基于动态类簇形成博弈的属性图聚类方法

卜 湛[1)] 王煜尧[2)] 马丽娜[3)] 蒋玖川[1)] 曹 杰[1)]

[1)](南京财经大学江苏省电子商务重点实验室 南京 210023)

[2)](南京理工大学计算机科学与工程学院 南京 210094)

[3)](云境商务智能研究院南京有限公司 南京 210003)

摘 要 以微博、微信为代表的社交网络不仅包含丰富的节点属性信息,还蕴含复杂的网络拓扑信息,这些社交网络通常可被建模为属性图.传统的图聚类方法假设节点属性与网络拓扑共享同一类簇结构.然而,在真实社交网络中,节点属性与网络拓扑所对应的类簇结构并非完全一致.譬如,通过社团发现技术分析新浪微博的好友关注列表能够直观地获取聚集在同一群组的用户集合;而借助文本挖掘技术分析同一群组的用户生成内容却会发现用户讨论话题的分布广泛,体现出差异化的用户偏好特征.如何有效融合属性与拓扑信息对属性图进行聚类是理解、分析和可视化大规模社交网络的关键难题之一.为此,本文将属性图聚类建模为多目标优化问题,提出一种基于动态类簇形成博弈的属性图聚类方法.首先定义一种新颖的中心性指标度量节点的影响力,并提出一种启发式方法初始化属性图类簇质心;其次在动态博弈理论框架下,提出一种贪心的局部搜索策略更新节点类簇标签,并严格证明该局部搜索策略可使类簇结构收敛至局部帕累托最优解;最后设计一种基于多智能体自治计算的属性图聚类算法,该算法无需预设初始类簇个数,且复杂度近似线性于边的数目.为验证本文所提算法的性能,我们依次从三个方面来对其进行测试和评估.首先我们在 Google+属性图上对所提算法进行了单独的收敛性分析.我们测试了算法中四个需要优化的目标函数(*K*-means 损失函数、Havrda-Charvat 生成熵、负模块度和负紧凑度)在三个不同的 Bregman 散度(欧氏距离平方、KL 散度距离和余弦距离)设置下的收敛性情况.实验结果表明,四个目标函数能在 50 轮迭代之后达到收敛状态.然后,我们在 4 个大规模属性图上分别从聚类精度和可扩展性两个方面将本文所提算法与 9 个基准方法作了充分对比.对比结果表明,本文所提算法在 NMI 指标下比其它算法所得最优结果高出 0.7%;而在 AvgF1 指标下比大多数算法所得的最优结果高出 0.2%.在可扩展性方面,本文所提算法即使在最大规模的 Google+属性图上也能在 1 个小时内计算出聚类结果.最后,我们在小规模 PolBK 数据集上进行了可视化分析.从可视化结果可以看出,在 14 轮迭代后本文所提算法就达到了稳定状态,与此同时找到了与真实情况接近的类簇结构.总体实验结果表明,本文方法能够准确发现大规模社交网络潜在的类簇结构,且同已有方法相比具备较好的有效性和高效性.

关键词 属性图聚类;多目标优化;动态类簇形成博弈;局部帕累托最优;自治计算

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2021.01824

Attributed Graph Clustering Approach Based on Dynamic Cluster Formation Game

BU Zhan[1)] WANG Yu-Yao[2)] MA Li-Na[3)] JIANG Jiu-Chuan[2)] CAO Jie[2)]

[1)](*Jiangsu Provincial Key Laboratory of E-Business*, *Nanjing University of Finance and Economics*, *Nanjing* 210023)

[2)](*School of Computer Science and Engineering*, *Nanjing University of Science and Technology*, *Nanjing* 210094)

[3)](*WinGin Business-Intelligence Academy Nanjing Co.*, *Ltd*, *Nanjing* 210003)

Abstract Except for rich node attribute information, there is the complex topological information in some modern online social networks, such as Sina Weibo and WeChat. Such types of social

收稿日期:2020-06-28;在线发布日期:2020-12-21.本课题得到国家重点研发计划(2019YFB1405000)、国家自然科学基金(71871109)、国家自然科学基金重点支持项目(92046206)资助.卜 湛,博士,教授,中国计算机学会(CCF)会员,主要研究领域为社会网络分析、数据挖掘、博弈论.E-mail:zhanbu@nufe.edu.cn.**王煜尧**,博士研究生,主要研究方向为数据挖掘、复杂网络.**马丽娜**,硕士,主要研究方向为商务智能、复杂网络.**蒋玖川**,博士,讲师,主要研究方向为多智能体系统、众包计算.**曹 杰**,博士,教授,主要研究领域为商务智能、推荐系统.

network can usually be represented as an attributed graph. Traditional graph clustering approaches are often based on an assumption that the node attributes and network topology share a same cluster membership. However, it does not always hold in many real-world social networks. Take Sina Weibo as an example, analyzing the follow lists of Weibo users through community detection techniques can directly obtain which users gather into a social group, while these users may produce diverse user-generated content, reflecting differentiated preference characteristics. How to effectively integrate attributive and topological information for clustering attributed graphs becomes a new challenge, which is also critical for understanding, analyzing as well as visualizing large-scale social networks. In this paper, we formulated the target problem as a multi-objective optimization problem, and proposed a dynamic cluster formation game based attributed graph clustering approach. First, we defined a new centrality index, called the influence of nodes, to measure the node influence and designed an effective heuristic method to initialize the cluster centroids of attribute graphs. Second, based on the dynamic game theory, a greedy local search strategy was proposed to update the cluster labels of nodes, and we strictly proved that such local search strategy can make the cluster structure converge to the local Pareto optimality. Third, an autonomy-oriented computing based attributed graph clustering algorithm was proposed, which does not need to specify the cluster number and its running time scales linearly with the total number of edges. Furthermore, we tested and evaluated the proposed approach's performance from three aspects. First, we performed a separate convergence analysis for the proposed approach on the Google+ attributed social network. We tested the convergence of four objective functions (i.e., K-means loss function, Havrada-Charvat generation entropy, negative modularity and negative compactness) that need to be optimized in the approach under three different Bregman divergence settings (i.e., Euclide distance squared, KL divergence distance and cosine distance). The results show that four objective functions can converge after 50 iterations. Then, we compared the proposed approach with 9 baseline methods in terms of accuracy and scalability on 4 large-scale attributed social networks. Experimental results of clustering accuracy showed that the proposed approach is at least 0.7% higher than other algorithms with best performance under NMI metric, and is at least 0.2% higher than most algorithms with best performance under AvgF1 metric. In addition, in terms of the test of scalability, the proposed approach can obtain final results within 1 hour even on the largest Google+ attributed social network. Finally, we performed a visualization analysis on a small PolBK network. The results showed that the proposed approach reached a stable state after 14 rounds of iteration, and the uncovered cluster structure was close to the ground-truth. Overall, extensive experiments shows that the proposed approach can accurately detect the hidden cluster structure in real-world attributed graphs. Compared with the state-of-the-art approaches of clustering nodes in attributed graphs, our approach has better effectiveness and efficiency.

Keywords　attributed graph clustering; multi-objective optimization; dynamic cluster formation game; locally Pareto optimality; autonomy-oriented computing

1　引　言

社交网络通常可被建模为属性图(如图 1 所示)，其中节点代表用户，边代表用户之间的好友关系，不同形状的节点具备不同的属性向量，表征用户差异化的特征属性. 真实属性图具备一些共性的统计特征，其中一个典型的统计特征为类簇结构-同一类簇内的节点属性向量相似且连接紧密，不同类簇间的节点属性向量差异显著且连接稀疏. 图 1 所示属性

图包含 3 个类簇(C_1,C_2 和 C_3). 显然,类簇 C_1 中的节点不仅形状(属性向量)差异显著,而且节点之间的连接较为稀疏;类簇 C_2 中的节点之间虽然连接紧密,但是形状(属性向量)仍存一定差异;类簇 C_3 中的节点不仅形状(属性向量)相近,而且彼此之间的连接较为紧密. 属性图聚类[1] 旨在将具有相似属性向量且连接紧密的节点划分到同一类簇,对分析社交网络拓扑结构、理解社交网络功能并预测演化趋势、制定品牌口碑营销策略、优化广告定向投放等具有重要意义.

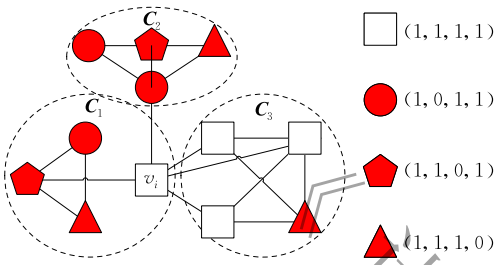


图 1 属性图类簇结构示例

传统的图聚类方法通常假设节点属性与网络拓扑共享同一类簇结构,进而主要考虑网络拓扑信息对属性图进行聚类. 然而在现实的属性图中,节点属性与网络拓扑所对应的类簇结构并非完全一致. 如何有效融合属性和拓扑信息是属性图聚类的关键难题之一,其主要挑战来自以下方面:

挑战 1. 数据异质性. 属性图既包含丰富的节点属性信息,又蕴含复杂的网络拓扑信息,它们是两个彼此独立(或弱相关)的异构数据[2]. 在真实属性图中,这两类异构数据的规模可能极不平衡. 譬如,在新浪微博中,每个用户可最多关联 10 个兴趣标签,而单个用户的好友关注上限可达 3000 个. 尽管近年来一些学者提出了多种基于信息融合的距离函数[1,3] 来度量属性图节点之间的接近性(相似性),但是如何在聚类过程中自适应调整属性和拓扑的影响权重仍是一个开放式问题. 此外,从有效性和高效性两个方面考虑,计算超大规模属性图中任意两个节点的接近性(相似性)也是不现实的.

挑战 2. 准则多样性. 数据异质性所带来的另一个挑战是,在评估属性图聚类结果时会存在多种评价准则. 譬如,在挖掘社交网络意见群组时,群组内部意见一致性和连接紧密性都可被用来作为评价准则. 近年来,一些学者引入概率图模型,通过融合多种评价准则来统一刻画属性图生成过程[4-5]. 然而,这类方法的有效性很大程度上依赖于对样本数据先验概率分布的估计,当属性图所包含异质数据的真实概率分布未知时,基于概率图模型的聚类方法有效性将大打折扣. 为解决这一问题,一些基于多目标优化的属性图聚类方法[6-7] 尝试引入帕累托最优这一概念来寻找多个评价准则的折中方案. 然而,面对超大规模属性图,帕累托最优解很难被发现,需要消耗大量时间在超高维解空间中搜索/比较海量的候选解.

挑战 3. 质心不确性. 经典的基于特征属性的聚类方法 K-means[8],假设类簇质心可以表征整个类簇. 这类方法首先随机初始化 K 个质心,然后采取一种二阶段迭代过程分别更新样本点类簇标签和类簇质心. 由于聚类结果的质量在很大程度上取决于初始质心的选取,经典聚类方法的鲁棒性较差. 为了避免因随机初始化类簇质心而引起的聚类不稳定问题,一些新颖的基于中心/密度的聚类方法被相继提出,譬如:基于亲和力传播的聚类方法[9]、基于密度的聚类方法 DBSCAN[10]、基于密度峰值的聚类方法[11] 等. 面对属性图聚类问题,由于引入了新的数据(网络拓扑),研究属性图类簇质心初始化的成果还相对较少.

综上所述,如何有效融合属性和拓扑信息是属性图聚类的关键;此外,我们还应综合考虑不同评价准则对类簇形成过程进行语义解释. 为解决以上问题,本文将属性图聚类建模为多目标优化问题,从一个新的视角重新审视了属性和拓扑信息在属性图聚类过程中的作用,提出了一种基于动态类簇形成博弈的属性图聚类方法(Dynamic Cluster Formation Game,DCFG). 在 DCFG 中,我们首先融合属性和拓扑信息定义了一种新颖的中心性指标度量每个节点的影响力,然后在此基础上提出了一种启发式方法,通过识别属性图中潜在的领袖节点来初始化类簇质心. 当属性图真实类簇数目 K 已知时,领袖节点识别问题可转化为影响力最大化问题;反之,领袖节点识别问题可转化为经典的顶点覆盖问题. 为了解释属性图类簇形成过程,我们提出了一种新颖的动态类簇形成博弈模型. 该模型遵循经典聚类方法的基础假设——类簇质心可以表征整个类簇. 区别于经典的基于特征属性的聚类方法,我们考虑了网络拓扑对类簇形成过程的影响,在动态博弈理论框架下探讨了节点和类簇的耦合影响,并基于模块度、紧密度、信息熵等类簇质量评价准则定义了节点的可行策略映射函数. 这样,在属性图聚类过程中,节

点的可行策略集将受到类簇结构评价准则的持续约束.最后,我们设计了一种高效的基于多智能体自治计算的属性图聚类算法,算法复杂度近似线性于边的数目.实验结果表明,本文方法能够准确发现大规模属性图的类簇结构;同已有方法相比,具备较好的有效性和高效性.

本文的主要贡献如下:

(1)融合属性和拓扑信息定义了一种新颖的中心性指标度量节点影响力,并提出了一种启发式方法初始化属性图类簇质心.

(2)设计了一种新颖的动态类簇形成博弈模型,并提出了一种贪心的局部搜索策略更新节点类簇标签,且严格证明了该局部搜索策略可使类簇结构收敛至局部帕累托最优解.

(3)提出了一种高效的基于多智能体自治计算的属性图聚类算法,该算法无需预设初始类簇个数,且复杂度近似线性于边的数目;并在4个真实属性图数据集上验证了算法的有效性和高效性.

本文第2节阐述和分析属性图聚类相关工作;第3节对属性图聚类问题进行定义和建模;第4节阐述属性图初始类簇质心选择方法;第5节介绍动态类簇形成博弈模型;第6节介绍基于多智能体自治计算的属性图聚类算法;第7节通过实验分析算法的质量和效率;第8节对全文进行总结和展望.

2　相关工作

国内外学者面对图聚类问题开展了大量的研究工作.本节将从社团发现、属性图聚类以及社团形成博弈等角度出发,对相关工作进行详细阐述.

2.1　复杂网络社团发现方法

传统的图聚类方法通常假设节点属性与网络拓扑共享同一类簇结构,进而主要考虑网络拓扑信息对属性图进行聚类.在复杂网络研究领域,这类方法亦被称作社团发现[12]方法.经典的社团发现方法主要包括图分割方法[13]、凝聚/分裂方法[14]、谱方法[15]、模块度最大化方法[16]、标签传播方法[17]等.上述方法的本质是将社团发现问题转化为基于特定社团结构评价函数的全局优化问题.譬如,Girvan和Newman基于随机图模型提出的模块度指标[14]是最具代表性的社团结构评价函数.在此基础上,学术界提出了很多高效的模块度最大化算法来发现网络中潜在的社团结构.然而,Fortunato和Barthelemy指出,模块度最大化方法存在着分辨率极限问题[18],即难以发现微观尺度的社团结构.为此,一些局部扩展方法[19]首先采用启发式策略挖掘网络中的局部社团,然后通过"组装"局部社团得到网络的全局社团结构.这类方法的关键在于如何合理定义社团质量评价指标.近年来,与网络表示学习相关的社团发现算法引起学术界的广泛关注[20],此类方法旨在将网络节点映射到低维空间,然后借助经典聚类方法将表示向量相似的节点划分到同一类簇.然而,这类方法的不足之处是割裂了节点表示和社团发现两个任务,缺乏对最终社团结构的语义解释.有鉴于此,一些学者尝试同时对节点和社团进行表示学习.譬如,涂存超等学者[21]提出了一种社团增强的网络表示学习算法,通过同时学习节点和社团的表示向量,可较好地完成复杂网络社团发现任务.

2.2　属性图聚类方法

属性图聚类的关键在于如何有效融合两种异质数据——节点属性和网络拓扑信息.一些学者提出了基于信息融合的距离函数[1,3]度量属性图中节点之间的接近性.如何在聚类过程中自适应调整属性和拓扑的影响权重仍是一个开放式问题.吴烨等学者[22]利用信息论中的最小描述长度原理对属性图聚类问题进行建模,提出了一种基于遗传算法的属性图聚类算法.金弟等学者[23]提出了一种矩阵分解框架,采用一个带先验的转移概率矩阵来刻画类簇和属性之间的关联性.此外,一些学者引入概率图模型,通过融合多种评价准则来统一刻画属性图生成过程[4-5,24].譬如,Pfeiffer等学者[24]提出的属性图模型(Attributed Graph Model,AGM)首先学习节点属性的相关性,并利用现有的图生成模型计算节点连边概率;然后结合属性相关性和连边概率对观测网络进行采样,其采样结果使得期望的边概率和度分布同观测网络保持一致.类似的研究工作还包括贝叶斯概率模型[4]和属性图生成模型(Communities from Edge Structure and Node Attributes,CESNA)[5]等.然而,这类方法的有效性很大程度上依赖于对样本数据先验概率分布的预先估计,当属性图所包含异质数据的真实概率分布未知时,基于概率图模型的图聚类方法的有效性将大打折扣.鉴于真实属性图中节点属性的高维稀疏性,一些学者尝试在属性图聚类前,删除一些冗余和无关属性,此类方法被称为子空间聚类.Gunnemann等学者[25]提出的GAMer算法,同时考虑类簇的边密度、规模以及

属性维度等，通过权衡这些质量特征来进行属性图聚类. 随后，他们又提出了一种通用的属性图聚类框架 EDCAR[26]，该方法有效结合了子空间聚类和稠密子图挖掘，可发现真实属性图中语义丰富的类簇. 然而，子空间聚类方法通常需要筛选海量的候选维度子集，具有较高的时间复杂度，难以被直接应用于大规模属性图聚类任务. 近年来，基于网络表示学习[27]的方法展现出强大的能力，一些学者尝试利用其解决属性图聚类问题. 譬如，Zhang 等学者[28]提出了一种用于属性图聚类的自适应图卷积方法(Attributed Graph Clustering，AGC)，该方法利用高阶图卷积来获取全局类簇结构，而且针对不同的属性图数据，可自适应选择合适的阶数. Wang 等学者[29]提出了一种深度注意力嵌入的属性图聚类框架 DAEGC，该方法利用注意力网络获取邻居节点对目标节点的重要度，将属性和拓扑信息编码成表示向量，并训练了一个解码器来重构属性图.

2.3 社团形成博弈模型

为了解释复杂网络中社团结构的形成过程，一些学者将节点看作理性玩家，将他们的决策(如社团标签更新)建模为一种博弈过程，其中单个节点的决策影响其它节点的决策. Torsello 等人在他们的早期工作中提出了一个基于非合作博弈的通用框架，对复杂网络社团发现问题进行了深入的研究[30]. 受这项工作启发，大量基于非合作博弈的社团发现方法被提出[31-32]. 此外，一些学者通过观测真实的网络数据发现，当若干节点履行一致的联盟协议时，社团可以更好地维持内部节点的收益. 因此，一些研究考虑群组玩家的收益，基于合作博弈发现社团结构. Jonnalagadda 等人[33]提出了一种基于合作博弈的社团发现方法，采用多数投票机制来揭示潜在的社团结构. 另外，他们还提出了基于合作博弈的有向网络社团发现算法[34]. Avrachenkov 等人[35]则分别提出了基于 Myerson 价值和 Hedonic 博弈的社团发现算法.

上述博弈模型以全新的视角解释了复杂网络社团形成机制，被广泛应用于解决复杂网络社团发现问题. 但是这些模型忽略了节点属性信息，难以被直接应用于属性图聚类问题. 在前期工作中，我们提出一种基于势博弈优化的图聚类框架(Graph cLustering framework based on potEntial gAme optiMization，GLEAM)[32]. 在该图聚类框架中，每个节点的效用函数由一系列局部线性的收益和损失函数构成，因此 GLEAM 本质上属于经典的势博弈，即存在一个可以反映所有节点收益变化的势函数. 相对于 GLEAM，本文方法的优势体现在以下三方面：(1) DCFG 考虑了节点属性信息对类簇形成过程的影响，且节点效用函数的定义更具一般性，可根据不同的“点-簇”距离函数进行扩展；(2) 除了考虑节点的效用，DCFG 还关注类簇的质量提升，借助贪心的局部搜索策略，我们可快速逼近类簇结构的局部帕累托最优解；(3) 在动态博弈过程中，节点可行策略取决于最近时间周期的类簇标签向量，且策略集规模的上限为节点度+1，本文方法的时间复杂度近似线性于边的数目.

3 问题描述

属性图 $\boldsymbol{G}$ 可表示为三元组 $\boldsymbol{G}=\langle \boldsymbol{V},\boldsymbol{F},\boldsymbol{A}\rangle$，其中 $\boldsymbol{V}=\{v_1,\cdots,v_n\}$ 表示 n 个节点的集合. $\boldsymbol{F}=(f_{i\tau})_{n\times d}$，$\forall f_{i\tau}\geq 0$ 表示非负属性矩阵，该矩阵的第 i 行 $\boldsymbol{f}_i=(f_{i1},\cdots,f_{id})$ 表示节点 v_i 的 d 维属性向量. $\boldsymbol{A}=(a_{ij})_{n\times n}$ 表示二值邻接矩阵，如果节点 v_i 和 v_j 之间存在无向无权边，则 $a_{ij}=1$，否则 $a_{ij}=0$；定义 $\boldsymbol{N}_i=\{v_j|a_{ij}=1\}$ 为节点 v_i 的邻居集合，$k_i=|\boldsymbol{N}_i|$ 表示节点 v_i 的度，则属性图中边的数目为 $m=\frac{1}{2}\sum_{i=1}^{n}k_i$.

属性图聚类旨在将 n 个节点划分到 K 个不相交的类簇，即每个节点 v_i 具有一个唯一的类簇标签 $x_i\in\{1,\cdots,K\}$，这样属性图的类簇结构可表征为 n 个节点类簇标签组成的向量 $\boldsymbol{x}=(x_1,\cdots,x_n)$. 给定类簇标签向量 $\boldsymbol{x}$，类簇 $\boldsymbol{C}_p$ 包含的节点集合为 $\boldsymbol{C}_p=\{v_i|x_i=p\}$，则属性图的节点分组可表示为 $\boldsymbol{P}=\{\boldsymbol{C}_1,\cdots,\boldsymbol{C}_K\}$，当且仅当满足以下条件：

(1) $\forall p\in\{1,\cdots,K\}$，$\boldsymbol{C}_p\subseteq\boldsymbol{V}$ 且 $\boldsymbol{C}_p\neq\varnothing$；

(2) $\bigcup_{p=1}^{K}\boldsymbol{C}_p=\boldsymbol{V}$；

(3) $\forall p,q\in\{1,\cdots,K\}$，$p\neq q$，$\boldsymbol{C}_p\cap\boldsymbol{C}_q=\varnothing$；

(4) 同一类簇中的节点属性向量相似，而不同类簇中节点属性向量相差较大；

(5) 同一类簇中节点连接紧密，而不同类簇中节点连接稀疏.

在属性图聚类过程中，为了满足条件(4)和(5)，本文考虑以下四种评价类簇结构的质量函数：

定义 1. K-means 损失函数[8].

$$Q^1(\boldsymbol{x})=\sum_{p=1}^{K}\sum_{v_i\in\boldsymbol{C}_p}\varpi(\boldsymbol{f}_i,\boldsymbol{c}_p)$$
$$\varpi(\boldsymbol{f}_i,\boldsymbol{c}_p)=\varphi(\boldsymbol{f}_i)-\varphi(\boldsymbol{c}_p)-(\boldsymbol{f}_i-\boldsymbol{c}_p)\otimes\nabla\varphi(\boldsymbol{c}_p) \tag{1}$$

其中，$\boldsymbol{c}_p=(c_{p1},\cdots,c_{pd})=\frac{1}{|\boldsymbol{C}_p|}\sum_{v_i\in\boldsymbol{C}_p}\boldsymbol{f}_i$ 表示类簇 $\boldsymbol{C}_p$ 的质心向量，$\varpi(\cdot,\cdot)$ 为布雷格曼散度（Bregman Divergence）函数，用于衡量属性向量 $\boldsymbol{f}_i$ 和 $\boldsymbol{c}_p$ 之间的差异大小；$\varphi(\cdot)$ 是一个严格凸二次可微函数，$\nabla\varphi(\boldsymbol{c}_p)$ 表示函数 $\varphi(\cdot)$ 在 $\boldsymbol{c}_p$ 处的梯度，$\otimes$ 表示向量的内积操作. 进而我们可以根据不同凸性质的 $\varphi(\cdot)$，衍生出不同形式的散度函数：

（1）欧式距离平方. 令 $\varphi(\boldsymbol{f}_i)=\|\boldsymbol{f}_i\|^2$，则 $\varpi(\boldsymbol{f}_i,\boldsymbol{c}_p)^{\mathrm{SE}}=\|\boldsymbol{f}_i-\boldsymbol{c}_p\|^2$；

（2）KL散度. 令 $\varphi(\boldsymbol{f}_i)=-\sum_{\tau=1}^{d}\left(\frac{f_{i\tau}}{\sum_{\rho=1}^{d}f_{i\rho}}\log\frac{f_{i\tau}}{\sum_{\rho=1}^{d}f_{i\rho}}\right)$，则 $\varpi(\boldsymbol{f}_i,\boldsymbol{c}_p)^{\mathrm{KL}}=\sum_{\tau=1}^{d}\left(f_{i\tau}\log\frac{f_{i\tau}}{c_{p\tau}}\right)-\sum_{\tau=1}^{d}f_{i\tau}+\sum_{\tau=1}^{d}c_{p\tau}$；

（3）余弦距离. 令 $\varphi(\boldsymbol{f}_i)=\|\boldsymbol{f}_i\|$，则 $\varpi(\boldsymbol{f}_i,\boldsymbol{c}_p)^{\mathrm{CD}}=\|\boldsymbol{f}_i\|-\frac{\boldsymbol{f}_i\otimes\boldsymbol{c}_p}{\|\boldsymbol{f}_i\|}$.

定义 2.　Havrda-Charvat 生成熵[36].

$$Q^2(\boldsymbol{x})=\sum_{p=1}^{K}(\boldsymbol{c}_p\otimes(1-\boldsymbol{c}_p)) \tag{2}$$

定义 3.　负模块度[16].

$$Q^3(\boldsymbol{x})=\sum_{p=1}^{K}\left(\left(\frac{\alpha_p}{2m}\right)^2-\frac{\beta_p}{m}\right) \tag{3}$$

其中，$\alpha_p=\sum_{v_i\in\boldsymbol{C}_p}k_i$ 表示类簇 $\boldsymbol{C}_p$ 中所有节点度的总和，$\beta_p=|\{(v_i,v_j)|a_{ij}=1\wedge v_i,v_j\in\boldsymbol{C}_p\}|$ 为类簇 $\boldsymbol{C}_p$ 中的实际边数.

定义 4.　负紧密度[19].

$$Q^4(\boldsymbol{x})=\sum_{p=1}^{K}\left(\gamma_p-\frac{2(n-|\boldsymbol{C}_p|)\beta_p}{|\boldsymbol{C}_p|}\right) \tag{4}$$

其中，$\gamma_p=|\{(v_i,v_j)|a_{ij}=1\wedge v_i\in\boldsymbol{C}_p\wedge v_j\notin\boldsymbol{C}_p\}|$ 表示类簇 $\boldsymbol{C}_p$ 中节点与外部节点的连边数.

基于上述定义，给定 n 个节点的类簇标签向量 $\boldsymbol{x}=(x_1,\cdots,x_n)$，$K$-means 损失函数和 Havrda-Charvat 生成熵越小，表明同一类簇中的节点属性向量相似，而不同类簇中节点属性向量相差较大；负模块度和负紧密度越小，表明同一类簇中节点连接紧密，而不同类簇中节点连接稀疏. 这样，属性图聚类问题可被建模为如下的多目标优化问题：

$$\min_{\boldsymbol{x}}\boldsymbol{Q}(\boldsymbol{x})=(Q^1(\boldsymbol{x}),Q^2(\boldsymbol{x}),Q^3(\boldsymbol{x}),Q^4(\boldsymbol{x})) \tag{5}$$

为了寻找四个评价准则的折中方案，本文试图求解上述多目标优化问题的局部帕累托最优解. 给定解空间中两个任意解 $\hat{\boldsymbol{x}}$ 和 $\check{\boldsymbol{x}}$，多目标优化问题的相关概念回顾如下：

定义 5.　支配关系.

若 $\hat{\boldsymbol{x}}$ 支配 $\check{\boldsymbol{x}}$（记为 $\hat{\boldsymbol{x}}\prec\check{\boldsymbol{x}}$），当且仅当 $\forall l\in[1,4]$，$Q^l(\hat{\boldsymbol{x}})\le Q^l(\check{\boldsymbol{x}})\wedge\exists\kappa\in[1,4]$，$Q^\kappa(\hat{\boldsymbol{x}})<Q^\kappa(\check{\boldsymbol{x}})$.

定义 6.　邻居解.

若 $\check{\boldsymbol{x}}=(\check{x}_1,\cdots,\check{x}_i,\cdots\check{x}_n)$ 是 $\hat{\boldsymbol{x}}=(\hat{x}_1,\cdots,\hat{x}_i,\cdots\hat{x}_n)$ 的邻居解（记为 $\check{\boldsymbol{x}}\in\boldsymbol{\Gamma}(\hat{\boldsymbol{x}})$），当且仅当属性图 $\boldsymbol{G}$ 中存在唯一节点 v_i 满足 $\check{x}_i\neq\hat{x}_i$，而其它节点 $\forall v_j\in\boldsymbol{V}\setminus\{v_i\}$ 满足 $\check{x}_j=\hat{x}_j$.

定义 7.　局部帕累托最优.

若 $\hat{\boldsymbol{x}}$ 满足局部帕累托最优，当且仅当 $\nexists\check{\boldsymbol{x}}\in\boldsymbol{\Gamma}(\hat{\boldsymbol{x}})$，$\check{\boldsymbol{x}}\prec\hat{\boldsymbol{x}}$.

4　类簇质心初始化

经典的基于特征属性的聚类方法，譬如 K-means[8]，假设类簇质心可以表征整个类簇. 这类方法首先随机初始化 K 个质心，然后采取一种二阶段迭代过程分别更新样本点类簇标签和类簇质心. 由于聚类质量在很大程度上取决于初始质心的选取，一些新颖的基于中心/密度的质心初始化方法被相继提出[9-11]以提升传统聚类方法的鲁棒性. 面对属性图聚类问题，引入了新的数据类型（如网络拓扑信息），本节将详细阐述如何融合属性和拓扑信息解决类簇质心初始化问题.

在社交网络中，意见领袖比其他用户更具影响力. 现有社交网络分离动力学研究[37]表明：如果网络中存在意见领袖，则围绕这些意见领袖会更容易形成意见群组. 为此，本文将借鉴社会学理论探究属性图中领袖节点识别问题，进而利用挖掘的领袖节点初始化类簇质心. 首先，我们给出以下假设：

A1：节点间拓扑相似性越大，则相互引力越大；

A2：节点间属性向量越相似，则相互引力越大，且引力会伴随属性向量距离的增加而迅速衰减；

A3：单个节点影响力取决于邻居节点的影响力贡献，且会根据引力为其邻居等比例贡献影响力.

基于上述假设，节点 v_i 的影响力定义为

$$r_i=\sigma\sum_{v_j\in\boldsymbol{N}_i}\left(\frac{\pi_{ij}r_j}{\eta_j}\right)+\frac{1-\sigma}{n} \tag{6}$$

其中，$0<\sigma\leq 1$ 是阻尼系数，$\pi_{ij}=\frac{|\boldsymbol{N}_i\cap\boldsymbol{N}_j|}{|\boldsymbol{N}_i\cup\boldsymbol{N}_j|}\exp(-\varpi(\boldsymbol{f}_i,\boldsymbol{f}_j))$表示节点 v_i 和 v_j 之间的引力，$\eta_j=\sum_{v_i\in\boldsymbol{N}_j}\pi_{ij}$ 表示节点 v_j 的引力和. 所有节点影响力向量 $\boldsymbol{r}=(r_1,\cdots,r_n)$ 可通过求解如下的特征值问题求解：

$$\boldsymbol{r}=\sigma\boldsymbol{H}\boldsymbol{\Phi}^{-1}\boldsymbol{r}+\frac{1-\sigma}{n}\boldsymbol{e},\ \text{s.t.},\ \|\boldsymbol{r}\|=1,\forall r_i\geq 0 \tag{7}$$

其中，$\boldsymbol{H}=(\pi_{ij})_{n\times n}$ 为引力矩阵，$\boldsymbol{\Phi}=(\eta_{ii})_{n\times n}$ 为节点引力和对角矩阵，$\boldsymbol{e}=(1,1,\cdots,1)$ 表示 n 维全 1 行向量. 我们可利用幂率迭代求解上述特征值问题：

$$\boldsymbol{r}(t+1)=\boldsymbol{M}\boldsymbol{r}(t)=\left(\sigma\boldsymbol{H}\boldsymbol{\Phi}^{-1}+\frac{1-\sigma}{n}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}\right)\boldsymbol{r}(t) \tag{8}$$

可以证明 $\boldsymbol{M}=\sigma\boldsymbol{H}\boldsymbol{\Phi}^{-1}+\frac{1-\sigma}{n}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$ 是满足正定、随机、不可约和非周期性质的马尔科夫转移矩阵. 给定任意的初始影响力向量 $\boldsymbol{r}(0)$，上述幂率迭代过程可收敛至唯一的正平稳向量 $\boldsymbol{r}^*=(r_1^*,\cdots,r_n^*)$.

令 $\boldsymbol{L}$ 为领袖节点集合，定义 $\boldsymbol{B}=\bigcup_{v_i\in\boldsymbol{L}}(\boldsymbol{N}_i\cup\{v_i\})$ 为 $\boldsymbol{L}$ 中所有领袖节点的覆盖区域. 当属性图真实类簇数目 K 已知时，领袖节点识别问题可转化为如下的影响力最大化问题：

$$\boldsymbol{L}^*=\underset{\boldsymbol{L}\subseteq\boldsymbol{V},|\boldsymbol{L}|=K}{\arg\max}\left(\sum_{v_j\in\boldsymbol{B}}r_j^*\right) \tag{9}$$

反之，当属性图真实类簇数目 K 未知时，领袖节点识别问题可转化为如下的顶点覆盖问题：

$$\boldsymbol{L}^*=\underset{\boldsymbol{L}\subseteq\boldsymbol{V},|\boldsymbol{B}|=n}{\arg\min}(|\boldsymbol{L}|) \tag{10}$$

令 $\boldsymbol{N}_i^{(2)}=\{v_k\,|\,v_k\neq v_i\wedge v_k\notin\boldsymbol{N}_i\wedge\exists v_j\in\boldsymbol{N}_i,a_{kj}=1\}$ 表示节点 v_i 的二阶邻居集合，$R_i=\sum_{v_j\in\boldsymbol{N}_i}r_j^*+r_i^*$ 表示节点 v_i 覆盖区域的影响力之和. 算法 1 给出了基于贪心思想的领袖节点识别过程. 不难发现：计算每条边上两个节点之间的引力以及每个节点引力和（算法 1 第 1 行）的时间复杂度为 $O(d\bar{k}m)$，其中 $\bar{k}$ 表示属性图的平均度；迭代计算节点影响力（算法 1 第 4 至 7 行）的时间复杂度为 $O(Zm)$，其中 Z 表示迭代次数；计算每个节点覆盖区域的影响力之和（算法 1 第 9 行）的时间复杂度为 $O(m)$，我们利用一个最大堆存储每个节点 R_i 值. 迭代识别领袖节点（算法 1 第 11～14 行）的主要开销来自更新每个领袖节点二阶邻居的 R_k，时间复杂度为 $O(|\boldsymbol{L}|\bar{k}^3+n\log n)$，其中 $O(n\log n)$ 表示最大堆的更新开销. 所以算法 1 的时间复杂度为 $O(d\bar{k}m+Zm+m+|\boldsymbol{L}|\bar{k}^3+n\log n)$. 在真实属性图中，$\log n\approx\bar{k}$，$|\boldsymbol{L}|\approx n/\bar{k}$，同时迭代次数 Z 一般可视为常数. 算法 1 的时间复杂度亦可表示为 $O(d\bar{k}m)$.

算法 1. 领袖节点识别算法（LIA）.

输入：属性矩阵 $\boldsymbol{F}$，邻接矩阵 $\boldsymbol{A}$，类簇数目 K（可选），收敛阈值 ε，阻尼系数 σ

输出：领袖节点集合 $\boldsymbol{L}$

1. $\forall\pi_{ij}\leftarrow\frac{|\boldsymbol{N}_i\cap\boldsymbol{N}_j|}{|\boldsymbol{N}_i\cup\boldsymbol{N}_j|}\exp(-\varpi(\boldsymbol{f}_i,\boldsymbol{f}_j))$；$\forall\eta_i\leftarrow\sum_{v_j\in\boldsymbol{N}_i}\pi_{ij}$；
2. $t\leftarrow 0$；$conv\leftarrow\varepsilon+10$；$\forall r_i\leftarrow 1/n$；
3. WHILE $conv>\varepsilon$ DO
4. $\quad conv\leftarrow 0$；$t\leftarrow t+1$；
5. $\quad\forall r_i\leftarrow\sigma\sum_{v_j\in\boldsymbol{N}_i}\frac{\pi_{ij}r_j}{\eta_j}+\frac{1-\sigma}{n}$；
6. $\quad\forall r_i\leftarrow r_i/\|\boldsymbol{r}\|_1$；
7. $\quad conv\leftarrow conv+\|\boldsymbol{r}(t)-\boldsymbol{r}(t-1)\|_1$；
8. END WHILE
9. $\boldsymbol{L}\leftarrow\varnothing$；$\boldsymbol{B}\leftarrow\varnothing$；$\forall R_i\leftarrow\sum_{v_j\in\boldsymbol{N}_i}r_j^*+r_i^*$；
10. WHILE $|\boldsymbol{L}|<K$（$|\boldsymbol{B}|<n$）DO
11. $\quad v_i\leftarrow\underset{v_i\in\boldsymbol{V}}{\arg\max}(R_i)$；$R_i\leftarrow 0$；$\boldsymbol{B}\leftarrow\boldsymbol{B}\cup\{v_i\}$；
12. $\quad\forall v_j\in\boldsymbol{N}_i$，$R_j\leftarrow 0$；$\boldsymbol{B}\leftarrow\boldsymbol{B}\cup\boldsymbol{N}_i$；
13. $\quad\forall v_k\in\boldsymbol{N}_i^{(2)}$，$R_k\leftarrow R_k-\sum_{v_j\in(\boldsymbol{N}_i\cap\boldsymbol{N}_k)}r_j$；
14. $\quad\boldsymbol{L}\leftarrow\boldsymbol{L}\cup\{v_i\}$；
15. END WHILE

5 动态类簇形成博弈

传统的社团形成博弈[29-31]模型通常假设网络中的节点是完全理性的：每个节点会根据其它节点的策略（社团标签），更新自己的策略以使其自身效用最大化. 这种社团发现方法仅考虑了节点的收益，忽略了社团的收益. 事实上，在真实社交网络中，一些群组（如豆瓣小组）会设置一些管理员，负责对申请入群的用户进行身份审核. 若申请入群的用户满足某些特定条件（如价值观认同），群组会接受该用户的入群申请，否则可能将其拒之门外. 受此启发，本节将综合考虑节点和类簇的收益，详细阐述动态类簇形成博弈的具体设置，以一种新的视角解释属性图中类簇结构的形成过程.

定义 8. 动态类簇形成博弈.

基于单个节点视角，本文的 DCFG 可被定义为一个六元组 $\boldsymbol{\Xi}=\langle t,\boldsymbol{V},\boldsymbol{x}^t,s_i(\cdot),u_i(\cdot,\cdot),\Theta(\cdot)\rangle$，其中：

$t=0,1,2,\cdots$ 表示离散时间周期索引；

$\boldsymbol{V}=\{v_1,\cdots,v_n\}$表示属性图的节点(玩家)集合；
$\boldsymbol{x}^t=(x_1^t,\cdots,x_n^t)$表示 t 周期的类簇标签向量；
$s_i(\cdot)$表示节点 v_i的可行策略映射函数；
$u_i(\cdot,\cdot)$表示节点 v_i的效用函数；
$\Theta(\cdot)$表示类簇结构转移函数.

为了求解式(5)中多目标优化问题的局部帕累托最优解，本文将 Havrda-Charvat 生成熵 $Q^2(\boldsymbol{x})$、负模块度 $Q^3(\boldsymbol{x})$和负紧密度 $Q^4(\boldsymbol{x})$视为次优化目标，进而定义每个节点的可行策略映射函数 $s_i(\cdot)$；同时将 K-means 损失函数 $Q^1(\boldsymbol{x})$视为主优化目标，据此定义每个节点的效用函数 $u_i(\cdot)$；在此基础上，本文设计一种高效的局部搜索策略更新节点类簇标签.

在 DCFG 中，我们为每个类簇设置一系列准入规则以防止因潜在节点的加入而导致类簇质量劣化. 任意节点 v_i的可行策略映射函数定义为

$$s_i(\boldsymbol{x}^t)=\{p\mid \forall\kappa\in\{2,3,4\},g_i^{\kappa}(p,x_i^t,\boldsymbol{x}_{-i}^t)\leq 0\} \tag{11}$$

其中，$p\in\{1,\cdots,K\}$为节点 v_i的候选类簇标签，$g_i^{\kappa}(p,x_i^t,\boldsymbol{x}_{-i}^t)=Q^{\kappa}(p,\boldsymbol{x}_{-i}^t)-Q^{\kappa}(x_i^t,\boldsymbol{x}_{-i}^t)$表示节点 v_i受当前类簇标签向量影响的第 κ 组约束方程. 其中，$\boldsymbol{x}_{-i}^t=(x_1^t,\cdots,x_{i-1}^t,x_{i+1}^t,\cdots,x_n^t)$表示 t 周期除节点 v_i外其它 $n-1$ 个节点的类簇标签向量. 给定当前属性图的类簇标签向量 $\boldsymbol{x}^t=(x_i^t,\boldsymbol{x}_{-i}^t)$，当其它节点类簇标签保持不动时，节点 v_i的类簇标签 x_i^t由 q 更新至 p 使得 $\forall\kappa\in\{2,3,4\},Q^{\kappa}(p,\boldsymbol{x}_{-i}^t)-Q^{\kappa}(q,\boldsymbol{x}_{-i}^t)\leq 0$，则基于不动点理论，式(5)中的子优化目标 $Q^2(\boldsymbol{x})$、$Q^3(\boldsymbol{x})$和 $Q^4(\boldsymbol{x})$将被同时优化. 我们在前期工作[2,19,35]中从启发式搜索角度充分探讨了上述三个目标函数性质，本文将直接给出 $g_i^{\kappa}(p,x_i^t,\boldsymbol{x}_{-i}^t)$的具体定义如下：

$$g_i^2(p,x_i^t,\boldsymbol{x}_{-i}^t)=\begin{cases}0, & p=x_i^t\\ -\dfrac{1}{d}\displaystyle\sum_{\tau=1}^{d}\dfrac{f_{i\tau}\,|\boldsymbol{C}_p^t|\,(2c_{pt}^t-1)}{(|\boldsymbol{C}_p^t|+1)^2}, & p\neq x_i^t\end{cases} \tag{12}$$

$$g_i^3(p,\underset{x_i^t=q}{x_i^t},\boldsymbol{x}_{-i}^t)=\frac{k_i}{m}\left(\frac{\alpha_p^t-\alpha_q^t}{2m}+\frac{|\boldsymbol{N}_i\cap\boldsymbol{C}_q^t|-|\boldsymbol{N}_i\cap\boldsymbol{C}_p^t|}{k_i}\right) \tag{13}$$

$$g_i^4(p,x_i^t,\boldsymbol{x}_{-i}^t)=\begin{cases}0, & p=x_i^t\\ -2n\dfrac{|\boldsymbol{C}_p^t|\,|\boldsymbol{N}_i\cap\boldsymbol{C}_p^t|-\beta_p^t}{|\boldsymbol{C}_p^t|\,(|\boldsymbol{C}_p^t|+1)}+k_i, & p\neq x_i^t\end{cases} \tag{14}$$

其中，$\boldsymbol{C}_p^t=\{v_i\mid x_i^t=p\}$表示 t 周期类簇 $\boldsymbol{C}_p$中的节点集合，$\boldsymbol{c}_p^t=(c_{p1}^t,\cdots,c_{pd}^t)=\dfrac{1}{|\boldsymbol{C}_p^t|}\displaystyle\sum_{v_i\in\boldsymbol{C}_p^t}\boldsymbol{f}_i$表示类簇 $\boldsymbol{C}_p^t$的质心向量，$\alpha_p^t=\displaystyle\sum_{v_i\in\boldsymbol{C}_p^t}k_i$表示 $\boldsymbol{C}_p^t$中节点度之和，$\beta_p^t=|\{(v_i,v_j)\mid a_{ij}=1\wedge v_i,v_j\in\boldsymbol{C}_p^t\}|$表示类簇 $\boldsymbol{C}_p^t$中实际边数，$|\boldsymbol{N}_i\cap\boldsymbol{C}_p^t|$表示隶属于类簇 $\boldsymbol{C}_p^t$的节点 v_i邻居数.

如图 1 所示，由于节点 v_i当前隶属于类簇 $\boldsymbol{C}_1$，则 $x_i^t=1$；根据式(12)～(14)可方便求得：$\forall\kappa\in\{2,3,4\},g_i^{\kappa}(1,x_i^t,\boldsymbol{x}_{-i}^t)=0,g_i^{\kappa}(2,x_i^t,\boldsymbol{x}_{-i}^t)>0,g_i^{\kappa}(3,x_i^t,\boldsymbol{x}_{-i}^t)<0$. 所以，节点 v_i的可行策略集为 $s_i(\boldsymbol{x}^t)=\{1,3\}$.

定理 1.　任意周期 t，任意节点 v_i的可行策略集是非空集，即 $\forall s_i(\boldsymbol{x}^t)\neq\varnothing$且 $\forall x_i^t\in s_i(\boldsymbol{x}^t)$.

证明.　$\forall t,\kappa\in\{2,3,4\},v_i\in\boldsymbol{V},g_i^{\kappa}(x_i^t,x^t)=0$，由式(11)可得，$\forall x_i^t\in s_i(\boldsymbol{x}^t)\Rightarrow\forall s_i(\boldsymbol{x}^t)\neq\varnothing$.　证毕.

定理 2.　任意周期 t，任意节点 v_i若要更新类簇标签 x_i^t，仅能选取其邻居节点的类簇标签，进而，$\forall s_i(\boldsymbol{x}^t)\subseteq\{x_j^t\mid v_j\in\boldsymbol{N}_i\}\cup\{x_i^t\}\Rightarrow\forall|s_i(\boldsymbol{x}^t)|\leq k_i+1$.

证明.　$\forall t,v_i\in\boldsymbol{V},p\in\{1,\cdots,K\}$，若 $p\neq x_i^t$且 $p\notin\{x_j^t\mid v_j\in\boldsymbol{N}_i\}$，则节点 v_i的所有邻居节点都不属于类簇 $\boldsymbol{C}_p^t$，即 $|\boldsymbol{N}_i\cap\boldsymbol{C}_p^t|=0$；根据式(14)可方便求得 $g_i^4(p,x_i^t,\boldsymbol{x}_{-i}^t)=2n\beta_p^t+k_i>0$，进而由式(11)可推导 $p\notin s_i(\boldsymbol{x}^t)$；所以，$\forall|s_i(\boldsymbol{x}^t)|\leq k_i+1$.　证毕.

已知 t 周期的类簇标签向量 $\boldsymbol{x}^t=(x_i^t,\boldsymbol{x}_{-i}^t)$，任意节点 v_i的效用函数定义为

$$u_i(p,\boldsymbol{x}^t)_{p\in s_i(\boldsymbol{x}^t)}=\varpi(\boldsymbol{f}_i,\boldsymbol{c}_p^t) \tag{15}$$

基于当前周期的类簇标签向量 $\boldsymbol{x}^t$，DCFG 将借助于转移函数 $\Theta(\cdot)$来更新每个节点的类簇标签，进而得到下一周期的类簇标签向量 $\boldsymbol{x}^{t+1}$. 具体地，类簇结构转移函数定义为

$$\boldsymbol{x}^{t+1}=\Theta(\boldsymbol{x}^t)=\underset{\forall x_i^{t+1}\in s_i(\boldsymbol{x}^t)}{\arg\min}\left(\sum_{v_i\in\boldsymbol{V}}u_i(x_i^{t+1},\boldsymbol{x}^t)\right) \tag{16}$$

DCFG 将 K-means 损失函数 $Q^1(\boldsymbol{x})$视为主优化目标. 经典的 K-means 聚类方法假设每个样本点可自由访问任意 K 个类簇，进而选择距离最近的类簇加入. 区别于经典聚类方法，本文考虑了网络拓扑结构对类簇结构形成过程的影响：在属性图聚类过程中，每个节点的候选类簇(可行策略)将受到类簇结构评价准则 $Q^2(\boldsymbol{x})$、$Q^3(\boldsymbol{x})$和 $Q^4(\boldsymbol{x})$的持续约束. 基于这种约束机制，最终聚类得到的类簇结构将会平衡不同评价准则的影响.

图 1 中节点 v_i可行策略集为 $s_i(\boldsymbol{x}^t)=\{1,3\}$；其中，类簇 $\boldsymbol{C}_1$的质心向量为(1，0.75，0.75，0.75)，类簇 $\boldsymbol{C}_3$的质心向量为(1，1，1，0.75). 基于欧式距离平

方定义的布雷格曼散度，节点 v_i 在下一周期将会加入类簇 $\boldsymbol{C}_3$.

定义 9. 优化路径.

给定任意的初始类簇标签向量 $\boldsymbol{x}^0$，类簇结构转移函数 $\Theta(\cdot)$ 得到的优化路径定义为如下序列：

$$\gamma^{\Theta(\cdot)}(\boldsymbol{x}^0)=\{\boldsymbol{x}^1,\boldsymbol{x}^2,\cdots\} \tag{17}$$

其中，$\forall t\geq 1,\boldsymbol{x}^t=\Theta(\boldsymbol{x}^{t-1})$. 若 $\gamma^{\Theta(\cdot)}(\boldsymbol{x}^0)$ 是有限序列，令 $\boldsymbol{x}^\psi$ 为该序列最终类簇标签向量，满足 $\boldsymbol{x}^\psi=\Theta(\boldsymbol{x}^\psi)$.

定理 3. 类簇结构转移函数 $\Theta(\cdot)$ 生成的优化路径 $\gamma^{\Theta(\cdot)}(\boldsymbol{x}^0)$ 使得 K-means 损失函数被持续优化，即 $\forall 1\leq t\leq\Psi,Q^1(\boldsymbol{x}^t)\leq Q^1(\boldsymbol{x}^{t-1})$.

证明. $\forall 1\leq t\leq\Psi$，令 $L(\boldsymbol{x}^{t-1},\mathbb{C}^{t-1})=Q^1(\boldsymbol{x}^{t-1})$ 表示 $t-1$ 周期基于类簇标签向量 $\boldsymbol{x}^{t-1}$ 的 K-means 损失函数值，其中 $\mathbb{C}^{t-1}=\{c_1^{t-1},\cdots,c_K^{t-1}\}$ 代表 $t-1$ 周期 K 个类簇质心向量的集合. 当每个节点 v_i 基于式(16)在其候选类簇集 $s_i(\boldsymbol{x}^{t-1})$ 中选择距离最近的类簇加入后，属性图的类簇标签向量将由 $\boldsymbol{x}^{t-1}$ 更新至 $\boldsymbol{x}^t$. 由定理 1 可得 $\forall x_i^{t-1},x_i^t\in s_i(\boldsymbol{x}^{t-1})$，进而 $L(\boldsymbol{x}^t,\mathbb{C}^{t-1})\leq L(\boldsymbol{x}^{t-1},\mathbb{C}^{t-1})$. 接下来，每个类簇将基于类簇标签向量 $\boldsymbol{x}^t$ 更新类簇质心向量，使得 $\forall p\in\{1,\cdots,K\},\boldsymbol{c}_p^t=\frac{1}{|\boldsymbol{C}_p^t|}\sum_{v_i\in\boldsymbol{C}_p^t}\boldsymbol{f}_i$. 对于任意其它的类簇标签 $\forall q\in\{1,\cdots,K\}\wedge q\neq p$，类簇 $\boldsymbol{C}_p^t$ 中节点分别距离 $\boldsymbol{c}_p^t$ 和 $\boldsymbol{c}_q^t$ 布雷格曼散度和的差异为

$$\begin{aligned}\Delta&=\sum_{v_i\in\boldsymbol{C}_p^t}\varpi(\boldsymbol{f}_i,\boldsymbol{c}_p^t)-\sum_{v_i\in\boldsymbol{C}_p^t}\varpi(\boldsymbol{f}_i,\boldsymbol{c}_q^t)\\&=-|\boldsymbol{C}_p^t|(\varphi(\boldsymbol{c}_p^t)-\varphi(\boldsymbol{c}_q^t))-\left(\overbrace{\sum_{v_i\in\boldsymbol{C}_p^t}\boldsymbol{f}_i}^{=|\boldsymbol{C}_p^t|\boldsymbol{c}_p^t}-|\boldsymbol{C}_p^t|\boldsymbol{c}_p^t\right)\otimes\\&\quad\nabla\varphi(\boldsymbol{c}_p^t)+\left(\overbrace{\sum_{v_i\in\boldsymbol{C}_p^t}\boldsymbol{f}_i}^{=|\boldsymbol{C}_p^t|\boldsymbol{c}_p^t}-|\boldsymbol{C}_p^t|\boldsymbol{c}_q^t\right)\otimes\nabla\varphi(\boldsymbol{c}_q^t)\\&=-|\boldsymbol{C}_p^t|(\varphi(\boldsymbol{c}_p^t)-\varphi(\boldsymbol{c}_q^t)-(\boldsymbol{c}_p^t-\boldsymbol{c}_q^t)\otimes\nabla\varphi(\boldsymbol{c}_q^t))\\&=-|\boldsymbol{C}_p^t|\varpi(\boldsymbol{c}_p^t,\boldsymbol{c}_q^t)\leq 0\end{aligned} \tag{18}$$

由式(18)可得，$\boldsymbol{c}_p^t=\arg\min_{c_u^t\in\mathbb{C}^t}\left(\sum_{v_i\in\boldsymbol{C}_p^t}\varpi(\boldsymbol{f}_i,\boldsymbol{c}_u^t)\right)$，所以 $L(\boldsymbol{x}^t,\mathbb{C}^t)\leq L(\boldsymbol{x}^t,\mathbb{C}^{t-1})\Rightarrow Q^1(\boldsymbol{x}^t)\leq Q^1(\boldsymbol{x}^{t-1})$.

证毕.

推论 1. 给定任意的初始类簇标签向量 $\boldsymbol{x}^0$，由类簇结构转移函数 $\Theta(\cdot)$ 生成的优化路径 $\gamma^{\Theta(\cdot)}(\boldsymbol{x}^0)$ 是有限序列，且该序列的最终类簇标签向量 $\boldsymbol{x}^\psi$ 是多目标优化问题(5)的局部帕累托最优解.

证明. 由于 K-means 损失函数的取值是有界的，因此优化路径 $\gamma^{\Theta(\cdot)}(\boldsymbol{x}^0)$ 一定是有限序列. 由式(16)可得，该优化路径的终止类簇标签向量满足：

$$\boldsymbol{x}^\psi=\arg\min\left(\sum_{\forall x_i^\psi\in s_i(\boldsymbol{x}^\psi)}\varpi(\boldsymbol{f}_i,\boldsymbol{c}_p^\psi)_{x_i^\psi=p}\right) \tag{19}$$

不失一般性，假设当前存在唯一节点 v_i 将其类簇标签由 $x_i^\Psi\in s_i(\boldsymbol{x}^\psi)$ 更新至 $y_i^\Psi\notin s_i(\boldsymbol{x}^\psi)$，使得 $Q^1(y_i^\Psi,\boldsymbol{x}_{-i}^\Psi)<Q^1(x_i^\Psi,\boldsymbol{x}_{-i}^\Psi)$，即节点 v_i 选择了不属于可行策略集的类簇标签，使 K-means 损失函数被优化. 由于 $y_i^\Psi\notin s_i(\boldsymbol{x}^\psi)$，所以 $\exists\kappa\in\{2,3,4\},g_i^\kappa(y_i^\Psi,x_i^\Psi,\boldsymbol{x}_{-i}^\Psi)>0$，即在子优化目标 $Q^2(\boldsymbol{x})$、$Q^3(\boldsymbol{x})$ 和 $Q^4(\boldsymbol{x})$ 中，至少存在一个目标函数满足 $Q^\kappa(y_i^\Psi,\boldsymbol{x}_{-i}^\Psi)>Q^\kappa(x_i^\Psi,\boldsymbol{x}_{-i}^\Psi)$. 根据定义 6，$(y_i^\Psi,\boldsymbol{x}_{-i}^\Psi)$ 是 $\boldsymbol{x}^\psi$ 的邻居解，所有 $\nexists(y_i^\Psi,\boldsymbol{x}_{-i}^\Psi)\in\boldsymbol{\Gamma}(\boldsymbol{x}^\psi),(y_i^\Psi,\boldsymbol{x}_{-i}^\Psi)\prec\boldsymbol{x}^\psi$. 根据定义 7，$\boldsymbol{x}^\psi$ 一定是多目标优化问题(5)的局部帕累托最优解.

证毕.

6 属性图聚类算法

基于动态类簇形成博弈模型，本文提出一种基于多智能体自治计算的属性图聚类算法(Attributed Graph Clustering based on Autonomy Oriented Computing，AGC-AOC). 首先，我们将属性图建模为离散时间多智能体系统 $MAS=\langle t,\boldsymbol{V},\boldsymbol{P}\rangle$. 其中，$t=0,1,2,\cdots$ 表示离散时间周期索引，$\boldsymbol{V}=\{v_1,\cdots,v_n\}$ 表示 n 个节点智能体的集合，$\boldsymbol{P}=\{\boldsymbol{C}_1,\cdots,\boldsymbol{C}_K\}$ 表示 K 个类簇智能体的集合. 在该多智能体系统中，每个节点智能体 v_i 的统计特征可由一个四元组 $\langle\boldsymbol{f}_i,\boldsymbol{N}_i,x_i^t,\boldsymbol{s}_i^t\rangle$ 描述，其中，$\boldsymbol{f}_i$ 表示节点智能体 v_i 的 d 维属性向量，$\boldsymbol{N}_i$ 表示节点智能体 v_i 的邻居集合，x_i^t 表示节点智能体 v_i 在 t 周期的类簇标签，$\boldsymbol{s}_i^t$ 表示节点智能体 v_i 在 t 周期的可行策略集合；类似的，系统中的每个类簇智能体 $\boldsymbol{C}_p$ 的统计特征也可由一个三元组 $\langle\boldsymbol{c}_p^t,\alpha_p^t,\beta_p^t\rangle$ 描述，其中，$\boldsymbol{c}_p^t$ 表示类簇智能体 $\boldsymbol{C}_p$ 在 t 周期的质心向量，α_p^t 表示类簇智能体 $\boldsymbol{C}_p$ 在 t 周期的类簇内部节点度之和，β_p^t 表示类簇智能体 $\boldsymbol{C}_p$ 在 t 周期的类簇内部实际边数. 多智能体系统的每个时间周期 t 中，任意节点智能体 v_i 仅可访问其邻居节点智能体的特征信息；每个类簇智能体共享属性图的网络拓扑信息，且可自由访问其内部节点智能体的特征信息.

AGC-AOC 方法的具体实现如算法 2 所示. 利用算法 1 获得属性图的领袖节点集合(算法 2 第 1 行)的时间复杂度为 $O(\bar{d}km)$；在初始化类簇标签向量过程中(算法 2 第 2～8 行)，我们将每个领袖节点

视为单一类簇，并更新相应统计特征，其它节点的初始类簇标签设置为 0，上述过程的时间复杂度为 $O(n+|\boldsymbol{L}|)$；迭代更新节点类簇标签的过程（算法 2 第 9～27 行）分为两个阶段：在第一阶段，每个节点首先计算可行策略集合（算法 2 第 12～16 行），然后从可行策略集合中选择距离最近的类簇更新自己的类簇标签（算法 2 第 17 行）；基于定理 2，每个节点只需考察所有邻居节点的类簇标签是否满足式(11)中的约束条件；具体地，式(12)～(14)所示的约束方程仅与节点和类簇的统计特征相关，因此每次迭代过程中，计算所有节点的可行策略集合的时间复杂度为 $O(dm)$. 在第二阶段，每个类簇基于内部节点的统计特征更新自身的统计特征（算法 2 第 19～22 行），其时间复杂度为 $O(dn+m)$. 由于 $n<m$，所以每轮迭代的时间复杂度近似为 $O(dm+dn+m)=O(dm)$. 上述迭代过程中，当所有节点的类簇标签都未发生改变，即类簇标签向量 $\boldsymbol{x}^{\psi}$ 未被更新，由推论 1 可得，$\boldsymbol{x}^{\psi}$ 是多目标优化问题(5)的局部帕累托最优解，进而整个算法终止迭代. 综上所述，算法 2 的总体时间复杂度为 $O(d\bar{k}m+n+|\boldsymbol{L}|+Tdm)=O(d(\bar{k}+T)m)$，其中 T 表示更新节点类簇标签的迭代次数. 可见，本文的属性图聚类算法时间复杂度独立于类簇数目 $|\boldsymbol{L}|$.

算法 2.　属性图聚类算法(AGC-AOC).

输入：属性矩阵 $\boldsymbol{F}$，邻接矩阵 $\boldsymbol{A}$，类簇数目 K(可选)，收敛阈值 ε，阻尼系数 σ，最大迭代次数 T

输出：类簇标签向量 $\boldsymbol{x}^{\psi}$

1. $\boldsymbol{L}\leftarrow LIA(\boldsymbol{F},\boldsymbol{A},K,\varepsilon,T,\sigma)$；$p\leftarrow 1$；$t\leftarrow 0$；
2. FOR $v_i\in\boldsymbol{L}$ DO
3. $\quad x_i^t\leftarrow p$；$\boldsymbol{C}_p\leftarrow\boldsymbol{C}_p\cup\{v_i\}$；$\langle\boldsymbol{c}_p^t,\alpha_p^t,\beta_p^t\rangle\leftarrow\langle\boldsymbol{f}_i,k_i,0\rangle$；
4. $\quad\boldsymbol{P}\leftarrow\boldsymbol{P}\cup\boldsymbol{C}_p$；$p\leftarrow p+1$；
5. END FOR
6. FOR $v_i\in\boldsymbol{V}-\boldsymbol{L}$ DO
7. $\quad\langle\boldsymbol{f}_i,\boldsymbol{N}_i,x_i^t,\boldsymbol{s}_i^t\rangle\leftarrow\langle\boldsymbol{f}_i,\boldsymbol{N}_i,0,\varnothing\rangle$；
8. END FOR
9. WHILE $t<T$ DO
10. $\quad$FOR $v_i\in\boldsymbol{V}$ DO
11. $\quad\quad\boldsymbol{s}_i^t\leftarrow\varnothing$；
12. $\quad\quad$FOR $p\in(\{x_j^t\mid v_j\in\boldsymbol{N}_i\}\cup\{x_i^t\})$ DO
13. $\quad\quad\quad$IF $p\neq 0$ AND $\forall g_i^{\kappa}(p,x_i^t,\boldsymbol{x}^t{}_{-i})\leq 0$ DO
14. $\quad\quad\quad\quad\boldsymbol{s}_i^t\leftarrow\boldsymbol{s}_i^t\cup\{p\}$；
15. $\quad\quad\quad$END IF
16. $\quad\quad$END FOR
17. $\quad\quad x_i^t=p\leftarrow\underset{q\in\boldsymbol{s}_i^t}{\arg\min}(\varpi(\boldsymbol{f}_i,\boldsymbol{c}_q^t))$；$\boldsymbol{C}_p\leftarrow\boldsymbol{C}_p\cup\{v_i\}$；
18. $\quad$END FOR
19. $\quad$FOR $\boldsymbol{C}_p\in\boldsymbol{P}$ DO
20. $\quad\quad\langle\boldsymbol{c}_p^t,\alpha_p^t,\beta_p^t\rangle\leftarrow\left\langle\sum_{v_i\in\boldsymbol{C}_p}\boldsymbol{f}_i\Big/|\boldsymbol{C}_p|,\sum_{v_i\in\boldsymbol{C}_p}k_i,\frac{1}{2}\sum_{v_i\in\boldsymbol{C}_p}|\boldsymbol{C}_p\cap\boldsymbol{N}_i|\right\rangle$；
21. $\quad\quad\boldsymbol{C}_p\leftarrow\varnothing$
22. $\quad$END FOR
23. $\quad t\leftarrow t+1$；
24. $\quad$IF $\boldsymbol{x}^t==\boldsymbol{x}^{t-1}$ DO
25. $\quad\quad$BREAK；
26. $\quad$END IF
27. END WHILE

7　实　验

本节将通过综合实验来测试各个算法在四个真实网络数据集上的性能表现. 所有算法将部署在一台配有 Linux 操作系统的服务器上执行，其配置为：主频 2.6 GHz 的 4 核 E5-2650v2 处理器，128 GB 内存，600 GB 的 SAS 硬盘和 240 GB 的固态硬盘. 下面，我们将依次阐述实验设置、收敛性分析、聚类性能对比、功能模块分析以及聚类过程可视化.

7.1　实验设置

实验所用四个数据集来源于斯坦福网络分析平台(Stanford Network Analysis Platform，SNAP)，分别采集于电子商务平台 Amazon 和三个社交平台(Facebook，Twitter 和 Google+). 我们删除了节点数少于 3 个的真实类簇，数据集统计信息见表 1.

表 1　实验数据集统计信息

数据集	节点数	边数	属性维度	真实类簇数目	类簇标签标注率/%
PolBK	105	441	3	3	100
FaceBK	4039	84243	89	193	100
Twitter	76245	1242397	33208	3170	29
Gplus	120100	12113501	610	438	23

PolBK 中的节点表示电子商务平台 Amazon 上出售的政治书籍或杂志，边表示两本书经常被买家同时购买. 每本书根据其内容被标注为“保守”、“自由”或“中立”. FaceBK 来源于在社交平台 Facebook 上收集到的调查问卷. 其中每一个节点表示社交用户，边表示用户间的好友关系. 该数据集中的每位用户包含 26 类不同的个人资料信息，如家庭住址、同事和政治立场等. Twitter 抽取自社交平台 Twitter 上的 1000 个自 ego 网络，其中每位用户的属性信息涉及该用户两周内发表内容的标签或主题. Gplus 收

集自社交平台 Google+，其中每个节点包含 6 类不同的个人资料信息，分别为性别、姓氏、职称、机构、大学和居住地.

除了本文提出的基于多智能体自治计算的属性图聚类算法 AGC-AOC，其它对比算法包括：(1) 基于随机游走模型和加权模块度优化的社团发现算法 Infomap[38]；(2) 基于社团聚类系数优化的社团抽取算法 SCD[39]；(3) 基于势博弈优化的社团发现算法 GLEAM[32]；(4) 基于结构/属性相似性的属性图聚类算法 SA-Cluster[1]；(5) 基于概率图模型的属性图聚类算法 CESNA[5]；(6) 基于子空间聚类的属性图聚类算法 EDCAR[26]；(7) 基于网络嵌入的算法 Node2Vec[40]；(8) 基于图卷积神经网络的属性图聚类算法 AGC-GCN[28]；(9) 基于深度注意力机制的属性图聚类算法 DAEGC[29]. 为了公平比较，本节所有实验均采用作者提供的源代码和默认参数. 对于网络嵌入方法 Node2Vec[40]，我们首先利用节点属性向量对节点嵌入向量初始化，再借助基于密度的聚类方法 DBSCAN[10] 对训练后的节点嵌入向量进行聚类，且设置聚类数目为属性图的真实类簇数. 对于深度学习方法 AGC-GCN[28] 和 DAEGC[29]，它们的聚类结果对于参数较为敏感，通常需花费较长时间调参才能获得理想的聚类效果；因此，实验中我们采用作者在原始论文中给出的实验参数设置.

考虑到所有实验数据集中真实类簇结构已知，本文使用评价 F1 分数(AvgF1)[41] 和标准化互信息(Normalized Mutual Information，NMI)[41] 作为属性图聚类结果的评价标准(AvgF1 和 NMI 是常用的度量两个类簇结构相似度的评价指标，它们的取值范围从 0 到 1，其值越大表示聚类效果越好).

7.2 收敛性分析

我们首先选取规模最大的 Gplus 数据集分析属性图聚类算法(AGC-AOC)的收敛性. 实验中，我们假设真实类簇数目未知，此时算法 1 第 10～15 行的迭代过程采用 $|\boldsymbol{B}|<n$ 的判定条件探测领袖节点，当遍历完所有节点后($|\boldsymbol{B}|=n$)，AGC-AOC 将自动识别属性图的类簇数目. 收敛阈值 ε 设置为 10^{-5}，阻尼系数 σ 设置为 0.8，最大迭代次数 T 设置为 50. 若某轮迭代中类簇标签向量未发生改变，则算法提前终止. 基于欧氏距离平方、KL 散度距离和余弦距离的变种算法分别表示为 AGC-AOC-SE、AGC-AOC-KL 和 AGC-AOC-CD. 图 2(a)～(d)分别展示了不同变种算法每轮迭代更新节点类簇标签过程中 K-means 损失函数、Havrda-Charvat 生成熵、负模块度和负紧密度的变化趋势. 其中，x 轴表示迭代轮次 t，y 轴表示每轮迭代后类簇结构相应的评价指标. 如图 2 所示，随着迭代轮次的增加，无论采用何种形式的布雷格曼散度，AGC-AOC 发现的类簇结构在上述评价指标上具有相似的收敛趋势：经过近 10 轮迭代，K-means 损失函数和 Havrda-Charvat 生成熵将收敛至相对稳定的值；而在负模块度和负紧密度两个评价指标上，AGC-AOC 需要近 20 轮迭代才能收敛至较稳定的值. 对比不同的变种算法，当采用余弦距离且达到设置的最大迭代次数 50 时，结果虽有轻微浮动，但基本已达稳定状态；而采用欧氏距离平方和 KL 散度距离时，AGC-AOC 分别迭代了 37 和 39 次找到了局部帕累托最优解 $\boldsymbol{x}^{\psi}$，进而提前终止了迭代. 这也说明，AGC-AOC 采用欧氏距离平方和 KL 散度距离具备较好的收敛性. 对比不同类簇结构评价指标的变化趋势，AGC-AOC-SE 发现的局部帕累托最优解具备最小的 K-means 损失函数和负紧密度，AGC-AOC-KL 发现的局部帕累托最优解具备最小的负模块度，AGC-AOC-CD 最终发现的类簇结构具备最小的 Havrda-Charvat 生成熵. 这也表明，采用欧氏距离平方的 AGC-AOC 算法可以更好地利用属性和拓扑信息，在不同类簇结构评价准则上取得更好平衡.

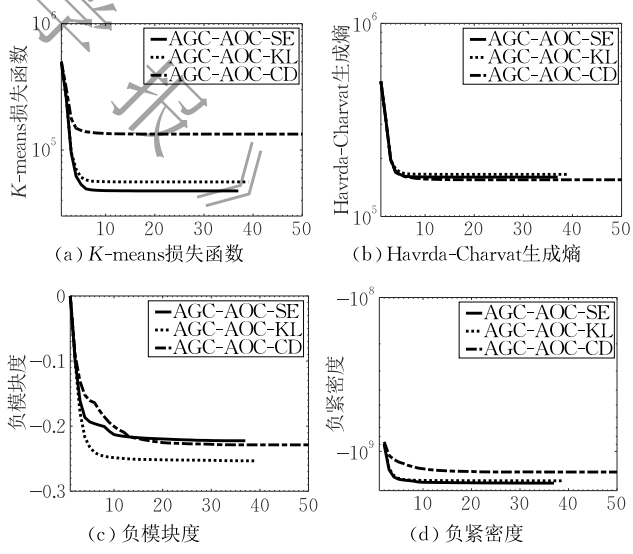


图 2 AGC-AOC 算法在 Gplus 数据集上的收敛性分析((a)表示算法中 K-means 损失函数选取三种不同距离下的收敛情况；(b)表示算法中 Havrda-Charvat 生成熵选取三种不同距离下的收敛情况；(c)表示算法中负模块度选取三种不同距离下的收敛情况；(d)表示算法中负紧凑度选取三种不同距离下的收敛情况)

在后续实验中，AGC-AOC 算法假设真实类簇数目未知，采用欧氏距离平方，收敛阈值 ε 设置为 10^{-5}，阻尼系数 σ 设置为 0.8，最大迭代次数 T 设置

为 50. 当属性图类簇数 K 已知时，算法 1 第 10～15 行的迭代过程采用 $|\boldsymbol{L}|<K$ 的判定条件探测领袖节点，当识别完给定数量的领袖节点后，算法 1 提前终止. 我们将这种变种算法标记为 AGC-AOC-K.

7.3　聚类性能对比

为了验证不同算法的聚类效果，我们使用 AvgF1[41] 和 NMI[41] 来量化分析不同算法所检测出的类簇结构. 实验过程中，针对基于网络嵌入和深度学习的属性图聚类算法，我们重复运行算法 30 次，取平均聚类效果. 不同算法的对比结果如表 2 和表 3 所示. 表中的数值均为百分比数值，每列数据集中最好结果用粗体标记出来. 不难发现，采用 AvgF1 度量，AGC-AOC-K 算法在 PolBK 数据集上具有最好的聚类效果；AGC-AOC 算法在 Twitter 和 Gplus 数据集上表现最好，基于深度学习的 AGC-GCN 算法在 FaceBK 数据集上获得了最优的聚类效果. 相比其它仅考虑网络拓扑结构的社团发现方法，我们的方法在 AvgF1 度量指标上具有显著优势. 传统的基于概率图模型的 CESNA 算法在 Twitter 和 Gplus 数据集取得了较为理想的聚类效果. 在 NMI 度量指标上，AGC-AOC-K 算法在 PolBK 数据集上获得最优聚类效果；在 FaceBK、Twitter 和 Gplus 数据集上，AGC-AOC 算法表现最优. 上述实验表明本文提出的基于动态类簇形成博弈的属性图聚类方法可较准确地捕捉真实属性图的类簇结构，且相比现有的图聚类方法，具备较高的识别精度.

表 2　不同算法的聚类性能对比（采用 AvgF1 度量）

精度/% 算法	类型	数据集			
		PolBK	FaceBK	Twitter	Gplus
Infomap	拓扑	62.3	40.5	15.1	10.7
SCD	拓扑	37.9	19.7	13.4	3.9
GLEAM	拓扑	72.1	31.2	19.2	14.4
SA-Cluster	结合	57.9	20.1	13.4	7.8
CESNA	结合	44.9	41.3	20.9	24.2
EDCAR	结合	35.9	25.1	18.5	10.5
Node2Vec	结合	65.6	35.4	15.8	13.2
AGC-GCN	结合	70.5	**44.3**	23.1	25.3
DAEGC	结合	73.5	41.6	22.8	24.9
AGC-AOC-K	结合	**77.5**	43.6	29.6	26.9
AGC-AOC	结合	51.5	44.1	**30.5**	**27.1**

表 3　不同算法的聚类性能对比（采用 NMI 度量）

精度/% 算法	类型	数据集			
		PolBK	FaceBK	Twitter	Gplus
Infomap	拓扑	50.8	69.5	60.0	38.6
SCD	拓扑	30.1	36.2	56.2	14.4
GLEAM	拓扑	54.2	52.0	66.5	49.4
SA-Cluster	结合	43.0	34.2	47.3	32.5
CESNA	结合	33.9	70.5	67.5	53.2
EDCAR	结合	27.1	40.7	63.2	41.6
Node2Vec	结合	40.5	65.6	67.8	45.6
AGC-GCN	结合	55.8	71.2	72.6	56.2
DAEGC	结合	54.4	71.0	72.4	54.6
AGC-AOC-K	结合	**56.8**	69.1	74.3	57.7
AGC-AOC	结合	45.4	**71.7**	**74.5**	**57.9**

在 Twitter 和 Gplus 数据集中，人工标记类簇标签的节点数量分别占节点总数的 29% 和 23%. AGC-AOC 算法将领袖节点识别问题转化为顶点覆盖问题. 在遍历完属性图所有节点后，AGC-AOC 算法识别的领袖节点更具代表性，所以最终聚类结果较预设类簇数目的 AGC-AOC-K 算法更为准确.

我们进一步以运行时间为单位来验证 AGC-AOC 算法的效率. 表 4 给出了实验结果，其中，每列数据集中最快的结果用粗体标记出来. 可以看出，在所有数据集中基于单一网络拓扑数据的算法的运行效率整体都要高于结合属性和拓扑信息的算法. 尤其是基于势博弈优化的社团发现算法 GLEAM 在 PolBK、FaceBK 和 Gplus 上具有最高的运行效率. 在同时考虑属性和拓扑信息的聚类算法中，本文所提算法的运行效率要显著优于其它方法. 尤其是在规模较小的数据集 PolBK 和 FaceBK 上，我们的方法和仅考虑网络拓扑信息的社团发现算法的运行时间同属一个数量级；在规模较大的数据集 Twitter 和 Gplus 上，相比其它属性图聚类算法，本文方法的运行时间要快近一个数量级，具有显著的效率优势. 基于网络嵌入或深度学习的属性图聚类算法尽管可以获得较理想的聚类效果，但是需要花费大量时间学习节点的表示向量，所以算法的总体运行效率要劣于我们的方法. 本文提出的 AGC-AOC-K 和 AGC-AOC 算法可识别不同数量的领袖节点（类簇）. 由表 4 所示，两个算法在四个数据集上的运行时间基本一致，这进一步验证了我们方法的时间复杂度独立于类簇数目.

表 4　不同算法的运行时间对比

时间/s 算法	类型	数据集			
		PolBK	FaceBK	Twitter	Gplus
Infomap	拓扑	**0.002**	1.000	11.12	95.54
SCD	拓扑	0.232	0.529	**4.688**	67.53
GLEAM	拓扑	**0.002**	**0.207**	5.547	**19.92**
SA-Cluster	结合	15.59	599.8	11287	15148
CESNA	结合	16.03	65.45	2065	15720
EDCAR	结合	26.03	587.4	4985	21594
Node2Vec	结合	1.456	30.23	1928	8992
AGC-GCN	结合	2.332	40.85	2496	15893
DAEGC	结合	4.324	145.3	3583	23242
AGC-AOC-K	结合	0.026	0.728	294.9	2987
AGC-AOC	结合	0.029	0.874	275.2	2865

7.4 类簇质心初始化模块分析

本文提出的属性图聚类方法需借助领袖节点识别算法 LIA 初始化类簇质心. LIA 算法首先基于幂率迭代过程计算节点影响力,然后将领袖节点识别问题转化为影响力最大化或顶点覆盖问题,并基于一种高效的贪心搜索策略识别属性图的领袖节点. 接下来,首先在四个真实数据集上对 LIA 算法进行收敛性分析. 实验中,收敛阈值 ε 设置为 10^{-5},阻尼系数 σ 设置为 0.8,节点间的引力函数是基于欧式距离平方的布雷格曼散度形式. 图 3 展示了 LIA 算法迭代计算节点影响力的收敛性分析. 其中,x 轴表示迭代轮次 t,y 轴表示前后两轮影响力向量差异 $\|\boldsymbol{r}(t)-\boldsymbol{r}(t-1)\|_1$. 正如我们看到的,在四个真实数据集上,我们的 LIA 算法具有相似的收敛趋势,且数据集规模越小,收敛速度越快,这表明 LIA 算法具备良好的收敛性能. 在规模较大的数据集 Twitter 和 Gplus 上,经过约 30 轮幂率迭代,节点影响力向量的前后差异降至 10^{-5} 以下,单个节点影响力更新前后的差异均值 $\frac{1}{n}\sum_{v_i\in\mathbf{V}}\|\boldsymbol{r}(32)-\boldsymbol{r}(31)\|_1$ 小于 10^{-10}. 这也说明了经过约 30 轮幂率迭代,Twitter 和 Gplus 数据集中所有节点的影响力值趋于稳定.

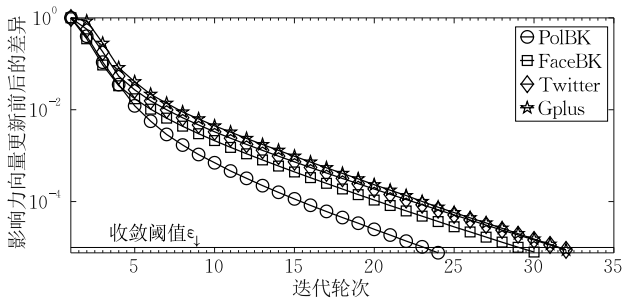


图 3 LIA 算法在四个真实属性图数据集上的收敛性分析

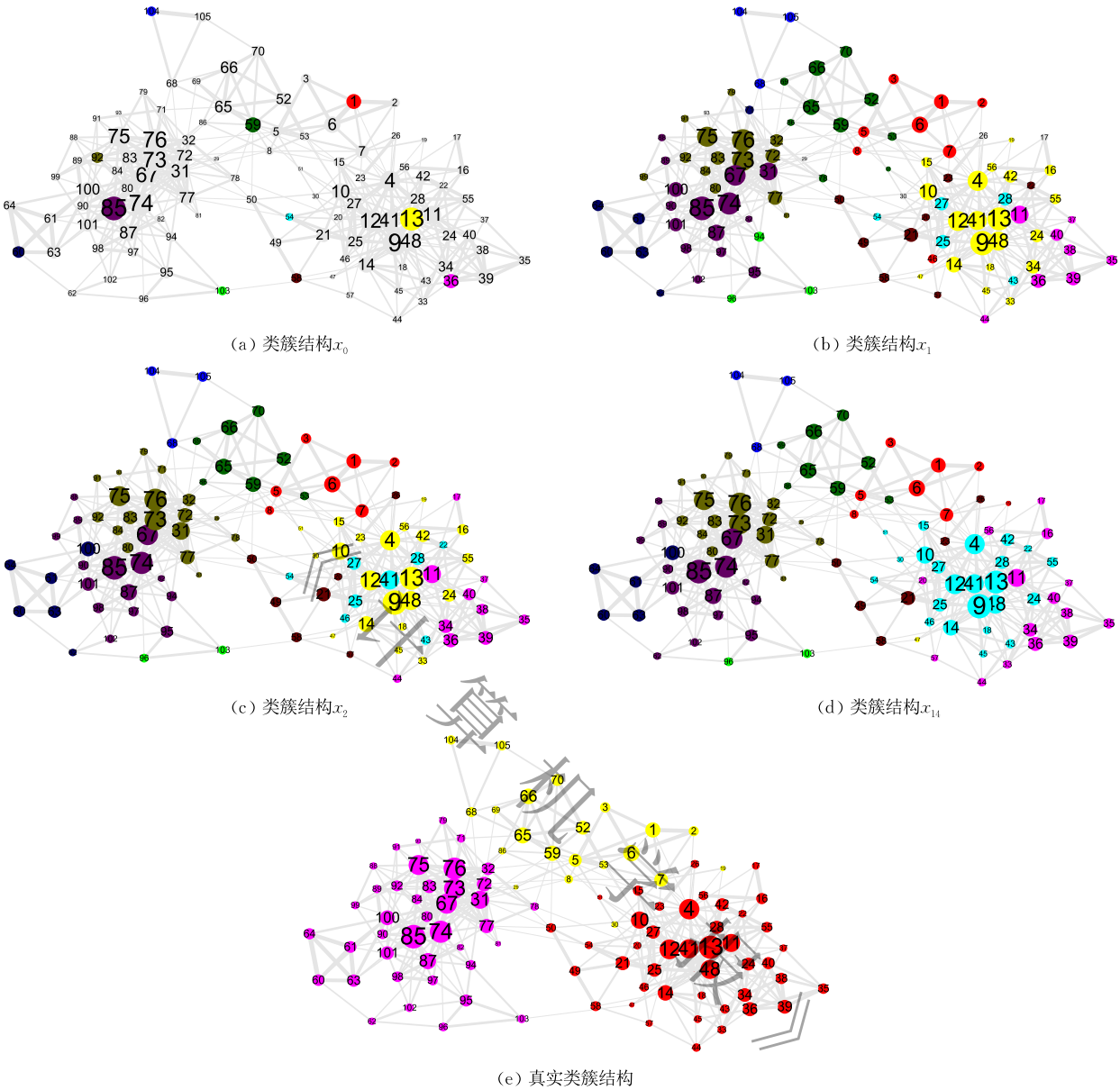
接下来,我们考虑 LIA 算法的四个变种:(1)基于随机初始化的领袖节点识别方法 LIA-Random-K(随机初始化 K 个类簇质心);(2)基于启发式遍历的领袖节点识别方法 LIA-Heuristic-K[2](首先对节点按照影响力降序排列,然后访问该有序列表中第一个未被访问的节点,视该节点为领袖节点,并将其所有邻居节点标记为已访问状态,迭代上述过程直至找到 K 个领袖节点);(3)自适应学习类簇数量的启发式遍历方法 LIA-Heuristic-Auto(重复 LIA-Heuristic-K 的迭代过程,直至遍历完属性图的所有节点,进而找到最佳类簇数量 K^*);(4)自适应学习类簇数量的领袖节点识别算法 LIA-Auto(重复 LIA-K 的迭代过程,直至遍历完属性图的所有节点,进而找到最佳类簇数量 K^*). 我们将 AGC-AOC 算法的第 1 行分别替换为上述四种领袖节点识别方法,采用 AvgF1 度量指标分析不同变种算法的聚类效果,对比结果如表 5 所示,其中 LIA-Random-K 算法取重复运行 30 次的平均聚类效果. 不难发现,采取随机初始化的 LIA-Random-K 算法在所有数据集上表现最为糟糕. 因此,类簇质心初始化对属性图聚类效果存在较大影响. 当属性图真实类簇数量已知时,相比于 LIA-Heuristic-K 算法,本文所提的 LIA-K 算法具备较强的竞争性. 当属性图真实类簇数量未知时,基于自适应类簇数量学习机制的 LIA-Heuristic-Auto 和 LIA-Auto 算法在遍历完整属性图后发现的领袖节点更具代表性,所以最终的聚类结果更为准确.

表 5 领袖节点识别算法的聚类性能分析(采用 AvgF1 度量)

精度/% 算法	数据集 PolBK	FaceBK	Twitter	Gplus
LIA-Random-K	45.0	33.9	19.9	20.3
LIA-Heuristic-K	56.9	39.5	22.9	25.2
LIA-K	**77.5**	43.6	29.6	26.9
LIA-Heuristic-Auto	59.1	41.7	24.8	25.2
LIA-Auto	51.5	**44.1**	**30.5**	**27.1**

7.5 PolBK 聚类过程可视化

最后,我们选择规模最小的 PolBK 数据集,可视化 AGC-AOC 算法的聚类过程. 其结果如图 4 所示,节点的直径刻画其影响力,节点的颜色表征其当前类簇标签,边的厚度反应相邻节点间的引力. 执行领袖节点识别算法后,我们得到 PolBK 数据集的 11 个领袖节点. 如图 4(a)所示,这些领袖节点彼此相隔较远,且在局部范围具有较大的影响力. 接下来,每个节点分别计算相应的可行策略空间,并选择可行策略空间中距离自己最近的类簇加入. 如图 4(b)所示,经过一轮迭代,近 95% 的节点更新了类簇标签. 重复上述过程,AGC-AOC 算法在 PolBK 数据集累计迭代了 14 轮并提前终止. 如图 4(d)所示,每个节点类簇标签在动态类簇形成博弈机制下趋于稳定,找到了满足局部帕累托最优的类簇结构. 观察 PolBK 数据集真实类簇结构(如图 4(e)所示),不难发现,AGC-AOC 算法可识别更小规模的类簇,具有较强的发现高分辨率类簇结构的能力.



(a) 类簇结构x_0　(b) 类簇结构x_1

(c) 类簇结构x_2　(d) 类簇结构x_{14}

(e) 真实类簇结构

图 4　AGC-AOC 算法在 PolBK 数据集上的聚类过程((a)表示领袖节点识别后的网络类簇结构；(b)表示算法第 1 轮迭代后的网络类簇结构；(c)表示算法第 2 轮迭代后的网络类簇结构；(d)表示算法第 14 轮迭代后的网络类簇结构；(e)表示 PolBK 网络真实的类簇结构)

8　总结与展望

本文提出了一种基于动态类簇形成博弈的属性图聚类方法. 首先，定义了一种新颖的中心性指标度量属性图中节点的影响力，在此基础上提出了一种启发式类簇质心初始化算法；然后，考虑节点属性和网络拓扑对类簇结构形成过程的影响，提出了一种贪心的局部搜索策略更新节点类簇标签，并严格证明该局部搜索策略可使类簇结构收敛至局部帕累托最优解；最后，设计了一种基于多智能体自治计算的属性图聚类算法，该算法无需预设初始类簇个数，且复杂度近似线性于边的数目. 我们在真实属性图数据集上验证了所提方法的有效性和高效性，同现有的主流图聚类方法相比，本文方法在聚类效果和执行时间上具备较强的竞争力. 然而本文所提出的属性图聚类方法目前仅适用于硬聚类(非重叠类簇发现)任务，如何改进本文方法以解决属性图的模糊聚类(重叠类簇发现)问题是我们下一步工作中的努力方向. 此外，真实属性图的节点属性和网络拓扑是动态演化的，如何扩展本文方法来发现动态属性图中的类簇结构，仍存在诸多有待解决的挑战，我们将在未来工作中加以改进.

参 考 文 献

[1] Zhou Y, Cheng H, Yu J X. Graph clustering based on structural/attribute similarities. Proceedings of the VLDB Endowment, 2009, 2(1): 718-729

[2] Bu Z, Li H, Zhang C, et al. Graph *K*-means based on leader identification, dynamic game, and opinion dynamics. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(7): 1384-1361

[3] Cheng H, Zhou Y, Yu J. Clustering large attributed graphs: A balance between structural and attribute similarities. ACM Transactions on Knowledge Discovery from Data, 2011, 5(2): 190-205

[4] Xu Z, Ke Y, Wang Y, et al. A model-based approach to attributed graph clustering//Proceedings of the ACM International Conference on Management of Data. Scottsdale, USA, 2012: 505-516

[5] Yang J, Mcauley J, Leskovec J. Community detection in networks with node attributes//Proceedings of the IEEE International Conference on Data Mining. Atlantic City, USA, 2013: 1151-1156

[6] Folino F, Pizzuti C. An evolutionary multiobjective approach for community discovery in dynamic networks. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1838-1852

[7] Li Z T, Liu J, Wu K. A multiobjective evolutionary algorithm based on structural and attribute similarities for community detection in attributed networks. IEEE Transactions on Cybernetics, 2018, 48(7): 1963-1976

[8] MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967: 281-297

[9] Frey B, Dueck D. Clustering by passing messages between data points. Science, 2007, 315(5814): 972-976

[10] Ester M, Kriegel H P, Sander J, Xu X W. A density-based algorithm for discovering clusters in large spatial databases with noise//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Portland, USA, 1996: 226-231

[11] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014, 344(6191): 1492-1496

[12] Fortunato S. Community detection in graphs. Physics Reports, 2010, 486(3-5): 75-174

[13] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs. The Bell System Technical Journal, 1970, 49(2): 291-307

[14] Newman M E J. Fast algorithm for detecting community structure in networks. Physical Review E, 2004, 69(6): 066133

[15] Donetti L, Munoz M A. Detecting network communities: A new systematic and efficient algorithm. Journal of Statistical Mechanics: Theory and Experiment, 2004, 10: P10012

[16] Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826

[17] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007, 76(3): 036106

[18] Fortunato S, Barthelemy M. Resolution limit in community detection. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41

[19] Bu Z, Wu Z A, Cao J, Jiang Y C. Local community mining on distributed and dynamic networks from a multiagent perspective. IEEE Transactions on Cybernetics, 2016, 46(4): 986-999

[20] Wang X, Cui P, Wang J, et al. Community preserving network embedding//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 203-209

[21] Tu C C, Zeng X K, Wang H, et al. A unified framework for community detection and network representation learning. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(6): 1051-1065

[22] Wu Ye, Zhong Zhi-Nong, Xiong Wei, et al. An efficient method for attributed graph clustering. Chinese Journal of Computers, 2013, 36(8): 1704-1713(in Chinese)
(吴烨，钟志农，熊伟等. 一种高效的属性图聚类方法. 计算机学报，2013, 36(8): 1704-1713)

[23] Jin Di, Liu Zi-Yang, He Rui-Fang, et al. A robust and strong explanation community detection method for attributed networks. Chinese Journal of Computers, 2018, 41(7): 1476-1489(in Chinese)
(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. 计算机学报，2018, 41(7): 1476-1489)

[24] Pfeiffer J J, Moreno S, Fond T L, et al. Attributed graph models: Modeling network structure with correlated attributes//Proceedings of the International World Wide Web Conference. Seoul, Korea, 2014: 831-842

[25] Gunnemann S, Farber I, Boden B, Seidl T. Subspace clustering meets dense subgraph mining: A synthesis of two paradigms//Proceedings of the IEEE International Conference on Data Mining. Sydney, Australia, 2010: 845-850

[26] Gunnemann S, Boden B, Farber I, Seidl T. Efficient mining of combined subspace and subgraph clusters in graphs with feature vectors//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Gold Coast, Australia, 2013: 261-275

[27] Wang Rui-Guo, Ye Ya-Ling, Bu Zhan. A community detection approach based on network embedding. Journal of Liaocheng University (Natural Science Edition), 2019, 1(1): 69-78(in Chinese)

(王瑞国，叶雅玲，卜湛. 一种基于网络嵌入的社区发现方法. 聊城大学学报(自然科学版)，2019，1(1)：69-78)

[28] Zhang X T，Liu H，Li Q M，Wu X M. Attributed graph clustering via adaptive graph convolution//Proceedings of the International Joint Conference on Artificial Intelligence. Macao，China，2019：4327-4333

[29] Wang C，Pan S，Hu R，et al. Attributed graph clustering：A deep attentional embedding approach//Proceedings of the International Joint Conference on Artificial Intelligence. Macao，China，2019：3670-3676

[30] Torsello A，Rota B S，Pelillo M. Grouping with asymmetric affinities：A game-theoretic perspective//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York，USA，2006：292-299

[31] Chen W，Liu Z M，Sun X R，Wang Y J. A game-theoretic framework to identify overlapping communities in social networks. Data Mining and Knowledge Discovery，2010，21(2)：224-240

[32] Bu Z，Cao J，Li H J，et al. GLEAM：A graph clustering framework based on potential game optimization for large-scale social networks. Knowledge and Information Systems，2018，55(3)：741-770

[33] Jonnalagadda A，Kuppusamy L. Overlapping community detection in social networks using coalitional games. Knowledge and Information Systems，2018，56(11)：637-661

[34] Jonnalagadda A，Kuppusamy L. Mining communities in directed networks：A game theoretic approach//Proceedings of the International Conference on Intelligent Systems Design and Applications. Delhi，India，2017：826-835

[35] Avrachenkov K E，Kondratev A Y，Mazalov V V，Rubanov D J. Network partitioning algorithms as cooperative games. Computational Social Networks，2018，5(11)：1-28

[36] Bu Z，Gao G L，Li H J，Cao J. CAMAS：A cluster-aware multiagent system for attributed graph clustering. Information Fusion，2017，37：10-21

[37] Henry A D，Prałat P，Zhang C Q. Emergence of segregation in evolving social networks. Proceedings of the National Academy of Sciences，2011，108(21)：8605-8610

[38] Rosvall M，Bergstrom C T. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences，2008，105(4)：1118-1123

[39] Prat-Pérez A，Dominguez-Sal D，Larriba-Pey J L. High quality，scalable and parallel community detection for large real graphs//Proceedings of the International World Wide Web Conference. Seoul，Korea，2014：225-236

[40] Grover A，Leskovec J. Node2vec：Scalable feature learning for networks//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. San Francisco，USA，2016：855-864

[41] Yang J，Leskovec J. Overlapping community detection at scale：A nonnegative matrix factorization approach//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome，Italy，2013：587-596

BU Zhan，Ph. D.，professor. His research interests include social network analysis，data mining and game theory.

WANG Yu-Yao，Ph. D. candidate. His research interests include data mining and complex network.

MA Li-Na，M. S. Her research interests include business intelligence and complex network.

JIANG Jiu-Chuan，Ph. D.，lecturer. His research interests include multiagent system and crowdsourcing computing.

CAO Jie，Ph. D.，professor. His research interests include business intelligence and recommendation system.

Background

Except for rich node attribute information，in some modern social networks，there exists the complex network topology information. Such types of social networks can usually be represented as attributed graphs. Traditional graph clustering approaches assume that the node attributes and network topology share the same cluster memberships. However，this assumption does not always hold in many real-world social networks. How to effectively integrate attributive and topological information for clustering attributed graphs becomes a new challenge，which is also critical for understanding，analyzing as well as visualizing large-scale social networks. In order to better reconcile the data with two different modalities，i. e.，node attributes and network topology，Yang et al. proposed CESNA—A generative probabilistic approach modeling the interaction between network topology and node attributes. Although CESNA performs well in many real-world attributed graphs，the choice of the priori distributions in the statistical models requires a non-trivial expertise. To understand the formation mechanism of real clusters with great individual diversity. In

2018, we presented a Graph cLustering framework based on potEntial gAme optimization (GLEAM). It first utilizes the cosine similarity to weight each edge in the original network. Then, an initial partition, including a number of clusters dominated by those potential leader nodes, is created by a fast heuristic process. Third, a potential game-based weighted Modularity optimization is used to improve the initial partition.

However, GLEAM still has the following problems. (1) It only considers the network topological structure but ignores the heterogeneous-nodes' attributive information. (2) Due to the introduction of new data (e. g., objects' attributes), how to adaptively determine the leader nodes in attributed graphs has not been addressed extensively. (3) Multiple proximity measures corresponding to different criteria may be required to detect certain types of cluster, how to find the trade-off solution among multiple clustering criteria is still missing. To solve the above issues, we proposed a dynamic cluster formation game based attributed graph clustering approach. First, we defined a new centrality index to measure the node influence and designed a heuristic method to initialize the cluster centroids of attribute graphs. Second, based on the dynamic game theory, a greedy local search strategy was proposed to update the cluster labels of nodes, and we strictly proved that such local search strategy can make the cluster structure converge to the local Pareto optimality. Third, an autonomy-oriented computing based attributed graph clustering algorithm was proposed, which does not need to specify the cluster number and its running time scales linearly with the total number of edges. Extensive experiments showed that the proposed approach can accurately detect the hidden cluster structure in real-world attributed graphs. Compared with the state-of-the-art graph clustering approaches, our approach has better effectiveness and efficiency.

This work was supported by the National Key R&D Program of China (No. 2019YFB1405000) and the National Natural Science Foundation of China (Grant No. 71871109). The first is about "The Research on Distributed Fusion Mechanism of Multi-Source Situational Data" and the second is about "The Research on the Formation and Evolution Mechanism of Consumer Groups in Social Networks".