基于弱边识别与有向传播机制的社区检测算法

陈 梅†，王 欢，付豪杰，周启辉，黄欣玥

(兰州交通大学，甘肃 兰州 730070)

摘 要: 针对复杂网络中社区边界模糊、结构不均衡以及局部信息缺失等因素对社区检测准确性与鲁棒性带来的挑战, 本文提出一种基于弱边识别与有向传播机制的社区检测算法 (Community Detection algorithm based on Weak edge identification and Directed propagation mechanism, CDWD). 该算法首先识别并剔除基于最少共同邻居准则的弱边, 使潜在社区边界得以显现, 每个连通子图由此形成初始社区结构; 接着, 进一步构建有向影响图, 通过局部相似性强化社区内部的结构联系, 提升信息传递的方向性与一致性; 最后, 依据节点与候选社区之间的拓扑关联强度, 动态判定其最优归属, 确保社区划分的完整性与合理性. 实验结果表明, CDWD 在多个真实网络、合成网络及由聚类数据集构建的图结构上均优于主流基线算法. 同时, 算法参数方便设置, 便于实际应用.

关键词: 复杂网络；社区检测；社区结构；弱边识别；有向传播机制

中图分类号: TP301 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0681

引用格式: 陈梅, 王欢, 付豪杰, 等. 基于弱边识别与有向传播机制的社区检测算法 [J]. 控制与决策

Community detection algorithm based on weak edge identification and directed propagation mechanism

CHEN Mei†，*WANG Huan*，*FU Hao-jie*，*ZHOU Qi-hui*，*HUANG Xin-yue*

(School of Electronic and Information Engineering，Lanzhou Jiaotong University，Lanzhou 730070，China)

Abstract: To address the challenges to community detection accuracy and robustness caused by fuzzy community boundaries, structural imbalance, and incomplete local information in complex networks, this paper proposes a community detection algorithm named CDWD, based on weak edge identification and directed propagation mechanisms. The algorithm first identifies and removes weak edges based on the minimum common neighbor criterion, revealing the boundaries of potential communities. Each connected subgraph thus forms an initial community structure. Next, a directed influence graph is constructed to enhance the internal structural connections within communities by leveraging local similarity, improving the directionality and consistency of information propagation. Finally, based on the strength of the topological association between nodes and candidate communities, dynamically determine their optimal affiliation to ensure the integrity and rationality of community division. Experimental results demonstrate that CDWD outperforms mainstream baseline algorithms on multiple real-world networks, synthetic networks, and graph structures constructed from clustered datasets. At the same time, the algorithm parameters are easy to set, making it convenient for practical applications.

Keywords: complex network； community detection； community structures； weak edge identification； directed propagation mechanism

0 引 言

社区结构广泛存在于各类复杂系统中[1], 反映了节点间紧密关联的功能性子集. 作为揭示网络潜在组织模式的重要手段, 社区检测对于理解网络的功能结构与优化信息传播具有重要意义[2]. 然而, 真实网络往往呈现出结构异质性强、社区边界模糊、尺度分布多样等特点[3], 导致现有方法在检测准确性与结果稳定性之间难以兼顾. 因此, 亟需设计一种更具鲁棒性和高识别精度的社区检测方法, 以更有效地应对复杂网络中的结构挑战.

收稿日期: 2025-06-26；**录用日期:** 2025-10-31.

基金项目: 国家自然科学基金项目 (62266029); 甘肃省重点研发计划项目 (24YFGA036); 甘肃省高等学校产业支撑计划项目 (2022CYZC-36).

†通信作者. E-mail: mei.chen.lzjtu@hotmail.com.

近年来, 社区检测在复杂网络分析中得到了广泛研究, 主要可分为四类: 基于模块度优化[4-7]、基于标签传播[8–17]、基于谱分析[18-20] 以及基于随机游走[21-23] 的方法. 其中, 基于模块度优化的方法[4-7] 通过模块度函数评估网络划分质量, 并以最大化模块度值为目标来识别社区结构. 典型的 Louvain 算法[4] 首先将每个节点初始化为独立社区, 然后迭代地将节点移动至能够最大化模块度增益的邻居社区, 逐步合并直至模块度不再显著提升, 从而完成社区划分. 在此基础上, 张等人[6] 对传统模块度密度方法进行改进, 提出了 Triangle_NMF (Triangular motif information) 算法, 该算法引入三角形子图结构, 同时结合边和三角形信息, 更全面地刻画社区内部的紧密连接性, 从而提升划分精度. 虽然此类方法在效率和可扩展性方面表现优异, 但由于过度依赖全局模块度优化, 常在社区规模不均衡或边界模糊的情况下出现误差, 导致小型或稀疏社区被合并到大型社区中, 社区边界难以精确识别.

基于标签传播的算法[8-17] 依托局部结构信息进行社区检测, 从而避免了对全局优化目标的依赖. 经典的标签传播算法[8] (Label Propagation Algorithm, LPA) 计算高效, 无需预设参数, 但对初始状态敏感, 在社区边界复杂的网络中容易产生误判. 为提升 LPA 算法的稳定性和准确性, 研究者提出了一些改进方法, 如基于中心节点选择与扩展的增强方法[16] (Central node selection and expansion, CNSE) 和多因素局部相似性标签选择与合并算法[17] (Local multi-factor node scoring and label selection-based algorithm, LMFLS). CNSE[16] 通过先识别网络中的核心节点, 再以这些核心节点为起点向外逐层扩散标签, 实现多级标签传播, 从而在一定程度上提升了标签传播的稳定性. LMFLS[17] 则结合多因素节点评分与基于相似性的标签选择与合并机制, 利用局部拓扑特征实现高效的社区检测. 然而, 由于这些方法仍主要依赖局部结构信息, 当网络中存在节点归属复杂或社区边界模糊的情况时, 标签传播容易受到局部噪声影响, 仍会导致部分节点划分错误.

基于谱分析[18-20] 和随机游走[21-23] 的方法更侧重于利用网络的全局结构或信息流特征. 基于谱分析的方法[18-20] 通过对图拉普拉斯矩阵进行特征分解, 将网络映射到低维空间, 再结合传统聚类算法实现社区划分. 这类方法能够较好地捕捉网络的全局结构特征, 因此在结构清晰、规模适中的网络中表现优异. 然而, 谱分析依赖复杂的矩阵分解运算, 计算开销较大, 且对噪声较为敏感, 难以直接应用于大规模或噪声较多的网络. 与此不同, 基于随机游走的方法[21-23] 通过模拟信息在网络中的传播轨迹来揭示潜在的社区结构. 典型的 Infomap 算法[21] 基于最小描述长度原理, 对信息路径进行压缩编码, 在稀疏网络或路径结构清晰的情况下能够取得显著效果. 除此之外, 多网络随机游走模型[23] (Random walk on multiple networks, RWM) 通过引入多个随机游走者, 从起始节点出发获取其相对局部邻近性, 不同游走者之间若在节点上的访问概率相近, 则会相互强化, 从而凸显社区内部的紧密性. 然而, 这类方法对游走策略和路径分布十分敏感. 在网络稠密时, 游走者频繁跨越不同社区, 容易导致社区被过度合并; 而在边界模糊时, 访问概率趋于平均化, 使得社区间的差异被削弱, 进而造成结构特征丢失和划分不准确.

尽管上述方法在社区检测领域已取得显著成果, 但在处理社区边界模糊、局部结构紧密或节点归属判定困难等复杂场景时, 仍存在一定局限性. 为此, 本文提出一种基于弱边识别与有向传播机制的社区检测算法 CDWD, 具体贡献总结如下: (1) 提出了一种融合"全局边界显性化"和"局部信息定向传播"的社区检测新范式. 通过识别并剔除结构脆弱边以显性化社区边界, 该方法降低了传统算法在社区边界区域的划分误差; 同时利用基于局部相似性的有向影响传播机制, 提升了社区内部信息一致性并避免了标签传播混乱. (2) 设计了一种基于弱连通性分析的社区划分机制. 通过在剔除弱边后识别弱连通分量以构建初步社区, 保证核心结构的稳定性; 并依据节点与社区间的连接强度对孤立节点分配社区, 从而兼顾社区划分的完整性与鲁棒性. (3) 构建了完整、可复现的 CDWD 算法流程, 系统融合弱边识别、有向图构建与社区划分三个核心步骤, 为缓解传统方法对全局信息依赖过强的问题提供了有效途径. (4) 在多种真实与合成网络上的实验结果表明, CDWD 算法可显著提升社区划分质量, 验证了算法的有效性. 跨任务实验进一步将聚类数据集转化为图结构, 并与典型聚类方法对比, 结果验证了算法的良好扩展性与跨任务适用性.

1 相关理论与算法实现

1.1 相关理论

本文研究对象为无向无权网络 (主要符号及描述见表 1), 研究范围限定于静态网络和非重叠社区检测问题.

定义 1 (无向无权网络) 一个无向无权网络可表示为$G(V, E)$, 其中V为节点集合, 节点数$n =$

表1　符号定义

符号	符号描述
$u \in V$	节点u
$(u,v) \in E$	连接, 节点u和v之间有边
$d(u)$	节点u的度
$\bar{d}$	网络平均度
$N(u)$	节点u的邻居集合
$cn(u,v)$	相邻节点u和v的共同邻居数
C	网络社区结构
$\|C\|$	社区数量

$|V|$, $E \subseteq \{(u,v)|u,v \in V, u \neq v\}$为边集合, 边数$m = |E|$. 若$(u,v) \in E$, 则必有$(v,u) \in E$, 且$(u,v) = (v,u)$. 网络中的所有边不区分方向, 也不附带权重.

定义 2 (社区检测)　社区检测的目标是将网络$G(V,E)$中的节点划分为若干社区集合$C = \{C_1, C_2, \ldots, C_s\}$. 具体而言, 对于任意社区$C_i \in C$和任意节点$v \in V$, 应满足$v \in C_i$且$C_i \subseteq V$. 并且任取$C_i, C_j \in C$, $C_i \cap C_j = \emptyset$.

1.2　算法实现

本文提出的 CDWD 算法结合弱边识别与有向传播机制, 通过逐步细化与重构社区结构, 实现高质量的社区划分, 整体流程如图 1 所示. 算法首先剔除网络中的弱边以明确社区边界; 随后基于节点局部相似性构建有向影响图, 强化社区内部结构连贯性; 最终结合弱连通性分析, 将每个弱连通分量视为一个社区, 并根据孤立节点与各社区的连接强度, 将其合理分配到对应社区, 从而完成整体的社区划分.

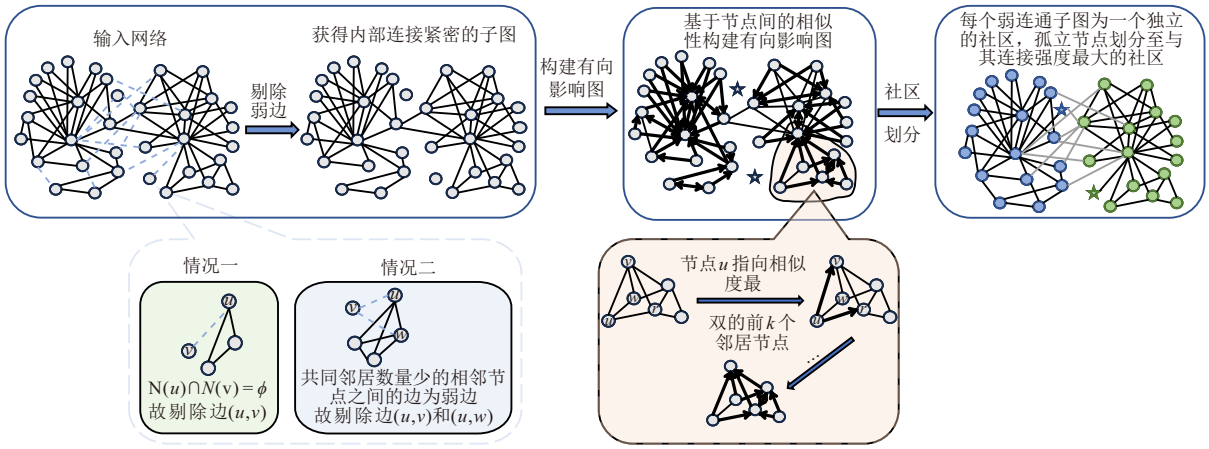

图1　CDWD 算法社区检测过程

1.2.1　弱边剔除

为揭示网络中的潜在社区边界, 本文在算法初始阶段设计了一种基于局部结构和共同邻居数的弱边剔除方法. 节点间共同邻居数作为衡量边强度的重要指标, 反映了节点之间的结构相似性和连接紧密度. 通常, 共同邻居数量越多, 边所处的局部结构越稳定, 表明该边更可能位于同一社区内部. 基于该思想, 算法通过识别并移除相邻节点共同邻居数较少的弱边, 将网络划分为多个内部连接紧密的子图. 具体而言, 本文将"弱边"定义为以下两类:

第一类: 无共同邻居的相邻节点之间的边. 若相邻节点u和v满足公式 (1), 即二者之间不存在任何共同邻居, 则边(u,v)被视为结构断裂性最强的弱边, 应予以剔除.

$$N(u) \cap N(v) = \emptyset \tag{1}$$

第二类: 当网络中不存在第一类边时, 将共同邻居数最少的相邻节点之间的边视为弱边. 若满足公式 (2), 则认为边(u,v)结构支持最弱, 在所有具有共同邻居的边中连接强度最低, 亦应予以剔除.

$$cn(u,v) = \min_{(u,v)\in E, N(u)\cap N(v)\neq\emptyset} |N(u) \cap N(v)|. \tag{2}$$

综上, 弱边剔除作为算法的首要步骤, 基于节点间共同邻居数识别并移除结构最脆弱的连接, 有效刻画了潜在的社区边界. 本文统一将相邻节点间共同邻居数最少的边定义为"弱边", 该策略无需引入额外参数, 具备较强的自适应性和良好的通用性. 记剔除这些弱边后的边集为E', 则原始网络被划分为若干内部结构紧密的子图$G' = (V, E')$, 为后续的有向影响图构建与社区划分提供输入结构.

1.2.2　构建有向影响图

在弱边剔除后, 网络被划分为多个局部紧密的子图. 为刻画节点间的导向作用, 本文基于相似性为每个节点保留指向最相似邻居的有向边, 构建有向影响图. 在该图中, "影响"表示节点对其邻居的结构引导作用: 边的方向指向节点认为最相似的邻居, 从而反映节点在局部结构中的主导关系. 此设计为社区划分提供了方向性依据, 使节点倾向于加入与其

局部相似度较高的社区, 从而增强社区内部的凝聚性与划分合理性. 具体地, 对于每个节点u, 计算其与所有邻居节点之间的相似度. 相似度指标可选用多种局部结构度量方法, 如共同邻居数、Jaccard 系数和 Adamic-Adar 等[24,25]. 鉴于算法在第一阶段已计算过相邻节点间的共同邻居数, 本文默认采用共同邻居数作为相似度的度量标准, 定义为公式 (3):

$$S(u,v)=|N(u)\cap N(v)|. \tag{3}$$

在此基础上, 为每个节点u选择与其相似度最高的前k个邻居节点, 并构建从u指向这些节点的有向边. 有向边集合可定义为公式 (4):

$$E^{\rightarrow}=\bigcup_{u\in V}\{(u\rightarrow v)\mid v\in N_k(u)\}. \tag{4}$$

其中, $N_k(u)$表示节点u相似度排名前k的邻居集合. 由此可构建得到一个包含所有节点的有向影响图$G^{\rightarrow}=(V,E^{\rightarrow})$, 其边的方向由局部结构相似性决定, 反映了节点间受相似度驱动的引导或影响关系.

该步骤基于节点间的相似度在网络中引入有向边, 以构建节点间的引导关系, 从而使原始网络具备方向性特征. 通过保留每个节点指向其相似度最高的少数邻居, 形成关键连接并减少冗余边的干扰. 由此得到的有向图在结构上突出社区内部的连接关系, 为后续社区检测提供参考.

1.2.3 社区划分与孤立节点归属

在构建完成有向影响图$G^{\rightarrow}=(V,E^{\rightarrow})$后, 算法通过识别其中的弱连通子图 (即忽略边的方向性后仍保持连通的子图), 将每个弱连通子图视为一个独立的社区. 对于未能形成任何连通结构的节点 (即在有向图中入度与出度均为零的节点), 算法基于其在原始网络中的连接关系, 判断其与已识别社区之间的归属. 具体地, 若u为某一孤立节点, C_j为某一候选社区, 定义节点u与社区C_j之间的连接强度如下:

$$Score(u,C_j)=\frac{|N(u)\cap C_j|}{|C_j|}+\frac{\sum_{v\in N(u)\cap C_j}d(v)}{|C_j|}. \tag{5}$$

其中, $N(u)$为节点u在原始网络中的邻居集合, $d(v)$为节点v在原始网络中的度. 上述评分函数综合考虑节点与社区之间的连接密度以及社区邻居的平均结构强度, 以此确定其最优归属. 最终, 将孤立节点逐步分配至得分最高的社区, 完成整个社区划分.

该步骤在社区结构通过弱连通子图自动识别的基础上, 为孤立节点设计了基于连接强度的归属机制. 通过在原始网络中计算节点与各候选社区之间的连接强度, 使节点划分依赖实际邻接关系而非随机划分, 从而提升划分结果的准确性与稳定性.

1.2.4 CDWD 算法伪代码和时间复杂度分析

算法 1 展示了 CDWD 算法的伪代码, 包括三个主要步骤, 本文将依次分析该算法的时间复杂度. 假设发现的社区数量为p, 网络节点数为n, 边数为m, 未分配社区节点数为n'. 在 Step 1 中, 算法需要遍历所有边, 时间复杂度为$O(m)$; 在 Step 2 中, 构建有向影响图涉及遍历所有节点及其邻居, 并计算相似度, 时间复杂度为$O(n\cdot\bar{d})$; 在 Step 3 中, 社区划分需识别弱连通子图, 每个节点和边仅访问一次, 时间复杂度为$O(n+m)$, 分配孤立节点的时间复杂度为$O(n')$. 综上所述, CDWD 算法的总时间复杂度为$O(n+2m+n\cdot\bar{d}+p\cdot n')$. 一般情况下, p、n'和$\bar{d}$都远小于n和m, 所以 CDWD 算法总的时间复杂度接近于$O(n+m)$.

算法 1 CDWD 社区检测算法

输入: 网络$G=(V,E)$, 参数k;
输出: 社区检测结果C;
Step 1: 弱边剔除
$E^{\text{Eliminate}}\leftarrow\varnothing$;
for each 每条相邻边$(v_i,v_j)\in E$ do
  根据公式(1)和公式(2)判断是否要剔除该边;
  将需要剔除的连接(v_i,v_j)放入$E^{\text{Eliminate}}$;
end for
$E'=E\setminus E^{\text{Eliminate}}$, 构造子图$G'=(V,E')$;
Step 2: 构建有向影响图
$E^{\rightarrow}\leftarrow\varnothing$;
for each $v_i\in V$ do
  for each $v_j\in N(v_i)$ do
    根据公式(3)计算v_i与v_j的相似度;
    根据公式(4)构建有向边集$E^{\rightarrow}$;
得到有向图$G^{\rightarrow}=(V,E^{\rightarrow})$;
Step 3: 社区划分与孤立节点归属
$C\leftarrow\varnothing$;
在$G^{\rightarrow}$中识别弱连通子图集合$C=\{C_1,C_2,\ldots,C_p\}$;
for each 未分配社区节点$u\in V\setminus\cup_{i=1}^{p}C_i$ do
  根据公式(5)计算其与各社区的得分;
  $C_k\leftarrow C_k\cup\{u\}$;
  $C\leftarrow C\cup C_k$;
return C

2 实验分析

2.1 对比算法和数据集介绍

为全面评估 CDWD 算法在社区检测任务中的有效性, 本文选取多种代表性的社区检测算法及三

种经典聚类算法作为对比, 所用数据集涵盖不同规模、密度及社区结构类型. 表 2 展示了真实网络的详细信息, 表 3 则展示了聚类数据集及其构建图数据时的参数配置. 表 2 中真实网络来自 Newman 教授的网站和斯坦福大学网站, 表 3 中数据集来自东芬兰大学的网站. 由于社区检测算法无法直接应用于聚类数据集, 本文采用ε-ball 方法[13,26] 对 5 个不同类型的数据集构建图结构, 使其适用于 CDWD 算法的社区划分, 以便与聚类方法进行性能对比. 社区检测算法包括基于模块度优化的 Louvain[4]、Leiden [5] 和 TE-MA[7], 基于标签传播的 LPA [8]、GCN [9]、LSMD [12]、LS [13]、RaidB [14]、LBLD [15] 和 LMFLS [17], 以及基于随机游走的 Infomap [21]. 此外, 还包括其他的社区检测算法 CDME [27]. 辅助对比的聚类方法包括基于划分的k-Means[28], 以及基于密度的 DPC[29] 和 R-MDPC[30]. 所有算法均在 Windows 11 操作系统、AMD Ryzen 7 6800H 3. 20 GHz CPU 和 16GB RAM 的计算环境下运行, 其获取方式已上传至 GitHub (地址链接: https://github.com/AQYH/CDWD.git).

表2　真实网络描述

类别	网络	$\|V\|$	$\|E\|$	$\|C\|$
有真实社区结构	Dolphin	62	159	2
	Football	115	613	12
	Karate	34	78	2
	Polbooks	105	441	3
	Riskmap	118	197	6
	Amazon	334 863	925 872	75 149
	YouTube	1 134 890	2 987 624	8 385
无真实社区结构	Power	4 941	6 594	–
	CA-GrQc	5 242	14 490	–
	CA-HepTh	9 877	25 985	–
	Netscience	1 589	2 742	–
	Email	1 133	5 451	–
	Facebook	4 039	88 234	–
	PGP	10 680	24 316	–
	Brightkite	58 228	214 078	–

表3　聚类数据集描述

数据集	样本数	维度	簇数	ε
Flame	240	2	2	0.925 ± 0.05
Aggregation	788	2	7	1.5 ± 0.04
R15	600	2	15	0.66 ± 0.02
Spiral	312	2	3	2.0 ± 0.1
Iris	150	4	4	2.0 ± 0.5

ε: 在构建图数据时, 若两点间的欧式距离不超过该阈值, 则在它们之间建立一条边.

2.2　评价指标

对于社区划分结果, 记真实的社区结构为$C = \{C_1, C_2, \ldots, C_s\}$, 其中$C_s$表示第$s$个真实社区; 算法检测到的社区划分为$P = \{P_1, P_2, \ldots, P_p\}$, 其中$P_p$表示第$p$个检测到的社区. 本文将采用以下四个指标对各算法结果进行评估.

2.2.1　模块度

模块度记作Q, 由 Newman 等人[31] 提出, 是一种用来衡量社区结构质量优劣的指标. 具体地, 模块度Q被定义为公式 (6):

$$Q = \frac{1}{2m} \cdot \sum (A_{ij} - \frac{k_i k_j}{2m}) \cdot \delta(c_i, c_j). \tag{6}$$

其中, A_{ij}表示网络中结点i和结点j之间的连接权重; k_i和k_j分别是结点i和结点j的度; m是网络的总边数. $\delta(c_i, c_j)$是一个指示函数, 当结点i和结点j属于同一社区时为 1, 否则为 0.

2.2.2　标准化的互信息量

标准化互信息 (Normalized Mutual Information, NMI)[32] 是一种常用的聚类评价指标, 用于衡量两个聚类结果之间的一致性. NMI 的取值范围为 [0, 1], 值越大表示两个聚类结果越接近, 聚类质量越高. 在社区检测任务中, 它被定义为公式 (7):

$$NMI(C, P) = \frac{-2\sum_{i=1}^{s}\sum_{j=1}^{p} |C_i \cap P_j| \log(\frac{|C_i \cap P_j| \cdot n}{|C_i| \cdot |P_j|})}{\sum_{j=1}^{p} |P_j| \log(\frac{|P_j|}{n}) + \sum_{i=1}^{s} |C_i| \log(\frac{|C_i|}{n})}. \tag{7}$$

2.2.3　F-score

F-score[33] 综合了检测结果在精确率 (公式 (8)) 与召回率 (公式 (9)) 两方面的表现, 定义为公式 (10):

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

$$Recall = \frac{TP}{TP + FN}, \tag{9}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{10}$$

其中, TP表示正确识别为同一社区的节点对数, FP表示错误识别为同一社区的节点对数, FN表示未能识别为同一社区的节点对数.

2.2.4　调整兰德指数

调整兰德指数 (Adjusted Rand Index, ARI)[34] 可以衡量检测结果与真实划分的一致性, 并对随机划分进行修正, 定义为公式 (11):

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}. \tag{11}$$

其中, $n_{ij}=|S_i\cap S_j|$, $a_i=|S_i|$, $b_i=|P_j|$. ARI 的取值范围为 [−1, 1], 值越接近 1 表示划分结果越好.

2.3 真实网络实验及结果

为评估 CDWD 算法在真实网络上的有效性, 本文选取多个对比算法进行对比实验. 对于已知社区结构的网络, 使用 NMI、F-score、ARI 和模块度Q评估算法性能; 对于未知社区结构的网络, 仅使用模块度Q. 对于 Amazon 和 YouTube 网络, 选取前 5000 个高质量社区计算 NMI、F-score 和 ARI. 表 4、表 5、表 6 和表 7 展示了各算法的评估结果, 最优值以加粗标注.

从表 4、表 5 和表 6 的结果可以看出, CDWD 算法在大多数网络上取得了最高的 NMI、F-score 和 ARI, 表现出较强的社区检测能力. 在 Dolphin 和 Karate 网络中, CDWD、LBLD、LSMD 和 CDME 算法均实现了完美识别, NMI、F-score 和 ARI 均为 1.00. 这类网络规模较小、社区边界清晰, 便于利用结构信息的算法准确识别; 相比之下, Louvain、Infomap、LPA 和 LS 等算法由于存在个别节点划分错误, 导致整体精度下降, 且输出的社区数与真实社区数不一致, 反映出其在边界处理或标签传播过程中的不稳定性. 在 Football 网络中, CDWD 依然取得最优结果 (NMI = 0.94, F-score = 0.91, ARI = 0.94). 该网络拓扑较为均匀且平均度较大, 因而基线算法也能取得相对较好的结果. 对于 Polbooks 网络, 虽然该网络社区之间边界模糊, 检测难度较大, 但 CDWD 依然保持领先, NMI、F-score 和 ARI 分别达到 0.68、0.82 和 0.69. 其优势得益于 CDWD 通过识别并剔除弱边, 有效显性化了社区边界, 并利用节点间的局部相似性构建有向传播网络, 增强信息在社区内的定向传递. 相比之下, 基于随机游走的 Infomap 在此类结构中的表现较弱 (NMI = 0.50, F-score = 0.45, ARI = 0.59), 说明其对边界模糊区域存在显著误判. LSMD 对相似度阈值较为敏感, LBLD 的社区合并策略存在局部最优问题, CDME 在边界识别上策略粗糙, 这些缺陷共同导致它们在该网络上的检

表4 7 个已知社区结构网络上的 NMI 比较

类型	模块度			随机游走	标签传播							其他	CDWD
网络	Louvain	Leiden	TE-MA	Infomap	GCN	LSMD	LBLD	RaidB	LS	LPA	LMFLS	CDME	
Dolphin	0.58	0.59	0.63	0.58	0.54	**1.00**	**1.00**	0.54	0.74	0.53	**1.00**	0.70	**1.00**
Karate	0.71	0.69	0.83	0.70	0.83	**1.00**	**1.00**	0.80	0.73	0.64	**1.00**	**1.00**	**1.00**
Football	0.85	0.85	0.86	0.92	0.87	0.93	0.91	0.64	0.46	0.90	0.90	0.93	**0.94**
Polbooks	0.53	0.57	0.54	0.50	0.53	0.59	0.60	0.60	0.57	0.50	0.60	0.58	**0.68**
Riskmap	0.84	0.86	0.84	0.89	0.75	0.76	0.81	0.73	0.67	0.84	0.85	0.72	**0.90**
Amazon	0.83	0.86	—	0.42	0.90	0.95	**0.97**	0.95	0.95	0.83	**0.97**	**0.97**	**0.97**
YouTube	0.49	0.46	—	0.50	0.23	0.19	0.57	—	—	—	0.50	—	**0.72**

表5 7 个已知社区结构网络上的 F-score 比较

类型	模块度			随机游走	标签传播							其他	CDWD
网络	Louvain	Leiden	TE-MA	Infomap	GCN	LSMD	LBLD	RaidB	LS	LPA	LMFLS	CDME	
Dolphin	0.44	0.55	0.39	0.54	0.44	**1.00**	**1.00**	0.78	0.88	0.56	**1.00**	0.73	**1.00**
Karate	0.84	0.82	0.63	0.87	0.95	**1.00**	**1.00**	0.89	0.88	0.42	**1.00**	**1.00**	**1.00**
Football	0.87	0.87	0.48	0.85	0.76	0.91	**0.92**	0.27	0.34	0.86	0.91	**0.92**	0.91
Polbooks	0.54	0.61	0.77	0.45	0.38	0.70	0.68	0.56	0.80	0.48	0.68	0.45	**0.82**
Riskmap	0.62	0.76	0.62	0.69	0.70	0.69	0.74	0.62	0.46	0.62	0.64	0.54	**0.79**
Amazon	0.20	0.28	—	0.02	0.80	0.84	0.90	0.84	0.69	0.76	**0.94**	0.80	0.88
YouTube	0.10	0.11	—	0.15	0.11	0.13	0.61	—	—	—	0.60	—	**0.63**

表6 7 个已知社区结构网络上的 ARI 比较

类型	模块度			随机游走	标签传播							其他	CDWD
网络	Louvain	Leiden	TE-MA	Infomap	GCN	LSMD	LBLD	RaidB	LS	LPA	LMFLS	CDME	
Dolphin	0.51	0.65	0.42	0.60	0.83	**1.00**	**1.00**	0.53	0.69	0.38	**1.00**	0.41	**1.00**
Karate	0.51	0.65	0.80	0.60	0.83	**1.00**	**1.00**	0.88	0.77	0.70	**1.00**	**1.00**	**1.00**
Football	0.74	0.80	0.67	0.76	0.87	0.88	0.56	0.49	0.31	0.82	0.82	0.89	**0.92**
Polbooks	0.65	0.52	0.64	0.59	0.53	0.64	0.64	0.58	0.64	0.63	0.57	0.66	**0.69**
Riskmap	0.65	0.57	0.65	0.73	0.66	0.70	0.72	0.54	0.43	0.65	0.58	0.54	**0.74**
Amazon	0.12	0.25	—	0.15	0.35	0.60	0.61	0.59	0.62	0.48	0.58	0.62	**0.65**
YouTube	0.15	0.10	—	0.41	0.03	0.47	0.52	—	—	—	0.42	—	**0.55**

表7　15 个真实网络上的模块度Q和社区数量比较

类型	模块度						随机游走		标签传播												其他		CDWD	
网络	Louvain		Leiden		TE-MA		Infomap		GCN		LSMD		LBLD		RaidB		LPA		LS		LMFLS		CDME			
	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q	$\|C\|$	Q
Dolphin	4	0.41	5	**0.52**	4	**0.52**	5	**0.52**	5	0.51	2	0.37	2	0.37	6	-0.01	3	0.41	3	0.38	2	0.37	5	0.34	2	0.37
Karate	4	**0.42**	4	**0.42**	3	0.39	4	0.40	2	0.39	2	0.37	2	0.31	3	0.02	3	0.31	2	0.31	2	0.37	2	0.37	2	0.37
Football	9	**0.60**	9	**0.60**	7	0.59	11	0.57	12	0.57	12	0.58	13	0.55	4	0.03	11	0.54	6	0.31	13	0.56	12	0.57	14	0.55
Polbooks	5	**0.53**	4	**0.53**	3	0.52	5	0.52	6	0.51	3	0.45	2	0.46	4	-0.004	3	0.45	2	0.44	2	0.46	3	0.49	4	0.44
Riskmap	5	0.61	7	**0.64**	5	0.62	7	0.60	5	0.56	6	0.58	6	0.62	4	0.54	5	0.62	4	0.56	5	0.57	5	0.50	6	0.62
Amazon	232	**0.93**	382	**0.93**	—	—	13	0.78	29 850	0.76	34 304	0.68	15 501	0.80	1 683	0.97	37 428	0.64	11 761	0.77	22 600	0.82	18 809	0.65	25 300	0.85
YouTube	7 365	0.72	4 039	**0.73**	—	—	945	0.69	40 255	0.65	9 636	0.42	25 169	0.45	—	—	—	—	—	—	18 838	0.34	—	—	20 309	0.45
Power	41	0.84	42	**0.94**	821	0.71	6	0.78	678	0.64	446	0.79	341	0.82	1 731	0.04	1 406	0.62	410	0.79	438	0.82	489	0.78	51	0.85
CA-GrQc	392	**0.86**	393	**0.86**	808	0.70	377	0.84	781	0.72	456	0.77	561	0.79	1 033	0.003	990	0.70	245	0.71	593	0.80	640	0.79	135	0.75
CA-HepTh	475	0.67	478	**0.77**	1 356	0.53	458	0.64	1 344	0.60	586	0.63	761	0.69	1 728	0.003	1 729	0.67	268	0.62	928	0.70	1 003	0.67	230	0.66
Netscience	277	0.94	278	**0.96**	322	0.92	268	0.93	317	0.88	275	0.94	303	0.94	366	0.25	343	0.91	176	0.94	309	0.94	1 461	0.94	166	0.94
Email	15	0.54	10	**0.57**	14	0.55	60	0.52	624	0.36	142	0.42	58	0.45	60	0.42	77	0.33	45	0.44	55	0.43	312	0.40	11	0.41
Facebook	12	0.83	17	**0.84**	14	0.79	1 530	0.71	152	0.60	10	0.68	9	0.71	17	0.13	21	0.66	11	0.69	10	0.69	8	0.67	9	0.73
PGP	96	**0.88**	95	**0.88**	108	0.84	3	0.13	911	0.79	643	0.59	358	0.82	260	0.56	1997	0.007	2 059	0.74	559	0.80	591	0.60	612	0.78
Brightkite	742	0.68	678	**0.69**	684	**0.69**	566	0.36	2 342	0.62	2 384	0.29	1 251	0.61	—	—	5 644	0.57	364	0.25	1 302	0.59	1 654	0.61	2 875	0.58

测精度不高. 在大网络 YouTube 上, CDWD 达到最优结果 (NMI = 0.72, F-score = 0.63, ARI = 0.55), 表明其在处理大规模稀疏网络时弱边剔除与有向传播机制依然有效.

从表 7 可以看出, 基于模块度优化的 Leiden 算法在各网络上均取得最高的模块度. 值得强调的是, 社区检测的核心目标是还原网络中的真实社区结构, 而非单纯最大化模块度[35]. 模块度存在"分辨率限制"问题, 即在算法处理大小不均或稀疏结构的网络时, 可能出现误判社区边界、过度合并或划分社区的情况, 从而导致检测精度下降. 以 Dolphin 网络为例, 其真实模块度为 0.37、对应社区数量为 2, 而 Leiden 和 TE-MA 算法的结果均为 0.52、社区数量为 5 或 4 个, 与真实情况存在较大偏差. 结合表 4、表 5 和表 6 可见, Leiden 在 Dolphin、Karate、Football、Polbooks 和 Riskmap 已知社区结构的网络上, 其 NMI、F-score 和 ARI 明显低于 CDWD, 且其得到的模块度与真实模块度之间存在较大偏差, 说明其所识别的社区结构与实际情况不一致. 在 Amazon 网络上, 该问题更加突出. Louvain 和 Leiden 分别检测出 232 和 382 个社区, 远低于真实社区数量 75 149, 说明这两种算法在大规模稀疏网络中将大量社区进行了合并. 相比之下, CDWD 检测出的社区数量为 25 300, 虽未完全还原真实划分, 但明显更接近真实结构, 且具备更合理的规模分布. 这进一步说明, CDWD 在保持较高检测精度的同时, 其模块度结果具有更强的解释性和结构一致性.

2.4　合成网络实验及结果

为验证 CDWD 算法在合成网络上的有效性, 本节基于 LFR 模型[36]生成了一组合成网络, 记为 $LFR_{\mu\in[0.1,0.8]}$, 用于评估算法在社区边界逐渐模糊条件下的性能表现. LFR 模型的参数及其描述见表 7. 实验中除混淆参数μ外, 其他参数配置为$n = 10\,000$, $k = 15$, $MaxK = 50$, $MinC = 20$, $MaxC = 100$, $\gamma = 2$, and $\beta = 1$. μ从 0.1 开始, 以 0.1 为步长递增至 0.8. 随着μ的增大, 跨社区连接逐渐增多, 社区边界变得模糊, 社区结构的可识别性随之降低, 从而增加了检测难度.

图 2 展示了各算法在$LFR_{\mu\in[0.1,0.8]}$网络上的 NMI 表现. 整体来看, CDWD 算法在不同混淆参数μ下均表现出最优性能. 随着μ的增加, 社区间连边增多、边界逐渐模糊, 各算法的 NMI 普遍呈现下降趋势, 但 CDWD 下降最为缓慢, 展现出良好的鲁棒性. 当$\mu \leqslant 0.5$时, 各算法均能较好地识别社区结构; 而在$\mu \geqslant 0.6$时, Louvain 和 Leiden 算法的性能显著下降. 这是由于它们高度依赖模块度的全局优化目标, 在社区边界模糊时易出现误判和过度合并. LPA 算法在$\mu = 0.7$ 和$\mu = 0.8$ 时几乎失效, 主要原因在于其缺乏全局控制机制, 容易陷入局部最优, 导致标签泛滥传播和边界节点划分混乱. Infomap 算法依赖随机游走路径识别社区, 在网络稀疏、边界清晰时效果良好, 但在高密度或高混合度场景中, 路径信息的判别力下降, 难以准确刻画社区结构. LBLD 与 LSMD 虽引入核心节点扩散机制以增强传播效果, 但其性能在高混合度下仍受到初始条件敏感性的限制, 难以稳定识别模糊边界. 相比之下, CDWD 通过剔除弱边以增强局部结构清晰度, 并引入基于相似性的有

向影响图, 有效克服了传统标签传播类算法在边界判定和全局一致性上的局限性, 始终保持较高的准确性, 验证了其在处理边界模糊社区时的优势.

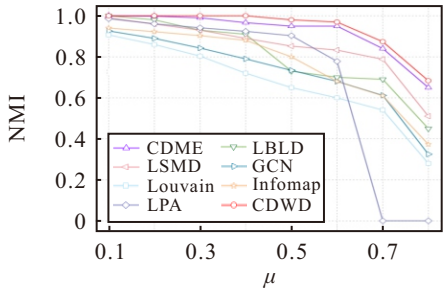

图2 各算法在$LFR_{\mu\in[0.1,0.8]}$网络上的 NMI 比较

2.5 聚类数据集实验及结果

为验证 CDWD 算法在非图结构数据上的适用性, 本节选取了 5 个不同类型的聚类数据集, 并采用 ε-ball 方法对原始数据构建图结构, 从而实现与聚类算法的公平对比. 所用聚类数据集及其构建图数据时的参数详见表 3. 对比算法包括 3 种经典聚类算法: k-Means, DPC 和 R-MDPC. 所有算法的划分结果均通过 NMI 进行评估, 具体结果列于表 9.

表8 LFR 合成网络参数及其描述

参数	参数描述
n	网络节点数
k	合成网络的平均度
$MaxK$	节点的最大度
$MinC$	网络社区中最小节点数
$MaxC$	网络社区中最大节点数
γ	节点度分布指数
β	社区规模分布指数
μ	混淆参数

表9 聚类数据集上的 NMI 比较

类型	基于密度		基于划分	CDWD
数据集	DPC	R-MDPC	k-Means	
Flame	0.97	0.80	0.39	**1.00**
Aggregation	**1.00**	0.69	0.88	**1.00**
Spiral	**1.00**	0.88	0.0007	**1.00**
R15	0.99	0.99	0.99	**1.00**
Iris	0.81	**0.86**	0.76	0.73

从表 9 可以看出, CDWD 在 Flame、Spiral、R15 和 Aggregation 四个数据集上均取得最优结果 (NMI = 1.00), 展现出卓越的簇结构识别能力. 特别是在 Spiral 这类高度非线性、包含螺旋簇的数据集中, k-Means 几乎完全失效 (NMI = 0.0007), 而 CDWD 凭借图建模与弱边剔除机制, 准确刻画了样本间的结构关系, 实现了完美划分. 在低维小样本的 Iris 数据集中, 尽管 CDWD 并未取得最高 NMI, 但整体表现依然稳定, 体现出良好的通用性与适应性. 相比之下, DPC 和 R-MDPC 对参数较为敏感, 在不同数据集上的表现波动较大; k-Means 则受限于对簇形状和数量的假设, 适应能力不足. 综上, CDWD 通过图结构转换与有向传播机制, 有效挖掘了数据的内在结构关系, 不仅在复杂形状和多尺度聚类中展现出明显优势, 也验证了其作为一种通用社区检测方法在聚类任务中的可迁移性和鲁棒性, 体现出优于传统聚类算法的综合性能.

2.6 消融实验

为验证 CDWD 算法中各核心模块的有效性, 本节在 6 个典型真实网络上进行消融实验, 其中 Karate、Polbooks 和 Dolphin 具有真实社区结构, Email、CA-GrQc 和 Netscience 无真实社区结构. 本实验设计三种方法进行对比: 完整方法 CDWD-Full、移除弱边识别模块的 CDWD-w/o Weak, 以及移除有向传播机制的 CDWD-w/o Directed.

表 10 展示了各方法在 6 个真实网络上的 NMI 和Q对比结果. 可以看出, CDWD-Full 始终取得最优性能; 与 CDWD-w/o Directed 的对比表明, 在缺少有向传播机制时, 算法性能显著下降. 原因在于未引入有向边时, 社区构建过程中对所有邻居节点一视同仁, 强相关邻居与弱相关邻居的作用难以区分, 从而削弱了社区内部的凝聚性并降低划分精度. 与 CDWD-w/o Weak 的对比显示, 一旦去除弱边识别模块, 算法在具有真实社区结构的网络上 NMI 几乎降至 0, 社区划分基本失效, 即所有节点被划入同一社区. 这是因为弱边往往跨越社区边界, 若未能识别并剔除这些边, 社区之间的隔离性就会消失, 算法将倾向于将整个网络视为一个整体.

表10 各方法在 6 个真实网络上的 NMI 和Q结果

指标	网络	CDWD-Full	CDWD-w/o Weak	CDWD-w/o Directed
NMI	Karate	**1.00**	0.00	0.33
	Polbooks	**0.68**	0.00	0.35
	Dolphin	**1.00**	0.00	0.26
	Netscience	**0.94**	0.93	-0.001
Q	Email	**0.41**	0.35	-0.002
	CA-GrQc	**0.75**	0.67	-0.0002

为了进一步检验网络中弱边占比对算法性能的影响, 本实验统计了各网络中弱边所占的比例, 结果如表 11 所示. 弱边占比在不同网络间差异显著 (Netscience 为 8%, CA-GrQc 高达 62%), 但无论比例高低, 移除该模块都会导致性能灾难性退化. 这表明即使仅占比 8% 的少量弱边, 其对社区划分也具有决定性作用. 同时, 在弱边占比较高的网络中 (如

CA-GrQc), CDWD-Full 依然保持优异性能, 说明该机制能够精准识别并剔除真正跨社区的冗余边, 而不会破坏网络的主体结构.

表11　6 个真实网络中的弱边统计

真实网络	Karate	Dolphin	Polbooks	Email	CA-GrQc	Netscience
弱边占比(100%)	14	47	52	41	62	8

综上, 弱边剔除与有向传播机制在 CDWD 中相辅相成: 前者确保社区边界的准确刻画, 后者提升节点信息传播与社区凝聚性. 二者结合显著增强了算法的稳健性与有效性.

2.7　CDWD 参数分析

为探究参数 k ($k \in \mathbb{Z}^+$) 的取值对 CDWD 算法性能的影响, 本节在 5 个具有真实社区结构的网络 (Karate、Dolphin、Football、Polbooks 和 Amazon) 和 5 个无真实社区结构的网络 (CA-GrQc、CA-HepTh、Netscience、Email 和 Facebook) 上进行实验分析.

图 3 展示了不同参数 k 设置下, CDWD 在各网络上的 NMI 结果. 整体来看, CDWD 在 $k \in [1,10]$ 范围内的某个子区间或取值下有着较高的 NMI. 在 Karate 和 Football 网络中, 算法在 $k=2$ 或 $k=3$ 时达到最优性能, 表明即便引导信息有限, CDWD 也能准确识别社区结构. 在 Polbooks 网络中, NMI 自 $k=3$ 起趋于稳定, 说明算法能够较好应对社区边界模糊所带来的挑战. 对于规模更大的 Amazon 网络, NMI 随 k 的增大逐步提升, 并在 $k \geq 5$ 时趋于稳定, 反映出算法在处理大规模网络时具备良好的扩展性. 值得注意的是, 在如 Karate 和 Football 等规模较小的网络中, 过大的 k 值可能导致引导信息冗余, 从而影响划分精度. 图 4 展示了 CDWD 在 5 个无真实社区结构网络上的模块度 Q 表现. 可以看出, CA-HepTh 和 Netscience 网络上算法在 $k=4$ 时 NMI 达到峰值, Email 在 $k=2$ 时表现最优, 而 Facebook 在 $k \in [4,5]$ 时表现最优. 尽管各网络的最优 k 值有所不同, 但整体分布高度集中于区间[2,5].

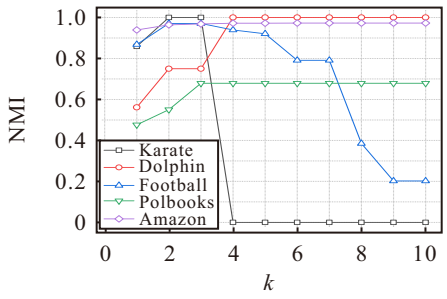

图3　CDWD 算法在 5 个具有已知社区结构网络上 $k \in [1,10]$ 时的 NMI 表现

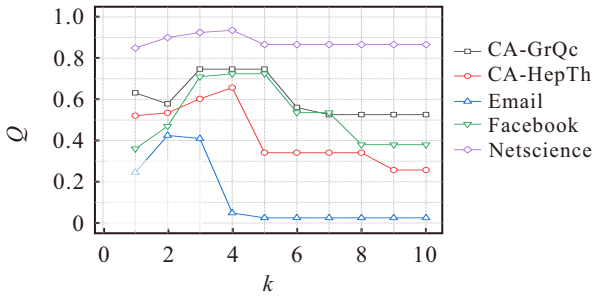

图4　CDWD 算法在 5 个未知社区结构网络上 $k \in [1,10]$ 时的 Q 表现

为进一步评估参数 k 的合理设置, 本实验对文中所使用的 15 个真实网络和 8 个合成网络进行了统计分析 (统计结果见表 12). 分析表明, 最优参数 k 的分布呈现出明显规律: 其下界与上界的中位数分别为 4 和 5, 且 78.3% 的网络其最优 k 范围与区间[2,7]存在重叠. 考虑到参数设置的效率与可靠性, 本文建议将 CDWD 算法应用于未知网络时, 优先将 k 的搜索范围设定为[3,6]. 该区间既是统计上高覆盖率区间[2,7]的紧凑核心, 也与中位数所体现的集中趋势高度一致. 对于社区结构清晰或规模较小的网络, 可将下界适当放宽至 2; 对于结构复杂或连边密集的网络, 则可考虑将上界扩展至 8 或 10 进行验证. 该统计结论与前述图中展示的个案趋势相互支持, 验证了 CDWD 在合理参数区间内的适用性.

表12　23 个网络上 CDWD 算法最优参数 k 的取值

网络	最优参数 k 范围	网络	最优参数 k 范围
Dolphin	[4,10]	Football	[2,3]
Karate	[2,3]	Polbooks	[3,10]
Riskmap	2	Amazon	[4,10]
YouTube	[3,8]	Power	[4,5]
CA-GrQc	[3,5]	CA-HepTh	4
Netscience	4	Email	2
Facebook	[4,5]	PGP	[3,6]
Brightkite	[4,8]	$LFR_{\mu=0.1}$	[2,6]
$LFR_{\mu=0.2}$	[2,6]	$LFR_{\mu=0.3}$	[3,5]
$LFR_{\mu=0.4}$	[2,6]	$LFR_{\mu=0.5}$	[2,8]
$LFR_{\mu=0.6}$	[4,8]	$LFR_{\mu=0.7}$	[4,7]
$LFR_{\mu=0.8}$	[5,10]		

2.8　无参统计分析

为验证 CDWD 算法性能优势的统计显著性, 本实验采用非参数的 Wilcoxon 符号秩检验对实验结果进行分析. 该检验不依赖于数据的正态分布假设, 适用于小样本情况下的算法性能比较. 检验基于 7 个已知社区结构的基准网络, 在 NMI、F-score 和 ARI 上分别将 CDWD 与每个基线方法进行配对比较. 对于部分基线方法在某些网络上无法获得结果

的情况, 本文遵循配对检验原则, 在相应比较中排除该网络. 显著性水平设定为$\alpha = 0.05$. Wilcoxon 符号秩检验结果汇总于表 13. 根据p值判断 CDWD 与基线方法之间的统计显著性: 若某比较的p值小于 0.05, 则认为 CDWD 在该指标上与该基线方法存在显著差异; 否则认为无显著差异.

表13 CDWD 与各基线方法的 Wilcoxon 符号秩检验结果

基线算法	NMI p值	F-score p值	ARI p值	有效比较的网络个数	统计结论
Louvain	0.018	0.018	0.018	7	显著优于
Leiden	0.018	0.018	0.028	7	显著优于
TE-MA	0.063	0.063	0.063	5	优势明显
Infomap	0.028	0.018	0.068	7	显著优于
GCN	0.018	0.018	0.018	7	显著优于
LSMD	0.068	0.068	0.068	7	竞争力相当
LBLD	0.128	0.345	0.345	7	竞争力相当
RaidB	0.043	0.600	0.116	6	显著优于
LS	0.018	0.028	0.068	6	显著优于
LPA	0.018	0.068	0.116	6	显著优于
LMFLS	0.068	0.345	0.600	7	竞争力相当
CDME	0.068	0.600	0.116	6	竞争力相当

注: 显著性水平$\alpha = 0.05$; "显著优于"表示至少一个指标上p值小于0.05.

根据表 13 可以看出, CDWD 算法在多数比较中显著优于主流基线方法. 具体而言, CDWD 在 NMI、F-score 和 ARI 三个指标上均显著优于 Louvain 和 Leiden; 与 LSMD、LMFLS 和 CDME (NMI $p = 0.068$) 等少数优秀方法相比, 虽然未达到统计显著性水平, 但表现出相当的竞争力. 该结果从统计角度验证了 CDWD 算法优势的普遍性和可靠性.

3 结 论

本文提出了一种基于弱边识别与有向传播机制的社区检测算法 CDWD. 算法通过剔除结构脆弱的边, 明确社区边界, 降低对全局结构信息的依赖; 随后构建基于局部相似性的有向影响图, 实现信息沿关键邻居方向高效传播, 强化社区内部结构连贯性; 最后结合弱连通性分析和节点与社区间连接强度, 对孤立节点进行合理分配, 优化社区划分的一致性. 实验结果表明, CDWD 在真实网络、LFR 合成网络及聚类图数据上均取得优异性能, 整体效果优于多种主流社区检测与聚类方法. 同时, 算法对参数设置不敏感, 具有良好的鲁棒性和适用性. 本文提出的 CDWD 算法研究对象为静态网络与非重叠社区检测, 动态网络和重叠社区的扩展将作为未来工作. 具体而言, 未来工作可从三个方向拓展: 将 CDWD 算法扩展至动态网络, 研究节点和边随时间演化对社区结构的影响; 扩展算法至重叠社区检测, 以处理节点可能属于多个社区的复杂网络; 结合更多节点特征信息, 进一步提升社区划分的准确性与可解释性.

参考文献 (References)

[1] Chen M, Chen Y X, Zhu H Y, et al. Analysis of pollutants transport in heavy air pollution processes using a new complex-network-based model[J]. Atmospheric Environment, 2023, 292: 119395.

[2] 武永亮, 窦世卯, 李景辉, 等. 融合异质性和动态性的社区发现研究综述[J]. 计算机工程与应用, 2024, 60(21): 55-72.
(Wu Y L, Do S M, Li J H, et al. A review of community discovery research incorporating heterogeneity and dynamics[J]. Computer Engineering and Applications, 2024, 60(21): 55-72.)

[3] Khawaja F R, Zhang Z P, Memon Y, et al. Exploring community detection methods and their diverse applications in complex networks: A comprehensive review[J]. Social Network Analysis and Mining, 2024, 14(1): 115.

[4] 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.

[5] Traag V A, Waltman L, van Eck N J. From Louvain to leiden: Guaranteeing well-connected communities[J]. Scientific Reports, 2019, 9: 5233.

[6] Zhang T, Lu P L. Detecting communities in complex networks using triangles and modularity density[J]. Physica A: Statistical Mechanics and its Applications, 2023, 613: 128504.

[7] Teng X Y, Luo X Y, Liu J. Transcending modularity: A memetic algorithm combining triangle motif and edge information for community detection[J]. Applied Soft Computing, 2025, 175: 113082.

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

[9] Tasgin M, Bingol H O. Community detection using boundary nodes in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2019, 513: 315-324.

[10] 邓琨, 李文平, 陈丽, 等. 一种新的基于标签传播的复

杂网络重叠社区识别算法[J]. 控制与决策, 2020, 35(11): 2733-2742.
(Deng K, Li W P, Chen L, et al. A novel algorithm for overlapping community detection based on label propagation in complex networks[J]. Control and Decision, 2020, 35(11): 2733-2742.)

[11] Aghaalizadeh S, Afshord S T, Bouyer A, et al. A three-stage algorithm for local community detection based on the high node importance ranking in social networks[J]. Physica A: Statistical Mechanics and its Applications, 2021, 563: 125420.

[12] Bouyer A, Roghani H. LSMD: A fast and robust local community detection starting from low degree nodes in social networks[J]. Future Generation Computer Systems, 2020, 113: 41-57.

[13] Shi D Y, Shang F, Chen B S, et al. Local dominance unveils clusters in networks[J]. Communications Physics, 2024, 7: 170.

[14] Sun P G, Wu X L, Quan Y N, et al. Rearranging 'indivisible' blocks for community detection[J]. IEEE Transactions on Knowledge and Data Engineering, 2022: 1.

[15] Roghani H, Bouyer A. A fast local balanced label diffusion algorithm for community detection in social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(6): 5472-5484.

[16] Zhao Z L, Zhang N N, Xie J Q, et al. Detecting network communities based on central node selection and expansion[J]. Chaos, Solitons & Fractals, 2024, 188: 115482.

[17] Li H X, Nasab S S, Roghani H, et al. LMFLS: A new fast local multi-factor node scoring and label selection-based algorithm for community detection[J]. Chaos, Solitons & Fractals, 2024, 185: 115126.

[18] Stephan L, Zhu Y Z. Sparse random hypergraphs: Non-backtracking spectra and community detection[J]. Information and Inference: A Journal of the IMA, 2024, 13: iaae004.

[19] Deng J Y, Huang D Y, Ding Y, et al. Subsampling spectral clustering for stochastic block models in large-scale networks[J]. Computational Statistics & Data Analysis, 2024, 189: 107835.

[20] 金红, 胡智群. 基于非负矩阵分解的稀疏网络社区发现算法[J]. 电子学报, 2023, 51(10): 2950-2959.
(Jin H, Hu Z Q. The non-negative matrix factorization based algorithm for community detection in sparse networks[J]. Acta Electronica Sinica, 2023, 51(10): 2950-2959.)

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

[22] Dabaghi-Zarandi F, KamaliPour P. Community detection in complex network based on an improved random algorithm using local and global network information[J]. Journal of Network and Computer Applications, 2022, 206: 103492.

[23] Luo D S, Bian Y C, Yan Y W, et al. Random walk on multiple networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(8): 8417-8430.

[24] Shan N, Li L J, Zhang Y K, et al. Supervised link prediction in multiplex networks[J]. Knowledge-Based Systems, 2020, 203: 106168.

[25] 李艳丽, 周涛. 链路预测中的局部相似性指标[J]. 电子科技大学学报, 2021, 50(3): 422-427.
(Li Y L, Zhou T. Local similarity indices in link prediction[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(3): 422-427.)

[26] Qian Y F, Expert P, Panzarasa P, et al. Geometric graphs from data to aid classification tasks with Graph Convolutional Networks[J]. Patterns, 2021, 2(4): 100237.

[27] 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.

[28] Mac Q J. Some methods for classification and analysis of multivariate observations[C]. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. California: University of California Press, 1967, 5: 281-298.

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

[30] Guan J Y, Li S, Zhu J H, et al. Fast main density peak clustering within relevant regions *via* a robust decision graph[J]. Pattern Recognition, 2024, 152: 110458.

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

[32] Amelio A, Pizzuti C. Is normalized mutual information a fair measure for comparing community detection methods? [C]. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. New York: ACM, 2015: 1584-1585.

[33] Chakraborty T, Dalmia A, Mukherjee A, et al. Metrics for community analysis: A survey[J]. ACM Computing Surveys, 2018, 50(4): 1-37.

[34] Steinley D, Brusco M J, Hubert L. The variance of the adjusted Rand index[J]. Psychological Methods, 2016, 21(2): 261-272.

[35] Kehagias A, Pitsoulis L. Bad communities with high modularity[J]. The European Physical Journal B, 2013, 86(7): 330.

[36] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E, 2008, 78(4): 046110.

作者简介

陈梅 (1973−), 女, 教授, 博士, 主要研究方向为人工智能、数据挖掘, E-mail: mei.chen.lzjtu@hotmail.com;

王欢 (2000−), 男, 硕士生, 主要研究方向为数据挖掘、复杂网络, E-mail: 1694608671@qq.com;

付豪杰 (2003−), 男, 硕士生, 主要研究方向为数据挖掘、复杂网络, E-mail: 1946531363@qq.com;

周启辉 (2002−), 男, 硕士生, 主要研究方向为数据挖掘、复杂网络, E-mail: 2021740486@qq.com;

黄欣玥 (2001−), 女, 硕士生, 主要研究方向为数据挖掘、复杂网络, E-mail: 973476240@qq.com.