文章编号:1008-1542(2021)03-0231-10

异质网络社区发现方法研究综述

周万珍,宋 健,许云峰

(河北科技大学信息科学与工程学院,河北石家庄 050018)

摘 要:社区结构是复杂网络研究中的重要领域,也是复杂网络的重要特征之一,发现网络中的社区结构在理解网络功能方面起着重要作用。通过对国内外异质网络社区发现文献进行深入研究,较为全面地对现有异质网络社区发现算法进行了归纳总结。首先,通过对国内外异质网络社区发现文献进行归纳,给出异质网络社区发现的基本概述,明确异质网络社区发现领域相关问题的基本定义。其次,介绍了异质网络社区发现算法及主要评价指标,利用不同网络结构以及算法对现有方法进行分类概述。最后,对异质网络社区发现算法的发展趋势进行了总结与展望,提出未来可以将研究重点集中在以下几个方面:1)探索基于异质网络的社区发现评价标准,以推动该领域的快速发展;2)设计更加通用的算法模型,解决由先验知识引起的未知社区数量问题;3)开展更多关于动态网络的研究。

关键词:计算机神经网络;复杂网络;社区发现;异质网络;图神经网络

中图分类号:TP311.13 **文献标识码**:A **doi**:10.7535/hbkd.2021yx03004

Survey of community discovery method of heterogeneous network

ZHOU Wanzhen, SONG Jian, XU Yunfeng

(School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China)

Abstract: Community structure is an important field in the research of complex networks, and it is also one of the important characteristics of complex networks. It is found that the community structure in the network plays an important role in understanding network functions. Through in-depth research on the literature of heterogeneous network community discovery at home and abroad, a more comprehensive summary of the existing heterogeneous network community discovery algorithm is carried out. Firstly, by summarizing the literature of heterogeneous network community discovery at home and abroad, the basic overview of heterogeneous network community discovery is given, and the basic definition of related issues in the field of heterogeneous network community discovery is defined. Then, it introduces the heterogeneous network community discovery algorithm and the main evaluation index, and classifies the existing methods by using different network structures and algo-

收稿日期:2020-11-26;修回日期:2020-12-31;责任编辑:张士莹

基金项目:中国留学基金委地方合作项目(201808130283);教育部人工智能协同育人项目(201801003011);河北科技大学校立课题(82/1182108);河北科技大学雾霾与空气污染防治科研项目(82/1182169)

第一作者简介:周万珍(1966—),男,河北张家口人,教授,博士,主要从事计算机网络及数据库应用技术方面的研究。

通讯作者:许云峰副教授。E-mail:hbkd_xyf@hebust.edu.cn

周万珍,宋健,许云峰.异质网络社区发现方法研究综述[J].河北科技大学学报,2021,42(3):231-240.

ZHOU Wanzhen, SONG Jian, XU Yunfeng.Survey of community discovery method of heterogeneous network[J].Journal of Hebei University of Science and Technology,2021,42(3):231-240.

rithms. Finally, the development trend of heterogeneous network community discovery algorithms is summarized and prospected, and the future research focus can be focused on the following aspects: 1) Exploring the evaluation criteria of community discovery based on heterogeneous networks to promote the rapid development of this field Development; 2) Design a more general algorithm model to solve the problem of the number of unknown communities caused by prior knowledge; 3) Carry out more research on dynamic networks.

Keywords: computer neural network; complex network; community discovery; heterogeneous network; graph neural network

现实世界中实体与实体之间的联系都可以被表示为复杂网络,比如人与人之间形成的社交网络、作者之间形成的引文网络、生物医学领域的蛋白质相互作用网络、化学分子网络等。对于复杂网络,钱学森先生给出了严格定义,将具有自组织、自相似、吸引子、小世界[1]、无标度特性[2]的网络称为复杂网络。

研究复杂网络的性质有助于人们对其加深理解和认识。例如,研究复杂网络中的路径长度、节点密度、节点的连通性等结构信息,有助于了解实体之间的关联性质;研究复杂网络中实体的属性信息,有助于了解实体的固有性质。社区结构是复杂网络中的另一个重要性质,从连通性和节点密度的角度出发,社区被称为节点的局部密集连通子图或聚类[3]。简单而普遍的理解是,社区由2个规则组成,一是社区中的节点紧密相连,二是不同社区的节点稀疏连接[3]。现实世界中,同一个社区中的节点可以共享公共属性或起类似的作用,这些共性几乎是所有社区发现策略的关键。社区发现,或者更特别的说法——基于相似特征或者特征集的聚类,可以帮助人们理解复杂网络的固有属性和功能。例如,在蛋白质-蛋白质相互作用(PPI)网络中的社区发现,揭示了具有相似功能的蛋白质[4];在Twitter和微博等社交网络中,具有共同兴趣的用户或者共同朋友的用户更有可能是同一个社区的成员[5];在引文网络中,社区发现可以确定研究主题的重要性和发展方向[6]。

根据结构,复杂网络可以分为同质网络和异质网络。早期,社区发现的相关研究多集中于同质网络,有许多不错的算法被提出。例如:2002年,GIRVAN等[3]提出了具有开创性的GN算法;2004年,NEWMAN等[7]提出了基于模块度优化的FastGN算法,CLAUSET等[8]提出了CNM算法;2007年,RAGHAVAN等[9]提出了LPA算法;2008年,ROSVALL等[10]提出了Infomap算法,BLONDEL等[11]提出了Louvain算法;2016年,XU等[12]提出了基于骨干度的重叠社区发现算法等。

异质网络具有高维度、多节点等特点,传统的同质网络社区发现方法无法应用于异质网络,使得异质网络社区发现研究成为一个新的挑战。作为一个逐渐兴起的研究领域,异质网络社区发现仍处于探索阶段,目前的研究方法还不是很成熟。现有的异质网络社区发现方法可以分为传统异质网络社区发现方法和基于深度学习的异质网络聚类方法2种。传统的异质网络社区发现方法大多针对特定网络结构,不能对一般拓扑结构的大规模异质网络进行社区发现。基于深度学习的异质网络聚类方法大多先采用网络表示学习方法对节点进行表示,然后进行特定的聚类任务。

目前,已有许多学者针对同质网络社区发现发表了一些综述性文献[13-14],但关于异质网络社区发现的综述较少。本文整理了当前比较热门的传统异质网络社区发现方法和新兴的基于深度学习的异质网络聚类方法,根据算法的适用性及原理对现有方法进行归纳总结,并对异质网络社区发现的未来进行展望。

1 异质网络社区概述

1.1 信息网络

信息网络[15](information networks)由有向图$G=(V,E)$所定义,V表示节点集,E表示边集,该有向图中包含对象类型映射函数$\tau:V\rightarrow A$和连接类型映射函数$\varphi:E\rightarrow R$,其中每一个对象$v\in V$都属于一个特定的对象类型$\tau(v)\in A$,每一个连接$e\in E$都属于一个特定的关系类型$\varphi(e)\in R$。当节点类型数量为1,即$|A|=1$,同时,边的类型数量也为1,即$|R|=1$时,称这个信息网络为同质网络。反之,当节点的类型数量大于1,即$|A|>1$,或者,边的类型数量大于1,即$|R|>1$时,则称这个信息网络为异质信息网络[16](以下简称异质网络)。具有4种节点类型的异质信息网络如图1所示[17]。

异质网络中也存在特殊情况,如多路网络(multiplex networks)、多模网络(multi-mode network)和多

维度网络(multi-dimensional network)。其中：多路网络即节点类型为 1、而关系类型为 N 的特殊异质网络，通常是针对异构网络中某种指定类型的中心节点而言的；多模网络则侧重于节点，指网络中存在不同类型的节点；多维网络侧重于边，指网络中存在的不同关系类型的边。

通常信息网络中使用元路径[15](meta path)对网络中不同的路径进行定义。元路径 Φ 是定义在模式 $S=(A,R)$ 上的一条路径，以 $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$ 的形式表示，其中 $R=R_1 \circ R_2 \circ \cdots \circ R_l$ 是定义在对象 $A_1, A_2, \cdots, A_{l+1}$ 中的一种复合关系，$\circ$ 表示关系上的组合运算符。

不同的元路径可以表达异质网络中不同的语义信息，根据这些语义信息可以将网络划分为不同的社区结构。图 2 中，“Author”既可以由元路径“Author-Paper-Author”(APA)连接，也可以由元路径“Author-Paper-Venue-Paper-Author”(APVPA)连接[18]。不同的元路径所包含的语义信息是完全不同的，比如元路径 APA 表示 2 位作者共同合作完成一篇文章，而元路径 APVPA 则表示 2 位作者在同一机构完成不同的文章。

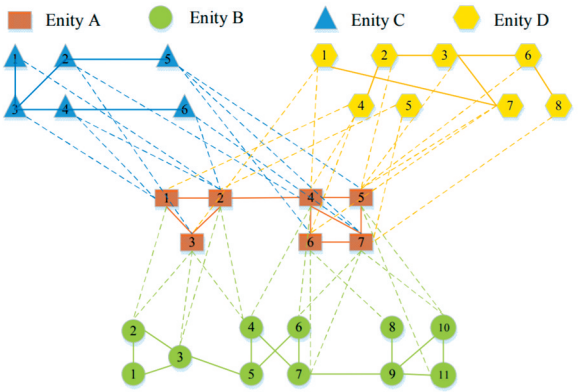


图 1　具有 4 种节点类型的异质信息网络

Fig. 1　Heterogeneous information network with 4 node types

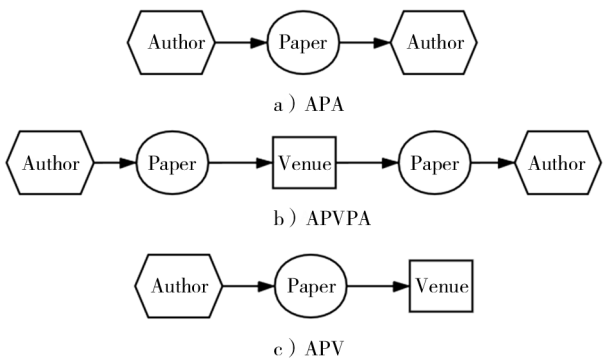


图 2　不同元路径所代表的不同含义

Fig. 2　Different meanings represented by different meta-paths

1.2　社区

所谓社区[13,19]，从连通性和密度的角度来看，社区被称为节点的局部连通子图或聚类，社区中的节点通常有着同样的特征或起着相同的作用。根据图论[20]，社区也被定义为集团(每个节点彼此相邻)或连接的组件(每对节点至少通过一条路径连接)。图 3 示出了 Zachary 空手道俱乐部网络，其中，不同颜色的符号表示在俱乐部的讲师(节点 1)和总裁(节点 34)之间存在分歧之后产生的 2 个社区[14]。

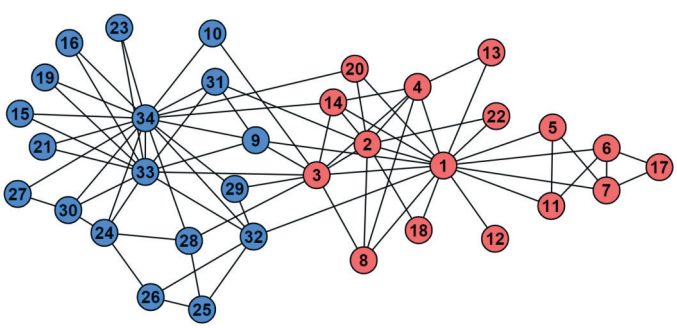


图 3　Zachary 空手道俱乐部网络

Fig. 3　Network of Zachary karate clubs

1.3　社区发现

社区发现(也被称为图聚类技术)是基于网络拓扑结构信息识别出的具有相似特征或起相同作用的节点的集合。在过去几十年中，社区发现受到了学者们的广泛关注，并取得了很多开创性的研究成果。例如：2002 年，GIRVAN 等[3]提出了 GN 算法，掀起了社区发现的研究热潮，引起了计算机、物理、生物等众多学科领域的关注；2004 年，NEWMAN[21]提出了模块度 Q 函数作为社区发现的评价指标，推动了社区发现领域的相关研究快速向前发展。但是，以上研究是基于非重叠社区结构的，即一个节点只属于一个社区(见图 4a))。而在现实生活中，更为常见的是重叠社区结构(见图 4b))，即一个节点属于多个社区。例如：生物

a）非重叠社区结构 b）重叠社区结构

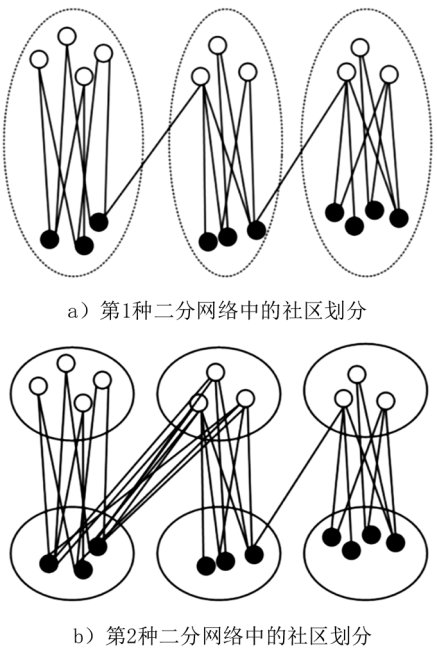
图4 2种社区结构

Fig. 4 Two community structures

信息学科学者发表的文章通常涉及生物以及信息2个领域；一名铁人三项运动员既属于跑步俱乐部，又属于自行车俱乐部。

相比于同质网络，现实生活中更为常见的是异质网络，即多类型节点和多类型关系组成的网络。异质网络不仅包括多种节点类型，还包括各种不同节点类型之间的复杂关系。例如：引文网络中的作者、论文、作者单位等节点之间的关系，作者与论文之间是写作关系，作者与作者之间是合作关系，作者与作者单位之间是隶属关系，论文与论文之间是引用关系等。异质网络中的社区结构一般被定义为同种类型节点组成的社区，社区间形成多对多关系[22]。

由于异质网络的复杂性，导致仅有结构最为简单的二分网络得到了一定的研究。二分网络仅由2种节点类型组成，不同类型的节点之间有较多的边连接，而同类型之间大多没有边。由于异质网络具有多分多维的特性，因而二分网络有2种形式的社区结构定义。第1种是基于同质网络的社区定义，将不同类型的节点划分到同一个社区，社区内不同类型节点间紧密连接，社区间形成一对一关系，如图5a)所示；第2种是基于社区内相似节点聚类来考虑的，将二分网络中同类型节点划分到一个社区，社区内的节点间没有连接，异质社区间形成多对多关系，如图5b)所示[22]。



a）第1种二分网络中的社区划分

b）第2种二分网络中的社区划分

图5 二分网络的社区划分

Fig. 5 Divided network community division

1.4 社区发现的实际应用

社区发现可以被应用到许多任务中，宽泛地分为2类：链路预测和社交影响力最大化。

1.4.1 链路预测

链路预测是数据挖掘领域的重要研究方向，是指通过已知节点的特征信息以及网络拓扑结构，预测尚未产生连接的节点对之间出现连边的可能性[23]。链路预测可以识别具有相似购买历史的用户组，进行更加有效的推荐。

1.4.2 社交影响力最大化

影响力最大化的应用场景十分丰富，包括病毒营销、信息扩散、推荐系统等。影响力最大化的目的是在特定的网络传播模型下找到一组节点，使得这组节点的最终影响力达到最大化。在社交网络中，由于社区结构的特性，社区内部节点通常具有相似的特征，因此信息在社区内部的传播速度较快。因此，一些科研人员尝试将社区发现与影响力最大化进行研究，取得了不错的效果。

2 异质网络社区发现方法

研究人员对同质网络提出了许多优秀的社区发现算法，例如谱方法[24]和标签传播方法[25]。但是这些算法并没有考虑网络的异质性，检测到的社区仅仅反映连接密度关系，无法反映异质网络中不同连接类型之间的语义关系。早期，受限于异质网络的复杂性与计算机的计算能力，传统异质网络社区发现方法多集中于结构较为简单的二分网络和多分多维网络。随着计算能力的提高以及深度学习的兴起，基于深度学习的异质网络聚类方法逐渐成为研究的主要方向。

2.1　传统异质网络社区发现方法

2.1.1　二分网络的异质网络社区发现方法

目前，二分网络社区发现主要有 2 种方法：一是将二分网络投影降维为同质网络，再使用同质网络社区发现算法对其进行社区划分，但该方法在投影降维过程中容易造成部分信息丢失；二是直接在二分网络上进行社区划分。

2007 年，BARBER[26]开创性地提出了二分模块度，并基于此提出了一种基于二分模块度的 Brim 算法，克服了同质网络模块度在异质网络中效果不佳的缺陷，但该方法无法有效识别出网络中的小规模社区。针对这一问题，XU 等[27]提出了基于密度的二分网络模块度，从网络连接密度出发，同时考虑二分网络社区内连接数量和节点数量，提升对小规模社区的识别效果。

2.1.2　多分多维的异质网络社区发现方法

2009 年，SUN 等[28-29]提出了基于概率模型的 RankClus 算法以及 NetClus 算法，采用排名和社区发现相互提升的思路，将排名问题和社区发现问题结合在一起。通过迭代“根据当前社区划分计算节点的等级分布”和“计算节点所属社区的后验概率，以及当前等级分布调整节点的社区划分”来进行社区发现，直到收敛为止。但这 2 种方法均是针对某一特定的网络结构类型提出的，RankClus 针对的是双类型异质网络，NetClus 针对的是包含多种实体类型的异构星形网络。为了解决这一问题，MING 等[30]提出了 RankClass 算法，该算法在 NetClus 的基础上进行改进，使其可以应用于具有常规拓扑结构的异质网络。具体地说，RankClass 通过建立一个基于图的排名模型，根据排名结果对网络结构进行调整，使得排名高的对象组成的子网络的重要性得以加强，最后通过估计各个对象属于每个类的后验概率来确定每个对象的最佳类别。QIU 等[31]注意到上述方法未考虑重叠社区问题，因此提出了 OcdRank 算法，该算法在有向异质的社交网络中将重叠社区发现和社区成员排名结合在一起，通过区分不同类型的节点来改善 RankClus，同时该方法支持增量更新的社区发现，这在现实场景中具有很大的应用价值。以上基于概率模型的异质网络社区发现算法具有较高的时间和空间复杂度，且由于需要丰富的先验知识以及预先设定社区数量，导致在大规模异质网络中社区发现的效果并不佳。

2011 年，COMAR 等[32]提出了基于非负矩阵分解的社区发现算法，使用邻接矩阵表示异质网络，通过将邻接矩阵分解成潜在因子进行社区发现。但是该算法只能用于 $A=2$，$R=3$ 的异质网络。针对上述算法只能用于特定网络结构这一问题，LI 等[33]提出了一个基于正则化和非负矩阵分解（regularized joint non-negative matrix factorization，简称 RJNMF）的框架，该框架在构建异质网络邻接矩阵时综合考虑链路信息和节点内容信息，引入调节功能以减少噪声数据对社区的影响，以此提高社区发现的效果。CHEN 等[34]将同质网络中基于矩阵分解的社区发现算法 Hom-SC 和 Hom-RSC 扩展至异质信息网络中，分别命名为 Het-SC 和 Het-RSC。具体地说，通过对给定图的拉普拉斯矩阵进行矩阵分解获得其特征向量，然后对不同节点类型对应的特征向量进行 K-means 聚类，获得社区划分结果。基于矩阵分解的算法可以有效识别异质网络的异质性，但是针对大规模网络进行矩阵分解具有较高的时间、空间复杂度，同时由于需要在进行社区划分之前对社区数量进行设定，因此对先验知识要求较高。

2016 年，LIU 等[35]提出了以种子为中心的社区发现算法 DenSeC，该算法根据节点密度对中心节点进行识别，然后将中心节点从紧密连接区域扩展到稀疏连接区域，并不断吸收新节点生成最终的社区划分。2018 年，LU 等[17]提出了用于多维社区发现的 Hete_MESE 算法，该算法可以从多维度检测异质网络中的重叠和异质社区，具有线性时间复杂度，因此可以应用于大规模异质网络。具体来说，首先将异质网络中的多个实体类型之一指定为社区中心节点类型，基于中心节点以及元路径提取多路网络，然后基于多路网络检测重叠的社区，并将其视为种子社区，吸收其他实体类型生成异质社区。

基于以上研究可以发现，传统异质网络社区发现方法遵循由特殊网络结构到一般拓扑结构、由非重叠社区发现到重叠社区发现这一发展路线；同时可以发现，随着网络规模的增大，传统算法近年的研究目标主要集中于如何降低时间、空间复杂度以及减少对先验知识的依赖方面。

2.2　基于深度学习的异质网络聚类方法

近年来，基于深度学习的网络表示学习方法是数据挖掘领域研究的热门方向[36]，该方法可以将信息网络表示为低维稠密的携带网络节点特征信息的实数向量，并应用于下游任务的输入，如聚类、社区发现等。TOMMASEL 等[37]的研究表明，仅关注网络拓扑结构的传统方法忽略了节点间的语义信息，而深度学习技

术可以同时考虑网络拓扑结构与网络上的语义信息。已有研究表明,深度学习技术可以提高社区发现的准确性[38]。例如,CHEN等[39]提出的线图神经网络模型(line graph neural network,简称LGNN),通过基于深度学习的聚类方法解决社区发现问题,取得了较为显著的效果。由此可见,基于深度学习的聚类方法可以帮助研究人员进行社区发现。

由于异质网络多节点类型、关系类型的特性,异质图神经网络通常会采用层次聚合的方式获得节点的表示信息:1)节点级别,获得节点在某种关系(即某条元路径)下的邻居节点并聚合邻居节点信息,获得节点在该关系下的表示;2)语义级别,对多种关系(即不同元路径)下的节点表示进行融合,获得最终的节点表示。

异质网络表示学习中,首先要解决的问题是不同节点类型以及不同关系类型对节点表示的影响。为解决此问题,WANG等[40]提出了基于注意力机制的异质图神经网络(heterogeneous graph attention network,简称HAN)。HAN采用经典的层次聚合方式对节点进行全面表示,通过设置节点级别注意力和语义级别注意力,判断不同邻居节点与元路径的重要性,最后通过相应的聚合操作对邻居节点的信息以及不同元路径的信息进行聚合以获得最终的节点表示。

HAN中节点级别的信息聚合如下:

$$e_{ij}^{\Phi}=\mathrm{att}_{\mathrm{node}}(h_i^{'},h_j^{'};\Phi),$$

$$z_i^{\Phi}=\sigma\Big(\sum_{j\in N_i^{\Phi}}\alpha_{ij}^{\Phi}h_j^{'}\Big)。$$

式中:e_{ij}^{Φ} 代表节点 i 在元路径 Φ 下邻居节点 j 的权重;z_i^{Φ} 即节点 i 在元路径 Φ 下的节点级别的表示。每个节点由多条元路径构成,给定一组元路径 $\{\Phi_0,\Phi_1,\cdots,\Phi_P\}$,可以得到一组节点级别的节点表示 $\{Z_{\Phi_0},Z_{\Phi_1},\cdots,Z_{\Phi_P}\}$。

HAN中语义级别的信息聚合如下:

$$(\beta_{\Phi_0},\beta_{\Phi_1},\cdots,\beta_{\Phi_P})=\mathrm{att}_{\mathrm{sem}}(Z_{\Phi_0},Z_{\Phi_1},\cdots,Z_{\Phi_P}),$$

$$Z=\sum_{i=1}^{P}\beta_{\Phi_i}Z_{\Phi_i}。$$

式中:$(\beta_{\Phi_0},\beta_{\Phi_1},\cdots,\beta_{\Phi_P})$ 代表各条元路径的权重;Z 为最终的节点表示。由于每个节点的最终表示 Z 融合了多条元路径及不同邻居节点信息,因此HAN可以为每个节点都获得更为全面的表示。

与HAN不同,ZHANG等[41]提出的HetGNN(heterogeneous graph neural network)则是在节点级别的信息聚合中使用LSTM作为聚合器,对某种关系下的邻居节点信息进行聚合以获得节点的表示,在语义级别的信息聚合中,使用注意力机制获得最终的节点表示。

HetGNN中,节点级别的信息聚合如下:

$$f_2^t(v)=\frac{\sum_{v'\in N_t(v)}[\overrightarrow{\mathrm{LSTM}}\{f_1(v')\}\oplus\overleftarrow{\mathrm{LSTM}}\{f_1(v')\}]}{|N_t(v)|}。$$

式中:$N_t(v)$ 是节点 v 在特定关系 t 下(即某条元路径)的邻居节点的集合;$f_1(v')$ 是邻居节点 v' 的初始节点表示;$f_2^t(v)$ 为节点 v 在节点级别的表示。

HetGNN中语义级别的信息聚合如下:

$$\varepsilon_v=\alpha^{v,v}f_1(v)+\sum_{t\in O_V}\alpha^{v,t}f_2^t(v)。$$

式中:ε_v 为聚合了多条元路径的节点 v 的最终表示;$\alpha^{v,v}$ 表示节点对自身信息的权重;$\alpha^{v,t}$ 表示节点 v 在特定关系 t 下的节点表示的权重;O_V 表示节点类型的集合。

以上2种适用于聚类任务的异质图神经网络虽然取得了不错的效果,但仍然存在一些缺陷,比如需要预先指定元路径,元路径的选择将直接影响到模型的效果,而元路径的选择又需要丰富的先验知识。这一缺陷极大影响了相关任务的效果。为解决这一问题,YUN等[42]提出了自适应选取元路径的GTN(graph transformer networks)模型,该方法从原始图结构中抽取2个图结构,通过矩阵乘法生成元路径,并通过多通道卷积学习不同类型的元路径,然后将GCN应用于每个通道对多个节点表示进行拼接,作为最终的节点表示。该方法虽然显著提升了节点分类任务的性能,但在聚类任务中表现并不佳。FU等[43]提出的MAGNN模型首先是将特定类型的线性变换应用于异构节点属性,将不同类型的节点属性投影到相同的潜在向量空

间,以此解决节点属性的异构性;然后,对于节点级别的信息聚合采用特殊的元路径实例编码器,将节点特征编码为向量,对于语义级别的消息则使用注意力机制进行聚合。

基于以上研究可知,异质图神经网络聚类方法的研究重点主要包括以下几方面:节点级别、语义级别信息聚合方式的选择,不同类型的节点属性处理方式,元路径的选择方式。异质图神经网络能够同时学习网络结构特征以及节点属性特征,显著提高了聚类任务的效果。同时,异质图神经网络适用于任意网络结构的特性也为后续的社区发现任务提供了极大便利。相比于传统社区发现算法,异质图神经网络模型的数据适应性更强,更能有效利用数据的特征。

2.3　方法总结与归纳

根据算法特点及适用的网络结构对上述算法进行了整理,详见表 1。

表 1　研究工作总结与特点比较

Tab.1　Research summary and feature comparison

算法模型	网络结构	算 法 特 点
Brim	二分网络	基于二分模块度的 Brim 算法,在大规模网络中效果不佳
基于密度的模块度	二分网络	使用基于密度的模块度,解决以往模块度无法识别小规模社区的缺点
RankClus	二分网络	采用排名和社区发现相互提升的思路,将排名问题和社区发现问题结合进行社区发现
NetClus	星型网络	与 NetClus 思想一致
RankClass	任意结构	对 NetClus 进行改进,使其可以应用在任意网络结构中
OcdRank	有向异质网络	在有向异质的社交网络中将重叠社区发现和社区成员排名结合进行社区发现,并且支持网络动态更新
Comar 框架	$A=2,R=3$ 的异质网络	使用邻接矩阵表示异质网络,并使用非负矩阵分解获取网络的潜在特征进行社区发现
RJNMF	任意网络	使用联合 NMF 同时分析连接和节点内容信息,并引入正则化获取社区结构
Het_RSC/Het_SC	任意网络	使用基于图的拉普拉斯矩阵分解方法识别异质网络结构,应用 K-means 进行聚类
DenSeC	任意结构	根据节点密度对中心节点进行识别,然后将中心节点从紧密连接区域扩展到稀疏连接区域进行社区发现
Hete_MESE	任意结构	从多维度检测异质网络中的重叠和异质社区,然后使用中心节点扩展方法进行社区发现
HAN	任意结构	使用注意力机制对节点信息以及语义信息进行聚合
HetGNN	任意结构	使用 LSTM 对节点信息进行聚合,注意力机制对语义信息进行聚合
GTN	任意结构	使用卷积网络生成新的图结构进行元路径的学习
MAGNN	任意结构	使用线性变换学习节点属性,使用特殊的实力编码器对节点级别的信息进行聚合

3　异质网络社区发现的评价指标及常用数据集

3.1　异质网络社区的模块度

2004 年,NEWMAN 等[7]创造性地提出了模块度 Q 函数,自此模块度 Q 成为衡量社区划分质量优劣的一个重要评价指标,现已成为常用的评价指标之一。该指标主要面向无向无权的同质网络,网络中不同的社区划分结果对应不同的模块度 Q,模块度 Q 越大,则社区划分结果越好,相反则越差。相比之下,由于异质网络的复杂性,目前仅有部分学者针对二分网络提出了相应的模块度。

2007 年,BARBER[26]率先提出了基于二分网络的模块度,将不同类型的节点划分到同一个社区中,计算方法如下:

$$Q_b=\frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{A(i,j)-\frac{t_i t_j}{m}\right\}\delta(C_i,C_j)。$$

式中:m 表示网络中的边数;$A(i,j)$ 表示邻接矩阵 $\boldsymbol{A}$ 中的元素,若节点相连,则 $A(i,j)=1$,否则 $A(i,j)=0$;C_i 与 C_j 分别表示节点 i 与节点 j 所属的社区,若节点 i 与节点 j 所属的社区相同,则 $\delta(C_i,C_j)=1$,反之则为 0。

2010 年,LIU 等[44]提出了一对多关系的二分模块度,用于表示 2 个社区对应关系的强弱度,计算方法如下:

$$Q_m=\sum\nolimits_l Q_{ml}=\frac{1}{M}\sum\nolimits_{i=1}(e_{tm}-a_l a_m)。$$

式中:l,m 分别对应 2 个异质网络社区;Q 值越大,表示二分网络的结构越强。

3.2　通用评价指标

标准化互信息(normalized mutual information,简称 NMI)[45],是目前广泛使用的一种社区划分评价指

标。计算方法如下：

$$\mathrm{NMI}(R,F)=\frac{-2\sum_{i=1}^{C_R}\sum_{j=1}^{C_F}N_{ij}\log\left(\frac{N_{ij}S}{N_{i*}\ N_{*j}}\right)}{\sum^{C_R}N_{i*}\log(N_{i*}/S)+\sum N_{*j}\log(N_{*j}/S)}。$$

式中：给定一个矩阵 $\boldsymbol{N}$，$\boldsymbol{N}$ 表示由真实社区与社区发现算法计算出来的社区对应的混合矩阵，N_{i*} 表示真实的社区，N_{*j} 表示社区发现算法所得到的社区；N_{ij} 表示同时出现在真实社区与社区算法所得社区中的节点数量；S 表示所有元素之和；C_R 表示真实社区的数量；C_F 表示算法所得到的社区数量。当算法所得的社区划分与真实社区完全一致时，NMI=1；当算法所得的社区划分与真实社区相互独立时，NMI=0。显然，NMI 值越大，社区发现算法效果越好。

3.3 异质网络社区发现的常用数据集

表 2 给出了 4 种常用测试算法效果的异质网络数据集，分别为引用网络 DBLP、美国最大的点评网站 Yelp 数据集、大数据挖掘与服务系统平台 Aminer、电影信息数据集 IMDB。对于每个数据集，给出了数据来源，并统计了该数据集的节点数、关系数、关系类型，方便研究者选择适合模型的数据集。

表 2 常用数据集整理

Tab.2 Summary of commonly used datasets

数据集	关系(A-B)	A类节点数量	B类节点数量	A-B的连边数量	数据集来源
DBLP	Paper- Author	14 328	4 057	19 645	https://github.com/Jhy1993/HAN
	Paper-Conference	14 328	20	14 328	
	Paper-Term	14 327	8 789	88 420	
Yelp	User-Business	16 329	14 284	198 397	https://www.yelp.com/dataset
	User-User	16 329	16 329	158 590	
	User-Compliment	16 329	11	76 875	
	Business-City	14 284	511	14 267	
	Business-Category	14 284	47	40 009	
Aminer	Paper-Label	127 623	10	127 623	https://www.aminer.cn/
	Paper-Author	127 623	164 472	355 072	
	Paper-Conference	127 623	101	127 632	
	Paper-Reference	127 623	147 251	392 519	
IMDB	Movie-Actor	4 780	5 841	14 340	https://github.com/Jhy1993/HAN
	Movie-Director	4 780	2 269	4 780	

4 研究展望

本文介绍了现有的异质网络社区发现方法，根据其特点及网络结构可分为 3 类，分别是二分网络社区发现方法、多模多维社区发现方法以及最近非常流行的基于深度学习的异质网络聚类方法，对比了所介绍模型的适用性与算法特点，给出了经典的异质网络数据集。

近年来，社区发现领域的研究发展迅速，但是在一些方面仍然存在挑战。例如，由于数据类型的多样性，不同领域的先验知识极大影响了社区划分的结果。因此，设计高效通用的社区发现算法非常重要。同时，在近两年崭露头角的基于深度学习的图神经网络方法展现出了巨大潜力，预示着在今后的研究中，图神经网络方法可能成为解决现有问题的一种有效途径。针对社区发现领域的研究现状，未来研究重点将会集中在以下几个方面。

1)评价标准

异质网络社区发现算法仍然缺少统一的评价标准，寻找一个科学的评价标准将是研究人员面临的第一个挑战。根据异质网络社区发现算法的发展趋势，新的评价标准不仅需要考虑同种类型节点间的社区连接强度，还需要考虑不同类型节点间的社区连接强度。面对这一挑战，同时考虑密度与节点扩张程度或许是一个新的方向。同时，面对不同的任务需求，异质网络中的重叠社区发现也变得极为重要，因此如何评价异质网络中的重叠社区划分也是一项新挑战。

2)未知的社区数量

由于从现实世界网络中提取的大多数数据都没有标签，因此无法提前预知社区数量，这会影响到许多需

要预先设定社区数量的算法效果。通常，社区数量的设定是由先验知识决定的，因此社区发现算法可以通过减少对先验知识的依赖解决这一难题，可以从节点属性中学习得到与先验知识相关的特征。尽管已经有部分研究者提出了解决方案，但是这个问题仍然没有得到充分解决，有待进一步研究。

3）动态网络

动态变化会影响网络拓扑结构以及节点特征，使得每个特征都必须以不同的方式进行处理。拓扑变化，如添加或删除边，会引起社区变化以及整个网络拓扑结构的变化。使用动态网络模型通常要用一系列快照进行重新训练，因此未来动态网络时间属性所面临的技术挑战在于对动态特征的提取。

4）节点的表示学习

随着深度学习的发展，基于神经网络的表示学习方法引起了复杂网络研究者的关注，该方法可以将信息网络表示为低维稠密的携带网络节点特征信息的实数向量，并应用于下游任务的输入。由于网络表示学习拥有强大的建模能力，因此将其与社区发现任务相结合将是未来研究的一个新方向。

参考文献/References：

[1] WATTS D J，STROGATZ S H.Collectivedynamics of 'small-world'networks[J].Nature，1998，393(6684)：440-442.

[2] BARABÁSI A L，ALBERT R.Emergence of scaling in random networks[J].Science，1999，286(5439)：509-512.

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

[4] CHEN J C，YUAN B.Detecting functional modules in the yeast protein-protein interaction network[J].Bioinformatics，2006，22(18)：2283-2290.

[5] YANG J，MCAULEY J J，LESKOVEC J.Community detection in networks with node attributes[C]//13th International Conference on Data Mining.[S.l.]：[s.n.]，2013：1151-1156.

[6] CHEN P，REDNER S.Community structure of the physical review citation network[J].Journal of Informetrics，2010，4(3)：278-290.

[7] NEWMAN M E J，GIRVAN M.Finding and evaluating community structure in networks[J].Physical Review E，2004，69(2)：1-16.

[8] CLAUSET A，NEWMAN M E J，MOORE C.Finding community structure in very largenetworks[J].Physical Review E，2004，70(6)：026132.

[9] RAGHAVAN U N，ALBERT R，KUMARA S.Near linear time algorithm to detect community structures in large-scale networks[J].Physical Review E，2007，76(3)：036106.

[10] ROSVALL M，BERGSTROM C T.Maps of random walks on complex networks reveal community structure[J].Proceedings of the National Academy of Sciences of the United States of America，2008，105(4)：1118-1123.

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

[12] XU Y F，XU H，ZHANG D W.A novel disjoint community detection algorithm for social networks based on backbone degree and expansion[J].Expert Systems with Applications，2015，42(21)：8349-8360.

[13] FORTUNATO S.Community detection in graphs[J].Physics Reports，2010，486：75-104.

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

[15] SUN Y Z，HAN J W.Meta-path-based search and mining in heterogeneous information networks[J].Tsinghua Science and Technology，2013(4)：329-338.

[16] SUN Y Z，HAN J W.Mining heterogeneous information networks：A structural analysis approach[J].Acm Sigkdd Explorations Newsletter，2013，14(2)：20-28.

[17] LU M L，QU Z H，WANG Z H，et al.Hete_MESE：Multi-dimensional community detection algorithm based on multiplex network extraction and seed expansion for heterogeneous information networks[J].IEEE Access，2018，6：73965-73983.

[18] SHI C，YU P S.Heterogeneous Information Network Analysis and Applications[M].[S.l.]：Springer，2017.

[19] LIU F Z，XUE S，WU J，et al.Deep learning for community detection：progress，challenges and opportunities[C]// International Joint Conference onArtificial Intelligence.[S.l.]：[s.n.]，2020：4981-4987.

[20] MALLIAROS F D，VAZIRGIANNIS M.Clustering and community detection in directed networks：A survey[J].Physics Reports，2013，533：95-142.

[21] NEWMAN M E J.Fast algorithm for detecting community structure in networks[J].Physical Review E，2004，69：066133.

[22] 赵卫绩，张凤斌，刘井莲.复杂网络社区发现研究进展[J].计算机科学，2020(2)：10-20.
ZHAO Weiji，ZHANG Fengbin，LIU Jinglian.Review on community detection in complex networks[J].Computer Science，2020(2)：10-20.

[23] GUIMERA R，SALES-PARDO M.Missing and spurious interactions and the reconstruction of complex networks[J].Proceedings of the National Academy of Sciences of the United States of America，2009，106(52)：22073-22078.

[24] CHENG J J,LI L J,LENG M W,et al.A divisive spectral method for network community detection[J].Journal of Statistical Mechanics Theory & Experiment,2016,2016(3):033403.

[25] SUN H L,HUANG J B,TIAN Y Q,et al.Detecting overlapping communities in networks via dominant label propagation[J].Chinese Physics B,2015,24(1):018703.

[26] BARBER M J.Modularity and community detection in bipartite networks[J].Physical Review E,2007,76(6):066102.

[27] XU Y C,CHEN L,LI B,et al.Density-based modularity for evaluating community structure in bipartite networks[J].Information Sciences,2015,317:278-294.

[28] SUN Y Z,HAN J W,ZHAO P,et al.Rankclus:Integrating clustering with ranking for heterogeneous information network analysis[C]//International Conference on Extending Database Technology.[S.l.]:[s.n.],2009:565-576.

[29] SUN Y Z,YU Y,HAN J W.Ranking-based clustering of heterogeneous information networks with star network schema[C]//Acm Sigkdd International Conference on Knowledge Discovery & Data Mining.[S.l.]:[s.n.],2009:797-806.

[30] MING J,HAN J W,DANILEVSKY M.Ranking-based classification of heterogeneous information networks[C]//Acm Sigkdd International Conference on Knowledge Discovery & Data Mining.[S.l.]:[s.n.],2011:1298-1306.

[31] QIU C H,WEI C,WANG T J,et al.Overlapping Community Detection in Directed Heterogeneous Social Network[M].[S.l.]:Springer International Publishing,2015:490-493.

[32] COMAR P,TAN PN,JAIN A K.A framework for joint community detection across multiple related networks[J].Neurocomputing,2012,76(1):93-104.

[33] LI Z,PAN Z S,ZHANG Y Y,et al.Efficient community detection in heterogeneous social networks [J].Mathematical Problems in Engineering,2016(12):1-15.

[34] CHEN Y G,SENGUPTA S.Spectral clustering in heterogeneous networks[J].Statistica Sinica,2015,25:1081-1106.

[35] LIU W,CHEN M,LIU K,et al.A Density-based seed-centric community detection algorithm[C]//IEEE Web Information Systems and Applications Conference.[S.l.]:[s.n.],2017:149-154.

[36] 鲁军豪,许云峰.信息网络表示学习方法综述[J].河北科技大学学报,2020,41(2):133-147.
LU Junhao,XU Yunfeng.A survey of information network representation learning[J].Journal of Hebei University of Science and Technology,2020,41(2):133-147.

[37] TOMMASEL A, GODOY D. Multi-view community detection with heterogeneous information from social media data [J]. Neurocomputing,2018,289:195-219.

[38] CAI B,WANG Y,ZENG L,et al.Edge classification based on convolutional neural networks for community detection in complex network [J].Physica A:Statistical Mechanics and Its Applications,2020,556:124826.

[39] CHEN Z D,LI X,BRUNA J.Supervised community detection with line graph neural networks[C]//International Conference on Learning Representations.[S.l.]:[s.n.],2019:1-15.

[40] WANG X,JI H Y,SHI C,et al.Heterogeneous graph attention network[C]//The World Wide Web Conference.[S.l.]:[s.n.],2019:2022-2032.

[41] ZHANG C X,SONG D J,HUANG C,et al.Heterogeneous graph neural network[C]//Acm Sigkdd International Conference on Knowledge Discovery & Data Mining.[S.l.]:[s.n.],2019:793-803.

[42] YUN S,JEONG M,KIM R,et al.Graph transformernetworks[C]//Neural Information Processing Systems.[S.l.]:[s.n.],2019:11960-11970.

[43] FU X Y,ZHANG J N,MENG Z Q,et al.MAGNN:Metapath aggregated graph neural network for heterogeneous graph embedding[C]//WWW'20:The Web Conference 2020 Taipei Taiwan.[S.l.]:[s.n.],2020:2331-2341.

[44] LIU X,MURATA T.Evaluating community structure in bipartite networks[C]//IEEE Second International Conference on Social Computing.[S.l.]:[s.n.],2010:576-581.

[45] 刘井莲,王大玲,赵卫绩,等.一种基于核心节点扩展的社区挖掘算法[J].山东大学学报(理学版),2016,51(1):106-114.
LIU Jinglian,WANG Daling,ZHAO Weiji,et al.A community mining algorithm based on core nodes expansion[J].Journal of Shandong University(Natural Science),2016,51(1):106-114.