基于非负矩阵分解的稀疏网络社区发现算法

金 红[1],胡智群[2]
(1. 湖北大学计算机与信息工程学院,湖北武汉 430062;2. 北京邮电大学信息与通信工程学院,北京 100088)

摘 要: 社区结构是复杂网络的重要特征之一,社区发现对研究网络结构有重要的应用价值. 基于非负矩阵分解(Non-negative Matrix Factorization,NMF)的社区发现方法是解决社区发现问题的一类基本方法,然而,大多数不能很好地扩展以适用于大型网络,并且在稀疏网络上往往会失败. 由于表达复杂网络拓扑结构特征的邻接矩阵在数据矩阵稀疏时,特征向量的局部化导致基于NMF的方法往往无法工作. 本文提出一种基于NMF的稀疏网络社区发现算法,尝试提高使用非负矩阵分解方法进行社区发现的准确性以及普适性. 本文提出从局部特征向量学习正则化矩阵用来表达原始网络拓扑结构特征,得到的特征矩阵能够很好地发掘数据矩阵隐含的全局结构有更强的特征表达能力. 与邻接矩阵相比,正则化数据矩阵克服了由于稀疏或噪声引起的特征向量(或奇异向量)的局部化问题. 在人工网络和现实网络中的实验结果显示:与经典的基于NMF的社区发现算法相比,该算法能够发现更准确的社区结构,同时,在稀疏网络上也有较好的表现.

关键词: 稀疏网络;社区发现;拓扑结构特征;非负矩阵分解;正则化矩阵
基金项目: 国家自然科学基金(No.61977021);湖北省教育厅科学研究计划项目(No.Q20211010)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)10-2950-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210950

The Non-negative Matrix Factorization Based Algorithm for Community Detection in Sparse Networks

JIN Hong[1], HU Zhi-qun[2]
(1. *School of Computer Science and Information Engineering, HuBei University, Wuhan, Hubei* 430062, *China*;
2. *School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing* 100088, *China*)

Abstract: Community structure is one of the important characteristics of complex networks. Community discovery has important application value in the study of network structure. Community discovery methods based on non-negative matrix factorization (NMF) are a kind of basic methods to solve the problem of community discovery. However, most of them can not be well extended to large networks, and often fail in sparse networks. Because of the localization of the feature vector when the data matrix of the adjacency matrix is sparse, the NMF based method is often unable to work. This paper proposes a NMF based sparse network community discovery algorithm, trying to improve the accuracy and fitness of NMF based community discovery methods. By learning the regularization matrix from local eigenvectors, it is able to express the topological features of original network. The obtained eigenmatrix can well explore the global structure implied in complex network and has stronger feature expression ability. Compared with adjacency matrix, regularized data matrix overcomes the localization problem of eigenvector (or singular vector) caused by sparse or noise. The experimental results in artificial network and real network show that compared with the classical NMF based community discovery algorithm, this algorithm can find more accurate community structure, and has better performance in sparse network.

Key words: sparse networks; community detection; topological structure features; non-negative matrix factorization; regularization matrix

Foundation Item(s): National Natural Science Foundation of China (No.61977021); Scientific Research Project of Education Department of Hubei Province (No.Q20211010)

收稿日期:2021-07-20;修回日期:2023-03-06;责任编辑:王天慧

1 引言

现实世界中大量复杂系统都可以建模为网络结构,例如,社会网络、蛋白质互作用网络、计算机因特网、生物疾病传播网络等都以网络的形式呈现[1]. 其中,网络节点表示各个实体,连边表示实体间的关系. 大量研究表明,复杂系统的网络结构表现出一种社团化的特征[2],即整个网络由多个社区结构组成,这些社区内部连接相对紧密,社区之间连接相对稀疏[3]. 然而,一个社区的物理意义很大程度上取决于网络所属的应用领域[4]. 例如,在社交网络的情况下,社区对应于具有相似的兴趣或偏好的团体[5]. 在新陈代谢网络中,它代表模块化结构,例如功能单元[6].

有效发现网络中的社区结构对分析复杂系统组成结构、了解系统内部互作用机制、揭示系统发展规律以及预测复杂系统行为具有重要的理论意义和实用价值[7]. 近年来,大量社区发现算法被提出,包括基于谱聚类的算法、基于模块化的方法、基于NMF的算法、基于模型的方法等. 最近的社区检测算法综述可以查阅相关的文献[8],在这些方法中我们专注于新近关注度较高的以NMF为基础的算法. 这类算法由于在社区发现中的良好表现,受到了广泛的关注. 文献[9]考虑到复杂网络包含层次和结构信息,如节点级相似性和社区级相似性,从深度自编码器得到启发,结合NMF提出一个类似于自动编码器的深度NMF来学习具有隐藏信息的分层特征,以更好地发现网络中的社区结构. 文献[10]针对目前大多数社区检测方法直接利用原始网络拓扑信息而忽略了固有的社区结构信息,它利用节点属性矩阵和社区结构嵌入矩阵信息,将属性社区检测表示为非负矩阵优化问题,以识别属性图中的所有社区.

然而,当网络规模大且稀疏时,现有的社区检测算法包括以NMF为基础的算法大多难以检测出准确的社区结构,有时甚至不能工作[11]. 针对网络的稀疏性问题,在社区发现研究领域近年来也有相关方法被陆续提出. 文献[12]提出了一种半监督社区检测方法,将深度学习技术与社交网络的拓扑特性相结合. 为了处理维度和稀疏性问题,并在非常大的稀疏矩阵上有效执行卷积,文献[12]对传统CNN卷积层进行修改,提出了一个针对高维稀疏矩阵之间的卷积计算而优化的卷积层,专门考虑非零值来优化稀疏矩阵的计算,从而大大减少了内存使用. 文献[13]提出了一种基于深度学习方法和社交网络拓扑属性的自动社区检测方法,该方法利用一个特定的卷积神经网络用于捕获和表示在线社交网络中典型用户的交互,将稀疏矩阵表示与复杂网络上的深度学习技术相结合. 文献[14]研究了4种不同的网络表示方法,选择最佳表示输入深度稀疏滤波网络中,以获得在每个节点上表示的社区特征映射. 其中,稀疏滤波是一种简单的双层学习模型,它可以处理高维的图数据,将高度稀疏的输入表征为低维特征向量. 以深度学习技术为基础模型的稀疏网络社区检测方法,也起到了较好的克服网络稀疏问题的作用. 只是在模型的参数化学习以及特征工程方面,需要进一步的降低算法的复杂度. 另外,对于潜在社区数目的确定仍然需要寻找自适应的方法. 因此,鉴于稀疏性限制,设计一种有效的社区检测算法仍然是一个具有挑战性的任务[15].

目前,基于NMF的社区检测算法主要从两个方面考虑. 一方面是算法的参数化,在基于NMF的算法中涉及到的大多数参数都具有合理的缺省值. 除了潜在因子的确定,具体到社区检测问题即社区数目的确定. 另一方面,是关于数据矩阵的构造,也称为特征矩阵,即在NMF模型中待分解的矩阵.

针对上述问题,进一步考虑网络的稀疏性,本文提出了基于非负矩阵分解的稀疏网络社区发现算法(the Non-negative Matrix Factorization based algorithm for community detection in Sparse networks, NMFS). NMFS算法克服了由于稀疏或噪声引起的数据矩阵特征向量(或奇异向量)的局部化问题,提高了基于NMF社区发现方法的准确性和适用性. 本文的主要贡献有如下3点:

(1)引入矩阵正则化变换以更好地表达网络的全局拓扑结构特征,缓解了在网络稀疏情况下数据矩阵特征向量的局部化问题.

(2)提出了基于正则化变换的非负矩阵分解稀疏网络社区发现方法,在不增加计算复杂度的情况下结合非回溯矩阵的谱分析来确定社区划分数.

(3)在多个数据集上进行的实验结果表明,本文提出的NMFS算法在稀疏网络上可以得到准确性更高的社区划分结果.

本文的组织结构如下. 首先给出了对称NMF和谱聚类目标函数的等价性证明. 在这个基础上,提出了基于正则化变换的用于稀疏网络社区发现的NMFS算法. 然后,通过在真实世界的稀疏网络和人工生成的稀疏网络上测试所提算法,去进一步分析它在处理稀疏网络上社区发现的性能. 最后,对该方法进行探讨分析,提出未来可扩展的方向.

2 对称非负矩阵分解与谱聚类目标函数的等价性

近年来,当给定数据可以表示为矩阵形式时,谱聚类方法作为一种数据聚类方法表现出较好的应用前景并维持了较高的研究热度[16]. 谱聚类的目标函数有3种类型:比率切割、归一化切割和最大最小化切割[17].

为了证明对称非负矩阵分解目标函数和谱聚类目标函数的等价性，这里重点讨论了比率切割形式的聚类目标函数.

假设$G=(V,E)$表示无向图，其中V表示图G的顶点集合，E代表连接节点的边集合. 假定图G是加权的，它的加权邻接矩阵记为$\boldsymbol{W}=(\boldsymbol{w}_{ij})_{i,j=1,2,\cdots,n}$. 其中，所有元素值都为非负值. 假如$w_{ij}=0$表示在顶点$v_i$和$v_j$之间没有边连接，$w_{ij}>0$表示顶点$v_i$和$v_j$之间的连边具有非负的权值[18]. 为了简便起见，记顶点$v_i\in V$的权值和为$O_i=\sum_{j=1}^{n}w_{ij}$.

注意，矩阵$\boldsymbol{U}$被定义为对角矩阵，其对角元素记为$u_1,u_2,\cdots,u_n$. 对于给定数量的K个子集$C_1,C_2,\cdots,C_k$，比率切割方法可以简化地表示如式(1)和(2)所示[19]：

$$\min \mathrm{RatioCut}(C_1,C_2,\cdots,C_K)=\frac{1}{2}\sum_{i=1}^{K}\frac{W\left(C_i,\overline{C}_i\right)}{\left|C_i\right|} \tag{1}$$

$$\mathrm{cut}\left(C_1,C_2,\cdots,C_k\right)=\frac{1}{2}\sum_{i=1}^{K}W\left(C_i,\overline{C}_i\right) \tag{2}$$

其中，C_i表示顶点集合的子集$C_i\in V$，它的补集记为$\overline{C}_i$. 并且$W(C_i,\overline{C}_i):=\sum_{i\in C_i,j\in\overline{C}_i}w_{ij}$，$\left|C_i\right|$表示子集$C_i$的大小即该子集当中所包含的顶点个数.

对于给定的K个子集$C_1,C_2,\cdots,C_k$，定义K个指示向量记为$(\boldsymbol{h}_{1,j},\boldsymbol{h}_{2,j},\cdots,\boldsymbol{h}_{n,j})^{\mathrm{T}},j=1,2,\cdots,K$，如式(3)所示：

$$h_{i,j}=\begin{cases}1\Big/\sqrt{\left|C_j\right|}, & \text{if } v_i\in C_j\\ 0, & \text{否则}\end{cases}\quad\left(i=1,2,\cdots,n;j=1,2,\cdots,K\right) \tag{3}$$

将这K个指示向量作为列向量形成新的矩阵记为$\boldsymbol{H}\in R^{n\times K}$，显然矩阵$\boldsymbol{H}$中的列向量是彼此正交的[20]，它使下面的式(4)成立：

$$\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{h}_i=\frac{\mathrm{cut}\left(C_i,\overline{C}_i\right)}{\left|C_i\right|} \tag{4}$$

进一步地，结合矩阵的迹，可以得到如下的式(5)：

$$\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{h}_i=\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right)_{ii} \tag{5}$$

其中，矩阵$\boldsymbol{L}$为非归一化拉普拉斯矩阵，记为$\boldsymbol{L}=\boldsymbol{U}-\boldsymbol{W}$，当$K$取任意值时比率切割聚类目标函数可以简化为式(6)：

$$\begin{aligned}\mathrm{RatioCut}\left(C_1,C_2,\cdots,C_K\right)&=\sum_{i=1}^{K}\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{L}\boldsymbol{h}_i\\&=\sum_{i=1}^{K}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right)_{ii}\\&=\mathrm{Tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right)\end{aligned} \tag{6}$$

假设tr表示矩阵的迹，最小化比率切割问题可以表示为式(7)所示：

$$\min_{C_1,C_2,\cdots,C_K}\mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right) \tag{7}$$

满足$\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}$，$\boldsymbol{H}$如式(3)所示.

通过允许矩阵$\boldsymbol{H}$的元素值取任意实值，可以得到该问题的松弛优化情形，如式(8)所示：

$$\min_{H\in R^{n\times K}}\mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right) \tag{8}$$

满足$\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}$.

由此可见，它已经转变为迹最小化问题的标准形式[21]. 由于K-均值和谱聚类的理论框架是统一的，所以用$\boldsymbol{H}$表示的谱聚类的解可以看作对应的K-均值聚类结果[22]. 给定n个数据点记为$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)^{\mathrm{T}}$，K－均值聚类的目标函数可以表达为式(9)：

$$\begin{aligned}\min J_K&=\sum_{k=1}^{K}\sum_{i\in C_k}\left\|x_i-m_k\right\|^2\\&=\sum_i\left\|x_i\right\|^2-\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,j\in C_k}x_i^{\mathrm{T}}x_j\end{aligned} \tag{9}$$

其中，$m_k=\sum_{i\in C_k}x_i/n_k$是聚类$C_k$的$n_k$个点中的聚类中心.

定义K个非负指示向量作为聚类的解决方案，可以表示为式(10)：

$$\begin{aligned}&Y=\left(y_1,y_2,\cdots,y_K\right)\\&\text{and}\quad y_K=\left(0,\cdots,0,\overbrace{1,\cdots,1}^{n_k},0,\cdots,0\right)^{\mathrm{T}}\Big/n_k^{1/2}\end{aligned} \tag{10}$$

显然，矩阵$\boldsymbol{Y}$表示的K-均值聚类的解和矩阵$\boldsymbol{H}$表示的比率切割聚类的解实际上是一致的[23]. 那么，式(9)可以被重写为式(11)所示：

$$\begin{aligned}J_K&=\sum_i\boldsymbol{x}_i^2-\sum_{k=1}^{K}\boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{h}_k\\&=\mathrm{tr}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)-\mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}\right)\end{aligned} \tag{11}$$

由于第一项是常数，式(9)可以被改写为式(12)：

$$\max_{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\boldsymbol{I},\boldsymbol{Y}\geqslant 0}J_W(\boldsymbol{H})=\mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}\right) \tag{12}$$

可以看出，除了在数据表示形式上比率切割是拉普拉斯矩阵$\boldsymbol{L}$，K-均值是矩阵$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$，它们在解的表达上是完全一致的. 综上所述，加权邻接矩阵$\boldsymbol{W}$可以被看作一般的数据表示形式. 在这种情况下，可以证明谱聚类的目标函数与对称NMF目标函数的等价性，如式(13)所示：

$$\begin{aligned}\boldsymbol{H}&=\underset{\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I},\boldsymbol{H}\geqslant 0}{\arg\max}\ \mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H}\right)\\&=\underset{\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I},\boldsymbol{H}\geqslant 0}{\arg\min}\ -2\mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H}\right)\\&=\underset{\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I},\boldsymbol{H}\geqslant 0}{\arg\min}\ \left\|\boldsymbol{W}\right\|^2-2\mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H}\right)+\left\|\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right\|^2\\&=\underset{\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I},\boldsymbol{H}\geqslant 0}{\arg\min}\left\|\boldsymbol{W}-\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right\|^2\end{aligned} \tag{13}$$

通过放宽约束$\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}$可以完成上述证明[24]. 经

由以上的分析,谱聚类直接与K-均值聚类相关. 其中,K-均值聚类与NMF具有等价性. 因此,NMF和谱聚类以一种简单的方式保持了一致性.

3　基于正则变换的NMF稀疏网络社区发现方法

3.1　NMF的社区发现基本原理

在概述基于NMF的社区发现方法原理之前,首先简要介绍经典社区发现方法的定义和数学形式. 社区发现的目的是将图 $G=(V,E)$ 的顶点集划分成 c 个不同的子集,使得该解决方案满足社区结构的基本特征. 假设 $V=\{1,2,\cdots,n\}$ 表示网络中的顶点集合,$n=|V|$ 表示顶点的总数量. 假设有 c 个子集的社区解决方案已经提前给出,然后用分区矩阵 $\boldsymbol{P}$ 表示网络的划分结果[25],如式(14)所示:

$$P_{ik}=\begin{cases}1, & \text{if 节点 } v_i \text{ 属于第 } k \text{ 个子集}\\ 0, & \text{否则}\end{cases}\qquad \text{s.t.}\ \sum_{k=1}^{c}P_{ik}=1\quad (1\leqslant i\leqslant n) \tag{14}$$

那么,第 k 个社区的大小可以表示为 $\sum_{i=1}^{n}P_{ik}$. 另外,对于一个有意义的社区,我们假设每个 k 满足条件 $0<\sum_{i=1}^{n}P_{ik}<n$(没有空的社区和包含所有节点的社区)[26]. 我们称这种划分为硬划分,因为它们构造的分区使每个节点都当且仅属于一个社区.

显然,应该构建一个有意义的分区,使类似的节点出现在相同的社区中[26]. 因此,我们可以定义相似性函数来评估节点之间的相似性如式(15)所示:

$$s\left(P,v_i,v_j\right)=\begin{cases}1, & \text{假如节点 } v_i \text{ 和 节点} v_j \text{ 是完全相似的}\\ 0, & \text{假如节点 } v_i \text{ 和节点 } v_j \text{ 是完全不相似的}\end{cases}$$
$$\text{s.t. } s\left(P,v_i,v_j\right)\in[0,1]$$
$$s\left(P,v_i,v_j\right) \text{ 对所有的 } P_{ij} \text{ 是连续可微的} \tag{15}$$

为了简单起见并从依赖关系 $\boldsymbol{P}$ 进行抽象,我们将上述相似函数简单地表示为 s_{ij}. 在考虑节点相似性方面,可以根据先验知识进行合理假设,在这里记为 $\widetilde{s_{ij}}$. 例如,在某种程度上,节点 v_i 和节点 v_j 之间有边意味它们之间的相似性,而没有边则意味着不相似性[27]. 基于此,我们可以通过计算实际相似度值与指定相似度值的接近程度来评估给定分区,如式(16)所示:

$$E(P)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\tilde{s}_{ij}-s_{ij}\right)^2 \tag{16}$$

为了通过矩阵来表达这一概念,我们引入矩阵 $\boldsymbol{S}(P)=[\boldsymbol{s_{ij}}]$ 和 $\tilde{\boldsymbol{S}}=[\widetilde{\boldsymbol{S_{ij}}}]$. 一般来说,我们可以把给定图的邻接矩阵 $\boldsymbol{A}$ 看作表示先验相似性的适当候选. 因此有 $\tilde{\boldsymbol{S}}=\boldsymbol{A}$. 由于邻接矩阵符合相似性假设,即对于有边连接的节点对,它的相似度为1;对于没有边连接的节点对,它的相似度为0[27]. 接下来,应该实例化指定的相似度函数,并满足上述公式列举的条件,可以表达为

$$s_{ij}=\sum_{k=1}^{c}P_{ik}P_{jk} \tag{17}$$

上述概念要求属于同一个分区的节点相似度等于1,否则等于0. 从矩阵的形式进行表达,可以表达为

$$S(P)=\left[s_{ij}\right]=\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}} \tag{18}$$

下面我们可以通过构造一个NMF问题来获得一个适当的分区矩阵 $\boldsymbol{P}$,使 $E(\boldsymbol{P})$ 最小化,其形式化表达如式(19)所示:

$$\begin{aligned}\min_{P\geqslant 0}E(P)&=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\tilde{s}_{ij}-s_{ij}\right)^2\\&=\left(A-\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}\right)^2\\&=\left\|A-\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}\right\|^2\end{aligned} \tag{19}$$

显然,期望的相似度矩阵以及社区的数目作为输入,在这里由邻接矩阵 $\boldsymbol{A}$ 和通过设置潜在因子 k 指定的社区数量.

3.2　基于正则变换的NMF稀疏网络社区发现算法原理

受以下事实的启发,即通过从局部特征向量学习得来的正则化矩阵作为网络的替代矩阵表示,谱方法在一般谱聚类会执行失败的稀疏网络上工作得很好[28]. 因此,尝试将这种正则化矩阵引入基于NMF的社区检测方法中.

首先,描述了如何生成上述正则化矩阵[29]. 由于局部化表达了网络系统的局部结构,它对这些局部化向量进行去局部化使反映全局结构信息的特征向量具有更大的特征值[30]. 在谱聚类当中,它使用反比参与率(Inverse Participation Ratio,IPR)去度量特征向量 $\boldsymbol{l}$ 的局部化程度[31]. 其中,IPR的计算式是 $\mathrm{IPR}(l)=\sum_{i=1}^{n}l_i^4$. 已有研究表明,较大的IPR($l$)值表明向量 $\boldsymbol{l}$ 反映1的局部化结构的程度更高. 显然,$I(l)$ 的值是从 $\frac{1}{n}$ 到1,它们分别对应向量 $\left\{\frac{1}{\sqrt{\boldsymbol{n}}},\frac{1}{\sqrt{\boldsymbol{n}}},\cdots,\frac{1}{\sqrt{\boldsymbol{n}}}\right\}$ 和 $\{0,\cdots,0,1,0,\cdots,0\}$.

解决方法是构建一个正则化矩阵 $\boldsymbol{L_Z}$ 命名为 Z-Laplacian,与邻接矩阵 $\boldsymbol{A}$ 具有相似的结构. 矩阵 $\boldsymbol{L_Z}$ 记为 $\boldsymbol{L_Z}=\boldsymbol{A}+\boldsymbol{Z}$,其中,$\boldsymbol{A}$ 表示数据矩阵或者它的变体. $\boldsymbol{Z}$ 可以通过下面的正则化学习过程获得[32],具体描述如算法1所示.

算法1 规则化学习

输入: 邻接矩阵A, 其他3个预设参数,即特征向量个数g、学习速率$\eta=O(1)$、阈值Δ.

输出: Z-Laplacian,记为L_Z, 它的主特征向量只揭示了A当中的全局结构

1. 将矩阵$\boldsymbol{Z}$的元素都初始化为0

2. 将L_Z的前g个最大特征值对应的特征向量集合记为

$\boldsymbol{E}=\left\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_g\right\}$

3. 在上述前g个特征向量中选择具有最大IPR值的特征向量记为l, 它可以表达为式子$l=\arg\max Z_{e\in E}\mathrm{IPR}(e)$

4. 假如$\mathrm{IPR}(l)<\Delta$,返回$L_Z=A+Z$. 否则,$\forall i, Z_{ii}\leftarrow Z_{ii}-\eta l_i^2$,然后转到步骤2

可以看出,上述正则化矩阵$\boldsymbol{Z}$是一个对角矩阵. 其中,所有对角元素都是从最局部化的向量中逐渐学习得来. 学习过程的原理是对局部特征向量进行惩罚,使局部特征向量的特征值被抑制[33]. 当所有g个主特征向量都去局部化之后,学习过程会停止. 在这种情形下,假定所获得的Z-Laplacian矩阵只显示矩阵$\boldsymbol{A}$的全局结构,而不显示它的局部结构.

基于上述正则化矩阵的NMF社区检测方法的思想可以描述为算法2[34].

算法2 基于正则化矩阵的NMF社区发现方法

输入: 正则化矩阵$\boldsymbol{L}_{\boldsymbol{Z}}$,社团数$\boldsymbol{k}$

输出: 划分矩阵$\boldsymbol{P}$,它揭示了图G中的社区结构

1. 利用非回溯矩阵估计社区数目的合适值$\boldsymbol{k}$

2. 对于计算对称非负矩阵分解$\boldsymbol{L}_z=\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}$,更新规则是

$$\boldsymbol{P}_{ij}\leftarrow\boldsymbol{P}_{ij}\left|1-\boldsymbol{\beta}+\boldsymbol{\beta}\frac{\left(\boldsymbol{L}_z\boldsymbol{P}\right)_{ij}}{\left(\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\right)_{ij}}\right.$$

其中$0<\boldsymbol{\beta}\leqslant 1$.实际上,我们发现$\boldsymbol{\beta}=\frac{1}{2}$是一个比较好的选择.

3. 在迭代的第一阶段可以使用快速算法,

$$\boldsymbol{P}\leftarrow\max\left(\boldsymbol{L}_z\boldsymbol{P}\left(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\right)^{-1},0\right)$$

4. 目标函数即式(19)收敛后,可得到划分矩阵$\boldsymbol{P}$

对于矩阵$\boldsymbol{P}$的每一个行向量$\boldsymbol{P}_i$,对$\boldsymbol{P}_i$进行规范化,使$\boldsymbol{P}_i$的所有元素的累加之后为1,即$\sum_{j=1}^{k}p_{ij}=1$,规范化后$\boldsymbol{P}$中的元素$\boldsymbol{P}_{ij}$表示节点i隶属于社区j的强度. 另外,在这里将节点i归属为隶属度最大的社区类别. 其中,关于算法的参数化部分,在下面的小节给出.

3.3 算法参数学习

如上所述,在这里选择一个简单快速的方法来估计基于非负矩阵分解的社区发现算法的社区数目k. 已有研究证明了它的有效性,它是基于某些图算子的谱性质,例如非回溯矩阵[35].

对于给定的复杂网络G,$\boldsymbol{A}$表示它的邻接矩阵,则节点k的度可以表示为$d_i=\sum_{j=1}^{n}A_{ij}$. 接下来,给出非回溯矩阵的定义,该矩阵将被用于估计社区的数目.

在这里,用m表示给定无向复杂网络的边数,相应的非回溯矩阵记为$\boldsymbol{B}$. 在建立矩阵$\boldsymbol{B}$时,用两个有向的边来表示节点i和节点j之间的边:一个是从节点i到节点j,另一个是从节点j到节点i. 然后规模为$2m\times 2m$的矩阵$\boldsymbol{B}$可以由式(20)定义:

$$B_{i\to p,q\to l}=\begin{cases}1, & \text{假如}p=q\text{并且}i\neq l\\ 0, & \text{否则}\end{cases}\tag{20}$$

矩阵$\boldsymbol{B}$的谱被证实由±1以及式(21)表示的规模为$2n\times 2n$的矩阵的特征值两个部分组成.

$$\tilde{B}=\begin{pmatrix}0_n & U-I_n\\ -I_n & A\end{pmatrix}\tag{21}$$

其中,0_n表示规模为$n\times n$并且所有元素为零的矩阵. $\boldsymbol{I}$表示$n\times n$的单位矩阵,$\boldsymbol{U}=\mathrm{diag}\left(\boldsymbol{d}_i\right)$表示$n\times n$的对角矩阵,对角元素为$d_i$. 有研究表明,假如网络具有$k$个社区,那么前$k$个最大特征值在矩阵$\tilde{\boldsymbol{B}}$的量级上是实值. 特别地,它们从其他特征值聚集的区域分离开来. 其他特征值聚集的区域包含在一个半径为$\left\|\tilde{\boldsymbol{B}}\right\|^{\frac{1}{2}}$的圆中,这$k$个特征值被认为是矩阵$\tilde{\boldsymbol{B}}$带有信息的特征值. 进一步地,已有研究表明非回溯矩阵的谱范数可以由式(22)来近似表示.

$$\tilde{d}=\left(\sum_{i=1}^{n}d_i\right)^{-1}\left(\sum_{i=1}^{n}d_i^2\right)-1\tag{22}$$

由于非回溯矩阵的信息特征值是实值的,并且与其他特征值坐落的半径为$\left\|\tilde{\boldsymbol{B}}\right\|^{\frac{1}{2}}$的圆形区域相分离. 可以通过对该圆形区域外的特征值数目进行计数来估计k的取值. 注意到,参数学习过程表现良好,尤其是当已知复杂网络的社区具有相似的大小和边缘密度时.

特别是,对于随机块模型生成的网络,圆外的实特征值的数目似乎是网络中存在的簇的数目的自然指标. 那么对于实际网络的划分而言,在某些网络中,圆外分布的大的实特征值可能对应于网络图中的小团.

4 实验结果分析

在这里,将提出的基于NMF的稀疏网络社区检测(NMFS)算法与现有的基于NMF模型的有代表性的社区检测算法,即Zhang等人[35]在稀疏环境下提出的有界非负矩阵三因子分解模型(overlapping community detection via Bounded Nonnegative Matrix Tri-Factorization, BNMTF)进行比较. 对于现实世界中的稀疏网络,选择以政治博客网络为例. 对于人工生成网络,使用随机块模型来生成具有预定义社区结构的稀疏网络. 通过对

这两组数据进行分析,结果表明本文提出的算法所发现的社区结构更接近实际情况. 同时,利用评价指标归一化互信息(Normalized Mutual Information, NMI)对几种典型稀疏网络的性能进行比较,验证了本文方法的优越性.

网络稀疏性的定性定义是网络中实际存在的边数要远小于最大可能的边数. 它的定量定义主要依赖以下两个方面的度量:网络密度以及平均度. 实验数据的相关统计信息如表1所示.

表1　数据集信息

数据集	数据集节点数	边数	社区个数	平均度	密度
Polbooks Network	105	441	3	8.40	0.080 770 0
PolblogsNetwork	1 222	16 714	2	27.37	0.022 403 9
Synthetic Network	500	2 442	5	9.78	0.019 571 5

4.1　现实世界网络与模拟网络测试

(1)政治博客网络

政治博客网络(Polblogs Network)是现实世界中具代表性的稀疏网络之一,它由1 222个节点和16 714条边组成[36]. 节点表示关注美国政治事件的博客,边表示这些博客之间的超链接. 在这里,忽略超链接的方向不考虑链接的方向性. 人工管理者手动标注了每个博客的观点倾向,即自由或者保守. 它们被认为是正确的社区标签,因此上述政治博客网可以划分为两个持不同政治观点的社团.

首先,能够通过相应的非回溯矩阵的谱特性来估计社区的数目. 图1显示了政治博客网对应的非回溯矩阵的谱分布情况,它表明政治博客网可以分为多于2个社区. 在前文中,已经提到对于实际网络它预示了网络中的小团结构. 显然,这里根据实际的标注情况,可以设置NMFS算法的关键参数社区数目$k=2$. 为了更加直观,在图2给出了与实际情况相符的社区划分结果. 然后利用提出的NMFS算法发现社区结构,结果如图3所示. 同时,通过BNMFS算法发现的社区结构在图4中给出.

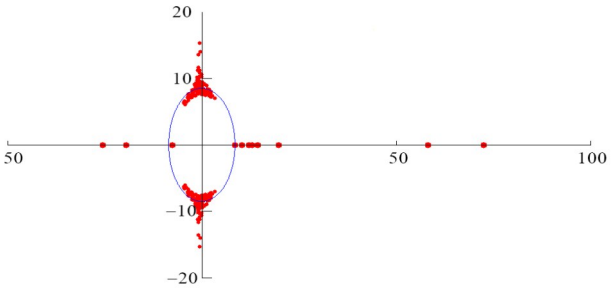

图1　政治博客网络的非回溯矩阵的谱分布情况

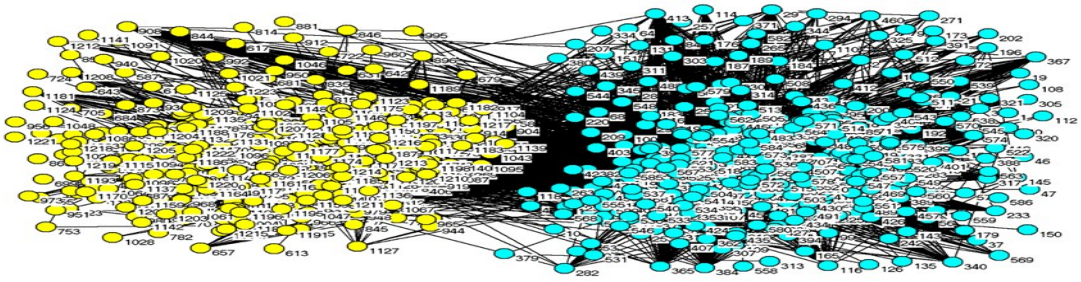

图2　政治博客网络的实际社区结构

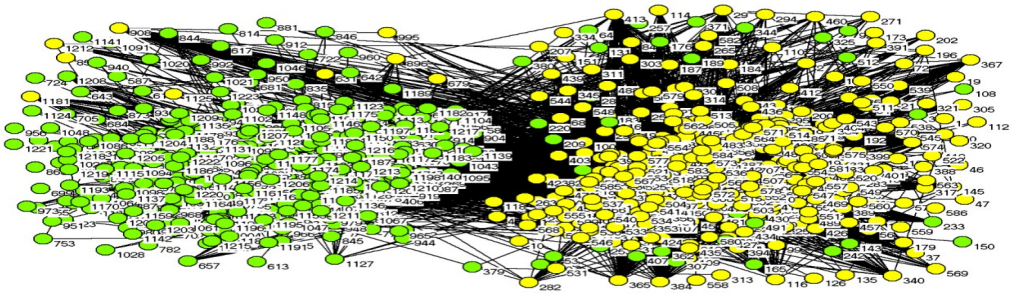

图3　由提出的NMFS算法发现的社区结构

与实际的情况相比较,图3中的分割结果与图2中实际情况大致相同,将政治博客网划分为自由与保守两个派系. 另外,由图可以看出它们的区别也是明显的.

特别地,对于BNMFS算法这里预先设置关键参数即社区数目为$k=2$. 它与实际的情况相一致,相应的社区发现结果如图4所示. 与本文提出的NMFS算法类似,由BNMFS算法发现的两个社区结构大致上与实际

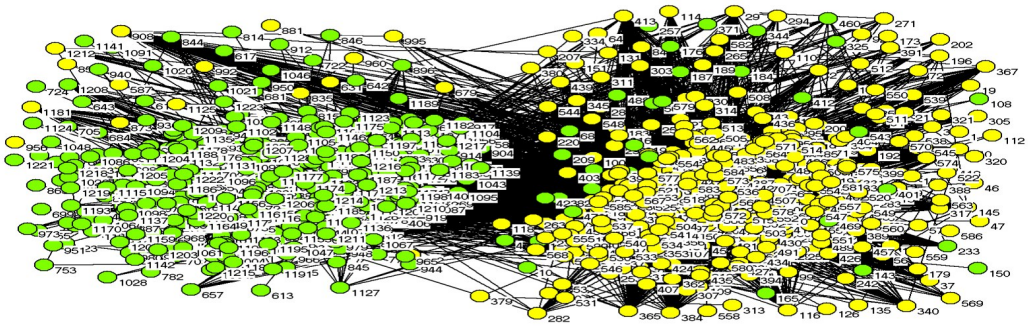

图4 由比较算法BNMTF发现的社区结构

背景是一致的. 但是,相比本文提出的NMFS算法,根据实际的情况而言它将更多的节点划入了不正确的社区. 关于节点误分的统计值将通过性能指标曲线在后文给出相应描述.

(2)合成网络

在这里,使用随机块模型来生成一个合成稀疏网络(Synthetic Network). 标准的随机块模型是用来生成具有内在社区结构网络的经典模型的. 考虑有N个节点的网络,它可以分成q个组. 每个节点i有隐藏的标签$l_i=\{1,2,\cdots,q\}$. 对于节点对(i,j),连接节点i和节点j的边以概率P_{l_i,l_j}独立产生. 因此,有$q\times q$矩阵$\boldsymbol{P}_{\mathrm{ab}}$称为关联矩阵. 当$\boldsymbol{P}_{\mathrm{ab}}=O\left(\frac{1}{N}\right)$时,生成的是稀疏网络. 所以有$c_{\mathrm{ab}}=N\boldsymbol{P}_{\mathrm{ab}}$,可以假设$c_{\mathrm{ab}}=O(1)$,当$N\to\infty$. 当组分配被确定时,可以使用该模型生成一个有500个节点的稀疏网络.

然后将两种算法的社区划分结果与上述合成稀疏网络的预定义社区进行比较. 类似地,首先通过分析模拟生成的稀疏网络相应的非回溯矩阵的谱来估计社团的合适数目. 如图5所示,由图可以看出该模拟生成的稀疏网络可以划分为5个社区. 因此,基于NMF模型的社区发现方法的关键参数即社区个数的值可以设置为$k=5$.

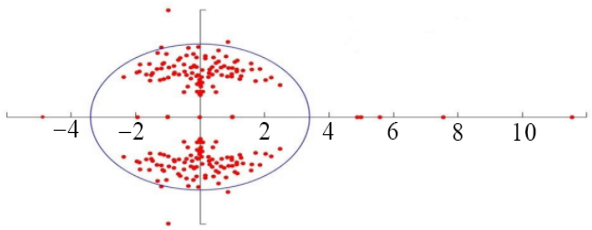

图5 合成稀疏网络的谱分布情况

为了方便对比,图6显示了模拟生成的稀疏网络的内在社区结构. 图7显示了提出的NMFS算法挖掘出的社区结构,而由比较算法发掘的社区结构如图8所示.

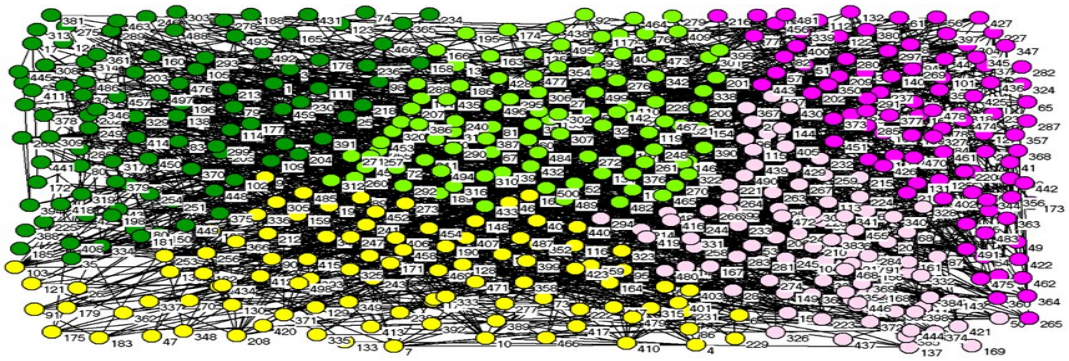

图6 合成稀疏网络的内在社区结构

依据图7可以看出,NMFS算法发掘出的社区结构与在该模拟生成的稀疏网络上预定义的社区结构基本一致. 对于BNMTF算法根据随机块模型的生成机理将社区个数值预设为$k=5$. 从图8中可以看出,与模拟生成的稀疏网络预定义的社区结构相比,由BNMTF算法发现的社区比本文提出的NMFS算法发现的社区具有更大的差异. 这种差异性在数值上的体现,将在后文通过性能指标曲线给出相应的描述.

4.2 准确性评价

通过计算归一化互信息NMI,比较提出的NMFS算法和BNMTF算法的性能. NMI指标被广泛运用于评测聚类算法的性能. 已有研究表明,NMI的值越大表示发现的社区结果越准确,NMI的形式化定义如式(23)所示:

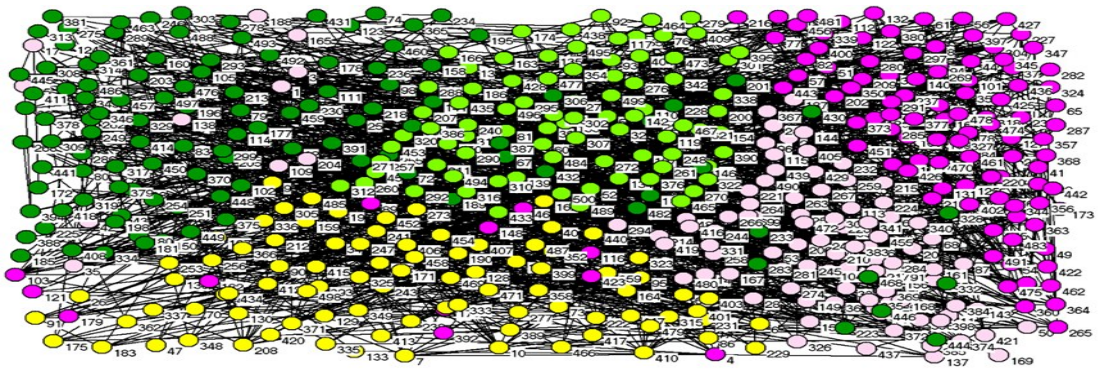

图7　由NMFS算法发现的社区结构

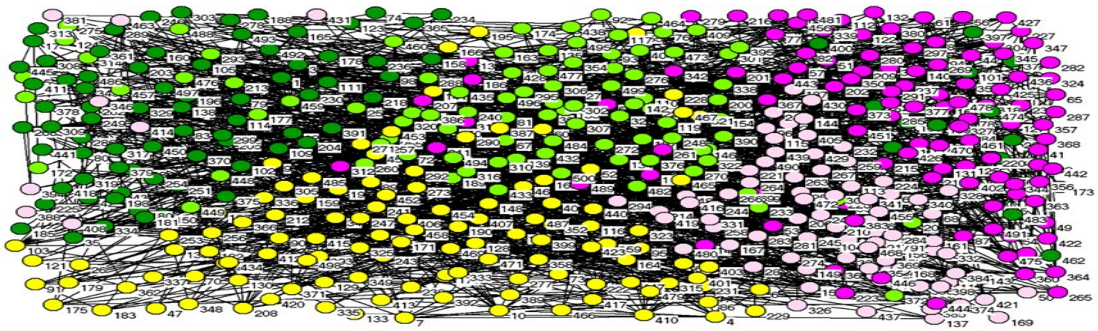

图8　由比较算法BNMTF发现的社区结构

$$\mathrm{NMI}(X,Y)=\frac{-2\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\boldsymbol{C}_{ij}\log\dfrac{\boldsymbol{C}_{ij}\boldsymbol{C}}{\boldsymbol{C}_i\boldsymbol{C}_j}}{\sum_i \boldsymbol{C}_i\log\dfrac{\boldsymbol{C}_i}{\boldsymbol{C}}+\sum_j \boldsymbol{C}_j\log\dfrac{\boldsymbol{C}_j}{\boldsymbol{C}}} \tag{23}$$

其中,X表示实际情况下的社区结构,Y表示由具体算法发掘的社区结构,N_X和N_Y分别表示X和Y的社区数目.特别地,$\boldsymbol{C}$表示混淆矩阵.C_{ij}表示由具体算法划分到社区j的节点数目,但事实上这部分节点在实际情况下应该是归属为社区i的.C_i表示矩阵$\boldsymbol{C}$第i行所有节点元素值的和,C_j表示矩阵$\boldsymbol{C}$第j列上所有元素值的和.

为了测试NMFS算法和BNMTF算法的性能,在前面两个数据集的基础上引入现实世界网络政治书籍网络(Polbooks Network).应用数据集包括两个现实世界的稀疏网络,如政治博客网络和政治书籍网络以及一个模拟生成的稀疏网络.对于上述数据集,NMI值被分别在两种算法上计算出来.如图9所示,可以看出NMFS算法在上述3个代表性稀疏网络上具有更高的NMI值.此外,BNMTF算法在网络上的社区发现结果过于依赖相关参数的预置值.

在社区发现方法的评价方面,NMI表明所发现的社区结构与实际情况的相似程度.NMI值越大,发现的社区结构越接近真实的社区结构.因此,从评价指标归一化互信息的角度来看,NMFS算法提高了在稀疏网络上发现社区结构的准确性.

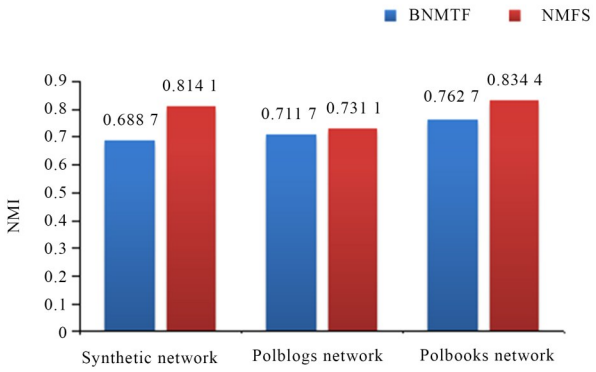

图9　NMFS算法和BNMTF算法在不同测试网络上的NMI值

5　总结

受解决基于矩阵表示的谱方法中局部化问题的一种方法的启发,本文提出了基于NMF的稀疏网络社区发现算法(NMFS).谱方法在给定数据可以表示成矩阵形式时,能够很好地发掘数据隐含的全局结构.然而,当数据矩阵稀疏或有噪声时,由于稀疏或噪声引起的特征向量(或奇异向量)的局部化,经典的谱方法往往无法工作.作为解决谱方法中局部化问题的一种方法,最近提出了一种从局部化特征向量中学习正则化矩阵的通用方法.

本文提出的NMFS算法,主要有两个方面的优势.首先,通过从局部特征向量学习正则化矩阵表达复杂网络拓扑结构,提高了算法在稀疏网络上发掘社区结

构的准确性和效用. 其次,对于基于NMF算法的关键参数,使用一种简单快速的方法来确定潜在因子值. 在未来的工作中,将针对稀疏网络上的社区发现问题做进一步的探索研究,并对基于NMF方法的关键参数设定问题设计自适应的方法,使参数的预设值适应性更强.

参考文献

[1] FORTUNATO S. Community detection in graphs[J]. Physics Reports, 2010, 486(3): 75-174.

[2] NADAKUDITI R R, NEWMAN M E J. Graph spectra and the detectability of community structure in networks[J]. Physical Review Letters, 2012, 108(18): 188701.

[3] ZHAO Y P, LEVINA E, ZHU J. Community extraction for social networks[J]. Proceedings of the National Academy of Sciences, 2011, 108: 7321 -7326.

[4] NEWMAN M E, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.

[5] CLAUSET A, NEWMAN M E J, Moore C. Finding community structure in very large networks[J]. Physical Review E, 2004, 70: 066111.

[6] HASTINGS M. Community detection as an inference problem[J]. Physical Review E, 2006, 74(3): 035102.

[7] LANCICHINETTI A, FORTUNATO S, RADICCHI F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E, 2008, 78(4): 046110.

[8] FORTUNATO S, HRIC D. Community detection in networks: A user guide[J]. Physics Reports, 2016, 659: 1-44.

[9] YE F H, CHEN C, ZHENG Z B. Deep autoencoder-like nonnegative matrix factorization for community detection [C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM, 2018: 1393-1402 .

[10] LI Y, SHA C F, HUANG X, et al. Community detection in attributed graphs: An embedding approach[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2018: 338-345.

[11] AMINI A A, CHEN A, BICKEL P J, et al. Pseudo-likelihood methods for community detection in large sparse networks[J]. Annals of Statistics, 2013, 41(4): 2097-2122.

[12] SANTO A D, GALLI A, MOSCATO V, et al. A deep learning approach for semi-supervised community detection in Online Social Networks[J]. Knowledge-Based Systems, 2021(6): 107345.

[13] SPERLI G. A deep learning based community detection approach[C]//Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. Limassol: ACM, 2019: 1107-1110.

[14] XIE Y, GONG M, WANG S, et al. Community discovery in networks with deep sparse filtering[J]. Pattern Recognition, 2018, 81: 50-59 .

[15] BICKEL P J, CHEN A. A nonparametric view of network models and Newman-Girvan and other modularities [J]. PNAS, 2009, 106(5): 21068-21073.

[16] CHAUDHURI K, CHUNG F, TSIATAS A. Spectral clustering of graphs with general degrees in the extended planted partition model[J]. Journal of Machine Learning Research, 2013, 23(35): 1-23.

[17] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: Analysis and an algorithm[J]. Advances in Neural Information Processing Systems, 2001, 14: 849-856.

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

[19] ZELNIK-MANOR L, PERONA P. Self-Tuning Spectral Clustering[J]. Advances in Neural Information Processing Systems, 2005, 17: 1601-1608.

[20] Gu M, Zha H, Ding C, et al. Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering[R/OL]. (2001)[2021]. https://citeseerx.ist.psu.edu/doc/10.1.1.10.2657.

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

[22] DING C, HE X, SIMON H D. On the equivalence of nonnegative matrix factorization and spectral clustering[C]// Proceedings of the 2005 SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, 2005: 606-610.

[23] SAADE A, KRZAKALA F, ZDEBOROVÁ L. Spectral Clustering of Graphs with the Bethe Hessian[J]. Advances in Neural Information Processing Systems, 2014, 1: 406-414.

[24] DHILLON I S, GUAN Y Q, KULIS B. Kernel k-means, spectral clustering and normalized cuts[C]//Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 551-556.

[25] ZHANG Z Y, WANG Y, AHN Y Y. Overlapping community detection in complex networks using symmetric binary matrix factorization[J]. Physical Review E, 2013, 87(6): 062803.

[26] NEPUSZ T, PETROCZI A, NEGYESSY L, et al. Fuzzy communities and the concept of bridgeness in complex networks[J]. Physical Review E, 2008, 77(1): 016107.
[27] BRUNET J P, TAMAYO P, GOLUB T R, et al. Metagenes and molecular pattern discovery using matrix factorization[J]. Proceedings of the National Academy of Sciences, 2004, 101 (12): 4164-4169.
[28] Saade A, Lelarge M, Krzakala F, et al. Clustering from sparse pairwise measurements[C]//2016 IEEE International Symposium on Information Theory (ISIT). Piscataway: IEEE, 2016: 780-784.
[29] ZHANG P. Robust spectral detection of global structures in the data by learning a regularization[EB/OL]. (2016)[2021]. https://arxiv.org/abs/1609.02906.
[30] KRZAKALA F, MOORE C, MOSSEL E, et al. Spectral redemption in clustering sparse networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(52): 20935-20940.
[31] COJA-OGHLAN A. Graph partitioning via adaptive spectral techniques[J]. Combinatorics, Probability and Computing, 2010, 19(2): 227-284.
[32] QIN T, ROHE K. Regularized spectral clustering under the degree-corrected stochastic blockmodel[EB/OL]. (2013)[2021]. https://arxiv.org/abs/1309.4111.
[33] KUANG D, YUN S, PARK H. SymNMF: Nonnegative low-rank approximation of a similarity matrix for graph clustering[J]. Journal of Global Optimization, 2015, 62 (3): 545-574.
[34] ZHANG Z Y, WANG Y, AHN Y Y. Overlapping community detection in complex networks using symmetric binary matrix factorization[J]. Physical Review E, 2013, 87(6): 062803.
[35] ZHANG Y, YEUNG D Y. Overlapping community detection via bounded nonnegative matrix tri-factorization[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012: 606-614.
[36] ADAMIC L A, GLANCE N. The political blogosphere and the 2004 US election: Divided they blog[C]//Proceedings of the 3rd International Workshop on Link Discovery. New York: ACM, 2005: 36-43.

作者简介

金　红　女. 1984 年 10 月出生,湖北咸宁人. 2006 年、2009 年、2013 年分别于武汉理工大学、武汉大学、武汉大学获工学学士、工程硕士和工学博士学位. 现为湖北大学计算机与信息工程学院讲师. 主要研究方向为图计算、深度学习、异质网络知识发现与推理.
E-mail: anya_1024@163.com

胡智群　女. 1989 年 9 月出生,湖北黄冈人. 2012 年、2018 年分别于湖北工业大学、北京邮电大学获工学学士和工学博士学位. 现为北京邮电大学信息与通信工程学院副研究员、博士生导师. 主要研究方向为环境感知与智能控制、车联网、无线通信.
E-mail: huzhiqun@bupt.edu.cn