动态属性网络的语义社区发现及演化分析方法

贺超波[1,2],成其伟[3],程俊伟[1],杨佳琦[1*],程 颢[1],汤 庸[1,2]
(1. 华南师范大学计算机学院,广东广州 510631;2. 琶洲实验室,广东广州 510335;3. 维沃移动通信有限公司,广东东莞 523859)

摘 要: 动态属性网络的语义社区发现及演化分析具有重要研究价值,其包含动态社区发现、社区语义解释及社区演化分析三个任务,但现有方法均难以同时实现. 针对该问题,提出一种基于联合非负矩阵分解的方法DAN-NMF(NMF for Dynamic Attributed Networks). DAN-NMF可以统一集成网络拓扑结构信息、节点属性信息及社区演化平滑约束信息,并利用最大最小化优化框架推导相关因子矩阵的迭代更新规则,从而可以直接获得动态社区发现、社区语义解释及社区演化分析结果. 在人工合成和真实的动态属性网络进行大量相关实验,结果表明DAN-NMF比最优的基准方法在准确性指标上至少提高了7.3%. 此外,在真实动态属性网络上的相关数据分析结果也表明DAN-NMF能够有效地发现动态社区的演化模式,并提供丰富的社区语义解释.

关键词: 动态属性网络;动态社区发现;社区语义解释;社区演化分析;非负矩阵分解
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)11-3757-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230239

A Community Discovery and Evolution Analysis Method for Dynamic Attributed Networks

HE Chao-bo[1,2], CHENG Qi-wei[3], CHENG Jun-wei[1], YANG Jia-qi[1*], CHENG Hao[1], TANG Yong[1,2]
(1. *School of Computer Science*, *South China Normal University*, *Guangzhou*, *Guangdong* 510631, *China*;
2. *Pazhou Lab*, *Guangzhou*, *Guangdong* 510335, *China*;
3. *Vivo Mobile Communication Co.*, *Ltd.*, *Dongguan*, *Guangdong* 523859, *China*)

Abstract: The topic of semantic community discovery and evolution analysis in dynamic attributed networks has important research value. It needs to simultaneously accomplish the tasks of dynamic community discovery, community semantic interpretation and community evolution analysis, but existing methods are difficult to achieve this goal. In view of this, this paper proposes a method DAN-NMF (NMF for Dynamic Attributed Networks) based on joint nonnegative matrix factorization. DAN-NMF can uniformly integrate network topology information, attribute information and smooth constraint information from community evolution, and derive iterative update rules of the related factor matrices using the majorization-minimization optimization framework, which helps it to directly obtain the results of dynamic community discovery, community semantic interpretation and community evolution analysis. Extensive experiments are conducted on multiple synthetic and real-world dynamic attributed networks. The results show that DAN-NMF has improved by at least 7.3% in term of accuracy metric, compared to the optimal baseline. Moreover, the data analysis results on real-world dynamic attributed networks also demonstrate that DAN-NMF can effectively discover the evolution patterns of dynamic communities and provide rich community semantic interpretations.

Key words: dynamic attributed networks; dynamic community discovery; community semantic interpretation; community evolution analysis; nonnegative matrix factorization
Foundation Item(s): National Natural Science Foundation of China (No.62077045)

收稿日期:2023-03-14;修回日期:2023-05-24;责任编辑:王天慧
*通讯作者:杨佳琦

1 引言

现实世界中的各种复杂网络(如社交网络、合著关系网络以及通信网络等)除了具有拓扑结构信息外,还常包含许多额外的属性信息(如节点关联的文本内容特征词、人口统计学属性以及文本标签等).此外,这些网络随着时间的推移,其拓扑结构信息和属性信息都会产生变化,即具备动态性.例如,在社交网络中,用户间的好友关系会动态地建立或消失,即拓扑结构发生改变.此外,如果用户不断发表评论,文本内容的增多会导致其属性信息动态变化.又如在合著关系网络中,作者不断建立合著关系及发表新的论文,也会导致网络拓扑结构及属性信息的变化.形式上,这些具有属性信息,又兼具有动态性的网络,被称为动态属性网络.

动态属性网络包含拓扑结构信息、属性信息以及时间信息,信息更为丰富,可以更好地表示现实世界中的各种复杂交互系统,与静态拓扑网络或者静态属性网络相比也更具有研究价值.语义社区发现及演化分析是动态属性网络的重要研究话题之一,它旨在检测节点间链接紧密、属性高度相似的聚簇,并对聚簇结果进行语义解释和分析其演化关系,即需要同时完成动态社区发现、社区语义解释及社区演化分析三个任务.近几十年来,复杂网络社区发现问题已得到广泛研究,并提出了许多不同类型的解决方法,文献[1~4]分别综述了目前流行的基于标签传播的方法、基于非负矩阵分解(Nonnegative Matrix Factorization,NMF)的方法、基于博弈论的方法以及基于深度学习的方法等.从这些综述文献的相关分析可以看出,尽管已有方法都展示出了一定的有效性,但都最多能同时完成动态社区发现、社区语义解释及社区演化分析中的两个任务,因此动态属性网络的语义社区发现及演化分析仍然需要研究一种更为有效的方法.

基于上述分析,本文提出一种面向动态属性网络的语义社区发现及演化分析方法DAN-NMF(NMF for Dynamic Attributed Networks).DAN-NMF首先构建联合NMF模型统一集成网络拓扑结构信息、节点属性信息以及社区演化平滑约束信息,然后推导乘性迭代更新规则对模型进行优化求解,最后通过分解得到的相应因子矩阵同时完成动态社区发现、社区语义解释及社区演化分析三个任务.

2 相关工作

2.1 动态社区发现

动态属性网络属于动态网络范畴,目前面向动态网络的社区发现(即动态社区发现)是复杂网络分析的研究热点,现有的相关方法主要包括三类: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.基于两阶段的方法核心思想是第一个阶段先对每一时刻的网络采用静态社区发现算法独立地进行社区划分,然后在第二个阶段对不同时刻的网络社区进行匹配,从而识别社区演化过程,代表性方法包括CPM(Clique Percolation Method)[5]、DCDA(Dynamic Community Detection Algorithm)[6]、ARIMA(AutoRegressive Integrated Moving Average)[7]以及MOCCD(Multi-Objective optimization based on Characteristics for Community Discovery)[8]等.基于增量学习的方法具有运行效率高的优势,因为其在获得初始社区结构后,只需要利用未来时刻网络的增量变化来改变部分节点的社区隶属情况,而无需再重新考虑整个网络.例如,Agarwal等人[9]提出的DyPerm(Dynamic Permanence)方法、Meng等人[10]提出的IDBLINK(Incremental Density-Based LINK)方法、郭昆等人[11]提出的IDCDDC(Incremental Dynamic Community Detection based on Density Clustering)方法以及Park等人[12]提出的CGC(Contrastive Graph Clustering)方法等都可以高效应用于较大规模的动态网络.基于演化聚类的方法是在假设网络结构变化缓慢条件下,通过演化聚类框架[13]同时优化两个目标:一是当前时刻网络的社区发现准确性,二是前一时刻与后一时刻网络社区划分结果的相似性.借助统一的优化策略,两个优化目标可以相互增强并提升整体性能.由于演化聚类框架更为可靠且易扩展,大量基于该框架的动态社区发现方法被提出,其中包括ADFLS(Attention Deep Forest based on Latent Structure)[14]、DGR-NMF(Dynamic Graph Regularized NMF)[15]和Cr-ENMF(Co-regularized Evolutionary NMF)[16]等方法.

早期的动态社区发现方法更多适用于无属性的动态网络,由于目前动态属性网络日益普遍,一些针对动态属性网络的社区发现方法已开始被陆续提出.例如,Bello等人[17]提出对动态属性网络进行社区发现,采用融合节点属性信息的Louvain方法[18]对每个时刻的网络进行社区划分.Appel等人[19]提出同时考虑链接、内容及时间信息的共享因子分解模型Chimera,以多维形式提取网络的潜在语义社区结构.Zhang等人[20]提出集成网络拓扑结构、文本及时间信息的生成模型DTCD(Dynamic Topical Community Detection),可以发现结构和主题上都有意义的社区.此外,Li等人[21]通过将时间与癌症基因属性信息集成到TANMF(Temporal Attributed NMF)模型,能够准确检测癌症动态属性网络存在的社区.

总的来说,现有动态社区发现方法最多能同时完成动态属性网络语义社区发现及演化分析涉及的其中两个任务.例如,DGR-NMF及Cr-ENMF只关注于利用拓扑结构信息进行动态社区发现.DTCD、Chimera及

TANMF等虽然集成了拓扑信息和属性信息进行动态社区发现及社区语义解释,但并没有考虑社区演化分析,CPM、DCDA及ARIMA等虽然能检测动态社区和追踪社区演化,但没有集成利用属性信息进行社区语义解释. 区别于已有的方法,本文提出的DAN-NMF方法能够同时完成动态社区发现、社区语义解释及社区演化分析三个任务.

2.2 NMF与社区发现

NMF是一种包含非负值约束的低秩矩阵分解方法[22]. 形式上,给定一个大小为$d_1 \times d_2$的非负矩阵$\boldsymbol{Y}$和一个正整数$k(k \ll \min(d_1, d_2))$,NMF可以将$\boldsymbol{Y}$分解为两个大小分别为$d_1 \times k$、$k \times d_2$的非负矩阵$\boldsymbol{P}$和$\boldsymbol{Q}$,且$\boldsymbol{P}$和$\boldsymbol{Q}$的乘积接近于$\boldsymbol{Y}$: $\boldsymbol{Y} \approx \boldsymbol{PQ}$,常用Frobenius范数的平方$\|\boldsymbol{Y}-\boldsymbol{PQ}\|_{\mathrm{F}}^2$度量其中误差. 文献[23]指出,当$d_1 = d_2$,即是对称矩阵时,NMF可以进一步转换为与聚类算法K-means等价的对称分解形式: $\boldsymbol{Y} \approx \boldsymbol{PP}^{\mathrm{T}}$,其中$\boldsymbol{P}$可以视为聚类结果指示矩阵. 由于对称NMF具有聚类能力,目前已被大量用于解决复杂网络的社区发现问题,此时$\boldsymbol{Y}$和$\boldsymbol{P}$分别作为拓扑结构矩阵和社区隶属强度表示矩阵,$\boldsymbol{PP}^{\mathrm{T}}$被视为基于两两节点的社区隶属强度表示向量重建网络的拓扑结构$\boldsymbol{Y}$. 文献[2]较为全面地归纳分析了基于对称NMF的各类社区发现方法,并指出对称NMF具有很好的可扩展性,可以通过联合对称NMF或者加入正则约束项处理各种网络的社区发现问题. 例如,DGR-NMF[15]、Cr-ENMF[16]及RSECD[24]等均是对称NMF的扩展方法. 本文提出的动态属性网络社区发现方法DAN-NMF也是基于对称NMF进行扩展,通过加入联合分解项及正则项集成更多相关信息以获得更为准确的动态社区发现结果.

3 语义社区发现及演化分析方法DAN-NMF

3.1 问题定义与符号说明

对于给定的动态属性网络,本文将其建模为一个包含T个时刻的快照属性网络集合,表示为$G=\{G_1, G_2, \cdots, G_T\}$,其中任何一个快照属性网络都建模为无向无权图$G=(V,E,X)$,其中$V=\{v_1, v_2, \cdots, v_n\}$为节点集合,$E=\{e_1, e_2, \cdots, e_l\}$为边集合,$X=\{x_1, x_2, \cdots, x_m\}$为节点属性集合. 对于节点集合$V$中的任意一个节点$v_i$,都关联一个$1 \times m$维的属性向量$\boldsymbol{S}_i=[s_{i1}, s_{i2}, \cdots, s_{im}]$,其元素$s_{ij}$取值为0或1,取值为1时表示节点$v_i$具备属性$x_j$,取值为0时则相反. 对于给定的一个动态属性网络$\mathcal{G}$,本文的问题在于同时完成其语义社区发现及演化分析涉及的三个任务:动态社区发现、社区语义解释以及社区演化分析,这三个任务的具体要求分别说明如下:

(1)动态社区发现:给定一个动态属性网络$G=\{G_1, G_2, \cdots, G_T\}$,将其每个时刻的属性网络$G_t \in G$划分为$k$个不相交的社区$C_t=\{C_{1,t}, C_{2,t}, \cdots, C_{k,t}\}$,其中对于$G_t$中的任意两个社区$C_{i,t}, C_{j,t} \in C_t$都有$C_{i,t} \cap C_{j,t} = \varnothing$. 此外,同一社区节点之间的链接相对于不同社区节点之间的链接应更为稠密,同一社区节点间所具有的节点属性相对于不同社区节点之间的节点属性也应更为相似.

(2)社区语义解释:给定第t时刻的第i个社区$C_{i,t}$,需要推断该社区$C_{i,t}$所对应的潜在主题$\boldsymbol{Z}_{i,t}$. 主题$\boldsymbol{Z}_{i,t}$可以使用一个向量$(\boldsymbol{Z}_{1i,t}, \boldsymbol{Z}_{2i,t}, \cdots, \boldsymbol{Z}_{mi,t})$表示,其中$\boldsymbol{Z}_{ji,t}$表示属性$j$隶属于主题$\boldsymbol{Z}_{i,t}$的强度.

(3)社区演化关系分析:给定相邻时刻的任意两个社区$C_{i,t-1}$和$C_{j,t}$,可以量化社区$C_{i,t-1}$转移到社区$C_{j,t}$的概率p_{ij}.

为方便后续介绍DAN-NMF方法的具体实现,在表1列出常用符号.

3.2 模型构建

(1)建模拓扑结构信息. 对任一给定的第t时刻快照属性网络G_t,定义一个表示节点社区隶属强度的矩阵$\boldsymbol{H}_t$,其元素$H_{ir,t}$表示为在时刻t中,节点i隶属于社区r的强度. $H_{ir,t}$与$H_{jr,t}$的乘积可以视为节点i与节点j在第r个社区中的交互概率. 通过对所有的k_t个社区进行求和:$\sum_{r=1}^{k_t} H_{ir,t} H_{jr,t}$,可得到节点$i$与节点$j$总的交互概率. 这种节点预期交互概率的计算过程表明如果两个节点具有相似的社区隶属关系则它们之间将有可能建立

表1　符号说明

符号	说明
$\boldsymbol{A}_t$	第t时刻属性网络快照的邻接矩阵,用于表示属性网络快照的拓扑结构信息,大小为$n \times n$
$\boldsymbol{S}_t$	第t时刻属性网络快照的节点属性矩阵,大小为$n \times m$
k_t	第t时刻属性网络快照的社区数和主题数
$\boldsymbol{H}_t$	第t时刻节点社区隶属强度矩阵,大小为$n \times k_t$
$\boldsymbol{B}_t$	第t时刻节点主题隶属强度矩阵,大小为$n \times k_t$
$\boldsymbol{W}_t$	第t时刻属性主题隶属强度矩阵,大小为$m \times k_t$
$\boldsymbol{M}_t$	第t时刻社区主题映射矩阵,大小为$k_t \times k_t$
$\boldsymbol{G}_t$	第$t-1$时刻到第t时刻的社区演化关系矩阵,大小为$k_{t-1} \times k_t$

边，即有 $A_{ij,t}\approx\sum_{r=1}^{k_t}H_{ir,t}H_{jr,t}$. 将该式扩展到所有节点对，可建模通过社区重建网络拓扑结构的损失为

$$L_{\text{link}}(\boldsymbol{H}_t)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij,t}-\sum_{r=1}^{k_t}H_{ir,t}H_{jr,t}\right)=\|\boldsymbol{A}_t-\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\|_{\mathrm{F}}^2 \tag{1}$$

(2)建模属性信息. 令 $\boldsymbol{B}_t$ 表示节点主题隶属强度矩阵，其元素 $B_{ip,t}$ 的含义为在时刻 t 中，节点 i 隶属于主题 p 的强度. 令 $\boldsymbol{W}_t$ 表示属性主题隶属强度矩阵，其元素 $W_{jp,t}$ 的含义为在时刻 t 中，属性 j 隶属于主题 p 的强度. 节点 i 具备属性 j 的概率可通过 $\sum_{p=1}^{k_t}B_{ip,t}W_{jp,t}$ 进行计算，该计算过程表明若节点 i 与属性 j 隶属于同一个潜在主题，那么节点 i 应倾向具备属性 j，即有 $S_{ij,t}\approx\sum_{p=1}^{k_t}B_{ip,t}W_{jp,t}$，扩展至所有的节点属性对，可建立如下损失项：

$$L_{\text{attribute}}(\boldsymbol{B}_t,\boldsymbol{W}_t)=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(S_{ij,t}-\sum_{p=1}^{k_t}B_{ip,t}W_{jp,t}\right)=\|\boldsymbol{S}_t-\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}}\|_{\mathrm{F}}^2 \tag{2}$$

(3)建模拓扑结构信息和属性信息的融合. 为了促进拓扑信息和属性信息相互融合，同时令模型能够具备一定的鲁棒性，以便应对社区与主题相互冲突的情况，引入社区主题映射对称矩阵 $\boldsymbol{M}_t$，其元素 $M_{ij,t}$ 表明在时刻 t 中，社区 i 映射到主题 j 的强度.

具体损失项的构造如下：

$$L_{\text{combine}}(\boldsymbol{H}_t,\boldsymbol{B}_t)=\|\boldsymbol{H}_t\boldsymbol{M}_t-\boldsymbol{B}_t\|_{\mathrm{F}}^2+\|\boldsymbol{M}_t-\boldsymbol{I}\|_{\mathrm{F}}^2 \tag{3}$$

其中，$\boldsymbol{I}$ 为单位矩阵，用于约束每一个社区都尽可能地对应一个主题.

(4)建模社区演化模式. 引入社区转移矩阵 $\boldsymbol{G}_t$ 描述第 $t-1$ 时刻到第 t 时刻的社区演化模式，其元素 $G_{ij,t}$ 表示第 $t-1$ 时刻的社区 i 在第 t 时刻转移到社区 j 的概率. 当 $\boldsymbol{H}_{t-1}$ 与 $\boldsymbol{G}_t$ 的乘积近似于 $\boldsymbol{H}_t$ 时，即 $\boldsymbol{H}_{t-1}\boldsymbol{G}_t\approx\boldsymbol{H}_t$，$\boldsymbol{G}_t$ 能表达相邻时刻社区间的转移概率，由此产生的损失可表示为

$$L_{\text{evolution}}(\boldsymbol{H}_t)=\|\boldsymbol{H}_{t-1}\boldsymbol{G}_t-\boldsymbol{H}_t\|_{\mathrm{F}}^2 \tag{4}$$

(5)建模社区演化的平滑性. 动态属性网络的节点社区隶属关系通常具有平滑性，即节点在相邻时刻的社区隶属关系变化是缓慢的. 为保留该平滑性，一般地在 $t-1$ 时刻的属性网络中两个节点的局部拓扑结构相似性越大，则认为在当前时刻 t 中这两个节点越可能隶属于同一个社区. 为提高计算效率，这里直接使用上一时刻表示节点间是否连接的0/1值作为相似性，由此可构造如下损失项：

$$\begin{aligned}L_{\text{smooth\_C}}(\boldsymbol{H}_t)&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij,t-1}\|\boldsymbol{H}_{i\cdot,t}-\boldsymbol{H}_{j\cdot,t}\|_{\mathrm{F}}^2\\&=\mathrm{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(D(\boldsymbol{A}_{t-1})-\boldsymbol{A}_{t-1}\right)\boldsymbol{H}_t\right)\end{aligned} \tag{5}$$

其中，$D(\boldsymbol{X})$ 以矩阵 $\boldsymbol{X}$ 为输入，输出如下对角矩阵：

$$D(\boldsymbol{X})_{ij}=\begin{cases}\sum_{p=1}^{n}X_{ip}, & \text{if } i=j\\ 0, & \text{if } i\neq j\end{cases} \tag{6}$$

同样地，节点主题隶属关系也具有平滑性，即在相邻时刻节点所属主题的变化是缓慢的. 可以认为在 $t-1$ 时刻的属性网络中，如果两个节点在主题状态空间中距离越近，那么在当前时刻中这两个节点越可能隶属于同一社区，对应损失项为

$$\begin{aligned}L_{\text{smooth\_T}}(\boldsymbol{H}_t)&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}}\right)_{ij}\|\boldsymbol{H}_{i\cdot,t}-\boldsymbol{H}_{j\cdot,t}\|_{\mathrm{F}}^2\\&=\mathrm{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(D\left(\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}}\right)-\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}}\right)\boldsymbol{H}_t\right)\end{aligned} \tag{7}$$

(6)统一模型. 以上损失项涉及的各个矩阵都具有非负值约束，因此可以基于联合NMF框架集成上述损失项构建如下面向动态属性网络的语义社区发现及演化分析模型：

$$\min L(\boldsymbol{H}_t,\boldsymbol{B}_t,\boldsymbol{W}_t,\boldsymbol{M}_t,\boldsymbol{G}_t)=\begin{cases}\|\boldsymbol{A}_t-\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\|_{\mathrm{F}}^2+\|\boldsymbol{S}_t-\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}}\|_{\mathrm{F}}^2\\ \quad+\alpha\left(\|\boldsymbol{H}_t\boldsymbol{M}_t-\boldsymbol{B}_t\|_{\mathrm{F}}^2+\|\boldsymbol{M}_t-\boldsymbol{I}\|_{\mathrm{F}}^2\right), & \text{if } t=1\\ \|\boldsymbol{A}_t-\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\|_{\mathrm{F}}^2+\|\boldsymbol{S}_t-\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}}\|_{\mathrm{F}}^2\\ +\alpha(\|\boldsymbol{H}_t\boldsymbol{M}_t-\boldsymbol{B}_t\|_{\mathrm{F}}^2+\|\boldsymbol{M}_t-\boldsymbol{I}\|_{\mathrm{F}}^2)\\ +\beta\|\boldsymbol{H}_{t-1}\boldsymbol{G}_t-\boldsymbol{H}_t\|_{\mathrm{F}}^2+\gamma\mathrm{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(F\left(\boldsymbol{A}_{t-1}\right)\right)\boldsymbol{H}_t\right) & \text{if } t>1\\ +\eta\mathrm{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(F\left(\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}}\right)\right)\boldsymbol{H}_t\right),\end{cases} \tag{8}$$

其中，α,β,γ,η 为相应损失项的权重参数，$F(\boldsymbol{X})=D(\boldsymbol{X})-\boldsymbol{X}$. 当 $t=1$ 时，动态属性网络只有1个快照，因此不需要建模社区演化模式及平滑性.

3.3 模型优化及算法分析

采用最大最小化框架[25]可以迭代优化求解式(8)目标函数的局部最优解. 该框架轮流迭代更新变量 $\boldsymbol{H}_t$、$\boldsymbol{B}_t$、$\boldsymbol{W}_t$、$\boldsymbol{M}_t$ 及 $\boldsymbol{G}_t$，在更新其中一个变量时，将其余变量看作常数，因此可以将式(8)模型的优化问题划分为5个变量的子优化问题. 具体地，由于每个变量都具有非负值约束，那么对每个变量的优化都可以视为受限求极值问题，可以通过拉格朗日乘数法和KKT(Karush-Kuhn-Tucker)条件推导其迭代更新规则，下面分别给出每个变量的优化过程.

(1)优化变量 $\boldsymbol{H}_t$

当 $t>1$ 时，将 $\boldsymbol{B}_t$、$\boldsymbol{W}_t$、$\boldsymbol{M}_t$、$\boldsymbol{G}_t$ 看作常量. 由于 $\|\boldsymbol{X}\|_{\mathrm{F}}^2=$

$\mathrm{tr}(\boldsymbol{XX}^{\mathrm{T}}),\mathrm{tr}(\boldsymbol{X})=\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}})$,式(8)可被重写为

$$\begin{aligned}L(\boldsymbol{H}_t)=&\mathrm{tr}(\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}-2\boldsymbol{A}_t\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}})\\&+\alpha\mathrm{tr}(\boldsymbol{M}_t^{\mathrm{T}}\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\boldsymbol{M}_t-2\boldsymbol{M}_t^{\mathrm{T}}\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{B}_t)\\&+\beta\mathrm{tr}(\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}-2\boldsymbol{H}_t\boldsymbol{G}_t^{\mathrm{T}}\boldsymbol{H}_{t-1}^{\mathrm{T}})\\&+\gamma\mathrm{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(D(\boldsymbol{A}_{t-1})-\boldsymbol{A}_{t-1}\right)\boldsymbol{H}_t\right)\\&+\eta\mathrm{tr}\left(\boldsymbol{H}_t^{\mathrm{T}}\left(D(\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}})-\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}}\right)\boldsymbol{H}_t\right)\end{aligned}\tag{9}$$

设$\boldsymbol{\Psi}$为约束$\boldsymbol{H}_t\geqslant 0$的拉格朗日乘子,式(9)的拉格朗日函数为

$$L(\boldsymbol{H}_t)=L(\boldsymbol{H}_t)+\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{H}_t^{\mathrm{T}})\tag{10}$$

$L(\boldsymbol{H}_t)$对$\boldsymbol{H}_t$求导可得:

$$\begin{aligned}\frac{\partial L(\boldsymbol{H}_t)}{\partial \boldsymbol{H}_t}=&2\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t-2\boldsymbol{A}_t\boldsymbol{H}_t+\alpha\left(\boldsymbol{H}_t\boldsymbol{M}_t\boldsymbol{M}_t^{\mathrm{T}}-\boldsymbol{B}_t\boldsymbol{M}_t^{\mathrm{T}}\right)\\&+\beta\left(\boldsymbol{H}_t-\boldsymbol{H}_{t-1}\boldsymbol{G}_t\right)+\gamma\left(D(\boldsymbol{A}_{t-1})-\boldsymbol{A}_{t-1}\right)\boldsymbol{H}_t\\&+\eta\left(D(\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}})-\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}}\right)\boldsymbol{H}_t+\boldsymbol{\Psi}\end{aligned}\tag{11}$$

根据KKT条件$\Psi_{ij}H_{ij,t}=0$可得:

$$\begin{aligned}&\Big[2\left(\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\right)_{ij}-2\left(\boldsymbol{A}_t\boldsymbol{H}_t\right)_{ij}+\alpha\left(\boldsymbol{H}_t\boldsymbol{M}_t\boldsymbol{M}_t^{\mathrm{T}}-\boldsymbol{B}_t\boldsymbol{M}_t^{\mathrm{T}}\right)_{ij}\\&+\beta\left(\boldsymbol{H}_t-\boldsymbol{H}_{t-1}\boldsymbol{G}_t\right)_{ij}+\gamma\left(\left(D(\boldsymbol{A}_{t-1})-\boldsymbol{A}_{t-1}\right)\boldsymbol{H}_t\right)_{ij}\\&+\eta\left(\left(D(\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}})-\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}}\right)\boldsymbol{H}_t\right)_{ij}\Big]H_{ij,t}=0\end{aligned}\tag{12}$$

由式(12)可得$H_{ij,t}$的乘性迭代更新规则:

$$H_{ij,t}=H_{ij,t}\frac{\left(2\boldsymbol{A}_t\boldsymbol{H}_t+\alpha\boldsymbol{B}_t\boldsymbol{M}_t^{\mathrm{T}}+\beta\boldsymbol{H}_{t-1}\boldsymbol{G}_t+\gamma\boldsymbol{A}_{t-1}\boldsymbol{H}_t+\eta\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}}\boldsymbol{H}_t\right)_{ij}}{\left(2\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t+\alpha\boldsymbol{H}_t\boldsymbol{M}_t\boldsymbol{M}_t^{\mathrm{T}}+\beta\boldsymbol{H}_t+\gamma D(\boldsymbol{A}_{t-1})\boldsymbol{H}_t+\eta D(\boldsymbol{B}_{t-1}\boldsymbol{B}_{t-1}^{\mathrm{T}})\boldsymbol{H}_t\right)_{ij}}\tag{13}$$

当$t=1$时,由上述过程同理可得$H_{ij,t}$的乘性迭代更新规则:

$$H_{ij,t}=H_{ij,t}\frac{\left(2\boldsymbol{A}_t\boldsymbol{H}_t+\alpha\boldsymbol{B}_t\boldsymbol{M}_t^{\mathrm{T}}\right)_{ij}}{\left(2\boldsymbol{H}_t\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t+\alpha\boldsymbol{H}_t\boldsymbol{M}_t\boldsymbol{M}_t^{\mathrm{T}}\right)_{ij}}\tag{14}$$

(2)优化变量$\boldsymbol{B}_t$

将$\boldsymbol{H}_t$、$\boldsymbol{W}_t$、$\boldsymbol{M}_t$、$\boldsymbol{G}_t$看作常量,式(8)在$t=1$和$t>1$时是等价的,其可被重写为

$$\begin{aligned}L(\boldsymbol{B}_t)=&\mathrm{tr}(\boldsymbol{W}_t\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}}-2\boldsymbol{S}_t^{\mathrm{T}}\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}})\\&+\alpha\mathrm{tr}(\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{B}_t-2\boldsymbol{M}_t^{\mathrm{T}}\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{B}_t)\end{aligned}\tag{15}$$

设$\boldsymbol{\Psi}$为约束$\boldsymbol{B}_t\geqslant 0$的拉格朗日乘子,式(15)的拉格朗日函数为

$$L(\boldsymbol{B}_t)=L(\boldsymbol{B}_t)+\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{B}_t^{\mathrm{T}})\tag{16}$$

$L(\boldsymbol{B}_t)$对$\boldsymbol{B}_t$求导可得

$$\frac{\partial L(\boldsymbol{B}_t)}{\partial \boldsymbol{V}_t}=\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{W}_t-\boldsymbol{S}_t\boldsymbol{W}_t+\alpha\left(\boldsymbol{B}_t-\boldsymbol{H}_t\boldsymbol{M}_t\right)+\boldsymbol{\Psi}\tag{17}$$

根据KKT条件$\psi_{ij}B_{ij,t}=0$可得

$$\Big[\left(\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{W}_t\right)_{ij}-\left(\boldsymbol{S}_t\boldsymbol{W}_t\right)_{ij}+\alpha\left(\boldsymbol{B}_t-\boldsymbol{H}_t\boldsymbol{M}_t\right)_{ij}\Big]B_{ij,t}=0\tag{18}$$

由式(18)可得$B_{ij,t}$的乘性迭代更新规则为

$$B_{ij,t}=B_{ij,t}\frac{\left(\boldsymbol{S}_t\boldsymbol{W}_t+\alpha\boldsymbol{H}_t\boldsymbol{M}_t\right)_{ij}}{\left(\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}}\boldsymbol{W}_t+\alpha\boldsymbol{B}_t\right)_{ij}}\tag{19}$$

(3)优化变量$\boldsymbol{W}_t$

将$\boldsymbol{H}_t$、$\boldsymbol{B}_t$、$\boldsymbol{M}_t$、$\boldsymbol{G}_t$看作常量,式(8)在$t=1$和$t>1$时是等价的,其可被重写为

$$L(\boldsymbol{W}_t)=\mathrm{tr}(\boldsymbol{W}_t\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}}-2\boldsymbol{S}_t^{\mathrm{T}}\boldsymbol{B}_t\boldsymbol{W}_t^{\mathrm{T}})\tag{20}$$

设$\boldsymbol{\Psi}$为约束$\boldsymbol{W}_t\geqslant 0$的拉格朗日乘子,式(20)的拉格朗日函数为

$$L(\boldsymbol{W}_t)=L(\boldsymbol{W}_t)+\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{W}_t^{\mathrm{T}})\tag{21}$$

$L(\boldsymbol{W}_t)$对$\boldsymbol{W}_t$求导可得

$$\frac{\partial L(\boldsymbol{W}_t)}{\partial \boldsymbol{W}_t}=\boldsymbol{W}_t\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{B}_t-\boldsymbol{S}_t\boldsymbol{B}_t+\boldsymbol{\Psi}\tag{22}$$

根据KKT条件$\Psi_{ij}W_{ij,t}=0$,可得

$$\Big[\left(\boldsymbol{W}_t\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{B}_t\right)_{ij}-\left(\boldsymbol{S}_t\boldsymbol{B}_t\right)_{ij}\Big]W_{ij,t}=0\tag{23}$$

由式(23)可得$W_{ij,t}$的乘性迭代更新规则为

$$W_{ij,t}=W_{ij,t}\frac{\left(\boldsymbol{S}_t\boldsymbol{B}_t\right)_{ij}}{\left(\boldsymbol{W}_t\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{B}_t\right)_{ij}}\tag{24}$$

(4)优化变量$\boldsymbol{M}_t$

将$\boldsymbol{H}_t$、$\boldsymbol{B}_t$、$\boldsymbol{W}_t$、$\boldsymbol{G}_t$看作常量,式(8)在$t=1$和$t>1$时是等价的,其可被重写为

$$\begin{aligned}L(\boldsymbol{M}_t)=&\mathrm{tr}(\boldsymbol{M}_t^{\mathrm{T}}\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\boldsymbol{M}_t-2\boldsymbol{M}_t^{\mathrm{T}}\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{B}_t)\\&+\mathrm{tr}(\boldsymbol{M}_t^{\mathrm{T}}\boldsymbol{M}_t-2\boldsymbol{M}_t^{\mathrm{T}})\end{aligned}\tag{25}$$

设$\boldsymbol{\Psi}$为约束$\boldsymbol{M}_t\geqslant 0$的拉格朗日乘子,式(25)的拉格朗日函数如下:

$$L(\boldsymbol{M}_t)=L(\boldsymbol{M}_t)+\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{M}_t^{\mathrm{T}})\tag{26}$$

$L(\boldsymbol{M}_t)$对$\boldsymbol{M}_t$求导可得到式(27):

$$\frac{\partial L(\boldsymbol{M}_t)}{\partial \boldsymbol{M}_t}=\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\boldsymbol{M}_t-\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{B}_t+\boldsymbol{M}_t-\boldsymbol{I}+\boldsymbol{\Psi}\tag{27}$$

根据KKT条件$\psi_{ij}M_{ij,t}=0$,可得

$$\Big[\left(\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\boldsymbol{M}_t\right)_{ij}-\left(\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{B}_t\right)_{ij}+M_{ij,t}+I_{ij}\Big]W_{ij,t}=0\tag{28}$$

由上式可得$W_{ij,t}$的乘性迭代更新规则为

$$M_{ij,t}=M_{ij,t}\frac{\left(\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{B}_t+\boldsymbol{I}\right)_{ij}}{\left(\boldsymbol{H}_t^{\mathrm{T}}\boldsymbol{H}_t\boldsymbol{M}_t+\boldsymbol{M}_t\right)_{ij}}\tag{29}$$

(5)优化变量$\boldsymbol{G}_t$

由于$\boldsymbol{G}_t$在$t=1$时不存在,因此只需考虑$t>1$的情

况. 将$\boldsymbol{H}_t$、$\boldsymbol{B}_t$、$\boldsymbol{W}_t$、$\boldsymbol{M}_t$看作常量,式(8)可被重写为

$$L\left(\boldsymbol{G}_t\right)=\mathrm{tr}\left(\boldsymbol{G}_t^{\mathrm{T}}\boldsymbol{H}_{t-1}^{\mathrm{T}}\boldsymbol{H}_{t-1}\boldsymbol{G}_t-2\boldsymbol{G}_t^{\mathrm{T}}\boldsymbol{H}_{t-1}^{\mathrm{T}}\boldsymbol{H}_t\right) \tag{30}$$

设$\boldsymbol{\Psi}$为约束$\boldsymbol{G}_t\geq 0$的拉格朗日乘子,式(30)的拉格朗日函数如下:

$$L\left(\boldsymbol{G}_t\right)=L\left(\boldsymbol{G}_t\right)+\mathrm{tr}\left(\boldsymbol{\Psi}\boldsymbol{G}_t^{\mathrm{T}}\right) \tag{31}$$

$L\left(\boldsymbol{G}_t\right)$对$\boldsymbol{G}_t$求导可得到式(32):

$$\frac{\partial L\left(\boldsymbol{G}_t\right)}{\partial \boldsymbol{G}_t}=\boldsymbol{H}_{t-1}^{\mathrm{T}}\boldsymbol{H}_{t-1}\boldsymbol{G}_t-\boldsymbol{H}_{t-1}^{\mathrm{T}}\boldsymbol{H}_t+\boldsymbol{\Psi} \tag{32}$$

根据KKT条件$\psi_{ij}G_{ij,t}=0$,可得

$$\left[\left(\boldsymbol{H}_{t-1}^{\mathrm{T}}\boldsymbol{H}_{t-1}\boldsymbol{G}_t\right)_{ij}-\left(\boldsymbol{H}_{t-1}^{\mathrm{T}}\boldsymbol{H}_t\right)_{ij}\right]G_{ij,t}=0 \tag{33}$$

由(33)可得$G_{ij,t}$的乘性迭代更新规则为

$$G_{ij,t}=G_{ij,t}\frac{\left(\boldsymbol{H}_{t-1}^{\mathrm{T}}\boldsymbol{H}_t\right)_{ij}}{\left(\boldsymbol{H}_{t-1}^{\mathrm{T}}\boldsymbol{H}_{t-1}\boldsymbol{G}_t\right)_{ij}} \tag{34}$$

使用DAN-NMF方法可对动态属性网络同时进行动态社区发现、社区语义解释及社区演化关系分析,其总的执行流程如算法1所示.

从算法1可以看出,DAN-NMF最耗时的部分来自于各个矩阵变量的迭代更新操作. 具体地,迭代更新$\boldsymbol{H}_t$、$\boldsymbol{M}_t$及$\boldsymbol{G}_t$的时间复杂度为$O\left(kn^2\right)$,而迭代更新$\boldsymbol{B}_t$和$\boldsymbol{W}_t$的时间复杂度为$O(knm)$,其中k为所有属性网络快照的平均社区个数. 假设每个属性网络快照的平均迭代更新次数为t_{iter},则算法总的时间复杂度为$O\left(t_{\mathrm{iter}}Tkn(3n+2m)\right)$.

算法1 DAN-NMF语义社区发现及演化分析

输入: T个时刻的属性网络邻接矩阵$\boldsymbol{A}_{\mathrm{s}}=(\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_T)$、节点属性矩阵$\boldsymbol{S}_{\mathrm{s}}=(\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_T)$、社区数$k_s=\{k_1,k_2,\cdots,k_T\}$、权重参数$\alpha,\beta,\gamma,\eta$

输出: T个时刻的社区划分结果$C_{\mathrm{s}}=(C_1,C_2,\cdots,C_T)$、社区对应主题$\boldsymbol{Z}_{\mathrm{s}}=(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_T)$、$T-1$个社区演化矩阵$\boldsymbol{G}_{\mathrm{s}}=(\boldsymbol{G}_2,\boldsymbol{G}_3,\cdots,\boldsymbol{G}_T)$

```
C_s = ( ) //保存社区划分结果的矩阵
Z_s = ( ) //保存社区对应主题的矩阵
G_s = ( ) //保存社区演化矩阵
FOR t = 1 to T
    随机初始化 H_t, B_t, W_t, M_t, G_t
    WHILE 目标函数式(8)不收敛 DO
        IF t = 1
            根据式(14)更新 H_t
        ELSE
            根据式(13)更新 H_t
            根据式(34)更新 G_t
        END IF
        根据式(19)更新 B_t
        根据式(24)更新 W_t
        根据式(29)更新 M_t
    END WHILE
    根据 H_t 获得时刻 t 的社区划分结果 C_t,添加到 C_s 中
    IF t > 1
        对 G_t 进行规一化,即 ∀l, Σ_{i=1}^{n} G_{li,t} = 1,然后添加到 G_s
    END IF
    //获取每个社区对应的主题
    FOR C_{i,t} in C_t
        p = argmax_l M_{il}
        Z_{i,t} = W_{·p}
    END FOR
    将 Z_t 添加到集合 Z_s 中
END FOR
RETURN C_s, Z_s, G_s
```

4 实验

4.1 数据集

(1)人工合成动态属性网络

目前人工合成动态网络有两种较为流行的生成工具:Dynamic LFR[26]和RDyn[27],但两种工具生成的网络都没有属性信息,为此需要为Dynamic LFR和RDyn的合成动态网络统一定义节点属性信息的生成过程. 首先为每个节点生成属于该节点的$1\times m$维属性向量,其中属性向量生成的规则为:假若节点的真实社区编号为x(编号从1开始),属性向量下标从$(x-1)\times h$至$x\times h-1$按照参数$p_{\mathrm{in}}\in[0,1]$的概率生成1,剩余的位置按照参数$p_{\mathrm{out}}\in[0,1]$的概率生成0. 最后随机交换网络$p_{\mathrm{mis}}\times 100\%$节点的属性向量,其中$p_{\mathrm{mis}}\in[0,1]$,这可以模拟属性信息存在噪声的情况. 接下来分别介绍Dynamic LFR和RDyn生成的动态属性网络.

(a)Dynamic LFR网络. Dynamic LFR根据网络演化事件的不同生成4组共8个网络,每个网络的参数设置如表2所示,其中n表示网络节点总个数,T表示网络快照个数,$l_{\max}$为各快照网络的最大边数,mu表示混合参数. 参数p在各个网络的含义有所不同,在Switch网络中p表示的是节点相邻时刻发生社区转移的概率. 在BirthDeath网络中p表示的是:每个时刻中,$p\times 100\%$的社区发生出生事件及$p\times 100\%$的社区发生死亡事件;在ExpandContract网络中p表示的是:每个时刻中,$p\times 100\%$的社区会发生扩张事件及$p\times 100\%$的社区发生收缩事件;在MergeSplit网络中p表示的是:每个时刻中,$p\times 100\%$的社区发生合并事件及$p\times 100\%$的社区发生分裂事件. 参数r则是ExpandContract网络特有的参数,表示社区扩张或收缩的比例. 通过表2的参数设置情况可以看出,对于同组的网络,编号为2的网络比编号为1的网络进行社区发现的难度要更大,因为编号为2的网络相比编号为1的网络:演

表2　Dynamic LFR网络参数设置

网络名称	n	T	l_{max}	mu	p	r	p_{in}	p_{out}	p_{mis}	h
Switch1	400	10	1524	0.3	0.2	—	0.5	0.05	0.1	20
Switch2	400	10	1927	0.3	0.4	—	0.5	0.1	0.2	20
BirthDeath1	400	10	1350	0.3	0.125	—	0.5	0.05	0.1	20
BirthDeath2	400	10	1505	0.3	0.25	—	0.5	0.1	0.2	20
MergeSplit1	400	10	1491	0.3	0.25	—	0.5	0.05	0.1	20
MergeSplit2	400	10	1526	0.3	0.25	—	0.5	0.1	0.2	20
ExpandContrat1	5000	10	51718	0.3	0.125	0.2	0.5	0.05	0.1	20
ExpandContrat2	5000	10	50404	0.3	0.25	0.4	0.5	0.1	0.2	20

化事件的发生更为频繁,节点属性信息存在更多的噪声以及属性信息与社区结构的匹配程度更低.

(*b*)RDyn网络. 动态网络生成器RDyn不仅能模拟交互动态,还能模拟社区动态,这里利用RDyn一共生成4个网络:Network1、Network2、Network3以及Network4,各网络的参数如表3所示,其中iter为RDyn生成网络过程的迭代次数,q表示电导质量阈值. q的取值越高,网络的社区特征就越模糊. 由表3所示的参数设置可以看出,Network1与Network2拥有同样的网络拓扑结构,但生成的节点属性向量不同(Network3与Network4同理).

表3　RDyn网络参数设置

网络名称	n	T	l_{max}	iter	q	p_{in}	p_{out}	p_{mis}	h
Network1	400	24	877	80	0.2	0.5	0.05	0.1	20
Network2	400	24	920	80	0.2	0.5	0.1	0.2	20
Network3	400	24	1106	80	0.3	0.5	0.05	0.1	20
Network4	400	24	1138	80	0.3	0.5	0.1	0.2	20

(2)真实动态属性网络

从DBLP(https://dblp.org)抽取论文相关信息来构建真实动态属性网络,这些论文源自2010年至2021年发表在以下领域的权威学术会议或学术期刊:

(*a*)计算机网络:SIGCOMM、MobiCom、NSDI、INFOCOM、JSAC

(*b*)数据库与数据挖掘:SIGKDD、ICDE、ICDM、VLDB、SIGMOD

(*c*)计算机视觉与模式识别:CVPR、ICCV、ECCV

收集以上相关论文信息后,首先提取至少有8年都发表过论文的作者作为网络节点,然后提取作者间的合著关系作为网络的边,最后每2年的网络数据作为一个快照,共有6个快照属性网络. 每个网络的节点数均统一为1591,边数依次为1 468、1 934、2 236、2 347、2 319和1 757. 对于网络节点属性的构建,则按如下流程:对论文标题进行分词并且去除停用词,过滤掉出现次数小于10的词项,剩余的词项作为属性集,最终属性数量为5 052. 如果某作者论文标题中的词项对应属性集中的某个属性,则认为该作者对应的节点包含相应的属性. 此外,把作者在相应时间段内发表论文数最多的领域作为其所在网络快照中的真实社区标签.

4.2　对比方法及评价准则

为了对比验证DAN-NMF的性能,选择5个具有代表性的方法作为基准,包括DGR-NMF[15]、Chimera[19]、TANMF[21]、RSECD[24]以及FaceNet[28],表4归纳了各方法的特性,包括所能够利用的信息种类以及能够完成的任务种类. 正如前面指出的是,目前现有方法基本还不能同时完成动态社区发现、社区语义解释及社区演化分析三个任务.

表4　对比方法的特性

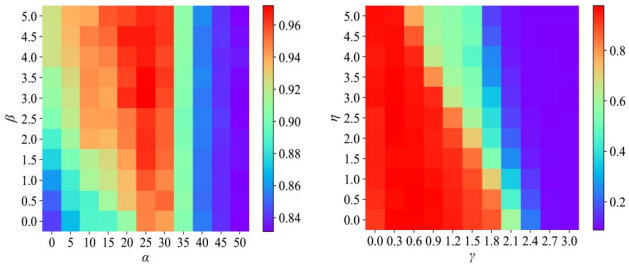
方法	是否利用拓扑结构信息	是否利用属性信息	是否利用社区演化平滑约束信息	能否检测动态社区	能否解释社区语义	能否分析社区演化
TANMF	√	√	×	√	√	×
RSECD	√	√	×	√	√	×
DGR-NMF	√	×	√	√	×	×
FaceNet	√	×	√	√	×	√
Chimera	√	√	√	√	√	×
DAN-NMF	√	√	√	√	√	√

由于所采用的人工合成及真实动态属性网络都具有真实社区标签，因此采用广泛使用的标准化互信息NMI(Normalized Mutual Information)[29]作为评价社区发现效果的性能指标，其值越大，则表示社区发现的性能越好.

4.3 实验结果与分析

(1)参数敏感性分析

DAN-NMF包含四个权重参数α、β、γ及η，需要对这些参数进行敏感性分析，以确定它们在实验中的最佳取值，这里选择Dynamic LFR的Switch1和BirthDeath1网络作为示例进行分析. 对于参数α和β，首先固定γ和η的取值为0，将α和β的取值范围分别设置为0到50和0到5，步长分别设置为5和0.5，然后计算10个时刻属性网络社区发现结果的平均NMI，最终结果如图1(a)和图2(a)所示. 可以看出当α取值为20至30并且β取值大于等于2时，DAN-NMF都能取得较大的NMI值. 同时需要指出的是，当α取值为0并且β取值为0时，NMI值较低，这说明融合节点拓扑结构与属性信息，以及考虑社区演化关系能够有效提高社区发现性能. 对于参数γ和η，首先固定$\alpha=20$和$\beta=2$，将γ和η的取值范围分别设置为从0到3和0到5，步长分别设置为0.3和0.5，然后计算10个时刻属性网络社区发现的平均NMI，最终结果如图1(b)和图2(b)所示. 可以看出γ取值为0.3至0.9且η取值大于0.5时，DAN-NMF都能取得较大的NMI值. 当γ取值为0且η取值为0时，NMI值相对较低，这说明考虑社区演化平滑约束信息有利于社区发现性能的提升.



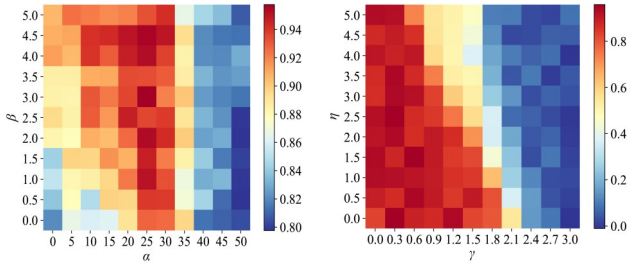
(a) $\alpha\beta$分析 (b) $\gamma\eta$分析

图1 Switch1网络上的参数敏感性分析

在其余动态属性网络进行相同的参数敏感性分析，也具有相似的发现，即α取值为20至30，β取值大于等于2，γ取值为0.3至0.9，η取值大于0.5时，DAN-NMF都能取得较大的NMI值. 在本文的所有实验中，都统一设置$\alpha=20$，$\beta=2$，$\gamma=0.3$及$\eta=1$.

(2)人工合成动态属性网络对比分析

(a)Dynamic LFR网络. 所有方法在Dynamic LFR的8个人工合成动态属性网络上分别重复运行20次，



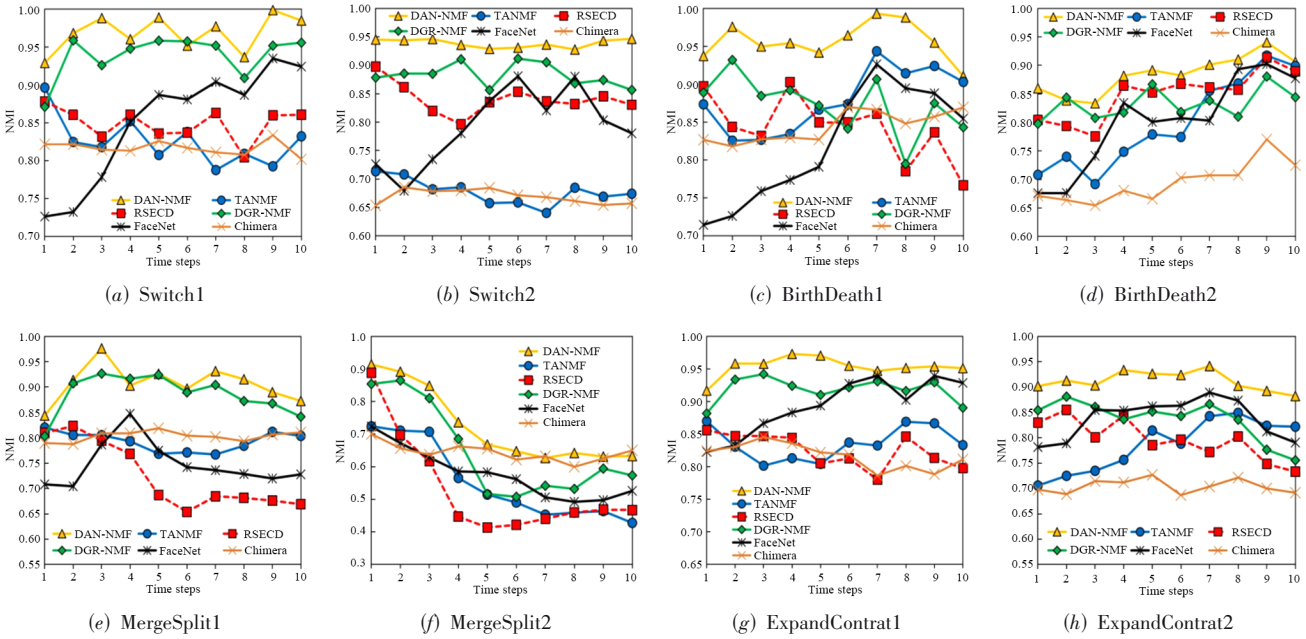
(a) $\alpha\beta$分析 (b) $\gamma\eta$分析

图2 BirthDeath1网络上的参数敏感性分析

各方法的NMI取平均值作为评价结果(图3). 从图3(a)~(h)可以看出，对比方法都存在这种问题：在面对某种类型网络时，社区发现性能会有较为明显的下降. 比如RSECD在Switch类型的网络中有着相对较好的性能，但面对MergeSplit类型的网络时表现最差. 又比如Chimera，当面对属性信息噪声增加及属性信息与社区结构匹配程度低的网络时，它的性能大多数情况下是最差的. 而本文提出的DAN-NMF方法无论应用在哪种类型的网络，都能在绝大部分时刻获得最好的NMI，即能够更准确地挖掘出网络潜在的社区结构. 这也说明DAN-NMF能够充分地利用动态属性网络中蕴含的拓扑结构信息、属性信息以及社区演化平滑约束信息，有效地提升社区发现的准确性及鲁棒性.

DAN-NMF可以进一步挖掘相邻时刻网络快照的社区演化模式. 为更直观地分析社区演化，可以对得到的社区演化矩阵$\boldsymbol{G}_t$进行可视化. 作为示例，选取Switch1和Swicth2网络各2个社区演化矩阵$\boldsymbol{G}_3$和$\boldsymbol{G}_7$进行分析. 由于社区演化矩阵$\boldsymbol{G}_t$的初始化是随机的，所以相同社区前一时刻的编号与后一时刻的编号可能并不相同. 为了便于分析，需将相同社区相邻时刻的编号调整为相同，具体调整过程为：对于$t-1$时刻社区编号为i的社区，它在时刻t中的相应编号通过$\underset{l}{\operatorname{argmax}}\, G_{il,t}$得到，然后交换$\boldsymbol{G}_t$的第$l$列与$\boldsymbol{G}_t$的第$i$列. 其余社区以此类推，从而令相邻时刻的同一社区编号也相同，对调整后的$\boldsymbol{G}_t$进行可视化的结果如图4所示.

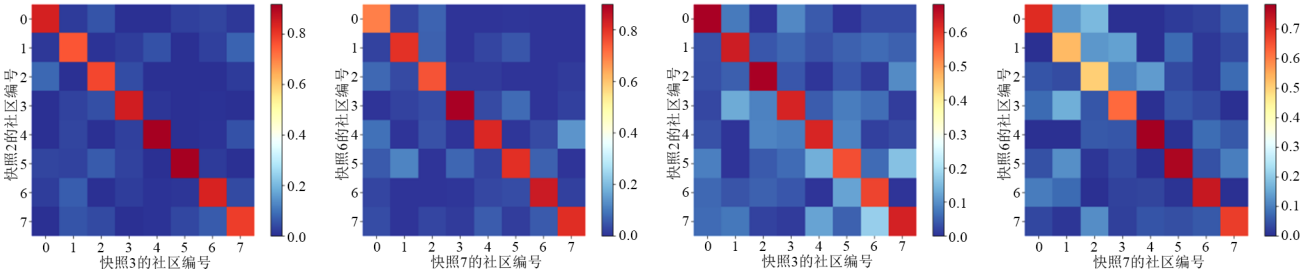
图4各个子图中第i行第j列格子的颜色深浅表示前一时刻快照的社区i转移到下一时刻快照社区j的概率. 颜色越深，社区i转移到社区j的概率就越大. 从图4(a)和图4(b)可以看到矩阵对角线基本都呈现深红色，说明Switch1网络的社区内大多数节点在下一时刻都会保留到当前社区，社区内只有少部分节点发生了社区转移. 而从图4(c)和图4(d)可以看到矩阵对角线所呈现的红色都相对较浅，说明Switch2网络中较多节点发生了社区转移. 回顾Switch1网络和Switch2网络所设置的演化过程：Switch1每个时刻都会随机挑选20%的节点改变其所属社区，剩余80%节点保留在原社区，而Switch2每个时刻都会随机



(a) Switch1　(b) Switch2　(c) BirthDeath1　(d) BirthDeath2

(e) MergeSplit1　(f) MergeSplit2　(g) ExpandContrat1　(h) ExpandContrat2

图3　Dynamic LFR网络的NMI比较



(a) Switch1 $\boldsymbol{G}_3$　(b) Switch1 $\boldsymbol{G}_7$　(c) Switch2 $\boldsymbol{G}_3$　(d) Switch2 $\boldsymbol{G}_7$

图4　社区演化矩阵$\boldsymbol{G}_t$的可视化

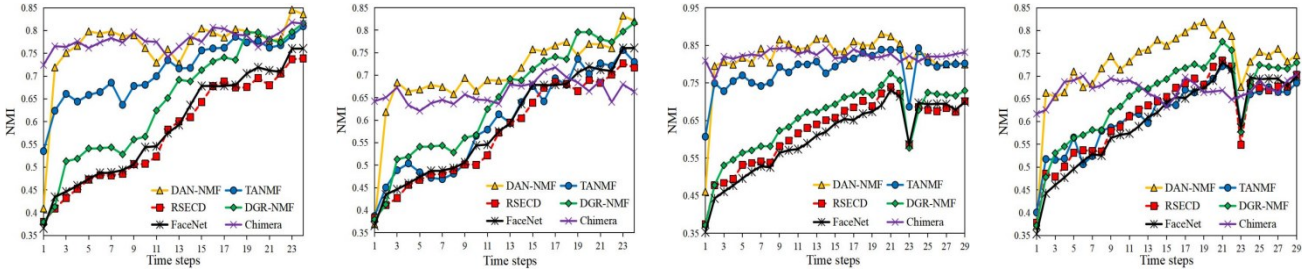
挑选40%的节点改变其所属社区,剩余60%节点保留在原社区. 根据前面的观察,不难发现这两个网络的演化过程与社区演化矩阵所表达的社区间转移概率高度吻合,这说明DAN-NMF的社区演化矩阵$\boldsymbol{G}_t$能准确表示相邻时刻社区间的转移关系.

(*b*)RDyn网络. 为进一步验证DAN-NMF的普适性,在另一类人工合成动态属性网络RDyn上进行对比实验. 所有方法在RDyn网络上重复运行20次,各方法的NMI取平均值作为评价结果(图5). 通过观察图5(*a*)和图5(*c*),可以看出在Network1和Network3中,DAN-NMF只在部分时刻的性能略低于Chimera,同时高于其余4个对比方法. 进一步地观察图5(*b*)和图5(*d*),可以看出在Network2和Network4中,DAN-NMF在绝大多数时刻都取得了最高性能. 这些结果表明DAN-NMF具备更强的鲁棒性,能够适应条件更为苛刻的网络. 需要指出的是,Chimera的NMI变化曲线波动较小,而其余对比方法(包括DAN-NMF)都存在较明显的波动(例如在Network3和Network4上). 这是因为Chimera会同时提取所有历史快照网络社区演化平滑约束信息来挖掘当前时刻网络的社区,从而使得其获得较为平滑的性能,当然也导致了更高的时间复杂度.

(3)真实动态属性网络对比分析

为了进一步验证DAN-NMF在真实网络同样具备有效性,在DBLP动态属性网络进行实验. 所有方法在DBLP网络上重复运行20次,各方法的NMI取平均值作为评价结果(图6). 通过图6可以看出,DAN-NMF在后四个时刻的网络快照都取得最高的NMI. 对于6个快照网络社区发现结果的平均NMI,最优的是DAN-NMF的0.561,次优的是Chimera的0.523,DAN-NMF比Chimera提高了7.3%. 以上结果表明DAN-NMF在真实网络上也具有良好的社区发现性能.

DAN-NMF不仅能够发现每个网络快照的社区,还能给出社区对应的语义解释. 具体地,DAN-NMF可以计算得到某个社区最关注的主题$Z=[Z_1,Z_2,\cdots,Z_m]$,Z_i表示词



(*a*) Network1 (*b*) Network2 (*c*) Network3 (*d*) Network4

图5 RDyn网络的NMI比较

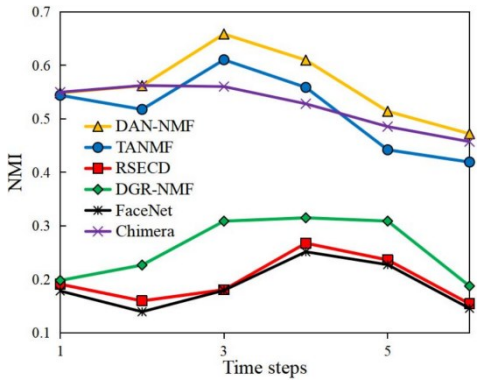


图6 DBLP网络的NMI比较

汇i与该主题关联的强度,通过对Z_i进行降序排序,并选取排序靠前的词汇作为主题的代表性词汇,最终这些词汇起到解释社区的作用. 以2018年至2019年的DBLP网络快照为例,各社区的语义解释如表5所示,词汇后面为其主题隶属强度. 通过表5可以看出,社区1的关键词汇包含目标检测(detection)、图像识别(recognition)及语义分割(segmentation),可以推测该社区的研究方向是计算机视觉与模式识别;社区2的关键词汇包含数据库查询(query)和图数据(graph),可以推测该社区的研究方向是数据库与数据挖掘;社区3的关键词汇包含移动通信(mobile)、无线网络(wireless)及云计算(cloud),可以推测该社区的研究方向是计算机网络.

进一步地,结合时间信息可以进一步分析社区语义的变化. 假设时刻t中社区i的代表性词汇集合为$\text{word}_{i,t}$,通过$\text{word}_{i,t}-\text{word}_{i,t-1}=\{x|x\in\text{word}_{i,t}\wedge x\notin\text{word}_{i,t-1}\}$来获得时刻$t$中社区$i$相对于时刻$t-1$的新增代表性词汇. 以表示计算机视觉与模式识别领域的社区为例,计算其在每一网络快照中的新增代表性词汇,结果如表6所示. 通过表6可以推断出如下信息:在2014年至2015年这个时间段,深度学习(deep learning)和卷积神经网络(convolutional neural networks)是热门话题;在2018年至2019年这个时间段,生成对抗网络(generative adversarial networks)和注意力机制(attention mechanism)变得流行;在2020年至2021年这个时间段,自监督学习(self-supervised learning)成为新兴热点.

表5 DBLP网络的社区语义解释

社区1	社区2	社区3
deep (26.74)	efficient (23.89)	networks (66.10)
image (24.57)	query (17.52)	mobile (22.92)
detection (22.99)	graph (15.01)	wireless (22.83)
video (21.20)	search (14.87)	system (15.38)
object (20.33)	system (13.26)	edge (13.23)
segmentation (19.51)	large (9.94)	distributed (12.36)
adversarial (18.13)	extend (9.62)	neural (12.11)
neural (17.58)	abstract (9.42)	scheduling (10.86)
visual (17.06)	processing (9.38)	communication (10.82)
estimation (16.93)	spatial (9.01)	information (10.78)
3D (16.25)	similarity (8.75)	joint (10.41)
recognition (16.04)	analytics (8.67)	analysis (10.16)
semantic (15.70)	management (8.63)	control (9.68)
networks (15.26)	dynamic (8.56)	social (9.36)
unsupervised (14.86)	framework (8.30)	adaptive (9.20)
feature (13.92)	discovery (8.17)	allocation (9.05)
convolutional (13.24)	distributed (7.86)	caching (8.91)
attention (12.68)	social (7.84)	resource (8.89)
pose (12.56)	fast (7.71)	cloud (8.25)
action (11.76)	top-k (7.63)	traffic (8.10)

表6 计算机视觉与模式识别社区各时间段的新兴热点词汇

2012—2013	2014—2015	2016—2017	2018—2019	2020—2021
mining segmentation information	convolutional neural deep semantic	unsupervised	generative adversarial attention	self-super- vised

综上所述,对于真实世界中的动态属性网络,DAN-NMF不但能够准确地挖掘网络的潜在社区,并对社区做出语义解释,还能够结合时间信息进一步地分析出社区语义的变化情况.

5 结论

由于现有方法难以同时完成动态属性网络语义社区发现及演化分析涉及的三个任务:动态社区发现、社区语义解释及社区演化分析,本文为此提出一种基于联合NMF

的方法DAN-NMF. DAN-NMF利用联合NMF框架集成网络拓扑结构信息、节点属性信息及社区演化平滑约束信息,并通过最大最小化优化策略同时学习社区隶属关系、社区隶属主题及社区演化模式. 在人工合成和真实动态属性网络进行了大量的对比实验,结果表明DAN-NMF不仅能高效地发现动态社区,还能挖掘社区演化模式及对社区进行语义解释. 由于DAN-NMF方法的时间复杂度近似为$O(n^2)$,这限制了它应用于大规模动态属性网络的能力. 在后续研究中,将从模型迭代更新算法优化设计及并行化实现这两个方向展开进一步研究,目的在于进一步提高DAN-NMF的运行效率.

参考文献

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

[2] 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.

[3] MOSCATO V, PICARIELLO A, SPERLÍ G. Community detection based on Game Theory[J]. Engineering Applications of Artificial Intelligence, 2019, 85: 773-782.

[4] JIN D, YU Z Z, JIAO P F, et al. A survey of community detection approaches: From statistical modeling to deep learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(2): 1149-1170.

[5] PALLA G, BARABÁSI A L, VICSEK T. Quantifying social group evolution[J]. Nature, 2007, 446: 664-667.

[6] DHOUIOUI Z, AKAICHI J. Tracking dynamic community evolution in social networks[C]//2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). Piscataway: IEEE, 2014: 764-770.

[7] 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. NewYork: ACM, 2015: 1509-1516.

[8] LI W M, ZHOU X K, YANG C, et al. Multi-objective optimization algorithm based on characteristics fusion of dynamic social networks for community discovery[J]. Information Fusion, 2022, 79: 110-123.

[9] AGARWAL P, VERMA R, AGARWAL A, et al. DyPerm: Maximizing permanence for dynamic community detection[M]// Advancesin Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2018: 437-449.

[10] MENG F R, ZHANG F, ZHU M, et al. Incremental density-based link clustering algorithm for community detection in dynamic networks[J]. Mathematical Problems in Engineering, 2016: 1873504.

[11] 郭昆, 彭胜波, 陈羽中, 等. 基于密度聚类的增量动态社区发现算法[J]. 模式识别与人工智能, 2018, 31(11): 965-978.
GUO K, PENG S B, CHEN Y Z, et al. Incremental dynamic community detection algorithm based on density clustering[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(11): 965-978. (in Chinese)

[12] PARK N, ROSSI R, KOH E, et al. CGC: Contrastive graph clustering for community detection and tracking[C]// Proceedings of the ACM Web Conference 2022. NewYork: ACM, 2022: 1115-1126.

[13] CHAKRABARTI D, KUMAR R, TOMKINS A. Evolutionary clustering[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2006: 554-560.

[14] 潘剑飞, 曹燕, 董一鸿, 等. 基于Attention深度随机森林的社区演化事件预测[J]. 电子学报, 2019, 47(10): 2050-2060.
PAN J F, CAO Y, DONG Y H, et al. The community evolution event prediction based on attention deep random forest[J]. Acta Electronica Sinica, 2019, 47(10): 2050-2060. (in Chinese)

[15] 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.

[16] 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.

[17] BELLO G A, HARENBERG S, AGRAWAL A, et al. Community detection in dynamic attributed graphs[M]// Advanced Data Mining and Applications. Cham: Springer International Publishing, 2016: 329-344.

[18] 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.

[19] APPEL A P, CUNHA R L F, AGGARWAL C C, et al. Temporally evolving community detection and prediction in content-centric networks[M]//Machine Learning and Knowledge Discovery in Databases. Cham: Springer In-

ternational Publishing, 2019: 3-18.
[20] ZHANG Y L, WU B, NING N W, et al. Dynamic topical community detection in social network: A generative model approach[J]. IEEE Access, 2019, 7: 74528-74541.
[21] LI D, ZHANG S, MA X. Dynamic module detection in temporal attributed networks of cancers[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(4): 2219-2230.
[22] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401: 788-791.
[23] DING C, HE X F, 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, PA: Society for Industrial and Applied Mathematics, 2005: 606-610.
[24] 金弟, 刘子扬, 贺瑞芳, 等. 面向带属性复杂网络的鲁棒、强解释性社团发现方法[J]. 计算机学报, 2018, 41(7): 1476-1489.
JIN D, LIU Z Y, HE R F, et al. A robust and strong explanation community detection method for attributed networks[J]. Chinese Journal of Computers, 2018, 41(7): 1476-1489. (in Chinese)
[25] CHOUZENOUX E, FEST J B. SABRINA: A stochastic subspace majorization-minimization algorithm[J]. Journal of Optimization Theory and Applications, 2022, 195(3): 919-952.
[26] 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.
[27] ROSSETTI G. RDYN: Graph benchmark handling community dynamics[J]. Journal of Complex Networks, 2017, 5(6): 893-912.
[28] 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.
[29] CHAKRABORTY T, DALMIA A, MUKHERJEE A, et al. Metrics for community analysis: A survey[J]. ACM Computing Surveys, 2018, 50(4): 1-37.

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.
Email: jung@m.scnu.edu.cn

杨佳琦 女, 2000 年 5 月出生于陕西省咸阳市. 现为华南师范大学计算机学院硕士研究生. 主要研究方向为知识图谱.
E-mail: 2022023277@m.scnu.edu.cn

程 颢 男, 2000 年 2 月出生于江西省上饶市. 现为华南师范大学硕士研究生. 主要研究方向是符号图数据挖掘和社交网络分析.
E-mail: haocheng@m.scnu.edu.cn

汤 庸 男, 1964 年 5 月出生于湖南省张家界市. 现为华南师范大学计算机学院教授, 博士生导师. 主要研究方向为数据智能与云服务、学术社交网络与教育大数据等.
E-mail: ytang@m.scnu.edu.cn