角色导向的网络表示学习综述

焦鹏飞[1] 潘 婷[2] 金 弟[2] 王文俊[2] 何东晓[2] 高梦州[1] 赵治栋[1]

[1](杭州电子科技大学网络空间安全学院 杭州 310018)

[2](天津大学智能与计算学部 天津 300350)

摘 要 网络表示学习是一种将网络节点映射到低维、连续的实值向量空间上的技术,它在网络分析中发挥着重要作用.社团导向的网络表示学习作为目前研究的主要分支之一,主张在学习的节点表示中保持自身的社团属性,如节点的邻近性,使得相近节点具有相似表示.这类方法虽然可以挖掘现实系统中具有明显聚集特征的实体集合,但因其未考虑节点结构上的相似性,导致它们无法识别扮演相同角色、发挥类似功能的实体.近些年,一些方法结合角色的概念,利用节点在网络中的连接模式来派生节点表示,这使得学习到的表示可以尽可能地保持原始网络中节点的结构相似性.尽管这种面向角色的网络表示学习对于现实场景的分析及网络科学的发展起到了一定推动作用,但是目前对该领域的研究仍然非常有限,已有工作缺乏统一的理论解释和实验比较.本文主要对近年来角色导向的网络表示学习工作进行了系统性综述:首先,本文结合相关概念及理论知识,分析了社团导向和角色导向网络表示学习的区别;接着,在总结现有角色导向网络表示学习方法的基础上,给出了一种全新的分类方式,以把握不同算法的本质原理;随后,本文在具有社团或角色标签的十个实验数据集上对基于社团或角色的算法进行了可视化、节点分类、聚类、鲁棒性分析和参数敏感性分析实验,以此横向比较社团与角色这两个重要概念的内在区别,纵向评估角色导向网络表示学习方法在不同学习机制下的性能差异;此外,为进一步推动该领域的深入发展,本文提供了一个集数据、算法、分析于一体的角色导向的网络表示学习平台,服务于该领域的后续研究;最后,本文对角色导向的网络表示学习面临的挑战和未来发展趋势进行了总结和展望.

关键词 网络角色;网络表示学习;结构相似性;分类体系;综合性实验

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2023.00274

A Survey on Role-Guided Network Representation Learning

JIAO Peng-Fei[1] PAN Ting[2] JIN Di[2] WANG Wen-Jun[2] HE Dong-Xiao[2]

GAO Meng-Zhou[1] ZHAO Zhi-Dong[1]

[1](*School of Cyberspace, Hangzhou Dianzi University, Hangzhou* 310018)

[2](*College of Intelligence and Computing, Tianjin University, Tianjin* 300350)

Abstract Network representation learning is a technique that maps network nodes into low-dimensional and continuous real vector space, which plays a critical role in network analysis. As one of the main branches of network representation learning research, community-guided methods advocate preserving the own community attributes of nodes in the learned representations, such as the proximity which makes adjacent nodes have similar representations. This class of methods can mine entity sets with obvious clustering characteristics in real-world systems, which however,

收稿日期:2022-02-16;在线发布日期:2022-08-12.本课题得到国家自然科学基金(61902278,62003120)、浙江省省属高校基本科研业务费专项(GK229909299001-008)、浙江省自然科学基金重大项目(LDT23F0101, LDT23F01012F01, LDT23F01015F01)资助.**焦鹏飞**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为复杂网络分析及其应用.E-mail: pjiao@hdu.edu.cn.**潘 婷**,硕士研究生,主要研究领域为角色导向的网络表示学习.**金 弟**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为网络表示学习.**王文俊**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为复杂网络分析和数据挖掘.**何东晓**,博士,副教授,主要研究方向为复杂网络分析.**高梦州**(通信作者),博士,讲师,主要研究方向为复杂网络分析与网络控制.E-mail: mzgao@hdu.edu.cn.**赵治栋**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能与信号系统.

without consideration of structural similarity, fail to identify entities that play the same role or perform the similar function in the network. In recent years, some works introduce the concept of roles to derive node representations by the connectivity patterns of nodes, which makes the learned representations preserve the structural similarity in original network as much as possible. The role-guided network representation learning promotes the development of network science and the analysis of real-world scenarios to a certain extent. However, the correlational research in this field is still limited and the existing works lack unified theoretical explication and experimental comparison. This paper mainly makes a systematic review of the role-guided network representation learning. First of all, based on relevant concepts and theoretical knowledge, this paper analyzes the differences between community-guided network representation learning and role-guided network representation learning. Next, on the basis of summarizing existing role-guided network representation learning methods, we propose a new taxonomy to grasp the essential principle of different algorithms. Furthermore, typical community-guided and role-guided methods are utilized to conduct comprehensive experiments including visualization, node classification, clustering, robustness analysis and parameter sensitivity analysis on ten common datasets with community labels or role labels. On the one hand, these experiments horizontally compare the intrinsic differences between the concept of community and role. On the other hand, they longitudinally evaluate the performance difference of role-guided network representation learning methods under different learning mechanisms. In addition, we build a role-guided network representation learning platform to further promote development of this field, which integrates common datasets, the state-of-the-art algorithms and various network analysis tasks to serve follow-up research. At last, we summarize the challenges and prospect the future development trends in role-guided network representation learning.

Keywords　network role; network representation learning; structural similarity; taxonomy; comprehensive experiments

1　引　言

复杂网络是一种用节点描述实体，用边刻画实体间交互关系的抽象模型[1]. 它与现实世界中各类复杂性系统如用户社交网络、论文引用网络、航空交通网络和生物代谢网络的研究存在密切关联[2]. 研究学者致力于从网络数据中捕获系统形成和演化的内在规律，以分析和预测实体的行为模式[3-5]. 然而，随着网络规模指数级的增长，传统基于邻接矩阵的网络表征方式因无法在大规模数据上扩展而不再适用，探索更为高效的方法成为必然[6].

近些年，网络表示学习(Network Representation Learning)以其强大且灵活的表征能力成为研究复杂网络的主要范式. 网络表示学习，又称作网络嵌入(Network Embedding)，它将网络表征视为机器学习任务本身，使用数据驱动的方法学习网络节点的编码表示. 直观来说，网络表示学习的目标是学习一种映射函数，该函数在将网络节点映射到低维、连续的实值向量空间上的同时，最大程度地保留原有网络中的结构特征和统一特性，以保证学习到的低维表示可以重构原始的网络结构或其它特征，并支持后续的网络分析任务，例如节点分类[7-8]、节点聚类[9]、链接预测[10-11]、异常检测[12]、网络对齐[13]和可视化分析[14-15]等. 网络表示学习一方面弥补了传统表征方式的不足，可以在较短时间内处理大规模网络，具有可扩展性和适应性；另一方面，它借助各种复杂或深度的算法模型实现对高度非线性结构的捕捉，为解决真实世界网络稀疏问题提供有力支撑. 近些年一系列基于矩阵分解[16-18]、随机游走[19]和图卷积神经网络[20]的网络表示学习研究方法和文献综述[21-25]被接连提出.

社团作为复杂网络最基本的特征之一，被定义为一组内部连接紧密、外部连接稀疏的节点集合，广泛地存在于各类现实系统中[26-27]. 例如在社交网络中，社团表示因工作主题或兴趣爱好相同而联系紧密的用户人群；在引文网络中，社团表示因内容相关而建立引用关系的论文集合；在基因调控网络中，社团表示因基因表达而产生协同作用的功能单元. 同属一个社团的节点不仅具有紧密的连接关系和邻近关系，还存在某种相似性关系. 基于社团的网络表示学习旨在结合“社团”的概念编码这种相似性. 换言之，社团导向网络表示学习倾向于在学习的低维表示中保留节点自身的社团属性，使得相近或连接紧密的节点能以相似的向量表示. 由于挖掘网络中的社团信息有助于发现节点间的作用关系，捕获异常的行为模式[28-29]，乃至预测未观察到的链接，面向社团的网络表示学习得到了快速发展.

然而，社团导向网络表示学习方法有其局限性，即它无法挖掘网络中更复杂的结构特征和异质信息. 这是因为社团导向的节点嵌入只考虑节点间的邻近性，忽略了远距离节点之间的作用关系[30-32]，造成全局信息的缺失. 同时，在高度异质的复杂网络中，节点特征与其邻居相关性很低——共享同一类特征的节点往往不相连，甚至不在同一个社团中，导致面向社团的方法不适用. 此外，由于社团的语义信息仅存在于特定网络，基于社团的嵌入算法不能用于迁移学习，极大地限制了算法的泛化能力. 为了解决这些问题，研究学者注意到了“角色”的重要性.

“角色”的概念虽然在社会学和社会心理学的研究中有着漫长的发展历史，但它在复杂网络中的应用仅有十几年的时间. 复杂网络中的角色[33-34]是一种重要的、与社团不同但互补的概念——社团关注节点的邻近性，而角色取决于节点连接模式的相似性. 其背后的假设认为网络中具有类似连接模式的节点往往在身份地位或功能作用上表现出某种相似性关系，例如具有较大度特征的中心节点、连接网络不同区域的桥节点等. 基于角色的网络表示学习主张在嵌入中编码这种相似性，使得扮演相同角色、发挥相同作用的节点能具有相似的向量表示. 这种由角色驱动的表征方式又被称为角色导向的网络表示学习或面向角色的网络表示学习. 基于角色的网络表示学习作为网络分析的一类新型技术，深化了网络科学的研究范畴：

(1) 角色的引入改变了以往基于邻近性挖掘网络数据信息的思维定势. 角色导向的研究通过比较节点连接模式的相似性来捕捉复杂网络中的节点差异和重要模式，反映现实系统中对应实体的功能作用或身份地位. 这类方法相比于社团导向的方法更适用于具有多类型对象的现实网络的探索性分析.

(2) 面向角色的网络表示学习泛化了网络的研究任务. 角色捕获了网络中更一般化的概念，使得学习到的网络表示学习模型可以在网络的不同部分，甚至是不同网络中泛化，让跨网络的迁移学习成为可能，这极大地丰富了分类、聚类、可视化、链接预测、异常检测、网络对齐等的实现方法和任务体系.

(3) 面向角色的网络表示学习与社会应用密切相关. 基于角色的网络表示允许我们从复杂的网络系统中识别节点的特定角色，进而指导现实世界的生产生活. 如识别社交网络中拥有大群粉丝的公众人物，对舆情的正确把控起着关键的引导作用；识别邮件网络中用户的专业角色，对于联系人的精准推荐具有重要的支持作用.

角色导向网络表示学习作为一种与社团导向网络表示学习相互补充的概念，也有其局限性. 由于角色导向的方法更注重于用网络拓扑结构来描述节点的连接模式，忽略了节点间的邻近性，导致绝大多数方法学习到的节点嵌入不适用于像链接预测这样的下游任务. 因此角色导向网络表示学习和社团导向网络表示学习在面对不同的应用场景和下游任务时，发挥着不同的重要作用.

尽管面向角色的网络表示学习具有多样化的研究价值和应用价值，但其目前仍处于发展的初步阶段，许多核心问题还未被突破，主要体现在以下几个方面：第一，基于角色的网络表示学习问题缺乏全面而有针对性的论述. 当前关于网络表示学习的工作被接连提出，但未有工作对角色导向网络表示学习进行完整梳理和明确论述；第二，基于角色的网络表示学习方法缺乏更深层次的分类方式. 当前对这类研究方法的划分依然参考社团导向的分类机制，然而由于角色信息的保持不仅要考虑邻近的节点信息，还涉及全局和高阶特征的捕捉，单从技术角度划分算法类型是不准确的；第三，基于角色的网络表示学习方法缺乏综合性实验评估. 近年来角色导向网络表示学习方法频出，但在许多工作中仍仅使用某几个典型的算法实例和有限的数据集进行比较，缺少与前沿算法的对照且没有统一全面的评估，导致我们不能直观地分析和认识这些算法的性能差异，

无法针对不同的任务提供有效的算法支持.

针对以上问题，本文围绕近些年的工作对角色导向网络表示学习研究进行了综述. 具体地，本文首先回顾了网络表示学习的相关概念，全面总结与分析了社团导向网络表示学习与角色导向网络表示学习的异同. 其次，本文从特征提取的角度出发，提出一种新的分类方式，作为理解和分析角色导向网络表示学习方法的理论依据，并在此基础上对当前一些代表性方法进行了系统性描述，展示了不同方法下的嵌入机制. 接着，本文详细介绍了该领域中常用的实验数据集，并提供了包括节点分类、聚类和可视化在内的完整性实验. 为深入探究不同方法的鲁棒性和参数敏感性，本文在不同的扰动机制和节点嵌入维度上做了进一步实验. 同时，为进一步阐述角色和社团二者之间的区别，本文还提供了具有社团标签的实验数据集和社团导向的实验方法，以得到更为直观的认识和见解. 此外，为推动该领域的深入发展，本文给出了一个集数据、算法、分析于一体的角色导向网络表示学习算法平台，服务于该领域的后续研究. 最后，本文从宏观层面阐述了当前角色导向网络表示学习的发展脉络，并对其面临的挑战和未来发展趋势进行了总结和展望，从而为今后的研究提供有价值的参考.

事实上，目前已有一些综述文献介绍了与角色或网络表示学习有关的发展情况，但这些工作与本文的侧重点存在明显差异. 2017 年张树森等人[35]就社会网络的角色识别问题做了详细概括，但该工作与网络表示学习不相关. 同年 Hamilton 等人[36]重点回顾了在机器学习和数据挖掘领域引起极大关注的节点嵌入算法和子图嵌入算法，并在此过程中开发统一的框架来描述这些方法，2019 年 Cui 等人[23]针对网络嵌入问题提供详细调查，并指出该领域的未来研究方向，但值得注意的是，这些文献对于结构或角色的方法论述寥寥无几. 2020 年 Zhang 等人[24]在对网络表示学习做系统性综述时给出了与角色有关的邻近性概念，并对其做出明确定义和说明，但他们在分类机制中未对社团和角色导向网络表示学习方法做出区分. 同年 Rossi 等人[37]对比分析了基于网络邻近性和基于结构或角色的网络嵌入方法，但是该工作对角色导向网络表示学习的论述依旧是不全面的，例如缺少对深度学习方法的讨论，也未对其复杂多样的符号、动机乃至模型进行形式化上的统一. 相对于以往的研究综述，本文关注于角色导向网络表示学习的内涵与方法体系，补充和完善了该领域的研究框架与思路.

本文的主要贡献如下：

(1) 本文是首个针对角色导向网络表示学习进行综述的中文论文，内容涵盖了全新的分类方式、综合性实验以及面临的挑战与未来展望. 在分类方式上，本文从特征提取的角度出发，将现有研究划分为基于局部特征、全局特征和高阶特征的方法，以解释不同嵌入下的生成机制.

(2) 本文提供综合性实验对主流的角色导向网络表示学习方法进行了多方面评估，并为探究这些方法的鲁棒性问题设计了不同层面的扰动机制来测试各类算法的性能. 同时为进一步阐述社团与角色这两个重要概念的不同，本文还增加了具有社团标签的实验数据集和相关算法，以供二者进行横向比较.

(3) 本文提供了一个集数据、算法、分析于一体的角色导向的网络表示学习算法平台，服务于该领域的后续研究.

2　问题定义

在本节中，我们就网络表示学习的相关概念给出明确的问题定义，然后从理论层面概述社团导向和角色导向网络表示学习之间的差异.

对于给定网络 $\mathcal{G}=(\mathcal{V},\mathcal{E})$，$\mathcal{V}=\{v_1,v_2,\cdots,v_{|\mathcal{V}|}\}$ 表示节点集合，$\mathcal{E}\subseteq\mathcal{V}\times\mathcal{V}$ 表示连边集合，$|\mathcal{V}|$ 是节点数，$|\mathcal{E}|$ 是连边数. 若存在 $e_{ij}=(v_i,v_j)\in\mathcal{E}$，那么节点 v_i 和 v_j 之间存在相应连边，权重大小记为 $\boldsymbol{W}_{ij}$. 在无权网络中，若节点 v_i 和 v_j 之间存在连边，则有 $\boldsymbol{W}_{ij}=1$，否则 $\boldsymbol{W}_{ij}=0$. 网络 $\mathcal{G}$ 的邻接矩阵用 $\boldsymbol{A}=(\boldsymbol{A}_{ij})_{|\mathcal{V}|\times|\mathcal{V}|}$ 表示，其中 $\boldsymbol{A}_{ij}=\boldsymbol{W}_{ij}$. 对于无向网络而言，邻接矩阵具有对称性，即满足 $\boldsymbol{A}_{ij}=\boldsymbol{A}_{ji}$，而在有向网络中则不然. 在已知网络邻接矩阵的情况下，可以得到节点的邻居信息和度信息. $\mathcal{N}_i^k$ 表示与节点 v_i 距离 k 步的节点集合，例如当 $k=1$ 时，$\mathcal{N}_i^1$ 表示与节点 v_i 直接相连的节点集合. 节点 v_i 的度 $\boldsymbol{D}_{ii}$ 表示与其直接相连的邻居节点的数量，即 $\boldsymbol{D}_{ii}=\sum_{j=1}^{|\mathcal{V}|}\boldsymbol{A}_{ij}$. 而在有向网络中，节点度又被细分为入度和出度. 入度 $\boldsymbol{D}_{ii}^-$ 是指向节点 v_i 的边数量，出度 $\boldsymbol{D}_{ii}^+$ 则表示从节点 v_i 出发的边数量. 相对应的，节点集合 $\mathcal{N}_i^k$ 也可被划分为 $\mathcal{N}_i^{k+}$ 和 $\mathcal{N}_i^{k-}$. 还有一些网络，具有节点的属性信息，将其表示为 $\boldsymbol{X}\in\mathbb{R}^{|\mathcal{V}|\times m}$，$m$ 是属性维度. 详细符号定义在表 1 中列出.

表 1 符号定义

符号	表示含义
$\mathcal{G}=(\mathcal{V},\mathcal{E})$	由点集$\mathcal{V}$和边集$\mathcal{E}$构成的网络
$\boldsymbol{W}_{ij}$	节点 v_i和 v_j的连边权重
$\mathcal{H}$	网络基元集合，$\mathcal{H}=\{\mathcal{H}_1,\mathcal{H}_2,\cdots,\mathcal{H}_T\}$
$\mathcal{H}_t=(\mathcal{V}_k,\mathcal{E}_k)$	网络基元$\mathcal{H}_t$，$\mathcal{V}_k\subset\mathcal{V}$，$\mathcal{E}_k\subset\mathcal{E}$
$\mathcal{W}_t\in\mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	网络基元$\mathcal{H}_t$的加权邻接矩阵
$\boldsymbol{I}\in\mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	单位矩阵
$\boldsymbol{A}\in\mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	邻接矩阵
$\boldsymbol{D}\in\mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	度矩阵，对角线元素为节点 v_i的度 $\boldsymbol{D}_{ii}$
$\boldsymbol{L}\in\mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	拉普拉斯矩阵，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$
$\boldsymbol{X}\in\mathbb{R}^{\lvert\mathcal{V}\rvert\times m}$	属性矩阵，m 为属性信息的维度
$\boldsymbol{F}\in\mathbb{R}^{\lvert\mathcal{V}\rvert\times f}$	特征矩阵，$\boldsymbol{F}_i$为节点 v_i的特征向量
$\boldsymbol{Z}\in\mathbb{R}^{\lvert\mathcal{V}\rvert\times d}$	嵌入矩阵，$\boldsymbol{Z}_i$为节点 v_i的嵌入向量
$\boldsymbol{S},\tilde{\boldsymbol{S}}$	相似性矩阵及其近似矩阵
$\mathcal{N}_i^k$	与节点 v_i距离 k 步的节点集合
$\mathcal{N}_{i,d}^k$	$\mathcal{N}_i^k$中度为 d 的节点集合
$\mathcal{R}_i^k$	与节点 v_i距离小于等于 k 的节点集合
$\mathcal{G}_i^k$	与节点 v_i距离小于等于 k 的邻域子图
$\boldsymbol{F}_i^k$	$\mathcal{N}_i^k$的统计特征向量，$\boldsymbol{F}_{i,l}^k$为第 l 个特征值
δ	不同跳距的聚合权重 $\delta\in(0,1]$
$\mathcal{P}_{i\to j}^k$	从节点 v_i到 v_j的路径，$v_j\in\mathcal{N}_i^k$
$\mathcal{L}$	损失函数

定义 1. 网络表示学习. 又称为网络嵌入. 对于给定网络$\mathcal{G}=(\mathcal{V},\mathcal{E})$，其目标是学习一个映射函数 $f:\mathcal{V}\to\mathbb{R}^d$，使得网络中的每个节点都被映射到一个低维稠密的 d 维嵌入向量 $\boldsymbol{z}$ 上，其中 $d\ll|\mathcal{V}|$. 网络表示学习的关键在于嵌入向量 $\boldsymbol{z}$ 应尽可能保持网络$\mathcal{G}$的拓扑结构和属性特征，使得原始网络中具有某种相似性关系的节点在映射空间中仍能保持相近的距离.

定义 2. 邻近性. 设$\mathcal{P}$是衡量节点邻近性的函数. 若$\mathcal{P}(v_i,v_j)>\mathcal{P}(v_i,v_k)$，则表示节点 v_i与 v_j在距离上更相近. 具体地，两个彼此相连的节点相比于不直接连通的节点具有更高的邻近性.

定义 3. 结构相似性. 设$\mathcal{S}$是衡量节点连接模式相似性的函数，若$\mathcal{S}(v_i,v_j)>\mathcal{S}(v_i,v_k)$，则表示节点 v_i和 v_j在结构上更相近. 例如两个具有大量邻居节点的中心节点相比于连接不同区域的桥节点具有更高的结构相似性.

定义 4. 社团导向的网络表示学习. 设$\mathcal{D}$是衡量节点在嵌入空间中距离远近的函数，$\boldsymbol{z}$ 是节点的嵌入向量. 社团导向的网络表示学习可以形式化定义为学习一个映射函数 $f:\mathcal{V}\to\mathbb{R}^d$，使得$\mathcal{P}(v_i,v_j)>\mathcal{P}(v_i,v_k)$时，满足$\mathcal{D}(z_i,z_j)<\mathcal{D}(z_i,z_k)$. 换言之，社团导向的网络表示学习的目标是在学习节点表征的同时，尽可能地保留原始网络中的邻近性，具体表现为同属一个社团的节点由更相似的嵌入向量表示.

定义 5. 角色导向的网络表示学习. 角色导向的网络表示学习可以形式化定义为学习一个映射函数 $f:\mathcal{V}\to\mathbb{R}^d$，使得$\mathcal{S}(v_i,v_j)>\mathcal{S}(v_i,v_k)$时，满足$\mathcal{D}(z_i,z_j)<\mathcal{D}(z_i,z_k)$. 换言之，角色导向的网络表示学习的目标是在学习节点表征的同时，尽可能地保留原始网络中的结构相似性，具体表现为扮演相同角色的两个节点由更相似的嵌入向量表示.

基于这些定义，我们以图 1 为例说明社团导向和角色导向的网络表示学习的内在差异. 在原始网络[34]中，我们用形状表示节点的社团标签，用颜色表示节点的角色标签. 然后我们分别以社团导向的 node2vec 方法[38]和角色导向的 GraphWave 方法[39]来学习该网络的节点嵌入，并将其可视化. 直观地，在社团导向的嵌入空间中，具有相同形状的节点表现出明显聚集性，而在角色导向的嵌入空间中，具有相同颜色的节点表现出明显聚集性. 也就是说，基于社团的嵌入方法将原始网络中连接紧密的节点嵌入到一起，从而保持节点的社团结构，而基于角色的嵌入方法则将原始网络中连接模式相似的节点嵌入到一起，从而保持节点的角色信息.

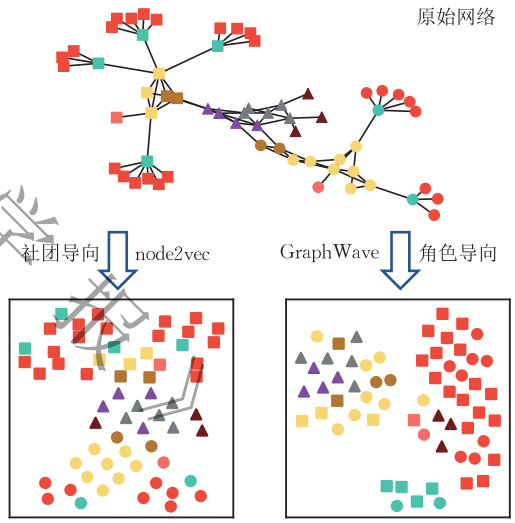

图 1 社团导向和角色导向网络表示学习示例

具体地，我们从以下四个方面总结了角色导向网络表示学习和社团导向网络表示学习的区别：

(1) 研究机制不同.

虽然社团和角色都可以看作是网络聚类的一种具体实例，但由于它们所关注的节点特征不同，使得划分的节点集合表现出显著差异. 社团是基于节点邻近性的概念，使得集合内部节点的连接关系相比于其与集合外部节点的连接关系更为紧密，例如集合内部节点的连边概率大于集合内部节点与外部节点间的连边概率. 而角色是基于节点结构相似性的概念，使得集合内部节点的结构相似性高于其与集合外部节点的结构相似性，例如集合内部节点可能

具有更相近的度大小或节点中心性.

(2)统计模式不同.

社团的提出依赖于网络的同质性假设,即节点的连接关系与其邻居节点存在高度的关联性,这使得基于社团的嵌入方法可以较好地拟合同质性网络的拓扑结构,但此方法在高度异质的网络上往往表现得不尽如人意.这是因为异质网络通常融合更多类型的对象,以及复杂的交互关系,使得直接相连的节点并不具有相似的属性特征.例如由医生、病人及其之间的看护关系构成的医疗网络,其链接由节点的功能角色派生,而与属性的相似性无关.角色的提出在于比较节点间的连接模式,而不受同质性假设的约束,使得它更适用于同质性较低甚至是异质的网络.

(3)方法设计不同.

社团和角色在研究机制和统计模式上的不同,决定了二者方法设计的不同.对于随机游走来说,基于社团的嵌入方法通常将目标节点的邻居节点视为上下文序列,用于保持节点间的邻近性关系,但基于角色的嵌入方法通常将与目标节点具有相似连接模式的其它节点视为上下文序列,进而保持节点的结构相似性关系.对于矩阵分解来说,社团导向的嵌入方法通常将网络的邻接矩阵或基于邻接矩阵派生的拉普拉斯矩阵以及概率转移矩阵作为待分解矩阵,从而获得含有社团信息的节点嵌入,而角色导向的嵌入方法往往通过构造节点结构特征的相似性矩阵来学习含有角色信息的节点嵌入.

(4)应用场景不同.

虽然社团和角色的目标都是在高度复杂的网络中挖掘有用的数据信息,但二者的应用场景存在本质差异.在网络分析中,社团旨在挖掘具有明显聚集性特征的实体集合,而角色旨在挖掘扮演相同角色、发挥类似功能的实体集合.例如在学术合作网络中,社团注重发现同属一个研究领域的相关学者,而角色倾向探索具有相似身份或影响力的相关学者;在社交网络中,社团注重分析具有共同爱好的用户,而角色更倾向于捕捉具有相似传播能力的用户.

3　角色导向的网络表示学习

在本节中,我们针对现有角色导向的网络嵌入方法提出了一种面向特征的分类方式,它将已有工作分为基于局部特征、全局特征和高阶特征的方法.所谓局部特征,是指节点在有限距离邻域内表现出的结构特征,这类特征通常反映了节点在局部环境中的功能或地位,例如企业员工在公司某部门中的职位;全局特征是指节点在全局视角下表现出的结构特征,它通常反映了节点在整个网络中的功能或地位,例如通信网络中连通不同区域的桥节点;高阶特征是对节点特定结构(如网络基元)的详细描述,比作为描述网络构成的基本单元的节点更能体现其功能性.基于这种分类体系,我们可以直观地理解每种角色导向网络表示学习方法的本质原理.我们在图 2 中总结了每种机制下的代表性方法,并在此基础上简要描述了这些方法的研究动机和核心设计.

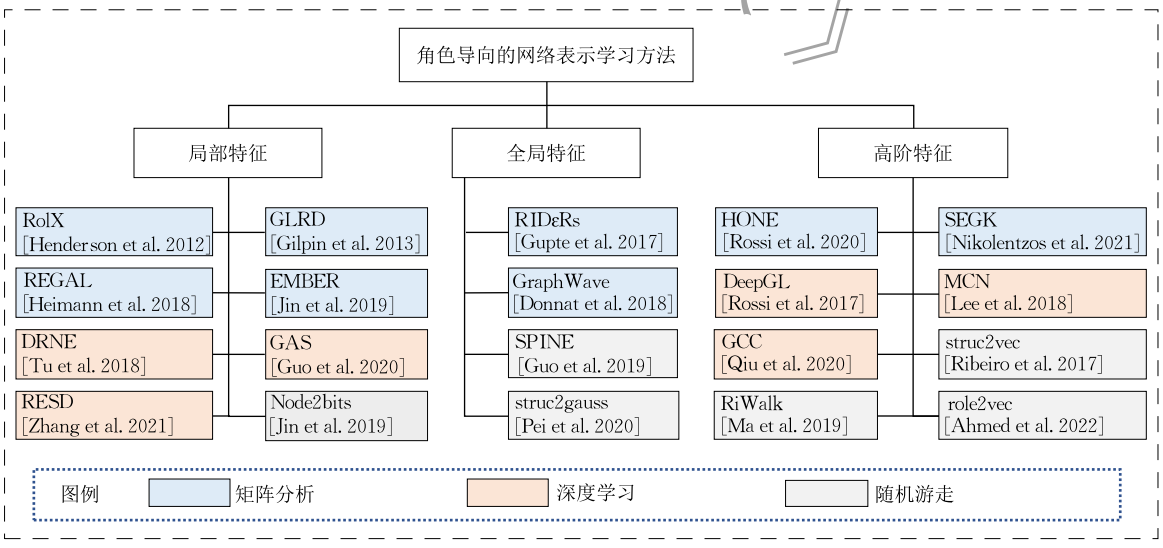

图 2　角色导向的网络表示学习方法的分类方式

3.1　基于局部特征的角色导向的网络表示学习方法

定义 6.　局部特征.局部特征派生于节点的局部拓扑结构.它通过抽取节点与其有限距离邻域(k阶邻域)的连接关系来简化表达高维空间中不同节点间的局部依赖关系.节点度作为最基本的局部特征之一,它度量了节点一阶邻域($k=1$)的数量.

现有角色导向的网络嵌入方法主要利用节点度或 ReFeX[40] 提取的结构特征来表征其局部拓扑结构. 基于矩阵分解的方法大多采用聚合操作捕获每个节点及其 k 阶邻域内的度信息,然后以此构造节点结构相似性矩阵. 基于深度学习的方法虽然利用各种神经网络模型来自动地学习节点间的依赖关系,但它们通常将节点度或 ReFeX 提取的结构特征作为约束信息,构建模型的损失函数.

RolX[41]. RolX 是一种利用 ReFeX 方法提取网络结构特征的无监督学习方法,它旨在挖掘一组可以表征网络节点结构行为的潜在角色. ReFeX 是一种预定义的特征提取技术. 它递归式地聚合基于节点的局部特征(如节点度)和基于自我中心网络的邻域特征(如自我中心网络的边数),使得中心节点的特征向量保持 k 阶邻域的结构信息. ReFeX 的输出是节点特征矩阵 $\boldsymbol{F}\in\mathbb{R}^{|\mathcal{V}|\times f}$,其每一行都包含了对应节点的 f 个特征值. 相比于非递归特征,ReFeX 不仅在各种图挖掘任务中表现出更好的性能,还易扩展到大规模网络中,故而被后来的方法广泛应用. RolX 则是第一个在该结构特征矩阵上利用非负矩阵分解技术(Non-negative Matrix Factorization,NMF)生成节点嵌入向量的方法,具体如式(1)所示:

$$\arg\min_{\boldsymbol{Z},\boldsymbol{M}}\|\boldsymbol{F}-\boldsymbol{Z}\boldsymbol{M}\|_{fro},\ \text{s.t.}\ \boldsymbol{Z}\geq 0,\ \boldsymbol{M}\geq 0 \tag{1}$$

其中 $\boldsymbol{Z}\in\mathbb{R}^{|\mathcal{V}|\times d}$ 为节点嵌入矩阵,也称为节点-角色矩阵,其每一行表征节点在 d 个角色上的分布情况,d 的大小由最小描述长度准则(Minimal Description Length,MDL)[42] 求解. $\boldsymbol{M}\in\mathbb{R}^{d\times f}$ 是角色-特征矩阵,表示不同特征对特定角色的贡献程度. RolX 的计算复杂度为 $O(|\mathcal{E}|f+|\mathcal{V}|fd)$.

GLRD[43]. Gilpin 等人认为以无监督的方法建模网络角色是不合理的,因此他们提出了第一个引导角色导向网络表示学习的探索性工作 GLRD. 具体来说,GLRD 在 RolX 基础上通过增加稀疏性、多样性和替代性约束来引导节点在不同角色上的分布. 首先,稀疏性约束如式(2)所示:

$$\begin{aligned}&\forall i\ \|\boldsymbol{Z}_{\cdot i}\|_1\leq\epsilon_{\boldsymbol{Z}},\\&\forall i\ \|\boldsymbol{M}_{i\cdot}\|_1\leq\epsilon_{\boldsymbol{M}}\end{aligned} \tag{2}$$

其中 $\epsilon_{\boldsymbol{Z}}$ 和 $\epsilon_{\boldsymbol{M}}$ 是约束上界. 直观地,稀疏性约束鼓励节点被分配给尽可能少的角色,角色由尽可能少的特征定义,这对于提高预测任务的准确性有着极其重要的作用. 其次,多样性约束如式(3)所示:

$$\begin{aligned}&\forall i,j\ \boldsymbol{Z}_{\cdot i}^{\mathrm{T}}\boldsymbol{Z}_{\cdot j}\leq\epsilon_{\boldsymbol{Z}},\ i\neq j,\\&\forall i,j\ \boldsymbol{M}_{i\cdot}^{\mathrm{T}}\boldsymbol{M}_{j\cdot}\leq\epsilon_{\boldsymbol{M}},\ i\neq j\end{aligned} \tag{3}$$

也就是说,GLRD 通过施加多样性约束来处理建模过程中可能出现的不同角色却具有相似成员分布以及不同角色却由相似特征分布定义的问题. 最后,替代性约束如式(4)所示:

$$\begin{aligned}&\forall i,j\ \boldsymbol{Z}_{\cdot i}^{*\mathrm{T}}\boldsymbol{Z}_{\cdot j}\leq\epsilon_{\boldsymbol{Z}},\\&\forall i,j\ \boldsymbol{M}_{i\cdot}^{*\mathrm{T}}\boldsymbol{M}_{j\cdot}\leq\epsilon_{\boldsymbol{M}}\end{aligned} \tag{4}$$

其中 $\boldsymbol{Z}^*$ 和 $\boldsymbol{M}^*$ 表示一组已知的解. 大规模的网络系统通常会包含一些复杂现象,其具体表现为 $\boldsymbol{Z}$ 和 $\boldsymbol{M}$ 可能存在许多组解. 为获得与已有解释不同的解,可以通过添加替代性约束来实现.

REGAL[13]. REGAL 是一种基于网络表示学习的网络对齐框架,其核心设计是利用 xNetMF 来学习角色信息保持的节点嵌入. xNetMF 首先线性地聚合了节点 K 跳邻域的度信息:

$$\boldsymbol{F}_i=\sum_{k=1}^{K}\delta^{k-1}\boldsymbol{F}_i^k \tag{5}$$

其中 δ 表示聚合权重,$\boldsymbol{F}_i^k$ 的第 l 个元素 $\boldsymbol{F}_{i,l}^k=|\{v_j\in\mathcal{N}_i^k\ |\lfloor\log_2\boldsymbol{D}_{jj}\rfloor=l\}|$. 然后依式(6)计算节点对 (v_i,v_j) 的相似性:

$$\boldsymbol{S}_{ij}=\exp[-\gamma_s\cdot\|\boldsymbol{F}_i-\boldsymbol{F}_j\|_2^2-\gamma_a\cdot dis(\boldsymbol{X}_i,\boldsymbol{X}_j)] \tag{6}$$

其中 $dis(\boldsymbol{X}_i,\boldsymbol{X}_j)$ 是不同节点属性距离的度量函数(如欧式距离),γ_s 和 γ_a 用于控制结构和属性对于身份识别的影响力.

由于在大规模网络中计算完整的相似性矩阵具有极高的复杂度,因此 xNetMF 提出利用低秩矩阵 $\tilde{\boldsymbol{S}}$ 做近似估计的方法. 具体步骤如下:

(1)随机选取 $p(p\ll|\mathcal{V}|)$ 个节点,计算图中所有节点与这 p 个节点的相似性矩阵 $\boldsymbol{S}\in\mathbb{R}^{|\mathcal{V}|\times p}$,并从矩阵 $\boldsymbol{S}$ 中获得 p 个节点间的相似性矩阵 $\boldsymbol{S}^*\in\mathbb{R}^{p\times p}$.

(2)低秩近似矩阵 $\tilde{\boldsymbol{S}}=\boldsymbol{S}\boldsymbol{W}^{\dagger}\boldsymbol{S}^{\mathrm{T}}$,$\boldsymbol{W}^{\dagger}$ 是 $\boldsymbol{S}^*$ 的伪逆矩阵. 由于最终目标是从近似矩阵 $\tilde{\boldsymbol{S}}$ 中隐式分解出节点的近似嵌入矩阵 $\tilde{\boldsymbol{Z}}$,该方法并不必显式计算近似矩阵 $\tilde{\boldsymbol{S}}$. 具体如式(7)所示:

$$\begin{aligned}&[\boldsymbol{U},\boldsymbol{\Sigma},\boldsymbol{V}]=SVD(\boldsymbol{W}^{\dagger}),\\&\tilde{\boldsymbol{Z}}=\boldsymbol{S}\boldsymbol{U}\boldsymbol{\Sigma}^{-\frac{1}{2}},\\&\tilde{\boldsymbol{Z}}=Normalize(\tilde{\boldsymbol{Z}})\end{aligned} \tag{7}$$

REGAL 复杂度为 $O(|\mathcal{V}|(K\boldsymbol{D}_{\mathrm{avg}}^2+pb+p^2))$,$b=\lceil\log_2\boldsymbol{D}_{\max}\rceil$,$\boldsymbol{D}_{\mathrm{avg}}$ 和 $\boldsymbol{D}_{\max}$ 是平均和最大节点度.

EMBER[44]. EMBER 是一种专门针对加权有向的电子邮件网络提出的节点嵌入算法,其目标在于利用网络中每个节点的局部特征实现专业角色推断. 在电子邮件网络中,不同专业角色的局部结构通常呈现出一些可观察的特征,例如高级职员往往会有更大的出度和入度,因为他可能存在更多的联系人,也会收到更多来自其他职员的电子邮件. 基于此,EMBER 首先聚合了节点 K 跳邻域内的度信息和边的权重信息,得到结构行为向量 $\boldsymbol{F}_i$:

$$\boldsymbol{F}_i=\Big[\sum_{k=0}^{K}\delta^k\boldsymbol{F}_i^{k+},\sum_{k=0}^{K}\delta^k\boldsymbol{F}_i^{k-}\Big] \tag{8}$$

其中 $\boldsymbol{F}_i^{k+}$ 和 $\boldsymbol{F}_i^{k-}$ 反映了节点 v_i 与不同方向上距离 k 步的节点的交互关系. 具体如式(9)所示：

$$\boldsymbol{F}_{i,l}^{k+}=\sum_{j\in\mathcal{N}_{i,l}^{k+}}\Big(\prod_{(u,v)\in\mathcal{P}_{i\to j}^{k+}}\boldsymbol{W}_{uv}\Big) \tag{9}$$

其中 $\mathcal{N}_{i,l}^{k+}=\{v_j\in\mathcal{N}_i^{k+}\mid\lfloor\log_2\boldsymbol{D}_{jj}\rfloor=l\}$，$\mathcal{N}_i^{k+}$ 为节点 v_i 在 k 步能够到达的节点集合，$\mathcal{N}_i^{k-}$ 为 k 步能够到达 v_i 的节点集合，(u,v) 为路径中存在相应连边的节点对. 类似地，可以计算得到 $\boldsymbol{F}_{i,l}^{k-}$. 此后与 xNetMF 一样，EMBER 随机选取 p 个节点计算相似性矩阵，然后通过隐式分解获得近似的节点嵌入矩阵 $\widetilde{\boldsymbol{Z}}$. EMBER 复杂度为 $O(|\mathcal{V}|(K\boldsymbol{D}_{\text{avg}}^2+p\log_2\boldsymbol{D}_{\max}+p^2))$.

DRNE[45]. Tu 等人在研究如何利用规则等价性学习节点表示时，提出了一种深度递归的网络嵌入方法 DRNE. 它以节点度为衡量依据，将节点的邻居节点排列成有序序列，然后利用归一化的 LSTM[46] 机制非线性地递归聚合来自邻域的节点表示，从而保留目标节点的局部信息. DRNE 实质上就是让目标节点的嵌入向量能够近似其一阶邻居嵌入向量的聚合，其重构损失如式(10)所示：

$$\mathcal{L}_1=\sum_{i\in\mathcal{V}}\|\boldsymbol{Z}_i-\widetilde{\boldsymbol{Z}}_i\|_{fro}^2 \tag{10}$$

其中 $\widetilde{\boldsymbol{Z}}_i=Agg(\{\boldsymbol{Z}_j\mid v_j\in\mathcal{N}_i\})$. 考虑到大多数网络具有重尾分布的特征，即仅有少数的节点拥有很大的度，DRNE 算法对度过大的节点的邻域进行采样以达到提高算法效率的目的，其策略是采样概率与节点度大小成正比. 此外，为了避免模型退化导致所得嵌入向量都为 $\mathbf{0}$ 的情况，DRNE 增加正则项 $\mathcal{L}_{reg}$，即把节点度信息作为弱引导信息，约束学习到的嵌入向量能够重构节点的度特征：

$$\mathcal{L}_{reg}=\sum_{i\in\mathcal{V}}\|\log(\boldsymbol{D}_{ii}+1)-MLP(\widetilde{\boldsymbol{Z}}_i)\|_{fro}^2 \tag{11}$$

因此，DRNE 整体的损失函数如式(12)所示：

$$\mathcal{L}=\mathcal{L}_1+\lambda\,\mathcal{L}_{reg} \tag{12}$$

其中 λ 是正则项的权重系数. 由于度作为弱引导信息仅起到辅助作用而非监督作用，因此它通常是一个较小的数值. 此外，虽然该工作论述了这一方法在不断递归过程中可以学习节点在全局上的结构信息，但实际上它与其它基于局部特征的方法类似，因此我们仍将其划分为了基于局部特征的方法. DRNE 迭代训练一次的时间复杂度为 $O(|\mathcal{V}|sd^2)$，其中 s 是邻域采样数量的上界，d 是节点嵌入维度.

GAS[47]. 当前大多数角色导向的网络嵌入方法都依赖于结构特征的设计，使得许多方法不具有普适性，如 EMBER. 图神经网络(Graph Neural Network，GNN)虽然在特征提取方面有着强大的潜力，但将其用于学习角色信息保持的节点嵌入向量时，损失函数又难以定义. 为解决这些问题，GAS 开创性地提出利用结构化特征来引导图神经网络的训练，这为面向角色的网络嵌入研究提供了新思路. 具体来说，GAS 首先修正图卷积网络(Graph Convolutional Network，GCN)[48] 的传播机制来编码网络的结构特征，获得节点的嵌入向量：

$$\boldsymbol{Z}^{(l)}=\sigma(\hat{\boldsymbol{A}}\boldsymbol{Z}^{(l-1)}\boldsymbol{\Theta}^{(l)}) \tag{13}$$

其中 $l=1,2,\cdots,L$，$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$，$\boldsymbol{\Theta}^{(l)}$ 是可训练的权重矩阵，$\sigma(\cdot)$ 是非线性的激活函数，$\boldsymbol{Z}^{(l)}\in\mathbb{R}^{|\mathcal{V}|\times d}$ 是第 l 层的输出，其每一行是对应节点的表示向量，$\boldsymbol{Z}^{(0)}=\hat{\boldsymbol{A}}$ 或其它可训练的初始化矩阵. 然后，GAS 利用多层感知器解码重构网络的结构特征矩阵 $\hat{\boldsymbol{F}}$：

$$\hat{\boldsymbol{Z}}^{(o)}=\sigma(\hat{\boldsymbol{Z}}^{(o-1)}\boldsymbol{W}^{(o)}+\boldsymbol{b}^{(o)}) \tag{14}$$

其中 $o=1,2,\cdots,O$，$\boldsymbol{W}^{(o)}$ 和 $\boldsymbol{b}^{(o)}$ 是第 o 层的权重矩阵和偏置向量，$\hat{\boldsymbol{Z}}^{(0)}=\boldsymbol{Z}^{(L)}$，$\hat{\boldsymbol{Z}}^{(O)}=\hat{\boldsymbol{F}}$.

最后，GAS 以 ReFeX 提取节点的结构特征 $\boldsymbol{F}$，并构建损失函数：

$$\mathcal{L}_g=\|\hat{\boldsymbol{F}}-\boldsymbol{F}\|_{fro}^2 \tag{15}$$

此外，为提高模型的鲁棒性，GAS 对模型参数添加了正则损失，如式(16)所示：

$$\mathcal{L}_{reg}=\sum_{l=1}^{L}\|\boldsymbol{\Theta}^{(l)}\|_{fro}^2+\sum_{o=1}^{O}\|\boldsymbol{W}^{(o)}+\boldsymbol{b}^{(o)}\|_{fro}^2 \tag{16}$$

GAS 最终的损失函数如式(17)所示：

$$\mathcal{L}=\mathcal{L}_g+\lambda\,\mathcal{L}_{reg} \tag{17}$$

其中 λ 是正则项的权重系数. GAS 每层图卷积网络的复杂度为 $O(|\mathcal{E}|w)$，w 为输入、输出维度的乘积.

RESD[49]. RESD 是当前将深度学习框架用于角色导向网络表示学习研究的最新方法之一. 它在特征提取上仍采用 ReFeX 方法实现，但不同的，也是最关键的是，RESD 通过引入变分自编码器(Variational Auto-Encoder，VAE)[50] 这一生成模型来建模这些结构特征的非线性关系，进而获得节点的向量表示. 变分自编码器的框架如下所示：

(1) 以 ReFeX 提取的特征为输入，编码节点结构特征：

$$\boldsymbol{Z}_j^l=\tanh(\boldsymbol{W}^l\boldsymbol{Z}_j^{l-1}+\boldsymbol{b}^l),\ l=1,\cdots,L \tag{18}$$

其中 $\boldsymbol{W}^l$ 和 $\boldsymbol{b}^l$ 为第 l 层的权重矩阵和偏置向量. 这里假设节点表示服从高斯分布，有：

$$\begin{aligned}\boldsymbol{\mu}_j&=\boldsymbol{W}^{\mu}\boldsymbol{Z}_j^l+\boldsymbol{b}^{\mu},\\ \log\boldsymbol{\sigma}_j&=\boldsymbol{W}^{\sigma}\boldsymbol{Z}_j^l+\boldsymbol{b}^{\sigma}\end{aligned} \tag{19}$$

其中 $\boldsymbol{W}^{\mu},\boldsymbol{b}^{\mu},\boldsymbol{W}^{\sigma},\boldsymbol{b}^{\sigma}$ 是可训练的参数.

(2) 利用重参数化技巧获得节点表示：

$$\boldsymbol{Z}_j=\boldsymbol{\mu}_j+\boldsymbol{\sigma}_j\odot\epsilon,\ \epsilon\sim N(0,1) \tag{20}$$

(3) 以多层感知机为解码器，重构节点结构特

征 $\hat{\boldsymbol{F}}$：

$$\hat{\boldsymbol{F}}_j^o=\tanh(\hat{\boldsymbol{W}}^o\hat{\boldsymbol{F}}_j^{o-1}+\hat{\boldsymbol{b}}^o),o=1,\cdots,L \quad (21)$$

其中 $\hat{\boldsymbol{W}}^o$ 和 $\hat{\boldsymbol{b}}^o$ 是可训练的参数，$\hat{\boldsymbol{F}}_j^0=\boldsymbol{Z}_j$.

(4) 损失函数如式(22)所示：

$$\mathcal{L}_{VAE}=\|\boldsymbol{F}-\hat{\boldsymbol{F}}\|_2^2 \quad (22)$$

RESD 本质上并不直接利用网络中提取来的特征，而是转向学习这些特征背后的某种生成机制，这使得学习到的节点嵌入对网络噪声具有健壮的鲁棒性. 此外，RESD 通过增加节点度正则来弥补编码过程中可能损失掉的统计信息：

$$\mathcal{L}_{reg}=\sum_{i\in\mathcal{V}}\|\log(\boldsymbol{D}_{ii}+1)-MLP(\boldsymbol{Z}_i)\|_{fro}^2 \quad (23)$$

模型的整体损失函数如式(24)所示：

$$\mathcal{L}=\mathcal{L}_{VAE}+\lambda\mathcal{L}_{reg} \quad (24)$$

其中 λ 是正则项的权重系数. RESD 的计算复杂度是 $O(f(|\mathcal{E}|+|\mathcal{V}|f+|\mathcal{V}|d))$.

Node2bits[51]. 该方法是针对用户缝合任务(即判断社交网络中哪些用户在现实中是属于一个人的)提出的，因此处理的是有向时序网络. 该方法首先定义了时序随机游走，即要求游走时后采样边的时间戳不能小于之前采样边的时间戳. 从节点 v_i 游走到 v_j 的概率定义如式(25)所示：

$$p_w(v_i,v_j)=\frac{\exp(-\tau_{ij}/T)}{\sum_{v_{j'}\in\mathcal{E}^{\tau_{ij'}}}\exp(-\tau_{i,j'}/T)} \quad (25)$$

其中 τ_{ij} 为边 e_{ij} 的某个时间戳.

对每条边生成多个不同长度的时序随机游走序列以捕获边的多个时间戳信息. 每个节点的局部特征可通过聚合其所在游走序列中的节点特征获得. 将以此得到的每个节点特征对数化并映射为 onehot 向量，然后利用 Simhash 算法[52]将其映射为二值向量并将其拼接. Node2bits的计算复杂度为 $O(|\mathcal{E}|rl+|\mathcal{V}|d)$，其中 r 是对每个边进行随机游走的次数，l 是随机游走的最大长度.

总结. 在上述方法中，基于矩阵分解的 RolX 和 GLRD 通过在表征节点连接模式的结构特征矩阵上应用非负矩阵分解技术来派生角色导向的节点表示. REGAL 和 EMBER 则在提取特征的基础上进一步计算了节点对之间的结构相似性矩阵. DRNE 和 GAS 作为深度学习的方法，将节点的局部特征作为辅助信息来引导角色导向的嵌入的生成，而 RESD 则更直接地转向学习产生这些局部拓扑结构的生成机制. 基于随机游走的 Node2bits 则通过聚合游走序列中节点的局部特征来生成含有角色信息的向量表示. 不难发现，基于局部特征的网络嵌入方法通常利用简单的统计特征来表示节点的局部拓扑结构，它们在一些结构简单的网络上具有良好的性能表现，并且可以快速地在大规模网络中扩展. 但是由于这类算法无法刻画节点的全局信息，缺乏对节点高阶特征的捕捉，使得它们不能很好地处理更为复杂的网络结构. 并且，因为这类算法依赖于节点的局部特征，使其对于网络中改变节点局部结构的噪声或者扰动十分敏感，即使是基于神经网络自动拟合节点间依赖关系的深度学习方法，也会因其损失函数的定义导致其性能随着扰动的增加而震荡.

3.2 基于全局特征的角色导向的网络表示学习方法

定义 7. 全局特征. 全局特征派生于节点的全局拓扑结构. 它通过抽取节点与其它所有节点之间的连接关系来简化表达其在高维空间中的全局位置信息.

现有角色导向的嵌入方法常利用能够捕获节点全局特征的方法来构造节点的特征矩阵，例如 εER 方法、Rooted PageRank 方法等，然后将其用于矩阵分解或随机游走中.

RIDεRs[53]. RIDεRs 是一种从全局视角编码图特征的角色发现算法. 它首先利用 εER[54] 方法将网络节点 $\{v_1,v_2,\cdots,v_{|\mathcal{V}|}\}$ 划分为 l 个不相交的节点集合 $\pi=\{\mathcal{C}_1,\cdots,\mathcal{C}_l\}$，这些节点集合满足以下条件，即对于所有 $1\leq i,j\leq l$，以及所有节点 $u,v\in\mathcal{C}_i$ 都有：

$$|deg(u,\mathcal{C}_j)-deg(v,\mathcal{C}_j)|\leq\varepsilon \quad (26)$$

其中 $deg(u,\mathcal{C}_j)=|\{v_k|(u,v_k)\in\mathcal{E},v_k\in\mathcal{C}_j\}|$ 表示集合 $\mathcal{C}_j$ 中与节点 u 相邻的节点个数.

实质上，εER 要求同属一个集合的节点要具有相似的连接模式，因此不同集合刻画了网络中的不同角色. 值得注意的是，在划分集合的过程中，εER 要考量每个节点与网络中所有其它节点的连接关系，所以这是一种基于全局特征的方法. 之后，RIDεRs 构建了节点特征矩阵 $\boldsymbol{F}$，其中 $\boldsymbol{F}_{ij}=|\mathcal{N}_i\cap\mathcal{C}_j|$ 表示集合 $\mathcal{C}_j$ 中与节点 v_i 存在连边的节点个数. 考虑到在较大的网络中选取较小的 ε 值会造成划分的角色数量大幅增加，因此需要对矩阵 $\boldsymbol{F}$ 进行对数装箱和修剪处理. 最后，通过拼接不同 $\varepsilon\in[1,\boldsymbol{D}_{\text{avg}}]$ 产生的特征矩阵并利用非负矩阵分解技术以获得节点的紧凑表示.

GraphWave[39]. 由于以节点为中心的谱图小波(spectral graph wavelets)在不同的拓扑结构上扩散会呈现出不同的状态，在具有相同角色的节点上会有相似的小波系数分布，因此 Donnat 等人提出了一种基于谱图小波扩散[55]学习节点表示的方法 GraphWave，并从数学上证明了这种方法的有效性，这是谱图小波第一次被用于该领域的研究. 对于给定

无向图$\mathcal{G}=(\mathcal{V},\mathcal{E})$，节点$v_i$的谱图小波$\boldsymbol{\Psi}_i$计算如下：

$$\boldsymbol{\Psi}_i=\boldsymbol{U}e^{-s\boldsymbol{\Lambda}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\delta}_i \tag{27}$$

其中，$\boldsymbol{U}$是拉普拉斯矩阵$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}$的特征向量，$\boldsymbol{\Lambda}=diag(\lambda_1,\cdots,\lambda_{|\mathcal{V}|})$是对应特征值构成的对角矩阵，$s$为缩放参数，$\boldsymbol{\delta}_i$是节点$v_i$的独热编码，$\boldsymbol{\Psi}_i$是一个$|\mathcal{V}|$维列向量，其中第$m$行元素$\boldsymbol{\Psi}_{mi}$反映了节点$v_m$对$v_i$的影响. 在GraphWave中，它将谱图小波看作是随机变量，求解经验特征函数ϕ_i：

$$\phi_i(t)=\frac{1}{|\mathcal{V}|}\sum_{m=1}^{|\mathcal{V}|}e^{it\boldsymbol{\Psi}_{mi}} \tag{28}$$

最后在等间隔的序列$\{t_1,\cdots,t_{d/2}\}$上对$\phi_i(t)$采样，进而获得d维的节点嵌入向量：

$$\boldsymbol{Z}_i=[\mathrm{Re}(\phi_i(t_k)),\mathrm{Im}(\phi_i(t_k))] \tag{29}$$

其中$k\in\{t_1,\cdots,t_{d/2}\}$，Re和Im分别代表复数的实部和虚部.

GraphWave区别于以往需要启发式定义节点结构特征再进行模型训练的方法，对细小的扰动具有更健壮的鲁棒性. 在计算复杂度上，由于GraphWave使用切比雪夫多项式来近似计算式(27)，所以模型复杂度为$O(K|\mathcal{E}|)$，K是多项式阶数.

SPINE[56]. SPINE是一种联合捕捉节点局部邻近性和结构化身份的归纳式网络嵌入框架，它在现实网络的分析任务上具有更一般的适用性. 该方法主要分为三个部分，首先是结构特征的生成. SPINE使用Rooted PageRank(RPR)[10]生成全局特征矩阵：

$$\boldsymbol{F}^{RPR}=(1-\beta_{RPR})(\boldsymbol{I}-\beta_{RPR}\boldsymbol{P})^{-1} \tag{30}$$

其中$\beta_{RPR}\in(0,1)$表示从当前节点随机游走到其他节点而不返回起始节点的概率，$\boldsymbol{P}$为节点间的转移概率矩阵，即$\boldsymbol{P}_{ij}=\dfrac{\boldsymbol{W}_{ij}}{\sum_{k=1}^{|\mathcal{V}|}\boldsymbol{W}_{ik}}$.

考虑归纳前提，SPINE采用蒙特卡洛近似来计算$\boldsymbol{F}_i^{RPR}$，并按降序排列选取前k个元素作为特征值表示节点的结构信息. 其次，SPINE融合节点结构和属性信息来生成节点嵌入：

$$\boldsymbol{Z}_i=\sigma\Big(\sum_{j=1}^{k}\boldsymbol{F}_{ij}^{RPR}\boldsymbol{X}_j\boldsymbol{W}_M\Big) \tag{31}$$

其中$\boldsymbol{X}_j\in\mathbb{R}^m$是对应节点的属性向量，$\boldsymbol{W}_M\in\mathbb{R}^{m\times d}$是多层感知机中的权重矩阵，$\sigma(\cdot)$是非线性激活函数.

最后，SPINE设计有偏的SGNS(Skip-Gram Negative Sampling)实现优化. 其采样策略是定义一个参数$\alpha\in(0,1)$控制基于局部邻近性和基于结构化身份的采样比率. 具体地说，该模型以α概率根据结构特征向量的相似性对节点v_i进行正采样，否则对随机游走的共现节点进行正采样.

struc2gauss[57]. 在struc2gauss提出之前，几乎所有角色导向的网络嵌入算法都是根据网络结构中的空间依赖关系学习确定性的点向量. 然而，由于现实世界中的复杂系统通常是嘈杂或不完整的，所以点向量并无法准确地刻画节点的角色信息，尤其是在度较小的节点上. 基于此，Pei等人提出了struc2gauss方法，该方法从高斯分布中学习节点的向量表示，以实现对网络不确定性的建模. 具体来说，struc2gauss首先利用RoleSim算法[58-59]计算节点v_i与其它节点的相似性，其复杂度为$O(w|\mathcal{V}|^2)$，w为其它相关参数. 其次，根据节点的相似性构造上下文序列，即选取l个最为相似的节点作为正样本集$\Gamma_+=\{(v_i,v_j)|j=1,\cdots,l\}$，同时随机采样相同个数的节点作为负样本集$\Gamma_-=\{(v_i,v_k)|k=1,\cdots,l\}$，该过程计算复杂度为$O(|\mathcal{V}|\log l)$. 然后，通过最小化式(32)中的损失函数来控制样本之间的相对距离，以优化节点嵌入：

$$\mathcal{L}=\sum_{v_i\in\mathcal{V}}\sum_{v_j\in\Gamma_+}\sum_{v_k\in\Gamma_-}\max(0,\varpi+\varepsilon(\boldsymbol{Z}_i,\boldsymbol{Z}_j)-\varepsilon(\boldsymbol{Z}_i,\boldsymbol{Z}_k)) \tag{32}$$

其中ϖ是超参数，$\varepsilon(\cdot,\cdot)$是不同节点表示在高斯分布上的相似性度量，通常采用期望似然和相对熵作为度量函数.

最后，为获得合理的向量$\boldsymbol{\mu}$和矩阵$\boldsymbol{\Sigma}$，struc2gauss在初始化和训练过程中对其添加了如下正则约束：

$$\begin{aligned}&\|\boldsymbol{\mu}_i\|\leq C,\ \forall i,\\&c_{\min}\boldsymbol{I}<\boldsymbol{\Sigma}_i<c_{\max}\boldsymbol{I},\forall i\end{aligned} \tag{33}$$

对于高斯嵌入部分，其计算复杂度为$O(|\mathcal{V}|)$. 也就是说，struc2gauss的可扩展性很大程度上受限于节点结构相似性的计算.

总结. 上述方法在学习节点嵌入时均考量了节点在网络中的全局信息. RIDεRs通过εER方法将全局网络中具有相似结构的节点划分到同一个集合. GraphWave通过谱图小波的扩散来捕获网络中所有其它节点对中心节点的影响. SPINE以Rooted PageRank方法构造的全局特征矩阵表示节点对之间的结构相似性. struc2gauss基于节点全局结构相似性生成节点的上下文序列. 直观来看，基于全局特征的网络嵌入方法相比基于局部特征的方法，可以更好地建模节点在全局网络中的角色信息. 但值得注意的是，这并不意味着这类方法的性能一定会优于第一类方法. 事实上，方法性能的好坏更大程度上取决于网络本身的角色分布情况，这在后续实验分析中有所论述. 此外，由于这一类方法不再只是关注于节点的局部特征，尤其是像GraphWave引入了谱

图小波和经验函数，struc2gauss 建模了网络的不确定性，它们对于网络中的细小扰动具有更健壮的鲁棒性.但也正是因为这些复杂技术的应用，使得这些方法的时间和空间复杂度会随着网络规模的增大而快速提升，这不利于方法的实际应用.

3.3 基于高阶特征的角色导向的网络表示学习方法

定义 8. 高阶特征.高阶特征派生于节点的高阶拓扑结构.它通过刻画节点在网络中的子图结构来表达节点在高维空间中的高阶依赖关系，例如对网络基元计数、对邻域子图使用图核创建的结构特征等.由于网络中的功能单元通常由两个或两个以上的节点构成，所以捕获节点的高阶特征有助于识别网络中的功能角色.

现有基于高阶特征的角色导向网络表示学习方法大都通过引入网络基元(network motif/graphlet)或邻域子图的概念来比较节点高阶连接模式的相似性.这一思想对学习面向角色的网络嵌入来说是有益的，并且在许多工作中得到了验证[60-61].此外，还有一些其它工作，它们扩展了传统随机游走的概念，通过为中心节点生成结构化上下文序列来学习角色信息保持的节点嵌入.

HONE[62]. Rossi 等人在捕捉网络高阶结构依赖和连接模式时，提出了一种基于网络基元的高阶网络表示学习框架 HONE.它在给定网络$\mathcal{G}=(\mathcal{V},\mathcal{E})$和网络基元$\mathcal{H}=\{\mathcal{H}_1,\cdots,\mathcal{H}_T\}$基础上，计算基元的加权邻接矩阵$\mathcal{W}=\{\mathcal{W}_1,\cdots,\mathcal{W}_T\}$.对于给定网络基元$\mathcal{H}_t$，有：

$$(\mathcal{W}_t)_{ij}=\text{节点 } v_i \text{ 和 } v_j \text{ 参与构成} \mathcal{H}_t \text{ 的实例数} \tag{34}$$

然后，HONE 利用基元矩阵函数$\Psi:\mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}\rightarrow\mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$将该加权邻接矩阵推广至其它矩阵，如表 2 所示，并派生$\mathcal{H}_t$的 k 阶矩阵$\boldsymbol{S}_t^{(k)}=\Psi(\mathcal{W}_t^{(k)})$.

最后通过优化式(35)得到局部嵌入向量$\boldsymbol{U}_t^{(k)}$：

$$\underset{\boldsymbol{U}_t^{(k)},\boldsymbol{V}_t^{(k)}\in C}{\arg\min}\ \mathbb{D}(\boldsymbol{S}_t^{(k)}\|\Phi\langle\boldsymbol{U}_t^{(k)}\boldsymbol{V}_t^{(k)}\rangle) \tag{35}$$

其中 $\mathbb{D}$ 为广义的 Bregman 散度，用于度量$\boldsymbol{S}_t^{(k)}$和$\Phi\langle\boldsymbol{U}_t^{(k)}\boldsymbol{V}_t^{(k)}\rangle$的距离，$\Phi$是描述低秩嵌入矩阵$\boldsymbol{U}_t^{(k)},\boldsymbol{V}_t^{(k)}$和 k 阶矩阵$\boldsymbol{S}_t^{(k)}$间线性或非线性关系的函数.

表 2 基元矩阵函数 Ψ

矩阵	函数定义
基元的加权图	$\Psi:\mathcal{W}_t\rightarrow\boldsymbol{I}\mathcal{W}_t$
基元的度矩阵	$\boldsymbol{D}=diag(\mathcal{W}_t\boldsymbol{e})$
基元的转移矩阵	$\boldsymbol{P}=\boldsymbol{D}^{-1}\mathcal{W}_t$
基元的拉普拉斯矩阵	$\boldsymbol{L}=\boldsymbol{D}-\mathcal{W}_t$
归一化的拉普拉斯矩阵	$\hat{\boldsymbol{L}}=\boldsymbol{I}-\boldsymbol{D}^{-\frac{1}{2}}\mathcal{W}_t\boldsymbol{D}^{-\frac{1}{2}}$
随机游走的归一化拉普拉斯矩阵	$\hat{\boldsymbol{L}}_{rw}=\boldsymbol{I}-\boldsymbol{D}^{-1}\mathcal{W}_t$

在此基础上，HONE 利用欧几里得范数对$\boldsymbol{U}_t^{(k)}$的列向量做归一化处理，得到拼接矩阵$\boldsymbol{Y}$：

$$\boldsymbol{Y}=[\boldsymbol{U}_1^{(1)}\cdots\boldsymbol{U}_T^{(1)}\cdots\boldsymbol{U}_1^{(K)}\cdots\boldsymbol{U}_T^{(K)}] \tag{36}$$

最后按照式(37)可以求解具有全局结构信息的节点嵌入矩阵$\boldsymbol{Z}$：

$$\underset{\boldsymbol{Z},\boldsymbol{H}\in C}{\arg\min}\ \mathbb{D}(\boldsymbol{Y}\|\Phi\langle\boldsymbol{Z}\boldsymbol{H}\rangle) \tag{37}$$

其中$\boldsymbol{H}$的每一列表示 k 阶基元的特征嵌入向量.式(37)常用 Frobenius 范数来最小化求解：

$$\min_{\boldsymbol{Z},\boldsymbol{H}}\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{H}\|_{fro}^2=\frac{1}{2}\sum_{ij}(\boldsymbol{Y}_{ij}-(\boldsymbol{Z}\boldsymbol{H})_{ij})^2 \tag{38}$$

直观地，因为基于网络基元的加权邻接矩阵刻画了中心节点高阶的连接模式，所以它可以保证矩阵分解后得到的节点表示保留了原始网络中的高阶结构相似性.HONE 的计算复杂度为$O(|\mathcal{E}|(\Delta_{ub}+KTD_l)+|\mathcal{V}|dKTd_l)$，其中$\Delta_{ub}$是最大度，$d_l$和 d 分别是局部和最终的嵌入维度.

SEGK[63]. SEGK 的核心思想是通过图核建立一个表征节点邻域子图结构相似性的核矩阵$\boldsymbol{S}$，然后对该矩阵做近似分解进而获得节点的向量表示.具体来说，SEGK 针对单一节点v_i，定义了一组不同尺度下的邻域子图集合$\{\mathcal{G}_i^1,\cdots,\mathcal{G}_i^K\}$，其中$\mathcal{G}_i^k$表示原始网络中以节点$v_i$为中心，距离$v_i$小于等于 k 的节点及其之间连边构成的子图，其计算复杂度为$O(|\mathcal{V}|+|\mathcal{E}|)$.在该集合上，按照如下公式定义节点间的核并作为核矩阵第 i 行第 j 列元素$\boldsymbol{S}_{ij}$：

$$\boldsymbol{S}_{ij}=s(v_i,v_j)=\sum_{k=1}^{K}\hat{s}_{\mathcal{G}}(\mathcal{G}_i^k,\mathcal{G}_j^k)\hat{s}_{\mathcal{G}}(\mathcal{G}_i^{k-1},\mathcal{G}_j^{k-1}) \tag{39}$$

其中$\hat{s}_{\mathcal{G}}$是$s_{\mathcal{G}}$的标准化版本，如式(40)所示，并有$\hat{s}_{\mathcal{G}}(\mathcal{G}_i^0,\mathcal{G}_j^0)=1$.$s_{\mathcal{G}}$通常采用最短路径核(Shortest Path kernel，SP)、WL 子树核(Weisfeiler-Lehman subtree kernel，WL)或 graphlet 核(Graphlet Kernel，GL)：

$$\hat{s}_{\mathcal{G}}(\mathcal{G}_i,\mathcal{G}_j)=\frac{s_{\mathcal{G}}(\mathcal{G}_i,\mathcal{G}_j)}{\sqrt{s_{\mathcal{G}}(\mathcal{G}_i,\mathcal{G}_i)s_{\mathcal{G}}(\mathcal{G}_j,\mathcal{G}_j)}} \tag{40}$$

由于在真实的网络上构造完整的核矩阵并对其进行奇异值分解是非常占用内存和消耗时间的，所以 SEGK 采用 Nyström 方法[64]，仅对核矩阵的某 p 列做矩阵分解，以获得满足$\boldsymbol{S}\approx\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$的嵌入矩阵$\boldsymbol{Z}$.该过程的计算复杂度为$O(|\mathcal{V}|pKt^{k_{\mathcal{G}}}+|\mathcal{V}|p^2)$，其中$t^{k_{\mathcal{G}}}$表示节点邻域子图为全图时的图核计算复杂度.在 SEGK 的方法设计中，它通过比较不同尺度的邻域子图，有效地解决了邻域子图同构但中心节点连接模式不相同以及节点一阶邻居可能大不相同，但其高阶邻域具有大规模相似性的问题.此外，SEGK 还可以处理属性图和标签图.

DeepGL[65]. DeepGL 作为一个通用的并且富有表现力的深度网络表示学习框架，它依赖于一些灵活可替换的组件，包括派生的基本特征、选取的关系特征算子、相关的图元素（例如节点、边）集合以及关系特征算子间的组合次数. 对于给定网络$\mathcal{G}=(\mathcal{V},\mathcal{E})$，DeepGL 首先利用局部子图分解方法[66-67]将$\mathcal{G}$划分为更小的网络基元，并将其计数和一些简单的统计特征如节点的出入度、自我中心网络特征等作为基本特征得到输入矩阵$\boldsymbol{F}_1$. 然后为学习更高阶的节点特征，DeepGL 递归式地利用前一层的特征和一组关系特征算子$\Phi=\{\Phi_1,\cdots,\Phi_K\}$来派生新的特征层，并通过特定的评价标准来实现重要特征的选取，其中关系特征算子如表 3 所示. 与以往工作不同，DeepGL 学习的关系函数使其可以在不同的网络中进行泛化，并有效地用于图的迁移学习任务中. 此外，该方法还引入了特征扩散的概念，使得每一层的特征矩阵都可以通过任意的扩散过程来平滑，例如$\overline{\boldsymbol{F}}^{(t)}=\boldsymbol{D}^{-1}\boldsymbol{A}\overline{\boldsymbol{F}}^{(t-1)}$. 由此产生的$\overline{\boldsymbol{F}}$可以直接替换已有的特征$\boldsymbol{F}$，也可以与$\boldsymbol{F}$进行拼接产生新的特征. DeepGL 本质上就是在基本特征的基础上，利用一系列的关系特征算子来拟合节点的高阶结构. 所以当两个节点在网络中具有相似的连接模式时，会因为它们的基本特征是相似的，而使得它们最终的节点表示也是相似的. DeepGL 计算节点表示的复杂度为$O(f(|\mathcal{E}|+|\mathcal{V}|f))$，其中$f$表示特征的数量.

表 3　关系特征算子

算子	定义
Hadamard	$\Phi\langle\mathcal{R},\boldsymbol{F}\rangle=\prod_{r_j\in\mathcal{R}}\boldsymbol{F}_j$
mean	$\Phi\langle\mathcal{R},\boldsymbol{F}\rangle=\frac{1}{\lvert\mathcal{R}\rvert}\sum_{r_j\in\mathcal{R}}\boldsymbol{F}_j$
sum	$\Phi\langle\mathcal{R},\boldsymbol{F}\rangle=\sum_{r_j\in\mathcal{R}}\boldsymbol{F}_j$
maximum	$\Phi\langle\mathcal{R},\boldsymbol{F}\rangle=\max_{r_j\in\mathcal{R}}\boldsymbol{F}_j$
weighted L^p	$\Phi\langle\mathcal{R},\boldsymbol{F}\rangle=\sum_{r_j\in\mathcal{R}}\lvert\boldsymbol{F}_i-\boldsymbol{F}_j\rvert^p$
RBF	$\Phi\langle\mathcal{R},\boldsymbol{F}\rangle=\exp\left(-\frac{1}{\sigma^2}\sum_{r_j\in\mathcal{R}}[\boldsymbol{F}_i-\boldsymbol{F}_j]^2\right)$

其中$\mathcal{R}$是相关的图元素集合，$\boldsymbol{F}_i$是特征向量.

MCN[68]. MCN 的核心是在传统 GCN 的基础上引入注意力机制以允许模型选择最佳的邻域来整合节点的高阶结构信息. 一方面，MCN 采用和 HONE 相同的方法描述节点在网络中的高阶连接模式，即为每个节点计算一组k阶的基元邻接矩阵；另一方面，MCN 在每一层图卷积神经网络上设计两个函数$f_l:\mathbb{R}^{S_l}\to\mathbb{R}^{\mathrm{T}}$和$f_l':\mathbb{R}^{S_l}\times\mathbb{R}^{\mathrm{T}}\to\mathbb{R}^K$，用于选择在给定节点$v_i$状态下与其最相关的网络基元$\mathcal{H}_t$和步长$k$，记其索引为$t_i$和$k_i$. 然后 MCN 依式(41)定义矩阵$\hat{\boldsymbol{A}}\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$代替原始的邻接矩阵进行层间的信息传播：

$$\hat{\boldsymbol{A}}=\begin{bmatrix}(\widetilde{\boldsymbol{A}}_{t_1}^{(k_1)})_{1,:}\\ \vdots\\ (\widetilde{\boldsymbol{A}}_{t_{|\mathcal{V}|}}^{(k_{|\mathcal{V}|})})_{|\mathcal{V}|,:}\end{bmatrix}\tag{41}$$

其中$\widetilde{\boldsymbol{A}}$是由基元邻接矩阵派生出的其它矩阵.

直观地，不同网络基元和步长的选取决定了模型对节点连接模式的描述，影响了图卷积神经网络中的特征扩散. 为此，MCN 除了对所有带标签的节点计算标准的交叉熵损失$\mathcal{L}_C$外，还基于强化学习设计了另一个损失函数$\mathcal{L}_A$用于注意力机制的训练，以保证模型在捕捉节点高阶结构信息方面的能力.

GCC[69]. GCC 是一种受自然语言处理和计算机视觉启发的自监督的图神经网络预训练框架，它认为即使在不同网络之间也存在通用的、可迁移的结构模式，因此只需要在给定的一组多样化的输入图上，利用自监督方式预训练一个通用的 GNN 模型，就可以经过微调使其在不同图上完成各种基于图的下游任务. 例如可以在 Facebook 社交网络和 DBLP 合作网络上预训练一个 GNN，然后将其用于美国航空网络的节点分类任务上. 该方法的核心是依据对比学习的思想，将子图实例判别作为预训练的自监督任务. 具体地，GCC 首先定义了节点v_i的k阶自我中心网络$\mathcal{G}_i^k$. 然后，在$\mathcal{G}_i^k$基础上通过匿名的重启随机游走进行两次图采样[70]，并将生成的子图样本作为一组相似的实例对(x^q,x^{k_+}). 相反的，如果两个子图样本源于不同的自我中心网络，则将二者构成的实例对(x^q,x^k)看作是不相似的. 此后，利用两个GIN(Graph Isomorphism Network)神经网络模型分别编码x^q和x^k的结构特征得到d维嵌入向量$\boldsymbol{z}^q$和$\boldsymbol{z}^k$. 最后应用 InfoNCE[71]损失来优化 GIN 编码器，具体如式(42)所示：

$$\mathcal{L}=-\log\frac{\exp(\boldsymbol{z}^{q\mathrm{T}}\boldsymbol{z}^{k_+}/\tau)}{\sum_{i=0}^{K}\exp(\boldsymbol{z}^{q\mathrm{T}}\boldsymbol{z}_i^k/\tau)}\tag{42}$$

其中τ是超参数.

对比损失的设置鼓励模型将相似实例对的节点表示映射到相近的位置，而让不相似的实例对嵌入到相对较远的位置，这尽可能地保证了输出的节点表示可以捕获这些子图的结构相似性. 对于下游任务的实现，GCC 提供了两种微调策略. 一种是将预训练的编码器看作是特征提取器，对待处理的网络进行特征提取，然后训练适合特定下游任务的分类

器;另一种是使用预训练的模型参数初始化新的图编码器,并让新的图编码器与分类器一起进行端到端的训练.

struc2vec[72]. 传统基于随机游走的模型,如 DeepWalk[73]和 node2vec[38],将随机游走生成的节点序列作为输入训练 Skip-gram 模型得到网络表示. 这类方法虽然有效,但是由于 Skip-gram 的窗口内只可见相互邻近的节点,导致它们不能捕获网络中的结构等价性. 基于这些问题,Ribeiro 等人提出了 struc2vec 方法. 在训练语言模型之前,利用层次结构来度量不同尺度的节点相似性,并通过构建多层加权图来实现结构相似性的编码和上下文序列的生成. 这一设计不仅有效地克服了传统随机游走的局限性,并为学习角色保持的节点嵌入向量拓宽了道路. 具体来说,struc2vec 启发式地定义递归函数 $f^k(i,j)$来计算节点对结构距离,如式(43)所示:

$$f^k(i,j)=f^{k-1}(i,j)+g(s(\mathcal{N}_i^k),s(\mathcal{N}_j^k)) \tag{43}$$

其中 $s(\mathcal{N}_i^k)$表示距离节点 v_i 为 k 的节点集合构成的有序度序列,$g(\cdot,\cdot)$表示两个有序度序列的距离度量函数,常采用动态时间规整算法(Dynamic Time Warping,DTW)[74].

然后,它利用该结构距离构建一个 K 层的加权图$\tilde{\mathcal{G}}$来衡量结构相似性. 图$\tilde{\mathcal{G}}$的每一层都是由原始网络中所有节点构成的完全图,层内节点连边权重、层间节点连边权重由式(44)计算:

$$\begin{aligned}&w^k(i,j)=e^{-f^k(i,j)},\ k=0,\cdots,K,\\&w(i^k,i^{k+1})=\log(\Gamma^k(i)+e),\ k=0,\cdots,K-1\\&w(i^k,i^{k-1})=1,\ k=0,\cdots,K\end{aligned} \tag{44}$$

其中 $\Gamma^k(i)$表示第 k 层中所有与节点 v_i 相连的边中权重大于该层平均权重的数量.

在得到加权图$\tilde{\mathcal{G}}$后,利用随机游走为每个节点生成相应的上下文序列. 其游走策略分为两步,首先以概率 p 停留在当前层,否则以式(45)选择跳转到上一层或下层:

$$\begin{aligned}&p^k(i^k,i^{k+1})=\frac{w(i^k,i^{k+1})}{w(i^k,i^{k+1})+w(i^k,i^{k-1})},\\&p^k(i^k,i^{k-1})=1-p^k(i^k,i^{k+1})\end{aligned} \tag{45}$$

其次,倘若停留在当前层,则以如下概率选择下一个遍历节点:

$$p^k(i,j)=\frac{e^{-f^k(i,j)}}{\sum\limits_{j\in\mathcal{V},i\neq j}e^{-f^k(i,j)}} \tag{46}$$

直观地,该方法的随机游走更倾向于跳转到与当前节点具有相似结构的其它节点上,而不依赖于该节点在原始网络中的标签和位置. 最后,将生成的上下文序列作为语料库来训练 Skip-gram 语言模型,进而获得保持结构相似性的向量表示. struc2vec 在未使用任何优化策略时的计算复杂度为 $O(K|\mathcal{V}|^3)$.

RiWalk[75]. RiWalk 是 Ma 等人为快速嵌入节点并保持结构化特征提出的一种灵活范式,它将结构化的网络嵌入问题解耦为角色识别和一般网络嵌入问题. 在角色识别过程中,它采用图核中重标记和子结构的思想来捕获上下文节点与中心节点之间的连接模式,并将这些模式编码为新的标识符以获得重新标记的子图结构. 具体地,RiWalk 提出了两种用于角色识别的方法,RiWalk-SP 和 RiWalk-WL. 前者可以看作是最短路径图核的一种近似,它对节点 v_i 的 k 阶邻域子图$\mathcal{G}_i^k$中的非 v_i 节点 v_j 做如下标记:

$$\varphi_i(v_j)=h(\boldsymbol{D}_{ii})\oplus h(\boldsymbol{D}_{jj})\oplus s_{ij} \tag{47}$$

其中 $\boldsymbol{D}_{ii}$ 表示节点 v_i 的度,s_{ij} 表示节点 v_i 和 v_j 的最短路径,$h(\cdot)$表示对数,即 $h(x)=\lfloor\log_2(x+1)\rfloor$,$\oplus$是串联运算符,计算复杂度为 $O(\boldsymbol{D}_{avg}|\mathcal{V}|^2)$.

然而,这种方法缺乏区分与节点 v_i 具有相同距离和相同度的节点的能力,但有时这些节点与 v_i 连接的紧密程度并不一致,这可能体现在它们到 v_i 的路径数不同. 因此,RiWalk 提出了 RiWalk-WL 方法,用于捕获更细粒度的连接模式. 该方法可以看作是 Weisfeiler-Lehman 图核的一种近似. 它对节点 v_i 的 k 阶邻域子图$\mathcal{G}_i^k$中的非 v_i 节点 v_j 做如下标记:

$$\varphi_i(v_j)=h(\boldsymbol{X}_{ii})\oplus h(\boldsymbol{X}_{ij})\oplus s_{ij} \tag{48}$$

其中 $\boldsymbol{X}_{ij}$ 的第 n 个元素表示节点 v_j 的邻居中与 v_i 最短距离为 n 的节点个数.

RiWalk-WL 的计算复杂度为 $O(\boldsymbol{D}_{\text{avg}}^2|\mathcal{V}|^2)$. 直观地,在重标记子图中,以相同方式与中心节点连接的上下文节点会被视为同一节点,即使它们属于不同的子图. 同时,结构上相似的中心节点会共享许多上下文节点,因此可以直接将其转化为一般的基于随机游走的网络嵌入问题,进而学习结构化的节点嵌入. 由于 RiWalk 是一个通用的策略,因此角色识别和网络嵌入的方法都可以针对下游任务做出相应的调整.

role2vec[76]. 通过传统随机游走的方法获得的上下文序列是由节点序号表示的,不仅无法刻画节点间的结构相似性[77],而且无法将节点属性或特征集成到网络表示中. 为解决这些问题,Ahmed 等人在 role2vec 框架中提出了一种基于特征的随机游走的概念,并将其作为泛化其它方法的基础. 该方法首先根据节点的拓扑结构(例如网络基元的计数)构造相应的特征向量,并对每一个元素做装箱处理. 其次它通过学习一个函数 $\Phi:f\to w$ 将 $|\mathcal{V}|$ 个节点依据各自的特征向量映射到角色集合 $\mathcal{W}=\{w_1,\cdots,w_M\}$ 中

的不同角色上，其中 $M \ll |\mathcal{V}|$. 然后，使用基于特征的随机游走方法遍历网络中的所有节点，即通过访问各个节点所属的角色来生成相应的上下文序列，用于最后概率模型的构建. 直观地，如果两个节点具有相似的拓扑结构，那么它们会被映射到同一个角色中，并在后续的学习中获得相近的节点嵌入向量.

总结. 在上述方法中，基于矩阵分解的 HONE 和 SEGK 方法分别以网络基元和邻域子图为核心，通过基元矩阵函数和图核方法派生节点的高阶特征. 这种方法一方面很大程度上依赖于基元矩阵函数和图核方法的选取，不具有普适性；另一方面它们通常面临计算效率的问题，不适合大规模网络的应用. 而基于深度学习的 MCN 方法利用注意力机制为节点自动选取更有用的网络基元，DeepGL 方法学习通用的关系函数拟合节点邻域子图的高阶特征，GCC 方法设计子图实例判别训练预训练模型，这些方法不仅可以自动地学习节点的高阶特征，并且可以有效地推广到不同领域的复杂网络中，以实现节点角色的挖掘. 基于随机游走的 struc2vec 依节点结构相似构造多层加权图，该方法虽然有效，但是由于完全图的构造，造成了较高的时间和空间复杂度. 而 role2vec 和 RiWalk 都旨在通过节点角色映射来学习一般的网络节点嵌入.

绝大多数的角色导向网络表示学习方法都专注于从结构相似性的角度来获得节点表示，而忽略了节点自身的属性信息. 然而，真实世界中的复杂网络通常包含丰富的属性信息，例如社交网络中用户的个人资料，合著网络中作者发表的文章数量，这些信息间接地反映了节点间可能存在的交互关系甚至是潜在的角色分布情况. 融合网络节点的属性信息有助于我们更好地学习节点表示并实现跨网络的任务. 例如 REGAL 在提取特征的基础上整合属性信息来缓解不同网络中相同节点拓扑结构不一致的问题，为网络对齐任务提供了新思路. SPINE、SEGK 和 DeepGL 等方法利用节点的属性信息来提高跨网络的节点分类、链接分类等迁移学习任务的性能，说明了属性信息在网络表示学习中的重要作用. 虽然结合属性的角色导向网络表示学习已经有了初步的进展，但这对于属性网络的研究来说还是远远不够的. 并且据我们所知，当前还没有公开的具有角色标签的属性网络被用于角色导向的网络表示学习的研究. 因此如何构建真实的属性网络并有效地融合节点的结构信息和属性信息来学习节点的低维向量表示将是未来研究的方向之一.

3.4　计算复杂度分析

计算复杂度是网络表示学习方法能否在大规模网络上应用的关键，因此我们在论述各种方法的嵌入机制时，还参考相关论文给出了模型的计算复杂度. 本节主要结合新的分类机制探讨和分析这些方法的可扩展性. 这主要是因为在现有的角色导向的网络表示学习方法中，节点结构特征的提取不仅影响了模型性能的好坏，其计算复杂度还限制了模型的可扩展性.

根据第 3.1 节的论述可以发现，基于局部特征的角色导向网络表示学习方法的计算复杂度往往与网络中的节点数或连边数成线性关系. 其中很重要的一个原因是这类方法在提取结构特征时仅需捕获节点与其有限距离邻域的连接关系，而不用考虑节点全局或高阶的拓扑结构，这极大地降低了模型的计算成本. 例如在真实网络中提取节点邻域特征的计算复杂度仅有 $O(|\mathcal{V}|)$[40]，ReFeX 的计算复杂度是 $O(f(|\mathcal{E}|+|\mathcal{V}|f))$，聚合节点 K 跳邻域内度特征的计算复杂度是 $O(|\mathcal{V}|K\boldsymbol{D}_{\text{avg}}^2)$. 反观第 3.2 节和3.3 节中基于全局特征、高阶特征的方法，由于它们在特征提取时需要考虑更全面或更复杂的网络拓扑结构，例如计算网络中所有节点对之间的结构相似性，计算邻域子图之间的相似性，构造结构化的上下文序列等，导致其特征提取的计算复杂度大幅提高，这在 struc2gauss、SEGK、RiWalk、struc2vec 等方法中表现得尤为明显，其计算复杂度是网络节点数的平方甚至是立方. 当网络规模较大时，这些方法将难以应用. 因此，优化特征提取过程对提高模型的可扩展性具有重要意义.

此外，角色导向的网络表示学习方法的计算复杂度还与其生成嵌入的方法有关. 目前，基于矩阵分解的方法大多采用低秩矩阵近似[13]的方法来降低模型复杂度，如 REGAL、EMBER、SEGK 等. 基于深度学习的方法则通过构建预训练模型来进一步提升下游任务的性能并降低模型训练难度，如 GCC.

4　实验及分析

在本节中，我们详细介绍了角色导向网络表示学习研究中常用的 8 个实验数据集及其基本统计指标，并从基于局部特征、全局特征和高阶特征的方法中选取了 10 个代表性方法进行综合性实验，以评估这些方法在捕获节点角色方面的潜力. 另外，本文还提供了 2 个具有社团标签的实验数据集以及 5 个相

关的算法模型来阐述角色和社团导向的网络表示学习方法在相同网络分析任务中的不同侧重.综合性实验包括:(1)可视化实验.通过绘制二维的嵌入空间来直观地分析不同导向或不同机制下的网络表示学习方法在捕捉节点依赖关系时的不同侧重,例如社团导向的网络表示学习倾向于捕捉节点的社团结构;(2)节点分类实验.本文利用数据集的真实标签和不同机制下生成的节点表示来训练相应的线性分类器,通过比较分类指标数值的高低来间接评估这些方法在该场景下的性能差异;(3)节点聚类实验.除了节点分类这一有监督的任务外,本文还利用 K-means 模型进行无监督的聚类实验,通过比较聚类指标的好坏来论述这些方法对节点的聚类能力;(4)鲁棒性实验.本文为探究不同机制下网络表示学习方法的稳定性,设计了四种不同的扰动来分析模型的鲁棒性;(5)参数敏感性实验.在机器学习与深度学习的任务中,特征维度是影响模型效果的重要因素,因此本文探究不同的节点表示维度对角色分类的影响.最后,本节介绍了一个角色导向的网络表示学习平台,为该领域的后续研究提供有力支撑.

4.1 真实数据集

本文在 8 个具有角色标签和 2 个具有社团标签的真实数据集上进行了可视化以外的其它实验.数据集的基本统计指标如表 4 所示,包括网络的节点数、连边数、类别数、平均度、密度、聚集系数和传递性,这些指标描述了不同网络的规模大小、稀疏程度以及结构化程度.数据集详情如下:

(1)航空网络[72].本文使用了巴西、欧洲和美国的航空网络(分别标识为 Brazil、Europe 和 America).在这些网络中,节点表示机场,连边表示机场间存在来往航班,节点类别根据机场的流量等级划分.

(2)电话通讯网络[78].该网络(标识为 Reality)记录了某一时期不同用户间的移动通话数据.其节点表示手机用户,连边表示用户之间的通话联系,节点类别依据用户的呼叫频率划分.

(3)影视网络.包括一个演员合作网络[75](标识为 Actor)和一个英文电影网络[79](标识为 Film).Actor 网络中节点表示演员,连边表示他们同时出现在一个维基百科页面中,节点类别根据该页面的单词数量划分.Film 网络中节点间连边仍表示共现关系,而节点角色包括电影、导演、演员以及编剧.

(4)作者合作网络[80].该网络表示作者间的合作关系.本文根据关键词提取了两个连通子图(分别标识为 Coauthor-Network 和 Coauthor-System).节点根据其 h-index 划分.

(5)引文网络[7].本文利用了两个引文网络(分别标识为 Cora 和 Citeseer),其节点表示文章,连边表示文章的引用关系,节点类别代表文章的研究领域.

表 4 数据集统计指标

数据集	节点数	边数	节点类别数	标签	网络平均度	网络密度/%	平均聚集系数	传递性
Brazil	131	1074	4	角色	16.3969	12.6130	0.6364	0.4497
Europe	399	5995	4	角色	30.0501	7.5503	0.5393	0.3337
America	1190	13599	4	角色	22.8555	1.9222	0.5011	0.4263
Reality	6809	7697	3	角色	2.2608	0.0332	0.0178	0.0024
Actor	7758	26733	4	角色	6.8917	0.0888	0.0790	0.0156
Film	27312	122706	4	角色	8.9855	0.0329	0.1180	0.0278
Coauthor-Network	67667	186863	3	角色	5.5230	0.0082	0.6234	0.2741
Coauthor-System	146092	400863	3	角色	5.4878	0.0038	0.6120	0.3979
Cora	2708	5278	7	社团	3.8981	0.1440	0.2407	0.0935
Citeseer	3279	4552	6	社团	2.7765	0.0847	0.1435	0.1301

在上述实验数据集中,前四类数据集的节点标签与其角色或功能相关,常被用于角色导向的网络表示学习的研究.而最后一类引文网络中的节点标签与其社团结构有关,常被用于社团导向的网络表示学习的研究.

4.2 实验设置

本文结合新的分类方式,选取 10 个面向角色的网络表示学习方法进行可视化、分类、聚类、鲁棒性和参数敏感性实验,其中 RolX[41]、DRNE[45]、GAS[47]和 RESD[49]属于基于局部特征的方法,RIDεRs[53]、GraphWave[39]和 struc2gauss[57]属于基于全局特征的方法,SEGK[63]、struc2vec[72]和 role2vec[76]属于基于高阶特征的方法.另外,本文还选择了 5 个社团导向的方法进行对比,包括 GraLSP[81]、SDNE[82]、node2vec[38]、GraphSTONE[83]和 VGAE[84].在实验过程中,若没有特殊说明,各个方法生成的节点表示维度均设置为 128(RIDεRs 和 GraphWave 除外,其维度与网络规模或其他模型参数相关),其它参数参考相关论文中的默认设置.

本文所有实验使用 Python 语言实现,运行于配

备 Intel(R) Xeon(R) CPU E5-2680 v4@2.40 GHz 和 125 GB RAM 的 Linux 服务器(Ubuntu 18.04.5 LTS).

4.3　可视化

本文首先在图 1 的样例网络上进行了可视化实验,其目的是为角色和社团概念提供直观的理解和认识.实验结果如图 3 所示,我们在第一个子图中展示了该网络真实的角色分布,其它子图的角色标签则通过对学得的节点表示应用 K-means 算法获得,不同颜色标记了不同节点簇.

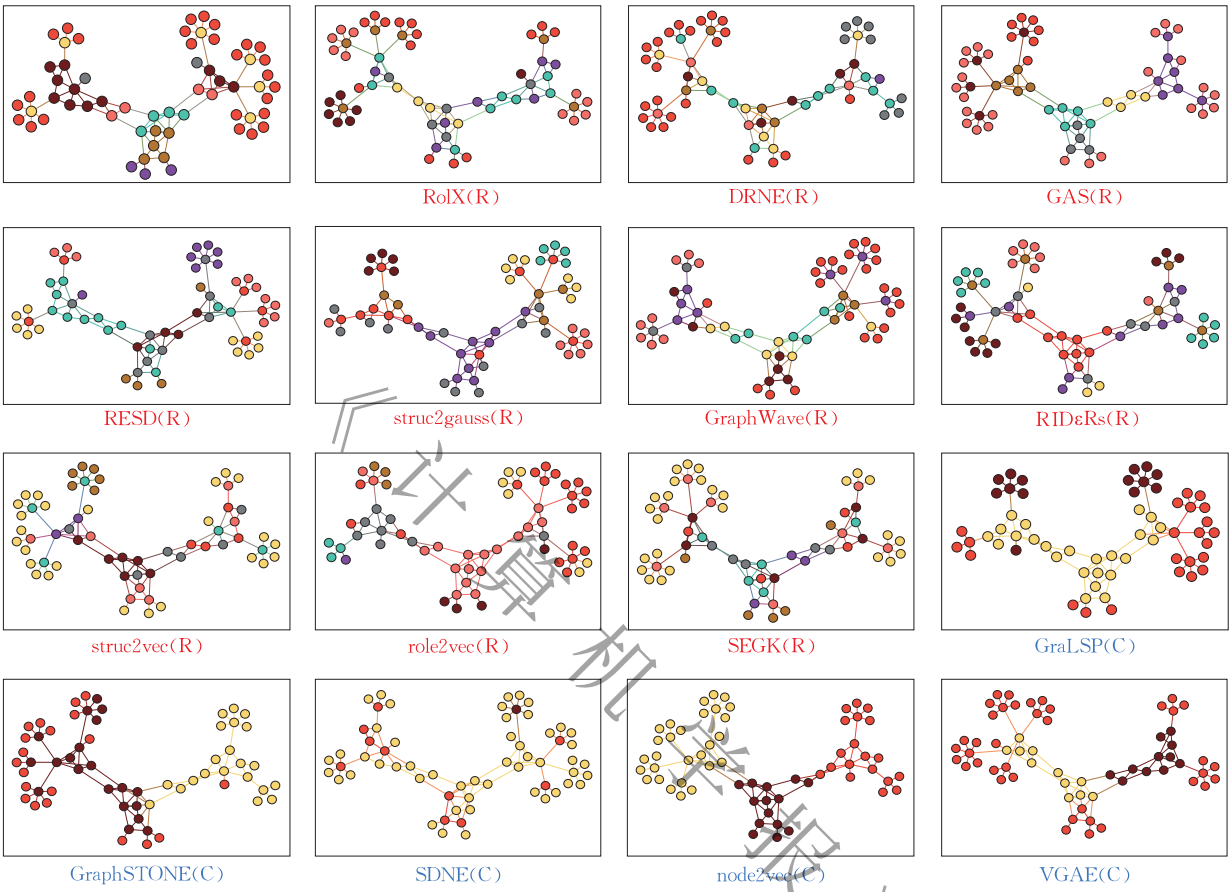

图 3　样例网络可视化

从第一个子图可以看出,网络中隶属同一角色的节点具有相同的功能或位置,如分配红色角色的节点是网络边缘的外围节点,分配黄色角色的节点是连接不同外围节点的中心节点.基于角色的网络嵌入方法则根据网络节点的结构特征捕获这样的角色信息.虽然图 3 中未展示该样例网络真实的社团结构,但 node2vec 的实验结果已反映了社团内部节点的聚集特性.

接着,本文在 Barbell 网络[72]和 Karate 网络[85]上对选取的 15 种网络嵌入方法做了进一步的可视化,旨在对比节点表示在嵌入空间中的分布情况来定性地评估各方法的性能.具体地,我们利用 PCA 模型将学得的节点表示降到二维空间中,实验结果如图 4 和图 5 所示.其中 Barbell 网络是人工合成的对称网络,包含 30 个节点和 101 条边,节点类别根据连接模式严格划分,与节点的功能角色相关联.Karate 网络是空手道俱乐部的内部交流网络,包含 34 个节点和 78 条边,节点类别划分借助模块度优化的社团发现算法 Louvain 实现,与节点的社团归属相关联.

(1)社团导向的网络表示学习方法

如图 4 和图 5 所示,社团导向的网络嵌入方法在以功能角色划分节点类别的 Barbell 网络上表现较差,而在以社团结构划分节点类别的 Karate 网络上表现出显著优势,这表明这类方法不具备捕获原始网络中节点结构相似性的能力,但是可以很好地保持节点间的邻近关系.这与其目标是相符合的,即社团导向的网络表示学习方法旨在为连接紧密或距离相近的节点学习相似的低维表示.但值得注意的是,SDNE 是一个特例.由于该方法主张联合利用一阶邻近性(假设两个相邻节点是相似的)和二阶邻近性(假设具有许多共同邻居的节点是相似的)来学

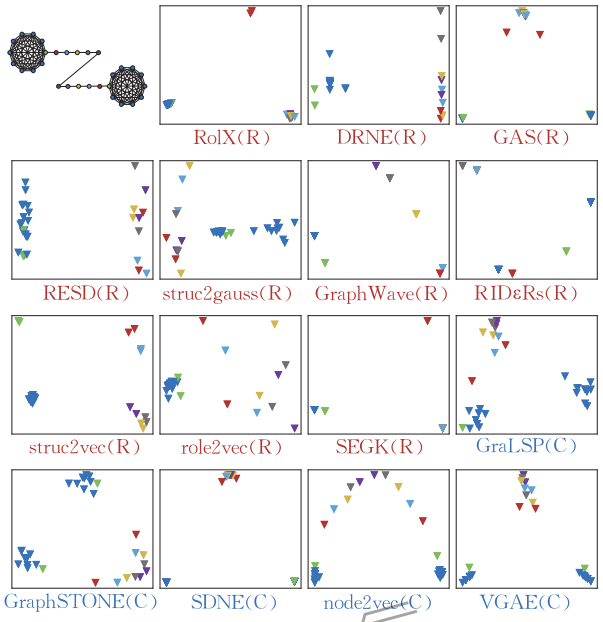

图 4 Barbell 网络可视化

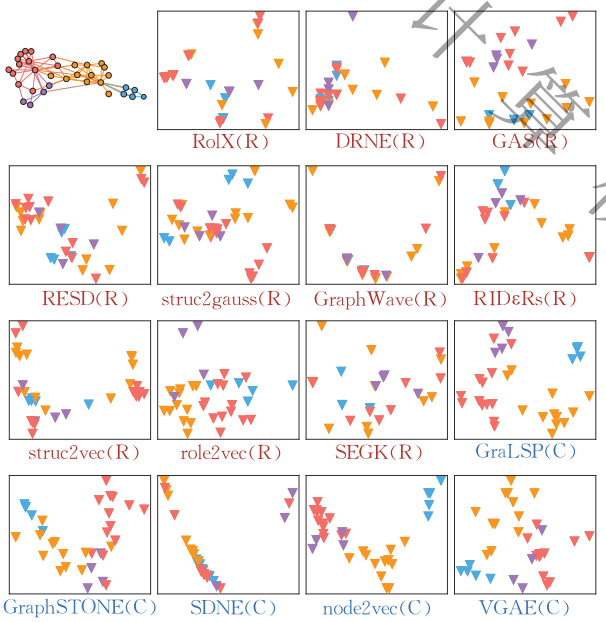

图 5 Karate 网络可视化

习节点嵌入，所以它在一定程度上可以保持节点结构上的相似性，如它将 Barbell 网络上的团节点嵌入到了同一位置. 但也因为 SDNE 弱化了对邻近性的保持，导致其在 Karate 网络上表现得并不好. 如何在学得的节点表示中同时保持邻近性和结构相似性是未来的研究热点和难点.

(2) 角色导向的网络表示学习方法

首先我们对比了角色导向的网络嵌入方法在 Barbell 网络和 Karate 网络上的表现，其结果与社团导向的方法呈现相反趋势，这表明这类算法更倾向于学习网络中的结构相似性，而非邻近性，与其目标相一致，即角色导向的网络表示学习方法旨在为具有相似连接模式的节点学习相近的低维表示. 此外，我们结合新的分类机制着重对比了不同方法在 Barbell 网络上的性能，分析如下：

① 基于局部特征的网络嵌入方法可以较好地识别 Barbell 网络中距离较远但同构的蓝色节点，如 RolX、DRNE 以及 RESD，但这类方法对不同桥节点的划分不够清晰，尤其是 RolX，它几乎只能识别该网络中的三类角色. 其主要原因是 Barbell 网络中不同桥节点的划分更依赖于节点在网络中的全局位置信息，但这通常被基于局部特征的方法所忽视.

② 基于全局特征的网络嵌入方法 GraphWave 在 Barbell 网络上表现出显著优势. 在其可视化结果中，相同颜色的节点被映射到相同位置，不同颜色的节点间距离符合预期，如角色相近的绿色节点与蓝色节点距离更近.

③ 基于高阶特征的网络嵌入方法如 struc2vec 虽然也尽可能地识别出了网络中所有的角色，但其投影间的距离与预期不一致，例如红色节点与黄色节点的距离比紫色节点更远. role2vec 和 SEGK 均在识别不同桥节点时表现较差.

基于以上分析可以看出，角色导向的网络表示学习方法可以有效地捕获远距离节点间的相似性关系. 此外，当节点的角色标签与节点在网络中的全局位置有关时，基于全局特征的网络嵌入方法相比于其它两类方法来说更适用.

4.4 节点分类

我们在 10 个数据集上评估了 15 种网络嵌入方法在节点分类任务上的性能. 具体地，在每一个实验数据集上，我们随机抽取 70% 的节点表示训练线性逻辑回归分类器，然后将其余节点表示留作测试，最后以 Micro-$F1$ 定量地评估分类结果的优劣. 不失一般性，本文在每个实验数据集上重复实验 20 次取其平均值. 实验结果如图 6 所示，其横坐标表示不同的实验方法，纵坐标表示每种方法的 Micro-$F1$ 得分，空值部分表示该方法无法在指定内存学习该网络的节点表示. 根据实验结果，我们有以下观察和分析：

(1) 角色导向和社团导向的网络表示学习方法在不同网络上存在较大的性能差异. 由图 6 可以看到，角色导向的网络表示学习方法在航空网络、电话通讯网络、影视网络以及作者合作网络上表现相对突出，社团导向的网络表示学习方法在引文网络上更占优势. 其原因在于前 4 类真实网络的节点标签与节点的连接模式高度相关(例如在以机场流量大

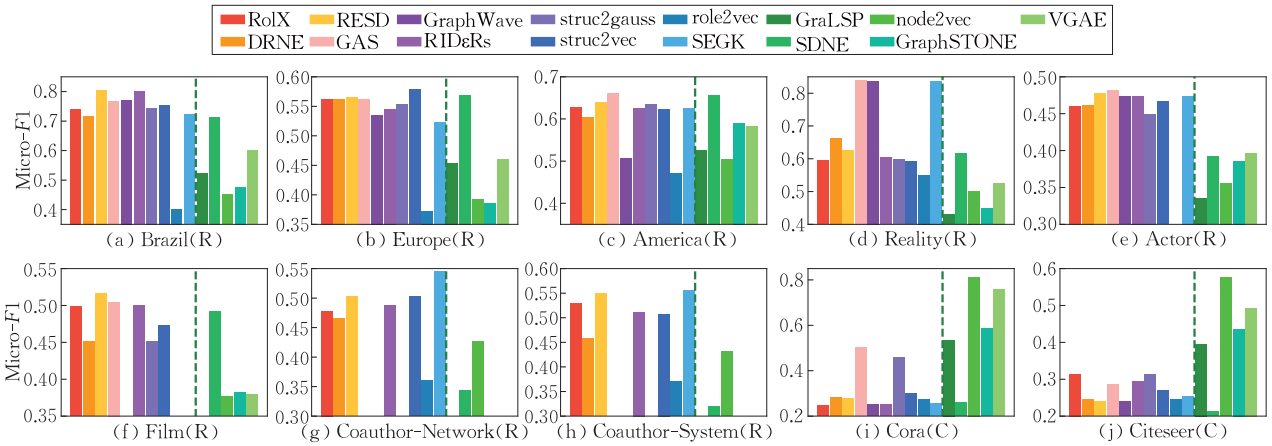

图 6　以 Micro-$F1$ 为指标的节点分类实验结果

小标注节点标签的航空网络中，与枢纽机场相对应的节点往往会表现出较大的度特征），所以角色导向的网络嵌入方法相较于社团导向的方法来说可以捕获更多的可用信息. 但是引文网络不同，在引文网络上隶属同一个类别的节点在其结构上并无明显的相似性关系，反而是在距离上表现出了显著的聚集性，因此对于这一类网络而言保持节点间的邻近性才能有效地提高节点分类的性能. 此外，我们观察了 SDNE 在前 8 个真实网络上的 Micro-$F1$ 得分，可以发现它在欧洲和美国航空网络上的分类效果甚至好于绝大部分的角色导向的网络嵌入方法，但在一些规模更大、结构更复杂的网络如演员合作网络和作者合作网络上表现不佳，这表明二阶邻近性并不足以表征复杂网络中节点连接模式的相似性.

（2）不同类别下的角色导向网络表示学习方法分类效果不一：

① 基于局部特征的方法在整体上表现出较好的分类效果，特别是网络结构相对简单的小型航空网络. 这是因为这些网络的节点角色通常可以简单地由节点的局部结构表征，例如节点度的大小在一定程度上可以反映相应机场在整个航空网络上的重要程度，而该重要程度又与其流量的大小密切相关. 基于深度学习的 RESD 和 GAS 相比于其它基于局部特征的方法总是能获得更高的 Micro-$F1$ 得分，这说明引入变分自编码器和以 ReFeX 提取的结构特征作为引导信息来学习节点低维嵌入是有效的. DRNE 因其仅考虑了节点度特征，所以它在相对复杂的英文电影网络上表现较差. 此外，我们还可以看到因为基于局部特征的方法相比于全局或高阶的方法具备更低的计算复杂度和空间复杂度，所以这类方法更易于在大规模网络如作者合作网络上扩展.

② 基于全局特征的方法在这些网络上并没有表现出非常显著的优势，其原因是这些网络中的节点标签很少依赖于节点的全局位置信息. 但值得注意的是，GraphWave 在电话通讯网络上获得了惊人的结果. 此外，基于全局特征的方法在英文电影网络以及作者合作网络等较大规模的网络上表现出局限性.

③ 基于高阶特征的方法中 role2vec 的分类结果最差（这可能是因为该方法所提出的基于特征的随机游走更适用于链接预测任务，而不适合角色发现），并且由于其需要对网络基元进行计数，导致该方法需要耗费更多的时间和空间成本. 事实上，struc2vec 和 SEGK 也存在较高的计算复杂度和空间复杂度，但二者在性能上相对较好，尤其是 SEGK 在电话通讯网络以及作者合作网络上表现极为突出，这表明融合节点的高阶特征是有助于学习更高质量的节点表示的.

从以上分析可以看出，基于局部特征的角色导向网络嵌入方法虽然在提取特征时会丢失网络的全局信息，但是当节点角色与其全局位置关联度不高时，这类方法会具有较好的性能，并且基于全局特征的方法也有可能因为忽视节点的局部结构而导致其性能不佳. 基于高阶特征的方法则因其捕获了更精细的节点连接模式，使得它在相对复杂的网络上有更好的适用性. 但是基于全局特征和高阶特征的方法都存在难以扩展的问题. 此外，我们还观察到基于局部特征的 GAS、基于全局特征的 GraphWave 以及基于高阶特征的 SEGK 在节点标签分布不均匀的电话通讯网络上有着一致的分类效果，这表明存在一些网络，我们难以推测其结构化行为是与局部或全局或高阶特征更相关.

4.5　聚类

社团检测和角色发现均可以看作是无监督的聚类任务. 我们利用 K-means 算法对各方法生成的节点表示进行聚类实验，其中 K-means 的类别数设置

为节点标签的类别数，之后利用 *NMI* 指标度量聚类结果与真实类别分布之间的相似性. 实验结果如表 5 所示，其中每个数据集上的最优结果均用粗体标出，OM 表示实验过程中超出内存限制.

表 5 以 *NMI* 为指标的节点聚类实验结果

方法	Brazil	Europe	USA	Reality	Actor	Film	Coauthor-Network	Coauthor-System	Cora	Citeseer
RolX	0.554	0.106	0.091	0.004	0.121	0.006	0.057	0.052	0.035	0.059
DRNE	0.462	**0.341**	0.218	0.015	0.108	0.049	0.079	0.058	0.055	0.031
RESD	**0.560**	0.312	**0.300**	**0.151**	0.170	**0.051**	0.098	0.090	0.010	0.032
GAS	0.407	0.259	0.270	0.134	0.107	0.003	OM	OM	**0.077**	0.045
GraphWave	0.503	0.320	0.268	0.122	**0.190**	OM	OM	OM	0.043	0.042
RIDeRs	0.558	0.283	0.234	0.005	0.101	0.046	0.044	0.023	0.032	**0.060**
struc2gauss	0.510	0.148	0.210	0.056	0.080	0.027	OM	OM	0.055	0.017
struc2vec	0.467	0.188	0.184	0.136	0.084	0.008	0.033	0.028	0.007	0.013
role2vec	0.136	0.057	0.064	0.005	OM	OM	0.000	0.001	0.018	0.004
SEGK	0.392	0.252	0.210	0.105	0.138	OM	**0.164**	**0.161**	0.025	0.024
GraLSP	0.216	**0.186**	**0.229**	0.001	0.028	OM	OM	OM	0.178	0.057
SDNE	0.263	0.177	0.202	0.002	**0.085**	**0.047**	**0.032**	0.002	0.113	0.013
node2vec	0.079	0.043	0.102	0.002	0.019	0.007	0.008	**0.024**	**0.448**	**0.229**
GraphSTONE	0.123	0.084	0.214	0.001	0.050	0.018	OM	OM	0.237	0.109
VGAE	**0.264**	0.098	0.194	**0.004**	0.062	0.004	OM	OM	0.346	0.162

(1) 基于角色或社团的网络表示学习方法学得的节点表示在聚类实验中也呈现出显著的性能差异. 与分类结果一致的是，角色导向的方法在前 8 个网络中获得了较高的 *NMI* 值，社团导向的方法在最后两个网络上表现出绝对的性能优势. 但特别的是，GraLSP 在航空网络上获得了与 SDNE 相近的 *NMI* 得分，该得分甚至优于少数几个角色导向的网络嵌入方法，这可能是因为 GraLSP 中的匿名随机游走在一定程度上可以捕获节点的局部结构信息.

(2) 综合比较前 8 个数据集上不同角色导向网络嵌入方法的 *NMI* 得分，有如下观察和分析：

① 基于局部特征的方法中，RESD 获得的聚类结果更趋近于真实的分布情况，这表明在未知任何样本类别信息时，该方法学得的节点表示更能反映节点间的相似性关系. 但值得注意的是，RESD 在更大规模的作者合作网络上表现得不如 SEGK，这与节点分类的结果相一致，也就说明基于高阶特征的网络嵌入方法可以更好地捕获大规模网络中的节点差异和重要模式.

② 基于全局特征的方法中，GraphWave 在各个网络上的 *NMI* 得分都与 RESD 更相近，甚至在演员合作网络上达到最优，这表明以谱图小波扩散代替手工提取的拓扑结构特征来建模结构化角色是可行且有效的.

③ 基于高阶特征的方法中，role2vec 的 *NMI* 得分最低，这与预期相符合. struc2vec 在电话通讯网络上表现较好，但在其它网络上效果不佳. SEGK 则在作者合作网络上表现出更大潜力.

从以上分析可以看出，聚类结果中存在很多与分类结果相类似的地方，这表明学习更高质量的节点表示是提升各种下游任务性能的关键. 但值得注意的是，同一方法在同一网络上学得的节点表示在不同任务上也可能会有不同的表现. 例如 GAS 在航空网络上表现出的分类能力相对突出，但聚类能力就较为一般. 而在电话通讯网络上分类结果一般的 RESD 和 struc2vec 却在节点聚类时表现极佳，甚至好于 GAS、GraphWave 和 SEGK 方法. 本文进行这些实验的目的是找到一般规律，但是具体情况还需具体分析.

4.6 鲁棒性

事实上，真实网络中总是存在偏差和噪声的，所以探究网络表示学习方法在不同网络攻击策略下的鲁棒性一直是该领域的重要研究问题[86-89]. 然而，目前针对角色导向的网络表示学习方法的鲁棒性研究十分匮乏. 基于此，本文设计了四种网络攻击策略来探究不同噪声或扰动对角色导向的网络表示学习方法造成的性能影响. 其攻击策略包括：(1) 根据网络中的节点数随机改变节点标签；(2) 根据网络中的连边数随机增加链接；(3) 依概率对网络中的链接进行随机重连；(4) 改变测试集中的类别分布.

在该实验中，我们仅选取 3 个航空网络和 2 个引文网络作为实验数据集，其原因是在大规模的数据集上反复训练这些方法的节点表示不仅要花费很高的时间成本，而且许多方法因为内存问题无法展示相应的实验结果，这对问题的分析无益. 此外，还需要说明的是，由于 role2vec 在欧洲和美国航空网络上学习节点表示需要大量时间，但其学习的节点表示又不适用于与角色有关的下游任务，所以在该实验中我们省略了一些 role2vec 的实验结果.

第一种攻击策略：我们随机改变网络中 2%、4%、6%、8%和 10%的节点类别标签，然后进行分类实验. 结果如图 7 所示，其中横坐标为 0 时的 Micro-*F*1 值表示在原始类别标签下的分类结果.

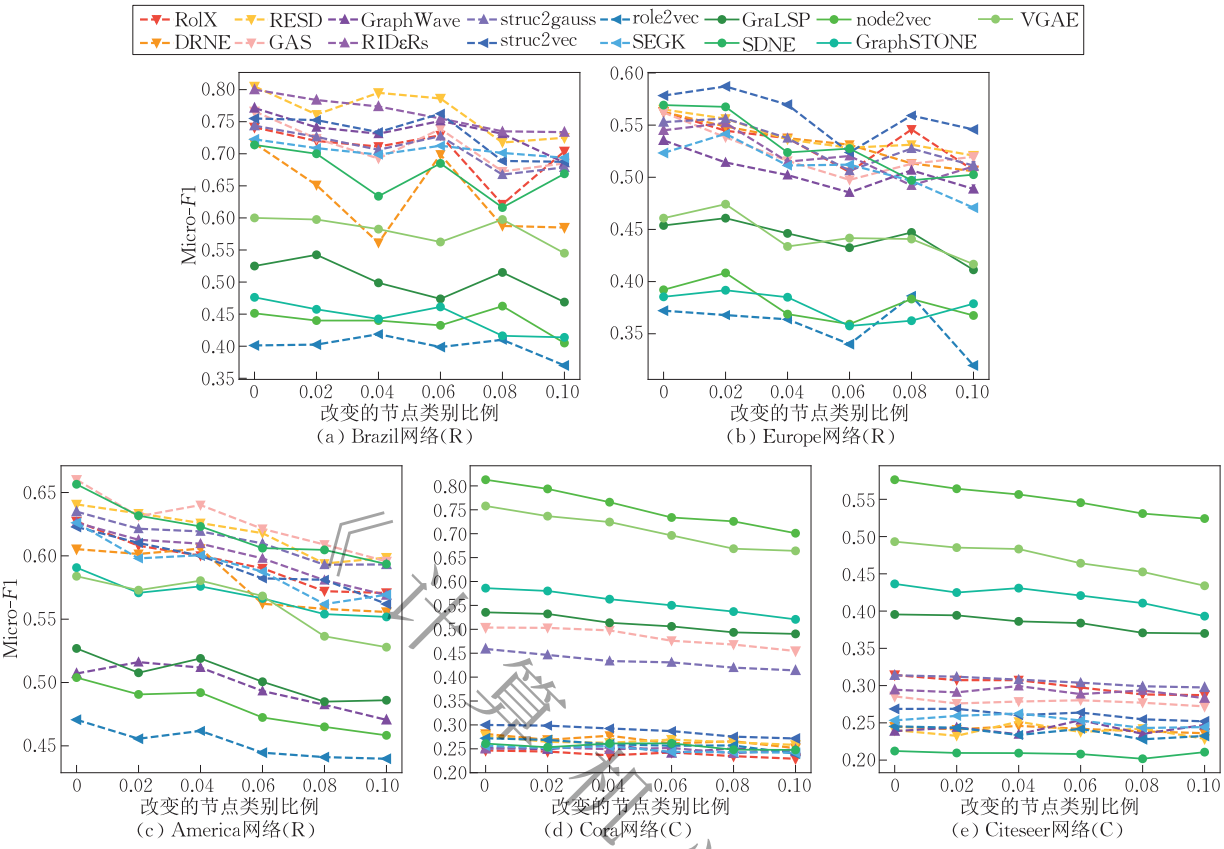

图 7　攻击策略 1：按比例随机改变节点标签

由图 7 可以看到，随着扰动比例的增加，角色导向的网络表示学习方法在三个航空网络上的分类结果呈现出明显且相似的下降趋势，其原因在于这种攻击策略并不影响节点表示的生成，而 Micro-*F*1 值的下降主要源于错误标签的引入. 但值得注意的是，角色导向的网络嵌入方法在两个引文网络上表现出较强的鲁棒性，这可能是因为这类方法本身就不适用于社团结构的挖掘，所以扰动的增加对其影响并不明显，这在其它攻击策略中也有类似的表现.

第二种攻击策略：我们在原始网络上随机增加 5%～30%的链接，然后学习新的节点表示并进行分类任务. 由图 8 可以看到，随着航空网络链接数的增加，大部分方法的分类曲线表现出明显的震荡趋势. 这是因为这种攻击策略改变了节点原本的拓扑结构，使得角色导向的网络嵌入方法难以根据现有的结构特征推断出节点原始的角色信息.

(1) 基于局部特征的方法受这种网络攻击的影响最大，如欧洲航空网络上 GAS 的分类曲线以及美国航空网络上 DRNE 的分类曲线. 其原因在于网络链接的增加首先会改变节点的局部拓扑，这使得扰动后的局部特征与原始的局部特征之间会产生较大的偏差，进而影响节点表示的生成. RESD 因为将节点特征映射到概率分布上，所以有效地提高了模型的鲁棒性.

(2) 基于全局特征的方法受这种网络攻击的影响相对较小，这是因为这类方法更关注于节点全局的连接模式，微小的扰动不会对其产生较大的影响. 此外，struc2gauss 在生成节点嵌入时建模了网络的不确定性.

(3) 基于高阶特征的方法中，role2vec 和 struc2vec 表现出较高敏感性，前者与其不适用性有关，后者与其利用节点度构建多层加权图有关. SEGK 在欧洲航空网络上表现出较好的鲁棒性.

此外，我们从图 8 中还可以观察到，无论是基于局部特征、全局特征还是高阶特征的角色导向网络表示学习方法，它们在引文网络上的表现几乎不受该网络攻击的影响.

第三种攻击策略：我们对每一个网络中的原始链接分别以 0.05～0.3 的概率进行随机重连. 实验结果如图 9 所示，我们发现这种攻击策略相比于前

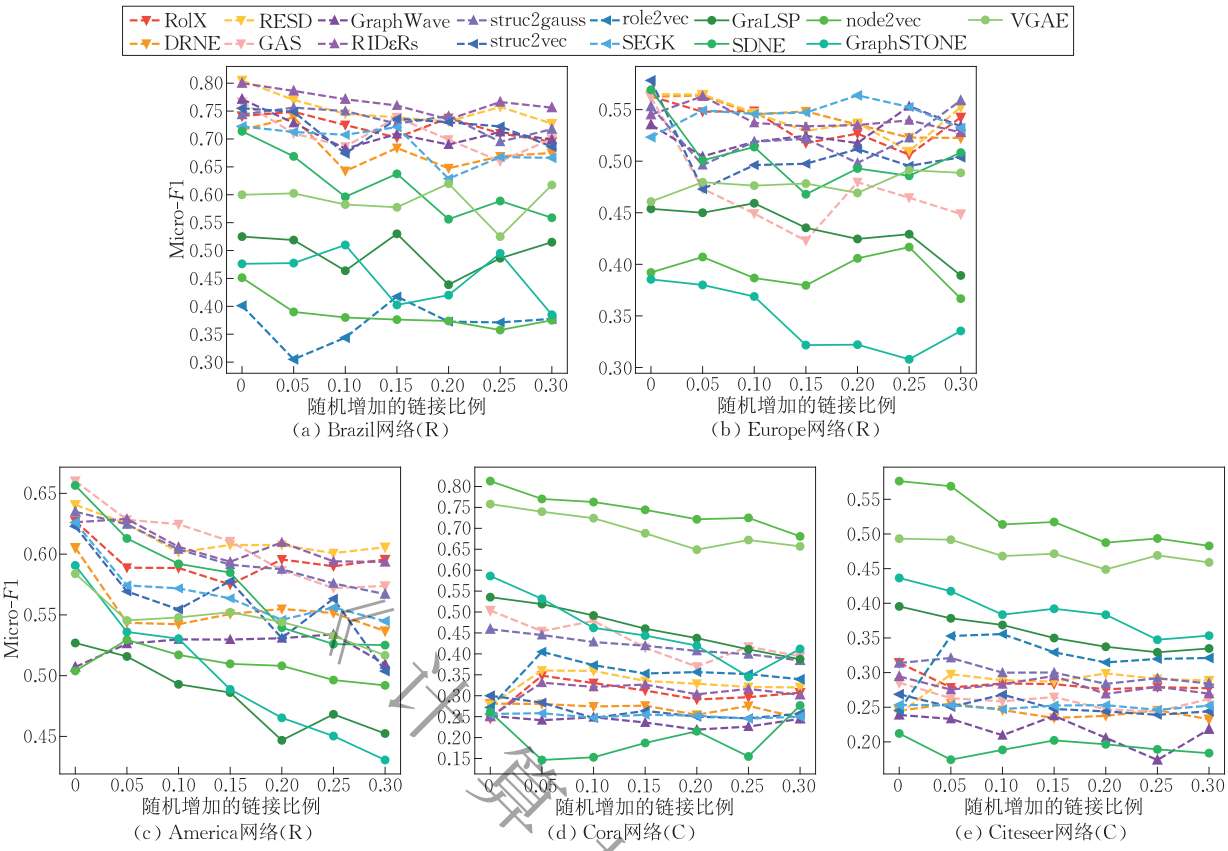

图 8 攻击策略 2:按比例随机增加链接

两种攻击策略对性能的影响更大. 这是因为第一种攻击策略实际上并未改变原始网络的拓扑结构,而第二种攻击策略虽然改变了网络的拓扑结构,但原有的连接模式并未发生改变. 然而,第三种攻击策略在不改变链接总数的情况下对已有链接进行随机重连,不仅会影响节点的局部特征,也会加剧节点全局特征的变化. 由图 9 可以看到,角色导向的网络表示学习方法在巴西航空网络以及美国航空网络上表现出了相近的变化趋势,只有在欧洲航空网络上基于局部特征的方法才表现出更大的波动. 并且值得注意的是,社团导向的网络表示学习方法也对这一网络攻击表现出极高的敏感性.

第四种攻击策略:我们改变测试集中的类别分布,即依次删除测试集中隶属同一标签的节点集合. 实验结果如表 6 和表 7 所示,其分别记录了在该攻击策略下 15 种网络表示学习方法在三个航空网络和两个引文网络上的分类结果.

从表 6 可以看出,对该攻击策略鲁棒的方法往往是更能准确捕获节点角色信息的方法. 例如在巴西航空网络上表现优异的 RESD 和 RIDεRs 在删除任何一类节点集合后依旧表现出较好的分类能力,变化趋势相对平缓,但性能较差的 role2vec 则对这一网络攻击表现出较大幅度的波动. 此外,我们还发现同一方法对删除不同类别的节点表现出不同的鲁棒性,而这种差异在不同方法中又表现出一致性. 例如在巴西航空网络上,删除标签为 0 的节点集合相比删除标签为 3 的节点集合对各种角色导向的方法造成的性能影响都是更大的. 这可能是由不同角色识别的难易程度不同导致的. 换言之,这些方法可能因为更容易捕获与角色 0 有关的信息,所以对这类节点的分类能力更为突出. 当这类节点被删去时自然会引起性能的大幅下降.

从表 7 可以看出,这种攻击策略对社团导向的方法的性能影响是微弱的,尤其是在 node2vec 和 VGAE 上. 我们推测这可能是因为不同的社团结构本质上都是连接紧密的节点集合,所以同一方法在捕获不同社团结构的能力上是相似的.

总结. 综合以上分析可以看出,在面对前两种网络攻击时,社团导向和角色导向的网络表示学习方法分别在航空网络和引文网络上表现出相似的敏感性. 但在面临网络链接的随机重连时,社团导向的

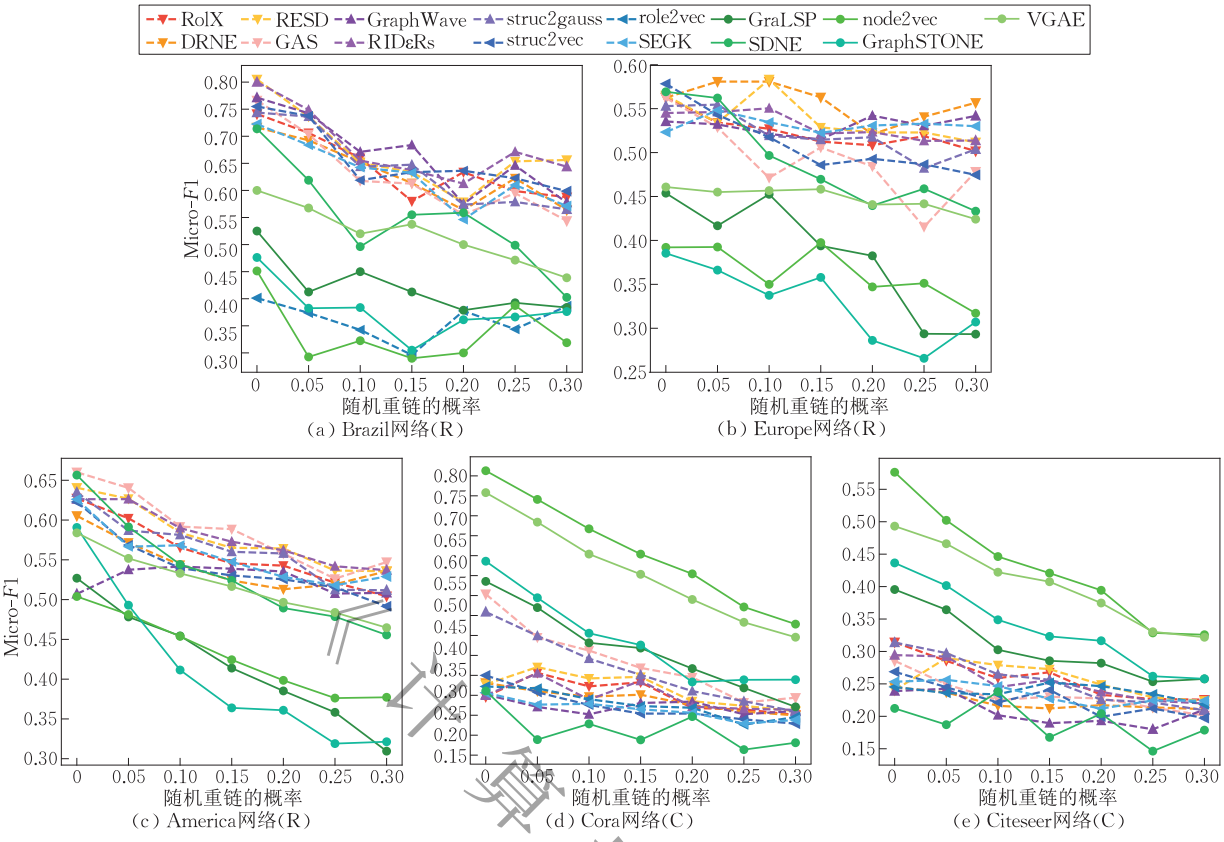

图 9　攻击策略 3：依概率对原始链接进行随机重连

表 6　攻击策略 4：依次删除航空网络测试集中隶属同一标签的节点集合

方法	Brazil 网络					Europe 网络					America 网络				
	原始	0	1	2	3	原始	0	1	2	3	原始	0	1	2	3
RolX	0.741	0.681	0.777	0.768	0.740	0.563	0.492	0.592	0.634	0.533	0.627	0.562	0.665	0.680	0.576
DRNE	0.716	0.671	0.735	0.789	0.712	0.563	0.468	0.606	0.626	0.500	0.605	0.524	0.683	0.649	0.559
RESD	0.805	0.793	0.809	0.825	0.797	0.565	0.503	0.587	0.650	0.525	0.641	0.588	0.692	0.702	0.619
GAS	0.766	0.699	0.793	0.781	0.745	0.562	0.462	0.568	0.620	0.533	0.660	0.585	0.694	0.738	0.603
GraphWave	0.771	0.703	0.774	0.808	0.749	0.535	0.405	0.600	0.634	0.455	0.507	0.454	0.658	0.590	0.389
RIDεRs	0.800	0.773	0.805	0.812	0.844	0.545	0.469	0.539	0.643	0.491	0.626	0.564	0.647	0.697	0.569
struc2gauss	0.744	0.700	0.772	0.820	0.694	0.553	0.502	0.556	0.616	0.533	0.635	0.559	0.682	0.696	0.597
struc2vec	0.755	0.708	0.790	0.797	0.755	0.578	0.527	0.609	0.631	0.577	0.623	0.568	0.669	0.706	0.582
role2vec	0.401	0.317	0.534	0.439	0.389	0.372	0.251	0.424	0.426	0.378	0.471	0.363	0.507	0.521	0.446
SEGK	0.723	0.681	0.708	0.733	0.720	0.523	0.460	0.560	0.612	0.527	0.626	0.553	0.636	0.697	0.566
GraLSP	0.525	0.455	0.631	0.602	0.494	0.454	0.377	0.550	0.518	0.417	0.527	0.428	0.576	0.613	0.477
SDNE	0.714	0.625	0.687	0.747	0.656	0.569	0.519	0.561	0.615	0.565	0.657	0.598	0.697	0.693	0.622
node2vec	0.451	0.391	0.472	0.493	0.410	0.392	0.350	0.388	0.422	0.393	0.504	0.453	0.528	0.537	0.534
GraphSTONE	0.476	0.485	0.446	0.529	0.417	0.385	0.358	0.452	0.442	0.326	0.591	0.515	0.644	0.648	0.552
VGAE	0.600	0.512	0.715	0.684	0.549	0.461	0.339	0.539	0.551	0.428	0.584	0.497	0.661	0.647	0.559

方法会受到更大的影响. 相反的，测试集中类别分布不均对角色导向的方法的性能影响更甚. 值得注意的是，在四种网络攻击下，角色导向的方法在引文网络上均表现出了极强的鲁棒性.

另外，我们从分类机制的角度讨论了角色导向的网络表示学习方法在四种攻击策略下的鲁棒性. 当网络攻击（如随机增加网络链接）显著改变节点的局部结构时，基于全局特征或高阶特征的方法会表现出相对较好的鲁棒性. 但是当网络攻击（如网络链接发生随机重连）加剧节点全局特征的变化时，这些方法也将具有较高的敏感性. 目前，已有一些角色导向的网络表示学习方法尝试建模网络中的不确定性，以此提高模型的鲁棒性，如 RESD、struc2gauss.

表 7 攻击策略 4:依次删除引文网络测试集中隶属同一标签的节点集合

方法	Cora 网络								Citeseer 网络						
	原始	0	1	2	3	4	5	6	原始	0	1	2	3	4	5
RolX	0.247	0.267	0.245	0.230	0.232	0.258	0.251	0.248	0.314	0.311	0.249	0.334	0.311	0.323	0.329
DRNE	0.281	0.295	0.268	0.258	0.272	0.294	0.290	0.283	0.245	0.265	0.203	0.261	0.241	0.262	0.237
RESD	0.277	0.282	0.282	0.251	0.264	0.290	0.284	0.281	0.240	0.258	0.191	0.264	0.243	0.240	0.247
GAS	0.504	0.526	0.512	0.474	0.488	0.519	0.531	0.518	0.286	0.225	0.235	0.340	0.267	0.306	0.276
GraphWave	0.252	0.290	0.263	0.255	0.184	0.268	0.271	0.283	0.239	0.256	0.152	0.285	0.220	0.233	0.270
RIDεRs	0.250	0.264	0.257	0.246	0.236	0.267	0.257	0.263	0.294	0.294	0.233	0.317	0.296	0.322	0.305
struc2gauss	0.459	0.476	0.464	0.418	0.477	0.468	0.479	0.473	0.314	0.292	0.262	0.345	0.327	0.345	0.331
struc2vec	0.300	0.310	0.299	0.267	0.297	0.304	0.313	0.318	0.269	0.254	0.224	0.285	0.269	0.280	0.283
role2vec	0.272	0.277	0.264	0.275	0.260	0.275	0.270	0.276	0.245	0.241	0.235	0.240	0.252	0.251	0.236
SEGK	0.256	0.256	0.253	0.231	0.285	0.282	0.261	0.266	0.253	0.251	0.217	0.288	0.263	0.262	0.268
GraLSP	0.536	0.537	0.526	0.498	0.569	0.541	0.546	0.541	0.396	0.372	0.363	0.410	0.405	0.428	0.406
SDNE	0.261	0.292	0.227	0.224	0.265	0.310	0.255	0.237	0.212	0.145	0.143	0.250	0.223	0.245	0.244
node2vec	0.813	0.822	0.814	0.790	0.812	0.814	0.811	0.812	0.576	0.544	0.553	0.576	0.595	0.597	0.573
GraphSTONE	0.586	0.616	0.578	0.559	0.588	0.600	0.606	0.584	0.437	0.417	0.393	0.479	0.440	0.436	0.434
VGAE	0.758	0.769	0.751	0.746	0.762	0.751	0.758	0.763	0.493	0.464	0.468	0.493	0.510	0.520	0.487

4.7 参数敏感性

为探究不同维度的节点表示对下游任务的影响,我们选取航空网络、演员合作网络以及引文网络进行敏感性分析,维度依次设置为 2、4、8、16、32、64 和 128. 实验结果如图 10 所示,横坐标表示不同嵌入维度的对数值. 需要说明的是,我们未对 GraphWave 和 RIDεRs 进行敏感性分析,这是因为这两种算法生成节点表示的维度与其它超参数有关.

由图 10 可以看到,各个方法的分类性能最先随着嵌入维度的增长而得到明显的改善,当嵌入维度达到最优阈值后,模型性能趋于平稳. 其原因是节点嵌入维度的增长有助于表征更多的网络结构信息. 但有时过高的嵌入维度容易导致过拟合问题,如部分算法在较高的嵌入维度上反倒表现不佳. 此外,我们按照分类方式比较了不同角色导向的网络表示学习方法随维度变化表现出的性能波动,并未发现显著的差异性.

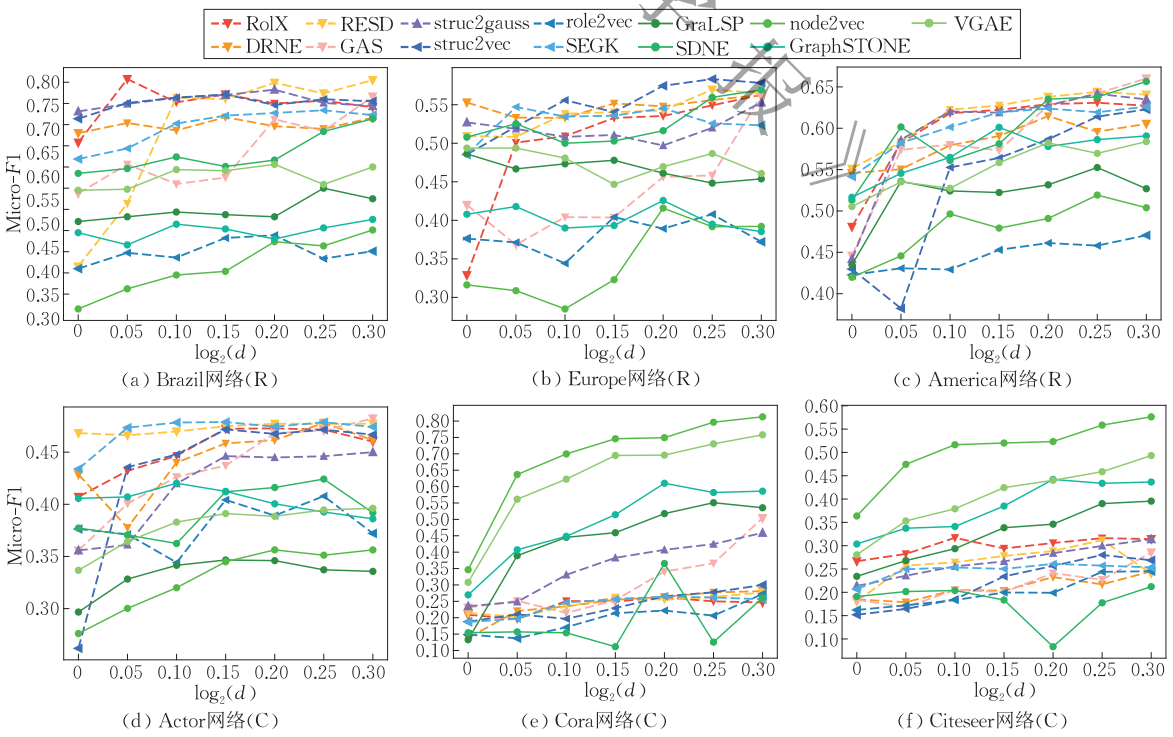

图 10 不同嵌入维度下的节点分类实验结果

图 11　角色导向的网络表示学习平台

4.8　平台介绍

除了综合性实验外，本文还搭建了一个角色导向的网络表示学习平台，如图 11 所示. 平台的主要功能和服务如下：

(1) 数据集. 该平台提供了本文涉及到的所有实验数据集. 用户可选择或下载已有数据集，也可上传其它数据集(需满足指定格式). 针对每一个数据集，该平台提供以下两种功能：一是网络可视化，节点颜色与其标签类别相对应；二是计算并展示网络的基本统计信息，如节点数、边数、聚集系数.

(2) 算法. 该平台提供了十种以上的角色导向的网络表示学习方法. 用户可根据自定义的模型超参数运行这些方法，并获得算法的实时运行日志以便使用与研究.

(3) 结果. 针对每个算法的运行结果，该平台提供两方面的功能：一是下载学得的节点表示. 下载的文件内容是一个 $|\mathcal{V}|$ 行 d 列的矩阵，其中 $|\mathcal{V}|$ 是网络的节点个数，d 是节点表示的维度；二是可以对学得的节点表示进行可视化、节点分类、节点聚类实验，并将其图形化结果展示在界面中.

5　研究挑战与展望

目前，角色导向的网络表示学习研究已表现出巨大的发展空间和应用潜力，具体如表 8 所示. 但是该领域的研究仍然存在许多不足和需要进一步完善的地方. 在此，我们针对角色导向的网络表示学习和相关研究领域的研究现状，对其面临的研究挑战和未来发展方向进行了分析和讨论.

5.1　角色的网络解释

传统的直接从网络出发的角色发现方法都基于网络上预定义的节点等价性[34]，如自同构等价性[90]、规则等价性[91-92]和随机等价性[90,93]等. 虽然很多角色导向的网络表示学习研究都声称其算法可以捕获某种等价性，但对于绝大多数算法，目前并不能解释由其判别为相似的节点在网络上是如何等价的. 然而，对网络中节点角色的含义进行解释必须要描述节点之间的等价关系. 因此，当前角色导向的网络表示学习无法对其捕获的关系或相似性在网络上进行解释，这是其发展道路上的首要难题. 如果这一难题得到解决，角色发现在离散的网络空间和连续的嵌入空间之间将不存在统一研究的壁垒，这将极大地促进网络科学的发展.

5.2　角色表示的鲁棒性

真实世界中的网络通常充斥着大量的噪声或扰动，数据所给出的节点交互关系未必反映了真实世界中对应实体间的真正关系[86-89]. 角色导向的网络表示学习方法因为依赖于节点的结构特征而极易受到网络噪声或扰动的影响. 如何提升角色导向的网络表示学习方法的鲁棒性是未来研究的重点之一. 本文在第 4.6 节中尝试给出 4 种不同的网络攻击策略来测试现有角色导向网络表示学习方法的鲁棒性，望提升读者对该问题的兴趣和注意力.

5.3　角色表示对图神经网络的增强

图神经网络是当下最热门的深度网络表示学习方法[20]. 几乎所有图神经网络都是通过边上的消息传播来更新节点表示的. 由于大多数研究都专注于消息传播机制的设计[31,48,99-101]，使得这些方法更注

表 8 角色导向的网络表示学习在图相关领域的任务和应用

任务名称	英文描述	目标	常用数据集	文献
可视化	Visualization	在节点表示上运行 t-SNE 或 PCA 算法可视化网络中的个体差异和重要模式	Barbell network, air-traffic network, diseasome network, ca-netscience network	[39],[45],[49],[57],[62],[63],[72],[94]
节点分类	Node classification	根据连接模式对具有相似功能作用(角色)的节点进行有监督的类别划分	air-traffic network, Reality-call network, Hospital network, Coauthor network, Expressway network, Actor network, Film network, Enron email network	[45],[49],[56],[57],[63],[65],[68],[69],[72],[75],[94],[95]
图分类	Graph classification	由一组向量(角色导向的节点表示)表示单个网络,并对该网络进行类别划分	MUTAG, ENZYMES, IMDB, COLLAB, REDDITBINARY, REDDIT-MULTI5K	[63], [69]
节点聚类	Node clustering	根据连接模式对具有相似功能作用(角色)的节点进行无监督聚类	air-traffic network, Enron email network	[39],[57],[63],[94]
链接预测	Link prediction	根据连接模式预测网络中缺失的边	soc-wiki-Vote, tech-routers-rf, soc-anybeat	[62], [65], [76]
相似性搜索	Top-k similarity search	挖掘网络中结构最为相似的 k 个节点,如机器人检测(Bot detection)	Wikipediatalk network, Coauthor network, ACL anthology network	[49], [63], [69]
身份解析	Identity resolution	根据连接模式的相似性匹配跨网络的节点身份,如网络对齐(Network alignment)、用户缝合(User Stitching)	Facebook, Arxiv, DBLP, PPI, Arenas, Coauthor network, citeseer, yahoo, wiki, digg, bitcoin, Web logs	[13],[43],[51],[62]
迁移学习	Transfer learning	提取网络 $\mathcal{G}$ 中的角色并在网络 $\mathcal{G}'$ 上学习同组角色	IP data, yahoo-msg, air-traffic network	[39],[41],[65],[75]
角色演化	Role transition	给定动态网络,分析网络角色的演化	IP trace network, Enron, Reality, Slashdot, Facebook; Political networks	[96],[97]
角色预测	Role prediction	给定动态网络,预测节点未来的成员分布,如预测政治网络上的党派分布	Enron, Reality, Slashdot, Facebook	[96],[97],[98]
异常检测	Anomaly detection	检测动态网络中的异常节点或行为	Enron	[96]

重节点间的邻近性而忽略了网络拓扑尤其是节点角色的作用.若将节点角色等网络拓扑信息引入图神经网络,能够增加其输入的信息总量,使得最终学到的表示包含更丰富的信息,具备更强的表征能力.此外,由于节点角色对节点间交互的决定性作用,在设计消息传播机制时考虑节点角色的影响能很大地提升图神经网络的性能,这一点在如 Geom-GCN[30]、GraLSP[81]、GraphSTONE[83]等一些尝试用角色表示对图神经网络增强的研究中得到了验证.因此,我们认为利用角色表示增强图神经网络将成为未来图神经网络研究的重点之一.但是,现有角色导向网络表示学习方法对角色的网络解释能力的缺乏极大地制约了这种研究的发展.

5.4 社团与角色的融合研究

社团隶属体现了节点的同质性,角色类型体现了节点的异质性.网络的形成与演化是在节点的同质性与异质性共同作用下发生的.例如微博社交网络中一个用户产生新的关注行为可能是由于其与被关注者存在共同的兴趣爱好,并且被关注者有很高的知名度.舆情在社交关系上的传播速度与形成的级联大小受传播者的身份影响很大.因此,同时融合节点的社团信息和角色信息来实现各种网络分析任务如链接预测、网络重构、信息传播等是十分必要的.REACT[102]和 SPaE[103]尝试用网络表示学习同时挖掘网络中的社团结构和结构化角色,但其未构建起社团和角色间的相互作用.目前有极少研究如 MMOR[104]尝试利用贝叶斯图模型同时刻画社团和角色的交互生成过程,但还未有工作基于网络表示学习进行相关研究.

5.5 更复杂网络上的角色导向网络表示学习

几乎所有现有的角色导向网络表示学习方法只针对一般的同质静态网络进行设计.然而,大多数真实世界中的网络的拓扑结构是动态变化的,且其节点和边的类型是繁杂的.目前相关研究中仅有 DBMM[96]和 DyNMF[97]在动态网络上扩展了基于矩阵分解的 RolX 方法,鲜有角色导向网络表示学习在其它更复杂网络上的应用.因此,角色导向网络表示学习亟需被扩展到符号网络[105]、动态网络[106]、异质信息网络[107]等更复杂的网络类型上,以提升其对真实网络任务的适用性和针对性.面对这些不同的网络,角色导向的网络表示学习亦面对不同挑战.例如在动态网络中,不仅要识别节点在各个时刻的角色,亦要刻画节点在相邻时刻的角色演化过程.而在异质信息网络中,捕获节点间结构相似性的同时应考虑到多样的节点类型.由不同类型节点构成的相同的子图结构会具有完全不同的内涵[108-109].

5.6 人类移动行为模式中的角色

挖掘人类移动行为模式一直是复杂系统研究的

热门话题[110-112]. 近些年，研究学者通过将人类的移动行为构建成复杂网络来分析人类移动行为模式[113-114]. 这类方法通常专注于捕获人类迁移网络中的共性模式，而忽略了节点个性（如角色）对移动行为的影响. 事实上，在人类迁移网络上研究角色问题存在两个难点：其一是难以刻画时间和空间信息；其二是难以知晓这些信息对不同角色节点的影响方式. 若这些问题得以解决，则可以实现对人类行为更个性化的预测，有助于政府政策和措施的制定以及宏观经济情况变化的预测.

6 总　结

角色是与社团不同但又互补的概念. 目前针对社团导向的网络表示学习方法的综合性研究已经非常充分，但仍缺乏对角色导向的网络表示学习方法的系统性研究. 本文辨析了社团导向和角色导向的网络表示学习的区别，并系统性地从特征提取的角度对现有角色导向的网络表示学习方法进行了分类. 本文在十个常用数据集上对 15 种社团导向或角色导向的网络表示学习方法进行了综合性实验，一方面深入探讨和分析了社团与角色这两个概念的内在区别，另一方面结合新的分类机制着重对比了不同的角色导向网络表示学习方法之间的性能差异. 另外，本文还特别提出了一套针对角色导向网络表示学习鲁棒性的评测方法. 此外，我们提供了一个集数据、算法、分析于一体的复杂网络分析与应用平台用于典型算法的实验与分析，以促进该领域的长足发展. 最后，本文对角色导向的网络表示学习面临的挑战和未来发展趋势进行了总结和展望.

参　考　文　献

[1] Brugere I, Gallagher B, Berger-Wolf T Y. Network structure inference, a survey: Motivations, methods, and applications. ACM Computing Surveys, 2018, 51(2): 1-39

[2] Song C, Havlin S, Makse H A. Self-similarity of complex networks. Nature, 2005, 433(7024): 392-395

[3] Papadopoulos F, Kitsak M, Serrano M Á, et al. Popularity versus similarity in growing networks. Nature, 2012, 489(7417): 537-540

[4] Benson A R, Gleich D F, Leskovec J. Higher-order organization of complex networks. Science, 2016, 353(6295): 163-166

[5] Grilli J, Barabás G, Michalska-Smith M J, et al. Higher-order interactions stabilize dynamics in competitive network models. Nature, 2017, 548(7666): 210-213

[6] Recanatesi S, Farrell M, Lajoie G, et al. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. Nature Communications, 2021, 12(1): 1417

[7] Sen P, Namata G, Bilgic M, et al. Collective classification in network data. AI Magazine, 2008, 29(3): 93-106

[8] McDowell L K, Gupta K M, Aha D W. Cautious collective classification. The Journal of Machine Learning Research, 2009, 10(96): 2777-2836

[9] Serrano M Á, Boguna M. Clustering in complex networks. I. general formalism. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 2006, 74(5): 056114

[10] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks//Proceedings of the 12th International Conference on Information and Knowledge Management. LA, USA, 2003: 556-559

[11] Martínez V, Berzal F, Cubero J-C. A survey of link prediction in complex networks. ACM Computing Surveys, 2017, 49(4): 1-33

[12] Akoglu L, Tong Hanghang, Koutra D. Graph based anomaly detection and description: A survey. Data Mining and Knowledge Discovery, 2015, 29(3): 626-688

[13] Heimann M, Shen H, Safavi T, et al. REGAL: Representation learning-based graph alignment//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino, Italy, 2018: 117-126

[14] Rossi R A, Ahmed N K, Zhou R, et al. Interactive visual graph mining and learning. ACM Transactions on Intelligent Systems and Technology, 2018, 9(5): 1-25

[15] Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(11): 2579-2605

[16] Li J, Wu L, Guo R, et al. Multi-level network embedding with boosted low-rank matrix approximation//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Vancouver, Canada, 2019: 49-56

[17] Qiu J, Dong Y, Ma H, et al. NetSMF: Large-scale network embedding as sparse matrix factorization//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 1509-1520

[18] Zhang Z, Cui P, Wang X, et al. Arbitrary-order proximity preserved network embedding//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London, UK, 2018: 2778-2786

[19] Xiong H, Yan J. BTWalk: Branching tree random walk for multi-order structured network embedding. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(8): 3611-3628

[20] Xu Bing-Bing, Cen Ke-Ting, Huang Jun-Jie, et al. A survey on graph convolutional neural network. Chinese Journal of Computers, 2020, 43(5): 755-780(in Chinese)

(徐冰冰，岑科廷，黄俊杰等. 图卷积神经网络综述. 计算机学报，2020，43(5)：755-780)

[21] Khosla M，Setty V，Anand A. A comparative study for unsupervised network representation learning. IEEE Transactions on Knowledge and Data Engineering，2021，33(5)：1807-1818

[22] Cai H，Zheng V W，Chang K C C. A comprehensive survey of graph embedding：Problems，techniques，and applications. IEEE Transactions on Knowledge and Data Engineering，2018，30(9)：1616-1637

[23] Cui P，Wang X，Pei J，et al. A survey on network embedding. IEEE Transactions on Knowledge and Data Engineering，2019，31(5)：833-852

[24] Zhang D，Yin J，Zhu X，et al. Network representation learning：A survey. IEEE Transactions on Big Data，2020，6(1)：3-28

[25] Qi Jin-Shan，Liang Xun，Li Zhi-Yu，et al. Representation learning of large-scale complex information network：Concepts，methods and challenges. Chinese Journal of Computers，2018，41(10)：2394-2420(in Chinese)
(齐金山，梁循，李志宇等. 大规模复杂信息网络表示学习：概念、方法与挑战. 计算机学报，2018，41(10)：2394-2420)

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

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

[28] López J A M，Arregui-Garca B，Bentkowski P，et al. Anatomy of digital contact tracing：Role of age，transmission setting，adoption and case detection. Science Advances，2021，7(15)：eabd8750

[29] Tian L，Bashan A，Shi D N，et al. Articulation points in complex networks. Nature Communications，2017，8(1)：1-9

[30] Pei H，Wei B，Chang K，et al. Geom-GCN：Geometric graph convolutional networks//Proceedings of the 8th International Conference on Learning Representations. Virtual Conference，Ethiopia，2020：1-12

[31] You J，Gomes-Selman J，Ying R，et al. Identity-aware graph neural networks//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Conference，Canada，2021：10737-10745

[32] You J，Ying R，Leskovec J. Position-aware graph neural networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach，USA，2019：7134-7143

[33] Mitra B，Sural S，Vaidya J，et al. A survey of role mining. ACM Computing Surveys，2016，48(4)：1-37

[34] Rossi R A，Ahmed N K. Role discovery in networks. IEEE Transactions on Knowledge and Data Engineering，2014，27(4)：1112-1131

[35] Zhang Shu-Sen，Liang Xun，Qi Jin-Shan. A review on role identification methods in social networks. Chinese Journal of Computers，2017，40(3)：649-673(in Chinese)
(张树森，梁循，齐金山. 社会网络角色识别方法综述. 计算机学报，2017，40(3)：649-673)

[36] Hamilton W L，Ying R，Leskovec J. Representation learning on graphs：Methods and applications. IEEE Data Engineering Bulletin，2017，40(3)：52-74

[37] Rossi R A，Jin D，Kim S，et al. On proximity and structural role-based embeddings in networks：Misconceptions，techniques，and applications. ACM Transactions on Knowledge Discovery from Data，2020，14(5)：1-37

[38] Grover A，Leskovec J. node2vec：Scalable feature learning for networks//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco，USA，2016：855-864

[39] Donnat C，Zitnik M，Hallac D，et al. Learning structural node embeddings via diffusion wavelets//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London，UK，2018：1320-1329

[40] Henderson K，Gallagher B，Lei L，et al. It's who you know：Graph mining using recursive structural features//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego，USA，2011：663-671

[41] Henderson K，Gallagher B，Eliassi-Rad T，et al. RolX：Structural role extraction & mining in large graphs//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing，China，2012：1231-1239

[42] Rissanen J. Modeling by shortest data description. Automatica，1978，14(5)：465-471

[43] Gilpin S，Eliassi-Rad T，Davidson I. Guided learning for role discovery(GLRD)：Framework，algorithms，and applications //Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago，USA，2013：113-121

[44] Jin D，Heimann M，Safavi T，et al. Smart roles：Inferring professional roles in email networks//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage，USA，2019：2923-2933

[45] Tu K，Cui P，Wang X，et al. Deep recursive network embedding with regular equivalence//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London，UK，2018：2357-2366

[46] Hochreiter S，Schmidhuber J. Long short-term memory. Neural Computation，1997，9(8)：1735-1780

[47] Guo X，Zhang W，Wang W，et al. Role-oriented graph auto-encoder guided by structural information//Proceedings of the 25th International Conference on Database Systems for Advanced Applications. Jeju-Island，Korea，2020：466-481

[48] Kipf T N，Welling M. Semi-supervised classification with graph convolutional networks//Proceedings of the 5th International Conference on Learning Representations. Palais des Congrès Neptune，Toulon，France，2017：1-14

[49] Zhang W, Guo X, Wang W, et al. Role-based network embedding via structural features reconstruction with degree-regularized constraint. Knowledge-Based Systems, 2021, 218(C): 106872

[50] Kingma D P, Welling M. Auto-encoding variational Bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada, 2014: 1-14

[51] Jin D, Heimann M, Rossi R A, et al. Node2bits: Compact Time- and attribute-aware node representations for user stitching//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Würzburg, Germany, 2019: 483-506

[52] Charikar M S. Similarity estimation techniques from rounding algorithms//Proceedings of the 34th Annual ACM Symposium on Theory of Computing. Montreal, Canada, 2002: 380-388

[53] Gupte P V, Ravindran B, Parthasarathy S. Role discovery in graphs using global features: Algorithms, applications and a novel evaluation strategy//Proceedings of 2017 IEEE 33rd International Conference on Data Engineering. San Diego, USA, 2017: 771-782

[54] Gupte P V, Ravindran B. Scalable positional analysis for studying evolution of nodes in networks//Proceedings of the SIAM Data Mining Workshop on Mining Networks and Graphs: A Big Data Analytics Challenge. Cardiff, UK, 2014: 1-13

[55] Hammond D K, Vandergheynst P, Gribonval R. Wavelets on graphs via spectral graph theory. Applied and Computational Harmonic Analysis. 2011, 30(2): 129-150

[56] Guo J, Xu L, Liu J. SPINE: Structural identity preserved inductive network embedding//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 2399-2405

[57] Pei Y, Du X, Zhang J, et al. struc2gauss: Structural role preserving network embedding via Gaussian embedding. Data Mining and Knowledge Discovery, 2020, 34(1): 1072-1103

[58] Jin R, Lee V E, Hong H. Axiomatic ranking of network role similarity//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 922-930

[59] Jin R, Lee V E, Li L. Scalable and axiomatic ranking of network role similarity. ACM Transactions on Knowledge Discovery from Data, 2014, 8(1): 1-37

[60] Yang C, Liu M, Zheng V W, et al. Node, motif and subgraph: Leveraging network functional blocks through structural convolution//Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Barcelona, Spain, 2018: 47-52

[61] Rossi R A, Zhou R, Ahmed N K. Estimation of graphlet counts in massive networks. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(1): 44-57

[62] Rossi R A, Ahmed N K, Koh E, et al. A structural graph representation learning framework//Proceedings of the 13th ACM International Conference on Web Search and Data Mining. Houston, USA, 2020: 483-491

[63] Nikolentzos G, Vazirgiannis M. Learning structural node representations using graph kernels. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(5): 2045-2056

[64] Williams C, Seeger M. Using the Nyström method to speed up kernel machines//Proceedings of the 13th International Conference on Neural Information Processing Systems. Denver, USA, 2000: 661-667

[65] Rossi R A, Zhou R, Ahmed N K. Deep feature learning for graphs. arXiv preprint arXiv:1704.08829, 2017

[66] Ahmed N K, Neville J, Rossi R A, et al. Efficient graphlet counting for large networks//Proceedings of the 2015 IEEE International Conference on Data Mining. Atlantic City, USA, 2015: 1-10

[67] Ahmed N K, Willke T L, Rossi R A. Estimation of local subgraph counts//Proceedings of the 2016 IEEE International Conference on Big Data. Washington, USA, 2016: 586-595

[68] Lee J B, Rossi R A, Kong X N, et al. Higher-order graph convolutional networks. arXiv preprint arXiv: 1809.07697, 2018

[69] Qiu J, Chen Q, Dong Y, et al. GCC: Graph contrastive coding for graph neural network pre-training//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Virtual Event, USA, 2020: 1150-1160

[70] Leskovec J, Faloutsos C. Sampling from large graphs//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 631-636

[71] Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018

[72] Ribeiro L, Saverese P, Figueiredo D R. struc2vec: Learning node representations from structural identity//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 385-394

[73] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 701-710

[74] Keogh E, Ratanamahatana C A. Exact indexing of dynamic time warping. Knowledge and Information Systems, 2005, 7(3): 358-386

[75] Ma X, Qin G, Qiu Z, et al. RiWalk: Fast structural node embedding via role identification//Proceedings of the 2019 IEEE International Conference on Data Mining. Beijing, China, 2019: 478-487

[76] Ahmed N K, Rossi R A, Lee J, et al. Role-based graph embeddings. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(5): 2401-2415

[77] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems, 2018, 151(1): 78-94

[78] Eagle N, Pentland A. Reality mining: Sensing complex social systems. Personal and Ubiquitous Computing, 2006, 10(4): 255-268

[79] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 807-816

[80] Stallings J, Vance E, Yang J, et al. Determining scientific impact using a collaboration index. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(24): 9680-9685

[81] Jin Y, Song G, Shi C. GraLSP: Graph neural networks with local structural patterns//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(4): 4361-4368

[82] Wang D, Cui P, Zhu W. Structural deep network embedding//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1225-1234

[83] Long Q, Jin Y, Song G, et al. Graph structural-topic neural network//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Virtual Event, USA, 2020: 1065-1073

[84] Kipf T N, Welling M. Variational graph auto-encoders//Proceedings of 2016 NIPS Workshop Bayesian Deep Learn. Barcelona, Spain, 2016: 1-3

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

[86] Cai Q, Alam S, Pratama M, et al. Robustness evaluation of multipartite complex networks based on percolation theory. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(10): 6244-6257

[87] Ma L, Zhang X, Li J, et al. Enhancing robustness and resilience of multiplex networks against node-community cascading failures. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(6): 3808-3821

[88] Fionda V, Pirrò G. Community deception or: How to stop fearing community detection algorithms. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(4): 660-673

[89] Chen J, Chen L, Chen Y, et al. GA-based Q-attack on community detection. IEEE Transactions on Computational Social Systems, 2019, 6(3): 491-503

[90] Holland P W, Leinhardt S. An exponential family of probability distributions for directed graphs. Publications of the American Statistical Association, 1981, 76(373): 33-50

[91] White D R, Reitz K P. Graph and semigroup homomorphisms on networks of relations. Social Networks, 1983, 5(2): 193-234

[92] Everett B. Relations, residuals, regular interiors, and relative regular equivalence. Social Networks, 1999, 21(2): 147-165

[93] Nowicki K, Snijders T. Estimation and prediction for stochastic block structures. Publications of the American Statistical Association, 2001, 96(455): 1077-1087

[94] Jiao P, Tian Q, Zhang W, et al. Role discovery-guided network embedding based on autoencoder and attention mechanism. IEEE Transactions on Cybernetics, 2021, 1(1): 1-14

[95] Zhang W, Guo X, Pan T, et al. Role-oriented network embedding based on adversarial learning between higher-order and local features//Proceedings of the 30th ACM International Conference on Information and Knowledge Management. Virtual Event, Australia, 2021: 3632-3636

[96] Rossi R A, Gallagher B, Neville J, et al. Modeling dynamic behavior in large evolving graphs//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome, Italy, 2013: 667-676

[97] Pei Y, Zhang J, Fletcher G H, et al. DyNMF: Role analytics in dynamic social networks//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 3818-3824

[98] Evans E, Guo W, Genctav A, et al. Role detection and prediction in dynamic political networks. Advances in Data Science, 2021, 26(1): 233-252

[99] Velikovi P, Cucurull G, Casanova A, et al. Graph attention networks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-11

[100] Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks?//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019: 1-17

[101] Li P, Wang Y, Wang H, et al. Distance encoding: Design provably more powerful neural networks for graph representation learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 4465-4478

[102] Pei Y, Fletcher G, Pechenizkiy M. Joint role and community detection in networks via $L_{2,1}$ norm regularized nonnegative matrix tri-factorization//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Vancouver, Canada, 2019: 168-175

[103] Shi B, Zhou C, Qiu H, et al. Unifying structural proximity and equivalence for network embedding. IEEE Access, 2019, 7(1): 106124-106138

[104] Chen T, Tang L A, Sun Y, et al. Integrating community and role detection in information networks//Proceedings of the 2016 SIAM International Conference on Data Mining. Miami, USA, 2016: 72-80

[105] Shen X, Chung F L. Deep network embedding for graph representation learning in signed networks. IEEE Transactions on Cybernetics, 2018, 50(4): 1556-1568

[106] Kazemi S M, Goel R, Jain K, et al. Representation learning for dynamic graphs: A survey. Journal of Machine Learning Research, 2020, 21(70): 1-73

[107] Yang C, Xiao Y, Zhang Y, et al. Heterogeneous network representation learning: A unified framework with survey and benchmark. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(10): 4854-4873

[108] Gu S, Johnson J, Faisal F E, et al. From homogeneous to heterogeneous network alignment via colored graphlets. Scientific Reports, 2018, 8(1): 1-16

[109] Rossi R A, Ahmed N K, Carranza A, et al. Heterogeneous graphlets. ACM Transactions on Knowledge Discovery from Data, 2020, 15(1): 1-43

[110] Zhou F, Yue X, Trajcevski G, et al. Context-aware variational trajectory encoding and human mobility inference// Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 3469-3475

[111] Yao Z. Exploiting human mobility patterns for point-of-interest recommendation//Proceedings of the 11th ACM International Conference on Web Search and Data Mining. Marina Del Rey, USA, 2018: 757-758

[112] Gallotti R, Bazzani A, Rambaldi S, et al. A stochastic model of randomly accelerated walkers for human mobility. Nature Communications, 2016, 7(1): 1-7

[113] Hu T, Xia Y, Luo J. To return or to explore: Modelling human mobility and dynamics in cyberspace//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 705-716

[114] Zha Y, Zhou T, Zhou C. Unfolding large-scale online collaborative human dynamics. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(51): 14627-14632

JIAO Peng-Fei, Ph. D., professor. His research interests include complex network analysis and its applications.

PAN Ting, M. S. candidate. Her research interest is role-based network representation learning.

JIN Di, Ph. D., associate professor. His research interest is network representation learning.

WANG Wen-Jun, Ph. D., professor. His research interests include complex network analysis and data mining.

HE Dong-Xiao, Ph. D., associate professor. Her research interest is complex network analysis.

GAO Meng-Zhou, Ph. D., lecturer. Her research interests include complex network analysis and network control.

ZHAO Zhi-Dong, Ph. D., professor. His research interests include artificial intelligence and signal systems.

Background

Network representation learning is the most popular research field in network. The research of network representation learning is closely coupled with the two most important node clustering problems in network: community detection and role discovery. The two branches of network representation learning, in which one is community-oriented and the other is role-oriented, are equally critical. While the community-oriented network representation learning is well studied and comprehensively surveyed by a great number of papers, there is limited attention on the role-oriented branch.

In this paper, we analyze and summarize the current works of role-oriented network representation learning both theoretically and experimentally: we discriminate the notions of role-oriented and community-oriented network embedding; the existing role-oriented methods are reviewed and grouped using a new taxonomy, by which the essence of these methods are clearly demonstrated; a series of experiments are conducted on real-world datasets for comprehensively evaluating the methods; specifically, we propose the idea of analyzing robustness of role-oriented methods, which previously remained in a vacuum; we additionally discuss the challenges of role-oriented network representation learning, which guides its development direction.

This work is funded by the special fund for basic scientific research business expenses of Zhejiang provincial colleges and universities (GK229909299001-008), Zhejiang Provincial Natural Science Foundation of China (LDT23F0101, LDT23F01012F01, LDT23F01015F01), and the National Natural Science Foundation of China (61902278, 62003120). This project focuses on complementing and improving the scope of network representation learning as well as advancing the development of network science and graph neural networks. Our group has been working on role-oriented network representation for several years. We have contributed a number of papers in this field.