

网络数据社区发现 谱聚类算法及其应用研究

2020级统计学院 陶雪然

1.1 网络数据社区发现

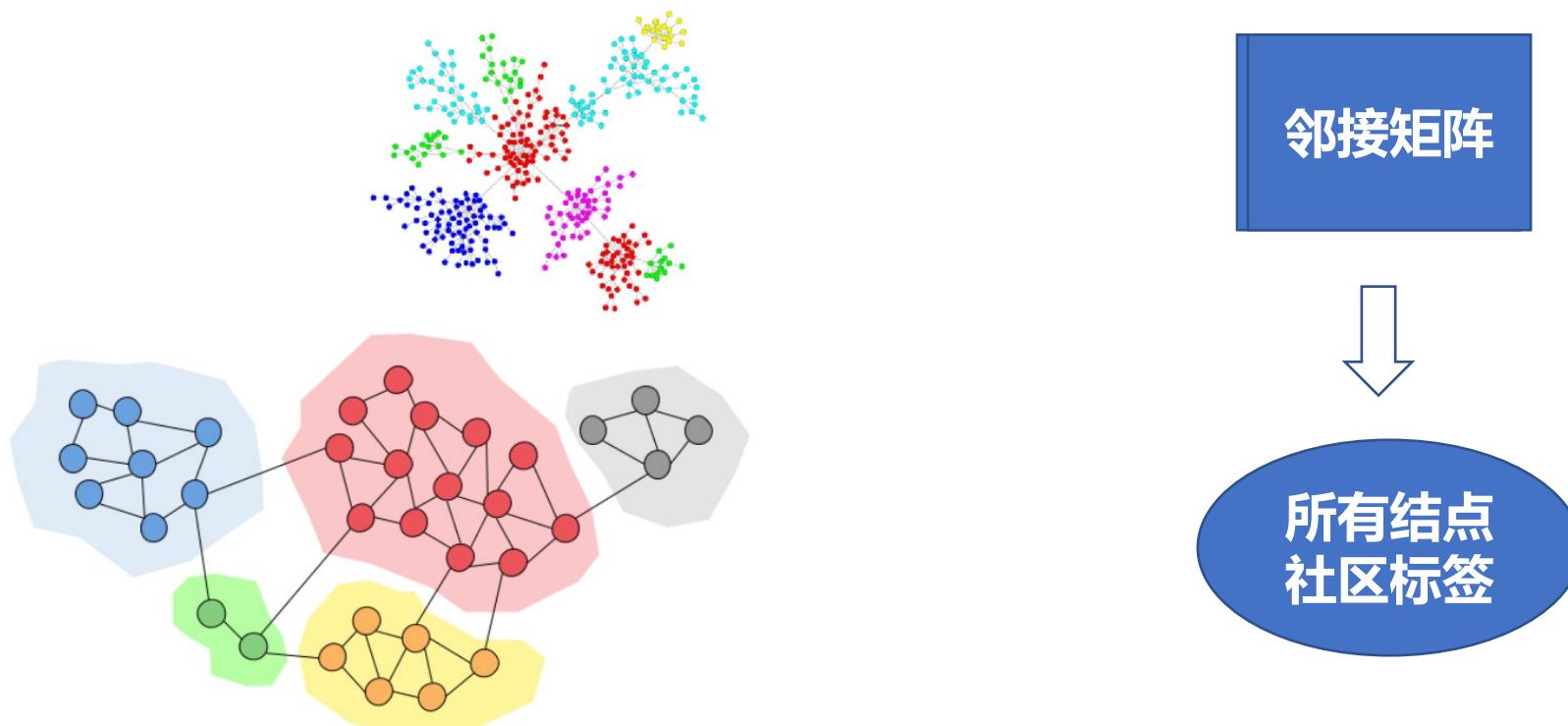
研究背景

- **社区**: 内部连接**紧密**, 外部连接**稀疏**的结构
- **社区发现**: 聚类, 将网络中的节点分配到社区中

研究方法

研究结果

研究结论



1.2 研究现状及问题

研究背景

研究方法

研究结果

研究结论

现有方法：

- 启发式算法：

层次聚类、模块度优化、GN算法、Louvain算法…

- 谱聚类算法：

计算效率高

较好拟合 SBM (随机块模型)，具有解释性

1.2 研究现状及问题

研究背景

研究方法

研究结果

研究结论

- ✓ **真实世界网络 节点异质性大、稀疏、存在噪声，谱聚类算法准确性较低**
——谱聚类算法修正
- ✓ **社区发现是无监督学习问题，难以准确评判与比较社区划分结果的好坏**
——用社区发现结果的可解释性、模块度的大小判断，选择较为稳健的算法

2 研究内容

研究背景

研究内容

模型设定

特征指标

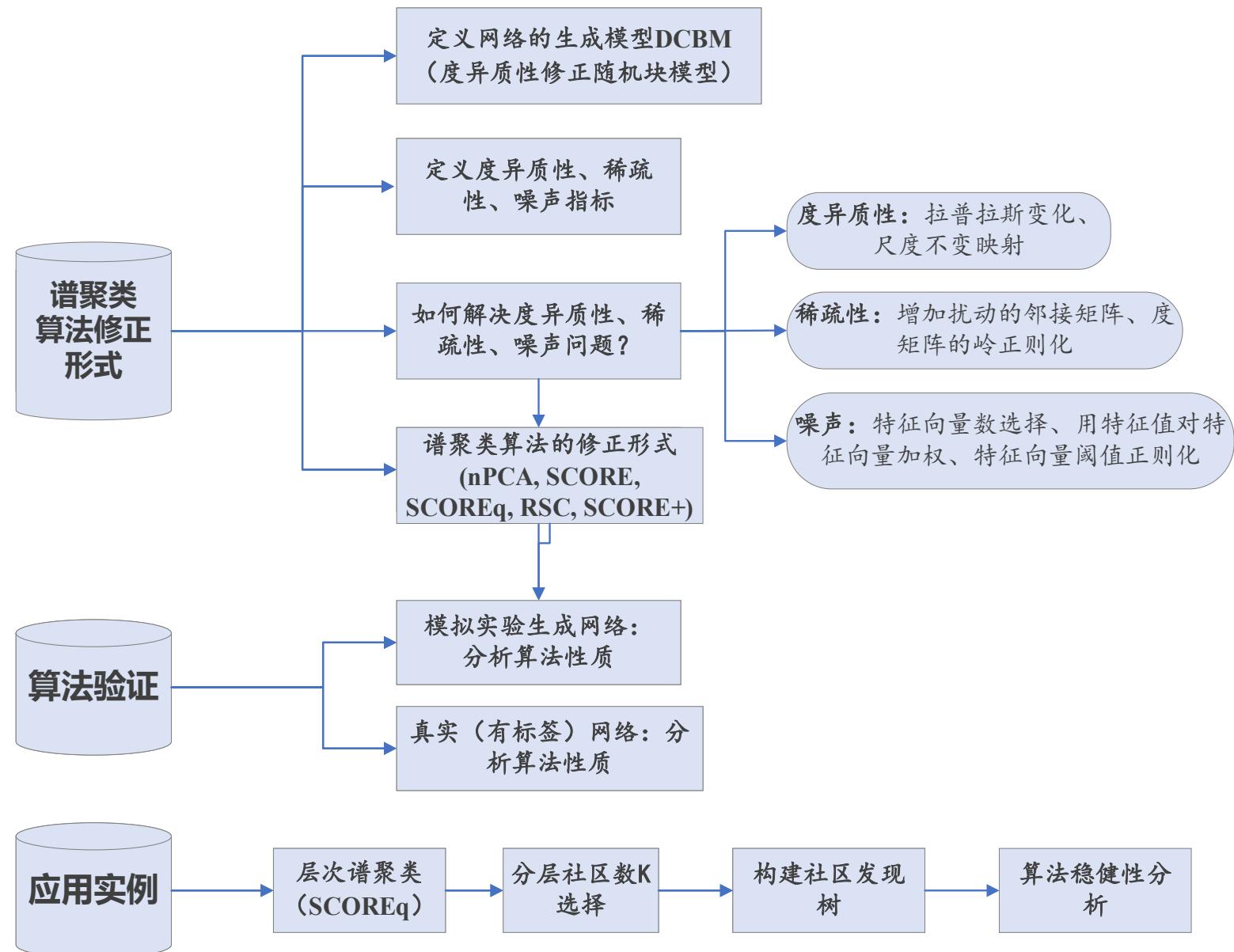
谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论

2.1 网络数据的生成模型



DCBM (度异质修正的随机块模型)

无向无权网络 $N = (V, E)$, 节点来自 K 个社区: $V = V^{(1)} \cup V^{(2)} \dots \cup V^{(K)}$, A 为邻接矩阵:

$$A = E(A) + W$$

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

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

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:

$$P(A(i,j) = 1) = \Omega(i,j) = \theta(i) \cdot \theta(j) \times \pi'_i P \pi_j$$

$$\Omega = \Theta \Pi P \Pi' \Theta$$

$$\Omega = \begin{pmatrix} \theta_1 & & \\ & \ddots & \\ & & \theta_n \end{pmatrix} \begin{pmatrix} \pi'_1 \\ \vdots \\ \pi'_n \end{pmatrix} \begin{pmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{pmatrix} (\pi_1 \quad \dots \quad \pi_n) \begin{pmatrix} \theta_1 & & \\ & \ddots & \\ & & \theta_n \end{pmatrix}$$

2.2 特征指标

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



节点异质性

- 节点度分布 (幂律分布?)
- $\alpha(\theta) = (\theta_{min}/\theta_{max}) \cdot \left(\|\theta\| / \sqrt{\theta_{max} \|\theta\|_1} \right) \in (0,1]$



稀疏性

- 网络密度、平均度
- $\|\theta\| \in (1, \sqrt{n})$



噪声

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



综合指标

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

2.3 谱聚类算法及其修正

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论

邻接矩阵
“归一化”变换

对于 $\hat{\Sigma}$ 的每一行 x_i ,
应用尺度不变映射 $M(x)$



$$\begin{aligned}\hat{\Sigma} &= [\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_K] \\ &= [x'_1, x'_2, \dots, x'_n]'\end{aligned}$$

$$\hat{R}^*(i, k) = \operatorname{sgn}(\hat{R}(i, k) \cdot \min\{T, |\hat{R}(i, k)|\})$$

2.3 谱聚类修正-节点度异质性

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



预处理——拉普拉斯变换

$$\tilde{A} = D^{-1/2} A D^{-1/2}$$



后处理——尺度不变映射 $M(ax) = M(x)$

$$\hat{\Xi} = [\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_K] = [x'_1, x'_2, \dots, x'_n]'$$

对前 K 个特征向量构成的 $n \times k$ 矩阵 $\hat{\Xi}$ 的每一行 x_i 做尺度不变映射，获得矩阵 \hat{R} 。

定义一个尺度不变映射 $M: W \rightarrow R^K$, 满足 $M(ax) = M(x)$ 。如：

- $W = \{x \in R^K, x(1) \neq 0\}, M(x) = x/x(1), x(1)$ 是向量 x 的第一个值 —— SCORE
- $W = R^K \setminus \{0\}, M(x) = x/\|x\|^q$, 其中 $q > 0$ 为常数。 —— SCOREq

2.3 谱聚类修正-稀疏性

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



预处理——增加微弱扰动

给所有的节点对添加“弱连接”：

$$\tilde{A} = A + c \cdot 1_n 1_n^T$$



预处理——为拉普拉斯变换的度矩阵D增加扰动

$$\tilde{A} = (D + \delta \bar{d} I_n)^{-1/2} A (D + \delta \bar{d} I_n)^{-1/2}$$

其中 δ 为扰动参数，在该情形下避免了度过小带来的问题。

2.3 谱聚类修正-信号与噪声

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



特征向量数选择(K或K+1)

由于 Ω 的秩为 K , 它只有 K 个非零特征值

- **强信号:** K 个 $|\hat{\lambda}_K - \hat{\lambda}_{K+1}| / |\hat{\lambda}_K|$ 较大: $\hat{\xi}_K$ 和 ξ_k 有高度相关性, 而 $\hat{\xi}_{K+1}$ 和 ξ_K 只有弱相关性。
- **弱信号:** $K+1$ 个 $|\hat{\lambda}_K - \hat{\lambda}_{K+1}| / |\hat{\lambda}_K|$ 较小, $\hat{\xi}_{K+1}$ 可能比 $\hat{\xi}_K$ 和 ξ_K 的关系更紧密。增加一个特征向量做K-Means很重要



使用特征值对特征向量加权

- 对于每一个 $\xi_k(i), 1 \leq i \leq n$, 信噪比随着 k 的增加而减小, 可以降低 ξ_k 的权重。
- 由于用 L_2 范数误差测量的 ξ_k 中的噪声水平近似成正比于 $1/\lambda_k$, 因此加权的合适选择是将每个 ξ_k 都乘以 λ_k 。



阈值T正则化

- 建议 $T = \log(n)$ (如果 n 相对较小, 则 $T = 2\log(n)$)
- 避免了社区发现后特征向量矩阵中极端值的干扰, 使得结果更稳定。

2.3 谱聚类的修正算法

研究背景

研究内容

模型设定

特征指标

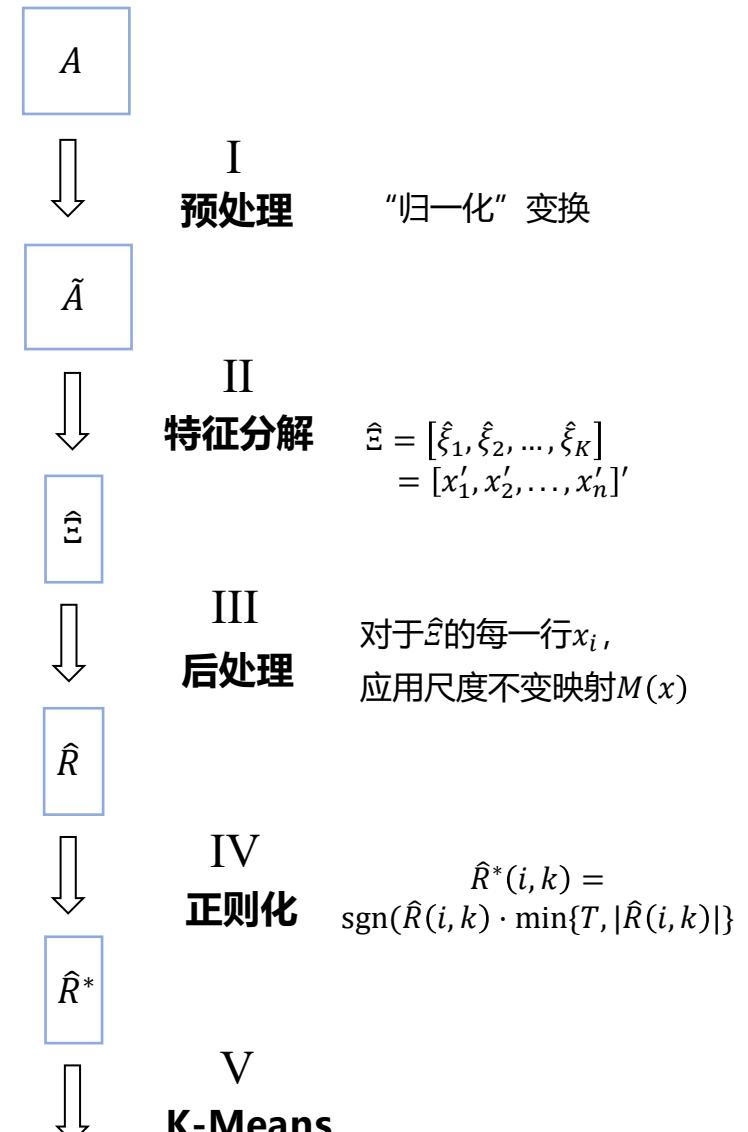
谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



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)	$(D + \delta \bar{d} I_n)^{-1/2} A (D + \delta \bar{d} I_n)^{-1/2}$	\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	$(D + \delta \bar{d} I_n)^{-1/2} A (D + \delta \bar{d} I_n)^{-1/2}$	$M(x) = x/x(1)$ $x(1)$ 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>i</i> 行, $M(x) = \hat{\lambda}_i x / \hat{\lambda}_1 x(1)$	$t = 0.10$ $\delta \in [0.05, 0.1]$

2.4 算法性能讨论-模拟实验

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



针对现实网络中存在节点度异质性、信号弱噪声大、网络规模大的问题，
比较谱聚类算法oPCA、nPCA、SCORE、SCORE+、SCOREq、RSC的性能

参数：大小 n 、社区数 K 、实验重复次数 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：

(1) 生成主要信息矩阵 $\Omega_{n \times n}$ ，满足：

$$\Omega(i, j) = \theta(i) \times P(l(i), l(j)) \times \theta(j)$$

(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，

$$P(W(i, j) = 1 - \Omega(i, j)) = \Omega(i, j),$$

$$P(W(i, j) = -\Omega(i, j)) = 1 - \Omega(i, j)$$

(3) 生成 $N(V, E)$ 的邻接矩阵 \tilde{A} ：

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

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

(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，得到错误率和方差

2.4 算法性能讨论-模拟实验

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



实验1：无异质性网络

参数：大小 n 、社区数 K 、实验重复次数 rep 、概率矩阵 $P_{K \times K}$ 、度异质向量 $\theta_{1 \times n}$ 、标签向量 $l_{n \times 1}$

$$(n, K, rep) = (1000, 2, 50), P_{K \times K} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \theta_{1 \times n} = (0.2, 0.2, \dots, 0.2), l_{n \times 1}: (l_i - 1) \sim Bernoulli(1/2)$$

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：SCORE、SCORE+、SCOREq2、RSC、SCOREq1、oPCA、nPCA**均能很好的划分社区，错误率均小于0.06。**

2.4 算法性能讨论-模拟实验

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



实验2：异质网络

参数：大小 n 、社区数 K 、实验重复次数 rep 、概率矩阵 $P_{K \times K}$ 、度异质向量 $\theta_{1 \times n}$ 、标签向量 $l_{n \times 1}$

$$(n, K, rep) = (1500, 3, 25), P_{K \times K} = \begin{pmatrix} 1 & 0.4 & 0.05 \\ 0.4 & 1 & 0.4 \\ 0.05 & 0.4 & 1 \end{pmatrix}, \theta(i) = 0.015 + 0.785 \times (i/n)^2,$$

$$l_{n \times 1}: l_i = 1, 2, 3, P(l_i = 1) = P(l_i = 2) = P(l_i = 3) = \frac{1}{3}$$

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：SCORE、SCORE+、SCOREq、RSC均能很好的划分社区，其中SCORE+错误率最小。而oPCA、nPCA表现欠佳。

2.4 算法性能讨论-模拟实验

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



实验3.1：多种异质性网络

参数：度异质向量 $\theta_{1 \times n}$ ：

- a. $\theta(i) = c_0 + (c_0 - d_0) \times i/n$
- b. $\theta(i) = c_0 + (c_0 - d_0) \times (i/n)^2$
- c. $\theta(i) = c_0 \mathbf{1}\{i \leq n/2\} + (c_0 - d_0) \mathbf{1}\{i > n/2\}$
 $(c_0, d_0) = (0.015, -0.77)$

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)	

结论3：随着度异质性的增加，社区划分错误率增加，SCORE、SCORE+、SCOREq、RSC对度异质性有较强的抑制作用，而nPCA对度异质性只有较弱的抑制作用。

2.4 算法性能讨论-模拟实验

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

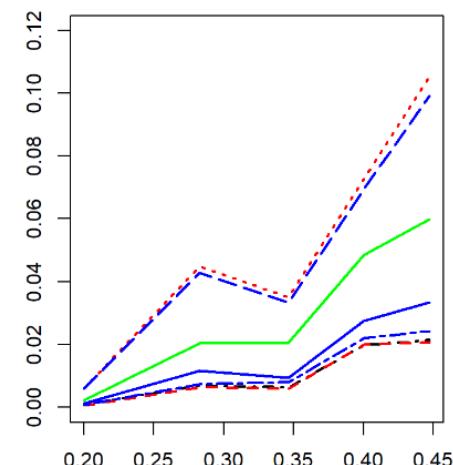
研究结果

研究结论

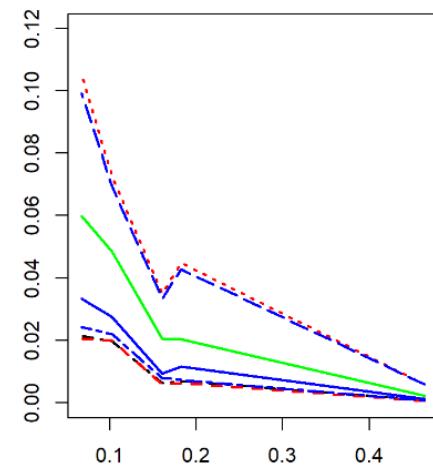


实验3.2：另一种异质网络

参数：度异质向量 $\theta_{1 \times n}$: $\log(\theta(i)) \sim N(0, \sigma^2)$, $\sigma = 0.2 \times [1, \sqrt{2}, \sqrt{3}, 2, \sqrt{5}]$, 然后归一化: $\theta = 0.9 \times \theta / \theta_{max}$
标签向量: $l(i) = 1\{i \leq n/4\} + 2 \times 1\{n/4 < i \leq n\}$



SCORE
SCORE+
SCORE_q2
RSC
oPCA
nPCA
SCORE_q1



SCORE
SCORE+
SCORE_q2
RSC
oPCA
nPCA
SCORE_q1

图左: y 轴为错误率, x 轴为参数 σ ; 图右: y 轴为错误率, x 轴为 S_n

结论3.2: **RSC、nPCA表现较好, SCOREq表现较差。oPCA、nPCA在此种设定下表现较好。**

2.4 算法性能讨论-模拟实验

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

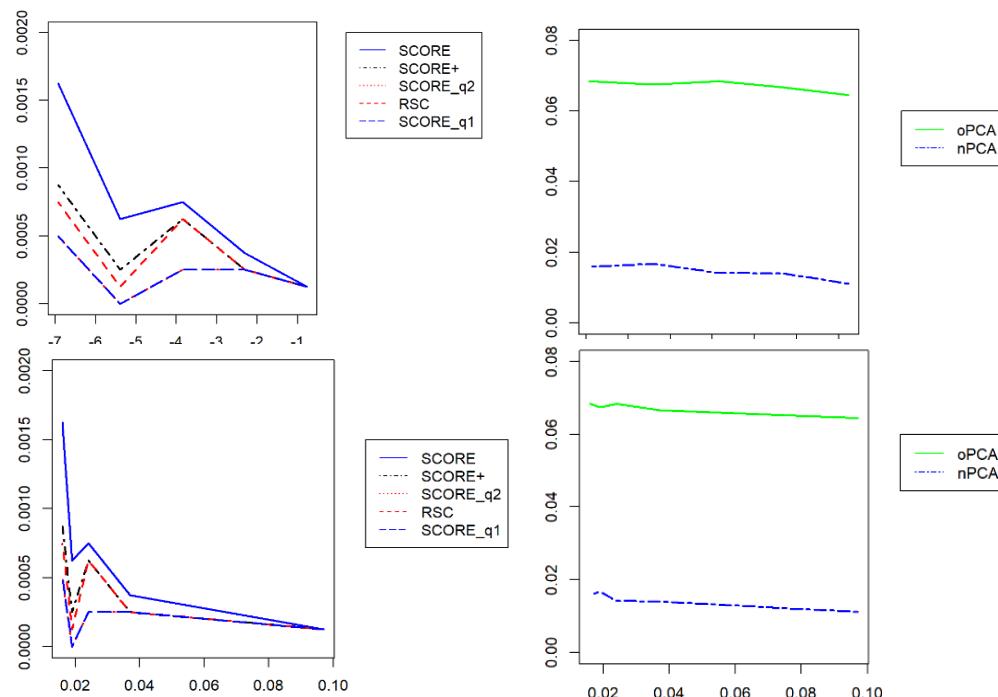
研究结果

研究结论



实验4：大规模、异质网络

参数: $(n, K, rep) = (4000, 2, 25)$, $P_{K \times K} = \begin{pmatrix} 3 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, 其余参数与3.1相同



上图: y 轴为错误率, x 轴为参数 d_0 ; 下图: y 轴为错误率, x 轴为 s_n

结论4: 该大规模异质网络中, **SCOREq表现比SCORE更好, oPCA、nPCA表现较差**

2.4 算法性能讨论-真实数据

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



在不同特征的真实数据集中，比较谱聚类算法oPCA、nPCA、SCORE、SCORE+、SCOREq、RSC的性能

数据集	节点数	社区数	边数	度最小值	度最大值	平均度	网络密度	异质性参数	数据特点
Karate	34	2	78	1	17	4.59	0.2781	0.0398	经典数据
Dolphins	62	2	159	1	12	5.12	0.1682	0.0628	经典数据
UKfaculty	79	3	552	2	39	13.97	0.1787	0.0290	非稀疏
Polbooks	92	2	374	1	24	8.13	0.1476	0.0034	异质性大
Football	110	11	570	7	13	10.36	0.1902	0.4834	社区多、同质性强
Caltech	590	8	12822	1	179	43.36	0.0751	0.0016	度异质性大、噪声大
Simmons	1137	4	24257	1	293	42.67	0.3583	0.0352	度异质性大、噪声大
Polblogs	1222	2	16714	1	351	27.35	0.0448	0.0014	网络大而稀疏

2.4 算法性能讨论-真实数据

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论



真实数据集下，各算法划分错误的情况

数据集	节点数	数据特点	SCORE	SCORE+	OPCA	RSC	nPCA1	nPCA0	SCOREq1	SCOREq2
Karate	34	经典数据	0	1	0	0	0	1	0	0
Dolphins	62	经典数据	0	2	12	1	0	0	1	1
UKfaculty	79	非稀疏	2	2	5	0	1	1	5	6
Polbooks	92	异质性大	1	2	4	3	2	2	3	3
Football	110	社区多、同质性强	5	6	5	5	6	6	5	4
Caltech	590	度异质性大、噪声大	183	98	224	170	204	174	177	178
Simmons	1137	度异质性大、噪声大	268	127	442	244	334	278	255	253
Polblogs	1222	网络大、稀疏	58	51	437	64	380	590	61	64

结论：SCORE、SCORE+、SCOREq、RSC在异质性、稀疏性、噪声等问题时表现比较好，其中SCORE+对具有明显噪声的数据表现较为优良

2.5 应用-统计学合作者网络

研究背景

研究内容

模型设定

特征指标

谱聚类与修正

算法性能讨论

应用实例

研究结果

研究结论

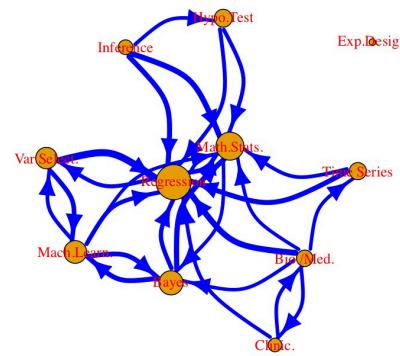


统计学论文合作者网络数据集

MADStat数据集：生物统计、概率论、机器学习....

期刊	作者数	年份	文章数
36	47331	1970-2015	83331

Zheng (Tracy) Ke Department of Statistics
Harvard University



分析网络特征（异质、稀疏、噪声）
结合前文算法验证结果，**选择算法**

2.5 研究内容—应用

研究背景

研究内容

模型设定

特征指标

谱聚类与修正
算法性能讨论

应用实例

研究结果

研究结论



社区发现：使用分层谱聚类，构建社区层次树

- **聚类** 将网络划分为 K_0 个子网络。聚类时，考虑运用性质优良的SCORE、SCOREq、RSC算法。
- **SgnQ假设检验** H_0 : 子网络中只有一个社区($K_0=1$)， H_1 : 子网络中有多个社区($K_0>1$)

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

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

H_0 下， $\psi_n \rightarrow N(0,1)$ ，若无法拒绝原假设， $p > 0.001$ ，或子网络中节点数小于等于250，则停止划分，否则继续划分。



每次社区发现选择社区数K：利用碎石图、模块度、连接结构



分析社区发现结果的可解释性

3.1 统计学合作者网络——网络描述

研究背景

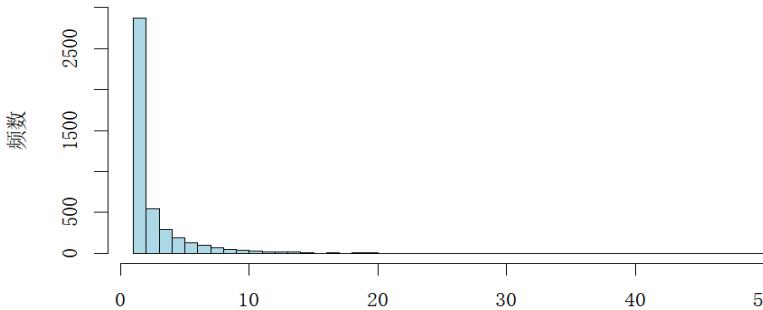
研究内容

研究结果

研究结论



度异质性



- 节点度异质性较强
- 近似幂律分布

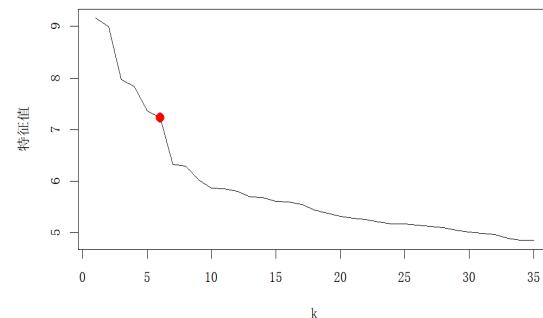


稀疏性

- 网络密度: 0.0006306262, 即0.0631%, 网络较稀疏。
- 网络的平均度为: 2.763404, 远小于 $\log|V|$, 认为该网络稀疏。



噪声



- $K = 6$ 时, 满足 $|\hat{\lambda}_K - \hat{\lambda}_{K+1}| / |\hat{\lambda}_K|$ 较大, 所以认为网络中的弱信号 (噪声) 问题并不十分严峻。

选择算法:
SCORE,
SCOREq,
RSC

3.2 统计学合作者网络-第一层K选择

研究背景

研究内容

研究结果

研究结论



模块度

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

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

结论：选择算法SCOREq，K=6

3.2 统计学合作者网络-第一层K选择

研究背景

研究内容

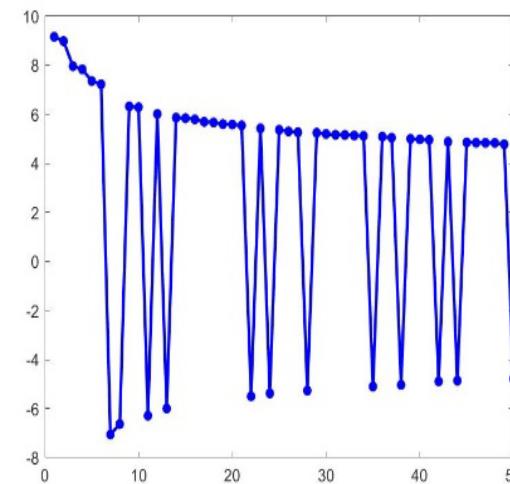
研究结果

研究结论

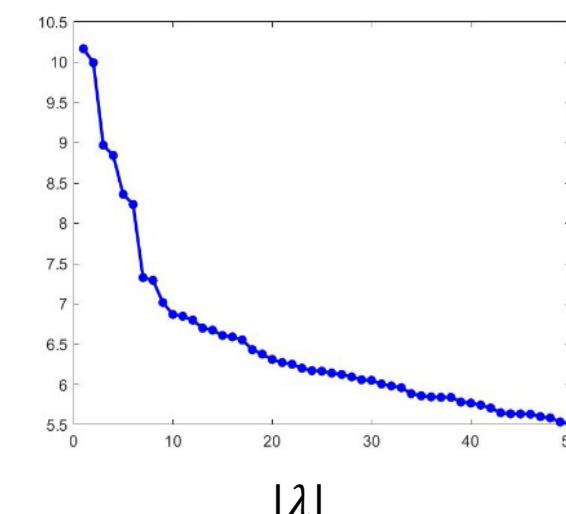
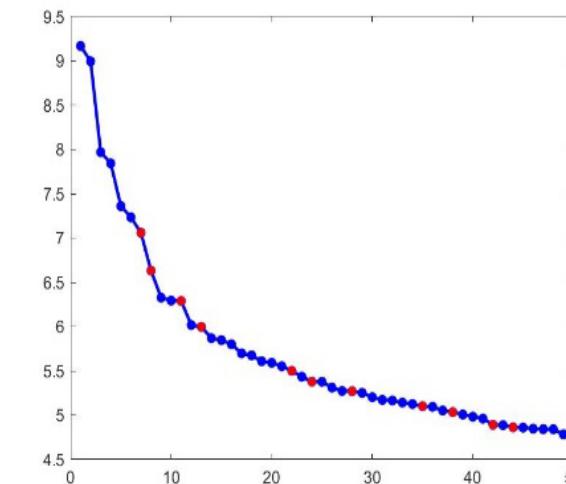
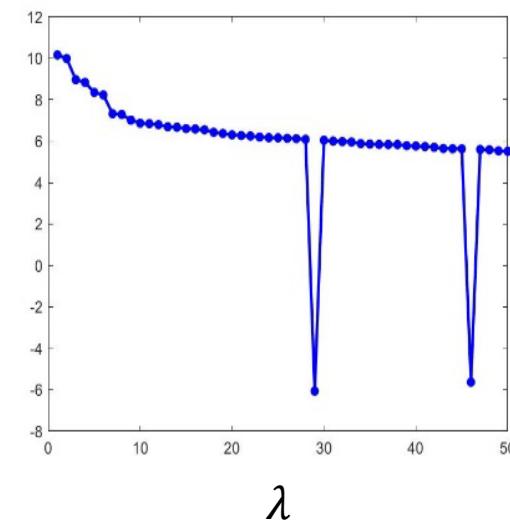


特征值碎石图

A 特征值



$A + I_n$ 特征值



选择 $K=6$

3.2 统计学合作者网络-第一层K选择

研究背景

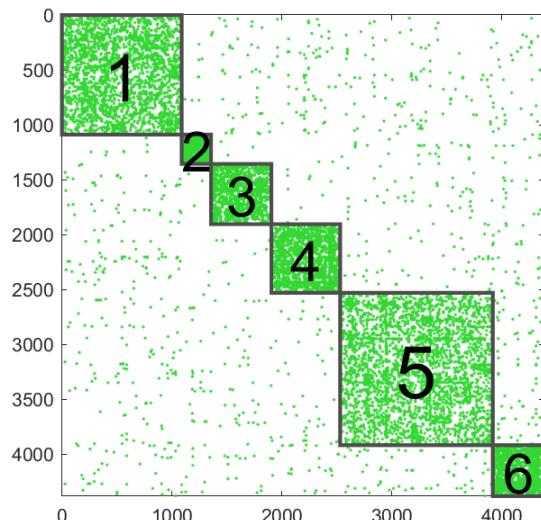
研究内容

研究结果

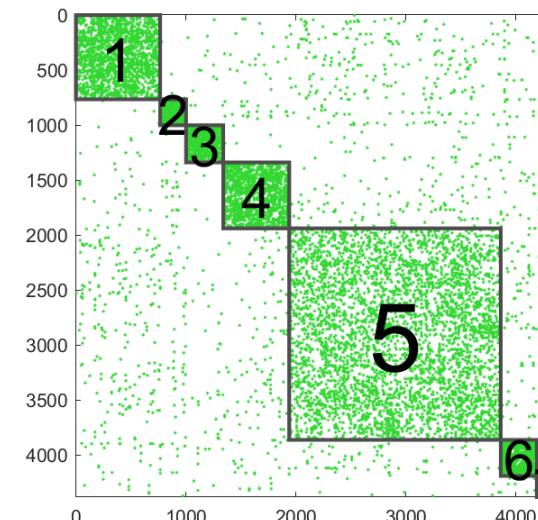


邻接矩阵

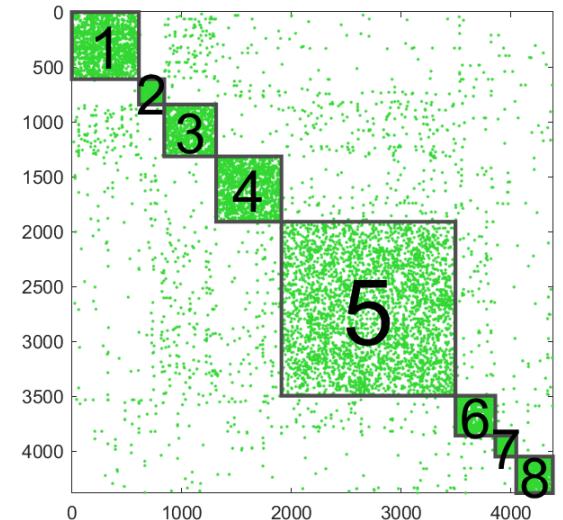
K=6



K=7



K=8



**注：绿色代表两个节点存在连边，框内为社区内，框外为社区间。
尽量使得框外连边较少**

研究背景

研究内容

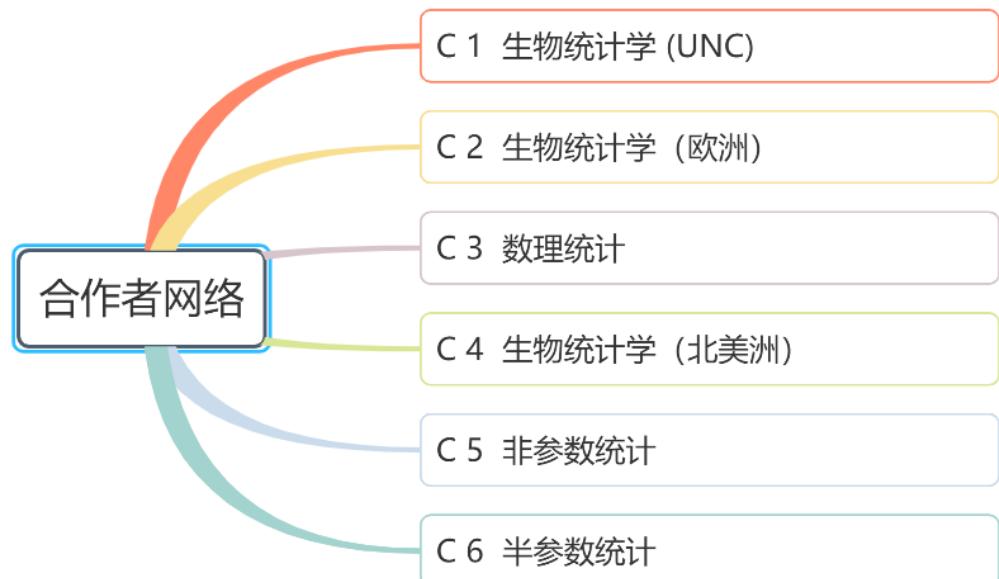
研究结果

研究结论

3.3 统计学合作者网络-结果



第一层社区发现结果



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	假设检验、统计计算、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

研究背景

研究内容

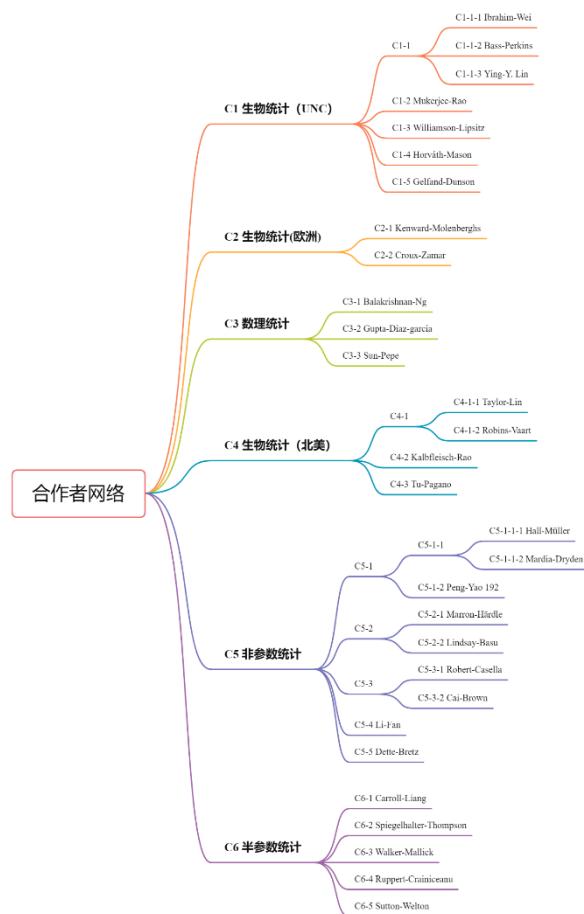
研究结果

研究结论

3.3 统计学合作者网络-结果



社区发现层次树



形成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。例如：

● 类似的研究兴趣

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)。

● 地理和文化因素

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。

● 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由 Jun Liu, Xiaotong Shen, 和Wing H Wong和三人的学生共同构成了这个子社区。

3.4 统计学合作者网络-稳健性分析

研究背景

研究内容

研究结果

研究结论



对比SCOREq与SCORE、RSC社区发现结果的差异（第一层）

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

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%	-

社区发现结果
较为接近。

4 研究结论与总结

研究背景

研究方法

研究结果

研究结论



对度异质性修正的**随机块模型DCBM**进行分析，并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**。



利用模拟数据、真实有标签数据集**验证谱聚类算法性质**，发现SCORE, SCOREq, RSC, SCORE+能较好地解决中大型稀疏异质网络社区发现问题，其中SCORE+算法对于解决噪声问题有出色表现。



基于统计学家合作者网络**实现了一个有效的真实数据社区发现范式**。包括层次谱聚类的社区发现方法、算法选择、社区数K的选择等操作，并通过算法进行比较，说明了社区划分的结果是较为稳健的。



提供了一个**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。该结果可以描述和可视化研究人员的研究概况，还为初级的研究人员选择研究主题、寻找资料、建立学术连结提供了一定指导。

感谢您的倾听
恳请老师批评指正！