计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(1):49−57 [doi: 10.15888/j.cnki.csa.009360]

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

基于非负矩阵分解的有向网络半监督社区检测①

杨士杰, 帅　阳, 韩　超, 张伟平

(中国科学技术大学 管理学院, 合肥 230026)
通信作者: 张伟平, E-mail: zwp@ustc.edu.cn

摘　要: 有向网络上的社区检测是网络科学领域一个重要的课题. 针对这一问题, 本文提出了一种基于非负矩阵分解的有向网络半监督社区检测算法, 首先利用先验信息重构邻接矩阵, 然后使用先验信息对节点的社区隶属度进行惩罚, 并通过行归一化消除节点度异质性的影响, 最后运用交替迭代更新给出了目标函数的求解方法. 在真实网络数据上的对比实验验证了算法的有效性, 相对于基于非负矩阵分解的现有方法, 本文方法能显著提高社区发现的准确性.

关键词: 非负矩阵分解; 有向网络; 社区检测; 先验信息

引用格式: 杨士杰,帅阳,韩超,张伟平.基于非负矩阵分解的有向网络半监督社区检测.计算机系统应用,2024,33(1):49–57. http://www.c-s-a.org.cn/1003-3254/9360.html

Semi-supervised Community Detection for Directed Network Based on Non-negative Matrix Factorization

YANG Shi-Jie, SHUAI Yang, HAN Chao, ZHANG Wei-Ping

(School of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: Community detection for directed networks is an important topic in network science. Thus, this study proposes a semi-supervised community detection algorithm for directed networks based on non-negative matrix factorization (NMF). First, prior information is adopted to reconstruct the adjacency matrix and then penalize the community membership of nodes. Meanwhile, the influence of node degree heterogeneity is eliminated by row normalization, and finally, the objective function is solved using alternating iterative updates. Comparative experiments on real network datasets demonstrate the effectiveness of the proposed algorithm. Compared to existing NMF-based methods, this method can significantly improve community detection accuracy.

Key words: non-negative matrix factorization (NMF); directed network; community detection; prior information

网络科学是一个融合了数学、生物、统计学、社会学等多个学科思想与技术的交叉学科领域. 网络是由节点两两之间连边所构成. 例如交通枢纽与航线构成了交通运输网络, Zhang 等[1]应用加入协变量的随机块网络模型分析了美国航空网络; 在生物上, 构建蛋白质之间的生化反应网络对于识别同源蛋白质或者蛋白质功能模块、蛋白质功能预测等方面具有十分重要的生物学意义[2]. 网络能够捕捉系统内部相互作用的模式, 是用来抽象并研究复杂系统的重要工具[3].

社区结构是网络数据最显著的特征之一, 是探索网络结构和功能的基本属性. 社区通常被定义为一个子网络, 子网络内部的节点之间连边较为密集, 不同子网络之间的连边相对比较稀疏[4]. 例如在文献引用网络中, 节点表示论文, 边表示论文之间的引用关系, 同一

① 基金项目: 国家自然科学基金 (12171450)
收稿时间: 2023-07-03; 修改时间: 2023-08-08; 采用时间: 2023-08-18; csa 在线出版时间: 2023-11-24
CNKI 网络首发时间: 2023-11-28

研究领域中的文献互相引用的频率往往比跨领域更高, 通过分析由相同研究领域的论文构成的社区结构, 可以帮助我们更好地了解当前的热门研究方向. 社区检测指将网络中的节点按照它们之间连边的密集程度进行聚类, 从而获得网络的社区结构, 它是近年来网络科学前沿领域十分热门的研究课题. Liu 等[5]对 fMRI 数据构建网络模型, 同时实现了网络结构的恢复与社区检测, Ren 等[6]则应用正则化网络嵌入的方法实现了对社区数目的相合求解, 并完成社区检测. 在实际中, 一方面, 社区检测可以作为揭示网络局部结构、挖掘相似模式的有力工具, 例如在对于消费者的购物信息构建的网络中进行社区检测可以发现具有相似购物倾向的消费者团体, 从而有利于企业进行针对性营销, 开发更高效的推荐系统[7]. 另一方面, 社区检测允许我们将大型网络有规律地划分为小型子网络, 并对它们分别进行研究. 大数据时代构建的网络规模常常会达到上万甚至上亿的规模, 对如此规模的网络进行整体研究是十分困难的事情, 因此社区检测能有效地将网络规模减小到可操作水平. 现有的大部分方法主要针对的是无向网络上的社区检测, 然而在实际应用中, 许多网络数据本质上是定向的, 例如在原核生物基因组序列数据中, 基因组之间的供体-受体关系是通过有向网络(横向基因转移网络-LGT) 来建模的. 因此在有向网络中寻找社区结构是一项重要的任务, 在多个领域中有着广泛的应用.

目前常见的社区检测算法大致可以分为两类[1]: 第 1 类是对网络建立概率模型, 运用似然的方法完成社区检测任务, 具有代表性的是随机块模型[8]、度修正的随机块模型[9]、混合成员随机块模型[10]等. 第 2 类是定义一个用以衡量网络划分准确性的目标函数, 然后对其进行优化来获得网络的社区结构. 在第 2 类方法中, Paatero 等[11]提出的非负矩阵分解 (non-negative matrix factorization, NMF) 以其计算简便、可解释性强等优点受到广泛关注. NMF 的基本思想是用两个低秩非负矩阵来逼近目标非负矩阵, 能够对原始高维数据给出低维潜空间表示. Lee 等[12]将 NMF 应用于图像处理, 取得了卓越的效果, 此后也被广泛应用于文档聚类[13,14]、推荐系统等[15,16]领域. Wang 等[17]将 NMF 应用于网络社区检测, 并分别针对无向网络、有向网络、复合网络给出了对应的模型与求解方法; 其中针对有向网络, Wang 等[17]提出采用非负矩阵三分解对标准 NMF 进行改进, 明确了社区之间的交互作用. 在现实世界的许多网络数据中, 来自专业领域知识的先验信息是可用的, 使用先验信息的社区检测方法称为半监督社区检测方法. 合理利用这些先验信息作为指导, 可以大大提升社区检测的准确性. Zhang[18]提出了一种半监督学习框架, 通过纳入 ML (must-link) 与 CL (cannot-link) 这两种先验信息来重建邻接矩阵以指导社区检测, 让社区结构更加清晰, 取得了较好的效果. 此后, Zhang 等[19]在文献[18]的基础上, 在重建邻接矩阵的步骤中加入逻辑推断, 使得先验信息得到了更充分的利用, 同时验证了 ML 先验信息比 CL 先验信息更加有效. 文献[18]与文献[19]所提出的算法简洁高效, 但主要针对无向网络, 并且对于先验信息的运用仅限于重构网络拓扑. Yang 等[20]通过在目标函数中增加了融入先验信息的图正则化项, 以约束具有 ML 先验信息的节点对在低维空间中的距离. 与文献[18]、文献[19]类似, 该工作主要考虑无向网络上的社区检测, 同时没有考虑节点度异质性对于目标函数的影响. Liu 等[21]则进一步在目标函数中引入度对角矩阵, 消除了节点度异质性的影响. Liu 等[21]所采用的对称 NMF 针对的是无向网络上的社区检测问题, 无法拓展到有向网络.

本文提出了一种基于非负矩阵分解的有向网络半监督社区检测算法, 与现有的基于非负矩阵分解的社区检测算法相比, 具有以下创新性.

(1) 本文采用非负矩阵三分解模型考虑有向网络上的社区检测问题, 与标准 NMF 相比, 可以发现社区之间的相互作用, 更类似于真实网络.

(2) 对先验信息的利用更加充分, 除了用先验信息重构邻接矩阵, 本模型还在目标函数中添加了融入先验信息的惩罚项, 从而使具有 ML 先验信息约束的节点对更有可能被分到同一社区.

(3) 本文考虑了网络中常见的节点度异质造成的影响, 通过对矩阵 $\boldsymbol{U}$ 施加行归一化约束提高了模型的适用性. 非负双奇异值分解初始化让算法计算效率更高.

后文组织如下: 第 1 节给出问题定义, 并介绍将非负矩阵分解用于社区发现的思路. 第 2 节阐述模型与求解方法. 第 3 节介绍实际数据模拟与实验结果分析. 第 4 节进行总结, 并给出下一步工作的想法.

1 社区检测与非负矩阵分解

本节介绍社区检测的问题定义以及如何利用非负

矩阵分解进行社区检测.

1.1 先验信息指导的有向网络社区检测问题

给定一有向网络 $G=(V,E)$, 其中 V 为网络节点集合, 大小为 n, E 为节点之间连边集合, 并设节点 v_i 对应编号 i, $i=1,2,\cdots,n$. 网络中存在从节点 v_i 到 v_j 的有向边当且仅当 $(i,j)\in E$. 记网络 G 的非对称邻接矩阵为 $\boldsymbol{A}\in\{0,1\}^{n\times n}$, 满足如下条件:

$$A_{ij}=\begin{cases}1, (i,j)\in E\\ 0, (i,j)\notin E\end{cases} \tag{1}$$

上述定义给出了网络的拓扑结构. 下面考虑用先验信息指导社区检测. 由于实际中 ML 先验信息比 CL 先验信息更容易获得, 且更加有效[19], 因此这里采用 ML 先验信息, 它是指节点对属于同一社区, 在进行划分时也要尽量把具有 ML 先验信息的节点对分到同一社区. 令 ML 先验信息集合为 C, 如果 $(i,j)\in C$, 那么节点 v_i 与 v_j 属于同一社区.

我们的任务目标是: 基于给定的有向网络拓扑 $\boldsymbol{A}$ 与先验信息集合 C, 求解网络 G 的社区结构.

1.2 利用非负矩阵分解进行社区检测

考虑非负矩阵 $\boldsymbol{X}\in\mathbb{R}_+^{n\times n}$, 它代表原数据矩阵. 非负矩阵分解旨在寻找两个低秩矩阵 $\boldsymbol{U}\in\mathbb{R}_+^{n\times k}$、$\boldsymbol{V}\in\mathbb{R}_+^{k\times k}$ 来近似 $\boldsymbol{X}$. 在社区检测问题中, 原数据矩阵为邻接阵 $\boldsymbol{A}$, k 表示社区数目, 且 $k\ll n$, 则问题转化为:

$$\min_{\boldsymbol{U}\geqslant 0,\boldsymbol{V}\geqslant 0}\left\|\boldsymbol{A}-\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right\|_F^2 \tag{2}$$

其中, $\|\cdot\|_F$ 表示 Frobenious 范数, $\boldsymbol{U}\geqslant 0$、$\boldsymbol{V}\geqslant 0$表示矩阵中的所有元素均非负.

该模型中, 矩阵 $\boldsymbol{U}$ 提供了节点所属社区信息, $\boldsymbol{V}$ 提供了社区之间的相互作用信息. $\boldsymbol{U}$ 的 n 行可以看作网络中的 n 个节点在低维潜空间中的表示, U_{ij} 表示节点 v_i 相对于社区 j 的隶属度, 即节点 v_i 被划分到社区 j 的倾向性, 因此我们将 v_i 分到社区 l 如果 $U_{il}=\max\{U_{ij}\,|\,j\in\{1,\cdots,k\}\}$. 注意到当 $\boldsymbol{U}$ 或 $\boldsymbol{V}$ 固定时, 式 (2) 对于这两者中的另一个矩阵是凸的, 因此可以采用交替迭代更新方法对目标函数进行优化.

根据文献[17], 在下述迭代更新规则下, 目标函数非增:

$$U_{ij}=U_{ij}\left(\frac{\left[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\boldsymbol{A}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\right]_{ij}}{\left[\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}\right]_{ij}}\right)^{1/4} \tag{3}$$

$$V_{ij}=V_{ij}\frac{\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\right]_{ij}}{\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right]_{ij}} \tag{4}$$

其中, 比例调整因子 1/4 是为了防止迭代更新过于激进. 根据式 (3)、式 (4) 可获得优化问题的解. 以上即为利用非负矩阵分解进行社区检测的基本思想. 下面我们将介绍本文所提出的算法.

2 基于非负矩阵分解的有向网络半监督社区检测方法

本节阐述本文所提出的模型, 首先利用先验信息重建网络拓扑, 然后在目标函数中加入由先验信息所指导的惩罚项, 同时对矩阵 $\boldsymbol{U}$ 进行行归一化来消除节点度异质性的影响, 最后给出求解算法.

2.1 利用先验信息重构网络拓扑

考虑有向网络 $G=(V,E)$, 我们获得如第 1.1 节中所定义的邻接矩阵 $\boldsymbol{A}$ 以及 ML 先验信息集合 C. 根据 C 可以构造 ML 先验指示矩阵 $\boldsymbol{M}\in\{0,\alpha\}^{n\times n}$, 满足:

$$M_{ij}=\begin{cases}\alpha, (i,j)\in C\\ 0, \text{其他}\end{cases} \tag{5}$$

其中, α 为权重参数, 它会影响模型中 ML 先验信息的重要性, α 越大, 先验信息越重要. 如果 α 过大, 模型将过分依赖先验信息而忽略网络自身结构, 而当 α 过小时, 先验信息无法有效指导社区检测过程. 根据经验, 本文采用 $\alpha=2$, 这在真实数据实验中取得了良好效果.

需要注意的是, 真实网络中的边往往非常稀疏, 仅使用观察到的节点之间的连接不足以表示节点之间的整体关联, 因此更好的做法是将先验信息整合到网络结构中, 使结果更清晰、更易于解释. 这里选择将 ML 先验信息整合到邻接阵 $\boldsymbol{A}$ 中, 获得重构邻接矩阵 $\boldsymbol{B}$, 如下所示:

$$B_{ij}=\begin{cases}M_{ij}, (i,j)\in C\\ A_{ij}, \text{其他}\end{cases} \tag{6}$$

由此在式 (2) 基础上, 问题转化为如下形式:

$$\min_{\boldsymbol{U}\geqslant 0,\boldsymbol{V}\geqslant 0}\left\|\boldsymbol{B}-\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right\|_F^2 \tag{7}$$

式 (7) 表示使用重构邻接矩阵 $\boldsymbol{B}$ 进行非负矩阵三分解. 与式 (2) 所使用的邻接阵 $\boldsymbol{A}$ 相比, 式 (7) 中的 $\boldsymbol{B}$ 融合了先验信息, 增强了网络社区内部节点之间的链接密集性, 因此分解获得的 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 所提供的节点隶属度信息和社区间的相互作用效应更加明确.

2.2 融入先验信息的惩罚项

仅重构邻接矩阵对于先验信息的使用是不充分的, 这里进一步利用先验信息对节点的社区隶属度矩阵 $\boldsymbol{U}$ 进行约束.

如上文所述, $\boldsymbol{U}$ 的行向量可以看作节点在低维潜空间中的表示, 其分量可以看作该行所对应的节点相对于各个社区的隶属度. 因此, 如果节点 v_i 和 v_j 属于同一社区, 那么 $\boldsymbol{U}$ 的第 i 行与第 j 行的最大值应该在同一列. 由这一想法, 我们可以在目标函数中添加融入先验信息的惩罚项:

$$\min_{\boldsymbol{U}\geqslant 0,\boldsymbol{V}\geqslant 0}\left\|\boldsymbol{B}-\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right\|_F^2+\lambda\sum_{i,j=1}^{n}M_{ij}\sum_{p=1}^{k}\sum_{q=1,q\neq p}^{k}U_{ip}U_{jq} \tag{8}$$

其中, 正则化参数 λ 控制矩阵分解项与惩罚项之间的平衡.

利用简单的代数运算可以将问题转化为如下形式:

$$\min_{\boldsymbol{U}\geqslant 0,\boldsymbol{V}\geqslant 0}\left\|\boldsymbol{B}-\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right\|_F^2+\lambda Tr\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{U}\boldsymbol{Q}\right) \tag{9}$$

其中, $\boldsymbol{Q}=\boldsymbol{E}_n-\boldsymbol{I}_n$, $\boldsymbol{E}_n$ 为元素全为 1 的 n 阶方阵, $\boldsymbol{I}_n$ 为 n 阶单位阵.

2.3 行归一化消除节点度异质性影响

上述模型对于先验信息的利用已经较为完善, 但忽略了节点度异质性的影响, 而节点度的信息存储在矩阵 $\boldsymbol{U}$ 的行范数中. 因此考虑对 $\boldsymbol{U}$ 的行进行归一化, 于是在式 (9) 的基础上, 问题转化为如下形式:

$$\min_{\boldsymbol{U}\geqslant 0,\boldsymbol{V}\geqslant 0}\left\|\boldsymbol{B}-\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right\|_F^2+\lambda Tr\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{U}\boldsymbol{Q}\right)+\eta\|U\mathbf{1}_k-\mathbf{1}_n\|_F^2 \tag{10}$$

其中, $\mathbf{1}_k$ 和 $\mathbf{1}_n$ 分别为元素全为 1 的 k 维列向量和 n 维列向量. 此即为本文的最终模型.

2.4 目标函数的求解

类似第 1.2 节, 固定 $\boldsymbol{U}$ 或者 $\boldsymbol{V}$ 时模型式 (10) 对这两者中的另一矩阵都是凸的, 因此采用交替迭代更新进行优化以获得局部最小值. 这里令 Γ 和 Θ 分别为约束条件 $\boldsymbol{U}\geqslant 0$ 和 $\boldsymbol{V}\geqslant 0$ 的拉格朗日乘子矩阵, 由此可定义如下拉格朗日函数 L:

$$L=\ell(\boldsymbol{U},\boldsymbol{V})+Tr\left(\Gamma\boldsymbol{U}^{\mathrm{T}}\right)+Tr\left(\Theta\boldsymbol{V}^{\mathrm{T}}\right) \tag{11}$$

其中, ℓ 是式 (10) 中的目标函数. L 相对于 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的偏导数分别为:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{U}}=&2\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+2\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}+2\eta\boldsymbol{U}\mathbf{1}_k\mathbf{1}_k^{\mathrm{T}}\\&+2\lambda\boldsymbol{M}\boldsymbol{U}\boldsymbol{Q}-2\boldsymbol{B}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}-2\boldsymbol{B}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}-2\eta\mathbf{1}_n\mathbf{1}_k^{\mathrm{T}}+\Gamma\end{aligned} \tag{12}$$

$$\frac{\partial L}{\partial \boldsymbol{V}}=2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}-2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{U}+\Theta \tag{13}$$

令式 (12) 与式 (13) 中的偏导数为 0, 并由 KKT 条件, 对任意的 i 和 j 都有 $\Gamma_{ij}U_{ij}=0$ 和 $\Theta_{ij}V_{ij}=0$, 得到以下关系:

$$\begin{aligned}0=&\left[\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}+\eta\boldsymbol{U}\mathbf{1}_k\mathbf{1}_k^{\mathrm{T}}+\lambda MUQ\right]_{ij}U_{ij}\\&-\left[\boldsymbol{B}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\eta\mathbf{1}_n\mathbf{1}_k^{\mathrm{T}}\right]_{ij}U_{ij}\end{aligned} \tag{14}$$

$$0=\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right]_{ij}\boldsymbol{V}_{ij}-\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{U}\right]_{ij}\boldsymbol{V}_{ij} \tag{15}$$

从而有更新规则:

$$U_{ij}=U_{ij}\left(\frac{\left[\boldsymbol{B}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\eta\mathbf{1}_n\mathbf{1}_k^{\mathrm{T}}\right]_{ij}}{\left[\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}+\eta\boldsymbol{U}\mathbf{1}_k\mathbf{1}_k^{\mathrm{T}}+\lambda\boldsymbol{M}\boldsymbol{U}\boldsymbol{Q}\right]_{ij}}\right) \tag{16}$$

$$V_{ij}=V_{ij}\left(\frac{\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{U}\right]_{ij}}{\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right]_{ij}}\right) \tag{17}$$

与式 (3) 类似, 为了防止对于 $\boldsymbol{U}$ 的更新过于激进, 可以对其乘法更新系数添加非线性调整比例因子 1/4. 因此最终的迭代更新规则如下:

$$U_{ij}=U_{ij}\left(\frac{\left[\boldsymbol{B}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\eta\mathbf{1}_n\mathbf{1}_k^{\mathrm{T}}\right]_{ij}}{\left[\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}+\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}+\eta\boldsymbol{U}\mathbf{1}_k\mathbf{1}_k^{\mathrm{T}}+\lambda\boldsymbol{M}\boldsymbol{U}\boldsymbol{Q}\right]_{ij}}\right)^{1/4} \tag{18}$$

$$V_{ij}=V_{ij}\left(\frac{\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{U}\right]_{ij}}{\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right]_{ij}}\right) \tag{19}$$

对于迭代初始矩阵, 这里采用基于奇异值分解的非负双奇异值分解方法对 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 进行初始化[22]. 综上所述, 我们提出了惩罚非负矩阵三分解 (PNMTF) 算法, 如算法 1 所示.

算法 1. PNMTF 算法

Input: n 阶邻接阵 $\boldsymbol{A}$, ML 先验信息集合 C, 社区数目 k, 迭代次数 $iter$

Output: 各节点 v_i 的社区标签 c_i

1) 根据 ML 先验信息集合 C 构建 ML 先验信息指示矩阵 $\boldsymbol{M}$;

2) 基于式 (6) 构建重构矩阵 $\boldsymbol{B}$;
3) 对 $\boldsymbol{U}$、$\boldsymbol{V}$ 进行初始化得到矩阵 $\boldsymbol{U}_0$ 和 $\boldsymbol{V}_0$;
4) **while** $t \leqslant iter$ **do**
基于式 (18) 更新 $\boldsymbol{U}$;
基于式 (19) 更新 $\boldsymbol{V}$;
5) **end**
6) $(v_i, c_i) = \arg\max_{j \leqslant k} U_{ij}$.

算法 1 具有以下特点: (1) 当先验信息较多时, 先验信息在算法中起主要作用, 而先验信息较少时, 主要基于网络本身的结构信息进行社区检测; (2) 我们采用的非负双奇异值分解初始化方法比通常使用的随机初始化方法计算效率更高, 只需要一次运行和较少次数的迭代就能保证算法收敛; (3) 式 (18)、式 (19) 中的迭代更新规则能够保证目标函数非增, 并收敛到使目标函数满足 KKT 条件的稳态点. 我们对该算法的时间复杂度进行了一个简单的估计. 在每次迭代中, 根据对更新规则式 (18) 和式 (19) 的分析, 可以得出时间复杂度为 $O(n^2k + nk^2)$ 的结论. 当邻接阵 $\boldsymbol{A}$ 具有稀疏性时, 时间复杂度可以降低到 $O(mk + nk^2)$, 其中 m 表示边数. 由于 k 通常比 n 小得多, 因此这种情况下可以认为 PNMTF 算法的时间复杂度与网络的规模接近线性.

3 实验分析

实验分为以下几部分: 第 3.1 节阐述 PNMTF 算法与其他方法的对比实验, 并分析消融实验结果; 第 3.2 节讨论 ML 先验信息比例对于社区发现结果的影响; 第 3.3 节进行参数调优, 分析惩罚项、行归一化项与矩阵分解项的平衡关系; 第 3.4 节分析算法的收敛性. 实验中, 我们使用了 5 个真实网络数据集: Cornell、Email、Texas、Washington 和 Wisconsin, 其中 Cornell、Texas、Washington、Wisconsin 来自 4 所大学的 WebKB 子网, 节点表示网页, 有向边表示网页之间的链接, 网页分为 5 类: 学生、教师、员工、课程、项目, Email 来自欧洲大型研究机构 42 个不同部门的 1 005 名员工之间的电子邮件交流网络, 节点表示员工, 有向边表示起始节点向末端节点发送了电子邮件. 表 1 给出了关于这 5 个网络的更详细信息.

3.1 方法对比与消融实验

我们将本文的 PNMTF 算法与以下的半监督方法 GNMF[20]、PSSNMF[21]、FSSNMF[19]和无监督方法 K-means、ANMF[17]进行了对比, 并将消融实验的结果一并呈现在对比图中. 对于无监督的 ANMF 和 K-means 方法, 先类似 FSSNMF 将 ML 先验信息嵌入到邻接阵中, 然后使用 ANMF 和 K-means 进行无监督社区检测. 在消融实验中, 当惩罚项、行归一化项均不包含时, 模型退化为上述 ANMF 算法; 仅包含惩罚项时, 对应 ABLATION 曲线; 仅包含行归一化项时, 由于节点度异质性不会对目标函数产生影响, 故这种情况不予考虑.

表 1 真实网络数据描述

数据集	节点个数	边数	社区数目
Cornell	195	304	5
Wisconsin	265	530	5
Texas	187	328	5
Washington	230	446	5
Email	1 005	25 571	42

本文在实验中采用了两种广泛使用的评价指标: 聚类精度 (clustering accuracy, AC) 与归一化互信息[23] (normalized mutual information, NMI) 来衡量算法性能. NMI 通过计算真实社区与由算法得到的社区之间的相似性来检验社区精度, 取值范围为[0, 1], NMI 值越大, 算法性能越好. AC 定义为社区归属划分正确的节点数与网络中节点总数之比, 取值范围为[0, 1], AC 值越大, 算法性能越好.

由于实验中不知道哪些节点对具有 ML 先验信息, 因此我们从网络中所有可能的 ML 先验信息中随机选择一定比例用于指导社区检测. 例如, 给定一个具有 n 个节点、k 个社区的网络, 社区 i 包含的节点数目为 n_i, 那么网络中所有可能的 ML 先验信息数量为:

$$N = \sum_{i=1}^{k} \frac{n_i(n_i - 1)}{2} \tag{20}$$

然后按一定比例从这 N 对中进行随机选择. 此外, Zhang 等[19]引入了能使先验信息得到更充分利用的逻辑推断, 即: 令 C 为 ML 先验信息集合, 如果 $(i, j) \in C$, $(j, k) \in C$, 那么节点 v_i 和 v_k 也应被分到同一社区, 这一推断是合理的, 我们在实验中进行采用.

模拟中, 指定迭代次数为 100 次, 当损失函数的变化量小于 10^{-5} 时, 迭代终止. 在每个数据集上, 设置先验信息的获取比例为 2%、4%、6%、8%、10%、12%、14%, 并对每种方法重复运行 10 次, 计算每次的 NMI 值与 AC 值, 并分别取均值, 最终结果如图 1 所示.

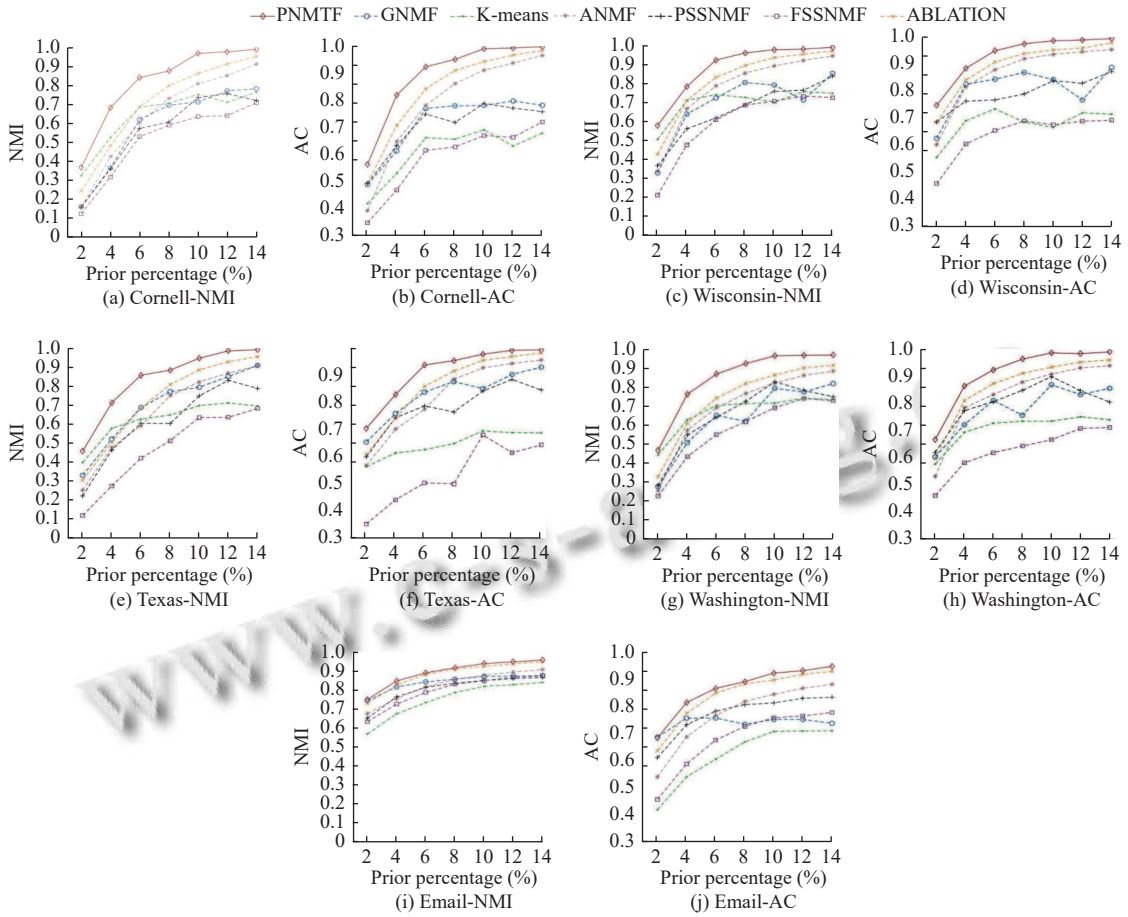


图 1　5 个数据集上不同方法及消融实验对比结果

不难看出, 在这 5 个数据集上所有的 ML 先验信息比例下, PNMTF 算法均优于其他 5 个对比方法. 在先验信息比例较少如 2%、4% 时, 可以看到 PNMTF 在 NMI 值和 AC 值上相对于其他算法均具有优势, 说明 PNMTF 对于先验信息的利用率最高. 而当先验信息的获取比例达到 10% 时, PNMTF 在 Cornell、Washington、Wisconsin 上的 NMI 值和 AC 值均逼近 1, 而在 Email 和 Texas 上的 NMI 值和 AC 值也都超过 0.9, 由此可以说明当先验信息比例达到 10% 的时候, 我们的方法已经具有较为出色的社区检测性能. 同时注意到当先验信息比例较大时, 非负矩阵分解及其拓展方法的效果普遍优于 K-means 算法, 而基于非负矩阵三分解的算法 (PNMTF、ANMF、ABLATION) 均优于非负矩阵二分解及其拓展算法 (GNMF、PSSNMF、FSSNMF), 这是因为在进行有向网络社区检测时, 二分解所获得的两个矩阵分别蕴含从节点生成社区的出边与入边信息, 仅使用其中一个矩阵进行社区检测会忽略掉另一方信息; 而三分解方法中, $\boldsymbol{V}$ 矩阵表示社区之间的相互作用, 已经蕴含连边的不对称性, 因此方向信息已经被分解出来, 再使用矩阵 $\boldsymbol{U}$ 进行社区检测会得到更准确的结果.

值得注意的是, 5 个数据集中 Email 属于大规模网络, 与其他数据相比, Email 数据包含的节点数和边数都显著增加, 因此算法的迭代更新速率较慢. 对比图 1(i)、图 1(j) 与图 1(a) 到图 1(h), 随着先验信息比例增加, PNMTF 在 Email 上获得的 NMI 值与 AC 值的增加速率也显著低于其他 4 个数据集. 这是因为 Email 网络包含的社区数目较多, 社区结构更复杂, 平均社区规模较小, 因此提高先验信息的比例时, 算法的提升效果不如其他 4 个数据集.

此外, 消融实验探究的是模型中惩罚项与行归一化项的必要性, 与 PNMTF 进行对比的是 ANMF 曲线和 ABLATION 曲线. 注意到在 5 个数据集中的所有先验信息比例下, ABLATION 的 NMI 值与 AC 值均优于 ANMF, 因为 ABLATION 在 ANMF 的基础上考虑了对节点的社区隶属度进行约束的惩罚项, 对先验信

息有更好的利用. 而 PNMTF 在社区检测性能上优于 ABLATION, 原因就是本文模型在目标函数中添加了 $\boldsymbol{U}$ 矩阵行归一化项, 解决了节点度异质性的问题, 最终呈现出不错的结果.

3.2 先验信息的作用

本节讨论不同的 ML 先验信息比例的影响. 我们不考虑网络本身的边, 仅基于先验信息来分析网络中的连接组件.

连接组件是指网络中一组可以通过路径互相连接的节点, 也就是说网络中任一连通子图所包含的节点都构成一个连接组件. 我们在这 5 个真实网络中计算了不同先验信息比例下由先验信息所导出的所有社区中连接组件的平均数量和平均规模, 结果如表 2 所示. 连接组件的平均数目定义为网络中连接组件个数除以社区数目, 表示平均每个社区中包含的连接组件数目. 连接组件的平均规模定义为所有连接组件包含的节点总数除以连接组件的个数, 表示平均每个连接组件包含的节点数目. 可以看出当先验信息比例较小时, 连接组件平均数量较大, 平均规模较小. 这意味着网络社区中包含许多由较少节点组成的小型连接组件, 这时 PNMTF 算法主要是决定合并哪些连接组件进行社区检测. Email 数据在 2% 先验信息比例时, 连接组件的平均数量和平均规模都比其他 4 个数据集小, 所以先验信息在 Email 上对结果的提升不如其他 4 个数据集.

表 2 5 个数据集上不同先验信息比例下所有社区中连接组件的平均数目和平均规模

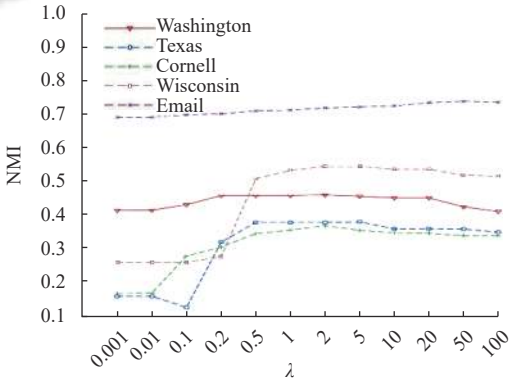
数据集	指标	2%	4%	6%	8%	10%	12%	14%
Cornell	平均数目	4.24	3.00	2.24	1.52	1.60	1.22	1.18
	平均规模	5.65	10.95	15.70	25.86	25.11	32.11	32.66
Wisconsin	平均数目	3.64	2.18	1.90	1.34	1.14	1.20	1.20
	平均规模	11.26	22.14	26.75	38.93	45.59	44.36	44.24
Texas	平均数目	3.18	2.34	1.66	1.30	1.06	1.00	0.88
	平均规模	8.02	14.08	22.75	27.98	35.24	37.50	42.68
Washington	平均数目	3.52	2.22	1.68	1.72	1.42	1.22	1.16
	平均规模	9.43	18.75	25.82	26.54	32.35	37.63	39.18
Email	平均数目	2.59	2.00	1.60	1.35	1.24	1.14	1.06
	平均规模	4.97	8.77	12.35	15.60	17.60	19.49	21.18

随着先验信息比例的增大, 连接组件平均数量减小, 平均规模变大, 这意味着社区中的连接组件存在融合, 也就是多个由少数节点组成的连接组件融合为包含许多节点的大型连接组件. 当先验信息足够多时, 从表 2 中可以看到各个数据集中平均每个社区包含的连接组件数量都接近 1, 而平均每个连接组件包含的节点数目都接近网络的平均社区规模, 这时网络中的社区结构更加清晰, 社区检测更加精确.

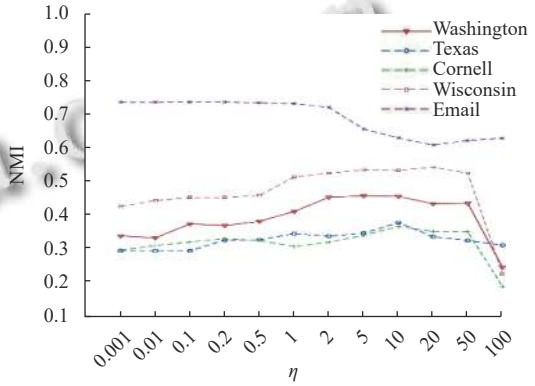
总的来说, 先验信息对社区检测具有较好的指导效果, 先验信息越多社区检测精度越高. 而先验信息在不同网络上起的作用受平均社区规模影响, 对比 Email 数据和其他 4 个数据可以发现, 平均社区规模越大, 先验信息对检测精度提升越明显.

3.3 调优参数分析

本节分析模型中的平衡参数 λ 和 η 对 PNMTF 算法的影响. 实验中, 我们考虑先验信息比例为 2%, 在 5 个实际数据集上计算了 PNMTF 在 λ 和 η 取 0.001–100 中的不同值时的 NMI 值, 结果如图 2 所示.



(a) 不同 λ 取值下 5 个数据集上的 NMI 值



(b) 不同 η 取值下 5 个数据集上的 NMI 值

图 2 5 个数据集上对于 λ 和 η 的调参结果

不难看出, 在这 5 个数据集上, 当固定其中一个参数而增大另一个参数值的时候, NMI 值均呈现出先增后降的趋势. 观察发现, 在 Cornell、Texas、Washington 和 Wisconsin 上, λ 的最佳取值在 0.2–5 之间, 而在 Email 上, λ 的最佳取值在 50 左右. λ 过大会导致惩罚过度, 使 NMI 下降. 结合表 1 中 5 个网络数据集的信息可知, λ 在较小的网络数据集上可以取适当小的值, 在较大的

网络上可以取适当大的值. 与此同时, 当η太大时, 由于社区检测功能主要由矩阵分解与对社区隶属度的惩罚所实现, 因此这会导致对 $\boldsymbol{U}$ 的归一化发挥过于突出的作用, 使 $\boldsymbol{U}$ 失真, 降低算法的社区检测性能.

3.4 收敛性分析

本节分析 PNMTF 算法的收敛性. 我们在这 5 个真实网络数据集上运行 PNMTF, 迭代次数为 100 次, 并记录每次迭代后目标函数的对数, 结果如图 3 所示.

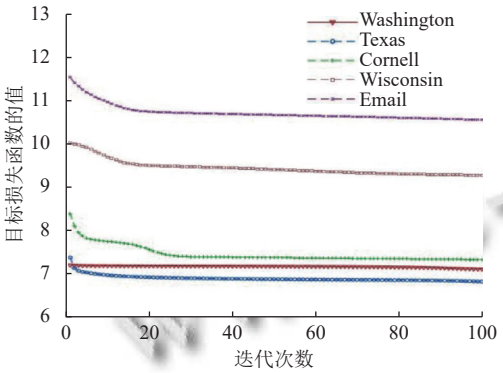


图 3 5 个数据集上目标函数的收敛情况

总体上, PNMTF 的目标函数曲线收敛到平稳值的速度较快. 可以看到在 Washington 数据集上, 目标函数在 5 次迭代后已经几乎没有变化, 而在另外 4 个数据集上的迭代次数达到 30 次以后, 目标函数曲线也都已经趋于平稳. 因此 PNMTF 算法无论是在规模较大还是较小的网络上, 都具有较好的收敛性.

4 结论与展望

本文针对有向网络上的社区检测问题, 提出了一种基于非负矩阵分解的有向网络半监督社区检测算法, 利用先验信息重构网络拓扑, 并在目标函数中添加由先验信息所指导的节点社区隶属度惩罚项, 同时考虑了节点度异质性影响, 实现了对有向网络社区结构的检测. 在多个真实网络数据集上的实验结果均证明了本文方法的有效性. 在接下来的工作中我们将考虑把该方法拓展到多层网络或者动态网络上, 基于多个网络结构提取出有说服力的部分结构信息作为先验信息来指导整体的社区检测. 此外, 在大规模矩阵上, 与所有基于非负矩阵分解的社区检测方法相同, 我们的算法仍然面临一定的挑战. 如果平均社区规模较小, 即平均每个社区包含的节点较少, 补充先验信息后, 会生成大量小型连接组件, 这时算法主要是对这些小的连接组件进行合并, 时间复杂度略有上升. 反之, 当平均社区规模较大时, 补充先验信息后会生成规模较大的连接组件, 使得网络稀疏性变小, 算法速率明显降低. 这时我们可以考虑采用并行或者分布式 NMF 算法[24,25], 这也将成为我们未来的工作方向. 同时, PNMTF 算法需要提前指定社区数目 k, 但在实际中 k 往往是未知的, 因此如何在 NMF 框架下弥补 k 的估计问题是一个值得讨论与研究的方向.

参考文献

1 Zhang Y, Chen KH, Sampson A, *et al*. Node features adjusted stochastic block model. Journal of Computational and Graphical Statistics, 2019, 28(2): 362–373. [doi: 10.1080/10618600.2018.1530117]

2 陈悦, 陈璟. 一种基于遗传算法的 PPI 网络全局比对算法. 小型微型计算机系统, 2022, 43(7): 1494–1498. [doi: 10.20009/j.cnki.21-1106/TP.2020-1089]

3 Newman M. Networks. 2nd ed., Oxford: Oxford University Press, 2018.

4 Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113. [doi: 10.1103/PhysRevE.69.026113]

5 Liu D, Zhao CW, He Y, *et al*. Simultaneous cluster structure learning and estimation of heterogeneous graphs for matrix-variate fMRI data. Biometrics. 2023, 79(3): 2246–2259. [doi: 10.1111/biom.13753]

6 Ren MY, Zhang SG, Wang JH. Consistent estimation of the number of communities via regularized network embedding. Biometrics. 2023, 79(3): 2404–2416. [doi: 10.1111/biom.13815]

7 Fernandes A, Gonçalves PCT, Campos P, *et al*. Centrality and community detection: A co-marketing multilayer network. Journal of Business & Industrial Marketing, 2019, 34(8): 1749–1762.

8 Holland PW, Laskey KB, Leinhardt S. Stochastic blockmodels: First steps. Social Networks, 1983, 5(2): 109–137. [doi: 10.1016/0378-8733(83)90021-7]

9 Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Physical Review E, 2011, 83(1): 016107. [doi: 10.1103/PhysRevE.83.016107]

10 Airoldi EM, Blei DM, Fienberg SE, *et al*. Mixed membership stochastic blockmodels. The Journal of Machine Learning Research, 2008, 9: 1981–2014.

11 Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error

estimates of data values. Environmetrics, 1994, 5(2): 111–126. [doi: 10.1002/env.3170050203]

12 Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788–791. [doi: 10.1038/44565]

13 马慧芳, 赵卫中, 史忠植. 基于非负矩阵分解的双重约束文本聚类算法. 计算机工程, 2011, 37(24): 161–163. [doi: 10.3969/j.issn.1000-3428.2011.24.054]

14 Li YT, Zhu RQ, Qu AN, *et al*. Topic modeling on triage notes with semiorthogonal nonnegative matrix factorization. Journal of the American Statistical Association, 2021, 116(536): 1609–1624. [doi: 10.1080/01621459.2020.1862667]

15 黄波, 严宣辉, 林建辉. 基于联合非负矩阵分解的协同过滤推荐算法. 模式识别与人工智能, 2016, 29(8): 725–734.

16 贾俊杰, 姚叶旺, 陈旺虎. 基于非负矩阵分解的群组推荐算法. 计算机工程与科学, 2022, 44(5): 933–943. [doi: 10.3969/j.issn.1007-130X.2022.05.020]

17 Wang F, Li T, Wang X, *et al*. Community discovery using nonnegative matrix factorization. Data Mining and Knowledge Discovery, 2011, 22(3): 493–521. [doi: 10.1007/s10618-010-0181-y]

18 Zhang ZY. Community structure detection in complex networks with partial background information. Europhysics Letters, 2013, 101(4): 48005. [doi: 10.1209/0295-5075/101/48005]

19 Zhang ZY, Sun KD, Wang SQ. Enhanced community structure detection in complex networks with partial background information. Scientific Reports, 2013, 3: 3241. [doi: 10.1038/srep03241]

20 Yang L, Cao XC, Jin D, *et al*. A unified semi-supervised community detection framework using latent space graph regularization. IEEE Transactions on Cybernetics, 2015, 45(11): 2585–2598. [doi: 10.1109/TCYB.2014.2377154]

21 Liu X, Wang WJ, He DX, *et al*. Semi-supervised community detection based on non-negative matrix factorization with node popularity. Information Sciences, 2017, 381: 304–321. [doi: 10.1016/j.ins.2016.11.028]

22 Boutsidis C, Gallopoulos E. SVD based initialization: A head start for nonnegative matrix factorization. Pattern Recognition, 2008, 41(4): 1350–1362. [doi: 10.1016/j.patcog.2007.09.010]

23 Danon L, Díaz-Guilera A, Duch J, *et al*. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(9): P09008.

24 He CB, Fei X, Li HC, *et al*. Improving NMF-based community discovery using distributed robust nonnegative matrix factorization with SimRank similarity measure. The Journal of Supercomputing, 2018, 74(10): 5601–5624. [doi: 10.1007/s11227-018-2500-9]

25 He CB, Li HC, Fei X, *et al*. A topic community-based method for friend recommendation in large-scale online social networks. Concurrency and Computation: Practice and Experience, 2017, 29(6): e3924. [doi: 10.1002/cpe.3924]

(校对责编: 牛欣悦)