计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(10):265−274 [doi: 10.15888/j.cnki.csa.009254]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

网络嵌入随机块模型的社区发现和链路预测①

刘　杰[1,2], 丁靖芝[1,2]

[1](中国科学技术大学 管理学院, 合肥 230026)
[2](中国科学技术大学 国际金融研究院, 合肥 230026)
通信作者: 丁靖芝, E-mail: dingjz1998@mail.ustc.edu.cn

摘　要: 社区发现与链路预测任务是网络数据研究中的热点问题, 兼顾网络传递性与区块结构有助于捕捉个体之间的有效关联、探测数据中蕴含的内在规律, 帮助研究者挖掘更多数据价值进而做出决策. 当前的算法与模型多侧重于网络传递性或区块结构单一层面的分析, 且依赖一定的假设条件. 本文提出网络嵌入随机块模型 (NE-SBM) 用于社区发现与链路预测. 搭建贝叶斯框架完成模型参数的正则化, 利用 Metropolis Hasting-Gibbs 算法获得节点嵌入表示的隐位置与社区隶属关系, 基于多维尺度变换算法解决隐位置可识别性问题. 本方法可解决传统启发式算法中过分依赖判断准则或评价函数的问题, 对各类型的数据都具有更好的适应性. 人工数据及真实数据的实验结果进一步验证了该方法在社区发现与链路预测中有更优的表现.

关键词: 网络嵌入; 随机块模型; 社区发现; 链路预测; 吉布斯采样; 多维尺度变换

引用格式: 刘杰,丁靖芝.网络嵌入随机块模型的社区发现和链路预测.计算机系统应用,2023,32(10):265–274. http://www.c-s-a.org.cn/1003-3254/9254.html

Community Detection and Link Prediction Based on Network Embedding Stochastic Blockmodel

LIU Jie[1,2], DING Jing-Zhi[1,2]

[1](School of Management, University of Science and Technology of China, Hefei 230026, China)
[2](International Institute of Finance, University of Science and Technology of China, Hefei 230026, China)

Abstract: Community detection and link prediction are hot issues in network data research. Taking into account both network transitivity and block structure can help capture the effective association between individuals and detect the inherent patterns in the data, thus helping researchers explore more data values and make decisions. Most of the current algorithms and models focus on single-level analysis of network transitivity or block structure, and they rely on certain assumptions This study proposes a network embedding stochastic blockmodel (NE-SBM) for community detection and link prediction. A Bayesian framework is built to regularize the model parameters, and the Metropolis Hasting-Gibbs algorithm is applied to obtain the hidden location and community affiliation represented by node embedding. The study also takes advantage of the multidimensional scaling algorithm to solve the hidden location identifiability problem. The proposed method can solve the problem of over-reliance on judgment criterion or evaluation function in traditional heuristic algorithms and has better adaptability to all types of data. In addition, the experimental results on artificial and real data further validate the superior performance of the method in community detection and link prediction.

Key words: network embedding; stochastic blockmodel; community detection; link prediction; Gibbs sampling; multidimensional scaling

① 基金项目: 国家自然科学基金 (71771201)
收稿时间: 2023-03-17; 修改时间: 2023-04-20; 采用时间: 2023-04-28; csa 在线出版时间: 2023-07-21
CNKI 网络首发时间: 2023-07-21

网络数据是复杂系统中个体之间相互关系的一种抽象表示, 网络数据研究的重点议题包括社区发现和链路预测. 社区发现利用高组内连通性和低组间连通性划分网络节点, 划分的节点类被称作社区; 链路预测作为社区发现的下游任务, 探测具有较高互动可能性的个体间的结构, 预测缺失或未来潜在连边. 如在学术合著网络[1] 中, 社区发现探寻作者合著团体与研究方向, 有利于揭示合著产生机制、促进新的合著; 链路预测挖掘潜在的合著可能, 有利于合著关系推荐. 社区发现能揭示网络数据结构形态, 在生物医疗、物流通讯等领域有着广泛的应用场景, 有助于提高人们的工作效率与生活质量. 链路预测可以帮助研究者挖掘网络数据中的潜在关联, 在社交推荐、资源分配等场景发挥重要作用, 为决策分析提供科学依据.

社区发现可被视作网络节点聚类问题, 经典启发式算法包括基于层次聚类的分裂算法 (GN)[2], 贪心算法 (FN)[3] 等. 链路预测算法主要以基于网络拓扑结构、基于动力学的算法为主, 如共同邻居算法 (CN)[4]、优先连接算法 (PA)[5] 等. 这类启发式算法依赖于现实网络形态的简单假设和离散的优化学习方式, 但难以描述更复杂的网络链路数据[6]. 在启发式算法的基础上, 生成模型利用具体概率分布对网络进行建模, 将社区发现和链路预测问题转化为统计推断问题.

随机块模型 (stochastic blockmodel, SBM)[7] 是网络分析中的经典模型, 能刻画不同类型的网络结构、还原网络生成过程. 此外随机块模型中的区块结构使其在社区发现任务上有天然的优势, 能够挖掘网络隐含结构与潜在联系, 常应用于推荐算法、风控反欺诈、传染病爆发范围预测等领域. 赵学华等[8] 提出精细随机块模型及对应快速学习方法降低时间复杂度. Yu 等[9] 提出混合隶属度随机块模型解决动态网络社区发现问题. Liu 等[10] 考虑加权随机块模型与块状学习算法分析大规模网络. Noroozi 等[11] 提出稀疏随机块模型, 加入惩罚项提高社区发现准确率. Chen 等[12] 基于随机块模型讨论多层网络中的社区发现问题. Legramanti 等[13] 利用无监督方法同时学习随机块模型中社区个数与节点-社区隶属关系. 然而随机块模型中同一区块内的节点等价, 因此无法描述网络中的传递性与相互性特征, 在链路预测中是粗糙的[14], 而网络嵌入思想的引入能有效解决这些问题.

网络嵌入 (network embedding) 将网络结构特征等信息映射为节点向量化表示, 基于此完成聚类、分类、预测等任务, 是网络表征学习的一种方法[15]. 网络嵌入为网络数据可视化提供一种新的视角, 更直观地展现了节点之间的连接紧密程度与相似程度, 为研究者决策分析、政策规划提供参考与指导. 隐空间模型是一类基于网络嵌入的生成模型, Hoff 等[16] 提出两类基于节点隐位置 (latent position) 的隐空间模型: 距离模型 (latent space distance model, LSM) 和投影模型 (latent space projection model, LPM). Handcock 等[17] 基于此提出隐空间聚类模型 (latent position cluster model, LPCM) 用于社区发现. Chang 等[18] 引入节点受欢迎程度, 简化计算步骤. Sewell 等[19] 建立动态隐空间模型用于社区发现与链路预测. Zhang 等[20] 考虑有向网络中节点的发送、接收模式, 推断社区与节点嵌入表示. Liu 等[21] 考虑多个隐空间中的网络嵌入探寻社区. Sosa 等[22] 和 MacDonald 等[23] 讨论多层网络下的隐空间模型与模型选择方法. 遗憾的是, 生物网络、脑网络等数据中常存在异配结构, 异配网络中有差异的节点之间更有可能建立连接, 而这类模型与算法倾向于连接具有相似结构的节点, 在刻画异配网络结构时有一定局限性[24], 进而影响社区发现的效果.

现有网络模型与算法通常集中在社区发现或链路预测的单一任务中, 仅关注网络区块结构或个体互动传递性的单一层面[25]. 而现实网络结构复杂, 往往跨越多个研究层面, 当前模型与算法不能完全刻画出现实网络的多样性, 为社区发现与链路预测任务带来挑战. 为了解决上述问题并满足现实需求, 我们在经典随机块模型算法的基础上借助网络嵌入的思想, 通过更具普适性的网络嵌入随机块模型 (network embedding stochastic blockmodel, NE-SBM) 构建网络, 采用 Metropolis Hasting-Gibbs 算法和可识别性调整算法得出节点-社区隶属关系完成社区发现任务, 并基于条件概率完成下游链路预测任务, 该方法在这两类任务中均有很好的表现.

本文有如下两方面的创新.

(1) 在网络生成方面, 网络建模的方式直接捕获网络结构信息, 还原网络生成过程. 本文将网络嵌入与随机块模型结合, 联合节点隶属社区关系与隐位置投影刻画有向网络连接模式. 本文提出的 NE-SBM 模型克服了传统随机块模型在链路预测、隐空间模型在社区发现任务中的不足, 可以视作两者的广义形式, 放宽了

对网络数据的要求, 更具普适性.

(2) 在算法方面, 首先构建贝叶斯框架完成模型参数正则化, 现实网络数据中常存在孤立节点, 传统基于极大似然估计的算法通常失效, 贝叶斯框架使得这些孤立节点具有与其他节点相连的正概率, 有利于链路预测任务; 其次, 利用 Metropolis Hasting-Gibbs 算法与增补变量法进行参数推断, 规避传统变分推断、EM 算法中部分参数无显式迭代表达式的问题; 最后, 利用加权多维尺度变换解决隐位置的可识别性问题. 实验证明了模型框架与算法的有效性. 相比之下, 启发式算法需要依据先验知识指定判断网络结构的准则或评价函数, 先验难以选取且直接影响最终分析结果.

1 网络嵌入随机块模型下的网络构建

本节通过网络嵌入随机块模型 (network embedding stochastic blockmodel, NE-SBM) 构建网络, 对节点的社区隶属关系的生成、节点嵌入以及边的连接模式做详细说明.

对于包含N个节点有向网络$G=(V,E)$, V 表示节点集合且$|V|=N$, E 表示边集合. N 维邻接矩阵 Y 以二维数组的形式描述节点之间的连接关系, $Y_{ij}=0$代表节点 i 和节点 j 之间没有连边; 反之, $Y_{ij}=1$代表存在由节点 i 发出, 指向节点 j 的有向边. 不考虑自连接的情形, 邻接矩阵的对角元均为 0.

节点-社区隶属关系分配: 假定网络中社区个数为 K, 向量$\gamma=(\gamma_1,\cdots,\gamma_N)$表示 N 个节点的社区隶属关系, 称为社区标签, 每个节点的社区标签相互独立, 服从多项分布, 如式 (1) 所示:

$$\gamma_i|\pi\sim\text{Multinomial}(1;\pi_1,\cdots,\pi_K),\quad i=1,\cdots,N \tag{1}$$

网络嵌入下节点隐位置表示: 利用网络嵌入的思想将节点映射到一个 d 维连续隐空间, 该隐空间能够体现原网络的结构、特征以及其他信息. 记节点对应的隐位置为Z_i, 假定每个节点的隐位置相互独立, 且服从于均值为 0 的多元正态分布:

$$Z_i\sim\text{MVN}\left(0,\left(\sigma_{\gamma_i}^2\right)I_d\right),\quad i=1,\cdots,N \tag{2}$$

社区间随机块模式: 如果节点 i 和节点 j 的社区标签不同, 两节点分别隶属于 k 社区和 l 社区, 假定产生连边的概率与两节点的社区标签及它们的发送、接收效应有关, 即:

$$Y_{ij}|\gamma_i=k,\gamma_j=l,s_i,r_j,k\neq l\sim\text{Bern}\left(\text{logit}^{-1}\left(\eta_{kl}+s_i+r_j\right)\right) \tag{3}$$

向量$s=(s_1,\cdots,s_N)$和$r=(r_1,\cdots,r_N)$为网络中 N 个节点的接收、发送效应, η为 K 维社区间系数矩阵, 其对角元都为 0. 定义:

$$\theta:=(\eta_{12},\cdots,\eta_{1K},\cdots,\eta_{K,K-1},s_1,\cdots,s_N,r_1,\cdots,r_N)^{\text{T}}$$

此时, 社区间边连接分布可以简化为:

$$Y_{ij}\,|\,\gamma_i=k,\gamma_j=l,s_i,r_j,k\neq l\sim\text{Bern}(\text{logit}^{-1}(X_{ij}^{\text{T}}\theta)) \tag{4}$$

矩阵 X 是所有块间节点对关于θ的系数矩阵, 其维数为$\sum_{i\neq j}I\left(\gamma_i\neq\gamma_j\right)\times\left(2N+K^2-K\right)$. X_{ij}表示节点对(i,j)对应的系数向量. 基于 Ng 等[24] 工作中有关可识别性的讨论, 我们对应地给出本模型的可识别性条件:

$$\sum_{i:\gamma_i=k}s_i=0,\quad\sum_{i:\gamma_i=k}r_i=0 \tag{5}$$

社区内网络嵌入模式: 如果节点 i 和节点 j 的社区标签相同, 隶属于社区 k, 假定产生连边的概率与节点隐位置和社区标签这两个因素有关, 利用节点隐位置投影刻画节点对产生连接的可能性, 连接概率的表达式可以写成如下形式:

$$\text{logit}\,P\left(Y_{ij}=1\,|\,\gamma_i=\gamma_j=k,Z_i,Z_j\right)=\beta_k+\frac{Z_i^{\text{T}}Z_j}{|Z_j|} \tag{6}$$

即: $Y_{ij}\,|\,\gamma_i=\gamma_j=k,Z_i,Z_j\sim\text{Bern}\left(\text{logit}^{-1}\left(\beta_k+\dfrac{Z_i^{\text{T}}Z_j}{|Z_j|}\right)\right)$.

随机块模型的区块结构揭示了社区发现结果, 下游的链路预测结果由概率矩阵得到, 但同一区块内节点连接概率相等, 难以刻画网络传递性与相互性, 因此链路预测结果是粗糙的; 隐空间模型结合网络嵌入的思想, 社区发现结果常由隐位置聚类得到, 链路预测结果则是依据节点隐位置间距离、投影等度量计算连接概率得到, 但由于缺少区块结构, 在社区发现任务中适用范围不及随机块模型广泛.

NE-SBM 模型的社区内网络嵌入模式克服了随机块模型难以刻画网络传递性和相互性的困难; 社区间随机块模式弥补了隐空间模型在刻画社区结构上的不足. 在不区分社区内与社区间网络连接模式的情况下, 随机块模型和隐空间模型可以视作 NE-SBM 模型的特殊形式, 因此 NE-SBM 模型更具普适性, 相比基础模型更加灵活, 平衡了简明性与丰富性, 可以适应于更多类型的网络数据, 在社区发现与链路预测任务中都能有较好的表现.

2 算法实现

本节主要对社区发现与链路预测算法实现做详细说明. 首先构建贝叶斯框架指定先验, 完成模型参数正则化, 之后利用 Metropolis Hasting-Gibbs 算法进行后验采样得到节点社区标签等参数的估计, 最后利用重新标注算法和加权多维尺度变换分别对节点社区标签与节点隐位置做可识别性调整, 得到社区发现结果, 并计算节点间存在边的条件概率得到链路预测结果.

2.1 贝叶斯框架的构建

针对 NE-SBM 的模型参数构建贝叶斯框架, 假设模型参数$\pi=(\pi_1,\cdots,\pi_K)$, $\beta=(\beta_1,\cdots,\beta_k)$, θ的先验分布如下:

$$\pi\sim \text{Dirichlet}\,(T,\cdots,T) \tag{7}$$

$$\beta_k\sim N\left(0,\sigma_\beta^2\right),\quad k=1,\cdots,K \tag{8}$$

$$\theta\sim \text{MVN}\left(0_{2N+K^2-K},\sigma_\theta^2 I_{2N+K^2-K}\right) \tag{9}$$

Nowicki 等[7] 提出, 选取较小的 T 会导致推断结果倾向于大小不平衡的社区结构, 本文设定$T=10$防止有大小几乎为 0 的社区产生. 此外, 实验配置与实证分析中我们设定先验分布标准差$\sigma_\beta=5$, $\sigma_\theta=5$, $\sigma_z=3$, 此参数设定先验分布下的模型有较好的预测表现.

模型的似然表达式为:

$$P(Y\mid\gamma,Z,\beta,\theta)=\prod_{i=1}^{N}\prod_{j\neq i}^{N}p_{ij}^{Y_{ij}}\left(1-p_{ij}\right)^{1-Y_{ij}} \tag{10}$$

基于上述先验分布的假设, 我们可以得到部分参数的后验条件分布:

$$\begin{aligned}&P\left(\gamma_i=k|\text{其他参数}\right)\\&\propto\pi_k\prod_{j\neq i}\{p_{ij}^{Y_{ij}}(1-p_{ij})^{1-Y_{ij}}\}\{p_{ji}^{Y_{ji}}(1-p_{ji})^{1-Y_{ji}}\}\end{aligned} \tag{11}$$

$$\pi\sim\text{Dirichlet}(T+n_1,\cdots,T+n_K) \tag{12}$$

其中, n_k为社区 k 中包含的节点数, $k=1,\cdots,K$, p_{ij}由式 (6) 和式 (3) 计算得出.

2.2 节点社区标签估计: Metropolis Hasting-Gibbs 算法

考虑到传统的 EM 算法与变分推断算法下, 部分参数不具有显式迭代表达式, 通过数值方法求解耗时较长. 我们采用 Metropolis Hasting-Gibbs 算法进行逐分量采样求解模型.

隐位置与社区标签的成对更新: 由于节点-社区隶属关系与其隐位置相关, 因此在采样的过程中, 对这两个变量成对更新. 给定节点 i, 从当前参数对(γ_i^t,Z_i^t)采样下一个参数对(γ_i^*,Z_i^*) 的提议分布形式为:

$$q\left(\left(\gamma_i^*,Z_i^*\right)|\left(\gamma_i^t,Z_i^t\right)\right)=q(\gamma_i^*|\gamma_i^t,Z_i^t)q(Z_i^*|\gamma_i^t,\gamma_i^*,Z_i^t) \tag{13}$$

$$\begin{aligned}&q(\gamma_i^*=k|\gamma_i^t,Z_i^t)=q(\gamma_i^*=k|Z_i^t)\\&\propto\pi_k\prod_{j\neq i}\left\{p_{ij}^{Y_{ij}}(1-p_{ij})^{1-Y_{ij}}\right\}\left\{p_{ji}^{Y_{ji}}(1-p_{ji})^{1-Y_{ji}}\right\}\end{aligned} \tag{14}$$

$$\begin{aligned}&q(Z_i^*\mid\gamma_i^t,\gamma_i^*,Z_i^t)=\\&\begin{cases}\text{MVN}(Z_i^*;Z_i^t,\delta_z^2I_d),\gamma_i^*=\gamma_i^t\\\text{MVN}(Z_i^*;\overline{Z_{\gamma_i^*}},\delta^2I_d),\gamma_i^*\neq\gamma_i^t\text{且}\sum\limits_j I_{(\gamma_j^t=\gamma_i^*)}>0\\\text{MVN}(Z_i^*;0_d,\sigma_z^2I_d),\gamma_i^*\neq\gamma_i^t\text{且}\sum\limits_j I_{(\gamma_j^t=\gamma_i^*)}=0\end{cases}\end{aligned} \tag{15}$$

其中, $\overline{Z_{\gamma_i^*}}$表示除去节点 i 外, 社区标签为γ_i^*的节点的隐位置均值. 给定节点 i, 当前的参数对(γ_i^t,Z_i^t)转移到(γ_i^*,Z_i^*)的概率为:

$$\alpha=\frac{P(Y_N|\gamma^*,Z^*,\theta,\beta)\,p(\gamma_i^*|\pi)p(Z_i^*|\sigma_z)q\left(\left(\gamma_i^t,Z_i^t\right)|\left(\gamma_i^*,Z_i^*\right)\right)}{P(Y_N|\gamma^t,Z^t,\theta,\beta)\,p(\gamma_i^t|\pi)p(Z_i^t|\sigma_z)q\left(\left(\gamma_i^*,Z_i^*\right)|\left(\gamma_i^t,Z_i^t\right)\right)} \tag{16}$$

为了保证算法效率, 使得隐位置 Z 对应的马氏链收敛更快, 在同时更新参数对(γ_i^t,Z_i^t)之后, 固定γ_i^t的值对Z_i^t再做一次随机游走迭代.

辅助变量下社区间连边参数的更新: 在模型设定和先验分布假设下, 并不能直接得出关于社区间连边参数θ的后验条件分布表达式, 因此我们利用增补变量法[26], 引入条件分布为 Pólya-Gamma 分布的辅助变量$\omega=(\omega_{ij})_{i\neq j:\gamma_i\neq\gamma_j}$, 以获得$\theta$的显式条件分布:

$$\omega_{ij}|\text{其他参数}\sim\text{PG}(1,X_{ij}(\gamma)'\theta) \tag{17}$$

定义$\Omega=\text{diag}(\omega_{ij})$, $u_{ij}=\left(Y_{ij}-\dfrac{1}{2}\right)\omega_{ij}^{-1}$, $u=(u_{ij})_{i\neq j}$, 此时有:

$$\theta|\omega\sim\text{MVN}(m,V) \tag{18}$$

其中,

$$m=VX^{\text{T}}\Omega u \tag{19}$$

$$V=\left(X^{\text{T}}\Omega X+\left(\sigma^\theta\right)^2I_{2N+K^2-K}\right)^{-1} \tag{20}$$

基于上述推导, NE-SBM 模型的 Metropolis Hasting-Gibbs 后验采样算法如算法 1 所示. 具体来说, 交替采样更新模型参数, 基于辅助变量更新社区间连边参数, Markov 链收敛至平稳分布后可得到后验样本.

算法 1. Metropolis Hasting-Gibbs 后验采样算法

(1) 对于γ_i, Z_i, $i=1,\cdots,N$:

1) 按照式 (13) 成对采样(γ_i^*, Z_i^*);

2) 按照式 (16) 转移概率接受$(\gamma_i^{t+1}, Z_i^{t+1})=(\gamma_i^*, Z_i^*)$, 否则$(\gamma_i^{t+1}, Z_i^{t+1})=(\gamma_i^t, Z_i^t)$;

3) 采样$Z_i^*\sim \mathrm{MVN}\left(Z_i^{t+1}, \delta_z^2 I_d\right)$, 以$\alpha=\frac{P(Y_N|\gamma, Z^*, \theta, \beta)p(Z_i^*|\sigma_z)}{P(Y_N|\gamma, Z^{t+1}, \theta, \beta)p(Z_i^{t+1}|\sigma_z)}$的转移概率接受$Z_i^{t+1}=Z_i^*$, 否则$Z_i^{t+1}=Z_i^t$.

(2) 对于β_k, $k=1,\cdots,K$:

1) 采样$\beta_k^*\sim N\left(\beta_k^t, \delta_\beta\right)$;

2) 按照$\frac{P(Y_N|\gamma^*, Z^*, \theta, \beta_k^*)p(\beta_k^*;\mu^\beta, \sigma_\beta)}{P(Y_N|\gamma^*, Z^*, \theta, \beta_k^t)p(\beta_k^t;\mu^\beta, \sigma_\beta)}$的转移概率接受$\beta_k^{t+1}=\beta_k^*$, 否则$\beta_k^{t+1}=\beta_k^t$.

(3) 按照式 (18) 采样更新θ, 做以下变换以保证满足式 (5) 中可识别性条件成立且式 (4) 值不变:

$$\eta_{kl'}=\eta_{kl}+\frac{1}{n_k}S_k+\frac{1}{n_l}R_l,\ s_{i'}=s_i-\frac{1}{n_k}S_k,\ r_{i'}=r_i-\frac{1}{n_k}R_k$$

(4) 按照式 (12) 采样更新π.

算法 1 提供了对 NE-SBM 模型参数进行后验采样的方法, 在得到后验样本之后, 计算后验均值得到模型中的参数估计, 计算后验众数得到节点社区标签估计, 经可识别性调整后即可得到社区发现结果.

2.3 可识别性调整: 重新标注算法与加权多维尺度变换算法

在得到模型参数的估计结果之后, 我们需要考虑节点社区标签的可识别性问题 (label switching problem). 事实上, 对节点社区标签做置换变换, 模型的似然值是不变的, 本文借助重新标注算法 (relabelling algorithm)[27] 来解决这种可识别性问题, 得到最终的社区发现结果.

此外, 隐位置的可识别性问题同样值得我们注意. 式 (10) 的值在旋转、反射这两种变换下不发生改变, 因此隐位置不唯一. 我们针对 Fosdick 等[25] 提出的基于加权多维尺度变换 (multidimensional scaling, MDS) 的算法加以改进, 将之拓展到有向网络的情形.

设后验样本的样本量为 M, 社区 k 中所有具有非零后验概率的节点集合为S_{k0}, 样本 m 中社区标签为 k 的节点集合为S_{km}. 基于算法 1 中得到的后验样本计算对应权重矩阵与距离矩阵, 利用 SMACOF 算法[28] 重构最佳相对位置, 依据后验样本中的社区标签利用 Procrustes 变换[16] 调整节点隐位置. 具体如算法 2 所示.

算法 2. 加权多维尺度变换下的隐位置可识别性调整

(1) 对于社区 k, $1\leqslant k\leqslant K$:

1) 计算权重矩阵W_{k0}: 对$i\neq j\in S_{k0}$, $(W_{k0})_{ij}$为后验样本中节点 i 和节点 j 社区标签都为 k 的样本数量;

2) 计算距离矩阵D_{k0}: 对$i\neq j\in S_{k0}$, $(D_{k0})_{ij}$为后验样本中节点 i 和节点 j 隐位置的平均距离, 若$(W_{k0})_{ij}=0$, 则$(D_{k0})_{ij}$缺失;

3) 基于 d 维平均距离矩阵D_{k0}, 利用 SMACOF 算法[28] 重构最佳相对位置, 得到重构后的隐位置集合Z_{k0}.

(2) 对每个后验样本 m, $1\leqslant m\leqslant M$:

1) 选取Z_{k0}中与集合S_{km}标签一致的子集$\tilde{Z}_{k0m}$;

2) 利用 Procrustes 变换[16] 得到调整后的节点隐位置$\tilde{Z}_{km}$:

$$\tilde{Z}_{km}=\arg\min_{TZ_{km}}\left\{tr\left(\tilde{Z}_{k0m}-TZ_{km}\right)^{\mathrm{T}}\left(\tilde{Z}_{k0m}-TZ_{km}\right)\right\}$$

其中, T 代表所有的反射变换、旋转变换.

算法 2 提供了调整节点隐位置可识别性的方法, 针对算法 2 中得到的隐位置估计, 结合式 (3) 和式 (4) 计算节点之间的连边概率即可得到链路预测结果.

3 实验及结果分析

3.1 实验设置

实验硬件环境为 AMD Ryzen 75800H 处理器, 运行内存为 16 GB, 操作系统为 Windows 10, 实验采用的语言是 R 4.1.3.

本节旨在验证 NE-SBM 模型及其算法在社区发现与链路预测任务上, 相比基准模型有更佳的表现效果. 具体地, 参照 Ng 等[24] 的实验设置, 利用不同参数设定下的 NE-SBM 模型、隐空间聚类模型 (LPCM)[17] 和随机块模型 (SBM)[7] 模拟生成 3 种人工网络数据集. 在每种数据集中, 设定网络节点个数$N=100$, 社区个数$K=3$, 隐空间维数$d=2$, 节点隶属于每个社区的概率均等, 重复 30 次以得到一般性的结论. 3 种人工网络数据集的生成方式如下.

数据集 1: 由 NE-SBM 模型生成的网络数据, 考虑社区间概率连接矩阵η中非对角元λ的不同取值.

数据集 2: 由 LPCM 模型生成的网络数据, 考虑隐空间距离参数β的不同取值[17].

数据集 3: 由 SBM 模型生成的网络数据, 考虑连接概率矩阵参数ω的不同取值[7].

3.2 社区发现实验

本部分评估 NE-SBM 模型与对应算法在社区发现任务中的表现, 与之对比的基准模型为 SBM[7] 和 LPCM[17] 模型. 采用的社区发现评价指标为聚类错误率 (clustering error)[20] 和标准化互信息 (NMI)[29]. 聚类错误率列举所有节点社区标签的排列, 其计算表达式为:

$$clust.err=\frac{2}{N(N-1)}\sum_{i<j}I\left(I\left(\gamma_i=\gamma_j\right)+I\left(\tilde{\gamma}_i=\tilde{\gamma}_j\right)=1\right) \tag{21}$$

其中, γ代表节点的真实社区标签, $\tilde{\gamma}$代表节点社区标签

的估计值. *NMI* 值从信息论角度衡量不同划分的一致程度, 取值在 0 和 1 之间, 越接近 1 意味着当前聚类结果与真实聚类结果一致性越高, 社区划分越准确. 其定义为:

$$NMI(\gamma,\tilde{\gamma})=\frac{\sum_{k=1}^{\tilde{K}}\sum_{l=1}^{K}N_{kl}\cdot\log\frac{N\cdot N_{kl}}{\tilde{N}_k\cdot N_l}}{\sqrt{\left(\sum_{k=1}^{\tilde{K}}\tilde{N}_k\log\frac{\tilde{N}_k}{N}\right)\left(\sum_{l=1}^{K}N_l\log\frac{N_l}{N}\right)}} \tag{22}$$

其中, 算法得到的社区个数为$\tilde{K}$, $\tilde{N}_k$代表算法得到的社区 k 包含节点的数目; 真实社区个数为 K, N_l代表真实分类下社区 l 包含的节点数目, N_{kl}代表同时属于算法分类下社区 k 和真实分类下社区 l 的节点数目.

图 1–图 3 展示了 NE-SBM, LPCM 和 SBM 模型在不同数据集中社区发现结果的聚类错误率和 *NMI* 值. 相对于基准模型, NE-SBM 模型及对应算法普遍具有更低的聚类错误率与更高的 *NMI* 值, 因此在社区发现上表现更加优越. LPCM 模型不包含区块结构, 社区间的随机块模式使得 NE-SBM 模型具备区块结构, 因此相较于 LPCM 模型在挖掘社区结构上表现更佳; 此外, SBM 模型刻画区块结构的方式单一, 而社区内网络嵌入模式的引入使得 NE-SBM 模型相较 SBM 模型应用范围更广, 在不同网络结构中均能有较好的社区发现效果.

值得注意的是, 在数据集 1 和数据集 3 中, 横坐标参数的值增大意味着网络结构更加明显, 社区发现的表现随之提升; 在数据集 2 中, 横坐标参数的值增大意味着网络稠密程度增大, 社区难以区分, 社区发现的表现随之下降.

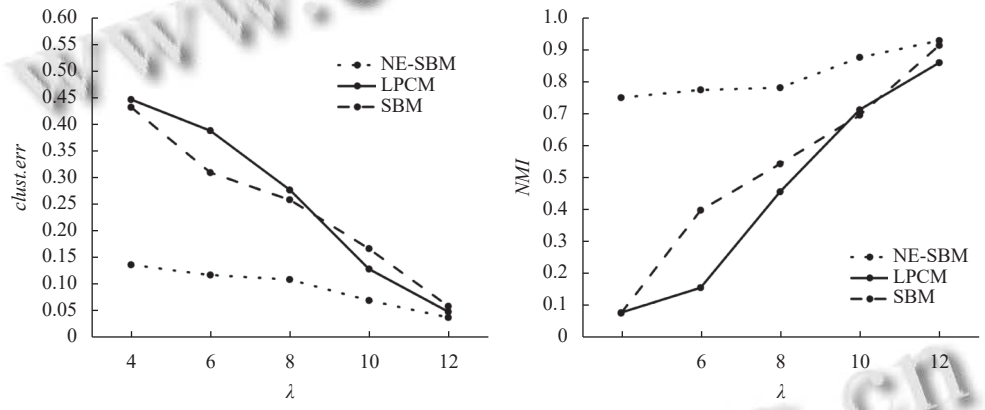

图 1 数据集 1 中 NE-SBM, LPCM, SBM 模型的社区发现结果对比

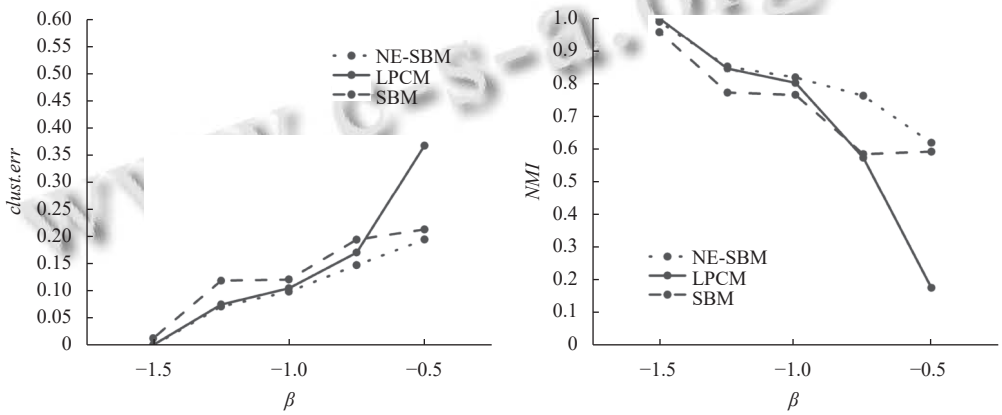

图 2 数据集 2 中 NE-SBM, LPCM, SBM 模型的社区发现结果对比

3.3 链路预测实验

为了评估 NE-SBM 模型与对应算法在下游链路预测任务中的表现, 我们从 3 种数据集中随机抽取 5% 的连边信息设置成缺失值, 利用不同模型拟合带有缺失值的网络数据, 针对缺失的连边信息进行链路预测[24]. 与 NE-SBM 模型对比的基准模型为 SBM[7]、LPCM[17]、LPM[16]、LSM[16] 模型, 评价指标为 AUC, 该值越接近 1 意味着预测效果越准确.

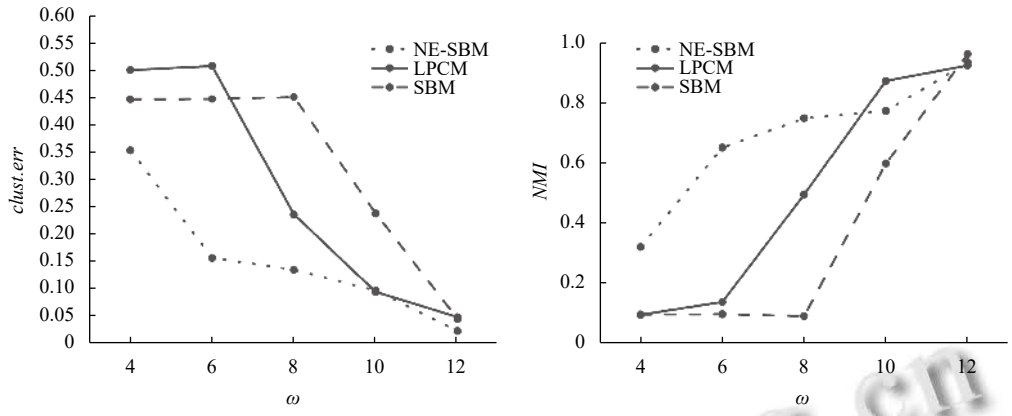

图 3 数据集 3 中 NE-SBM, LPCM, SBM 模型的社区发现结果对比

表 1 展示了不同数据集中各模型链路预测的 AUC 均值与标准差. 相对于基准模型, NE-SBM 模型及其算法有更低的 AUC 均值和较小的标准差, 因此在链路预测上有更优越的表现. 在隐空间模型中, 通过隐位置距离度量节点间连接概率的 LSM 和 LPCM 模型不能准确地刻画有向网络连接模式, LPM 模型不能结合节点隶属社区关系进行链路预测, 因此隐空间模型相较于具有区块结构且利用隐位置投影度量连接概率的 NE-SBM 模型表现欠佳; SBM 模型的同个区块内节点连接概率等同, 难以刻画网络传递性和相互性, 社区内网络嵌入模式的引入使得 NE-SBM 模型在链路预测中表现更好.

表 3 3 种数据集中链路预测 AUC 均值及标准差对比

模型	数据集1			数据集2			数据集3		
	$\lambda=4$	$\lambda=8$	$\lambda=12$	$\beta=-0.5$	$\beta=-1$	$\beta=-1.5$	$\omega=4$	$\omega=8$	$\omega=12$
NE-SBM	**0.9260** (0.0426)	**0.9444** (0.0292)	**0.9280** (0.0365)	**0.8839** (0.0734)	**0.9085** (0.0456)	**0.9120** (0.0508)	**0.9349** (0.0165)	**0.9505** (0.0213)	**0.9519** (0.0243)
LPCM	0.7566 (0.0230)	0.8130 (0.0225)	0.8834 (0.0202)	0.8691 (0.0292)	0.8685 (0.0161)	0.8726 (0.0181)	0.7674 (0.0355)	0.8173 (0.0188)	0.8858 (0.0165)
SBM	0.7401 (0.0231)	0.7545 (0.0287)	0.8236 (0.0310)	0.8648 (0.0246)	0.8748 (0.0144)	0.8711 (0.0164)	0.8446 (0.0263)	0.8089 (0.0312)	0.8255 (0.0249)
LSM	0.7552 (0.0233)	0.8129 (0.0210)	0.8833 (0.0202)	0.8686 (0.0287)	0.8669 (0.0176)	0.8726 (0.0177)	0.7668 (0.0357)	0.8171 (0.0189)	0.8852 (0.0171)
LPM	0.7578 (0.0248)	0.7769 (0.0232)	0.8400 (0.0345)	0.8720 (0.0264)	0.8711 (0.0156)	0.8754 (0.0188)	0.8146 (0.0337)	0.8307 (0.0219)	0.8787 (0.0230)

值得注意的是, 链路预测作为社区发现的下游任务, 其表现受到社区发现结果影响, 因此与社区发现的结论有相似之处. 具体地, 数据集 1 和数据集 3 中, 链路预测表现会随网络结构的明显而提升; 在数据集 2 中, 链路预测的表现会随网络稠密程度增加而变差.

4 实际数据分析

为了进一步论证 NE-SBM 模型及算法的有效性, 我们选取经典的 World Trade 网络数据集进行分析. 该数据集来自国际贸易组织的贸易方向年鉴, 描述了 1994 年 80 个国家之间金属制品的贸易量, 参考文献 [19] 中的预处理工作, 将国家视作网络中的节点, 国家之间的进口交易视作节点之间的有向连边以分析各国之间的贸易联系, 交易价值低于该国总进口量 1% 的进口数据被省略[30]. 我们旨在探索贸易体系的国家构成以及背后的世界分工现象.

我们采取 NCV (network cross-validation) 方法[31] 来确定该网络中的社区个数, 该方法依据社区划分来切割节点对并进行交叉验证, 得到最优社区个数为 5. 从社区划分来看, 5 个社区可归纳为: (1) 美国-南美洲贸易区, (2) 北美洲贸易区, (3) 亚洲贸易区, (4) 以德国、意大利为中心的欧洲-非洲贸易区, (5) 以法国、新西兰为中心的欧洲贸易区. 由此可见, 地理因素在贸易体系的形成中发挥重要作用. 全球分工体系中, 发达的

核心国家往往充当中心节点, 从事资本密集型和高科技生产, 从财富落后的外围国家进口原材料与粗加工产品, 将其加工成高科技产品出口至其他国家. 因此核心国家间贸易较多, 而外围国家间贸易较少.

参照 Zhang 等[32] 的分析思路, 计算各模型聚类结果相对于 80 个国家所属大洲的分类结果的 *NMI* 值, 可以得到 NE-SBM, LPCM, SBM 模型对应的 *NMI* 值分别为 0.522 1, 0.100 4, 0.379 4. 验证了社区划分结果与地理位置的相关性, NE-SBM 模型对应的 *NMI* 值更大, 相对其他模型能更好地刻画这一点.

NE-SBM 模型中网络嵌入的思想为网络结构可视化提供了一种新思路, 对经过可识别性调整后的节点隐位置的后验估计进行标准化处理, 使其位于单位圆上[16], 节点隐位置分布以及连边情况如图 4 所示. 标记为同种形状的节点隶属于同一个社区, 两节点在圆周上距离越近意味着它们与圆心的夹角越小, 即对应隐位置夹角越小, 节点间更有可能产生连边. 具体地, 图 4 反映出同一社区内节点连接较为稠密, 但也存在一定的异配现象, 如社区 5 中的节点在圆周上相对分散, 对应子网络密度为 0.199 3, 但是社区 5 与社区 1 中的节点连接更为紧密, 对应子网络密度为 0.251 8. 从实际意义上看, 社区 1 主要涉及欧洲核心国家, 在 1994 年进出口贸易频繁. 而社区 5 涉及少部分欧洲国家以及多数非洲国家, 多数非洲国家作为外围国家内部贸易相对较少. 图 4 的可视化方式可以清晰地展现网络节点之间的靠近程度与其社区标签的关系, 印证参数估计中体现的异配网络结构, 研究者可利用这些信息进一步分析网络结构, 优化国际贸易政策、促进经济增长和国际合作.

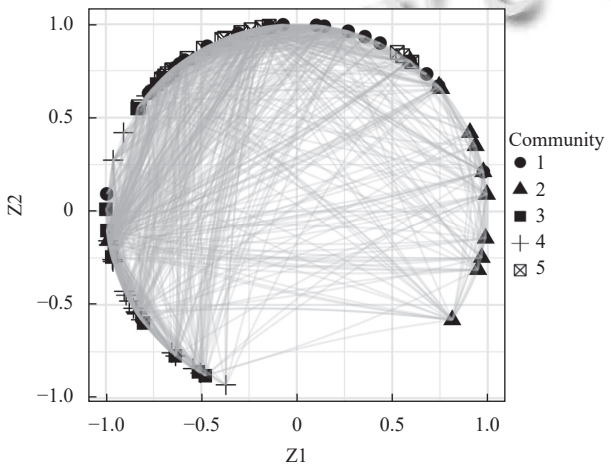

图 4 网络嵌入的可视化展示

最后, 评估 NE-SBM 模型在该数据上的链路预测表现. 针对数据进行不同随机数种子下的 10 折交叉验证, 计算链路预测结果的 AUC 平均值, 并将 NE-SBM 模型的预测结果与 SBM、LPCM、LPM、LSM 模型进行对比, 箱线图结果如图 5 所示.

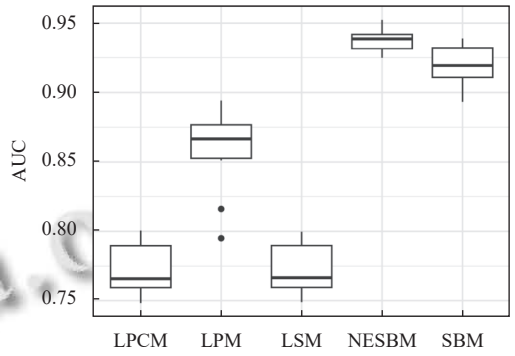

图 5 10 折交叉验证下各方法链路预测 AUC 值箱线图

NE-SBM 模型在链路预测的表现明显优于基准模型, AUC 较高并且表现稳定. SBM 模型难以刻画网络传递性, 因而表现次之; LPCM 和 LSM 模型由于采用隐位置距离刻画连接概率, 不适用于有向网络, 对应 AUC 值相对较小; LPM 模型在刻画网络异配结构上有所欠缺, 因此表现不如 NE-SBM 模型. 结合图 5 可知, NE-SBM 模型在链路预测任务中表现优越, 可以帮助研究者在探寻世界贸易体系结构的基础上, 进一步挖掘国家地区之间潜在的贸易关系, 为国家间未来的贸易合作往来提供参考.

5 结论与展望

本文在随机块模型的基础上引入网络嵌入的思想, 提出一种网络嵌入随机块模型下的社区发现与链路预测算法. 一方面, 弥补隐空间模型等基于网络嵌入的传统模型与算法在拟合异配网络结构数据中的不足, 且在社区发现任务上表现更优; 利用节点隐位置投影刻画网络连接模式, 将算法适用范围拓展到有向网络的情形, 弥补随机块模型刻画网络传递性与相互性特征的不足, 在下游链路预测任务中有更好的表现. 另一方面, 建立贝叶斯框架以应对孤立节点的干扰, 引入辅助变量, 利用 Metropolis Hasting-Gibbs 算法逐分量快速更新参数, 并针对隐位置的可识别性问题给出基于加权多维尺度变换的解决方案. 实验结果和实证分析的可视化展示均验证了该算法的有效性.

本文仍然有许多值得继续研究拓展的地方. NE-SBM 模型和对应算法可以拓展到有权网络的情形, 只需要改变节点间的连接概率分布以及相应的联系函数. 其次, 针对大型稀疏网络, 可以参照 Raftery 等[33] 的思路, 分层抽样不同度的节点, 对似然函数进行近似以降低时间复杂度至O(*N*). 此外, 网络嵌入中隐空间维数的选择也是一个值得进一步探讨的问题.

参考文献

1 张金柱, 胡一鸣. 利用链路预测揭示合著网络演化机制. 情报科学, 2017, 35(7): 75–81. [doi: 10.13833/j.cnki.is.2017.07.014]

2 Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821–7826. [doi: 10.1073/pnas.122653799]

3 Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004, 69(6): 066133. [doi: 10.1103/PhysRevE.69.066133]

4 Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019–1031. [doi: 10.1002/asi.20591]

5 Barabási AL, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509–512. [doi: 10.1126/science.286.5439.509]

6 王寒蕊, 丁岱宗, 张谧. 基于梯度的重叠式层次社区检测. 计算机系统应用, 2021, 30(8): 207–212. [doi: 10.15888/j.cnki.csa.008016]

7 Nowicki K, Snijders TAB. Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, 2001, 96(455): 1077–1087. [doi: 10.1198/016214501753208735]

8 赵学华, 杨博, 陈贺昌. 一种高效的随机块模型学习算法. 软件学报, 2016, 27(9): 2248–2264. [doi: 10.13328/j.cnki.jos.004855]

9 Yu Z, Pietrasik M, Reformat M. Deep dynamic mixed membership stochastic blockmodel. Proceedings of the 2019 IEEE/WIC/ACM International Conference on Web Intelligence. Thessaloniki: IEEE, 2019. 141–148. [doi: 10.1145/3350546.3352511]

10 Liu XY, Yang B, Chen HC, *et al*. A scalable redefined stochastic blockmodel. ACM Transactions on Knowledge Discovery from Data, 2021, 15(3): 46. [doi: 10.1145/3442589]

11 Noroozi M, Pensky M, Rimal R. Sparse popularity adjusted stochastic block model. The Journal of Machine Learning Research, 2021, 22(1): 193.

12 Chen Y, Mo DX. Community detection for multilayer weighted networks. Information Sciences, 2022, 595: 119–141. [doi: 10.1016/j.ins.2021.12.011]

13 Legramanti S, Rigon T, Durante D, *et al*. Extended stochastic block models with application to criminal networks. The Annals of Applied Statistics, 2022, 16(4): 2369–2395. [doi: 10.1214/21-AOAS1595]

14 黄丹阳. 大规模网络数据分析与空间自回归模型. 北京: 科学出版社, 2022.

15 王兴源. 基于图嵌入和 GRU 的兴趣点推荐模型. 计算机系统应用, 2021, 30(10): 40–47. [doi: 10.15888/j.cnki.csa.008161]

16 Hoff PD, Raftery AE, Handcock MS. Latent space approaches to social network analysis. Journal of the American Statistical Association, 2002, 97(460): 1090–1098. [doi: 10.1198/016214502388618906]

17 Handcock MS, Raftery AE, Tantrum JM. Model-based clustering for social networks. Journal of the Royal Statistical Society Series A: Statistics in Society, 2007, 170(2): 301–354. [doi: 10.1111/j.1467-985X.2007.00471.x]

18 Chang XY, Huang DY, Wang HS. A popularity-scaled latent space model for large-scale directed social network. Statistica Sinica, 2019, 29(3): 1277–1299. [doi: 10.5705/ss.202017.0103]

19 Sewell DK, Chen YG. Latent space approaches to community detection in dynamic networks. Bayesian Analysis, 2017, 12(2): 351–377. [doi: 10.1214/16-BA1000]

20 Zhang JN, He X, Wang JH. Directed community detection with network embedding. Journal of the American Statistical Association, 2022, 117(540): 1809–1819. [doi: 10.1080/01621459.2021.1887742]

21 Liu SH, Wang B, Liu B, *et al*. Multicommunity graph convolution networks with decision fusion for personalized recommendation. Proceedings of the 26th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Chengdu: Springer, 2022. 16–28. [doi: 10.1007/978-3-031-05981-0_2]

22 Sosa J, Betancourt B. A latent space model for multilayer network data. Computational Statistics & Data Analysis, 2022, 169: 107432. [doi: 10.1016/j.csda.2022.107432]

23 MacDonald PW, Levina E, Zhu J. Latent space models for multiplex networks with shared structure. Biometrika, 2022, 109(3): 683–706. [doi: 10.1093/biomet/asab058]

24 Ng TLJ, Murphy TB, Westling T, *et al*. Modeling the social

media relationships of Irish politicians using a generalized latent space stochastic blockmodel. The Annals of Applied Statistics, 2021, 15(4): 1923–1944.

25 Fosdick BK, McCormick TH, Murphy TB, *et al*. Multiresolution network models. Journal of Computational and Graphical Statistics, 2019, 28(1): 185–196. [doi: 10.1080/10618600.2018.1505633]

26 Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Pólya-Gamma latent variables. Journal of the American Statistical Association, 2013, 108(504): 1339–1349. [doi: 10.1080/01621459.2013.829001]

27 Stephens M. Dealing with label switching in mixture models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2000, 62(4): 795–809. [doi: 10.1111/1467-9868.00265]

28 de Leeuw J, Mair P. Multidimensional scaling using majorization: SMACOF in R. Journal of Statistical Software, 2009, 31(3): 1–30.

29 Strehl A, Ghosh J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research, 2003, 3: 583–617.

30 De Nooy W, Mrvar A, Batagelj V. Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software. 3rd ed., Cambridge: Cambridge University Press, 2018.

31 Chen KH, Lei J. Network cross-validation for determining the number of communities in network data. Journal of the American Statistical Association, 2018, 113(521): 241–251. [doi: 10.1080/01621459.2016.1246365]

32 Zhang Y, Levina E, Zhu J. Community detection in networks with node features. Electronic Journal of Statistics, 2016, 10(2): 3153–3178. [doi: 10.1214/16-EJS1206]

33 Raftery AE, Niu XY, Hoff PD, *et al*. Fast inference for the latent space network model using a case-control approximate likelihood. Journal of Computational and Graphical Statistics, 2012, 21(4): 901–919. [doi: 10.1080/10618600.2012.679240]

(校对责编: 孙君艳)