0 引言

近年来,网络在很多领域应用广泛,自然界中诸多系统都可以用网络来描述,如社交网络、计算机网络、人类疾病基因网等.网络的一个普遍特征是具有社区结构,一般意义上的社区指的是网络节点的子集合,位于该集合内部的节点之间连接紧密,集合间的节点连接疏松.给定一个网络图,找出其社区结构的过程就是社区发现,它是网络分析中的一个基本问题.社区发现能够揭示网络中节点之间的交互关系,是一个非常有意义的课题.已有大量的学者从构建统计模型和设计社区发现评价标准等方面进行社区发现问题的研究. 统计模型包括潜在位置聚类模型[1](latent position cluster model), 随机分块模型[2](stochastic block model,SBM)及其扩展,如带有权重边的随机分块模型[3]、考虑重叠社区的随机分块模型[4-6]和度奇异的随机分块模型[7].在随机分块模型框架下,组内边的密度要大于组间边的密度.另一方面,部分学者从社区发现问题和聚类分析问题之间的关系,通过构造节点的相似度,将社区发现问题转化为聚类问题,提出不同的衡量标准,比如归一化切分[8](normalized cut, Ncut),模度块[9-10](modularity) 等.上述绝大多数的方法仅从网络的拓扑结构信息寻找网络节点的社区结构,有关此方面更多的工作请参考文献[11-13].

现实数据的形式多样,本文考虑两种形式的数据:①刻画网络拓扑结构的以邻接矩阵形式表现的数据,邻接矩阵 A 中的元素表示相邻节点之间有无结构关系,如果相邻的节点之间有结构关系,在网络图中通常以连边表示,那么邻接矩阵中对应的元素非零,否则,该元素为零;②刻画节点属性特征以多维向量 X 表示的数据. 基于拓扑结构的网络以带有节点及其连边情况的图的形式呈现.本文考虑一类不带自环或者重边的简单图,因此邻接矩阵是无权重对称矩阵.现实中同一社区内的节点不仅在结构上联系紧密而且这些节点也具备相似的特征信息,比如在社交网络中,同一个社区内的人们不仅具有朋友关系(结构关系),同时也具有较为相似的教育背景、居住环境、业余爱好等(节点属性).现实中对网络数据的抓取存有不足,比如在社交网络中会抓取大量噪声链接,同时也会丢失部分有用的链接,仅依靠网络的拓扑结构很难划分出有意义的社区.将两类信息结合会使得划分社区的精度进一步提高.

近年来,已有部分学者将网络的拓扑结构和节点属性结合起来研究网络的社区结构问题. Wang 等[14]提出了一类整合 K 均值(K-means)聚类和拉普拉斯聚类的算法,即对网络节点的特征属性用 K 均值聚类,对拓扑结构进行谱聚类,用一个调节参数平衡这两类信息在聚类过程中的作用.Zhou 等[15]给出了一种结合结构相似度和属性相似度的算法—SA 算法.SA 算法首先将分类型特征离散化为虚拟的特征节点,若网络节点具有某特征,则该节点与虚拟的特征节点之间有连边. 在计算距离时 SA 算法考虑了不同特征属性在社区划分过程中的贡献并通过迭代估计其权值.Yang 等[16]考虑了一类生成模型,假设网络结构数据和节点属性是相依的且由不同的概率分布生成,用最大似然法估计节点的类标签. Binkiewicz 等[17]给出了协变量辅助谱聚类方法(CASC),假设 L 是邻接矩阵 A 对应的拉普拉斯矩阵,Y 是特征属性 X 对应的协变量矩阵,CASC 算法将邻接矩阵 A 的拉普拉斯矩阵 L 推广为 $\widetilde{L}=LL+\alpha Y^{\mathrm{T}}Y$,后续算法与标准谱聚类相同,其中 α 为调节参数. Zhang 等[18]介绍了一类新的结合两类信息进行社区发现的算法 (JCDC),构造了一个凸目标函数,并估计出不同社区内不同节点的权重.该算法的目标函数中除了待估的特征权值还含有两个参数,其中一个调节结构信息和属性信息在划分社区过程中的相对重要性. 模拟数据表明该参数越大,结构信息越有用,反之,特征信息在划分社区过程中贡献度大.这就要求人们在使用该算法时需预先了解网络数据的结构信息和节点属性信息的相对价值, 实践中是不好推测的.更多相关文献可参考文献[19-20].

一般来说,节点的特征属性较多,不是所有的属性特征在划分节点时都能提供有用的信息,如考虑朋友圈关系,身高特征对划分朋友圈是没有用的.基于此,本文提出一类以谱聚类为基础结合结构信息和节点属性信息进行社区划分的算法(spectral clustering with structural and attributes information,SCSA).

与 Zhang 等[18]不同,SCSA 算法以标准切(Ncut)函数作为目标函数,由于目标函数非凸,提出一种新的估计特征权值的方法,不同社区内特征权值的大小取决于该特征在社区内的离散程度和该社区与其他社区之间在该特征上的离散程度.模拟实验和实例数据分析表明在结合两类信息进行社区划分时 SCSA 算法是非常有效的.

1　基于谱聚类的结合结构信息和特征属性的社区发现算法

1.1　准备工作

首先给出本文要用到的符号. $G=(V,E)$ 表示网络图,其中 $V=\{1,2,\cdots,n\}$ 是点集合, $E\subseteq V\times V$ 是边集合, $X=(X_1,\cdots,X_n)\in\mathbb{R}^{n\times p}$ 为节点特征属性矩阵, $X_i=(X_{i1},\cdots,X_{ip})$ 是节点 i 的 p 维特征向量.本文关注的是协调社区结构,即社区内部节点之间连接紧密而社区之间的节点连接疏松.令 $A=(A_{ij})$ 为邻接矩阵,若节点 i 和节点 j 之间无连边,则 $A_{ij}=0$,否则 $A_{ij}=1$.假设社区个数 K 已知.令 $\vec{c}=\{c_1,\cdots,c_n\}$ 为待估的社区标签(也称为类标签), $C_k=\{i:c_i=k,1\leqslant i\leqslant n\}$ 为待估社区, $\vec{\nu}=(\nu_1,\nu_2,\cdots,\nu_n)$ 表示网络节点的真实分类标签, $\Delta_k=\{i;\nu_i=k\}$, $1\leqslant k\leqslant K$ 为相应的社区. 向量 $\vec{\nu}$ 服从多项分布,即 $\Pr(\nu_i=k)=\pi_k$.图的划分问题就是寻找图的最小切问题,但是最小切函数会出现将所有节点划分为一个社区的极端情况,为了使得社区大小更均衡,将标准切(Ncut)函数作为图划分的目标函数(详见文献[8,21]),定义为

$$\mathrm{Ncut}(A)=\sum_{k=1}^{K}\frac{\sum\limits_{i\in C_k}\sum\limits_{j\notin C_k}A_{ij}}{\sum\limits_{i\in C_k}\sum\limits_{j=1}^{n}A_{ij}}\tag{1}$$

注意到 Ncut(A)中仅含结构信息,本文考虑节点的特征信息,需重新定义 Ncut.我们将节点的特征信息赋值在节点之间的连边上,如果相邻两点有连边($A_{ij}=1$),那么这两节点之间存在带有权值的连边,定义其权值为 $w_{ij}=A_{ij}f(X_i,X_j)$,其中 f 是非负函数,其具体形式将在下节给出.也就是说,如果两节点在结构上有连边,那么它们的特征信息会改变结构边的权值,从而影响节点的最终归属的社区.

1.2　算法

如上所述,本文不仅在划分社区过程中综合考虑结构信息和节点特征,还注意到不同社区内节点特征属性的差异性.定义带有权重的邻接矩阵 $W=\{w_{ij},1\leqslant i,j\leqslant n\}$,

$$w_{ij}=A_{ij}\exp\left\{-\frac{\langle\vec{\varphi}_{c_i},D_{ij}\rangle}{2\sigma^2}\right\}\tag{2}$$

式中, $\langle\cdot,\cdot\rangle$ 表示内积, $\vec{\varphi}_{c_i}=\{\varphi_{c_i1},\cdots,\varphi_{c_ip}\}$ 是 c_i 社区内不同特征的权值, $D_{ij}=(D_{ij1},D_{ij2},\cdots,D_{ijp})$, D_{ijl} 是节点 i 和节点 j 在第 l 个特征上的距离, σ 是高斯核函数控制邻接点范围的参数.如果节点的特征属性是数值型变量,那么 $D_{ijl}=|X_{il}-X_{jl}|$,若节点的特征属性是分类型变量,那么 $D_{ijl}=I(X_{il}\neq X_{jl})$, I 为示性函数. 重新定义式(1)中的 Ncut 为

$$\mathrm{Ncut}(\vec{e},\vec{\varphi})=\sum_{k=1}^{K}\frac{\sum\limits_{i\in C_k}\sum\limits_{j\notin C_k}w_{ij}}{\sum\limits_{i\in C_k}\sum\limits_{j=1}^{n}w_{ij}}\tag{3}$$

注 1.1　von Luxburg[21] 给出了几种估计 σ 方法,包括用图的最小生成树的最长边取值作为 σ 的估计,同时指出这些估计方法缺乏严格的理论证明, σ 的估计值仍依赖于数据类型. 本文拟用由 X 生成的图的最小生成树的最长边取值作为 σ 的估计值.

本文的目标是关于标签 $\vec{c}$ 和权值 $\vec{\varphi}$ 最小化 $\mathrm{Ncut}(\vec{c},\vec{\varphi})$.由式(2)可知, w_{ij} 和 w_{ji} 一般是不相等的,因此转化后的图就变成了有向图,即 $W=(w_{ij})$ 不是对称的.本文暂不考虑有向图,进一步的工作正在讨论中.常用的将有向图转换为无向图的方法是构造一个对称矩阵 $\widehat{W}$,如 $\widehat{W}=\frac{1}{2}(W+W^{\mathrm{T}})$,其中 W^{T} 是 W 的转置矩阵.

优化目标函数(3)可分拆成下面两个循环迭代问题:

问题 E :给定权值 $\hat{\vec{\varphi}}$,关于类标签 $\vec{c}$ 优化 $\mathrm{Ncut}(\vec{c},\hat{\vec{\varphi}})$;

问题 φ :给定类标签 $\hat{\vec{c}}$,关于权值 $\vec{\varphi}$ 优化 $\mathrm{Ncut}(\hat{\vec{c}},\vec{\varphi})$.

由于 $\mathrm{Ncut}(\vec{c},\vec{\varphi})$ 关于 $\{\vec{\varphi}\}$ 是非凸的,为了估计 $\{\vec{\varphi}\}$,需要将非凸函数(3)凸化或者找到 $\{\vec{\varphi}\}$ 一组“好”的初值,这都是不易实现的.本文不再局限于目标函数的凸化,而是采用一种自适应的权值调节方法估计节点特征的权值. 假设节点的结构信息对应的潜在的类标签与节点特征的潜在类标签是一致的.在算法有效的情况下,如果社区内节点之间在某个特征属性上的距离小,而相应的社区之间距离大,那么可以认为该节点特征对于社区划分是有用的,在社区划分过程中其对应的权值较大. 记 $a_{kl}=\sum\limits_{i\in C_k}\sum\limits_{j\notin C_k}A_{ij}D_{ijl}$, $b_{kl}=\sum\limits_{i\in C_k}\sum\limits_{j\in C_k}A_{ij}D_{ijl}$, a_{kl} 为第 k 个社区内的点与其余社区内相连点之间在特征 l 上的总距离, b_{kl} 为第 k 个社区内的相连节点之间在特征 l

上的距离之和.那么在第 $(t+1)$ 次迭代中,第 k 个社区内特征 l 的权值满足 $\varphi_{kl}^{(t+1)}=\varphi_{kl}^{(t)}+\Delta\varphi_{kl}^{(t)}$,其中 $\Delta\varphi_{kl}^{(t)}$ 为调整幅度函数,可用下式估计[22]

$$\Delta\varphi_{kl}=\frac{a_{kl}/b_{kl}}{\sum_{l=1}^{p}(a_{kl}/b_{kl})} \tag{4}$$

假设在每次迭代中社区内节点特征的权值满足 $\sum_{l=1}^{p}\varphi_{kl}^{(t)}=1,k=1,2,\cdots,K$,为此,需要对 $\varphi_{kl}^{(t)}$ 标准化,我们引入一个常用的标准化函数 $f(x_n)=\frac{x_n}{\sum_{i=1}^{n}x_n}$,那么标准化后的 $\varphi_{kl}^{(t+1)}$ 为(仍用 $\varphi_{kl}^{(t+1)}$ 表示)

$$\varphi_{kl}^{(t+1)}=f(\varphi_{kl}^{(t+1)})=\frac{\varphi_{kl}^{(t)}+\Delta\varphi_{kl}^{(t)}}{\sum_{l=1}^{p}\varphi_{kl}^{(t)}+\sum_{l=1}^{p}\frac{a_{kl}/b_{kl}}{\sum_{l=1}^{p}(a_{kl}/b_{kl})}}=\frac{1}{2}(\varphi_{kl}^{(t)}+\Delta\varphi_{kl}^{(t)}) \tag{5}$$

SCSA 算法详述如下:

①初始化 $\vec{\varphi}^{(0)}$ 并计算 $W^{(0)}$;$t=1,2,\cdots$;

②令 $\widetilde{W}^{(t)}=\frac{1}{2}(W^{(t)}+(W^{(t)})^{\mathrm{T}})$,其对应的拉普拉斯矩阵 $L^{(t)}=(D^{-1/2})^{(t-1)}\widetilde{W}^{(t)}(D^{-1/2})^{(t)}$;

③计算 $L^{(t)}$ 的前 K 个特征向量 $v_1^{(t)},\cdots,v_K^{(t)}$,归一化为 $Nv_1^{(t)},\cdots,Nv_K^{(t)}$;

④对特征向量矩阵 $[Nv_1^{(t)},\cdots,Nv_K^{(t)}]\in\mathbb{R}^{n\times K}$ 作用 K 均值算法,估计社区标签 $\vec{c}$;

⑤由式(4)和式(5)估计 $\vec{\varphi}^{(t)}$;

⑥重复步骤 ②~⑤直到目标函数(3)收敛.

注 1.2 一般地,在社区划分之前特征属性的信息是未知的,不妨将特征权值的初始值取成一样的,即 $\varphi_{kl}^{(0)}=\frac{1}{p}$.给定一个包含 n 个节点 $|E|$ 条边 p 个节点特征的网络图,SCSA 算法的实现过程如下:初始化 W 所需时间复杂度为 $O(|E|p)$,更新类标签 $\vec{c}$ 需用谱聚类,谱聚类算法中最耗时的是计算特征向量,时间复杂度为 $O(n^3)$,更新特征权值 $\vec{\varphi}$ 时间复杂度为 $O(|E|p)$,因此实现算法 SCSA 总的时间复杂度为 $O(n^3+|E|p)$.当 n 较大时,SCSA 算法的复杂度偏高,但现实中,当 n 充分大时,网络数据对应的邻接矩阵是稀疏的,那么计算特征向量的时间复杂度降为 $O(n^{3/2})$[8],有关稀疏网络数据结合节点特征进行社区划分的问题留待进一步探讨.

1.3 算法的收敛性

SCSA 算法是以谱聚类算法为基础的,而关于谱聚类算法收敛性的结论已经很成熟,在一般条件下,vou Luxburg 等[23]证明随着节点数目的增大,正则化的谱聚类是收敛的.与[15]类似,本文在随机分块模型下证明 SCSA 算法的收敛性,同时假设节点的特征属性 X 的分布有界.令 $\Pr(A_{ij}=1)=\gamma_n P_{\nu_i\nu_j}$,其中 $n\gamma_n\to\infty$,$n\to\infty$.还需要以下假设:

A1 对任意的 $l\in\{1,2,\cdots,K\}$,$\tilde{c}_0 n\geqslant|E_l|\geqslant c_0 n$,$0<c_0<\tilde{c}_0<1$.

A2 存在正常数 M_d ,$\widetilde{M}_d$ 使得对任意的节点 i ,j ,$\widetilde{M}_d\leqslant\sum_{l=1}^{p}D_{ijl}\leqslant M_d$.

A3 存在正常数 α_1 ,α_2 使得 $0<\alpha_1\leqslant\frac{(K-1)\max\{P_{kl}\}}{\min\{P_{kk},P_{ll}\}}\leqslant\alpha_2<1$,$(1-\epsilon)\leqslant\frac{P_{kl}}{P_{k'l}},\frac{P_{kk}}{P_{ll}}\leqslant(1+\epsilon)$,其中 $\epsilon>0$,$1\leqslant k<k'<l\leqslant K$.

假设 A1 保证了社区规模适度均衡,当节点的特征属性分布有界且有限维时,假设 A2 是合理的(暂不考虑高维稀疏的情形),由于本文考虑的是协调的随机分块模型,因此假设 A3 是合理的.定义

$$d(\vec{e},\vec{\nu})=\min_{\sigma(e_i)\in\Omega_k}\frac{1}{n}\sum_{i=1}^{n}I\{\sigma(e_i)\neq\nu_i\},$$

式中,Ω_k 是 $\{1,2,\cdots,K\}$ 所有排列的集合.

定理 1.1 在假设 A1~A3 下,对任意的 $\varepsilon>0$ 有

$$\Pr\left(d\left(\arg\min_{\vec{c}}\left(\min_{\vec{\varphi}}N(\vec{e},\vec{\varphi})\right),\vec{\nu}\right)>\varepsilon\right)\to0,n\to\infty.$$

证明见附录.

2 实验

为了验证 SCSA 算法的有效性,将 SCSA 与其他几种算法分别在模拟数据和实际数据上做对比试验,对比的算法有 JCDC,CASC,谱聚类(SPC)和 K 均值(K-means)算法.算法 SCSA,JCDC 和 CASC 都用到了结构信息和节点属性特征,而谱聚类(SPC)仅依靠结构信息进行社区划分,K 均值算法仅依靠节点特征信息对社区进行划分.

2.1 模拟数据

2.1.1 模拟一

第一个实验比较各种算法在给定节点特征的条件下比较上述几种算法在不同的结构信息下检测社

区的有效性.假设带有度校正的随机分块模型有两个类,每个类包含 100 个节点.若节点 i 和节点 j 在同一个社区内,则它们之间连边的概率为 $\theta_i\theta_j p$,若节点 i 和节点 j 不在同一个社区,则它们之间连边的概率为 $\theta_i\theta_j rp$,$0<r<1$. 假设每个社区内 95% 的节点的度校正参数为 $\theta_i=1$,余下 5% 的节点的度校正参数为 $\theta_i=10$.假设每个节点有 10 个特征,在第一个社区内,服从多元正态分布 $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I})$,在第二个社区内,服从多元正态分布 $\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I})$,其中 $\boldsymbol{\mu}=(1,0.75,0.5,0,0,0,0,0,0,0)$,$\boldsymbol{I}$ 为单位方阵. 由于高斯核函数中的参数 σ 在 SCSA 算法中有非常重要的作用,因此,首先说明注 1.1 给出的估计 σ 方法的可行性.图 1(a)给出了 SCSA 算法检测的社区结果与 σ 之间的关系,图中 $d=2\sigma^2$,$r=0.55$,垂直线对应的 x 值即为最小生成树的最长边的值对应的 d 值,记为 d_0. 由图 1(a)可看出在以 d_0 中心的区间内,SCSA 算法的估计精度相差不大,这也再次印证了 σ 取值的灵活性[21].

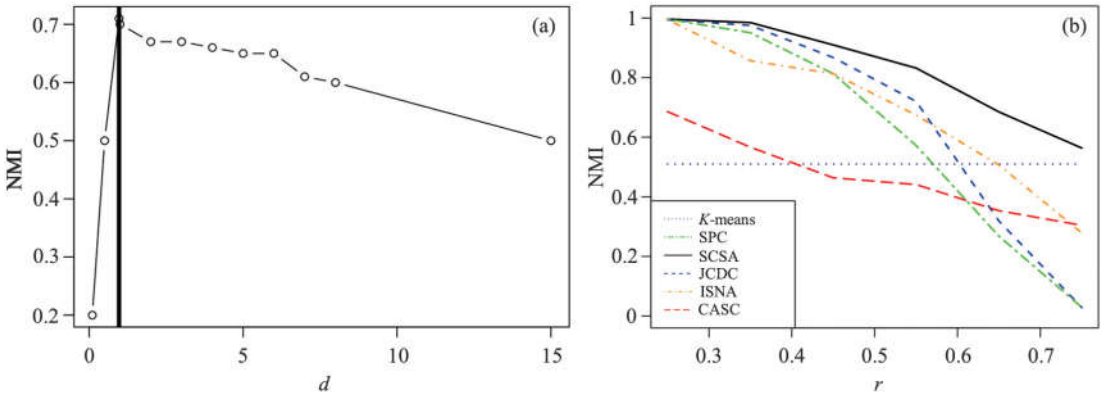


图 1　(a)$r=0.55$ 时,改变 $d=2\sigma^2$ 对应的 SCSA 算法 NMI 值.(b)改变 r 值对应各类算法的 NMI 值

Fig.1　(a) When $r=0.55$, NMI values of SCSA as $d=2\sigma^2$ varies.(b) The NMI values of different methods as r varies

下面考察在不同的结构信息下各种算法检测社区的有效性,通过变动 r 的值改变结构信息的强弱. SCSA 算法强调特征权值在算法实施过程中的作用,为了进一步说明,在本试验中考虑 SCSA 算法的一个特例—同社区内所有特征权值相等(即 $\varphi_{kl}=\frac{1}{p}$),也就是说在 SCSA 算法中不需要循环($t=0$). 为了记号的方便,将此算法记作 ISNA(integrated structural and node attributes information).度量算法精度的指标很多,本文用正则互信息(NMI)度量算法的有效性. NMI 取值越大,则该算法的精度越高.每次实验重复 20 次,实验结果如图 1(b)所示.

由 w_{ij} 解析式易见只有连接节点的特征信息在社区划分过程才起作用,由图 1(b)可见,当数据的结构信息较强时,ISNA 算法精度较高,但由于噪音特征的干扰,在网络数据的结构信息变弱时,ISNA 算法估计的 NMI 值远远低于 SCSA.由于 K 均值算法仅依赖于节点特征,对应的 NMI 值不会随着 r 的改变而改变. 谱聚类依赖于网络结构,因此它对应的 NMI 值随着 r 的改变变化明显.CASC 算法仅通过一个调节参数权衡类信息,即将节点的所有特征属性看成同样重要的.在噪音特征较多且结构信息较弱的情形下,CASC 算法划分社区的结果不是很好.JCDC 在结构信息强度和特征信息强度较大时划分社区的精度很高,但随着结构信息的减弱,JCDC 划分社区的能力骤减. JCDC 算法用梯度下降法估计节点特征的权重, 对节点特征的相似函数做了标准化处理,在估计特征的权值时会将部分无用的特征看成是“有用的”.我们随机选取了一组结构信号较弱时的 SCSA 和 JCDC 算法的实验结果,此时谱聚类对应的 NMI 值为 0.0006,如表 1 所示,SCSA 更善于发现特征的相对重要性.

表 1　$r=0.75$ 时,算法 SCSA 和 JCDC 估计结果

Tab.1　The estimated results returned by SCSA and JCDC when $r=0.75$

算法	NMI	权　值
SCSA	0.39	组 1:(0.15,0.13,0.1,0.09,0.09,0.08,0.08,0.09,0.08,0.08)　组 2:(0.22,0.14,0.07,0.06,0.07,0.06,0.06,0.06,0.07,0.06,0.06,0.06)
JCDC	0.002	组 1:(0.2,0.4,0.23,0.03,0.26,0.4,0.5,0.28,0.32,0.27)　组 2:(0.19,0.07,0.31,0.42,0.3,0.3,0.27,0.43,0.15,0.3)

2.1.2 模拟二

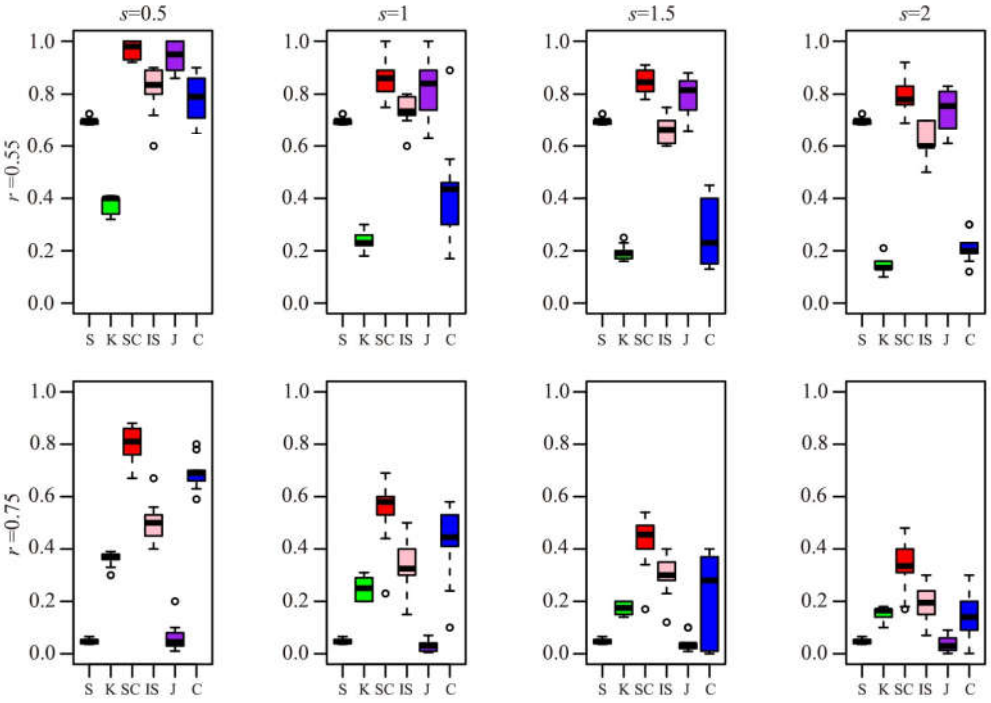
本组实验考察给定结构信息变动特征信息时各种算法的有效性.与第一组实验类似,在随机模块模型下,假设组内边链接概率为 $\theta_i\theta_j p$,组间边链接概率为 $\theta_i\theta_j r\ p$,$0<r<1$.假设每个节点有 $p=12$ 维特征,对于前 10 个特征,在第一个类内,它们的分布函数为多元正态分布 $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$,在第二个类内服从多元分布 $\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{\Sigma})$;剩下的两个特征服从均匀分布.令 $\mu=(1.25,0.75,0.5,0,0,0,0,0,0,0)$,$\boldsymbol{\Sigma}$ 为对角阵,前 3 个对角元素为 s ,余下 7 个对角元素全为 1,其中 s 分别取 4 个值:$s=0.5,1,1.5,2$. s 的值越大,对应的特征信号越弱.为了让实验更具一般性,在两组不同结构信息强度下考察随着 s 的变化各类算法的有效性,令 $r=0.55$,$r=0.75$,r 越大,结构信息越弱.

每组实验重复 20 次,实验结果如图 2 所示.当 $r=0.55$ 时,SPC 对应的 NMI 值显示结构信号很强.随着 s 增大,算法 SCSA,JCDC 和 ISNA 波动较小,它们对应的 NMI 值仍然较高,SCSA 显得尤为突出.当 $r=0.75$ 时,由 SPC 对应的 NMI 值可知结构信息较弱不足以得到有意义的社区结构,即使如此,SCSA 仍然可以有效利用特征信息得到较好的分组结果.随机选取两组实验:① $r=0.55$,$s=1.5$,对应的结构信息和特征信息都较强,谱聚类的 NMI 值为 0.57,K 均值算法对应的 NMI 值为 0.6;② $r=0.75$,$s=2$,结构信息和特征信息都较弱,谱聚类对应的 NMI 值为 0.02,K 均值算法对应的 NMI 值为 0.28.表 2 和表 3 中分布记录了两类算法 SCSA 和 JCDC 估计的 NMI 值和特征权值,SCSA 能更好地甄别有用的特征信息,JCDC 在结构信息和特征信息都较弱时将所有的特征信息都估计为“无用的”,此时它的社区划分结果不理想.

表 2 $s=1.5$,$r=0.55$ 时,算法 SCSA 和 JCDC 估计结果

Tab.2 The estimated results returned by SCSA and JCDC when $s=1.5$, $r=0.55$

算法	NMI	权	值
SCSA	0.81	组 1:(0.26,0.13,0.07,0.07,0.06,0.06,0.05,0.06,0.06,0.06,0.06,0.06)	组 2:(0.22,0.14,0.07,0.06,0.07,0.06,0.06,0.06,0.07,0.06,0.06,0.06)
JCDC	0.68	组 1:(0.24,0.05,0,0,5×10^{-9},0,0,0,0,0,0,0)	组 2:(0.1,0.11,0,0,0,0,0,0,0,0,0,0)



S:谱聚类;K:K-means;SC:SCSA;IS:ISNA;J:JCDC;C:CASC

图 2 改变 s 时各类算法 NMI 值的箱线图

Fig.2 The boxplots of NMI values of different methods as s varies

表 3　$s=2, r=0.75$ 时，算法 SCSA 和 JCDC 估计结果

Tab.3　The estimated results returned by SCSA and JCDC when $s=2, r=0.75$

算法	NMI	权值	
SCSA	0.38	组 1：(0.24, 0.09, 0.08, 0.07, 0.06, 0.06, 0.07, 0.06, 0.07, 0.07, 0.06, 0.06)	组 2：(0.24, 0.1, 0.07, 0.07, 0.07, 0.06, 0.06, 0.06, 0.07, 0.07, 0.06, 0.06)
JCDC	0.06	组 1：(3×10^{-9}, 0, 0, 2×10^{-9}, 0, 0, 0, 0, 0, 0, 0, 0.009, 0, 0)	组 2：(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

2.2　实际数据

2.2.1　世界贸易网络数据

De Nooy 等[24]用网络数据描述了 80 个国家之间的贸易往来情况.每个国家代表一个节点，如果两国之间的贸易满足一定限额则两国之间有连边.每个国家有若干特征信息，包括所属洲（非洲、亚洲、欧洲、南北美洲、大洋洲），各国在 1980 年的世界体系中地位（中心、强边缘、弱边缘、外围）及各国在 1994 年世界体系中地位（中心、边缘、外围）.由于真实的类标签是不知道的，我们分别以上述的特征分类标签作为参考的类标签，在数据的结构信息对应的邻接矩阵上作用谱聚类得到相应的 NMI 值，如表 4 所示.注意谱聚类过程中类别的个数是与对应的参考标签一致，即若以洲的分类标签为参考标签，对应的社区数是 $K=3$，其他类似.另外由于大洋洲的 3 个国家与其他国家连边极少，因此在社区划分过程不予考虑.

表 4　作用于世界贸易网络邻接矩阵的 SPC 算法的检测结果与各特征标签比较的 NMI 值（类别个数）

Tab.4　The NMI value (cluster number) of the SPC algorithm applied to the adjacency matrix of the world trade network compared with different attribute labels

节点特征	1994 年世界体系	1980 年世界体系	所属洲
NMI（类别个数）	0.07($K=3$)	0.17($K=4$)	0.47($K=5$)

由表 4 知以洲的分类标签作为参考标签得到的 NMI 值较高，也就是说洲的分类标签与数据的结构信息较为吻合，因此我们以洲的分类标签作为整组实验的“真实”标签，余下的两个信息（各国在 1980 年和 1994 年在世界体系中的地位）作为节点的特征属性，比较各类算法关于 77 个国家的贸易划分.仍以 NMI 值衡量各类算法识别社区的有效性.图 3 给出了各类算法的估计结果，SCSA 估计的 NMI 值最高.图 4 比较了 SCSA 和 JCDC 估计的各特征权值，由于 SCSA 算法要求在同一社区内节点特征权值和为 1，因此 SCSA 估计的权值要比 JCDC 对应的权值略大，但是各特征的相对重要性是一致的. 图 4 表明两类算法都能较好地区别各节点特征在社区划分过程中的相对重要性，其中 SCSA 估计的权值与表 4 所示各节点特征的重要性是吻合的.

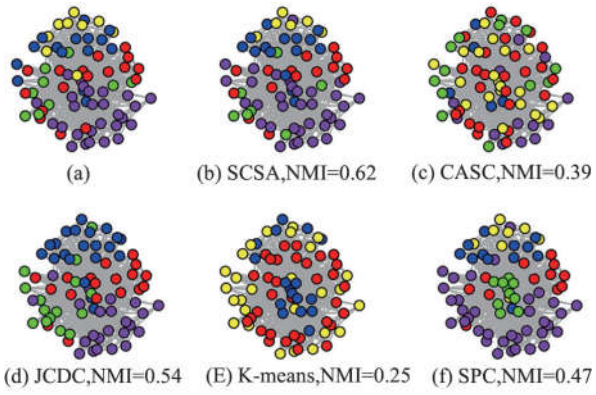
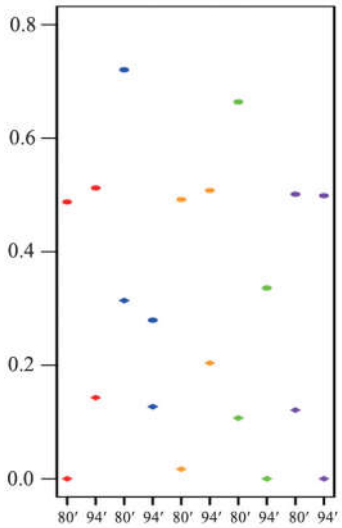


图 3　(a)贸易网络数据的“真实标签”，(b)～(f)5 种算法对贸易网络数据的社区检测结果

Fig.3　(a) The “true labels” of world trade network, (b)～(f) the community detection results of the world trade network returned by the five methods



○为 SCSA 估计的权值，◇为 JCDC 估计的权值，不同的颜色代表不同的社区. 记号 80′：1980 年世界体系中地位；94′：1994 世界体系中地位.

图 4　算法 SCSA 和 JCDC 估计的贸易网络数据节点特征权值

Fig.4　Attribute weights of the world trade network returned by SCSA and JCDC

2.2.2 律师朋友圈网络

第二个实例出自 Lazega[25]，每个节点代表一位律师，邻接矩阵描述了 71 位律师之间的朋友关系. 每位律师带有 7 个属性特征，包括其在公司的地位（合伙人，聘用律师），办公地点（波士顿，哈特福特，普罗维登斯），性别，实践（诉讼，企业），毕业院校（哈佛，耶鲁，其他），年龄及工龄. 由于真实的分类标签是未知的，我们采用与第一个案例类似的方法，在邻接矩阵上作用谱聚类，分别以上述 7 个特征分类标签作为参考类标签估计相应的 NMI 值. 以不同的特征标签作为参考标签，谱聚类中的类别个数 K 是不同的，如表 5 所示. 通过比较 NMI 值发现，律师在公司中地位的分类标签与朋友圈结构信息较为一致，因此选取地位的分类标签作为“真实”标签，其余 6 个特征作为节点的属性特征，图 5 给出了不同算法在上述假设下的估计结果. 图 6 给出了 SCSA 和 JCDC 算法估计的各特征的权值，因为节点特征较多，SCSA 给出的权值较 JCDC 估计的对应的权值略小. SCSA 给出的权值估计值与表 5 的结果是吻合的，SCSA 的估计权值和 JCDC 估计权值表明二者能够很好地甄别节点属性的重要性. 综上所述，SCSA 算法不仅可以较好地甄别律师所在朋友圈，同时可以发现同一个朋友圈内的律师们具有的相似特征.

表 5 作用于律师朋友圈邻接矩阵的 SPC 算法的检测结果与各特征标签比较的 NMI 值（类别个数）

Tab.5 The NMI value (cluster number) of the SPC algorithm applied to the adjacency matrix of the lawyer friendship network compared with the different attribute labels

地位	性别	实践	毕业院校	办公地点	年龄	工龄
0.44	0.005	0.087	0.12	0.12	0.22	0.25
(3)	(2)	(3)	(2)	(2)	(2)	(2)

2.2.3 威德尔海数据

最后一个实例是关于威德尔海的生物链数据[8]，这是一组关于捕食者-猎物的数据模型，包含 489 种生活在威德尔海的生物. 每一种生物具备下列特征属性：进食方式，代谢类型，移动类型，居住环境，平均体重. 与前两例讨论类似，对邻接矩阵作用谱聚类，以上述 5 个特征的标签作为“真实标签”，对应的 NMI 值分别为 0.42，0.3，0.29，0.2，0.12，显然以进食方式的

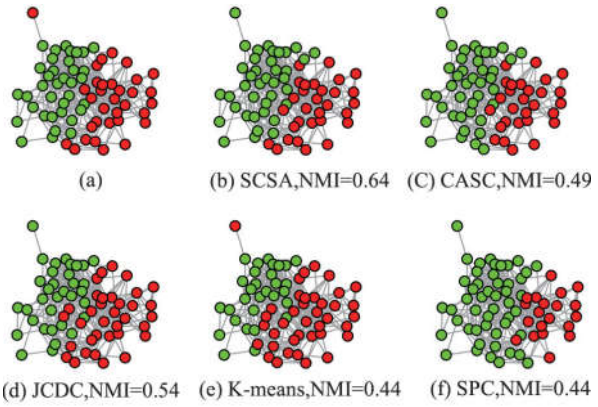
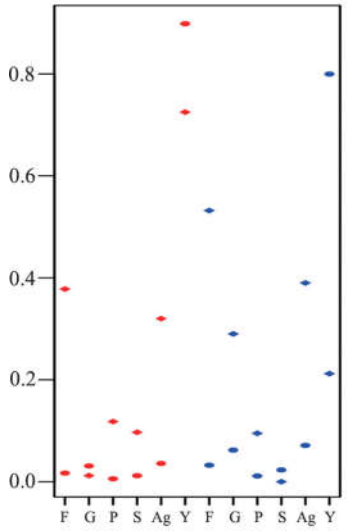


图 5 （a）律师朋友圈数据的“真实标签”，（b）～（f）5 种算法对律师朋友圈数据的划分结果

Fig.5 (a) The “true labels” of the lawyer friendship network, (b)～(f) the community detection results of the lawyer friendship network returned by the five methods



○为 SCSA 估计的权值，◇为 JCDC 估计的权值，不同的颜色代表不同的社区. F：办公地点；G：性别；P：实践；S：毕业院校；Ag：年龄；Y：工作年限.

图 6 算法 SCSA 和 JCDC 估计的律师朋友圈数据节点特征权值

Fig.6 Attribute weights of the lawyer friendship network returned by SCSA and JCDC

标签作为结构信息的“真实标签”更合理. 将余下的 4 个特征作为节点的属性特征. 图 7 给出了不同算法的估计结果. 图 8 给出了 SCSA 和 JCDC 估计的各特征的权值，SCSA 算法估计的权值与上述以不同特征标签为“真实标签”的划分结果对应的 NMI 值相吻合. JCDC 算法在计算节点之间特征相似度时做标准化处理，使得取值分布较为集中的节点信息的作用弱化，因此估计结果（图 7(c)）不甚理想.

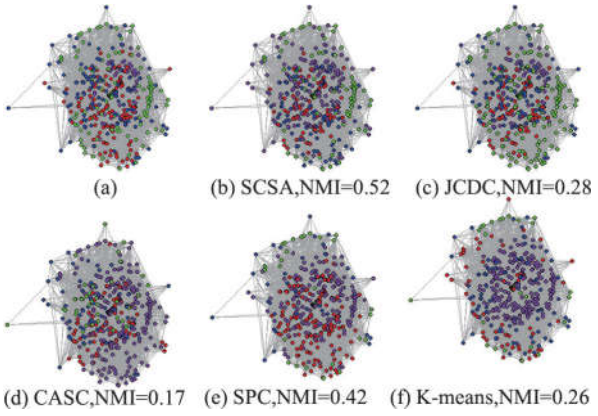
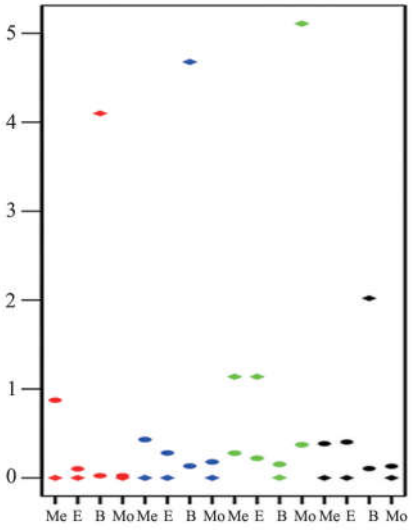


图 7　(a)威德尔海数据的"真实标签"，(b)～(f)5 种算法对威德尔海数据的划分结果

Fig.7　(a) The "true labels" of the Weddell sea dataset, (b)～(f) the community detection results of the Weddell sea dataset network returned by the five methods



○为 SCSA 估计的权值，◇为 JCDC 估计的权值，不同的颜色代表不同的社区.Me：代谢类型；E：居住环境；B：平均体重；Mo：移动类型.

图 8　算法 SCSA 和 JCDC 估计的威德尔海数据节点特征权值

Fig.8　Attribute weights of the Weddell sea dataset returned by SCSA and JCDC

3　结论

现实中，大量的网络数据包含节点的特征信息和网络的结构信息，在很多情形下，二者信息是关联的，将二者结合起来能够提高社区发现的精度. 基于谱聚类，本文提出了一类结合两类信息的社区发现算法 SCSA.有别于大部分结合两类信息的算法，算法 SCSA 在谱聚类过程中嵌入了特征权值校正方法，可估计出不同特征在社区划分过程中的不同作用，有效地利用信号强的特征信息，弱化不相关的特征信息.即使在结构信息较弱的情况下，SCSA 也能挖掘有用的特征信息来提高社区发现的精度.通过多组模拟数据和实例数据验证了算法的有效性.但是在节点属性较多的情形下，本算法的合理有效性还有待于进一步的探讨.

参考文献(References)

[1] HANDCOCK M S, RAFTERY A E, TANTRUM J M. Model-based clustering for social networks [J]. Journal of the Royal Statistical Society Series A, 2007, 170: 301-354.

[2] HOLLAND P W, LASKEY K B, LEINHARDT S. Stochastic blockmodels: First steps [J]. Social Networks, 1983,5: 109-137.

[3] MARIADASSOU M, ROBIN S, VACHER C. Uncovering latent structure in valued graphs: A variarion approach [J]. The Annals of Applied Statistics, 2010, 4(2): 715-742.

[4] AIROLDI E M, BLEI D M, FIENBERG S E, et al. Mixed-membership stochastic blockmodels [J]. Journal of Machine Learning Research, 2008,9: 1981-2014.

[5] LATOUCHE P, BIRMELÉ E, AMBROISE C. Overlapping stochastic block models[DB/OL]. arXiv.org: arXiv:0910.2098, 2009.

[6] ZHANG Y, LEVINA E, ZHU J. Detecting overlapping communities in networks using spectral methods[DB/OL]. arXiv.org: arXiv:1412.3432, 2014.

[7] KARRER B, NEWMAN M E. Stochastic blockmodels and community structure in networks [J]. Physical Review E, 2011, 83: 016107.

[8] SHI J, MALIK J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[9] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69: 026113.

[10] NEWMAN M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103: 8577-8582.

[11] FIENBERG S E. Introduction to papers on the modeling and analysis of network data[J]. Annals of Applied Statistics, 2010, 4: 1-4.

[12] FIENBERG S E. A brief history of statistical models for network analysis and open challenges[J].Journal of Computational and Graphical Statistics, 2012, 21: 825-839.

[13] 高启航，景丽萍，于剑，等. 基于结构和适应度的社区发现[J]. 中国科学技术大学学报，2014，44(7)：563-569.
GAO Qihang, JING Liping, YU Jian, et al.

Community detection based on structure and fitness [J]. Journal of University of Science and Technology of China, 2014, 44(7):563-569.

[14] WANG F, DING C, LI T. Integrated KL (K-means-Laplacian) clustering: A new clustering approach by combining attribute data and pairwise relations[C]// Proceedings of the 2009 SIAM International Conference on Data Mining. SIAM, 2009.

[15] ZHOU Y, CHENG H, YU J X. Graph clustering based on structural/attribute similarities [J]. Proceedings of the VLDB Endowment, 2009, 2: 718-729.

[16] YANG J, MCAULEY J, LESKOVEC J. Community detection in networks with node attributes[C]//2013 IEEE 13th International Conference on Data Mining. IEEE, 2013: 1151-1156.

[17] BINKIEWICZ N, VOGELSTEIN J T, ROHE K. Covariate assisted spectral clustering[DB/OL]. arXiv.org: arXiv:1411.2158, 2014.

[18] ZHANG Y, LEVINA E, ZHU J. Community detection in networks with node features[J].Electronic Journal of Statistics, 2016, 10: 3153-3178.

[19] GUNNEMANN S, FARBER I, RAUBACH S. Spectral subspace clustering for graphs with feature vectors[C]// 2013 IEEE 13th International Conference on Data Mining. IEEE, 2013: 231-240.

[20] NEWMAN M E J, CLAUSET A. Structure and inference in annotated networks [J]. Nature Communications, 2016, 7: 11863.

[21] VON LUXBURG U. A tutorial on spectral clustering [J]. Statistics and Computing, 2007, 17(4):395-416.

[22] TSAI C Y, CHIU C C. Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm [J]. Computational Statistics and Data Analysis, 2008, 52: 4658-4672.

[23] VON LUXBURG U, BELKIN M, BOUSQUET O. Consistency of spectral clustering[J]. The Annals of Statistics, 2008, 36(2): 555-586.

[24] DE NOOY W, MRVAR A, BATAGELJ V. Exploratory Social Network Analysis with Pajek[M]. Cambridge, MA: Cambridge University Press, 2011.

[25] LAZEGA E. The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership[M]. New York: Oxford University Press, 2001.

附录

与 Zhang 等[25]证明类似,为了证明定理 2.1,需要以下引理.

引理 A.1 在假设 A1～A3 下,若 $\gamma_n n \to \infty$,则有

$$\Pr(\max_{\vec{e},\vec{\varphi}} |\, \mathrm{Ncut}(\vec{e},\vec{\varphi}) - E[\mathrm{Ncut}(\vec{e},\vec{\varphi})]\,|) = O_P((\gamma_n n)^{-\frac{1}{2}}) \tag{A.1}$$

证明 令 $0<\epsilon<1$, $\Delta>0$, 可得

$$\Pr\Big(\max_{\vec{\varphi}} |\,\mathrm{Ncut}(\vec{e},\vec{\varphi}) - E[\mathrm{Ncut}(\vec{e},\vec{\varphi})]\,| > K\delta\Big) \leqslant$$

$$\sum_{k=1}^{K}\Pr\Big(\max_{\varphi_k}\left|\frac{\sum_{i\in E_k}\sum_{j\notin E_k}A_{ij}e^{-\langle D_{ij},\varphi_k\rangle}}{\sum_{i\in E_k}\sum_{j\in E_k}A_{ij}e^{-\langle D_{ij},\varphi_k\rangle}} - \frac{\sum_{i\in E_k}\sum_{j\notin E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}{\sum_{i\in E_k}\sum_{j\in E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}\right| > \delta\Big) \leqslant$$

$$\sum_{k=1}^{K}\left\{\Pr\Big(\max_{\varphi_k}\left|\frac{\sum_{i\in E_k}\sum_{j\notin E_k}(A_{ij}-\gamma_n P_{\nu_i\nu_j})e^{-\langle D_{ij},\varphi_k\rangle}}{(1-\epsilon)\sum_{i\in E_k}\sum_{j\in E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}\right| > \frac{\delta}{2}\Big) + \Pr\Big(\max_{\varphi_k}\left|\frac{\epsilon\sum_{i\in E_k}\sum_{j\notin E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}{(1-\epsilon)\sum_{i\in E_k}\sum_{j\in E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}\right| > \frac{\delta}{2}\Big)\right\} = I_1 + I_2 \tag{A.2}$$

由假设 A1 可得,

$$\frac{\epsilon\sum_{i\in E_k}\sum_{j\notin E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}{(1-\epsilon)\sum_{i\in E_k}\sum_{j\in E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}} \leqslant \frac{\epsilon\,|E_k|\,(n-|E_k|)}{(1-\epsilon)\,|E_k|^2\gamma_n e^{-M_d}\min\limits_{k,l}P_{kl}} \leqslant \frac{\epsilon}{4(1-\epsilon)\gamma_n e^{-M_d}\min\limits_{k,l}P_{kl}}.$$

由 ϵ 的任意性可得当 $\epsilon\to 0$ 时, $I_2\to 0$.

令 $\Psi_\epsilon = \left\{\left(\frac{a_1\epsilon}{\sqrt{p}},\frac{a_2\epsilon}{\sqrt{p}},\cdots,\frac{a_p\epsilon}{\sqrt{p}}\right), a_1,a_2,\cdots,a_p \in \{0,1,2,\cdots,[\frac{\sqrt{p}}{\epsilon}],\frac{\sqrt{p}}{\epsilon}\}\right\}$ 是 $[0,1]$ 上的 ϵ 网格,在 Ψ_ϵ 内, $\hat{\varphi}_k(\Psi_\epsilon)$ 是 φ_k 最优估计.则有

$$\max_{\varphi_k}\left|e^{-\langle D_{ij},\varphi_k\rangle}-e^{-\langle D_{ij},\hat{\varphi}_k(\Psi_\epsilon)\rangle}\right|\leqslant\max_{\varphi_k}\left|\frac{\partial e^{-\langle D_{ij},\varphi_k\rangle}}{\partial\varphi_k}\right|\left|\varphi_k-\hat{\varphi}_k(\Psi_\epsilon)\right|\leqslant 2M_d e^{-\tilde{M}_d\epsilon}\tag{A.3}$$

将(A.3)代入 I_1,由 Hoeffding 不等式可得

$$I_1\leqslant\sum_{k=1}^{K}\left\{\Pr\left(\max_{\varphi_k}\left|\frac{\sum\limits_{i\in E_k}\sum\limits_{j\notin E_k}\left|A_{ij}-\gamma_n P_{\nu_i\nu_j}\right|\left|e^{-\langle D_{ij},\varphi_k\rangle}-e^{-\langle D_{ij},\hat{\varphi}_k(\Psi_\epsilon)\rangle}\right|}{(1-\epsilon)\sum\limits_{i\in E_k}\sum\limits_{j\in E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}\right|>\frac{\delta}{4}\right)+\right.$$

$$\left.\Pr\left(\max_{\varphi_0\in\Psi_\epsilon}\left|\frac{\sum\limits_{i\in E_k}\sum\limits_{j\notin E_k}(A_{ij}-\gamma_n P_{\nu_i\nu_j})e^{-\langle D_{ij},\hat{\varphi}_0(\Psi_\epsilon)\rangle}}{(1-\epsilon)\sum\limits_{i\in E_k}\sum\limits_{j\in E_k}\gamma_n P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}\right|>\frac{\delta}{4}\right)\right\}\leqslant$$

$$K\Pr\left(\frac{2\epsilon(1+\epsilon)\,M_d\,|E_k|\,(n-|E_k|)}{(1-\epsilon)\gamma_n\min\limits_{k,l}P_{kl}\,|E_k|\,(|E_k|-1)}>\frac{\delta}{4}\right)+$$

$$\sum_{k=1}^{K}\Pr\left(\max_{\varphi_0\in\Psi_\epsilon}\left|\sum_{i\in E_k}\sum_{j\notin E_k}(A_{ij}-\gamma_n P_{\nu_i\nu_j})e^{-\langle D_{ij},\hat{\varphi}_0(\Psi_\epsilon)\rangle}\right|>\frac{c_2\delta\,|E_k|^2\gamma_n}{4}\right)\leqslant$$

$$I_3+2K\,|\Psi_\epsilon|\exp\left\{-\frac{(c_2\delta\gamma_n)^2c_0^4n^2}{8\tilde{c}_0(1-c_0)}\right\}\tag{A.4}$$

式中,$c_2=(1-\epsilon)e^{-M_d}\min\limits_{k,l}P_{kl}$.由 ϵ 的任意性可得 $I_3\to 0$.综上可得

$$\Pr\left(\max_{\vec{e},\vec{\varphi}}\left|\mathrm{Ncut}(\vec{e},\vec{\varphi})-E[\mathrm{Ncut}(\vec{e},\vec{\varphi})]\right|>K\delta\right)\leqslant 2K\,|\Psi_\epsilon|\exp\left\{-\frac{(c_2\delta\gamma_n)^2c_0^4n^2}{8\tilde{c}_0(1-c_0)}+p\ln\left(\frac{\sqrt{p}}{\epsilon}\right)\right\}\tag{A.5}$$

因为 $\gamma_n n\to\infty$,令 $\Delta=(\gamma_n n)^{-\frac{1}{2}}$ 可证得引理.

令 $\chi=(\chi_{kl})$ 为 K 阶方阵,其中 $\chi_{kl}=\dfrac{1}{n}\sum\limits_{i=1}^{n}I\{e_i=k,\nu_i=l\}$.令 $\mathcal{D}$ 为对角矩阵,其对角元素满足 $\mathcal{D}_l=\dfrac{1}{n}\sum\limits_{i=1}^{n}I\{\nu_i=l\}$,$l=1,\cdots,K$.易见当 $\vec{e}=\vec{v}$ 时,$\chi=\mathcal{D}\mathcal{O}$,其中 $\mathcal{O}$ 是置换矩阵.由此可得

$$d(\vec{e},\vec{\nu})=\frac{1}{2}\min_{\mathcal{O}\in\mathcal{O}_k}\|\chi(\vec{e})-\mathcal{D}\mathcal{O}\|_1=\frac{1}{2}\min_{\mathcal{O}\in\mathcal{O}_k}\sum_{l=1}^{n}\sum_{k=1}^{n}\left|((\chi(\vec{e})-\mathcal{D}\mathcal{O})_{kl}\right|.$$

为了估计 $d(\vec{e},\vec{\nu})$,引入 χ 的连续函数 $F_k(\chi)=\dfrac{1}{n}\sum\limits_{k=1}^{K}\dfrac{\sum\limits_{k'\neq k}\sum\limits_{l=1}^{K}\sum\limits_{l'=1}^{K}\chi_{kl}\chi_{k'l'}P_{ll'}}{\sum\limits_{l=1}^{K}\sum\limits_{l'=1}^{K}\chi_{kl}\chi_{kl'}P_{ll'}}$.对任意的 $\vec{e}$,

$$F_k(\chi(\vec{e}))=\frac{1}{n}\sum_{k=1}^{K}\frac{\sum\limits_{i\in E_k}\sum\limits_{j\notin E_k}P_{\nu_i\nu_j}}{\sum\limits_{i\in E_k}\sum\limits_{j\in E_k}P_{\nu_i\nu_j}}.$$

引理 A.2　在假设 A1～A3 下,存在正常数 c 使得

$$\max_{\vec{e},\vec{\varphi}}\left|\frac{E[\mathrm{Ncut}(\vec{e},\vec{\varphi})]}{n}-F_k(\chi(\vec{e}))\right|\leqslant\frac{c}{n}\tag{A.6}$$

证明　由假设 A2,存在正常数 e_0 使得对任意的 i,j,k,有 $e_0\leqslant e^{-\langle D_{ij},\varphi_k\rangle}\leqslant 1$,那么

$$\max_{\vec{e},\vec{\varphi}}\left|\frac{E[\mathrm{Ncut}(\vec{e},\vec{\varphi})]}{n}-F_k(\chi(\vec{e}))\right|=\frac{1}{n}\max_{\vec{e},\vec{\varphi}}\sum_{k=1}^{K}\left|\frac{\sum\limits_{i\in E_k}\sum\limits_{j\notin E_k}P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}{\sum\limits_{i\in E_k}\sum\limits_{j\in E_k}P_{\nu_i\nu_j}e^{-\langle D_{ij},\varphi_k\rangle}}-\frac{\sum\limits_{i\in E_k}\sum\limits_{j\notin E_k}P_{\nu_i\nu_j}}{\sum\limits_{i\in E_k}\sum\limits_{j\in E_k}P_{\nu_i\nu_j}}\right|\leqslant$$

$$\frac{1}{n}\sum_{k=1}^{K}\left|\frac{|E_k|\,(n-|E_k|)}{e_0\min\limits_{k,l}P_{kl}\,|E_k|^2}-\frac{\min\limits_{k,l}P_{kl}\,|E_k|\,(n-|E_k|)}{|E_k|^2}\right|=\frac{1}{n}\left|\frac{1-e_0\min\limits_{k,l}P_{kl}^2}{e_0\min\limits_{k,l}P_{kl}}\right|\sum_{k=1}^{K}\frac{(n-|E_k|)}{|E_k|}\leqslant$$

$$\frac{1}{n}\left|\frac{1-e_0\min\limits_{k,l}P_{kl}^2}{e_0\min\limits_{k,l}P_{kl}}\right|\frac{K(1-c_0)}{c_0}=\frac{c}{n}\tag{A.7}$$

式中，$c=\left|\dfrac{1-e_0\min\limits_{k,l}P_{kl}^2}{e_0\min\limits_{k,l}P_{kl}}\right|\dfrac{K(1-c_0)}{c_0}$.

定理 2.1 的证明 对任意的 $1\leqslant k\leqslant K$，当 $n\to\infty$ 时，有 $\sum\limits_{l=1}^{K}\chi_{kl}\to\pi_k$，$\mathscr{D}_k\to\pi_k$.故对任意的 $0<\epsilon<1$，

$$F_k(\mathscr{D})-F_k(\chi)=\frac{1}{n}\sum_{k=1}^{K}\left[\frac{\sum\limits_{k'\neq k}\mathscr{D}_k\mathscr{D}_{k'}P_{kk'}}{\mathscr{D}_k{}^2P_{kk}}-\frac{\sum\limits_{k'\neq k}\sum\limits_{m=1}^{K}\sum\limits_{m'=1}^{K}\chi_{km}\chi_{k'm'}P_{mm'}}{\sum\limits_{l=1}^{K}\sum\limits_{l'=1}^{K}\chi_{kl}\chi_{kl'}P_{ll'}}\right]\leqslant$$

$$\frac{1}{n}\sum_{k=1}^{K}\left[\frac{(1+\epsilon)\sum\limits_{k'\neq k}\sum\limits_{m=1}^{K}\sum\limits_{m'=1}^{K}\chi_{km}\chi_{k'm'}P_{kk'}}{(1-\epsilon)\sum\limits_{l=1}^{K}\sum\limits_{l'=1}^{K}\chi_{kl}\chi_{kl'}P_{kk}}-\frac{\sum\limits_{k'\neq k}\sum\limits_{m=1}^{K}\sum\limits_{m'=1}^{K}\chi_{km}\chi_{k'm'}P_{mm'}}{\sum\limits_{l=1}^{K}\sum\limits_{l'=1}^{K}\chi_{kl}\chi_{kl'}P_{ll'}}\right]=$$

$$\frac{1}{n}\sum_{k=1}^{K}\frac{\sum\limits_{k'\neq k}\sum\limits_{l=1}^{K}\sum\limits_{l'=1}^{K}\sum\limits_{m=1}^{K}\sum\limits_{m'=1}^{K}\chi_{km}\chi_{k'm'}\chi_{kl}\chi_{kl'}\left((1+\epsilon)P_{ll'}P_{kk'}-(1-\epsilon)P_{kk}P_{mm'}\right)}{(1-\epsilon)\sum\limits_{l=1}^{K}\sum\limits_{l'=1}^{K}\chi_{kl}\chi_{kl'}P_{kk}\sum\limits_{l=1}^{K}\sum\limits_{l'=1}^{K}\chi_{kl}\chi_{kl'}P_{ll'}}\tag{A.8}$$

对任意的 k，将(A.8)的分子划分为如下几部分：

$$\sum_{k'\neq k}\sum_{m=1}^{K}\sum_{m'=1}^{K}\chi_{km}\chi_{k'm'}\chi_{kl}^2\left((1+\epsilon)P_{ll}P_{kk'}-(1-\epsilon)P_{kk}P_{mm'}\right)+$$

$$\sum_{k'\neq k}\sum_{l=1}^{K}\sum_{l'\neq l}\sum_{m=1}^{K}\sum_{m'=1}^{K}\chi_{km}\chi_{k'm'}\chi_{kl}\chi_{kl'}\left((1+\epsilon)P_{ll'}P_{kk'}-(1-\epsilon)P_{kk}P_{mm'}\right)=$$

$$\sum_{k'\neq k}\chi_{km}\chi_{k'm}\chi_{kl}^2\left((1+\epsilon)P_{ll}P_{kk'}-(1-\epsilon)P_{kk}P_{mm}\right)+$$

$$\sum_{k'\neq k}\sum_{m=1}^{K}\sum_{m'\neq m}\chi_{km}\chi_{k'm'}\chi_{kl}\chi_{kl'}\left((1+\epsilon)P_{ll}P_{kk'}-(1-\epsilon)P_{kk}P_{mm'}\right)+$$

$$\sum_{k'\neq k}\sum_{l=1}^{K}\sum_{l'\neq l}\sum_{m=1}^{K}\sum_{m'=1}^{K}\chi_{km}\chi_{k'm'}\chi_{kl}\chi_{kl'}\left((1+\epsilon)P_{ll'}P_{kk'}-(1-\epsilon)P_{kk}P_{mm'}\right)=J_1+J_2+J_3\tag{A.9}$$

由假设 A3，对任意的 $k'\neq k$，$l'\neq l$，有

$$(1+\epsilon)P_{ll}P_{kk'}-(1-\epsilon)P_{kk}P_{mm}<0\tag{A.10}$$

$$(1+\epsilon)P_{ll'}P_{kk'}-(1-\epsilon)P_{kk}P_{mm'}<0\tag{A.11}$$

因此 $J_1\leqslant 0$，$J_3\leqslant 0$.

现在估计 J_2.

$$J_2\leqslant\epsilon(3+\epsilon^2-3\epsilon)\sum_{k'\neq k}\sum_{m=1}^{K}\sum_{m'\neq m}\chi_{km}\chi_{k'm'}\chi_{kl}\chi_{kl'}P_{ll}P_{kk'}\tag{A.12}$$

由 ϵ 的任意性可得

$$F_K(\mathscr{D})-F_K(\chi)\leqslant 0\tag{A.13}$$

由(A.9)～(A.13)可见，对任意的 k,m,l，(A.13)中等号成立当且仅当 $\chi_{kl}\chi_{km}=0$ 时成立.此时，$\chi=\mathscr{D}\mathscr{O}$.

令 $a_n=(\gamma_n n)^{-\frac{1}{4}+\frac{a}{2}}$，$b_n=\max\limits_{\chi:F_K(\mathscr{D})-F_K(\chi)\leqslant a_n}\min\limits_{\mathscr{O}\in\mathscr{O}_K}\|\chi-\mathscr{D}\mathscr{O}\|_1$，那么，$a_n\to 0$，$n\to\infty$.易见 $b_n\to 0$，若 $b_n>0$，那么由连续型可知(A.13)中等号在 $\chi\neq\mathscr{D}\mathscr{O}$ 时成立，矛盾.

由引理 A.1 和引理 A.2 可得

$$\max_{\vec{e},\vec{\varphi}}|\mathrm{Ncut}(\vec{e},\vec{\varphi})-F_K(\chi(\vec{e}))|=O_P((\gamma_n n)^{-\frac{1}{2}})\tag{A.14}$$

从而，$\Pr\left(|\mathrm{Ncut}(\hat{\vec{e}},\hat{\vec{\varphi}})-F_K(\chi(\hat{\vec{e}}))|>a_n\right)\to 0$，$\Pr\left(|\mathrm{Ncut}(\vec{\nu},\hat{\vec{\varphi}})-F_K(\mathscr{D})|>a_n\right)\to 0$，其中 $\hat{\vec{e}},\hat{\vec{\varphi}}$ 包括类标签的转置.由此可得，当 $n\to\infty$ 时，$\Pr\left(F_K(\mathscr{D})-F_K(\chi(\hat{\vec{e}}))\leqslant a_n\right)\to 1$.故当 $n\to\infty$ 时，$d(\hat{\vec{e}},\vec{\nu})=\dfrac{1}{2}\min\limits_{\mathscr{O}\in\mathscr{O}_K}\|\chi(\hat{\vec{e}})-\mathscr{D}\mathscr{O}\|\leqslant b_n\to 0$. 定理 2.1 得证.

《中国科学技术大学学报》征稿简则

《中国科学技术大学学报》由郭沫若、华罗庚和严济慈等一大批老一辈科学家亲手于 1965 年在北京创刊，先后有 39 位院士担任编委.由中国科学院主管，中国科学技术大学主办，为综合性自然科学国家重点级核心学术期刊(月刊，国内外公开发行)，主要刊登具有创新性、高水平的学术论文以及由科学大家或知名教授撰写的反映学科前沿的评述，并且开辟专家论坛，就一些重大科学研究问题进行有益的讨论.

欢迎国内外学者投稿，中英文稿均可.

1　栏目

本刊设研究论文、研究快报、特约评述和专家论坛等栏目.

1.1　*研究论文*　介绍某一课题高水平研究成果.来稿要求内容充实，推论严谨，数据可靠、完整，文字精练，结论正确.欢迎发表系列论文和团队论文.

1.2　*研究快报*　简要、快速报道某一研究工作取得重要进展或突破的创新性、高水平的阶段性成果和主要结论.要求方法从简、数据完整，结论明确，篇幅不超过 4000 字.

1.3　*特约评述*　综述某一重要研究领域的代表性成果，评论研究现状，提出尚待解决的问题，并指明今后研究方向.一般约请科学大家或知名教授撰写，作者亦可向编辑部自荐.

1.4　*专家论坛*　就科学研究重大基础问题、前沿问题或热点问题提出解决问题的新思路，发表不同的见解或进行必要的有益讨论.

2　投稿要求和注意事项

2.1　*正文书写顺序*　标题(一般不超过 20 个汉字)、作者姓名、作者单位，所在城市及邮政编码、中文摘要、关键词(3～8 条)、中图分类号(数学稿还须提供 AMS Subject Classification，参照 http://just.ustc.edu.cn)、与中文相对应的英文标题、作者姓名(汉语拼音，姓前名后，姓全大写，名首字母大写)、作者单位译名、英文摘要、英文关键词、正文、参考文献.若为英文稿，题名不超过 100 个字符，书写顺序同上.

在文稿首页地脚处注明基金资助项目名称及项目号(*将作为论文评审时参考的重要背景资料*)，并对第一作者(姓名，性别，出生年，学位/职称，目前主要从事的研究方向及 E-mail)与通讯作者(姓名，学位(博士以上才注)/职称(教授以上才注)，E-mail 及必要的联系电话)简要介绍.通讯作者是课题负责人或导师，要负责及时对读者的问题给予解答.

2.2　*对摘要的要求*　摘要内容应包括有与论文同等量的主要信息，应说明研究目的、采用的方法、研究成果及结论四个部分.中英文摘要需对应.中文摘要约 250 个汉字，英文摘要约 1500 个字符.请参照 EI，SCI 要求，避免使用“This paper，in this paper(本文)”或“I(我)”等，用词要客观，尽量减少不必要的修饰.

2.3　*对量、单位及符号的要求*　文中物理量、计量单位及符号的使用必须符合国际标准和国家标准(GB3100-93～GB3102-93).正确书写易混淆的外文字母的文种、大小写、正斜体、黑白体及上下角标.

2.4　*对图、照片、表的要求*　文中图要直观、简明、清晰.图中的文字、符号、纵横坐标必须写清，并与正文保持一致.

图版、照片必须图像清晰，层次分明，请提供矢量图或线条图，不接收扫描图；可根据作者需要印刷彩页.

表的格式采用三线表，必要时可加辅助线，所用文字、符号、单位要与正文一致.

图、图版、照片、表均要求提供中、英文对照的图题、表题.

2.5　*对参考文献的要求*　参考文献必须标全并注意引用国内外及本刊的最新文献.按在文中出现的先后次序列于文后，用数字加方括号表示，如[1]，[2]，…，与正文中的指示序号一致.

本刊执行国家标准《文后参考文献著录规则》(GB/T 7714-2005)和《中国学术期刊(光盘版)检索与评价数据规范》，并参照 EI，SCI 要求.中文参考文献请先按下列格式完整列出其英文，然后完整著录出中文.各类参考文献条目的编排格式如下：