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 detection, 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 detection of communities in sparse networks with degree heterogeneity. We utilize both synthetic and real datasets for community detection 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 detection in a statistical co-authorship network that is both heterogeneous and sparse, we propose an effective analytical paradigm. We adopt a hierarchical community detection approach, employing the SgnQ statistic as a hypothesis testing criterion for stopping divisions. The SCOREq algorithm is applied for community detection 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 detection 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 Detection, Spectral Clustering, DCBM Model, Statistical Co-authorship Network

网络数据社区发现谱聚类算法及其应用研究

摘要：大多数网络数据都呈现出社区结构，发现这些结构有助于挖掘社区内部的潜在信息。谱聚类算法是解决社区社区发现问题的经典方法，以其准确性、稳定性、高效性和获得全局最优解的能力而著称。然而，传统的谱聚类算法对于大型稀疏网络数据可能效果不佳。本文通过模拟具有节点度异质性的稀疏网络中的社区发现来解决这一局限性。我们利用生成数据集和真实数据集进行社区发现，并提出了改进的谱聚类算法，即 SCORE、SCORE+、SCOREq 和 RSC。这些算法的性能优于传统的谱聚类算法，例如 oPCA 和 nPCA，尤其是在具有节点度异质性的稀疏网络中，为度校正块模型 (DCBM) 提供了更好的解释。针对社区发现应用于具有异质性和稀疏性的统计学家合作网络，我们提出了一种有效的分析范式。我们采用分层社区发现方法，以 SgnQ 统计量作为停止划分的假设检验准则。在每一层，我们都应用 SCOREq 算法进行社区发现。我们使用树状图、模块度分析以及社区内/社区间边分析等可视化工具来辅助确定最佳社区数量（K），最终得到一个四层的分层社区发现树。随后，我们分析了每个已识别社区中的中心性排名前两位作者的研究方向，并赋予社区有意义的名称以解释其重要性。最后，通过与 SCORE 和 RSC 算法的比较分析，我们观察到结果具有一定的相似性，表明社区划分结果的稳健性。

关键词：社区发现，谱聚类，DCBM 模型，统计学家合作网络

一 引言

近年来，随着海量数据存储和处理能力的飞速发展，网络数据已成为统计学、数学、信息科学、物理学和生物学等诸多领域的研究热点。网络数据将自然界中的实体抽象为节点，并将这些实体之间的关系表示为边。

许多研究表明，网络数据通常呈现出社区结构，其中同一社区内的节点连接紧密，而社区之间的连接则相对稀疏。发现这些社区能够简化研究过程，并帮助研究人员挖掘潜在的有用信息。社区发现是一个典型的聚类问题，其目标是将研究网络中的每个节点映射到至少一个社区。尽管以往的研究提出了各种社区发现算法，例如层次聚类和模块度优化，但在现实世界的网络中发现社区仍然面临挑战。

现实世界网络社区发现的两大主要难点在于：一是节点数量庞大，且连接稀疏、异构，这使得网络结构复杂，对现有算法构成挑战；二是网络数据中缺乏真实的社区标签，使得社区划分结果的准确评估和比较变得复杂。

本研究聚焦于基于统计推断的网络特征描述方法，旨在将生成模型拟合到网络数据，从而编码社区结构。谱聚类算法已被证明能够有效拟合随机块模型，并在聚类误差率方面展现出收敛特性。

本文分析了经典的谱聚类算法，探讨了它们的扩展，并讨论了各种变体和改进方法。通过模拟和真实数据分析，本研究考察了这些算法在不同数据特征下的性能，为在真实网络中选择社区发现算法提供了参考。

此外，本文还探讨了社区发现的实际应用，尤其是在统计学家合作网络中发现社区的背景下。该网络将作者表示为节点，将作者之间的合作关系表示为边。分析该网络的社区结构有助于理解统计学家之间的合作模式、特征和研究主题，从而为文献研究领域做出贡献。鉴于该网络规模庞大、稀疏且节点度异质性，

本研究考虑了前几节的分析结果，以指导算法选择，并从多个角度评估社区发现结果的合理性和可解释性。

二 基本概念

对于大规模网络结构数据，显著的节点度异质性、稀疏的网络结构以及关键信息易受噪声影响是常见的挑战。为了探究网络中的这些特征，本文考虑以下指标：

1) 度异质性

节点的度表征网络中单个节点的属性，而网络的度表征网络的整体属性。节点的度高表明其在网络中具有广泛的连接，反映了其影响力。在度分布相对均匀的网络中，节点表现出同质性，节点之间的关系也相对平缓。相反，在度分布不均匀的网络中，节点之间的异质性更强[17]。在大规模网络数据中，节点度通常遵循幂律分布，表明存在显著的异质性。基于 DCBM 模型，我们可以进一步定义节点度异质性的指标：

$$\alpha(\theta) = (\theta_{min}/\theta_{max}) \cdot \left(\|\theta\|/\sqrt{\theta_{max}\|\theta\|_1}\right) \quad \in \quad (0,1] \tag{3.3.1}$$

$\alpha(\theta)$越小，度异质性越严重。当$\theta_{max} \asymp \theta_{min}$时，$\alpha(\theta)$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，$\alpha(\theta)$接近于 0。

2) 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$\log(|V|)$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：

$$\|\theta\| \tag{3.3.2}$$

在该模型下，其取值范围为$\left(1, \sqrt{n}\right)$

3) 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：

$$\lambda_K/\sqrt{\lambda_1} \tag{3.3.3}$$

DCBM 模型下，λ_k是邻接矩阵Ω的第 K 大的特征值。如果θ, P, π依赖于 n，则 DCBM 足以对弱信号情况进行建模，其中$|\lambda_k|$可能远小于$|\lambda_1|$，$1 < k \leq K$。具体解释为：

（1）$\sqrt{\lambda_1}$视为稀疏性水平和噪声水平（即噪声矩阵 W 的谱范数）

（2）$|\lambda_k|$视为信号强度，因此$|\lambda_k|/\sqrt{\lambda_1}$是信噪比

4) 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：

$$s_n = \alpha(\theta) \cdot \lambda_K/\sqrt{\lambda_1} \tag{3.3.4}$$

三 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$N=(V,E)$，V 为节点集合，E 为边的集合，节点来自 K 个社区，即：$V=V^{(1)}\cup V^{(2)}\ldots\cup V^{(K)}$，设$A$为 N 的$n\times n$大小的邻接矩阵，

无向无权网络$N=(V,E)$中有 K 个社区，则 DCBM 模型为：

$$A=E(A)+W \tag{4.2.1}$$

$$E(A)=\Omega-diag(\Omega) \tag{4.2.2}$$

即：

$$A=\Omega-diag(\Omega)+W=\text{“主要信息”}+\text{“次要信息”}+\text{“噪音”} \tag{4.2.3}$$

$$P(A(i,j)=1)=\Omega(i,j)=\theta(i)\theta(j)P(k,l),\text{若}i\in V^k,j\in V^l \tag{4.2.4}$$

引入指示社区$K\times 1$的向量π_i：$\pi_i(k)=1$，若$i\in V^k$，其余元素为 0。则有

$$P(A(i,j)=1)=\Omega(i,j)=\theta(i)\cdot\theta(j)\times\pi_i'P\pi_j \tag{4.2.5}$$

令$\Theta=diag\big(\theta(1),\theta(2),\ldots,\theta(n)\big)$，$\Pi=[\pi_1,\pi_2,\ldots,\pi_n]'$，则矩阵形式为

$$\Omega=\Theta\Pi P\Pi'\Theta \tag{4.2.6}$$

即：

$$\Omega=\begin{pmatrix}\theta_1 & & \\ & \ddots & \\ & & \theta_n\end{pmatrix}\begin{pmatrix}\pi_1' \\ \vdots \\ \pi_n'\end{pmatrix}\begin{pmatrix}\mathrm{p}_{11} & \cdots & \mathrm{p}_{1k} \\ \vdots & \ddots & \vdots \\ \mathrm{p}_{k1} & \cdots & \mathrm{p}_{kk}\end{pmatrix}(\pi_1 \quad \ldots \quad \pi_n)\begin{pmatrix}\theta_1 & & \\ & \ddots & \\ & & \theta_n\end{pmatrix} \tag{4.2.7}$$

其中 P为描述社区结构的$K\times K$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，$\max\limits_{1\le i,j\le K}P(i,j)=1$，$0\le\theta_{min}\le\theta_{max}\le g_0$，$g_0$为$(0,1)$区间的常数。$E(A)$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：Ω元素$\Omega(i,j)$对应非矩阵形式的节点 i 与 j 的连边概率，由于$A(i,j)$取值为 0 或 1，自然$E(A(i,j))=\Omega(i,j)$，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$E(A)=\Omega-diag(\Omega)$。而 W 为中心伯努利分布，

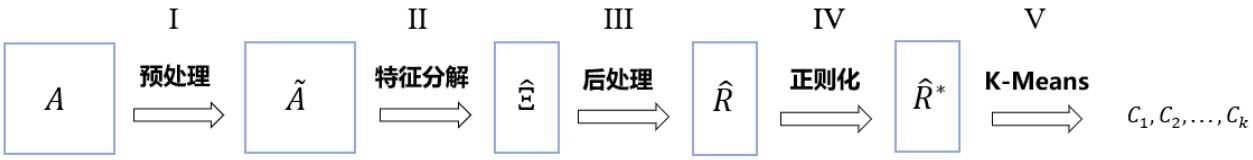
$$\begin{cases}W(i,j)=1-\Omega(i,j), & A(i,j)=1\\ W(i,j)=-\Omega(i,j), & A(i,j)=0\end{cases} \tag{4.2.8}$$

即

$$\begin{cases}P\big(W(i,j)=1-\Omega(i,j)\big)=\Omega(i,j)\\ P\big(W(i,j)=-\Omega(i,j)\big)=1-\Omega(i,j)\end{cases} \tag{4.2.9}$$

2. 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$\tilde{A}$。

2) 特征分解（可以视作降低噪声）

对矩阵$\tilde{A}$特征分解，获得前 K 个最大特征值对应特征向量$\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$，对应的特征值

$$记\hat{\Xi} = [\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K] = [x_1', x_2', \ldots, x_n']'$$

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$(\hat{\lambda}_K^* - \hat{\lambda}_{K+1}^*)/\hat{\lambda}_K^* \le t$时，设$M = K + 1$，否则$M = K$。

3) 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$M(x)$，满足$M(ax) = M(x)$。

对于$\hat{\Xi}$的每一行x_i，应用这种映射，获得矩阵$\hat{R}$。

4) 正则化

对于一个给定的阈值 $T\,(T > 0)$，设$\hat{R}^*$是$\hat{R}$的正则化版本，其中$\hat{R}^*(i,k) = \mathrm{sgn}(\hat{R}(i,k) \cdot \min\{T, |\hat{R}(i,k)|\}$，$1 \le k \le K-1, 1 \le i \le n$。

5) 聚类

基于$n \times (K-1)$矩阵$\hat{R}^*$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$\tilde{A}$	后处理$\hat{R}$	超参数取值
OPCA (传统谱聚类)	A	$\hat{R}$(无)	
nPCA （拉普拉斯变换谱聚类）	$\left(D+\delta\bar{d}I_n\right)^{-1/2}A\left(D+\delta\bar{d}I_n\right)^{-1/2}$	$\hat{R}$(无)	$\delta=0$ 还可尝试其他取值 如 0.05
SCORE （度异质性修正谱聚类）	A	$M(x)=x/x(1)$ $x(1)$是x的第一项， 最后删除第一项	
SCOREq	A	$M(x)=x/\|\|x\|\|^q$	$q=1,2$
RSC[13]	$\left(D+\delta\bar{d}I_n\right)^{-1/2}A\left(D+\delta\bar{d}I_n\right)^{-1/2}$	$M(x)=x/x(1)$ $x(1)$是x的第一项	$\delta=0.05$ 还可尝试其他取值
SCORE+	$(D+\delta d_{max}I_n)^{-1/2}A(D+\delta d_{max}I_n)^{-1/2}$	加权：对于第 i 行， $M(x)=\hat{\lambda}_i x/\hat{\lambda}_1 x(1)$	$t=0.10$ $\delta\in[0.05,0.1]$

四　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$P_{K\times K}$、度异质向量$\theta_{1\times n}$、标签向量$l_{n\times 1}$

实验步骤为：

(1) 生成主要信息矩阵$\Omega_{n\times n}$，满足

$$\Omega(i,j) = \theta(i) \times P\big(l(i), l(j)\big) \times \theta(j) \tag{6.2.1}$$

(2) 生成噪音矩阵W

对角线为 0 的对称矩阵，上对角矩阵为以$\Omega(i,j)$为参数的中心伯努利分布，即$j > i$时：

$$P\big(W(i,j) = 1 - \Omega(i,j)\big) = \Omega(i,j), \tag{6.2.2}$$

$$P\big(W(i,j) = -\Omega(i,j)\big) = 1 - \Omega(i,j) \tag{6.2.3}$$

(3) 生成$N(V,E)$的邻接矩阵$\tilde{A}$

$$\tilde{A} = \Omega - diag(\Omega) + W \tag{6.2.4}$$

即

$$P\big(\tilde{A}(i,j) = 1\big) = \Omega - diag(\Omega) \tag{6.2.5}$$

(4) 求 N 的最大连通分量：$N_0(V_0, E_0)$，n_0为N_0的大小。

(5) 应用谱聚类算法进行社区发现，得到每个节点的社区标签。

(6) 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$P_{K\times K} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$
- 度异质向量：$\theta_{1\times n} = (0.2, 0.2, \dots, 0.2)$
- 标签向量：$l_{n\times 1}: (l_i - 1) \sim Bernoulli(1/2)$

在这种参数设定下，其网络结构的指标为：$\alpha(\theta) = 1, ||\theta|| = 6.325, SNR = 1.820, s_n = 1.820$。注意，度异质性越弱，$\alpha(\theta)$越大，当没有度异质性时，$\alpha(\theta) = 1$；噪声越强，SNR 越小；当度异质性弱，信号强时，s_n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。

2)　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：$P_{K\times K} = \begin{pmatrix} 1 & 0.4 & 0.05 \\ 0.4 & 1 & 0.4 \\ 0.05 & 0.4 & 1 \end{pmatrix}$
- 度异质向量：$\theta(i) = 0.015 + 0.785 \times (i/n)^2$
- 标签向量：$l_{n\times 1}: l_i = 1,2,3\ , P(l_i = 1) = P(l_i = 2) = P(l_i = 3) = \frac{1}{3}$

在这种参数设定下，其网络结构的指标为：$\alpha(\theta) = 0.014, ||\theta|| = 14.046, SNR = 2.909, s_n = 0.042$。

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。

3)　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：$P_{K\times K} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$
- 度异质向量：比较三种形式的度异质性设定形式如下：

$$\text{a.}\ \theta(i) = c_0 + (c_0 - d_0) \times i/n$$

$$\text{b.}\ \theta(i) = c_0 + (c0 - d_0) \times (i/n)^2$$

$$\text{c.}\ \theta(i) = c_0 1\{i \le n/2\} + (c_0 - d_0) 1\{i > n/2\}$$

$$\text{其中} (c_0, d_0) = (0.015, -0.77)$$

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：$l_{n\times 1}: (l_i - 1) \sim Bernoulli(1/2)$

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	0.015(0.000)	0.058(0.000)	0.013(0.000)	
b	0.074(0.000)	0.250(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：$\alpha(\theta)=[0.016,0.014,0.019],||\theta||=[14.8,11.5.17.6],SNR=[4.25,3.29,5.07],s_n=[0.068,0.048,0.099]$。

可以看到此种模型设定的度异质性无法仅用$\alpha(\theta)$或$\theta_{max}/\theta_{min}$来度量，显然三种设定下均有$\theta_{max}=2c_0-d_0,\theta_{min}=c_0$，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，s_n指标也失效。

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。

4)　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：$P_{K\times K}=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}$
- 度异质向量：先生成$log(\theta(i))\sim N(0,\sigma^2),\sigma=0.2\times[1,\sqrt{2},\sqrt{3},2,\sqrt{5}]$,然后归一化：$\theta=0.9\times\theta/\theta_{max}$
- 标签向量：$l(i)=1\{i\le n/4\}+2\times 1\{n/4<i\le n\}$

在不同参数设定下($\sigma=0.2\times[1,\sqrt{2},\sqrt{3},2,\sqrt{5}]$)，其网络结构的指标为：

$$\alpha(\theta)=[0.180,0.091,0.073,0.050,0.035],||\theta||=[13.8,11.0,12.0,10.0,10.0],$$

$$SNR=[2.57,2.01,2.20,2.04,1.88],s_n=[0.464,0.182,0.160,0.102,0.067]。$$

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σ变化造成，因此$\alpha(\theta)$能够很好地度量度异质性程度。

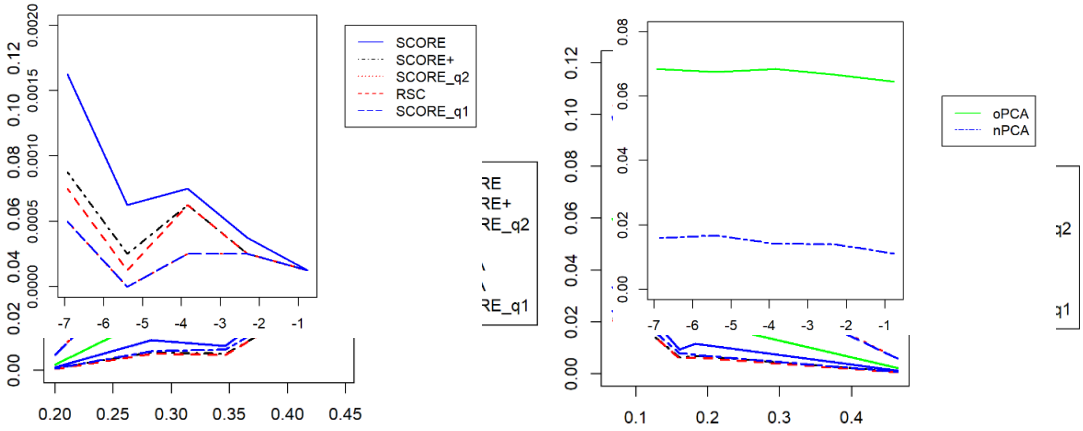


图 1 实验 3.2 中不同算法的节点分配错误率

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

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s_n是算法性能的一个预示，s_n越大(对应σ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。

5) 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：$P_{K\times K} = \begin{pmatrix} 3 & 0.5 \\ 0.5 & 1 \end{pmatrix}$
- 度异质向量：首先生成

$$\theta\big(1:(n/4)\big) = c_0 + (c_0 - d_0) \times 4i/n,$$
$$\theta\big((1+n/4):n\big) = c_0 + (c_0 - d_0) \times (4i/3n)^2,$$
$$\text{其中} c_0 = 0.015, d_0 = -0.77 \times [1,3,5,7,9]$$

然后对θ值除以l_1范数$|\theta|$归一化，并令$\theta = 0.8 \times \theta/\theta_{max}$

- 标签向量：$l(i) = 1\{i \le n/4\} + 2 \times 1\{n/4 < i \le n\}$

在不同参数设定下$(d_0 = -0.77 \times [1,3,5,7,9])$，其网络结构的指标为：$\alpha(\theta) = [0.0098,0.0037,0.0024,0.0019,0.0016], ||\theta|| = [24.7,24.6,24.6,24.6,24.6], SNR = [9.97,9.95,9.95,9.95,9.94], s_n = [0.097,0.037,0.024,0.019,0.016]$。

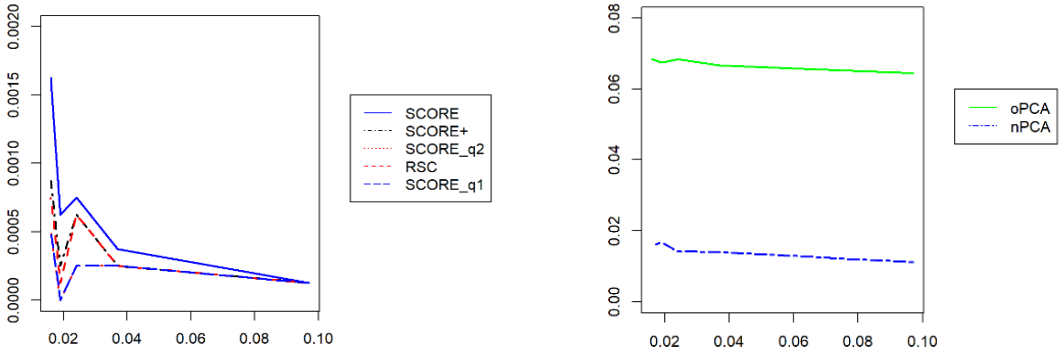
在此种模型设定下，由于度异质性参数的均由线性函数产生，其异质性的不同是由参数d_0变化造成,因此$\alpha(\theta)$能够很好地度量度异质性程度。

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d_0；下图：y 轴为节点分配错误率，x 轴为信噪比s_n；

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。

五 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]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）；

[1] 数据网址：http://zke.fas.harvard.edu/MADStat.html

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：

2. 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。

2) 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

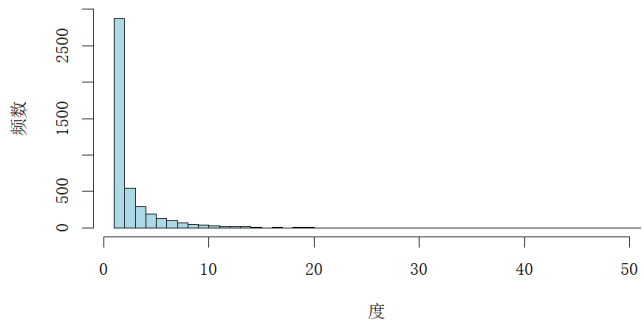


表 7 展示了统计学家合作者网络中度最高的 10 个作者。排名第一的作者是 Peter Hall，它的度为 51，代表其合作者共有 51 个。

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

3) 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

4) 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

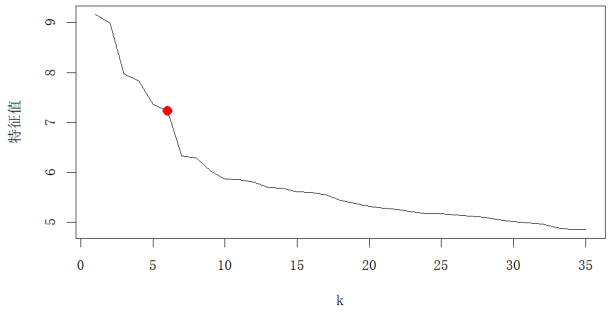
排名	介数中心性	统计学家姓名

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

5) 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$|\hat{\lambda}_K - \hat{\lambda}_{K+1}|/|\hat{\lambda}_K|$取值较大。在图像上表示，可以视作是一个明显的拐点，$\hat{\lambda}_K$取值较大，而$\hat{\lambda}_{K+1}$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$|\hat{\lambda}_K - \hat{\lambda}_{K+1}|/|\hat{\lambda}_K|$较大的 K 值。（$\hat{\lambda}_K$显著大于$\hat{\lambda}_{K+1}$但与$\hat{\lambda}_{K-1}$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）。

3. 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K_0个子网络，其中$K_0 < K$，K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（H_0）为子网络中只有一个社区($K_0 = 1$)，备择假设（H_1）为子网络中有多个社区($K_0 > 1$)。

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

A 是子网络对应的邻接矩阵，令$\hat{\eta} = \frac{1}{\sqrt{\mathbf{1}'_n A \mathbf{1}_n}} A\mathbf{1}_n \in \mathbb{R}^n$且$A^* = A - \hat{\eta}\hat{\eta}' \in \mathbb{R}^{n,n}$

$$\psi_n = \frac{1}{\sqrt{2}}\left(\frac{\sum_{i_1,i_2,i_3,i_4(\text{互不相同})} A^*_{i_1 i_2} A^*_{i_2 i_3} A^*_{i_3 i_4} A^*_{i_4 i_1}}{2\left(||\hat{\eta}||^2 - 1\right)^2} - 1\right). \tag{7.3.1}$$

若原假设成立，在温和条件下，有$\psi_n \to N(0,1)$。因此可以通过$1 - \Phi(\psi_n)$得到近似 p 值，其中Φ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。

（3）当零假设在每个子网络中都被接受时，算法停止。

（4）算法的输出是一个层次树，树每个叶节点对应一个社区。

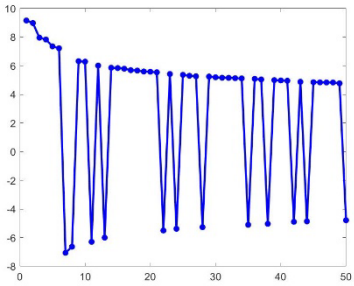
2) *K* 的选择准则

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$K \in \{4,5, \ldots 11\}$。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$A + I_n$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$K \in \{4,5,\ldots,11\}$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）

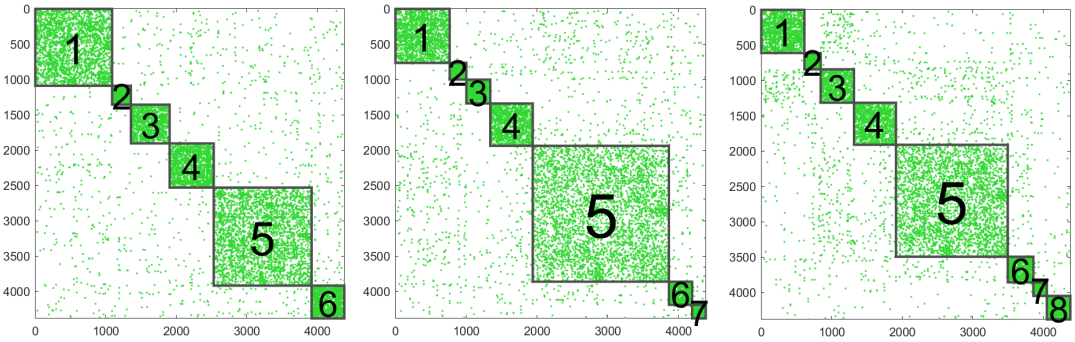
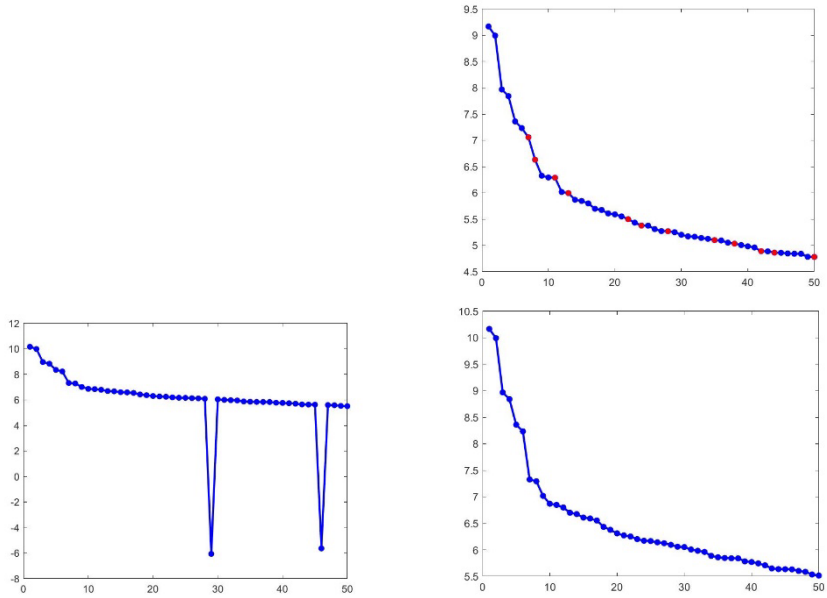


图 6 中每个散点代表一个作者，每个方块代表一个社区，方框内的绿色亮点代表社区内对应的两个节点存在连边，方框外的点代表该点与不同社区存在连边，即社区间有连边。基于本文的分析目的与“无混合成员”的假设，希望示意图中方框外的点尽可能的少，即社区间的连边尽可能地少。直观地，当 K=7，8 时，不同社区间连边相对多于 K=6，并且当 K>8 时也有类似的结果，不再作展示。



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。

4.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）。

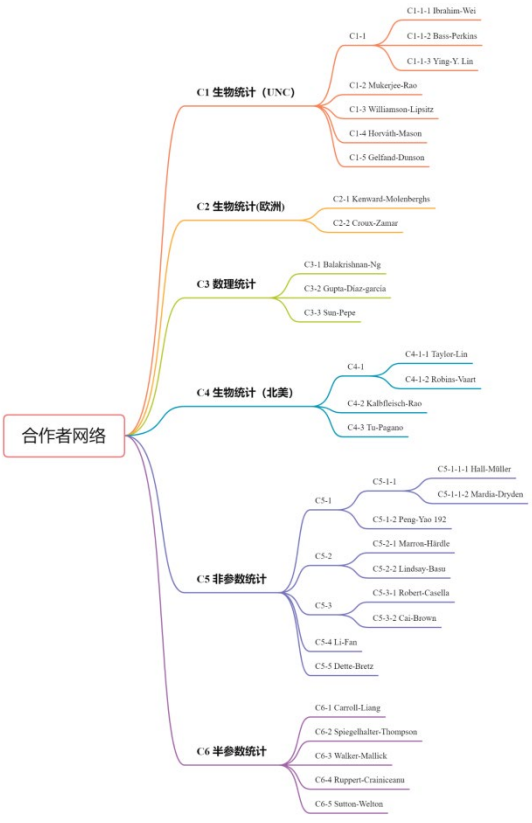
例 2：地理和文化因素

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



5.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。

六 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

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] Jin J. Fast community detection by SCORE[J]. The Annals of Statistics, 2015, 43(1): 57-89.

[9] 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.

[10] Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416.

[11] 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.

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

[13] Qin, T. K. Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel[J]. Advances in Neural Information Processing Systems, 2013: 3120–3128.

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

[15] 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, 2021: 1-61.

[16] 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.

[17] 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.

[18] 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.

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