current" \t "_blank), 2021: 1-61.

Jin J, Ke Z T, Luo S. Optimal adaptivity of signed-polygon statistics for network testing[J]. The Annals of Statistics, 2021, 49(6): 3408-3433.

黄丹阳. 大规模网络数据分析与空间自回归模型[M]. 北京: 科学出版社, 2022.2.

Amini, A. A., A. Chen, P. J. Bickel, E. Levina. Pseudo-likelihood methods for community detection in large sparse networks[J]. The Annals of Statistics, 2013, 41 (4):2097–2122.

1. Ji, P., J. Jin. Coauthorship and citation networks for statisticians (with discussion)[J]. Annals of Applied Statistics, 2016, 10 (4):1779–1812.

1. 数据网址：http://zke.fas.harvard.edu/MADStat.html [↑](#footnote-ref-0)