计算机科学
COMPUTER SCIENCE

复杂网络社团发现综述

曹金鑫, 许伟忠, 金弟, 丁卫平

引用本文

曹金鑫, 许伟忠, 金弟, 丁卫平. 复杂网络社团发现综述[J]. 计算机科学, 2023, 50(11A): 230100130-11.

CAO Jinxin, XU Weizhong, JIN Di, DING Weiping. Survey of Community Discovery in Complex Networks [J]. Computer Science, 2023, 50(11A): 230100130-11.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230100130

复杂网络社团发现综述

曹金鑫[1]　许伟忠[1]　金　弟[2]　丁卫平[1]
1 南通大学信息科学技术学院　江苏 南通 226019
2 天津大学智能与计算学部　天津 300350
(alfred7c@ntu.edu.cn)

摘　要　现实世界中许多复杂系统均被建模成复杂网络,如社交网络、科学家协作网络等。复杂网络的研究吸引了不同领域的诸多研究者的广泛关注。挖掘社团结构,即将网络划分到具有类内链接稠密、类间链接稀疏的不同社团,是复杂网络研究的问题之一。研究复杂网络社团检测对分析复杂网络中潜在结构、规律以及社团的形成有着至关重要的意义,并且有着广泛的应用前景。鉴于复杂网络中同时包含了网络拓扑与节点内容,结合节点内容的社团检测研究将成为该研究领域的新趋势之一。文中介绍了复杂网络社团检测的研究背景和研究意义;并从基于网络拓扑、基于节点内容和基于网络拓扑和节点内容融合3个方面,较为全面地对社团检测研究现状进行了梳理以及对其面临的问题进行了分析。从3类社团检测方法中选择了10种具有代表性的算法,对它们进行性能对比和时间复杂度分析,以期望描绘关于社团发现新趋势的清晰轮廓。

关键词:复杂网络;社交网络;社团发现;节点内容;信息融合
中图法分类号　TP182

Survey of Community Discovery in Complex Networks

CAO Jinxin[1],XU Weizhong[1],JIN Di[2] and DING Weiping[1]
1 School of Information Science and Technology,Nantong University,Nantong,Jiangsu 226019,China
2 College of Intelligence and Computing,Tianjin University,Tianjin 300350,China

Abstract　Many complex systems in the real world can be modeled as complex networks,such as social networks and scientist collaboration networks. The study of complex networks has attracted the attention of many researchers in different fields. The mining of community structure,division of a network into different communities of nodes with dense intra-community links and sparse inter-community links,is one of the main problems in the study of complex network. Research on community detection in complex networks is of vital importance to the analysis of the potential structure,laws,and the formation of communities in complex networks,and has a wide range of application prospects. Since complex networks contain both network topology and node content,the study of community detection combining node content will become one of the new trends in this field. This paper introduces the research background and significance of community detection in complex networks. And from three aspects based on network topology,node content,and network topology and node content integration,we comprehensively sort out the research status of community detection and analyze the problems it faces. We select 10 representative algorithms from the mentioned three types of community detection methods,and compare their performance of identifying communities and analyze time complexity of these algorithms,hoping to draw a clear outline of the new trend of community discovery.

Keywords　Complex networks,Social networks,Community discovery,Node content,Information fusion

1　引言

现实世界存在大量复杂系统,其潜在的规律和功能具有重要的研究价值。这些复杂系统通常可以建模为复杂网络[1],如通信系统中的电话网络和因特网、社会人际系统中的作家协作网络、生物系统中蛋白质交互网络和基因网络等。基于数学角度,网络可以认为是一类用于建模并描述现实世界中的复杂系统的模型,其中节点对应复杂系统中个体(或是某一单元)、节点之间的链接,表示个体之间的关系。

关于复杂网络分析,一些研究者运用图论、矩阵论、信息论、控制论、统计物理等理论对复杂网络进行定量刻画。1736年,Euler提出了图论[2],以研究小规模的复杂网络,他利用符号描述复杂系统,给予不同研究领域的学者一个较为统一的复杂网络描述性语言。Erdös等[3]于1961年提出了ER随机图理论,对复杂网络研究具有里程碑的贡献。该理论发现,随着网络规模的变大,随机图都具有某种特性。虽然不同的复杂网络可以描述不同的复杂系统,但它们具有一些相似的特性,如复杂网络具有较小平均路径长度和较大聚类系数特性的"小世界效应"[4]和复杂网络中节点的度分布呈现幂律分布统计特性的"无标度效应"[5]。总的来说,上述理论及其特性

基金项目:国家自然科学基金面上项目(61876128,61976120);江苏省自然科学基金面上项目(BK20191445);江苏省高等学校自然科学研究面上项(21KJB520018)
This work was supported by the National Natural Science Foundation of China(61876128,61976120),Natural Science Foundation of Jiangsu Province(BK20191445) and Natural Science Foundation of the Higher Education Institutions of Jiangsu Province(21KJB520018).
通信作者:金弟(jidi@tju.edu.cn)

的研究标志着复杂网络分析成为了一种愈来愈重要的研究领域。

除了小世界、无标度等特性之外，社团结构是复杂网络的又一个重要统计特性[6]。一般地，具有“同一群簇内节点连接相对紧密、不同群簇间节点连接相对稀疏”的节点及边的集合被称为社团(见图 1)，这种结构在社交网络[7-11]、生物信息网络[12-14]中普遍存在。复杂网络中的社团结构的分析，对于获取网络的主要结构、挖掘复杂网络中的隐藏规律以及预测复杂网络行为有着重要的理论意义和广泛的应用价值。因此，复杂网络中社团结构分析吸引了物理、计算机科学、社会科学等越来越多领域的研究者广泛关注，已成为多个领域学科以及交叉学科的研究热点[7,12]。

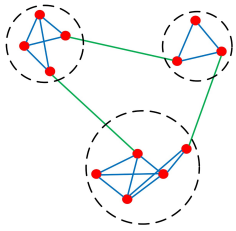


图 1 复杂网络中的社团结构[6]

Fig. 1 Community structure in complex network[6]

自从 2002 年 Girven 等提出社团结构这一概念[6]，大量社团检测理论以及方法不断地被提出。它们已被应用于电话网络的线路布控预测、社交网络的兴趣群落识别、社会人际网络的组织结构分析、生物信息网络的蛋白质功能预测、主控基因识别等，并取得了不错的成绩。根据研究方法不同，传统社团发现方法一般分为：基于划分聚类[15-16]的社团发现、基于层次聚类[17-18]的社团发现、基于模块度[19]的社团发现、基于动力学理论[20]的社团发现、基于统计推理[21-24]的社团发现等。近几年，网络表征[25]、基于神经网络[26-27]的社团检测研究崭露头角，并逐步形成新一类社团发现方法。上述研究被广泛地发表在多个领域的著名国内外期刊和顶级会议，包括国际顶级期刊《Science》《Nature》《Proceedings of the National Academy of Sciences》，CCF 推荐顶级期刊《IEEE Transactions on Pattern Analysis and Machine Intelligence》《IEEE Transactions on Knowledge and Data Engineering》《The ACM Transactions on Knowledge Discovery from Data》，物理领域期刊《Physics Review E》《Physics Review Letter》《Physica A》，国内著名核心期刊《自动化学报》《计算机学报》《软件学报》《计算机研究与发展》《计算机科学》；数据挖掘领域重要会议 SIGKDD，ICDE，CIKM，WSDM，ECML PKDD，人工智能顶级领域会议 AAAI，IJCAI，万维网会议 WWW 等。

传统的社团检测方法往往关注网络中的链接信息，人们意识到许多复杂网络还蕴含了大量的内容信息，如在电话网络中，每个电话节点具有号码、用户姓名、地址等内容；在社交网络中，用户节点具有账号相关信息；在社会人际网络中，个人节点具有人物背景信息；蛋白质交互网络中，蛋白质节点具有其编码等内容。2015 年，Newman[28] 发表文章于《Nature Communication》，并阐述了复杂网络同时包含链接和一些注释信息，如节点上、链接上的文本信息、图片信息等，其中注释信息能辅助链接信息，以改善社团发现的质量。同时，融合拓扑与内容识别的社团结构同时具有真实场景的普遍性和重要的应用价值。鉴于此，越来越多领域的研究人员关注融合网络拓扑和内容信息的社团发现，这使得该类型社团检测迅速成为复杂网络分析领域的新研究热点。

在上述研究背景下，在本文分别从基于网络拓扑、基于内容信息、融合网络拓扑和内容信息这 3 个方侧面，综述了复杂网络社团发现的研究现状，讨论了其所面临的问题以及其发展趋势，尝试描述社团发现研究较为清晰的轮廓，以给出该研究领域有益的参考。

2 复杂网络社团发现研究现状

2002 年，Newman 等[6] 在《Proceedings of the National Academy of Sciences》发表论文并提出了社团结构的概念。随之，许多研究者了提出诸多社团发现方法[16,21,23,29]，并成功地将它们应用到不同学科领域。根据使用拓扑信息和内容信息的不同程度，将这些方法梳理为以下 3 类：传统(基于网络拓扑)社团发现算法、基于内容信息的社团发现算法、融合拓扑信息和内容信息的社团发现算法。算法类别归纳树形图如图 2 所示。本文依次介绍这 3 类社团检测方法的研究现状，并分析其发展趋势。

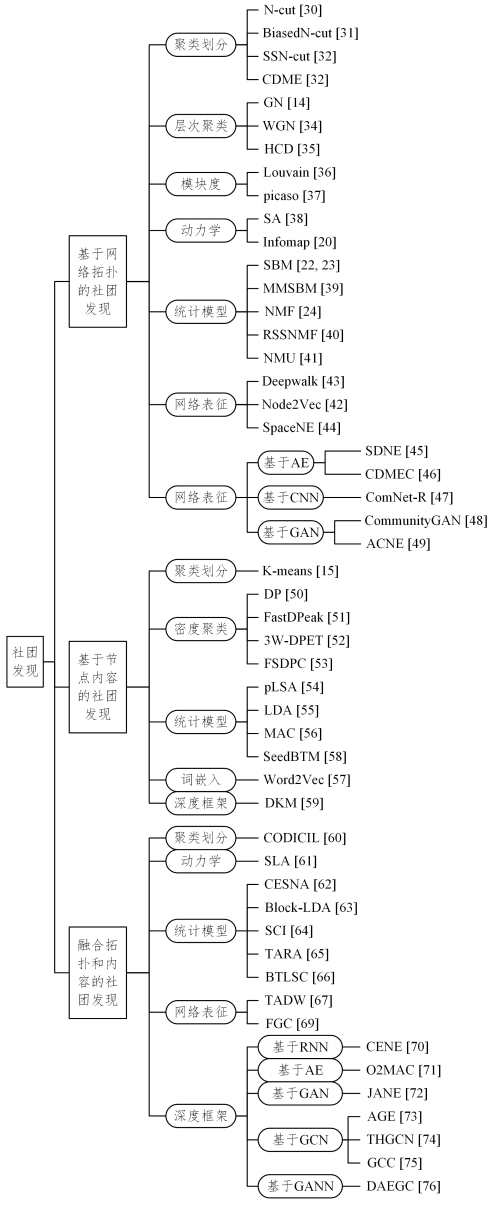


图 2 社团发现算法分类

Fig. 2 Categories of community discovery algorithms

2.1 基于网络拓扑的社团发现

一般地,社团为一组具有紧密链接的节点。因此,一些早期的聚类算法可以用于社团发现,例如划分聚类算法[6]和层次聚类算法[16]。

划分聚类的主要思想是,属于同一类别的节点之间距离足够近,属于不同类别的节点之间距离足够远。Shi 等[30]于2000年提出了一种图切割的方法,即标准化切割(Normalized cut,N-cut),其主要思想可以形式化为广义矩阵特征值分解问题。该方法中 $cut(A,B)$ 函数可描述为,通过移除图的链接将图划分为 A 和 B 两个互不相交的分组,运用删除链接的权重总和来计算分组 A 和 B 之间的差异程度。然后,他们将图中所有节点依次分配到 A 或 B,将总权重添加到 cut 函数,形成 $Ncut(A,B)$,即标准化切割的目标函数。最小化 $Ncut(A,B)$的值以实现图的划分,但是该最小化过程是 NP-complete 问题。为了解决这个问题,Shi 等将目标图划分为 2 个分组改为划分为 k 个分组,同时将指示向量进行松弛,标准化切割的优化问题可运用广义矩阵特征值分解的方式实现,进而可以有效地挖掘网络中的社团结构。N-cut 方法类似于谱聚类,一般适用于均衡分类问题,对于各分类中样本个数相差悬殊的网络识别度不高。针对此问题,Maji 等[31]、Chew 等[32]基于 N-cut 算法添加约束项,分别提出了 Biased N-cut,SSN-cut。

考虑到现实复杂网络中大多数集中于社团内部、社团之间的节点具有相反的情况,Sun 等[33]受马太效应的启发,设计了一种基于图划分的社团检测方法 CDME,解决了 N-cut 中预设社团个数的问题。该方法首先初始化每个节点都是一个独立的社区;接着,基于节点之间的吸引力挑选出核心组,该吸引力依靠节点的度 D_u 以及与不同节点之间的结构相似性 J_{uv},那么节点可能会吸引相邻节点加入其社区,形成核心组,而且迭代计算新的核心组;然后,使用设定迭代方法模拟社区吸引节点的马太效应过程,核心组不断吸引周围节点加入,形成更大的社区结构,并使用归一化互信息 NMI 指标来评估每个社区分区的质量;最后,由于拓扑驱动的影响,网络结构达到稳定状态,以得到社区的最佳划分。他们的方法使用到批量处理和增量技术检测网络社团,很显然,这种方法的精度以牺牲时间复杂度为代价。

而层次聚类的主要思想是,根据某一规则迭代运算样本集合的层次结构,将距离邻近的样本点(或小集合)合并到同一个小集合(或大集合),构建为一棵层次聚类树,对样本点集合进行层次化分类。2002 年,Girvan 等[14]提出了经典层次聚类的社团检测算法 GN,他们基于社团间链接的边介数(Edge betweenness)应当大于社团内链接的边介数这一规则,通过迭代计算"网络中任意两节点经过该链接的最短路径条数",来识别并删除社团之间的链接,建立一个层次聚类树,基于聚类树的分化情况,运用某一划分度量方法实现社团发现。该方法对真实网络中的边权重和节点的度存在刻画限制。Sun 等[34]为区分社团内部链接和社团之间的链接,对两种链接赋予不同权重,对 GN 方法进行链接稳定性改进,所提方法简称为 WGN。

鉴于传统的、基于层次聚类的社团检测方法均需要停止规则考量,Li 等[35]基于二叉树随机模块提出了 HCD 的社团检测模型框架。该框架的迭代过程如下:(1)Li 等基于二叉树随机块模型(BTSBM)将社团嵌入到一棵二叉树中,其中每一个二进制字符串视为表示树的每一个层级,例如第一个数字对应于树顶部的第一个拆分节点,定义为一个 $K\times K$ 矩阵概率的矩阵 $\boldsymbol{B}$;(2)基于矩阵 $\boldsymbol{B}$ 的特征向量进行分解或是谱聚类分析,获取次大特征值对应的特征向量 $\boldsymbol{u}_2$,迭代步骤(2),直至 $K=1$;(3)对特征向量 u_2 进行 K-means 聚类标注社团标签。HCD 模型框架具有无模型差异、时间复杂度低等优点。

2004 年,Newman 等[21]提出了模块度函数 Q,用 Q 来评价所识别网络中社团结构的好坏。不同研究领域的科学家相继提出了基于模块度函数及对其改进的函数,并将其作为优化目标的社团发现方法。其中,Louvain 算法[36]就是一种基于模块度优化的社团检测算法。该算法具体分两个部分:(1)不断地遍历网络中的节点,并将某一节点划分到能最大提升模块度 Q 值的社团,直到不再对节点进行划分;(2)将小社团合并为超节点并构建新网络,然后迭代以上两部分,直到 Louvain 算法收敛。该方法引入边权重,有效降低了时间复杂度,并且解决了解析度问题,其局限性为不适于大规模网络。

针对复杂网络的规模越来越大,Qiao 等[37]提出了一种基于近似优化的并行社团检测算法 picaso,该算法集成了 Mountain 模型和用于模型推导的 Landslide 算法两种新技术,并运用 GraphX 分布计算框架,以实现复杂网络上的并行社区检测。该算法分为以下 4 步:首先,基于模块度函数 Q 设计的模块度增量 ΔQ,运用 GraphX 分布计算框架初始化网络 G;其次,计算每个链组的 ΔQ,并建立 Mountain 模型;然后,近似 ΔQ 并选择多个链组形成新社团,并更新 ΔQ;最后,并行化 picaso 模型,以发现复杂网络中的社区结构。该方法虽然在处理大规模网络时用时短,但是其社团检测的精度优势不大。

基于分子动力学原理,Kirkpatrick 等[38]提出了经典的模拟退火算法 SA,基于"能量越大,分子和原子越不稳定"的物理学原理构建算法。模拟退火算法分为两个步骤,即 Metropolis 算法和退火过程,解决了聚类的局部最优解问题。SA 算法的局限性在于高鲁棒性牺牲了收敛速度,不适用于规模大的网络。

2007 年,Rosvall 等[20]基于动力学理论提出了 Infomap 算法,其主要思想是在一个图上进行不限步数的随机游走,并用 Huffman 编码来刻画游走产生的路径,所得图中每个节点的编码不一样,其中具有重复编码的节点划分为同一社团。由于算法存在随机采样,Infomap 算法对于同一网络的检测精度会出现震荡的情况。

与聚类算法和模块度模型不一样,为探究网络结构的生成方式,一些研究者[22-23]提出了一类统计模型,即随机块模型(Stochastic Block Model,SBM)。2008 年,Airoldi 等[39]基于随机块模型提出了混合隶属度随机块模型(Mixed Membership Stochastic Block Model),用于解决节点可以隶属于多个社团的问题。这类模型具有较为灵活地刻画不同网络拓扑结构的能力,但是其用于建模的模型参数数量也会按指数级增长,因此模型时间复杂度一般较大。

2012 年,Jin 等[24]基于随机块思想,运用非负矩阵分解(Non-negative Matrix Factorization,NMF)技术将邻接矩阵分解为两个非负值矩阵,然后对提出的模型进行优化,将学习所得的非负矩阵作为社团隶属度矩阵,以挖掘网络中的社团结构。2019 年,He 等[40]提出了一种融合先验信息的半监督

NMF 方法，并添加 2-1 范数来提高模型的鲁棒性。此外，2021 年，Luo 等[41]为了处理基于 SNMF 的模型中比例因子不可调的问题，设计了一种 NMU 方案，用于比例因子自调节，同时保证模型的收敛以及较高的社团检测精确度。该类模型往往适用于聚类数目预设的社团检测任务场景。

随着网络表征学习的发展，网络嵌入(Network Embedding)算法被广泛应用于社团检测及其相关研究领域[25]，基于网络嵌入的算法层出不穷。2016 年，Leskovec 等[42]提出了经典网络嵌入算法 Node2Vec，其主要思想是基于随机游走方式的 Deepwalk 算法[43]进行扩展。Deepwalk 算法随机选取随机游走序列中的下一个节点，而 Node2Vec 算法设置两个参数，分别以深度优先和广度优先策略选择随机游走序列中的节点。然后，他们利用构造节点在网络上的随机游走路径模仿文本生成的过程，进而学习每个节点在低维空间中的表征。最终，通过节点表征的聚类来实现社团发现。可以发现，上述模型采用随机过程，模型的精度会出现波动的情况。

Long 等[44]利用网络中成对节点的邻近性描述社区，从而形成由社区组成的层次结构，提出了一种网络嵌入框架 SpaceNE。他们提出的网络表征学习方法通过子空间、具有灵活维数的流形保持社区形成的层次结构。此外，子空间能够解决表示分层社区中稀疏性和空间扭曲等问题。他们对要降噪的子空间维度进行约束，并进行联合优化，以提高方法的效率。该模型需要计算不同节点之间的相似度，其使用场景多为规模较小的网络。

鉴于深度学习的蓬勃发展，亦有诸多基于深度学习的社团检测方法被提出。为了能保留网络拓扑的局部和全局结构信息，Wang 等[45]利用自动编码器(Autoencoder)同时优化一阶和二阶相似度图的图嵌入算法 SDNE 算法。他们使用模块最小化节点的隐空间距离建模一阶相似度图，使用邻接矩阵和节点邻居结构相似矩阵建模二阶相似度图，分别运用自动编码器和谱聚类作为框架，以构建社团检测的联合模型，获取的网络表征亦可用于社团发现。同样基于 Autoencoder 框架，Xu 等[46]将 4 种不同的复杂网络相似性表示应用于社区检测问题，这些相似性表示可以充分描述和考虑网络拓扑中节点之间的充分局部信息，提出的 CDMEC 框架结合了转移学习和堆叠式自动编码器，以获得复杂网络的高效低维特征表示。基于自动编码器的模型需要逐层训练，其检测社团结构的能力以牺牲时间复杂度为代价。

基于卷积神经网络(CNN)框架，Cai 等[47]提出了一种 ComNet-R 方法。他们设计了一个 E2I 模型，将复杂网络的边结构转换为图像结构，并将其作为训练数据输入到基于 CNN 的模型(ComNet)，经过训练的 ComNet 模型足够有效，可以识别网络中的边缘是社团内部的还是社团之间的；随后分割网络并将初步得到的社团与局部模块化 R 函数合并，以优化社区结构，从而获得最终社团划分结果。该类模型一般适用于规模较小的网络。

而基于对抗人工网络(GAN)框架，Jia 等[48]为了处理图表征学习方法难以处理重叠社团结构的问题，提出了一种社区检测框架 CommunityGAN，它联合解决了重叠社区检测和图表示学习问题。首先，与传统的图表示学习算法的嵌入不同，传统的图表达学习算法的向量输入值没有特定的含义，CommunityGAN 的嵌入表明了垂直于社区的成员强度。其次，采用专门设计的 GAN 来优化这种嵌入。motif 生成器和鉴别器之间的极大极小竞争提高了模型的性能，并使得提出的模型输出更好的社区结构。Chen 等[49]考虑网络嵌入难以处理重叠社团的问题，提出了结合网络嵌入和重叠社团检测的方法 ACNE；采用带感知的游走策略来获得包含更多可能边界顶点的路径，借助简单分类器的软社团分配作为监督来更新 ACNE 的权重，通过以上两步实现顶点分类和重叠社团检测。由于 GAN 不适合离散数据，这类模型难以处理属性网络。

虽然上述算法的类别不尽相同，但是它们具有相似之处，即仅依靠网络中的拓扑信息(或链接信息)挖掘网络中的社团结构。当网络的拓扑信息存在噪音或是丢失时，这类方法识别网络中社团结构的精度将会大打折扣。

2.2 基于节点内容的社团发现

这类社团发现算法所分析的复杂网络可以被刻画为，节点上存在内容信息、节点之间无链接的网络。一般地，如此“复杂网络”可视为数据样本集合，节点表示样本个体，节点内容表示样本个体上的内容信息。那么，挖掘这类网络中社团结构，即将内容信息相似的样本个体划分到同一类别。这类算法不完全属于社团发现算法，可以认为是数据聚类算法，且该算法运行的基本前提是网络拓扑所蕴涵的社团结构和节点内容所蕴涵的类簇结构两者具有对应关系。

1967 年，MacQueen 等[15]提出了一种经典的划分聚类算法 K-means，其主要思想是，将数据节点通过特定计算公式映射到某一种度量空间，依据不同节点之间的距离，将距离相近的节点划分到同一类别，以获得数据聚类。K-means 对样本噪音敏感，依赖于原始样本的真实情况。

Rodrigues 等[50]于 2014 年在《Science》上提出了基于密度峰值的聚类算法 DP，该算法基于以下两个假设：(1)每个类别中心节点(具有密度峰值的节点)的局部密度大于它邻居节点的局部密度；(2)不同类别的中心节点之间距离相对较远。在该算法中，他们首先计算每个节点 i 附近的密度 p_i；然后，计算每个节点 i 与较高局部密度的其他节点之间的最短距离 δ_i，取 p_i 值和 δ_i 值均高的节点作为类别中心节点；最后，将其余节点划分到局部密度高且距离近的类别，即实现聚类。DP 算法时间复杂度高，这限制了其应用于大规模数据的聚类。

针对 DP 算法的时间复杂度高的问题，Chen 等[51]基于 kNN-密度等价于局部密度的思想，识别局部密度峰值和非局部密度峰值，加快了计算 p_i 的速度。他们还构建了一个包含较高密度节点的层次树，运用三角不等式策略来加速计算 δ_i。所提算法 FastDPeak 的时间复杂度为 $O(n\log n)$，低于 DP 算法的时间复杂度 $O(n^2)$。Yu 等[52]针对 DP 中高局部密度的邻居数据点引发聚类标签错误传播的问题，提出了一种基于证据理论的三向密度峰聚类方法 3W-DPET，使用三向聚类表示将聚类形成为区间集，包括 3 个不相交的正区域(POS)、边界区域(BND)和负区域(NEG)，有效地发现了数据集中聚类的类别。Xu 等[53]考虑到构建决策图所需时间开销大的问题，提出了一种快速稀疏搜索密度峰值聚类(FSDPC)算法，用于增强 DP，设计了一种新颖的稀疏搜索策略，用于测量每个数据点的最近邻居之间的相似性，以提高算法效率。DP 类的算法聚类功能易于实现，但时间复杂度高一直是其短板。

在自然语言处理领域，大量文本聚类算法被提出。这些算法多属于统计模型，其中流行的算法大多基于经典的话题

模型,其主要思想是将具有相似"主题"属性的文档归置到同一类别。1999 年,Hofmann 等[54]利用概率生成模型分析文本集合并将其归类到不同的话题。他们在模型 pLSA 中引入话题这个隐变量,基于文本生成话题、话题生成单词的生成过程,进而获得单词-文本共现矩阵。然后,他们运用期望最大化算法推导隐变量基于最大似然度的解,以实现文本集合的聚类。

2003 年,Blei 等[55]在 pLSA 算法的基础之上添加狄雷克雷先验分布,以解决模型过拟合问题,提出了 LDA 算法。该模型是一种三层的贝叶斯概率模型,由文档、主题和词 3 层构成。其中,文档的主题概率分布、主题的词分布均服从于狄雷克雷分布。他们运用吉布斯采样的策略来实现模型推导,以聚类属于不同主题的文本。主题模型是基于生成框架构建的,其参数量呈指数级,解空间大,模型精度依靠较高的时间复杂度。

2009 年,Frank 等[56]提出了一种面向布尔型样本数据聚类的统计概率模型 MAC。首先,他们基于生成框架对样本数据 $\boldsymbol{x}^S \in \{0,1\}^{N\times D}$ 非负矩阵分解成为二值指示矩阵 $\boldsymbol{z} \in \{0,1\}^{N\times K}$,指示矩阵的元素由伯努利分布生成;然后,运用面向维度噪音(Dimension-wise Noise)和面向目标噪音(Object-wise Noise)的扩展混合噪音模型(Mixture Noise Model),构建第一步所描述的模型,以实现布尔型数据的多隶属类别聚类。

Mikolov 等[57]提出了词向量的概念,嵌入型表示技术在自然语言处理领域得到广泛应用。他们提出的工具 Word2Vec 可训练出低维、稠密且具有上下文语义信息的词向量。Word2Vec 是一个基于连续词袋模型(Continuous Bag of Words,CBOW)或是跳字模型(Skip-gram)的单层神经网络。CBOW 根据某一词的前面或是后面 n 个连续的词来计算当前词出现的概率,而 Skip-gram 根据当前词分别计算出它前后出现 n 个词的概率。CBOW 和 Skip-gram 通过层次 SoftMax 或负采样进行模型推导以获取词向量,词向量进一步可用于文本聚类。Word2Vec 算法是一种处理文本词-向量的单一方式,难以处理文本重叠分类的场景。

在 Dataless 短文本分类任务中,Yang 等[58]提出了一种处理在线短文本的种子双项主题模型(SeedBTM),融合主题模型和词嵌入,实现了带有 seed-word 的 Dataless 短文本分类。SeedBTM 利用主题模型中的词共现信息和词嵌入中的类别词相似度作为先验的主题集合,不仅短文本分类精度得到提升,而且较好地处理了具有重叠话题的短文本数据分类。

Fard 等[59]联合聚类和表征学习,提出了一种基于 Autoencoder 深度网络框架的算法 DKM。该深度框架包含两部分:(1)需要聚类的数据 $\boldsymbol{X}$,运用原始 Autoencoder 的数据重构思想学习处数据 $\boldsymbol{X}$ 的低维表征 $h(x)$;(2)$\boldsymbol{r}_k$ 为 K-means 中类别为 k 的低维表征,设计类别 k 中其他数据点的低维表征 $h(x)$ 与 $\boldsymbol{r}_k$ 的约束相似,并将该约束以正则项的方式添加到第一部分 Autoencoder 目标函数中。由以上两部分所构建的 Autoencoder 即可实现 K-means,并基于深度框架学习各个节点的表征以实现聚类。该模型基于 Autoencoder 框架构建,需要逐层训练,其精度以牺牲时间复杂度为前提。

这些数据聚类算法能成功应用于社团发现算法,这充分说明内容信息也蕴含了社团,有助于网络拓扑挖掘网络中的社团结构。

2.3 融合网络拓扑和节点内容的社团发现

众所周知,属于复杂网络的社交网络同时包含拓扑信息和内容信息。例如,微博网络包含了用户之间关注与被关注所表示的拓扑信息;亦包含用户背景信息、发布微博等文本所表示的内容信息;Flickr 网络具有用户之间的联系,以表示网络拓扑,同样包含了用户交互图片所表示的内容信息;科学家协作网络同时带有作者合作关系表示的网络拓扑和作者研究兴趣、论文列表等文本表示的内容信息。根据 2.2 节的分析,可以得知内容信息有助于社团发现。已有研究者提出了不少融合拓扑和内容信息的社团发现算法,并取得了不错的效果。按照不同方法,我们梳理的社团检测算法如下。

基于图划分的思想,Ruan 等[60]提出了 CODICIL 算法。他们在当前节点与其 k 个具有最相似内容的邻居节点之间构建内容型链接 E_c。然后,将内容型链接 E_c 和原始网络的拓扑型链接 E_t 融合,以形成联合链接 E_u 集合,通过"边采样"的策略保留联合链接集合中最相关性的 E_u。最后,基于 E_u 和原始网络中的节点构造图并进行图划分聚类,进而实现融合拓扑信息和节点内容信息的社团发现。该算法简单易行,但其构造联合链接的时间复杂度 $O(n^2 \log n)$ 较高,难以处理大规模的属性网络。

基于动力学理论,Bu 等[61]处理属性图聚类(AGC)的问题,将其理解为动态集群形成博弈(DCFG),其中的每一个节点的可行动作集可以收到离散时间动态系统中每个集群的约束。他们采用东塔社会博弈(DSG)与 DCFG 关联,并验证了 AGC 的平衡解可以通过求解相关 DCFG 中的有限耦合静态纳什平衡问题来找到,Bu 等将上述思路总结为一种自学习算法 SLA,该算法从任意初始集群配置开始,使有限纳什均衡序列在 DSG 中收敛,最终找到对应 AGC 平衡解,即获取所有节点和集群所满足的集群配置。可以发现 SLA 模型需要基于纳什均衡来获取最优解,因此该模型限制于大规模属性网络任务。

基于统计模型的思想,Yang 等[62]基于这样一种假设:"网络中社团同时生成了网络拓扑和节点属性,它们的生成过程相互独立"。他们利用伯努利分布分别构建基于链接和基于内容的概率模型,然后构建联合概率模型 CESNA,以结合这两种信息,最后通过推导 CESNA 最大化似然来获得节点的社团隶属度,进而挖掘网络中的社团结构。由于 CESNA 模型基于相互独立逻辑模型进行构建,该模型依赖节点属性的类型、节点属性个数应小于网络节点个数等。

Balasubramanyan 等[63]认为,文本实体之间的链接可以为挖掘文本中的主题提供线索,联合建模在网络结构和文本之间共享有关潜在主题的信息,从而产生更连贯的主题。基于此,他们提出了一种融合拓扑链接和文本的模型 Block-LDA,他们运用混合隶属度随机块模型建模链接,运用主题模型(Topic model)建模文本信息,将主题模型中的潜在主题(Topic)和随机块模型中的块结构(Block)合二为一,以构建联合模型。然后,运用吉布斯采样完成模型推导,以获取文本数据集的话题分析。

Wang 等[64]考虑到属于同一社团的节点具有链接稠密,也具有相似特征的节点属性,提出了 SCI 算法,以实现融合网络拓扑与节点属性。此外,他们还考虑节点属性应当可以语义解释所发现的社团结构。基于此,Wang 等运用生成框架、非负矩阵分解技术对网络拓扑建模,然后构建社团-属性隶属

度矩阵，基于“属于当前社团的节点具有与社团相近的属性”，去拟合社团隶属度矩阵。通过模型优化，他们所获得的社团-属性隶属度矩阵不仅可以识别网络中的社团结构，也可以基于节点属性描述所识别的社团。为了描述语义社团结构，对语义与社团结构存在偏移的属性网络效果不佳。

2019 年，Hu 等[65]认为，虽然基于模型的算法能够避免特定度量的局限性，但只适用于属性信息为二进制形式的属性网络。于是，他们引入了“节点-属性-值”三层的层次结构，以灵活和可解释的方式描述属性信息，并提出了一种贝叶斯模型 TARA，用于模拟复杂网络的生成过程。在 TARA 中，属性信息是通过层次结构生成的，而成对节点之间的链接是由随机块模型生成的，然后运用变分方法实现模型推导。针对大规模复杂网络，他们通过并行化使 TARA 模型高效地执行社团检测任务。

Jin 等[66]发现，网络内容中的词往往体现出了层次化的语义结构，忽略这种语义结构通常会导致描绘网络的内容不够准确。为了解决这个问题，他们提出了一种新的贝叶斯概率模型 BTLSC，通过将单词与背景主题或某些两级主题（即一般主题和专业主题）区分开来。该模型不仅可以更好地利用网络内容来帮助寻找社区，还可以提供更清晰的多元语义社区解释。他们给出了一种用于模型推理的有效变分算法，以提高算法的效率。上述研究基于层次结构建模语义信息，模型的解空间呈现指数级，运用变分算法有效提高了模型的收敛速率。

网络嵌入迅猛发展，鉴于现实世界网络中包含了丰富的内容信息未被表征学习方法充分利用，一些研究者也提出了融合网络拓扑和节点内容的网络嵌入算法进行社团检测。2015 年，Yang 等[67]提出了具有文本关联的 TADW 算法，他们证明了 Deepwalk 网络表征算法与低秩矩阵分解的理论等价。在矩阵分解的框架下，他们运用矩阵补全（Inductive Matrix Completion）的思想[68]将节点的文本特征引入到网络表征学习。为了提高计算效率，他们的观测数据设置为顶点 i 在 t 步随机走中到顶点 j 的对数化平均概率，进一步简化为 2 步随机游走矩阵 $\boldsymbol{M}$。在 TADW 算法中，他们按照非负矩阵分解将 $\boldsymbol{M}$ 矩阵分解为 $\boldsymbol{W}$ 和 $\boldsymbol{H}$；然后，运用矩阵补全的思想，将上述分解转化为“$\boldsymbol{M}$ 矩阵分解为 $\boldsymbol{W}$ 和包含文本特征矩阵 $\boldsymbol{T}$ 的 $\boldsymbol{HT}$”。那么，他们通过模型推导获取的网络表征 $\boldsymbol{H}$ 就融合了网络拓扑和文本信息，运用 K-means 等方法在网络表征 $\boldsymbol{H}$ 上进行聚类，以实现社团检测。该算法使用的 2 步随机游走矩 $\boldsymbol{M}$ 对网络的广度和深度两个特性描述具有局限性，不适用稀疏的大规模属性网络。

Kang 等[69]提出了一种属性图聚类算法 FGC，旨在自动构建相似图矩阵。他们探索属性图聚类的边界节点特征和结构信息。FGC 模型学习基于卷积特征的相似图，该图需要最佳逼近初始高阶关系（细粒度图）。通过细粒度图的图聚类获取的网络嵌入信息产生高质量的聚类。该模型适用于属性网络的非重叠社团检测。

与 TADW 算法的隐式建模、无序文本信息的引入相比，Sun 等[70]考虑到了文本的上下文信息，基于深度学习框架提出了融合单词级别文本的 CENE 算法。他们将网络刻画为节点-节点型链接和节点-内容型链接。然后，他们分别运用随机游走算法和 3 种句子表示学习组合模型（词嵌入、RNN、BiRNN）构造节点-节点子模型和节点-内容子模型，基于负采样的策略分别构建节点-节点损失函数 L_m 和节点-内容损失函数 L_{nc}，通过平衡参数融合 L_m 和 L_{nc} 为联合目标函数 L，以形成融合拓扑和内容的 CENE 模型。最后，他们运用随机梯度下降算法对 CENE 实现模型优化，以获得表征并进行社团检测。由于在建模过程中运用了随机游走方式，CENE 算法的精度会出现震荡的情况。

基于 Autoencoder 框架，Fan 等[71]将深度学习技术用于属性多视图的图聚类，并提出了一种任务引导的自编码器聚类框架 O2MAC。该图形自动编码器能够通过使用一个信息丰富的图形视图和内容数据来重建多个图形视图以学习节点嵌入，可以很好地捕获多图的共享特征表示。此外，他们提出了一种自训练聚类目标，以迭代改进聚类结果。通过将自训练和自编码器的重构集成到一个统一的框架中，提出的模型可以联合优化适合图聚类的簇标签分配和嵌入。该模型适用于多视图属性图的聚类。

基于 GAN 神经网络结构，Yang 等[72]针对对抗正则化网络嵌入存在直接比较嵌入和高斯先验生成的样本且无法捕获语义信息的缺点，提出了一种联合对抗网络嵌入的算法 JANE，以联合网络拓扑、节点特征和网络嵌入进行组合而不是比较嵌入与高斯分布的样本。JANE 由 3 个可插拔组件组成：嵌入模块（E）、生成器模块（G）和鉴别器模块（D）。在 JANE 中引入了嵌入模块，以生成网络嵌入，而不是生成新数据（属性网络）。其次，JANE 与 GAN 中完全基于高斯分布获得的样本生成伪属性网络不同，将真实属性网络的嵌入结果与高斯分布生成的样本相结合，生成伪嵌入，然后生成伪属性网络。

Cui 等[73]分析现有的基于图卷积神经网络（GCN）的方法存在 3 个主要缺点：（1）图卷积滤波器和权重矩阵的纠缠会损害性能和鲁棒性；（2）图卷积滤波器是广义拉普拉斯平滑滤波器的特殊情况，不能保持最佳低通特性；（3）现有算法的训练目标通常是恢复邻接矩阵或特征矩阵。为了解决这些问题，他们提出了自适应图编码器（AGE），AGE 由两个模块组成：拉普拉斯平滑滤波器模块和自适应编码器模块。AGE 首先使用拉普拉斯平滑滤波器处理数据特征，然后运用自适应编码器对其进行迭代增强，以获得更好的节点嵌入。AGE 需要计算网络中不同节点对的相似度，一般适用于小规模的属性网络。

Zheng 等[74]处理异构的、动态变化的图数据，提出了时间异构图卷积网络（THGCN），使用一组时间异构图的学习特征表示来检测社区。在 THGCN 模型中，他们首先设计了一个异构 GCN 组件，用于表示异构图在每个时间步的特征。然后，提出了一个残差压缩聚合组件，用于学习从两个连续异构图中提取的时间特征表示，以进行异构网络的社团检测。

Fettal 等[75]同样基于 GCN，在一个统一的框架中提出了一个同时考虑聚类进而表征学习这两个任务的目标函数，模型命名为 GCC。基于简单图卷积网络的一种变体，GCC 模型通过最小化卷积节点表示与其重建的聚类类别代表之间的差异来进行聚类。

此外，基于图的注意力机制神经网络（GANN）的方法，Wang 等[76]考虑到，图嵌入不是目标导向的，即为特定的聚类任务而设计的，提出了一种目标导向的深度学习方法，即深度注意力嵌入式图聚类（DAEGC）。该方法侧重于属性图，以充分探索图中信息的两个方面。通过使用注意力网络来捕获相邻节点对目标节点的重要性，DAEGC 算法将图中的拓扑结构

和节点内容编码为紧凑表示,在该表征上训练内积解码器以重建图结构。此外,基于图嵌入生成的软标签来监督自训练图聚类过程,以迭代地细化聚类结果。该模型需要计算不同节点对的度量,因此其深度模型的训练时间复杂度也相对较高。

由于这类方法大都基于传统的社团发现方法扩展得到,它们均受到传统方法所带来的局限,例如基于统计推理的方法难以刻画真实网络中的非二值的内容属性等。此外,拓扑和内容信息与社团结构的关系一般是未知的,上述方法大多依靠人工方式调整这两种信息的比重,来实现识别社团结构的质量提升。

3 实验分析

为了定量地分析不同社团检测算法,我们按照基于网络拓扑、基于节点内容、结合网络拓扑和节点内容的分类选择了10种不同的代表性社团检测方法,并进行了比较实验。实验环境为Dell品牌R740系列刀片服务器,具体配置如下:Intel Exon Silver 4210R CPU @ 40核心2.4 GHz、内存125.5 GB、硬盘1 TB、操作系统Ubuntu 18.04.5 LTS。本文中,对比算法的代码都从所属作者处获得,各算法中的参数均为默认设置。

我们先给出测试本文方法的数据集描述。该数据集是社团发现领域中常用的人工生成网络,由Lancichinetti等[77]于2008年提出的LFR基准数据集,具有"节点度和社团规模的幂律分布"的异构特性,该特征比较贴近真实网络。在LFR基准数据集中,控制如下参数变换网络的特性:参数 n 表示节点数目的网络规模,参数 C 表示网络中社团个数,参数 m_{avg} 表示网络中节点的平均度,参数 $c_{\min}$ 和 $c_{\max}$ 分别表示社团规模中最小值和最大值,混合参数 μ 表示节点 v_i 与社团外部节点的链接数与节点 v_i 的度之比,参数 α 表示节点的幂律分布系数,参数 β 表示社团规模的幂律分布系数。与Lancichinetti等设置实验参数类似,我们在本文中设置网络规模 $n=1\,000$ 或5 000,社团规模 $c_{\min}$ 和 $c_{\max}$ 均为50,社团个数 $C=20$ 或100,节点平均度 $m_{avg}=16$,混合参数 μ 取值以0.05为间隔从0.1变化到0.6,节点度和社团规模的幂律分布系数(α,β)为(2,1),(2,2),(3,1),(3,2)。根据以上参数设置产生8组不同的网络,以不同网络规模将8组网络归为2类网络规模分别为 $n=1\,000$ 和 $n=5\,000$ 的网络组。

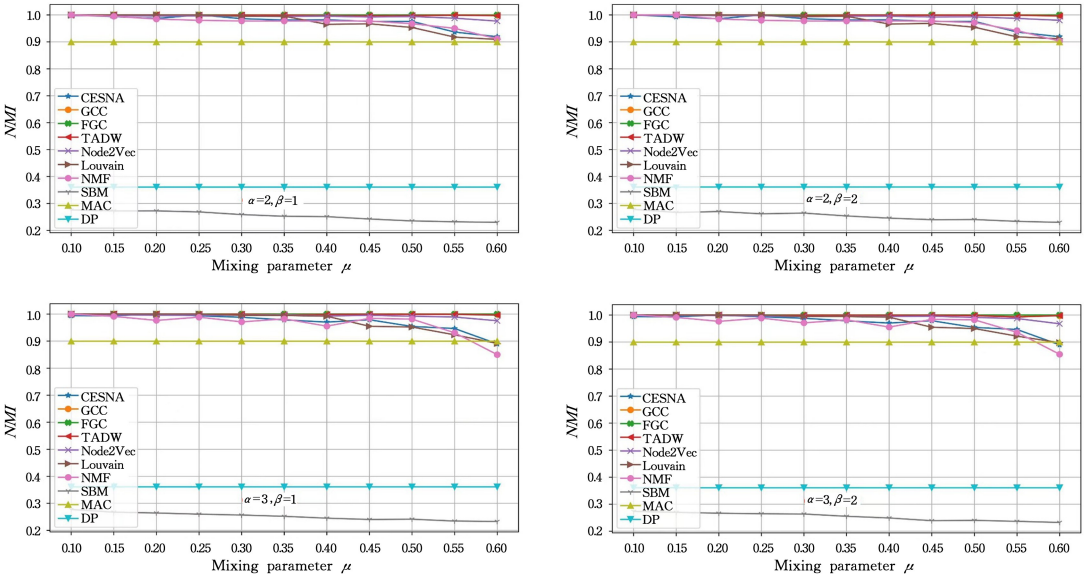
由于现有人工网络均未包含节点内容信息,考虑到定量分析结合拓扑与节点内容的算法,我们构建一种全新的、具有节点内容信息的LFR基准数据集。新的基准数据集的网络拓扑按照上述LFR基准数据生成,节点内容的生成步骤如下:(1)LFR包含的每个社团包含1维属性,并设置参数 $\varphi=32$ 和 $\varphi_{in}=8$,参数 p 表示属性向量的维度;(2)第 i 个社团中每个节点的属性向量中第$((i-1)\times\varphi+1)$到第$(i\times\varphi)$元素值基于系数为 $\Omega_{\text{in}}=\varphi_{\text{in}}/\varphi$ 的二项式分布生成;(3)属性向量中其余元素值基于系数为 $\Omega_{\text{out}}=(\varphi-\varphi_{\text{in}})/(p-\varphi)$ 的二项式分布生成。这样便构建了一种新型、具有节点内容的LFR基准数据集。

为了评估不同社团检测方法的性能,我们将这些方法运行于上述带有节点内容信息的LFR基准数据集,并运用基于信息熵的评价方法标准互信息熵(Normalized Mutual Information,NMI)[78]作为算法检测社团的度量标准。其中,NMI值越高,测试算法识别的社团结构与网络中真实的社团结构之间的相似程度就越高。我们还分析每种测试算法的时间复杂度,在本文运用n,m,C分别表示网络中节点数、链接数、社团数,以便形式化描述复杂度。

3.1 基于网络拓扑的社团发现

本文选择了4种具有代表性的基于网络拓扑社团发现算法作为测试算法,它们分别是Louvain算法[36]、SBM算法[23]、NMF算法[24]、Node2Vec算法[42]。

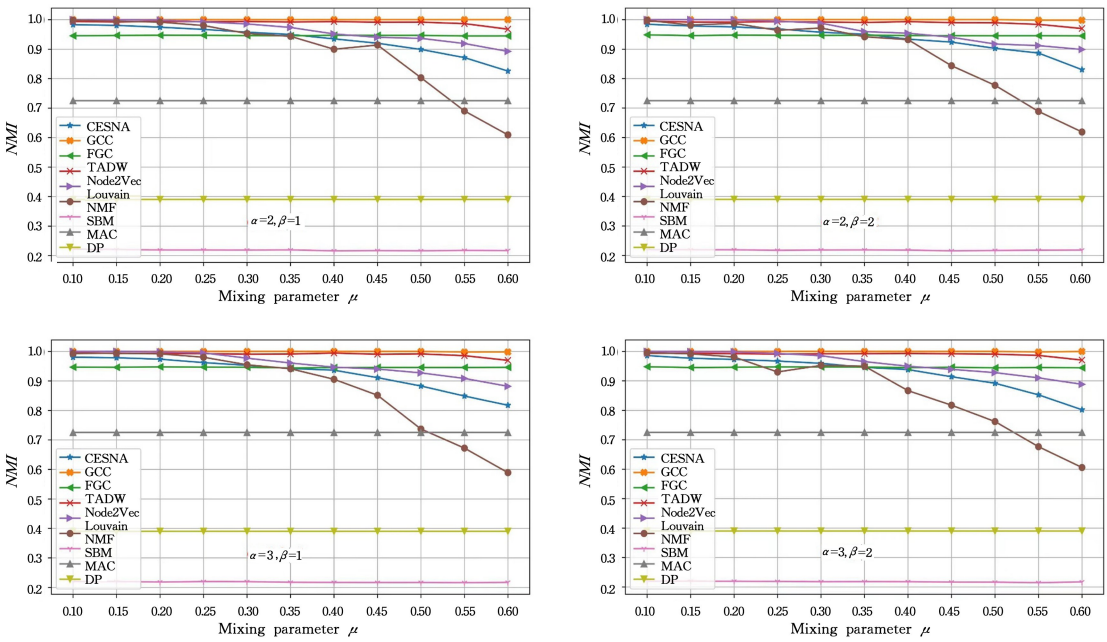
通过理解生成LFR网络的算法原理,混合参数 μ 值逐渐增大,人工网络中社团之间的链接逐渐增多。那么,人工网络中社团结构也随之越来越模糊,这会使得测试算法的识别网络中社团结构的性能逐渐下降。如图3、图4所示,不同算法的检测社团结构的NMI精度曲线均出现不同程度的下降。需要注意的是,NMI精度曲线下降速度缓慢的测试算法具有更为优越的社团检测能力。



注:网络规模 $n=1\,000$,参数 α 和 β 分别表示生成网络算法中度分布和社团规模分布的指数。

图3 不同方法在带有节点内容LFR网络上的NMI曲线

Fig. 3 Performance comparison of the different methods on LFR benchmark with node contents in terms of NMI



注：网络规模 $n=5000$，参数 α 和 β 分别表示生成网络算法中度分布和社团规模分布的指数。

图 4 不同方法在带有节点内容 LFR 网络上的 NMI 曲线

Fig. 4 Performance comparison of the different methods on LFR benchmark with node contents in terms of NMI

如图 3 所示，本节中涉及的基于网络拓扑的算法中，Node2Vec 算法是网络嵌入型算法，可获取更为丰富的网络拓扑信息，其表现比其他 3 种算法更好。Louvain 算法是基于著名的模块度函数 Q 设计的，会受社团规模影响；NMF 算法基于数据降维思想，对网络中的节点度敏感。从实验结果可以看出，当网络中节点度和社团规模变化时，Louvain 算法与 NMF 算法检测网络中社团结构的表现均出现波动。SBM 算法基于生成思想构建，其原理是社团基于概率分布生成网络中的链接，很少涉及网络中节点度和社团规模等信息的建模。因此，SBM 算法在带有节点内容的 LFR 基准数据集上检测社团也表现不佳。

图 4 给出了不同算法在较大规模（$n=5000$）的网络上检测社团结构的表现。Node2Vec 算法的时间复杂度为 $O(n)$，即与网络规模呈线性关系。Louvain 算法的时间复杂度为 $O(nC+m)$，即与网络规模和节点度呈线性关系。这两种算法均具有较低的算法复杂度，更适于处理大规模网络。NMF 算法和 SBM 算法的时间复杂度分别是 $O(nmC)$ 和 $O(n^2C^2)$，即均接近网络规模的二次方。从实验结果可以看出，Node2Vec 和 Louvain 算法处理大规模网络（$n=5000$）的表现与在网络规模 $n=1000$ 的带有节点内容的 LFR 数据集上的表现相比变化较小，Node2Vec 算法的表现更胜一筹。而 NMF 算法和 SBM 算法的检测社团的表现均因网络规模变大而受到影响，SBM 算法表现所受的影响更为明显。

3.2 基于节点内容的社团发现

本文选择了 MAC 算法[56]、DP 算法[50]作为只运用内容信息进行社团发现的测试算法。

该类算法可视为数据聚类算法，其主要任务是挖掘出数据中聚类的类别。我们设计的生成内容算法依照 LFR 基准数据集中的社团进行节点内容的生成，不同 LFR 基准数据所包含的内容信息均为同一系列数据，不会因链接的设置而变化。如图 3、图 4 所示，MAC 算法和 DP 算法的 NMI 精度曲线表现为一条直线。MAC 算法是基于生成模型构建的，而网络中节点内容生成算法与其原理相近。因此，MAC 算法的表现优于 DP 算法。DP 算法表现不佳的原因是，其对随机产生的数据不敏感。此外，MAC 算法的时间复杂度为 $O(nmC)$，这里 m 为数据中非零元素的个数，m 远远大于 n，那么 MAC 的时空复杂度大于 $O(n^2C)$。DP 算法的时间复杂度为 $O(n^2)$。因此，MAC 算法在大规模数据集上（$n=5000$）的聚类精度下降相对明显，同时 DP 算法的聚类表现有所波动。另外，从图中可以看到的是，基于节点内容的方法结果基本低于基于网络拓扑的方法，这主要是因为实验中拓扑和内容都是独立生成的，且生成的内容信息在社团结构上的模糊程度低于拓扑信息，这与现实世界中真实网络上的情形是基本一致的。

3.3 融合网络拓扑和节点内容的社团发现

为了定量地分析融合网络拓扑和内容信息社团发现算法，本小节选择了 5 种特色的算法进行测试，分别为 CESNA 算法[62]、SCI 算法[64]、TADW 算法[67]、FGC 算法[69]、GCC 算法[75]。需要说明的是，在 $n=5000$ 的网络上，我们运行的 GCC 算法存在内存溢出，无法在图 4 中展示 GCC 算法检测社团的性能曲线。

如图 3 所示，本文涉及融合网络拓扑与节点内容的算法中，TADW 算法为基于网络嵌入、融合节点内容的算法。其能够挖掘更为丰富的拓扑结构信息，融合节点内容信息后，具有更强的表征社团结构的能力，进而检测社团的性能更佳。观察图 3 中不同算法的社团检测精度曲线，TADW 算法比本文测试的所有算法的表现更优。与同属于网络嵌入类型算法 Node2Vec 相比，TADW 融合了内容信息，进而提升了检测社团的性能。这表明了，融合网络拓扑与节点内容信息能够提升社团的检测性能。虽然，在人工网络上，SCI 算法、CESNA 算法的社团检测性能比 Node2Vec 算法稍显逊色，但是比基于网络拓扑的 SBM 算法、Louvain 算法、NMF 算法以及基于

节点内容的 MAC 算法、DP 算法的检测社团能力更优。这同样说明,融合网络拓扑和节点内容更能准确地检测网络中的社团结构。

本文分析了 CESNA 算法、TADW 算法的时间复杂度,分别是 $O(n^2C^2)$,$O(mnC+nC^2)$。这两种结合网络拓扑与节点内容算法的时间复杂度比基于网络拓扑 Node2Vec 算法的时间复杂度 $O(n)$高很多。如图 4 所示,在大规模网络上($n=5000$),与 Node2Vec 算法相比,CESNA 算法、TADW 算法的识别网络中社团结构的性能均有所下降。由于 CESNA 算法的时间复杂度高于 Louvain 算法,CESNA 算法与 Louvain 算法相比性能下降更多。但是,CESNA 算法、TADW 算法比其他的算法,如 SBM 算法、NMF 算法、MAC 算法、DP 算法,仍具有一定的竞争力。这也揭示了融合网络拓扑与节点内容信息能提高模型的识别网络中社团结构的准确性。

从总体来看,基于网络拓扑的方法的性能优于基于节点内容的方法,但融合了网路拓扑和节点内容两种信息之后,方法的性能并非介于两者之间,而是得到了进一步的提升。这是因为模型能够对这二元信息中的有益信息进行充分结合,从而实现社团检测性能的进一步提升。

4 结论以及未来研究方向

本文分别从以下两部分对已有的社团检测研究进行了综述:(1)按照基于网络拓扑的社团检测方法、基于节点内容的社团检测方法、基于结合网络拓扑与节点内容的社团检测方法,进行分类并介绍它们的基本原理,进一步选择了流行的、代表性的算法进行详细介绍;(2)从介绍的 3 类算法中选取一些典型算法,在人工网络上基于检测社团精度的定量对比、算法复杂度分析,进行算法性能对比。

虽然社团检测已经取得不错的成绩,并被成功地应用到不同的研究领域,但社团检测还有诸多问题未被解决。本文给出了一些当前需解决的问题:

(1)关于社团结构具有各种定义,如社团内链接稀疏的观点、社团间链接稠密的广义社团观点,未有统一的界定,诸多社团检测的优劣较难判断。因此,各类社团检测算法需要一种具有统一评价的且有较为清晰界定社团定义的理论框架。

(2)真实网络中社团存在重叠现象,网络拓扑与内容信息不一定具有同质性。目前,既能融合网络拓扑与内容信息又能发现重叠社团结构的社团检测算法鲜有提出。需要一系列综合考虑社团与网络拓扑、内容信息之间内在联系的理论研究,给出了一种新型社团结构。此外,不同类型社团检测算法利用网络中不同源的信息越来越多,基于不同信息或是信息融合的社团检测质量的更为有效地评价算法也有待提出。

参 考 文 献

[1] WANG X F, LI X, CHEN G R. Complex network theory and its applications[M]// Beijing: Tsinghua University Press, 2006.

[2] EULER L. The seven bridges of Königsberg[J]. The world of mathematics, 1956, 1: 573-580.

[3] ERDŐS P, RÉNYI A. On the strength of connectedness of a random graph[J]. Acta Mathematica Hungarica, 1961, 12(1/2): 261-267.

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

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

[6] 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.

[7] HILL R A, DUNBAR R I M. Social network size in humans[J]. Human Nature, 2003, 14(1): 53-72.

[8] FORTUNATO S. Community detection in graphs [J]. Physics Reports, 2010, 486(3/4/5): 75-174.

[9] JIN D, YOU X X, LIU Y S, et al. Stuctural Feature Enhanced Markov Random Field for Community Detection in Large-Scale networks[J]. Chinese Journal of Computer. 2019, 40(12): 2822-2835.

[10] ZHANG L, LIU Q, YANG S S, et al. A Dual Representation-Based Multi-Objective Evolutionary Algorithm for Overlapping Community Detection [J]. Acta Electronica Sinica, 2021, 49(11): 2101-2107.

[11] DOURISBOURE Y, GERACI F, PELLEGRINI M. Extraction and classification of dense communities in the web[C]// Proceedings of the 16th International Conf. on World Wide Web. ACM, 2007: 461-470.

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

[13] RIVES A W, GALITSKI T. Modular organization of cellular networks[J]. Proceedings of the National Academy of Sciences, 2003, 100(3): 1128-1133.

[14] SPIRIN V, MIRNY L A. Protein complexes and functional modules in molecular networks[J]. Proceedings of the National Academy of Sciences, 2003, 100(21): 12123-12128.

[15] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297.

[16] GLARIA F, HERNÁNDEZ C, LADRA S, et al. Compact structure for sparse undirected graphs based on a clique graph partition[J]. Information Sciences, 2021, 544: 485-499.

[17] DONETTI L, MUNOZ M A. Detecting network communities: a new systematic and efficient algorithm[J]. Journal of Statistical Mechanics: Theory and Experiment, 2004, 2004(10): P10012.

[18] WU C R, PENG Q L, LEE J, et al. Effective hierarchical clustering based on structural similarities in nearest neighbor graphs [J]. Knowledge-Based Systems, 2021, 228: 107295.

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

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

[21] FAYSAL M A M, ARIFUZZAMAN S. Distributed community detection in large networks using an information-theoretic approach[C]// Proceedings of 2019 IEEE International Conference on Big Data. IEEE, 2019: 4773-4782.

[22] HOLLAND P W, LASKEY K B, LEINHARDT S. Stochastic blockmodels: First steps[J]. Social Networks, 1983, 5(2): 109-137.

[23] KARRER B, NEWMAN M E J. Stochastic blockmodels and community structure in networks[J]. Physical Review E, 2011, 83(1): 016107.

[24] JIN D, CHEN Z, HE D X, et al. Modeling with node degree preservation can accurately find communities[C]// Proceedings of the 29th AAAI Conf. on Artificial Intelligence. 2015.

[25] CUI P, WANG X, PEI J, et al. A survey on network embedding [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(5): 833-852.

[26] LIU F Z, XUE S, WU J, et al. Deep learning for community detection: progress, challenges and opportunities[J]. arXiv: 2005.08225, 2020.

[27] SU X, XUE S, LIU F Z, et al. A comprehensive survey on community detection with deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33: 6737-6748.

[28] NEWMAN M E J, CLAUSET A. Structure and inference in annotated networks [J]. Nature Communications, 2016, 7(1): 1-11.

[29] 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): 1147-1170.

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

[31] MAJI S, VISHNOI N K, MALIK J. Biased normalized cuts [C]// Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition. 2011: 2057-2064.

[32] CHEW S E, CAHILL N D. Semi-supervised normalized cuts for image segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1716-1723.

[33] SUN Z J, SUN Y N, CHANG X F, et al. Community detection based on the Matthew effect[J]. Knowledge-Based Systems, 2020, 205: 106256.

[34] SUN P G. Weighting links based on edge centrality for community detection[J]. Physica A: Statistical Mechanics and Its Applications, 2014, 394: 346-357.

[35] LI T X, LEI L H, BHATTACHARYYA S, et al. Hierarchical community detection by recursive partitioning[J]. Journal of the American Statistical Association, 2022, 117(538): 951-968.

[36] 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, 2008(10): P10008.

[37] QIAO S J, HAN N, GAO Y J, et al. A fast parallel community discovery model on complex networks through approximate optimization[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1638-1651.

[38] KIRKPATRICK S, GELATT C D, VECCHI M P. Optimization by simulated annealing[J]. Science, 1983, 220(4598): 671-680.

[39] AIROLDI E M, BLEI D M, FIENBERG S E, et al. Mixed membership stochastic blockmodels[J]. Journal of Machine Learning Research, 2008, 9: 1981-2014.

[40] HE C B, ZHANG Q, TANG Y, et al. Community detection method based on robust semi-supervised nonnegative matrix factorization[J]. Physica A: Statistical Mechanics and its Applications, 2019, 523: 279-291.

[41] LUO X, LIU Z G, JIN L, et al. Symmetric nonnegative matrix factorization-based community detection models and their convergence analysis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(3): 1203-1215.

[42] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.

[43] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations [C]// Proceedings of the 20th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining. 2014: 701-710.

[44] LONG Q Q, WANG Y M, DU L, et al. Hierarchical community structure preserving network embedding: A subspace approach [C]// Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 409-418.

[45] WANG D X, CUI P, ZHU W W. Structural deep network embedding[C]// Proceedings of the 22nd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining. 2016: 1225-1234.

[46] XU R B, CHE Y, WANG X M, et al. Stacked autoencoder-based community detection method via an ensemble clustering framework[J]. Information sciences, 2020, 526: 151-165.

[47] CAI B, WANG Y P, ZENG L N, 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.

[48] JIA Y T, ZHANG Q Q, ZHANG W N, et al. Communitygan: Community detection with generative adversarial nets[C]// Proceedings of the World Wide Web Conference. 2019: 784-794.

[49] CHEN J Y, GONG Z G, MO J Q, et al. Self-training enhanced: Network embedding and overlapping community detection with adversarial learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33: 6737-6748.

[50] RODRIGUEZ A, LAIO A. Clustering by fst search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[51] CHEN Y W, HU X L, FAN W T, et al. Fast density peak clustering for large scale data based on kNN[J]. Knowledge-Based Systems, 2020, 187: 104824.

[52] YU H, CHEN L Y, YAO J T. A three-way density peak clustering method based on evidence theory [J]. Knowledge-Based Systems, 2021, 211: 106532.

[53] XU X, DING S F, WANG Y R, et al. A fast density peaks clustering algorithm with sparse search[J]. Information Sciences, 2021, 554: 61-83.

[54] HOFMANN T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999: 50-57.

[55] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[56] STREICH A P, FRANK M, BASIN D, et al. Multi-assignment clustering for boolean data[C]// Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 969-976.

[57] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed rep-

resentations of words and phrases and their compositionality [C]// Proceedings of Advances in Neural Information Processing Systems. 2013:3111-3119.

[58] YANG Y,WANG H G,ZHU J Q,et al. Dataless short text classification based on biterm topic model and word embeddings [C]// Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. 2021:3969-3975.

[59] FARD M M,THONET T,GAUSSIER E. Deep k-means:Jointly clustering with k-means and learning representations[J]. Pattern Recognition Letters,2020,138:185-192.

[60] RUAN Y,FUHRY D,PARTHASARATHY S. Efficient community detection in large networks using content and links [C]// Proceedings of the 22nd International Conference on World Wide Web. 2013:1089-1098.

[61] BU Z,LI H J,CAO J,et al. Dynamic cluster formation game for attributed graph clustering[J]. IEEE transactions on cybernetics,2017,49(1):328-341.

[62] YANG J,MCAULEY J,LESKOVEC J. Community detection in networks with node attributes[C]// Proceedings of IEEE 13th International Conf. on Data Mining. IEEE,2013:1151-1156.

[63] BALASUBRAMANYAN R,COHEN W W. Block-LDA:Jointly modeling entity-annotated text and entity-entity links[C]// Proceedings of the 2011 SIAM International Conf. on Data Mining. Society for Industrial and Applied Mathematics. 2011:450-461.

[64] WANG X,JIN D,CAO X C,et al. Semantic community identification in large attribute networks[C]// Proceedings of the AAAI Conf. on Artificial Intelligence. 2016.

[65] HU L,CHAN K,YUAN X H,et al. A variational Bayesian framework for cluster analysis in a complex network[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(11):2115-2128.

[66] JIN D, WANG K Z, ZHANG G, et al. Detecting communities with multiplex semantics by distinguishing background,general, and specialized topics[J]. IEEE Transactions on Knowledge and Data Engineering,2019,32(11):2144-2158.

[67] YANG C,LIU Z Y,ZHAO D L,et al. Network representation learning with rich text information[C]// Proceedings of the 24th International Joint Conference on Artificial Intelligence. 2015.

[68] NATARAJAN N,DHILLON I S. Inductive matrix completion for predicting gene-disease associations [J]. Bioinformatics, 2014,30(12):i60-i68.

[69] KANG Z,LIU Z Y,PAN S R,et al. Fine-grained Attributed Graph Clustering[C]// Proceedings of the 2022 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. 2022:370-378.

[70] SUN X F,GUO J,DING X,et al. A general framework for content-enhanced network representation learning[J]. arXiv:1610.02906,2016.

[71] FAN S H,WANG X,SHI C,et al. One2multi graph autoencoder for multi-view graph clustering[C]// Proceedings of The Web Conference 2020. 2020:3070-3076.

[72] YANG L,WANG Y Y,GU J H,et al. JANE:Jointly Adversarial Network Embedding [C]// Proceedings of IJCAI. 2020: 1381-1387.

[73] CUI G Q,ZHOU J,YANG C,et al. Adaptive graph encoder for attributed graph embedding[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020:976-985.

[74] ZHENG Y P,ZHANG X F,CHEN S Y,et al. When Convolutional Network Meets Temporal Heterogeneous Graphs:An Effective Community Detection Method[J]. IEEE Transactions on Knowledge and Data Engineering,2023,35(2):2173-2178.

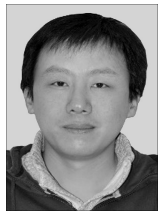
[75] FETTAL C,LABIOD L,NADIF M. Efficient Graph Convolution for Joint Node Representation Learning and Clustering [C]// Proceedings of the 15th ACM International Conference on Web Search and Data Mining. 2022:289-297.

[76] WANG C,PAN S R,HU R Q,et al. Attributed graph clustering:A deep attentional embedding approach[J]. arXiv:1906.06532,2019.

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

[78] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.

CAO Jinxin, born in 1987, Ph.D, lecturer, is a member of China Computer Federation. His main research interests include data mining, machine learning, community detection.



JIN Di, born in 1981, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include data mining, complex network analysis, machine learning.