一种融合节点变化信息的动态社区发现方法

贺超波[1],成其伟[2],程俊伟[1*],刘星雨[1],余 鹏[1],陈启买[1]
(1. 华南师范大学计算机学院,广东广州 510631;2. 维沃移动通信有限公司,广东东莞 523859)

摘 要: 动态社区发现旨在检测动态复杂网络中蕴含的社区结构,对于揭示网络的功能及演化模式具有重要研究价值. 由于相邻时刻网络的社区结构具有平滑性,前一时刻网络的社区划分信息可以用于监督当前时刻网络的社区划分过程,但已有方法均难以有效提取这些信息来提高动态社区发现性能. 针对该问题,提出一种融合节点变化信息的动态社区发现方法(Semi-supervised Nonnegative Matrix Factorization combining Node Change Information, NCI-SeNMF). NCI-SeNMF首先采用k-core分析方法提取前一时刻社区网络的degeneracy-core,并选取degeneracy-core中的节点构造社区隶属先验信息,然后对相邻时刻网络的节点局部拓扑结构变化程度进行量化,并将其用于进一步修正社区隶属先验信息,最后通过半监督非负矩阵分解模型集成社区隶属先验信息进行动态社区发现. 在多个人工合成动态网络和真实世界动态网络上进行大量对比实验,结果表明,NCI-SeNMF比现有动态社区发现方法在主要评价指标上至少提升了4.8%.

关键词: 动态社区发现;半监督非负矩阵分解;k-core分析;社区网络;复杂网络
基金项目: 国家自然科学基金(No.62077045)
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112(2024)08-2786-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221142

A Dynamic Community Discovery Method via Fusing Node Change Information

HE Chao-bo[1], CHENG Qi-wei[2], CHENG Jun-wei[1*], LIU Xing-yu[1], YU Peng[1], CHEN Qi-mai[1]
(1. *School of Computer Science*, *South China Normal University*, *Guangzhou*, *Guangdong* 510631, *China*;
2. *Vivo Mobile Communication Co. , Ltd. , Dongguan*, *Guangdong* 523859, *China*)

Abstract: Dynamic community discovery aims to detect community structure in dynamic complex networks, and has important research value for revealing the functions and evolution patterns of networks. Because the community structure of the adjacent snapshot networks is smooth, the community discovery result of the previous snapshot network can be used to supervise the community discovery process of the current snapshot network. However, existing methods are difficult to effectively extract these information to improve the performance of dynamic community discovery. In view of this, a dynamic community discovery method named NCI-SeNMF (Semi-supervised Nonnegative Matrix Factorization combining Node Change Information) is proposed, which can fuse node change information. NCI-SeNMF firstly uses k-core analysis method to extract the degeneracy core of every community network at the previous snapshot, and selects the nodes in the degeneracy core to construct the prior community membership information. Then, it quantifies the change degree of the local topology structure of the nodes in the adjacent snapshot networks, and applies it to further improve the prior community membership information. Finally, it integrates the prior community membership information through semi-supervised nonnegative matrix factorization model to discover dynamic communities. Extensive comparative experiments have been conducted on several synthetic and real-world dynamic networks, and the results show that NCI-SeNMF improves at least 4.8% in term of core evaluation metrics comparing with the existing dynamic community discovery methods.

Key words: dynamic community discovery; semi-supervised nonnegative matrix factorization; k-core analysis; community network; complex networks
Foundation Item(s): National Natural Science Foundation of China (No.62077045)

收稿日期:2022-10-11;修回日期:2023-01-02;责任编辑:王天慧
*通讯作者:程俊伟

1 引言

现实世界的复杂网络(例如社交网络、合著关系网络及蛋白质交互网络等)不仅具有隐式的统计特性(例如小世界和无标度[1]),而且还具有可直接观察的显式特性:社区结构. 社区是指复杂网络中的节点集簇,表现为相同社区的节点间链接紧密,不同社区的节点间链接稀疏[2,3]. 对复杂网络进行社区发现不仅有利于理解网络的功能和结构特征,而且还具有重要的应用价值. 例如,社交网络的社区被解释为具有相同兴趣爱好的用户群体,可以用于精准的产品推荐[4]. 论文合著关系网络的社区被理解为具有相同研究兴趣的科研团队,可以为促进更广泛的科研合作推荐团队或研究人员[5]. 蛋白质交互网络的社区被定义为蛋白质功能模块,可以用于辅助疾病诊断与治疗[6].

作为一个涉及社会科学、物理学、计算机科学等多个学科的研究话题,社区发现已吸引了大量关注,各类社区发现方法已被大量提出,其中包括基于模块度优化的方法[7]、基于概率生成模型的方法[8]、基于标签传播的方法[9]、基于博弈论的方法[10]、基于非负矩阵分解的方法[11]以及基于深度学习的方法[12]等. 需要指出的是,大部分现有社区发现方法都只能适用于静态复杂网络. 然而,现实中的复杂网络通常是动态变化的,节点与边会随着时间的推移而生成或消失,潜在的社区结构也会随之不断演化[13]. 适用于动态网络的社区发现(以下简称动态社区发现)方法不仅能发现网络在各个时刻的社区结构,还能进一步分析社区演化模式和驱动因素,最终可进一步拓展社区发现的实际应用范围(如预测网络的演化趋势[14]). 为此,研究有效的动态社区发现方法具有更迫切的现实需求和更高的应用价值.

目前,动态社区发现已成为复杂网络分析领域的研究热点,并提出了一些代表性方法,这些方法通常可以划分为三类:基于两阶段的方法、基于增量学习的方法以及基于演化聚类的方法. 基于两阶段的方法(如CPM(Critical Path Method)[15]和OCDA(Open Compound Domain Adaptation)[16])核心思想是第一个阶段先对每一个时刻的网络采用静态社区发现算法独立地进行社区划分,第二个阶段再在不同的时刻上对社区进行匹配,从而可以捕捉社区动态变化过程. 基于增量学习的方法可以在初始时刻网络应用聚类算法检测社区,后续时刻网络只需要分析增量部分对社区结果进行动态更新. 这类方法具有效率优势,如代表性方法Lime(Low-cost and incremental learning for dynamic heterogeneous information networks)[17]、DISCAN(Distributed and Incremental Structural Clustering Algorithm for Networks)[18]和DyPerm(Permanence for Dynamic community detection)[19]等都具有近似线性的时间复杂度. 基于演化聚类的方法采用双目标优化框架,第一个优化目标为最大化当前时刻的社区发现准确性,第二个优化目标为最小化前后时刻的社区划分差异. 由于能够灵活集成更多有用的约束信息,这类方法往往具有更好的社区发现准确性,同时易采用各种聚类优化模型进行扩展实现,如基于粒子群优化的DYN-MODPSO(Multi-Objective Discrete Particle Swarm Optimization for DYNamic network)[20]、基于非负矩阵分解的DGR-NMF(Dynamic Graph Regularized symmetric Nonnegative Matrix Factorization)[21]和Cr-ENMF(Co-regularized Nonnegative Matrix Factorization)[22]等. 虽然已有的动态社区发现方法都具有不同程度的有效性,但它们未能有效利用隐含的先验信息进一步提升性能. 事实上,社区在动态演化的过程中具有平滑性,相邻时刻的网络存在稳定的节点社区隶属信息. 例如,论文合著关系网络中的科研团队在演化过程中通常具有较为稳定的核心成员. 显然,这些信息可以很自然地作为先验信息用于监督当前时刻网络的社区发现过程. 虽然已有方法尝试集成这些社区隶属先验信息来提高性能,如DGR-NMF(Dynamic Graph Regularized symmetric Nonnegative Matrix Factorization)[21]和Cr-ENMF[22],但如何有效地提取并集成利用仍然需要进一步研究. 具体而言,目前大部分已有方法都无法有效地解决以下两个问题:

(1)由于无论使用何种方法都很难保证社区发现结果是完全正确的,因此从前一时刻网络的社区划分结果提取先验信息时,很容易包含过多错误的社区划分信息.

(2)某些节点随着网络的演化,可能会发生社区转移. 因而即使这些节点在前一时刻网络的社区划分正确,但如果其社区隶属关系在当前时刻发生了转移,与这些节点相关的社区隶属先验信息将是无效的.

显然,错误和无效的社区隶属先验信息将会极大地误导当前时刻网络的社区发现过程并降低性能. 基于此,本文提出了一种融合节点变化信息的半监督非负矩阵分解动态社区发现方法NCI-SeNMF(Semi-supervised Nonnegative Matrix Factorization combining node Change Information). 为有效处理社区隶属先验信息,NCI-SeNMF首先引入k-core分析方法提取前一时刻各个社区网络的degeneracy-core. 接着使用degeneracy-core的节点来构造先验信息,从而过滤潜在的错误社区划分节点. 然后对节点相邻时刻网络的局部结构变化程度进行量化,将其用于修正已发生社区转移节点的先验信息. 最后使用演化聚类框架设计NCI-SeNMF动态社区发现模型,通过图正则非负矩阵分解集成利用社区隶属先验信息以实现动态社区发现性能的提升.

2 社区隶属先验信息处理

本节首先介绍k-core分析[23]相关概念的定义，然后基于k-core对社区隶属节点进行特征分析，最后提出前一时刻网络社区隶属先验信息的提取及修正方法.

2.1 基于k-core的社区隶属节点特征实证分析

定义1 给定一个图$G=(V,E)$，一个非负整数k，$S=(V',E')$为G的子图，若S同时满足以下两个约束，则称S为G的k-core：

约束1 对于任意的$v\in V'$，都有$\text{degree}(v)\geqslant k$，其中$\text{degree}(v)$表示节点$v$的度.

约束2 S是满足约束1的最大子图，即不存在任何S的超图S'满足约束1.

定义2 给定一个图$G=(V,E)$，节点$v\in V$，一个非负整数i，若节点v存在于图G的i-core，但不存在于图G的$(i+1)$-core，则称i为节点v的coreness.

定义3 给定一个图$G=(V,E)$，一个非负整数k，若G的k-core存在，但G的$(k+1)$-core不存在，则称k-core为G的degeneracy-core.

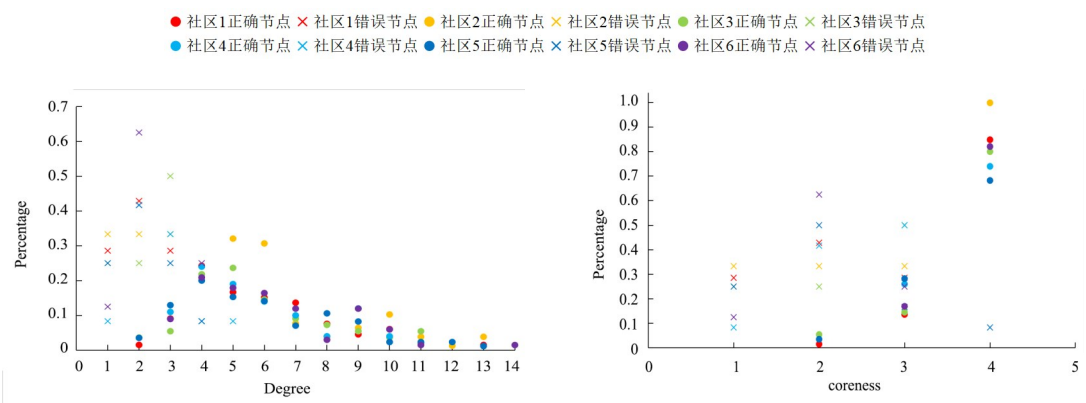
基于k-core相关概念，采用人工合成网络生成工具LFR[24]生成一个网络G，并对其社区隶属节点的特征进行实证分析. G的节点个数为500、社区个数为6、混合参数mu为0.4. 首先使用常用的Louvain算法[25]对G进行社区检测，最终检测的6个社区分别记作$C_1,C_2,\cdots,C_6$. 特别地，将每一个社区C_x包含的节点和其节点之间的边看作一个独立的网络，并表示为社区网络G_{c_x}. 由于LFR生成的网络包含所有节点的真实社区标签，所以根据是否被划分到正确的社区，社区网络中的节点可以被分为两类：正确节点与错误节点. 对每一个社区网络的正确节点集合与错误节点集合分别进行度分布统计，其中社区网络G_{c_x}正确节点集合的度分布定义为

$$P_{\text{correct}}(d,C_x)=\frac{Q_{\text{correct}}(d,C_x)}{N_{\text{correct}}(C_x)} \tag{1}$$

其中，$Q_{\text{correct}}(d,C_x)$表示社区网络$G_{c_x}$正确节点集合中度为$d$的节点个数，$N_{\text{correct}}(C_x)$表示社区网络$G_{c_x}$正确节点集合的节点总数. 错误节点集合的度分布可以按类似的方法计算.

最终G中各个社区网络的度分布计算结果如图1(a)所示. 通过图1(a)可以直观地看出：对于错误节点集合，节点的度大部分都为1~3，极少部分度大于等于4；而对于正确节点集合，度大于等于4的节点则占了该集合的绝大部分. 可以认为：错误节点大概率为小度数节点，而大度数节点几乎不可能为错误节点. 需要说明的是，在其他网络上也发现了该现象，为避免重复，这里忽略了相类似的统计结果. 本文认为产生该现象的原因是：节点在社区网络中的度越大，则该节点与同一社区其他节点存在的链接就越多. 进而大度数节点相互构成的网络具备高稠密性，符合社区内部链接紧密的特征，因此大度数节点的划分结果会趋于正确. 而小度数节点与同一社区其他节点的连接十分稀疏，其组成的网络不能较好地满足社区链接紧密的特征，所以错误节点几乎都是小度数节点.

文献[21]指出节点的coreness与节点的度数呈强正相关，并且节点的度数为节点coreness的上界，因此容易推测错误节点具备小coreness，而正确节点具备大coreness. 为验证该推测，根据式(1)，对每一个社区网络的正确节点与错误节点进行节点coreness的分布统计，此时$Q_{\text{correct}}(d,C_x)$则是表示社区网络$G_{c_x}$的正确节点集合中coreness为$d$的节点个数. 最终得到的结果如图1($b$)所示，从该图中可以看出，两类节点的coreness分布相对于度分布变得更为清晰. 其中所有社区网络正确节点的coreness大部分都为4，同时4是所有社区网络节点的最大coreness值，即正确节点大部分具有所处社区网络的最大coreness值. 相比之下，大部分错误节点的coreness值都是小于等于3.



(a) 度分布 (b) coreness分布

图1 正确和错误节点的度分布与coreness分布

综上分析并结合定义2与定义3,可以发现:degeneracy-core的节点都具有最大的coreness,所以可以认为degeneracy-core中的节点基本都为正确节点.为进一步验证该结论的正确性,对G的6个社区网络从k-core分层的角度进行节点社区划分准确率统计,结果如表1所示.

表1　各个社区网络的k-core节点社区划分准确率　　单位:%

社区网络	1-core	2-core	3-core	4-core
社区网络1	90.41	92.96	97.01	100
社区网络2	92.86	95.12	97.50	100
社区网络3	93.22	94.83	96.30	100
社区网络4	89.29	90.09	94.34	100
社区网络5	87.63	90.43	96.47	98.31
社区网络6	89.33	90.54	97.10	100

通过表1可以看出,随着k-core的k值增大,所包含节点的社区划分准确率随之提升,并且5个社区网络的degeneracy-core(即4-core)的节点划分准确率提升到了100%,另外1个社区的准确率也提升到了98.31%.由此可知,通过k-core分解提取社区网络的degeneracy-core可以有效地过滤错误节点,并尽可能地保留正确节点.进一步地,文献[26]指出,节点的corness越大,该节点离开网络的概率越小,因而还可以认为degeneracy-core的节点在网络动态变化的过程中离开社区的概率很小,即degeneracy-core的节点同时还具有比较稳定的社区隶属关系.

2.2　先验信息提取

Must-link约束常被用作先验信息,用于监督社区发现过程以提高社区划分性能.若两个节点属于must-link约束,则认为这两个节点隶属于同一个社区.根据前一节的分析可知,degeneracy-core的节点基本都具有正确且稳定的社区隶属关系,因此可以提取前一个时刻各个社区网络degeneracy-core的节点社区隶属关系构建must-link约束,并用于监督当前时刻网络的社区发现过程.具体而言,如果在上一个时刻两个节点同时在同一个社区网络的degeneracy-core中,则认为这两个节点在当前时刻也会隶属于同一个社区,从而它们可以构成一个must-link约束.所有的must-link约束可以表示为一个约束矩阵$\boldsymbol{M}_t$,其中分量定义如下:

$$M_{ij,t}=\begin{cases}1, & \text{if } v_i\in \text{degeneracy}\left(G_{c_x,t-1}\right) \text{ and}\\ & \quad v_j\in \text{degeneracy}\left(G_{c_x,t-1}\right)\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

其中,$\text{degeneracy}\left(G_{c_x,t-1}\right)$表示第$(t-1)$时刻社区网络$G_{c_x}$的degeneracy-core.

处于相同社区的节点应具有更为相似的社区隶属关系表示,为此对于任意节点v_i和v_j,它们的must-link约束应能限制它们在t时刻对应的社区隶属关系表示向量$\boldsymbol{H}_{i\cdot,t}$和$\boldsymbol{H}_{j\cdot,t}$之间的距离,即有

$$\begin{aligned}R\left(\boldsymbol{H}_t\right)&=\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|M_{ij,t}\boldsymbol{H}_{i\cdot,t}-\boldsymbol{H}_{j\cdot,t}\right\|_F^2\\&=\sum_{i=1}^{N}\sum_{j=1}^{N}M_{ij,t}\left(\boldsymbol{H}_{i\cdot,t}\boldsymbol{H}_{i\cdot,t}^{\mathrm{T}}+\boldsymbol{H}_{j\cdot,t}\boldsymbol{H}_{j\cdot,t}^{\mathrm{T}}-2\boldsymbol{H}_{i\cdot,t}\boldsymbol{H}_{j\cdot,t}^{\mathrm{T}}\right)\\&=2\sum_{i=1}^{N}\mathcal{D}\left(\boldsymbol{M}_t\right)\boldsymbol{H}_{i\cdot,t}\boldsymbol{H}_{i\cdot,t}^{\mathrm{T}}-2\sum_{i=1}^{N}\sum_{j=1}^{N}M_{ij,t}\boldsymbol{H}_{i\cdot,t}\boldsymbol{H}_{j\cdot,t}^{\mathrm{T}}\\&=2\mathrm{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\mathcal{D}\left(\boldsymbol{M}_t\right)\boldsymbol{H}_t\right)-2\mathrm{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{M}_t\boldsymbol{H}_t\right)\end{aligned} \tag{3}$$

其中,N为节点数量;$\mathrm{tr}(\cdot)$表示矩阵的迹;函数$\mathcal{D}\left(\boldsymbol{M}_t\right)$以矩阵$\boldsymbol{M}_t$作为输入,输出一个对角矩阵:

$$\mathcal{D}\left(\boldsymbol{M}_t\right)_{ij}=\begin{cases}\sum_{j=1}^{N}M_{ij,t}, & \text{if } i=j\\ 0, & \text{if } i\neq j\end{cases} \tag{4}$$

2.3　先验信息修正

尽管$\boldsymbol{M}_t$包含高度可靠的先验信息,但仍可能存在由于节点发生转移而导致相关先验信息失效的问题.即对于前后两个相邻时刻,若某个节点发生了社区转移,即使该节点相关的先验信息在上一时刻是正确的,但却不适用于当前时刻.以图2所示的网络为例,在时刻t中,编号1至5的节点属于1号社区,编号6至10的节点属于2号社区.在时刻$(t+1)$中,2号节点从原先的1号社区转移到了2号社区中.假设提供了符合上一时刻的先验信息,即2号与1、3、4、5号节点同属于1号社区,那么可能会导致在时刻$(t+1)$由于先验信息的监督作用错误地将2号节点划分到社区1中.基于此,有必要减弱发生社区转移的节点对当前时刻社区划分所产生的消极影响.

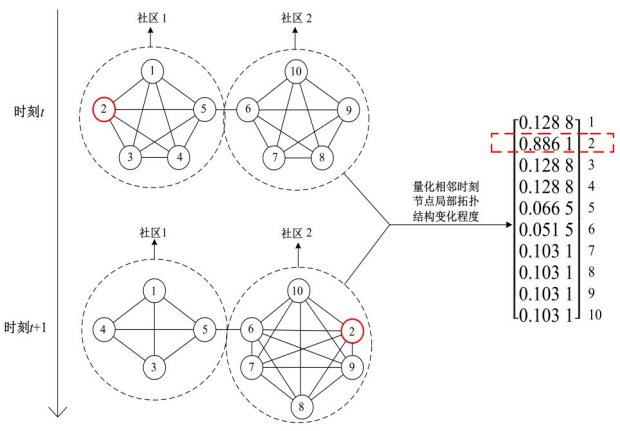


图2　相邻时刻节点发生社区转移的网络

直观地，若一个节点发生了社区转移，则当前时刻以该节点为中心的局部网络结构相比于前一个时刻会产生剧烈的变化，而变化程度可以使用下式进行计算：

$$\text{change}(i,t)=1-\text{similarity}\left(\boldsymbol{s}_{i,t-1},\boldsymbol{s}_{i,t}\right) \tag{5}$$

其中，$\boldsymbol{s}_{i,t}$是一个关于节点i的相似度向量，其元素$s_{ij,t}$表示在时刻t中节点i与节点j的相似度. similarity$(\cdot)$则是度量两个向量相似度的函数.

本文选择RA相似度[27]来构造节点的局部相似性向量：

$$s_{ij,t}=\sum_{v\in\Gamma(i)\cap v\in\Gamma(j)}\frac{1}{\text{degree}(v)} \tag{6}$$

其中，$\Gamma(i)$表示节点i的直接邻居，degree(v)表示节点v的度. 对于similarity函数则选择余弦相似度：

$$\text{similarity}\left(\boldsymbol{s}_{i,t-1},\boldsymbol{s}_{i,t}\right)=\frac{\sum_{k=1}^{N}\left(s_{ik,t-1}\times s_{ik,t}\right)}{\sqrt{\sum_{k=1}^{N}\left(s_{ik,t-1}\right)^2}\times\sqrt{\sum_{k=1}^{N}\left(s_{ik,t}\right)^2}} \tag{7}$$

通过式(5)，可得到相邻时刻节点局部网络结构的变化程度. 若变化程度越大，则表明该节点有很大概率发生了社区转移. 若变化程度较小，则表明该节点发生社区转移的概率很小. 以图2左边的网络为例子，使用式(5)计算各个节点的局部结构变化程度向量，得到的结果如图2右边的向量所示，可以看出节点2的局部结构变化程度为0.886 1，远大于其他节点，说明该节点大概率已经发生了社区转移. 将相邻时刻节点局部网络结构的变化程度用于修正先验信息，若节点局部结构变化程度越强，则说明发生社区转移的可能性就越大，因此应尽量削弱与该节点相关的先验信息，反之若变化程度越弱，则说明该节点发生社区转移的可能性就越小，应尽量保留与该节点相关的先验信息. 为此，结合式(5)将式(3)重写为

$$\begin{aligned}R(\boldsymbol{H}_t)=&\sum_{i=1}^{N}\sum_{j=1}^{N}M_{ij,t}S_{ii,t}S_{jj,t}\left(\boldsymbol{H}_{i\cdot,t}\boldsymbol{H}_{i\cdot,t}^{\mathrm{T}}\right.\\&\left.+\boldsymbol{H}_{j\cdot,t}\boldsymbol{H}_{j\cdot,t}^{\mathrm{T}}-2\boldsymbol{H}_{i\cdot,t}\boldsymbol{H}_{j\cdot,t}^{\mathrm{T}}\right)\\=&2\sum_{i=1}^{N}\mathcal{D}\left(\boldsymbol{S}_t\boldsymbol{M}_t\boldsymbol{S}_t\right)_{ii}\boldsymbol{H}_{i\cdot,t}\boldsymbol{H}_{i\cdot,t}^{\mathrm{T}}\\&-2\sum_{i=1}^{N}\sum_{j=1}^{N}M_{ij,t}S_{ii,t}S_{jj,t}\boldsymbol{H}_{i\cdot,t}\boldsymbol{H}_{j\cdot,t}^{\mathrm{T}}\\=&2\text{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\mathcal{D}\left(\boldsymbol{S}_t\boldsymbol{M}_t\boldsymbol{S}_t\right)\boldsymbol{H}_t\right)-2\text{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{S}_t\boldsymbol{M}_t\boldsymbol{S}_t\boldsymbol{H}_t\right)\\=&2\text{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(\mathcal{D}\left(\boldsymbol{S}_t\boldsymbol{M}_t\boldsymbol{S}_t\right)-\boldsymbol{S}_t\boldsymbol{M}_t\boldsymbol{S}_t\right)\boldsymbol{H}_t\right)\end{aligned} \tag{8}$$

其中，$\boldsymbol{S}_t$是一个对角矩阵，定义如下：

$$S_{ij,t}=\begin{cases}\text{similarity}\left(\boldsymbol{s}_{i,t-1},\boldsymbol{s}_{i,t}\right), & \text{if } i=j\\0, & \text{if } i\neq j\end{cases} \tag{9}$$

3 动态社区发现方法NCI-SeNMF

本节首先介绍NCI-SeNMF基于演化聚类框架的动态社区发现模型，然后推导模型的迭代优化规则，最后设计相应的动态社区发现算法，并进行时间复杂度分析.

3.1 动态社区发现模型

基于演化聚类的动态社区发现框架形式上可表示为

$$\text{Cost}=\alpha\text{CS}+(1-\alpha)\text{CT} \tag{10}$$

其中，CS是当前时刻网络的社区划分损失项，用于衡量社区划分结果拟合当前时刻网络结构的好坏程度，而CT是网络的时变损失项，用于衡量当前时刻的社区结构与前一个时刻社区结构的相似程度，参数α则用于控制CS和CT两个损失项的贡献. 基于演化聚类的动态社区发现框架本质上是一种半监督学习框架，CT可以作为监督项用于集成来自于前一个时刻网络的先验信息.

基于演化聚类框架，考虑到对称非负矩阵分解SNMF[11,28]具有良好的网络数据聚类能力且易扩展，因此本文选择SNMF构建损失项CS，具体步骤：

(1) 给定T个时刻的网络快照序列$G=\{G_1,G_2,\cdots,G_T\}$，令$\boldsymbol{A}_t$表示第t时刻网络G_t的邻接矩阵，对于$\forall A_{ij,t}\in\boldsymbol{A}_t$，如果在第$t$时刻节点$i$与节点$j$之间存在边，$A_{ij,t}=1$，否则$A_{ij,t}=0$.

(2)应用SNMF将$\boldsymbol{A}_t$分解成低秩矩阵因子$\boldsymbol{H}_t$与其转置矩阵的乘积，即$\boldsymbol{A}_t\approx\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}$，分解损失表示为

$$\min L\left(\boldsymbol{H}_t\right)=\|\boldsymbol{A}_t-\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\|_F^2 \tag{11}$$

其中，$\|\cdot\|_F$为Frobenius范数；$\boldsymbol{H}_t$可以作为社区隶属关系表示矩阵，对于$\forall H_{ij,t}\in\boldsymbol{H}_t$，其表示在第$t$时刻节点$i$隶属于社区$j$的强度.

(3)使用$L\left(\boldsymbol{H}_t\right)$作为当前时刻网络的损失项，即CS$=L\left(\boldsymbol{H}_t\right)$.

对于时变损失项CT，由于其可以视为集成先验信息的监督项，因此可以很自然地把$R\left(\boldsymbol{H}_t\right)$视为时变损失项CT. 结合$L\left(\boldsymbol{H}_t\right)$和$R\left(\boldsymbol{H}_t\right)$，最终可得到融合节点变化信息的半监督非负矩阵分解动态社区发现模型：

$$\begin{aligned}\min\mathcal{O}\left(\boldsymbol{H}_t\right)=&\alpha\|\boldsymbol{A}_t-\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\|_F^2+(1-\alpha)\text{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}(\mathcal{D}\left(\boldsymbol{S}_t\boldsymbol{M}_t\boldsymbol{S}_t\right)\right.\\&\left.-\boldsymbol{S}_t\boldsymbol{M}_t\boldsymbol{S}_t\right)\boldsymbol{H}_t\right)\\&\text{s.t.}\quad H_{ij,t}\geqslant 0,\forall i,j\end{aligned} \tag{12}$$

需要注意的是，当$\alpha=1$或$t=1$时，由于缺乏先验约束，可以直接使用SNMF进行社区发现.

3.2 优化求解

通过最小化目标函数$\mathcal{O}(\boldsymbol{H}_t)$可获得$\boldsymbol{H}_t$的近似解,从而获得社区划分结果.由于$\boldsymbol{H}_t$带有非负值约束,最小化$\mathcal{O}(\boldsymbol{H}_t)$可以转化为受限约束求极值问题,这可以应用拉格朗日乘数方法进行优化求解.设$\boldsymbol{\Psi}$为约束$\boldsymbol{H}_t \geqslant 0$的拉格朗日乘子,式(12)的拉格朗日函数可表示为

$$\begin{aligned}\mathrm{L}(\boldsymbol{H}_t)=&\alpha\|\boldsymbol{A}_t-\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\|_F^2+(1-\alpha)\operatorname{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(\mathrm{D}(\boldsymbol{Q}_t)-\boldsymbol{Q}_t\right)\boldsymbol{H}_t\right)\\&+\operatorname{tr}(\boldsymbol{\Psi}\boldsymbol{H}_t)\end{aligned} \tag{13}$$

其中,$\boldsymbol{Q}_t=\boldsymbol{S}_t\boldsymbol{M}_t\boldsymbol{S}_t$.

根据矩阵Trace与Frobenius范数的关系:$\|\boldsymbol{X}\|_F^2=\operatorname{tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$,式(13)可以重写为

$$\begin{aligned}\mathrm{L}(\boldsymbol{H}_t)=&\alpha\left(\operatorname{tr}(\boldsymbol{A}_t\boldsymbol{A}_t^{\mathrm{T}})-2\operatorname{tr}(\boldsymbol{A}_t\boldsymbol{A}_t^{\mathrm{T}})+\operatorname{tr}(\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}})\right)\\&+(1-\alpha)\operatorname{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(\mathrm{D}(\boldsymbol{Q}_t)-\boldsymbol{Q}_t\right)\boldsymbol{H}_t\right)+\operatorname{tr}(\Psi\boldsymbol{H}_t^{\mathrm{T}})\end{aligned} \tag{14}$$

$\mathrm{L}(\boldsymbol{H}_t)$对$\boldsymbol{H}_t$求偏导可得到下式:

$$\begin{aligned}\frac{\partial \mathrm{L}}{\partial \boldsymbol{H}_t}=&-4\alpha\boldsymbol{A}_t\boldsymbol{H}_t-2(1-\alpha)\boldsymbol{Q}_t\boldsymbol{H}_t+4\alpha\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\\&+2(1-\alpha)\mathrm{D}(\boldsymbol{Q}_t)\boldsymbol{H}_t+\boldsymbol{\Psi}\end{aligned} \tag{15}$$

根据KKT条件$\Psi_{ij}H_{ij,t}=0$,可得:

$$\begin{aligned}&\Big[-2\alpha\left(\boldsymbol{A}_t\boldsymbol{H}_t\right)_{ij}-(1-\alpha)\left(\boldsymbol{Q}_t\boldsymbol{H}_t\right)_{ij}+2\alpha\left(\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\right)_{ij}\\&+(1-\alpha)\left(\mathrm{D}(\boldsymbol{Q}_t)\boldsymbol{H}_t\right)_{ij}\Big]H_{ij,t}\\&=0\end{aligned} \tag{16}$$

由上式可得$H_{ij,t}$的乘性迭代更新规则为

$$H_{ij,t}=H_{ij,t}\frac{\left(2\alpha\boldsymbol{A}_t\boldsymbol{H}_t+(1-\alpha)\left(\boldsymbol{Q}_t\boldsymbol{H}_t\right)\right)_{ij}}{\left(2\alpha\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t+(1-\alpha)\left(\mathrm{D}(\boldsymbol{Q}_t)\boldsymbol{H}_t\right)\right)_{ij}} \tag{17}$$

注意,当$T=1$时,由上述过程同理可得$\boldsymbol{H}_1$的乘性迭代更新规则:

$$H_{ij,1}=H_{ij,1}\frac{\left(\boldsymbol{A}_1\boldsymbol{H}_1\right)_{ij}}{\left(\boldsymbol{H}_1\boldsymbol{H}_1^{\mathrm{T}}\boldsymbol{H}_1\right)_{ij}} \tag{18}$$

得到社区隶属强度矩阵$\boldsymbol{H}_t$后,可通过$\underset{l}{\operatorname{argmax}}\, H_{il,t}$获得节点$i$隶属的社区编号.

3.3 社区发现算法及时间复杂度分析

使用NCI-SeNMF方法对动态社区发现的流程可通过算法1进行描述.

算法1 NCI-SeNMF社区发现

输入:T个时刻的网络序列$G=\{G_1,G_2,\cdots,G_T\}$,平衡参数α

输出:T个时刻的社区划分结果序列$C=\{C_1,C_2,\cdots,C_T\}$

```
1. communities = [] // 保存每个时刻社区划分结果的集合
2. 随机初始化 H_1
3. WHILE O(H_1)不收敛 DO
4.     根据式(18)更新 H_1
5. END WHILE
6. 根据 H_1 获得时刻1的社区划分结果 C_1,添加到communities集合中
7. FOR t = 2 to t = T
8.     FOR ∀community in C_{t-1}:
9.         根据community构造社区网络 G
10.        i ← 1 //该变量表示k-core的k值
11.        degenercy – core ← null
12.        WHILE G的节点个数不为0 DO
13.            WHILE G不满足k值为i的k-core DO
14.                删除度数小于i的节点及该节点关联的边
15.                更新其余节点的度数
16.            END WHILE
17.        END WHILE
18.        degenercy – core ← G
19.        i ← i + 1
20.        根据式(2)对矩阵 M 赋值
21.    End FOR
22.    根据式(6)、式(7)和式(9)构造矩阵 S_t
23.    随机初始化 H_t
24.    WHILE O(H_t)不收敛 DO
25.        根据式(17)更新 H_t
26.    END WHILE
27.    根据 H_t 获得时刻t的社区划分结果 C_t,添加到communities集合中
28. END FOR
29. RETURN communities
```

可以分析,算法1的时间消耗主要集中在三个部分.第一部分(第12~17行)是提取社区网络的degeneracy-core,其时间复杂度为$O(n+m)$,其中n和m分别为网络的节点个数和边数.第二部分(第22行)是构造节点相邻时刻的局部结构相似度矩阵S_t,其时间复杂度为$O(n^2)$.第三部分是迭代更新社区隶属矩阵H_t(第24~26行),其时间复杂度为$O(t_{\mathrm{iter}}cn^2)$,其中,c为平均社区个数,t_{iter}为平均迭代次数.那么对于总共有T个时刻的网络,总的时间复杂度为$O\left(T\left(n+m+n^2+t_{\mathrm{iter}}cn^2\right)\right)$.

4 实验研究

为验证NCI-SeNMF的有效性,本文选取人工合成动态网络和真实动态网络进行实验对比分析.实验环境为64位Windows 10操作系统、3.5 GHz Intel Core i9 PC-11900K CPU和64 GB RAM,所有方法都使用Python 3.7编程实现.

4.1 评价指标

为评价社区发现方法的性能，对于具有真实社区标签的网络，采用标准化互信息NMI(Normalized Mutual Information)作为评价指标，对于不具有真实社区标签的网络，采用模块度(Modularity)作为评价指标. NMI和Modularity分别定义如下：

$$\mathrm{NMI}(C,C')=\frac{-2\sum_{i=1}^{q}\sum_{j=1}^{l}F_{ij}\log\left(\frac{F_{ij}N}{\boldsymbol{F}_{i\cdot}\boldsymbol{F}_{\cdot j}}\right)}{\sum_{i=1}^{q}\boldsymbol{F}_{i\cdot}\log\left(\frac{\boldsymbol{F}_{i\cdot}}{N}\right)+\sum_{j=1}^{l}\boldsymbol{F}_{\cdot j}\log\left(\frac{\boldsymbol{F}_{\cdot j}}{N}\right)} \tag{19}$$

其中，C和C'分别表示为检测的社区集合和真实的社区集合，$\boldsymbol{F}$是一个混淆矩阵，其元素F_{ij}表示同时属于C_i和C_j'的节点总个数，$F_{i\cdot}$表示混淆矩阵$\boldsymbol{F}$的第i行元素值的和，$F_{\cdot j}$表示混淆矩阵$\boldsymbol{F}$的第j列元素值的和，q和l分别表示检测的社区个数和真实的社区个数.

$$\mathrm{Modularity}=\frac{1}{2m}\sum_{i}\sum_{j}\left[A_{ij}-\frac{d_id_j}{2m}\right]\delta_{ij} \tag{20}$$

其中，d_i表示节点i的度数；δ_{ij}用于判断节点i与节点j是否属于同一社区，若属于同一社区，$\delta_{ij}=1$，否则$\delta_{ij}=0$.

NMI用于评价社区划分的准确性，而Modularity用于评价社区结构的链接紧密度. 对于NMI和Modularity，取值越高，则表明社区划分性能越好.

4.2 对比方法

为了对比验证NCI-SeNMF的有效性，选择下列5个具有代表性的动态社区发现方法进行对比.

(1) DNMF (Deep Nonnegative Matrix Factorization)[29]：DNMF应用NMF检测当前时刻网络的社区，并通过图正则项约束当前时刻网络的社区划分结果与上一时刻网络的社区划分结果保持相似.

(2) DGR-NMF[21]：DGR-NMF基于演化聚类框架，通过引入网络的几何结构来表示短时间内的网络平滑度，同时应用图正则化SNMF来检测动态社区.

(3) FaceNet (Face recognition and clustering Networks)[30]：FaceNet是一种经典的基于生成模型的动态社区发现方法，它将社区检测和社区演化集成到一个统一的概率模型中，使得社区演化模式能够约束当前时刻网络的社区检测过程.

(4) ESPRA (Evolutionary clustering based on Structural Perturbation and Resource Allocation similarity)[31]：ESPRA是一种结合了结构扰动和拓扑特征的演化聚类方法，它首先通过融合结构扰动和网络拓扑信息来构造相似度，然后通过密度聚类来发现社区结构.

(5) Lime[17]：Lime基于增量学习框架，其首先通过共享特征策略学习节点的增量式表示，然后应用k-means聚类算法获得各个时刻网络的社区.

在实验中，NCI-SeNMF的参数α设置为0.7. 为公平比较，所有基准方法的超参都设置为最优，并且每次实验都重复运行10次并取平均值作为最终结果.

4.3 人工合成动态网络的对比分析

(1) Dynamic GN网络

Dynamic GN是由Lin等人[30]开发的一种基准动态网络生成工具. 使用该工具生成动态网络需要设置5个参数，分别是节点数n，社区数q，网络序列数T，控制社区噪音程度的混合参数z、表示节点社区转移概率的参数nc. z取值越大，表明社区结构越模糊. nc取值越大，表明相邻时刻的网络结构变化越剧烈. 本文主要通过变更z和nc两个参数生成4个人工合成动态网络，具体参数的设置如表2所示.

表2 Dynamic GN网络

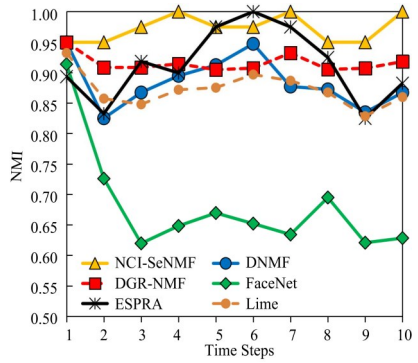
Networks	n	q	T	z	nc
GN1	128	4	10	4	10%
GN2	128	4	10	4	30%
GN3	128	4	10	5	10%
GN4	128	4	10	5	30%

将NCI-SeNMF与基准方法在4个人工合成动态网络进行性能对比，使用的评价指标为NMI，结果如图3所示. 可以看出，在每一个Dynamic GN网络，相对于其他4个基准方法，NCI-SeNMF除了在个别时刻上取得的NMI略低于最佳NMI外，大多数时刻取得的NMI都要高于其余对比方法的NMI，这说明NCI-SeNMF在面对社区结构模糊程度不同和变化剧烈程度不同的动态网络都有着更好的社区检测性能. 此外，在稳定性方面，可以观察到NCI-SeNMF的性能变化曲线是相对平稳的，而其余4个基准方法的性能随着时间的推移都有着不同程度的波动，说明NCI-SeNMF还具有更好的稳定性. 需要指出的是，在每个网络的时刻1中NCI-SeNMF、DNMF和DGR-NMF具有几乎相同的性能. 这是因为它们在时刻1中由于缺乏前一时刻网络的先验信息监督，都退化为基于SNMF的静态社区发现方法，但它们均优于FaceNet、ESPRA及Lime.

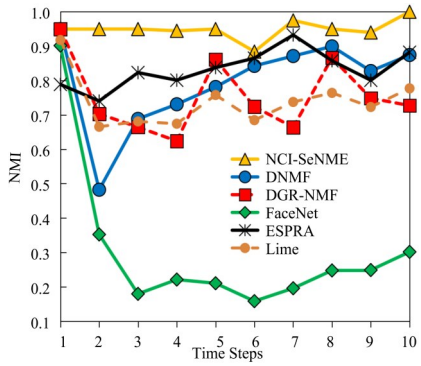
(2) Dynamic LFR网络

Dynamic LFR是由Greene等人[32]开发的一种基于LFR的动态网络生成工具，该工具可以生成不同网络演化事件驱动的无向无权网络序列. 本实验选择的网络演化事件包括以下4类：节点隶属社区转移(Switch)、社区合并与分裂(MergeSplit)、社区出生与死亡(Birth-Death)以及社区扩张与收缩(ExpandContrat)，在实验中为每类网络演化事件都生成2个人工合成动态网络，最终生成共计8个人工合成动态网络，各个网络的参数设置如表3所示.

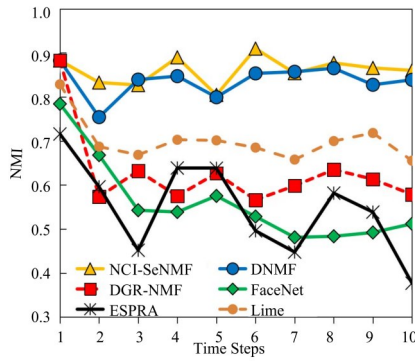
每一类演化事件的网络生成过程说明如下：对于



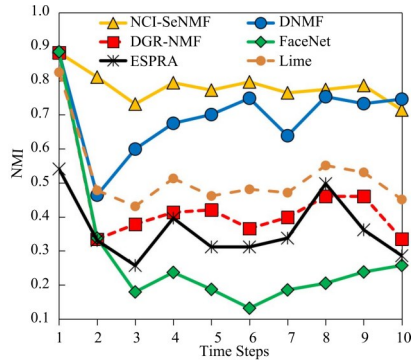
(a) GN1



(b) GN2



(c) GN3



(d) GN4

图3　在Dynamic GN网络上的性能比较

表3　Dynamic LFR网络

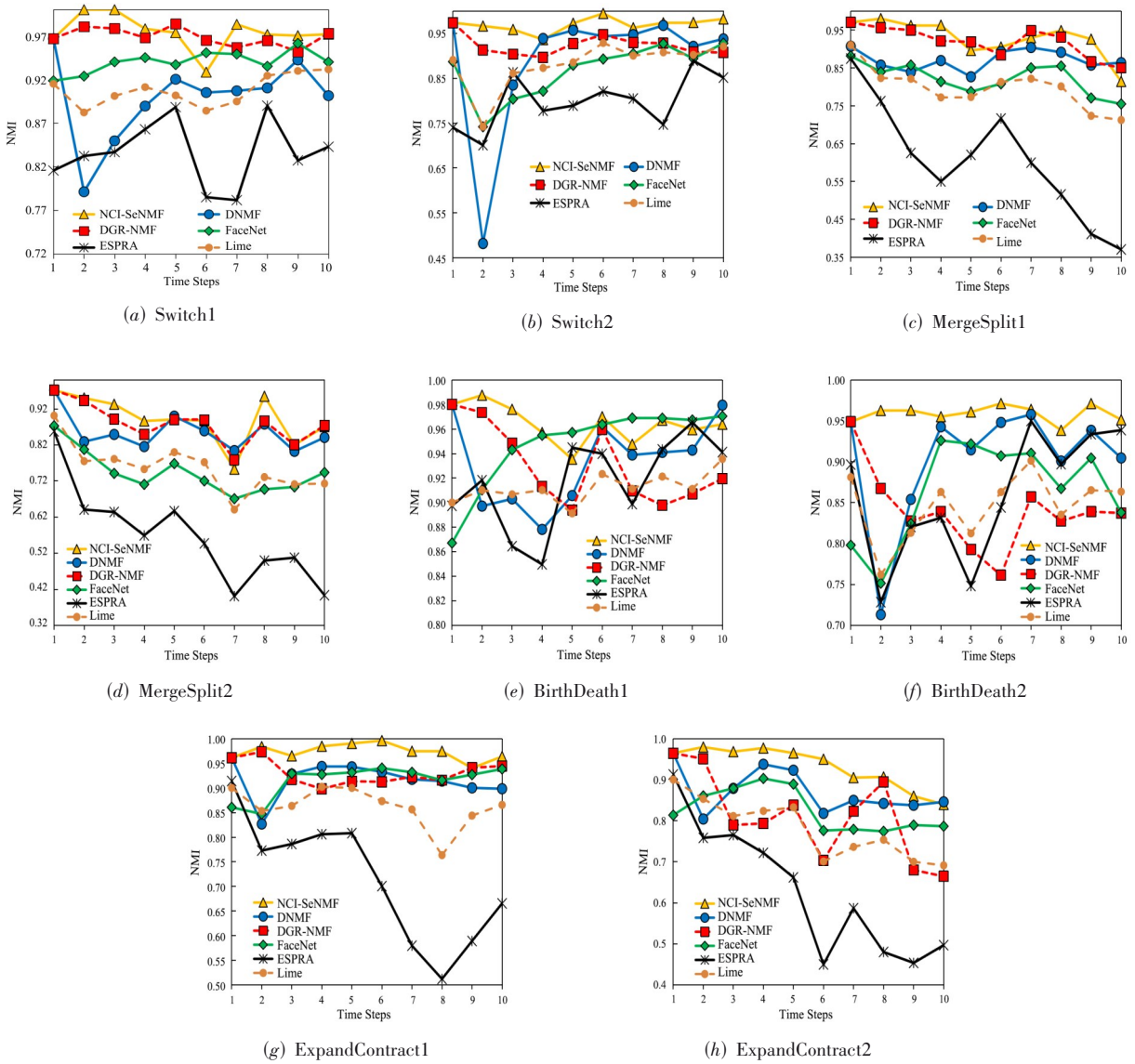
Networks	n	t	mu	p	r
Switch1	500	10	0.4	0.2	—
Switch2	500	10	0.4	0.5	—
MergeSplit1	500	10	0.4	0.1	—
MergeSplit2	500	10	0.4	0.2	—
BirthDeath1	500	10	0.4	0.1	—
BirthDeath2	500	10	0.4	0.2	—
ExpandContrat1	500	10	0.4	0.2	0.2
ExpandContrat2	500	10	0.4	0.2	0.4

节点隶属社区转移事件,每个时刻会随机选择$p\times100\%$的节点更改其隶属社区. 对于出生与死亡事件,每个时刻会从其他社区中移除节点来创建为$p\times100\%$的新社区,并随机删除$p\times100\%$的社区. 对于扩张与收缩事件,每个时刻会随机选择$p\times100\%$的社区以$r\times100\%$的大小进行扩张或收缩操作. 对于合并与分裂事件,每个时刻会随机选取$p\times100\%$的社区进行分裂,随机选择$p\times100\%$的社区进行合并. 从表3可以发现,对于基于相同演化事件生成的两个网络,编号为2的网络比编号为1的网络设置有更大的参数p或r,因此前者比后者发生演化事件的概率更大或者社区变化程度更大,从而导致社区检测难度通常也更高.

图4给出了各方法在8个Dynamic LFR网络的评价结果. 首先通过图4(a)~(d)可以看出,在面对发生节点隶属社区转移事件的网络以及发生社区合并与分裂事件的网络时,NCI-SeNMF和DGR-NMF

有着相似的社区发现性能,且两者的社区发现性能在大多数时刻都高于其他3个对比方法. 进一步观察图4(e)~(h),可以发现在面对发生社区出生与死亡事件的网络以及发生社区扩张与收缩事件的网络时,DGR-NMF的社区发现性能变得相对较低,而NCI-SeNMF却依旧能保持较好的性能. 不仅如此,通过观察图4各方法性能曲线的平稳程度可以发现,NCI-SeNMF不论在具有哪种演化事件的网络,其性能都能够保持相对稳定,而其余方法的性能常常会出现较为剧烈的波动.

为进一步比较各方法在Dynamic LFR网络上的性能稳定性,首先计算各方法在每个网络所有时刻的平均NMI,然后再计算在相同演化事件的两个网络下各方法平均NMI的下降率,最终结果如图5所示. 从图5可以看出,当演化事件发生概率增加(p或r增大)时,DGR-NMF、FaceNet、ESPRA及Lime的社区发现性能下降较快,NCI-SeNMF和DNMF则呈现相对较低的下降率. 例如,在Switch事件网络中,NCI-SeNMF和DNMF的性能下降率分别为0.47%和0.88%,而DGR-NMF、FaceNet、ESPRA及Lime分别为4.65%、7.66%、4.54%及2.89%. 需要指出的是NCI-SeNMF的下降率仅在MergeSplit事件类型网络



(a) Switch1 (b) Switch2 (c) MergeSplit1

(d) MergeSplit2 (e) BirthDeath1 (f) BirthDeath2

(g) ExpandContract1 (h) ExpandContract2

图4 Dynamic LFR网络上的性能比较

中略高于DNMF,说明NCI-SeNMF的性能总体更为稳定.

4.4 真实动态网络的对比分析

人工合成动态网络往往只能模拟出某一类型的社区演化事件,而现实世界的网络往往能同时发生多种类型的社区演化事件.因此为了进一步验证NCI-SeNMF的有效性,选取了如下2个真实世界的动态网络进行实验分析:

(1)Enron邮件网络[13].该网络来自于美国企业Enron的电子邮件集合,选取了2001年151名员工的252 759封电子邮件数据,每个月的数据构成一个时刻的网络快照,共有12个时刻的网络快照序列.网络节点代表员工,边代表两名员工之间有邮件通信.

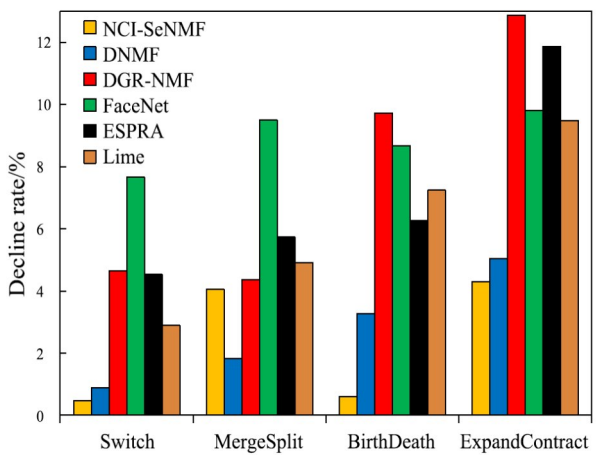


图5 各方法的平均NMI下降率

(2)SFHH会议交互网络[33]. 该网络描述了2009年法国尼斯SFHH会议6月4日至5日间405名与会者的面对面互动情况,它通过设置相同时间间隔将互动等分为25份,然后合并每一份数据,共生成6个时刻的网络快照序列.

由于上述网络都不具备真实的社区标签,因此采用的评价指标是模块度. 各方法在Enron和SFHH动态网络上的性能评价结果分别如表4和表5所示.

表4　Enron网络上的性能比较

T	NCI-SeNMF	DNMF	DGR-NMF	FaceNet	ESPRA	Lime
1	0.584 2	0.584 2	0.584 2	**0.617 1**	0.600 2	0.584 2
2	**0.645 1**	0.629 4	0.635 6	0.529 0	0.624 8	0.600 2
3	**0.635 7**	0.588 7	0.604 1	0.576 2	0.634 7	0.591 7
4	0.629 1	0.604 1	0.618 8	0.581 5	**0.642 2**	0.610 5
5	**0.542 3**	0.522 3	0.526 5	0.516 9	0.471 3	0.500 6
6	**0.667 3**	0.612 0	0.651 2	0.645 9	0.611 0	0.602 3
7	0.640 1	0.619 2	0.621 9	0.591 9	**0.646 9**	**0.611 3**
8	0.516 5	**0.535 9**	0.527 5	0.507 2	0.507 9	0.501 5
9	**0.608 6**	0.597 0	0.601 1	0.588 9	0.585 3	0.576 2
10	**0.562 9**	0.536 5	0.544 7	0.532 2	0.552 9	0.532 9
11	0.572 4	0.561 4	0.563 6	0.565 2	**0.581 1**	0.551 3
12	**0.609 7**	0.573 2	0.579 3	0.583 6	0.561 0	0.573 7

注:粗体表示最好结果,下划线表示次优结果.

通过表4可以看出,不考虑没有先验信息监督的第一个时刻,NCI-SeNMF在剩余11个时刻中有7个时刻都能取得最高的模块度且有3个时刻取得次优的模块度,而DNMF、DGR-NMF、FaceNet、ESPRA和Lime分别有1个、0个、1个、3个和1个时刻取得最高模块度. 这说明在Enron网络中,NCI-SeNMF相对于另外4个对比方法有着更好的性能. 同样地,从表5中可以看出,不考虑没有先验信息监督的第一个时刻,NCI-SeNMF在剩余的5个时刻都能取到最高的模块度,并且相对于同一时刻的次优模块度有着较大的提升,分别提升了12.4%、4.8%、9.8%、15.4%及9.2%.

综上分析可知,通过在人工合成动态网络和真实动态网络进行对比实验,结果表明NCI-SeNMF可以有效处理具有各种复杂演化事件的网络,有着更好且稳定的社区发现性能.

表5　SFHH网络上的性能比较

T	NCI-SeN-MF	DNMF	DGR-NMF	FaceNet	ESPRA	Lime
1	0.3066	0.3066	0.3066	0.2786	**0.3090**	0.3012
2	**0.3868**	0.2704	0.3441	0.3204	0.3031	0.2928
3	**0.2894**	0.2690	0.2761	0.2099	0.1863	0.2219
4	**0.2527**	0.2302	0.2052	0.2179	0.1572	0.1973
5	**0.3605**	0.3094	0.2759	0.3124	0.2815	0.2976
6	**0.4261**	0.3846	0.3903	0.3842	0.3837	0.3743

注:粗体表示最好结果,下划线表示次优结果.

4.5　时间效率分析

为测试NCI-SeNMF的运行效率,选择GN1、Switch1、Enron及SFHH网络计算其运行时间,并与其他方法进行对比分析,结果如表6所示. 从表6中可以看出,基于增量学习的Lime方法具有较为明显的运行效率优势,NCI-SeNMF与其他同样基于NMF的方法DNMF和DGR-NMF相比具有相近时间成本,甚至可以在个别网络(如Enron)获得次优的运行效率.

表6　时间效率比较　　单位:s

数据集	NCI-SeNMF	DNMF	DGR-NMF	FaceNet	ESPRA	Lime
GN1	22	20	19	20	21	**17**
Switch1	77	78	81	67	70	**46**
Enron	43	45	52	49	53	**25**
SFHH	63	65	71	61	65	**31**

注:粗体表示最好结果,下划线表示次优结果.

4.6　α敏感度分析

NCI-SeNMF社区发现模型的超参α用于平衡网络快照损失项和时变损失项的贡献. 明显地,α设置为0或者1时对于动态社区发现都不具有实际意义. 为分析其合理的取值,选择人工合成动态网络GN1和真实动态网络SFHH作为示例,通过从0.1~0.9改变α来检测社区,NMI和Modularity分别作为评价指标,最终结果分别如图6和图7所示. 从图6可以看出当$\alpha \geq 0.7$时,所有时刻取得的NMI都几乎大于0.95. 特别地,当$\alpha=0.7$时,NMI在10个时刻中有9个时刻都能够取到最大值,只在时刻9时略低于$\alpha=0.9$取得的NMI值. 在图7中,也有相似的发现,即当$\alpha=0.7$时,NCI-SeNMF在绝大部分时刻可以取得更好的性能. 根据以上观察,在本文的所有实验中,α都统一设置为0.7.

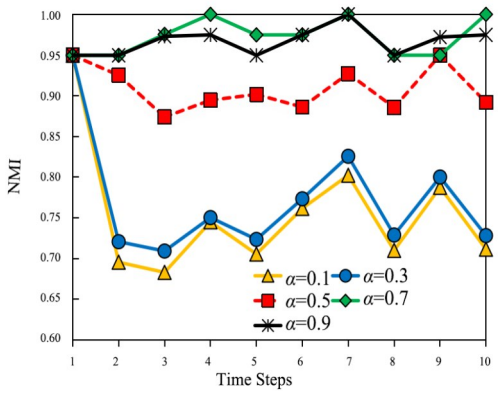


图6　GN1网络不同α值的性能变化

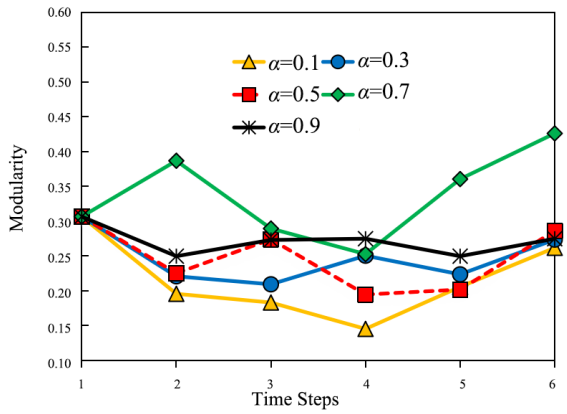


图7 SFHH网络不同α值的性能变化

5 结论

动态复杂网络中的前一个时刻网络的社区隶属信息可以用于监督当前时刻网络的社区发现过程,但已有的动态社区发现方法无法有效提取这些信息. 为此本文提出了一种融合节点变化信息的动态社区发现方法NCI-SeNMF. NCI-SeNMF引入degeneracy-core节点和社区转移节点的识别策略,以提取高度可靠的社区隶属关系先验信息,并用于监督当前时刻网络的社区发现过程. 在人工合成动态网络和真实动态网络进行了大量实验分析,结果表明NCI-SeNMF具有更稳定且更优的社区发现性能. 虽然NCI-SeNMF在小型网络可以获得满意的运行效率,但正如第3.3节的分析,NCI-SeNMF的时间复杂度近似为$O(n^2)$,这导致它无法有效应用于大规模动态网络. 为此在后续研究中,将重点对节点局部网络结构相似度计算算法和社区发现模型迭代更新算法进行效率优化,目的在于进一步提高NCI-SeNMF的整体运行效率.

参考文献

[1] WANG X F, CHEN G R. Complex networks: Small-world, scale-free and beyond[J]. IEEE Circuits and Systems Magazine, 2003, 3(1): 6-20.

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

[3] MANSOUREH N, MOHAMMAD H F Z, SUSAN B. Fuzzy community detection on the basis of similarities in structural/attribute in large-scale social networks[J]. Artificial Intelligence Review, 2022, 55: 1373-1407.

[4] LIU S Y, WANG S H. Trajectory community discovery and recommendation by multi-source diffusion modeling[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(4): 898-911.

[5] HU Z W, LIN A, WILLETT P. Identification of research communities in cited and uncited publications using a co-authorship network[J]. Scientometrics, 2019(118): 1-19.

[6] DILMAGHANI S, BRUST M R, RIBEIRO C H C, et al. From communities to protein complexes: A local community detection algorithm on PPI networks[J]. PLoS One, 2022, 17(1): e0260484.

[7] 潘剑飞, 董一鸿, 陈华辉, 等. 基于结构紧密性的重叠社区发现算法[J]. 电子学报, 2019, 47(1): 145-152.
PAN J F, DONG Y H, CHEN H H, et al. The overlapping community discovery algorithm based on compact structure[J]. Acta Electronica Sinica, 2019, 47(1): 145-152. (in Chinese)

[8] HE D X, LIU H X, FENG Z Y, et al. A joint community detection model: Integrating directed and undirected probabilistic graphical models via factor graph with attention mechanism[J]. IEEE Transactions on Big Data, 2022, 8(4): 994-1006.

[9] GARZA S E, SCHAEFFER S E. Community detection with the label propagation algorithm: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2019, 534: 122058.

[10] MOSCATO V, PICARIELLO A, SPERLI G. Community detection based on game theory[J]. Engineering Applications of Artificial Intelligence, 2019, 85: 773-782.

[11] HE C B, FEI X, CHENG Q W, et al. A survey of community detection in complex networks using nonnegative matrix factorization[J]. IEEE Transactions on Computational Social Systems, 2022, 9(2): 440-457.

[12] SU X, XUE S, LIU F Z, et al. A comprehensive survey on community detection with deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 4682-4702.

[13] ROSSETTI G, CAZABET R. Community discovery in dynamic networks: A survey[J]. ACM Computing Survey, 2019, 51(2): 35.

[14] 李赫, 印莹, 李源, 等. 基于多目标演化聚类的大规模动态网络社区检测[J]. 计算机研究与发展, 2019, 56(2): 281-292.
LI H, YIN Y, LI Y, et al. Large-scale dynamic network community detection by multi-objective evolutionary clustering[J]. Journal of Computer Research and Development, 2019, 56(2): 281-292. (in Chinese)

[15] DHOUIOUI Z, AKAICHI J. Tracking dynamic community evolution in social networks[C]//2014 IEEE/ACM In-

ternational Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). Piscataway: IEEE, 2014: 764-770.

[16] İLHAN N, ÖĞÜDÜCÜ Ş G. Predicting community evolution based on time series modeling[C]//Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. New York: ACM, 2015: 1509-1516.

[17] PENG H, YANG R Y, WANG Z, et al. Lime: Low-cost and incremental learning for dynamic heterogeneous information networks[J]. IEEE Transactions on Computers, 2022, 71(3): 628-642.

[18] WISSEM I, SABEUR A, HAITHEM M, et al. A distributed and incremental algorithm for large-scale graph clustering[J]. Future Generation Computer Systems, 2022, 34: 334-347.

[19] AGARWAL P, VERMA R, AGARWAL A, et al. DyPerm: Maximizing permanence for dynamic community detection [C]//PHUNG D, TSENG V, WEBB G, et al. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2018: 437-449.

[20] YIN Y, ZHAO Y H, LI H, et al. Multi-objective evolutionary clustering for large-scale dynamic community detection[J]. Information Sciences, 2021, 549: 269-287.

[21] WANG S H, LI G P, HU G Y, et al. Community detection in dynamic networks using constraint non-negative matrix factorization[J]. Intelligent Data Analysis, 2020, 24(1): 119-139.

[22] MA X K, ZHANG B H, MA C Z, et al. Co-regularized nonnegative matrix factorization for evolving community detection in dynamic networks[J]. Information Sciences, 2020, 528: 265-279.

[23] SHIN K, ELIASSI-RAD T, FALOUTSOS C. CoreScope: Graph mining using k-core analysis—Patterns, anomalies and algorithms[C]//2016 IEEE 16th International Conference on Data Mining (ICDM). Piscataway: IEEE, 2016: 469-478.

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

[25] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10, 10008.

[26] MALLIAROS F D, VAZIRGIANNIS M. To stay or not to stay: Modeling engagement dynamics in social graphs [C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management - CIKM'13. New York: ACM, 2013: 469-478.

[27] ZHOU T, LÜ L, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[28] 成其伟, 陈启买, 贺超波, 等. 基于改进对称二值非负矩阵分解的重叠社区发现方法[J]. 计算机应用, 2020, 40(11): 3203-3210.
CHENG Q W, CHEN Q M, HE C B, et al. Overlapping community detection method based on improved symmetric binary nonnegative matrix factorization[J]. Journal of Computer Applications, 2020, 40(11): 3203-3210. (in Chinese)

[29] JIAO P F, LYU H D, LI X M, et al. Temporal community detection based on symmetric nonnegative matrix factorization[J]. International Journal of Modern Physics B, 2017, 31(13): 1750102.

[30] LIN Y R, CHI Y, ZHU S H, et al. Facetnet: A framework for analyzing communities and their evolutions in dynamic networks[C]//Proceedings of the 17th International Conference on World Wide Web. New York: ACM, 2008: 685-694.

[31] WANG P, GAO Z L, MA X K. Dynamic community detection based on network structural perturbation and topological similarity[J]. Journal of Statistical Mechanics: Theory and Experiment, 2017, 2017(1): 013401.

[32] GREENE D, DOYLE D, CUNNINGHAM P. Tracking the evolution of communities in dynamic social networks [C]//2010 International Conference on Advances in Social Networks Analysis and Mining. Piscataway: IEEE, 2010: 176-183.

[33] 吴旻诚. 多层网络中结构与动力学研究[D]. 浙江: 浙江大学, 2021.
WU M C. Research on the Structure and Dynamics in Multilayer Networks[D]. Zhejiang: Zhejiang University,

2021. (in Chinese)

作者简介

贺超波 男,1981年9月出生于广东省河源市. 现为华南师范大学计算机学院教授、博士生导师,主要研究方向为图数据挖掘与智能教育.
E-mail: hechaobo@foxmail.com

成其伟 男,1997年8月出生于广东省江门市. 现工作于维沃移动通信有限公司,主要研究方向为社区发现.
E-mail: chengqiwei13@163.com

程俊伟 男,1997年9月出生于福建省宁德市. 现为华南师范大学计算机学院博士研究生,主要研究方向为图数据挖掘.
E-mail: jung@m.scnu.edu.cn

刘星雨 女,1998年12月出生于广东省湛江市. 现为华南师范大学计算机学院硕士研究生,主要研究方向为图数据挖掘.
E-mail: liuxingyu@m.scnu.edu.cn

余 鹏 男,2000年7月出生于湖北省武汉市. 现为华南师范大学计算机学院硕士研究生,主要研究方向为图数据挖掘.
E-mail: yupeng@m.scnu.edu.cn

陈启买 男,1965年5月出生于湖南省衡阳市. 现为华南师范大学教师教学发展中心教授,主要研究方向为智能教育理论与技术.
E-mail: chenqimai@m.scnu.edu.cn