



上海交通大学硕士学位论文

基于主动学习框架的交互式图社区搜索研究

姓 名：周家纬
学 号：123120910040
导 师：王凯
院 系：安泰经济与管理学院
学 科/专 业：管理科学与工程
申 请 学 位：管理学硕士

2026 年 5 月 15 日



**A Dissertation Submitted to
Shanghai Jiao Tong University for the Degree of Master**

**AN INTERACTIVE COMMUNITY SEARCH
FRAMEWORK VIA GRAPH ACTIVE LEARNING**

Author: Jiawei Zhou

Supervisor: Kai Wang

Antai College of Economics & Management

Shanghai Jiao Tong University

Shanghai, P.R. China

May 15th, 2026

摘要

近年来,图数据上的社区搜索 (Community Search, CS) 任务,受社交网络分析、欺诈检测与生物信息学等下游应用需求的驱动,取得了显著进展。其中,查询依赖型社区搜索作为识别与特定查询节点相关的结构紧密社区的重要工具,受到了广泛关注。相比传统的静态社区检测方法,基于学习的交互式社区搜索 (Interactive Community Search, ICS) 将搜索过程建模为包含人机交互的多轮优化流程,从而在鲁棒性方面展现出更强的实用价值。然而,现有的学习型交互式社区搜索方法在应对大规模图上的复杂查询任务时仍面临两方面挑战:其一,用于缩小搜索空间的现有策略或过度依赖查询语义,或依赖固定的拓扑结构,导致在复杂的大规模图上进行在线社区查询时鲁棒性不足;其二,交互式社区搜索普遍缺乏高效的交互策略,即算法未能有效向用户推荐最具信息量(如高不确定性)的节点以供标注,从而阻碍了搜索质量的快速迭代优化。为解决上述问题,本文提出 *COMET*, 一种面向大规模图的交互式社区搜索框架。*COMET* 包含三个关键模块:第一,基于 Personalized PageRank (PPR) 提出社区感知的子图挖掘方法,为每个查询自适应生成候选子图以兼顾语义与结构相关性;第二,将交互式社区搜索形式化为一系列二分类任务,在每一轮中使用图神经网络 (GNN) 在候选子图内传播标签信息;第三,设计基于主动学习的节点选择机制,综合利用 GNN 输出的预测熵与 PPR 得分,动态确定最具信息价值的节点以增强交互效率。大量实验证明, *COMET* 在八个真实大规模数据集上的性能显著优于当前最先进的基于学习的社区搜索与交互式社区搜索方法,体现了其在准确性、效率及交互效果方面的综合优势。

关键词: 交互式社区搜索, 图局部划分, 主动学习, 图神经网络

Abstract

In recent years, community search (CS) on graph-structured data has achieved significant progress, driven by the demands of downstream applications such as social network analysis, fraud detection, and bioinformatics. Among these, query-dependent community search, as a vital tool for identifying structurally cohesive communities related to specific query nodes, has garnered widespread attention. Compared to traditional static community detection methods, learning-based interactive community search (ICS) models the search process as a multi-round optimization flow involving human-in-the-loop interaction, thereby demonstrating enhanced practical value in terms of robustness. However, existing learning-based ICS methods still face two primary challenges when addressing complex query tasks on large-scale graphs: First, current strategies for narrowing the search space either heavily rely on query semantics or depend on fixed topological structures, leading to insufficient robustness during online community querying on complex, large-scale graphs. Second, ICS methods generally lack efficient interaction strategies; that is, the algorithms fail to effectively recommend the most informative nodes (e.g., those with high uncertainty) to the user for labeling, thereby hindering the rapid iterative optimization of search quality. To address the aforementioned issues, this paper proposes *COMET*, an interactive community search framework designed for large-scale graphs. *COMET* comprises three key modules: First, a community-aware subgraph mining method based on Personalized PageRank (PPR) is proposed to adaptively generate a candidate subgraph for each query, balancing both semantic and structural relevance. Second, ICS is formulated as a series of binary classification tasks, employing a Graph Neural Network (GNN) in each round to propagate label information within the candidate subgraph. Third, an active-learning-based node selection mechanism is designed, which synthesizes the prediction entropy from the GNN output and the PPR scores to dynamically identify the most informative nodes, thereby enhancing interaction efficiency. Extensive experiments demonstrate that *COMET*'s performance on eight real-world large-scale datasets significantly surpasses that of current state-of-the-art learning-based CS and ICS methods, highlighting its comprehensive advantages in accuracy, efficiency, and interaction effectiveness.



Key words: Interactive Community Search, Local Partitioning, Active Learning, Graph Neural Network

目 录

第 1 章 绪论	1
1.1 研究背景与研究问题.....	1
1.1.1 传统基于结构驱动的社区搜索	1
1.1.2 学习驱动与交互式社区搜索	2
1.2 现有方法的关键挑战.....	4
1.2.1 候选子图提取的可扩展性与稳健性不足	5
1.2.2 交互策略的信息增益不足	5
1.3 技术路线: <i>COMET</i> 框架概述	6
1.4 本文研究意义.....	6
1.5 文章结构安排.....	7
1.6 本章小结.....	8
第 2 章 文献综述	9
2.1 图神经网络.....	9
2.1.1 GNN 范式简述	9
2.1.2 GNN 的可扩展性	11
2.1.3 GNN 自监督学习	13
2.1.4 GNN 在复杂图结构上的扩展	14
2.1.5 总结	15
2.2 社区检测.....	15
2.2.1 范式演进	16
2.2.2 卷积网络	16
2.2.3 生成对抗网络	18
2.2.4 总结	19
2.3 社区搜索.....	20
2.3.1 传统社区搜索	20
2.3.2 属性感知与学习驱动的社区搜索	23
2.3.3 交互式社区搜索	25

2.3.4 总结	26
2.4 本章小结.....	27
第 3 章 方法论	28
3.1 问题描述.....	28
3.2 框架概述.....	30
3.2.1 面向低延迟的候选子图收缩	31
3.2.2 基于不确定性采样的主动学习优化	31
3.3 社区感知的子图构建.....	32
3.3.1 PPR 及其高效近似	32
3.3.2 PPR 作为子图提取的实践依据	34
3.4 基于学习算法的在线社区搜索优化.....	36
3.4.1 学习驱动的社区搜索建模	36
3.4.2 子图局部化建模与社区输出	37
3.5 用于交互式社区搜索的节点选择.....	37
3.5.1 主动学习目标与可实现近似	37
3.5.2 基础节点采样策略	39
3.5.3 自适应聚合采样	40
3.6 <i>COMET</i> 的整体流程	41
3.7 本章小结.....	42
第 4 章 实验评估	44
4.1 数据集.....	44
4.2 实验设置.....	46
4.2.1 基线方法	46
4.2.2 <i>COMET</i> 实现	47
4.2.3 评估指标	48
4.2.4 社区和查询设置	49
4.3 有效性评估.....	49
4.3.1 性能评估	50
4.3.2 效率评估	51

4.4 消融实验.....	52
4.5 敏感性分析.....	54
4.5.1 子图构建	54
4.5.2 基于 PPR 的子图构建的参数选择	56
4.5.3 子图表示	57
4.5.4 GNN 架构.....	58
4.5.5 节点选择	58
4.6 合成图模拟.....	59
4.6.1 模拟设置	60
4.6.2 模拟图结果	60
4.7 本章小结.....	62
第 5 章 COMET 扩展与案例分析.....	63
5.1 有关有向图扩展的可行性分析.....	63
5.2 案例研究.....	67
5.2.1 在 Karate 数据集上的交互式社区搜索	67
5.2.2 在 Amazon-CN 数据集上的虚假评论检测	68
5.3 本章小结.....	69
第 6 章 全文结论	71
6.1 主要结论.....	71
6.2 应用价值.....	71
6.3 未来研究展望.....	72
6.4 总结.....	73
参考文献.....	75
致 谢.....	86
学术论文和科研成果目录.....	87
个人简历.....	88

符号对照表

图拓扑结构

V, E	节点集合与边集合
$G(V, E)$	图结构，除非另行说明，默认为无向图
n, m	图 G 的节点数与边数
A	图 G 的邻接矩阵
X	图 G 的特征矩阵
D	图 G 的度矩阵
v_i	图中第 i 个节点
$\mathcal{N}(v_i)$	节点 v_i 的一阶邻居集合
$ \cdot $	取集合或图中元素（如节点）数量的操作

评估指标

$O(\cdot)$	计算复杂度，除非另行说明，默认为时间复杂度
$H(\cdot)$	熵函数
$I(\cdot; \cdot)$	互信息函数
ρ_{local}	局部密度
$\bar{d}_c^{\text{sub}}, \bar{d}_c^G$	社区 c 的平均度（基于 subgraph 与 graph）
μ_c	社区混合水平
C_{local}	聚类系数

PPR 计算

$\text{pr}(\alpha, s)$	原始 PPR 向量
$\text{apr}^{(n)}(\alpha, s)$	第 n 次迭代近似的 PPR 向量
α	传送（跳转）概率
ϵ	近似/幂迭代的误差阈值（收敛判据）
s	个性化向量（通常由指示函数 χ_{Q^+} 定义）
W	归一化转移矩阵， $W = A^\top D^{-1}$

社区建模

Q	查询节点集合 $Q \subset V$
Q^+, Q^-	查询集合中的正样本与负样本节点, $Q = Q^+ \cup Q^-$
B	预算节点集合 $B \subset V$
C_Q	查询集合 Q 的真实社区
CS_Q^t	在第 t 轮节点选择前的候选子图
$\hat{C}_Q^t$	在第 t 轮节点选择后的生成社区
$\hat{C}_Q$	查询集合 Q 的最终输出社区
M^t	第 t 轮的图神经网络 (GNN) 模型
$y(v_i Q)$	节点 v_i 相对于 Q 的标签
$p(v_i M^t)$	节点 v_i 相对于模型 M^t 的社区得分
τ	分类 (社区判定) 的决策边界 (阈值)
$\phi(v_i CS_Q^t)$	主动学习的采样函数 ($\phi(C_q)$ 不加说明指代导通度)

第 1 章 绪论

1.1 研究背景与研究问题

图结构 (Graph) 是一类能够刻画实体及其相互关系的基础数据载体, 广泛存在于社交网络^[1-2]、金融交易网络^[3]、生物分子互作网络^[4] 等复杂系统中。围绕图的结构组织与演化规律的研究, 为刻画网络整体与局部的功能特征提供了重要工具。其中, **社区 (Community)** 通常指代在拓扑上更加紧密连接的节点子集 (即密集子图), 对这些社区结构的深入分析和挖掘, 能够为图的动态演化和局部语义提供关键见解, 进而在诸多下游应用中创造巨大价值, 例如用户购买行为分析^[5]、金融欺诈等异常检测^[6-7] 以及网络信息排名^[8-9]。

在图分析领域, 识别社区结构的核心任务主要沿着两条路径发展。第一条是传统的社区发现 (Community Detection, CD)^[10-12]。社区发现的目标通常是对整个图进行一次性的全局划分, 找出图中所有潜在的社区。然而, 这种全局性方法存在根本局限: 首先, 它试图发现所有社区, 这在动辄亿级节点的大规模网络上往往计算代价高昂且不切实际; 其次, 社区发现算法通常生成固定粒度的簇, 这种基于粒度的划分方式无法满足用户个性化的、即时的查询需求。

鉴于传统社区发现的局限性, 社区搜索^[13] 作为一种以用户为中心的范式应运而生。与社区发现不同, 社区搜索旨在根据用户提供的特定“查询” (一个或一组节点), 在线、实时地识别出包含该查询且高度相关的特定社区。以用户组装电脑为例, 如图 1.1 所示: 当用户查询显卡 (GPU) 和处理器 (CPU) 时, 社区搜索能够返回一个结构紧密且语义相关的高性能 PC 组件社区 (如主板、内存、SSD); 而社区发现则可能将 GPU 和 CPU 划分到两个各自的、粒度固定的显卡簇和处理器簇, 这无法满足用户构建统一解决方案的特定需求。因此, 社区搜索强调局部结构和查询语义的统一, 为用户提供了更灵活、更精确的图探索工具。

社区搜索旨在从图中高效地找出与一个 (或一组) 查询节点相关的内聚子图。根据其核心驱动范式与交互模式的演进, 相关方法大致可分为三个阶段, 如图 1.2 所示。

1.1.1 传统基于结构驱动和社区搜索

早期的社区搜索方法以显式的局部内聚结构约束为主要建模手段。这类方法的核心思想是: 一个“好”的社区, 其成员在拓扑上必须紧密相连, 为此定义了多种严

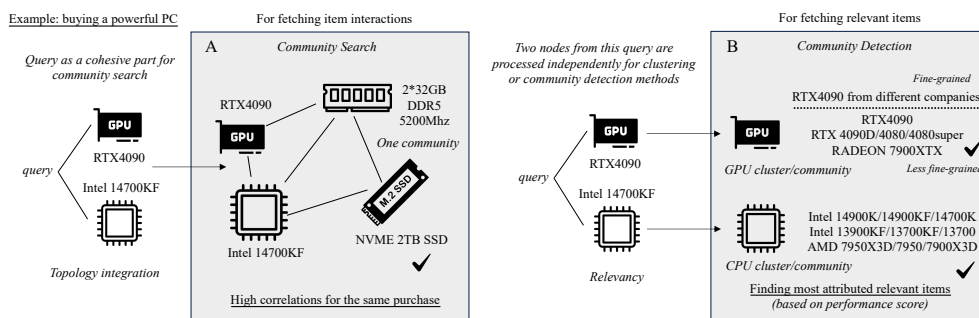


图 1.1 社区发现与社区搜索的区别，以用户商品购买行为为例

Figure 1.1 Comparison of Community Detection and Community Search in the Context of Consumers' Product Purchase Behavior

社区发现 (CD) 与社区搜索 (CS) 的区别：社区发现先计算所有社区，再匹配查询；社区搜索则将查询（如两个节点）视为一个整体，仅搜索由该查询扩展出的社区。

格的结构度量。

最经典的度量包括 k -core（要求子图内每个节点的度至少为 k ）及其相关模型（如 CST/CSM^[14]），以及 k -truss（要求子图内每条边至少被 $k - 2$ 个三角形共同支持）及其索引技术（如 TCP-index^[15]、CTC^[16]、EquiTruss^[17]）。图 1.2 (A) 所展示的便是基于 k -core 的传统社区搜索范例，它通过计算图的核心度来圈定查询驱动的内聚子图。此外，一些局部聚类方法，如基于模块度（Modularity）优化的^[18]或基于 PPR 的随机游走^[19]，亦可在一定程度上挖掘与查询相关的结构。

该类方法的优势在于其定义清晰、可解释性强，并且通常能发展出高效的索引和剪枝技术，适用于快速的在线查询。然而，它们共同的根本缺陷在于缺乏灵活性。它们严重依赖于刚性的、预定义的结构模式。现实世界的社区形态各异，一个在语义上高度一致的真实社区（例如一个兴趣小组或一个学术合作网络）在拓扑上未必满足 k -core 或 k -truss 的严格约束。这种刚性约束导致了一个难以调和的困境：降低阈值（例如 k ）虽然可以提高召回率，确保包含目标社区的更多潜在成员，但极易引入大量语义不相关的噪声节点，导致社区边界模糊。提高阈值虽然能保证挖掘出的社区具有极高的结构纯度，但又可能过度收缩，遗漏掉处于社区“边缘”或扮演“桥梁”角色的关键成员，导致对真实社区边界的刻画失真。

1.1.2 学习驱动与交互式社区搜索

为突破传统结构范式的离散、刚性约束，近年来出现了学习驱动和社区搜索方法，如图 1.2 (B) 所示。其核心思想是利用 GNN 强大的表征学习能力，将节点的拓扑

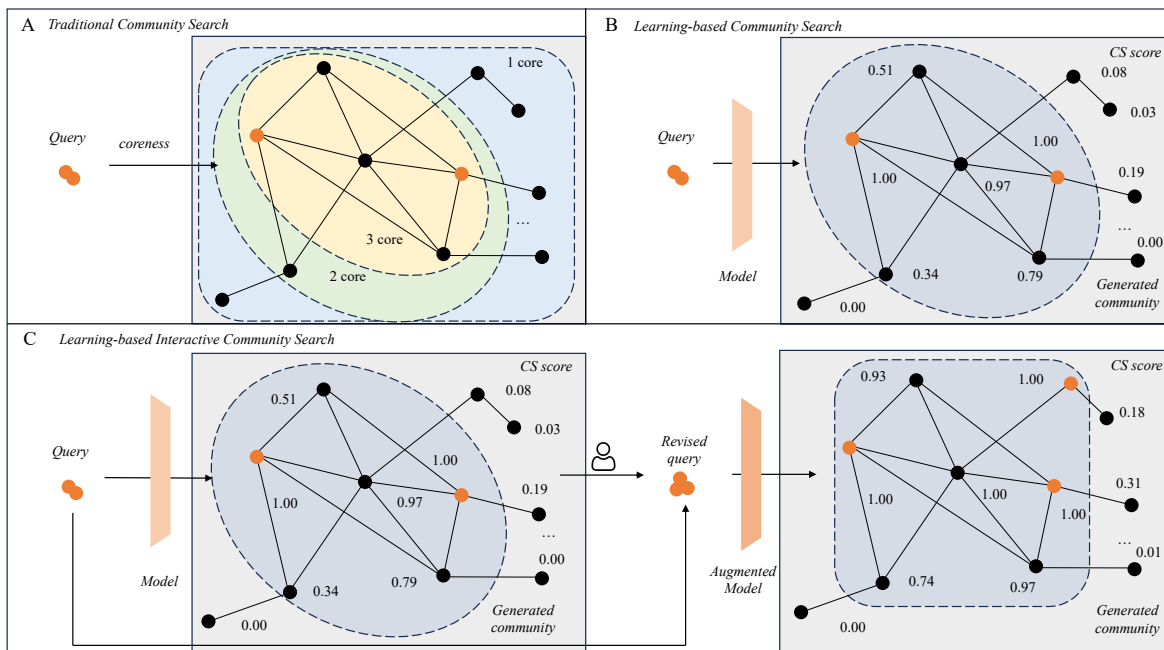


图 1.2 三类社区搜索划分

Figure 1.2 Taxonomy of Three Types of Community Search

A 部分表示基于 K-core 的传统社区搜索方法，其使用图的核心度来查找查询驱动的内聚子图。B 部分表示基于学习的社区搜索方法。C 部分表示一个两轮的、基于学习的交互式社区搜索方法。

结构信息与（可选的）属性特征统一整合到低维、连续的向量空间中。GNN 能够以数据驱动的方式，学习到比 k -core 或 k -truss 更复杂、更灵活的社区结构模式。通过这种方式，社区判别的边界变成了在表征空间中连续可微的优化问题，模型得以适应多样的社区形态与图数据中的噪声分布。代表性工作包括 QD-GNN^[20] 和 COCLEP^[21] 等。

在此基础上，部分工作进一步将社区搜索建模为一个交互式的多轮过程（ICS），如图 1.2 (C) 所示。其动机在于，单一、静态的查询（如一个节点）往往不足以精确表达用户的真实意图。交互式框架允许“人-机”协同：当初始搜索结果未能满足用户期望时，用户可以提供小量的正负样本反馈（例如指出某个节点“是”或“不是”目标社区成员），模型据此迭代优化查询表征或社区边界，从而使输出社区逐渐贴合用户的真实需求。ICS-GNN^[22] 是这一方向的代表性工作。

图 1.3 进一步详细对比了不同学习式架构（特别是 SOTA 方法）在实现机制上的差异。这些差异集中体现在它们如何平衡离线预处理与在线推理，以及它们如何处理图的规模与交互信息：COCLEP^[21] 采用一种“离线-在线”混合策略。它在离线阶段

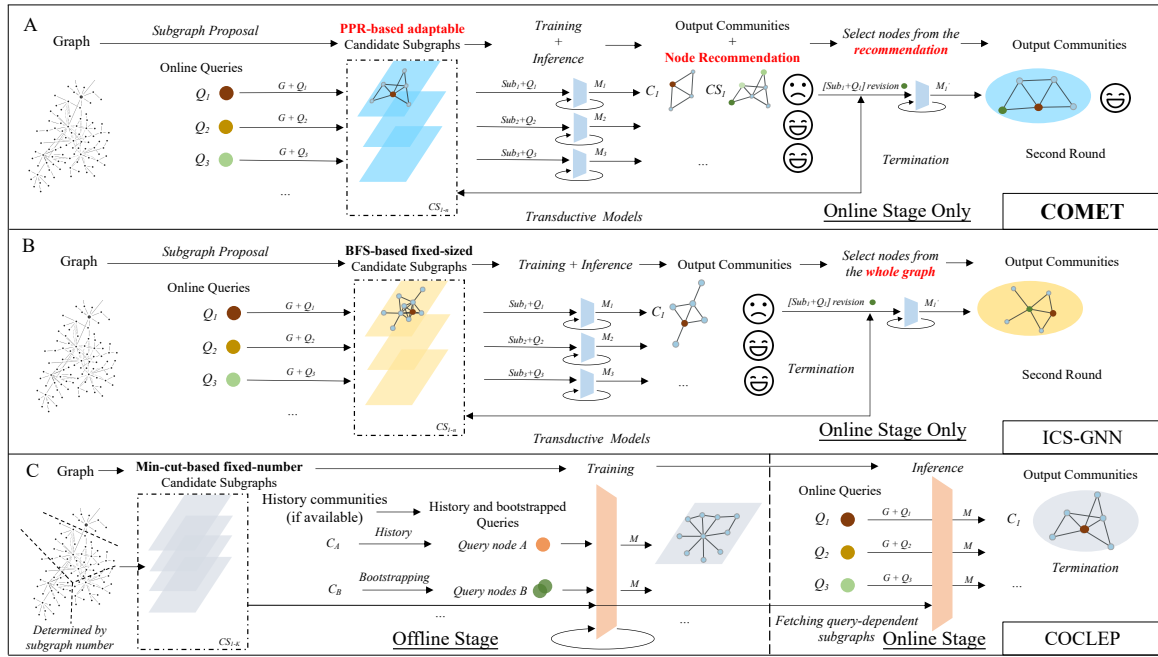


图 1.3 基于学习的社区搜索不同架构的比较

Figure 1.3 Comparison of Alternative Architectures for Learning-Based Community Search

ICS-GNN 基于 BFS 生成定规模子图，依赖在线多轮交互完成社区更新；COCLEP 在离线阶段通过最小割子图划分与历史社区训练，实现高效推理但缺乏查询自适应性；而本文提出的 *COMET* 框架融合了二者优势，采用 PPR 自适应子图采样与节点推荐式反馈机制，在仅在线阶段即可实现动态可扩展的交互式社区优化，从而显著提升社区搜索的灵活性与交互效率。

通过最小割算法（如 METIS）对全图进行子图划分，并利用历史社区数据进行预训练。在线推理时，它能高效地返回结果，但代价是牺牲了对查询自适应的能力——子图划分在查询前已固定，无法根据查询节点动态调整。ICS-GNN^[22] 则是一个纯在线框架。它从查询节点出发，通过广度优先搜索（BFS）生成一个固定规模的子图，然后在该子图上依赖多轮在线交互来迭代更新社区成员。这种方式具有良好的查询自适应性，但其子图提取策略（固定大小 BFS）和交互模式（被动接收反馈）带来了新的挑战。

1.2 现有方法的关键挑战

尽管学习驱动的（交互式）社区搜索在灵活性与鲁棒性上取得进展，但在大规模、在线场景中仍面临两个核心瓶颈：

1.2.1 候选子图提取的可扩展性与稳健性不足

在大规模图上（例如数十亿节点/边的社交网络或知识图谱），直接运行复杂的 GNN 推理是不现实的，这既受限于计算预算（时延与显存），也可能涉及用户隐私约束。因此，所有在线方法都必须先行快速收敛到一个与查询高度相关的候选子图上。

然而，现有方法的子图提取策略普遍缺乏自适应性与稳健性。它们常依赖较为固定的结构来确定子图的粒度。例如，ICS-GNN^[22] 依赖严格固定大小（如 $k = 3000$ 个节点）的 BFS 采样子图（没有考语义信息，也没有特别关注到局部拓扑结构）；而 COCLEP^[21] 则依赖最小割算法（METIS^[23]）将大图划分为固定数量（如 M 个）的子图。此类固定粒度或固定子图数量的方案，虽然实现简单、易于控制计算开销，但它们忽视了现实世界中社区本身在大小、密度、拓扑形态上的巨大异质性。这种策略在面对不同查询时会造成严重的欠/过覆盖问题：当目标社区规模较小但高度密集时，一个固定大小的 BFS 子图可能会引入大量无关的二跳或三跳邻居，造成严重的噪声污染。当目标社区规模庞大但结构稀疏（如一个跨学科的研究领域）时，一个固定大小的 BFS 子图可能只切割了社区的冰山一角，导致社区结构不完整，后续 GNN 无论如何优化也无法恢复完整的社区边界。METIS 在 BFS 的基础上考虑了拓扑结构，但固定的子图数量 M 需要先验知识，难以在不同查询场景中通用。较为详细的架构对比如图 1.3 (B, C) 所示，这种对社区异质性缺乏自适应能力的子图提取阶段，构成了模型性能的第一个关键瓶颈。

1.2.2 交互策略的信息增益不足

ICS 的核心优势在于“人”作为搜索的引导。然而，在昂贵的人工标注成本约束下，交互效率（即每轮交互带来的模型提升）至关重要。现有交互框架（如 ICS-GNN^[22]）大多采用一种被动的交互模式：模型给出一个初步结果，等待用户被动地修订查询（例如自行在结果中寻找并追加正样本）。这种模式并未回答主动学习（Active Learning, AL）领域的核心问题：为了最快提升模型性能，下一步应该标注哪个节点？

缺少对高价值节点（例如模型高度不确定的“边界”节点、或具有高影响力的“枢纽”节点）的主动推荐，将显著降低每轮交互的信息增益和效率收益。这种盲目的、被动的交互模式，迫使用户陷入低效的试错循环：用户可能付出了高昂的标注成本（需要标记很多节点），但模型的搜索质量收敛缓慢，严重阻碍了交互式社区搜索框架的实用价值。

1.3 技术路线：COMET 框架概述

为系统缓解以上两类瓶颈，本文提出 **COMET** (Interactive Community Search via Graph Active Learning) 框架，将社区感知的候选子图挖掘与主动学习驱动的交互统一纳入一个端到端流程（总体示意图 1.3 的 A 部分）：

为解决候选子图的鲁棒性问题，**COMET** 提出了一种基于 **Personalized PageRank (PPR)** 的社区感知子图挖掘方法。PPR^[24] 能够通过模拟随机游走来评估图中所有节点相对于查询节点的重要性和相关性。因此，它天然具备社区感知能力：PPR 修改随机游走概率以偏向个性化节点，通过识别关键节点及其相关子图来辅助图划分，这有助于在大型图中进行内聚的社区搜索。基于以上观察，**COMET** 利用近似 PPR 算法，根据查询自适应地生成候选子图。其核心思想是，在排序和对数转换后的 PPR 向量中，其数值的“急剧下降”点（即最大一阶差分）往往对应着目标社区的“天然边界”。通过这种方式生成的子图，既兼顾了查询语义，又尊重了图的拓扑结构，远比固定大小的 BFS 或固定的图划分更为鲁棒和精确。

为解决交互效率低下的问题，**COMET** 创新性地将交互式社区搜索形式化为一个主动学习过程。**COMET** 不再被动等待用户反馈，而是主动引导交互过程。在每一轮交互中，**COMET** 会通过一个动态聚合评分函数识别出那些模型最不确定但在结构上最重要的节点，并主动推荐给用户进行标记。该评分函数综合考虑了两个维度：1) 节点的拓扑重要性（基于上述模块计算的 PPR 得分）和 2) GNN 模型的预测不确定性（基于模型熵）。通过 log-sum-exp 权重函数和时间调整策略，**COMET** 在交互早期侧重于探索拓扑相关节点 (Exploration)，在后期侧重于修正不确定的社区边界 (Exploitation)。这种主动的节点选择机制，旨在以最小的用户标注代价，实现最快的搜索质量迭代优化，从而极大提升交互效率。

1.4 本文研究意义

本文围绕交互式社区搜索提出 **COMET**，主要贡献概括如下：

- 提出一个面向大规模图的交互式社区搜索统一框架。**COMET** 将近似 PPR 的社区感知子图构建与主动学习的节点选择耦合为闭环流程。在保证搜索社区结构的同时，**COMET** 在统一配置下表现稳健，显著压缩在线搜索空间，提升鲁棒性与时效性。

- **面向交互式社区搜索的主动学习策略。**构建融合拓扑重要性与模型熵的不确定性评分，并通过时间调整实现探索—利用的动态权衡。*COMET* 支持多节点选择以适配不同标注预算，在多轮交互中持续提升质量。
- **广泛的实验评估。**在多个真实大规模数据集上与代表性基线（含传统与学习性社区搜索）对比，*COMET* 在准确性、效率与交互收益上均取得显著优势；并通过敏感性分析、消融实验与合成图模拟验证其稳健性与可迁移性。
- **有向图扩展性验证与案例分析。**将 *COMET* 扩展至有向图情形，并给出详细案例研究，展示其在不同图类型与应用情境下的适用性与可解释性。

1.5 文章结构安排

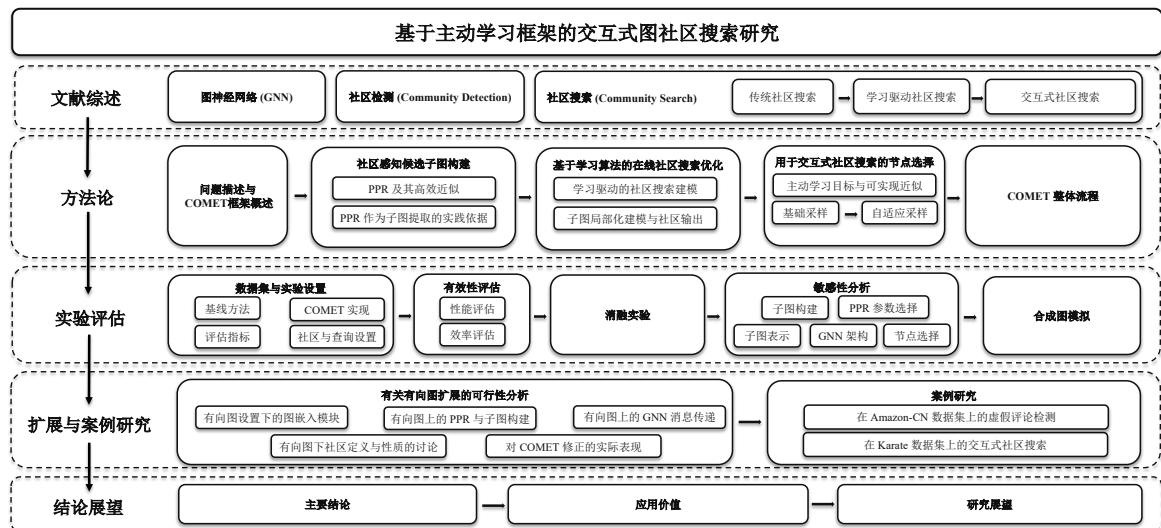


图 1.4 文章结构安排
Figure 1.4 Organization of the Thesis

全文结构可参考图1.4。第2章系统回顾图神经网络 (GNN)、社区发现 (CD) 与社区搜索 (CS) 等相关研究；第3章形式化定义问题并给出 *COMET* 的三大模块（社区感知候选子图构建、用于在线社区搜索的 GNN、主动学习驱动节点选择）；第4章给出实验设置与结果，包括性能与效率对比、消融与敏感性讨论以及合成图模拟；第5章展示 *COMET* 在有向图与实际案例中的扩展与落地；第6章总结全文并讨论未来工作。

1.6 本章小结

本章首先阐述了图数据分析中社区搜索的研究背景，明确辨析了其与传统社区发现任务的本质区别，并强调了以用户为中心的在线社区搜索在个性化挖掘任务中的重要价值。随后，本章系统梳理了社区搜索技术的发展脉络，从早期的结构约束方法（如 k -core）演进至近期的图神经网络驱动及交互式方法，并深入剖析了现有技术在大规模在线场景下所面临的候选子图鲁棒性缺失与交互效率低下的双重挑战。针对上述痛点，本章引出了本文的核心解决方案——基于主动学习的交互式社区搜索框架 *COMET*，简要概述了其利用 PPR 自适应子图构建与主动节点推荐策略来提升搜索性能的技术路线。最后，本章总结了本文的主要研究贡献，并给出了后续章节的组织安排。

第2章 文献综述

围绕图社区结构的理解与利用,对于研究主要沿两条脉络发展:社区检测 (Community Detection, CD) 与社区搜索 (Community Search, CS)。二者的核心差异在于目标与输出范围:社区检测从全局视角在给定粒度上发现所有社区^[25-26]; CS 面向查询驱动的局部发现,仅识别包含给定查询节点 (或节点集合) 的个性化社区^[13,27]。近年来,以图神经网络 (Graph Neural Networks, GNNs) 为代表的图表示学习技术,使得结构-属性联合建模成为现实,在社区检测与社区搜索的灵活性、扩展性与精度上带来系统性跃迁^[28-29]。本节系统回顾图神经网络、社区检测与社区搜索的关键理论与方法,并总结实践挑战与研究趋势。

2.1 图神经网络

图神经网络 (Graph Neural Networks, GNNs) 已发展成为处理图结构化数据的核心方法,并为社区检测 (Community Detection, CD) 与社区搜索 (Community Search, CS) 等任务提供了统一的“结构-属性”协同建模框架。与图像、语音等欧氏数据不同,图数据具有典型的非欧结构:节点邻域大小不固定、拓扑连接不规则,且不存在天然的全局顺序。这使得卷积神经网络 (CNN) 中依赖规则网格与平移不变性的经典卷积算子难以直接迁移到图域。因此,早期 GNN 研究的核心问题在于:如何在图上定义有效的局部表示学习算子,并进一步实现可泛化、可扩展的端到端训练。

围绕这一问题,图神经网络的发展大体经历了从谱方法 (spectral methods) 到空间方法 (spatial methods) 的演进。谱方法借助图信号处理理论,在图拉普拉斯矩阵的特征空间中定义卷积;空间方法则直接在顶点域上设计邻域聚合规则,从而避免对固定图谱的依赖。随着研究的推进,空间方法因其更强的灵活性、归纳能力与工程可扩展性,逐渐成为主流;而大量空间 GNN 又可进一步统一到消息传递神经网络 (Message Passing Neural Network, MPNN) 的建模框架之下。

2.1.1 GNN 范式简述

谱方法 GNN 的早期探索始于谱方法,其基本思想源于图信号处理。给定图的归一化拉普拉斯矩阵

$$L_{sym} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \quad (2.1)$$

由于 L_{sym} 是对称半正定矩阵，其特征分解

$$L_{sym} = U\Lambda U^T \quad (2.2)$$

给出了图的谱基 U 及对应频率 Λ 。据此，图信号 x 的图傅里叶变换可写为 $\hat{x} = U^T x$ ，逆变换为 $x = U\hat{x}$ 。在这一框架下，Bruna 等人^[30] 最早将卷积定义到图谱域中，其基本形式为

$$x' = U g_{\theta}(\Lambda) U^T x \quad (2.3)$$

其中 $g_{\theta}(\Lambda)$ 表示谱域滤波器。该定义为图卷积提供了严格的数学基础，但也带来了明显局限：其一，特征分解计算代价高，难以扩展到大图；其二，谱滤波器依赖于特定图的特征基，缺乏跨图泛化能力；其三，早期谱卷积的空间局部性并不显式。

ChebNet 在此基础上作出了关键推进。其核心思想是不再显式学习任意谱滤波器，而是利用 K 阶切比雪夫多项式 T_k 对其进行近似：

$$g_{\theta}(\Lambda) \approx \sum_{k=0}^K \theta_k T_k(\tilde{\Lambda}) \quad (2.4)$$

这里 $\tilde{\Lambda}$ 是缩放的特征值。从而可将卷积写为

$$x' \approx \sum_{k=0}^K \theta_k T_k(\tilde{L}) x \quad (2.5)$$

其中 $\tilde{L} = \frac{2}{\lambda_{max}} L - I$ 。同时，由于多项式滤波器对应有限跳数邻域，卷积也获得了明确的空间局部性。ChebNet 因此被普遍视为连接谱方法与后续实用 GNN 的重要桥梁。

空间方法 在 ChebNet 的基础上，GCN^[31] 可被看作谱卷积的一阶近似。通过进一步的近似与参数简化，Kipf 和 Welling 得到了经典传播公式

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}\right) \quad (2.6)$$

其中 $\tilde{A} = A + I$ 表示加入自环后的邻接矩阵， $\tilde{D}$ 是 $\tilde{A}$ 对应的度矩阵。GCN 的重要性在于，它虽然源于谱理论，却呈现出清晰的空间解释：每一层本质上都是对节点自身及其一阶邻域特征进行归一化聚合。也正因为这一简洁形式，GCN 成为现代 GNN 最具代表性的基础模型之一。

进一步地，空间方法逐渐摆脱了对固定传播算子的依赖，转向直接在顶点域定义可学习的邻域聚合函数。GraphSAGE^[32] 是这一方向的重要里程碑。与 GCN 相比，它

不再依赖整图上的固定归一化传播,而是学习一个可共享的聚合器(如 Mean、Pooling 或 LSTM Aggregator),从而具备更强的归纳能力,并可通过邻居采样实现 mini-batch 训练。随后, GAT^[33] 又将自注意力机制引入图邻域聚合,通过学习节点间的注意力系数自适应地分配邻居权重,进一步提升了模型的表达灵活性。由此,空间 GNN 从“固定传播规则”逐步发展为“可学习的局部消息聚合机制”。

统一的消息传递视角 尽管具体模型形式各异,绝大多数现代 GNN 都可以统一到消息传递神经网络 (MPNN)^[28-29,34] 框架下理解。MPNN 将一层 GNN 抽象为三个基本步骤:首先,节点根据自身特征、邻居特征及边信息构造消息;其次,对来自邻域的消息进行置换不变的聚合;最后,结合节点历史状态完成表示更新。该框架表明,谱方法与空间方法在实践中的分野,最终都可归结为对“消息构造函数”“聚合算子”与“更新规则”的不同设计。

围绕这一统一框架, GNN 的设计通常体现在若干关键维度上。其一,聚合器的选择决定了模型的表达偏好与数值性质,例如 Sum^[35]、Mean^[32]、Max^[32] 以及 Attention^[33] 等聚合方式分别对应不同的归纳偏置。其二,网络深度决定了感受野范围,但更深的消息传递也会带来过平滑^[36] 等训练与表达问题,因此出现了诸如残差连接 (Residuals)^[36-37]、DropEdge^[38] 和层归一化 (LayerNorm)^[37] 等缓解技术。其三,当图中包含边类型、关系语义或更复杂结构时,模型还需要进一步扩展其消息传递机制,以适应异构关系、动态图演化等更复杂的数据形态。

除模型结构本身外, GNN 在真实应用中还面临两类重要约束:一类来自图规模的急剧增长,另一类来自监督信号的稀缺。前者推动了可扩展 GNN 的发展,后者则促成了 GNN 自监督学习 (GSSL) 的兴起。

2.1.2 GNN 的可扩展性

GNN 在真实世界中的应用面临一个核心瓶颈:图规模往往极大,工业场景中常见十亿级节点与百亿级边。每次直接更新全图带来性能与效率上的挑战。基于此延伸出了 *full-batch* 和 *mini-batch* 上的训练目标。前者每一次参数更新都在整张图上执行完整的 L 层消息传递,并基于全部训练节点计算梯度;而 *mini-batch* 则对每一步只针对一部分监督目标 $B \subset V$ 计算损失,并仅构造这部分目标所需的局部计算子图 $\mathcal{G}_B$,从而得到随机梯度 $\hat{\mathcal{L}}_B(\theta)$ 。因此,可扩展 GNN 的核心目标是避免每次更新都在整张图上进行 full-graph 传播,转而通过局部化的计算子图实现 mini-batch SGD。

现有方法大体可以归纳为两类:邻居采样 (neighbor sampling) 与子图采样 (Sub-

graph Sampling)。前者以节点为中心,通过逐层采样种子节点的多跳邻域构造计算子图;后者则直接以局部子图作为计算单元,其中既包括随机子图采样方法,也包括基于图划分的簇级训练方法。

邻居采样 (Neighbor Sampling) 邻居采样由 GraphSAGE^[32] 开创,是一类典型的节点驱动 (node-wise) mini-batch 方法。其基本思想是:先选取一小批输出端的种子节点 (seed nodes) V_B ,再从这些节点出发,对其 L 层依赖进行逐层反向展开;在第 ℓ 层,为每个节点最多采样 S_ℓ 个邻居,最终形成该批节点对应的计算子图,随后再按消息传递方向自底向上完成聚合。该类方法的优点是单步计算量可以通过采样数 $\{S_\ell\}_{\ell=1}^L$ 控制,其感受野规模上界约为 $\prod_{\ell=1}^L S_\ell$ 。然而,当层数增加时,依赖展开会带来显著的“邻居爆炸” (neighbor explosion) 问题:一方面 I/O 与访存开销迅速增加,另一方面不同种子节点的重叠邻域往往会被重复计算,从而造成明显的冗余。此外,随机采样本身也会引入梯度方差并影响最终精度。VR-GCN^[39] 等工作进一步通过复用历史表征、构造方差缩减估计器来缓解这一问题。

子图采样 (Subgraph Sampling) 与邻居采样不同,基于子图的方法不再围绕种子节点逐层展开依赖,而是直接构造一个局部子图作为 mini-batch,并将 GNN 的全部 L 层传播限制在该子图内部。该类方法能够避免逐层扩张带来的邻居爆炸,并通常具有更好的计算局部性。根据子图构造方式的不同,这一路径又可进一步分为随机子图采样与基于图划分的簇级训练两类。

以 GraphSAINT^[40] 为代表的随机子图采样方法,在每一步训练前直接从原图 G 中采样一个子图 G_s (例如基于节点采样、边采样或随机游走采样),并在该子图内部完成完整的 L 层传播。其优势在于避免了邻居爆炸,并且采样得到的子图通常具有较好的连通性与局部结构保真性,更有利于保留原图中的社区结构;其挑战则在于,采样子图通常并非原图训练目标的天然无偏代表,因此需要引入额外的归一化与重加权机制来修正采样偏差。

以 Cluster-GCN^[41] 为代表的簇级训练方法,则首先利用 METIS^[23,42] 等图划分算法对原图进行离线划分,将大图分解为若干簇;训练时,每个 mini-batch 由一个或多个簇所诱导的子图构成,并将消息传递限制在这些簇对应的子图内部。与随机子图采样相比,该方法的 batch 结构更加规整,通常具有更高的吞吐量和更好的硬件利用率;但其代价是被切断的簇间边无法在当前 batch 中参与传播,从而产生边界效应,影响跨簇信息交互。为了解决上述问题,其提出了一种随机多簇聚合方法,在簇接之间进

行合并，并减少 batch 间的差异。

2.1.3 GNN 自监督学习

GNN 在现实应用中的另一项重要挑战是标签稀缺。为缓解对人工标注的依赖，GNN 自监督学习（Graph Self-Supervised Learning, GSSL）试图从图数据自身的结构信息与属性信息中构造监督信号，从而学习具有迁移性和泛化能力的节点或图表征。现有研究较为丰富，本文仅从两条具有代表性的技术路线进行概述：生成式方法与对比式方法。

生成式方法 生成式方法通常通过设计重构目标来学习图表征，其中图自动编码器（GAE）及其变分形式 VGAE^[43] 是最具代表性的工作。其基本思想是：利用 GNN 编码器将图 G 映射为节点表征 $Z = \mathcal{M}(A, X)$ ，再通过解码器根据 Z 重构图结构，例如采用内积解码器得到 $\hat{A} = \sigma(ZZ^T)$ 。模型训练的监督信号来自重构结果与原始图结构之间的差异。该类方法形式直观、实现简单，并在链路预测等任务上表现良好；但其局限也较为明显：一方面，重构目标通常更偏向于保留局部连接模式，所学习的表示未必能够充分捕获下游任务所需的高级语义信息；另一方面，当重构对象直接涉及邻接矩阵时，计算与存储开销会随图规模迅速增长，从而限制其在大规模图上的适用性。

对比式方法 相比生成式方法，对比学习已成为近年来更主流的 GSSL 范式。其核心思想是通过构造正样本对与负样本对，引导模型学习判别性更强的图表征。早期代表工作 DGI^[44] 采用“局部-全局”对比策略，通过最大化节点表示与全局图摘要之间的互信息来进行无监督训练。该方法表明，对比式目标能够有效提升节点表示的质量，并为后续研究奠定了基础。

后续工作进一步借鉴了计算机视觉中 SimCLR^[45] 等思想，将图对比学习发展为基于多视图增强的“实例-实例”对比范式。以 GRACE^[46] 和 GCA^[47] 为代表，这类方法通常先对原图施加随机扰动，生成两个语义相关但形式不同的增强视图；随后利用共享参数的 GNN 编码器分别提取两种视图中的节点表示，并将同一节点在不同视图下的表示作为正样本对、不同节点的表示作为负样本对；最后通过 InfoNCE 等对比损失，使模型在拉近正样本的同时区分负样本。相比早期的局部-全局对比，这种基于数据增强的实例级对比通常能够学习到更具鲁棒性与泛化性的图表示，因此已成为现代 GSSL 的主流框架之一。

2.1.4 GNN 在复杂图结构上的扩展

标准 GNN 范式主要面向同质且静态的图，即默认所有节点与边具有统一类型，且图结构与节点属性在训练和推理过程中保持不变。然而，现实世界中的图数据往往具有更复杂的结构特征，例如多类型实体及关系共存，或者图连接与节点状态随时间持续演化。为适应这些更贴近真实场景的数据形态，GNN 需要在建模对象与消息传递机制上作进一步扩展。本文主要从两个具有代表性的方向进行概述：异构图与时变图。

异构图 (Heterogeneous Graphs) 异构图中通常包含多种类型的节点与边，不同类型实体往往具有不同的语义角色和特征空间。因此，标准 GNN 中“所有节点共享统一参数、所有边使用统一传播规则”的假设不再成立。围绕这一问题，现有异构图 GNN 大致形成了两类代表性思路。

第一类方法直接对不同关系进行区分建模。以 R-GCN^[48] 为代表，这类方法为不同关系引入关系特定的消息传递参数，从而使模型能够感知边类型差异；同时，为避免关系种类过多带来的参数膨胀，通常还会配合参数共享或低秩分解等机制。

第二类方法则通过显式建模异构图中的高阶语义关系来进行学习，其中元路径方法最具代表性。以 HAN^[49] 为代表，这类方法首先基于预定义的元路径提取不同语义视角下的邻域关系，再利用注意力机制对不同语义信息进行融合，从而获得适合下游任务的节点表示。总体来看，异构图 GNN 的核心目标是在统一框架下同时刻画“关系差异性”与“语义多样性”，这也是该方向持续演进的关键。

时变图 (Temporal / Dynamic Graphs) 与静态图不同，时变图中的节点属性、边连接乃至交互事件都会随时间变化。因此，模型不仅需要学习图的结构依赖，还必须进一步刻画时间演化规律。围绕这一问题，动态图 GNN 一般可以概括为两类主流范式。

一类是基于离散时间快照的建模方法。该类方法将动态图表示为一系列按时间排序的图快照，并结合 GNN 与序列模型对其进行编码。例如，GCN-LSTM^[50] 等方法先在每个时间步上利用 GNN 提取结构表示，再利用 RNN、LSTM 或 GRU 等时序模块建模跨时间依赖。该范式实现相对直接，适合处理具有明显时间粒度划分的场景。

另一类是基于连续时间事件流的建模方法。相比快照模型，这类方法直接将动态图表示为带时间戳的交互事件序列，能够更细粒度地刻画异步演化过程。以 TGN^[51] 为代表，这类方法通常为节点维护随时间更新的状态表示，并在新事件到来时结合历

史状态、当前交互及时间信息进行动态更新。连续时间模型因而能够更自然地处理不规则交互、长期依赖与实时演化等问题，已成为动态图表示学习中的重要发展方向。

2.1.5 总结

综上所述，GNN 的发展历程，是深度学习对非欧几里得数据结构一次深刻的卷积泛化探索。从早期基于图谱理论、计算昂贵且受限于 Transductive 设置的谱方法（如 GCN 的前身），到凭借邻域聚合思想、实现归纳学习与高效计算的空间方法（如 GraphSAGE, GAT），GNN 最终被统一在灵活的消息传递（MPNN）框架之下。这一框架通过迭代的“聚合-更新”操作，实现了节点结构与属性的协同编码。然而，GNN 的强大能力也伴随着诸多挑战：为了将其应用于工业界十亿级节点、百亿级边的超大规模图，邻居采样（GraphSAGE）、子图采样（Cluster-GCN）等可扩展技术应运而生，它们的核心是在精度与效率之间进行权衡。同时，为解决现实世界中普遍存在的标签稀缺问题，以图自动编码器（GAE）和图对比学习为代表的自监督范式，通过挖掘数据自身的结构与属性信息构造监督信号，学会了鲁棒且可迁移的图表征。最后，面对数据日益增长的复杂性，GNN 进一步拓展出 R-GCN、HAN 等异构图模型以处理多类型实体与关系，并发展出 TGN 等时变图模型以捕捉动态演化模式。GNN 及其异构、时变扩展，已成为推荐系统（如 PinSage^[2]）、反欺诈^[52]、舆情分析^[53] 等关键领域中不可或缺的可复用骨干技术，持续推动着图智能的边界。

2.2 社区检测

社区发现（CD），亦常称为社区检测，是网络科学与图挖掘领域长期以来的核心问题之一^[25]。其目标是在给定网络中识别若干“内密外疏”的节点子集，即社区^[54-55]。通常而言，同一社区内的节点在连接模式、属性特征或功能角色上具有更高的相似性，而与社区外节点则存在较为显著的差异^[56]。这一现象在社会网络中常被视为“同质性”（Homophily）原理的体现，即“物以类聚，人以群分”^[57]。

形式化地，给定网络 G ，社区检测的目标是求得节点集合 V 上的一个社区结构表示 $C = \{C_1, \dots, C_K\}$ 。根据社区之间是否允许共享节点， C 可以分为非重叠（disjoint）与重叠（overlapping）两类：前者要求任意两个社区满足 $C_i \cap C_j = \emptyset$ ；后者则允许节点同时属于多个社区，从而更贴近现实场景中个体具有多重角色或多重兴趣归属的事实，例如社交网络用户可能同时参与多个兴趣群体^[54]。

2.2.1 范式演进

在深度学习兴起之前,社区检测研究主要围绕网络拓扑结构展开,并形成了较为系统的经典方法谱系,包括图划分、统计推断、层次聚类、动力学方法、谱方法、基于密度的算法以及质量函数优化等方向。这些方法共同奠定了社区检测研究的基础。具体而言,图划分方法(如 Kernighan–Lin 算法^[10])与谱聚类方法^[58]主要利用图的代数与矩阵性质刻画社区边界;统计推断方法,尤其是随机块模型(Stochastic Block Model, SBM)及其变体^[59-61],从生成式建模角度对社区结构进行解释与估计;层次聚类方法(如 Girvan–Newman 算法^[26])强调社区的多层嵌套关系;动力学方法(如标签传播算法 LPA^[62]和 InfoMap^[63])则借助信息传播或随机游走过程识别网络中的结构模块;此外,以模块度(Modularity)最大化为代表的优化方法^[64]及其高效实现(如 Louvain^[12])在大规模网络分析中也取得了广泛应用。

尽管上述传统方法在社区检测领域具有奠基性意义,并在许多简单或中等规模网络上取得了良好效果,但当面对现实世界网络中普遍存在的复杂拓扑、高维属性、异质关系以及非线性依赖时,其表达能力与适应性往往受到限制^[56]。特别地,若仅依赖网络连接关系进行建模,容易忽略节点属性、语义信息及高阶结构模式,从而导致检测结果次优;与此同时,一些传统方法在大规模图上的计算开销也较为显著。

深度学习的发展为克服上述局限提供了新的研究路径,并推动社区检测逐步从“基于手工设计准则的结构划分”转向“基于表征学习的社区建模”^[65]。这类方法的核心思想在于:通过学习得到低维且信息密集的节点表示,再在嵌入空间中刻画社区结构。相较于传统方法,深度学习驱动社区检测通常具有以下优势:其一,能够建模复杂的非线性关系^[56];其二,能够将高维网络结构压缩为低维嵌入,并在压缩过程中保留关键的结构信息^[66-68];其三,可以自然融合节点属性等辅助信息,以提升对社区归属的判别能力^[69-71];其四,具备更强的可扩展性,从而更适用于大规模、高稀疏、异构及动态网络场景^[72-76]。

基于上述考虑,本文在本节中不再对传统社区检测方法和高阶结构驱动方法作展开式综述,而是将重点放在基于学习的社区检测范式,尤其是深度学习驱动的现代方法上。

2.2.2 卷积网络

在基于学习的社区检测方法中,卷积网络是最早形成系统范式的一类方法。其基本思想是通过局部邻域聚合与层间特征变换,学习能够表征社区结构的节点表示。该

范式最初源于面向规则网格数据的卷积神经网络 (CNN)，随后被推广到非欧式图结构，发展为图卷积网络 (GCN) 及进一步引入邻域重要性建模的图注意力网络 (GAT)。由于传统 CNN 对输入结构具有较强的规则性要求，将其直接用于社区发现往往需要额外的图到网格映射，适用性有限，因此现有研究主要集中于以 GCN 和 GAT 为代表的图卷积模型。基于此，下面重点讨论这两类网络在社区检测中的应用与演进。

图卷积网络 GCN^[31] 可视为将 CNN 的局部卷积思想推广到图结构上的代表模型，其核心可由谱图卷积的一阶近似得到。通过堆叠图卷积层，GCN 能够不断聚合节点邻域信息，从而学习同时包含局部连接模式与更高阶结构线索的节点表示，这为社区检测提供了统一的表示学习框架。

围绕 GCN 的社区检测研究，大体可以从任务设定、模型改进与场景扩展三个方面理解。首先，在任务设定上，GCN 既可用于 (半) 监督的社区分类，也可用于无监督的表示学习与后续聚类。在监督方向上，LGNN^[77] 对 SBM 进行了改进，将非回溯算子与置信传播的消息传递机制结合起来，并通过交叉熵损失进行训练；在半监督场景中，针对标准 GCN 并非专为社区发现设计的问题，MRFaSGCN^[66] 将马尔可夫随机场 (MRF) 扩展为新的卷积层 (eMRF)，以社区感知的方式对 GCN 输出进行平滑与细化。对于重叠社区检测，NOCD^[78] 将 Bernoulli-Poisson 概率模型与两层 GCN 结合，以学习节点对多个社区的归属向量，从而适应更符合现实场景的重叠社区结构。

其次，在模型改进层面，GCN 在社区检测中的一个核心困难是过平滑问题，即随着层数加深，节点表示趋于同质，导致社区边界被削弱。为缓解这一问题，GCLN^[79] 借鉴 U-Net 的对称结构，通过收缩路径与扩展路径融合上下文信息和局部细节，以在更深网络中维持社区区分性。除过平滑外，如何使卷积操作更贴合社区结构本身也是一个重要方向。CayleyNets^[80] 提出高效的 Cayley 谱滤波器，侧重于窄带尤其是低频信息的建模，从而更有效地提取与社区结构相关的平滑模式。

再次，在复杂场景扩展方面，当网络同时包含拓扑结构与节点属性时，社区检测不再仅依赖邻接关系，而需要对结构信号和属性信号进行联合建模。围绕这一点，AGC^[69] 融合低通滤波思想，并通过 k 阶图卷积自适应选择合适的邻域范围；AGE^[81] 则以平滑滤波为核心，自适应执行节点相似性度量与 t 次拉普拉斯平滑，以提升属性图上的社区划分效果。进一步地，一些工作开始突破“先表示学习、后聚类”的两阶段范式，转向社区表示学习与目标函数的联合优化。例如，SENet^[82] 将谱聚类损失直接纳入 GCN 的输出学习过程，CommDGI^[83] 则通过联合最大化节点-社区互信息与图模块度，同时优化表示学习与社区划分。总体而言，GCN 类方法构成了卷积式社

区检测的基础框架，其演进主线体现为：从一般图表示学习走向面向社区结构的任务定制、结构改进与目标联合优化。

图注意力网络 在 GCN 采用固定归一化邻接权重进行聚合的基础上，GAT^[33] 进一步引入可训练的注意力机制，以自适应评估不同邻居对目标节点表示的贡献。对于社区检测而言，这意味着模型能够在聚合过程中突出与社区归属更相关的邻域信息，并抑制噪声连接带来的干扰。因此，相较于 GCN，GAT 尤其适合处理关系类型更丰富、语义异质性更强的复杂网络。

现有基于 GAT 的社区检测研究主要集中在多路复用网络与异构信息网络两类场景。在多路复用网络中，不同类型的关系边共同决定社区结构，关键在于如何区分各关系层的重要性。DMGI^[67] 为不同关系类型分别学习节点嵌入，并结合对比学习与注意力机制强调跨关系共享的全局特征，从而削弱不重要关系的干扰；考虑到该方法对节点属性等内在信号利用不足，HDMI^[70] 进一步通过语义注意力设计融合模块，同时建模内在信号、外在关系及其交互。

在异构信息网络 (HIN) 中，由于节点与边类型多样，社区结构往往依赖更高层次的语义路径。围绕这一特点，许多方法以元路径为核心建模单元。MAGNN^[75] 在元路径内部采用节点级注意力，并在不同元路径之间引入元路径级注意力，以同时刻画异构拓扑与语义差异；HeCo^[84] 则联合建模网络模式与元路径，并借助改进的对比学习与注意力机制提升表示质量。然而，元路径设计通常依赖较强的领域知识，这会限制模型的通用性。针对这一问题，CP-GNN^[85] 不再依赖预定义元路径，而是通过“非预定义上下文路径”学习节点表示，并以注意力机制区分不同关系的重要性，从而降低了先验知识依赖。

总体来看，GAT 类方法通过引入可学习的权重分配机制，进一步增强了社区检测模型在多关系、异构和语义复杂网络中的表达能力，也推动卷积网络范式从均匀邻域聚合迈向更细粒度的结构选择性建模。

2.2.3 生成对抗网络

生成对抗网络 (GAN)^[86] 通过生成器 (M_g) 与判别器 (M_d) 之间的博弈式训练，同时提升样本生成能力与判别能力。将 GAN 引入社区检测的核心动机，在于利用对抗学习增强社区表示的区分性与模型鲁棒性；但与此同时，由于社区检测任务中可用监督信号通常有限，GAN 也可能面临训练不稳定及过拟合等问题。

在社区检测场景下，对抗学习中的“真实样本”与“伪样本”并不局限于单一形

式，而可以定义在不同的结构粒度上，这也构成了相关方法的主要差异。针对稀疏网络中邻近性信息不足的问题，ProGAN^[71] 在节点局部结构层面进行建模，通过实例化三元组 (triplets) 来编码节点邻近性，从而增强表示学习对局部关系的刻画能力。对于重叠社区检测，CommunityGAN^[87] 将对抗过程提升至高阶结构层面，在特定 motif (如团, cliques) 上构造生成器与判别器，以更好地捕获重叠社区中的高阶连接模式。进一步地，SEAL^[88] 将生成粒度提升至社区层面：其以种子节点为起点，利用增量更新的图指针网络 (iGPN) 作为生成器生成种子感知社区，并采用图同构网络 (GIN) 作为判别器对生成结果进行评估。

除直接用于社区生成或结构判别外，GAN 还常与其他模型结合，以应对社区检测中的特定挑战。例如，针对社区类别分布不平衡问题，DR-GCN^[89] 引入条件 GAN 机制，并设计潜在分布对齐与类条件对抗两类正则项，其中前者以 KL 散度衡量潜在分布差异；而 CANE^[90] 则将 LDA 型社区检测纳入对抗式表示学习框架，以增强社区表示与生成过程之间的一致性。

总体而言，GAN 类方法并非社区检测中的主流框架，但其优势在于能够通过对抗训练强化社区表示的判别性，并在稀疏网络、重叠社区及类别不平衡等场景下提供补充性的建模能力。

2.2.4 总结

总体而言，社区发现研究已经完成了从传统结构划分范式向深度表示学习范式的重要转变。早期方法主要依赖模块度优化^[26]、随机块模型 (SBM)^[59] 等手工设计准则或统计生成机制来刻画社区结构；而近年来，以 GCN^[31]、GAT^[33] 为代表的深度学习方法，则通过端到端表示学习显著提升了对复杂非线性结构、多源异构信息以及高维属性特征的建模能力。总体来看，深度学习的引入使社区发现方法在表达能力、鲁棒性与适用范围等方面均获得了明显提升^[56]。

尽管如此，深度学习驱动社区发现仍然面临若干关键挑战。首先，在表示与建模层面，如何在无监督条件下自适应地确定未知社区数量 K ，仍然是一个基础而开放的问题；与此同时，现有方法大多以节点表示学习为中心，如何进一步学习更具解释性的社区级表示^[91]，以及如何统一刻画层次化社区结构^[92]，仍有待深入研究。其次，在数据形态层面，未来研究需要超越简单图建模范式，进一步面向多层网络 (Multi-layer Networks)^[93]、异构信息网络 (HINs)^[75] 以及属性多视图网络 (Attributed Multi-view Networks)^[94-95] 等复杂场景开展研究。尤其值得注意的是，许多现有方法仍隐含依赖同质性假设，而如何有效建模异质性 (Heterophily) 网络^[96]，已成为提升

模型泛化能力与现实适用性的关键问题。

最后，在系统与应用层面，可扩展性与动态性仍是深度社区发现方法走向实际应用的主要瓶颈。一方面，如何将现有模型有效扩展至数十亿节点的大规模网络^[72]，在计算复杂度、存储开销与训练效率之间取得平衡，仍然具有明显挑战；另一方面，对于动态网络^[97]，如何高效捕捉社区结构随时间的演化规律，而非依赖代价高昂的重复训练，也是未来研究的重要方向。

2.3 社区搜索

社区搜索是图社区挖掘中的一个重要分支，但其问题设定与经典的全局社区发现存在本质区别。社区发现通常面向整个图 G ，试图依据某种全局准则对所有节点进行划分；而社区搜索则是一个查询驱动的局部分析任务：给定图 G 中的一个或一组查询节点 $Q \subset V$ ，目标是返回一个包含 Q 的局部社区 $\hat{C}_Q \subset V$ ，使其在结构、属性或语义上具有较强的内聚性^[27]。

与全局社区发现相比，社区搜索并不试图解释整个图 G 的全局组织结构，而是围绕查询节点 Q 在其局部邻域内快速生成与当前任务相关的结果。因此，社区搜索更强调局部性、个性化与在线响应能力。这类能力使其在许多实际场景中具有直接应用价值，例如金融风控中的可疑账户团伙挖掘、供应链网络中的核心企业圈识别、社交媒体中的关系圈定位，以及情报分析中的关联对象扩展等^[27]。概括而言，社区搜索的发展主线可以理解为：在查询效率、社区质量与结果灵活性三者之间不断寻求更合理的折中。

2.3.1 传统社区搜索

在表示学习和图神经网络方法引入之前，传统社区搜索主要沿着两条技术路径演进。第一类方法以内聚子图约束 (cohesive subgraph constraints) 为核心，通过预先定义严格的结构性质来刻画社区；第二类方法则以局部扩散与局部聚类 (local diffusion and local clustering) 为基础，从查询节点 Q 出发，通过局部相关性传播与质量评估逐步确定社区边界。前者强调结构可解释性与查询效率，后者则更强调结果的灵活性与对复杂社区形态的适应能力。

基于内聚子图约束的方法 这一类方法的基本思想可以概括为“先定义，再搜索”：首先给出一个可验证的图论约束，以刻画何谓“紧密社区”；然后在所有包含查询节

点集合 Q 的候选社区中，搜索满足该约束的结果，并进一步结合附加准则选出最终社区 $\hat{C}_Q$ 。必要时，还会额外最小化社区直径、控制社区规模，或优化某种紧凑性指标，以消除多个可行解之间的歧义。

其中最基础且最有代表性的模型是 k -core^[14,98]。若某个候选社区 $C_Q \subset V$ 满足

$$\forall v_i \in C_Q, \quad |N(v_i) \cap C_Q| \geq k \quad (2.7)$$

则称 C_Q 为一个 k -core 社区。这里约束的是节点在社区 C_Q 内部的度，而非其在整个图 G 中的度。换言之， k -core 刻画的是社区成员在社区内部至少具有 k 个邻居，从而反映了一种最基本的局部参与强度。该模型的突出优势在于计算高效：全图的核心分解可以预先完成，并支持后续在线查询，因此在大规模图上的应用非常广泛。

但是， k -core 对社区内部“紧密性”的刻画仍然偏粗糙。它只要求每个节点拥有足够多的内部邻居，却不进一步约束这些邻居之间是否形成更高阶的闭合结构。为增强这一点，后续研究提出了 k -truss^[15,17]。该模型要求候选社区 C_Q 中的每一条边，在社区内部至少参与 $k-2$ 个三角形。与 k -core 相比， k -truss 不再仅依赖节点局部度，而是通过边的三角支撑度刻画更强的局部团簇性，因此往往能得到更加“抱团”的社区结果。类似地，truss decomposition 及相关索引结构（如 EquiTruss^[16]）也使得这一类方法可以兼顾较强内聚性与在线查询效率。

沿着“提高结构约束强度”的方向，进一步的工作涉及 clique-based 社区模型与连通鲁棒性模型。例如，基于 k -clique 及其推广（如 clique percolation）的方法^[99] 通过要求社区由大量重叠的完全子图构成，刻画极强的局部紧密性；另一类方法则从结构鲁棒性的角度定义社区，例如 k -ECC (k -edge-connected component) 要求社区在删除少量边之后仍保持连通，从而强调结果对边扰动的稳定性^[100-101]。

总体而言，这一路线的优势主要体现在三个方面。其一，定义清晰：社区由明确的图论条件刻画，因此结果具有较强的可解释性。其二，查询高效：很多结构约束可以通过预处理与索引显著降低在线查询开销。其三，结果稳定：在固定参数下，方法往往具有较好的可重复性。然而，这类方法也存在一个根本局限，即结构假设过于刚性。真实网络中的查询相关社区未必总表现为高核度、高三角闭合或高连通鲁棒性的局部结构。一个在语义、功能或行为上高度一致的节点群体，在拓扑结构上可能是稀疏的、弱闭合的，甚至近似树状的。此时，若仍强制采用 k -core、 k -truss 或 clique 等硬约束，则可能导致返回社区规模过小、召回不足，甚至不存在满足约束的可行解。因此，这一类方法以牺牲表达灵活性为代价，换取了较高的可解释性与效率。

基于局部扩散与局部聚类的方法 与基于硬约束的思路不同，这一类方法通常不预设唯一的结构模式，而是从查询节点集合 Q 出发，在图 G 上执行局部扩散，为节点 $v_i \in V$ 赋予相对于 Q 的相关性分数，再据此构造候选社区并评估其质量。其核心思想是：一个优质社区不一定满足某种固定的拓扑模板，但应当体现出“内部联系更紧、外部联系更弱”的局部聚类特征。

这一类方法中最具代表性的质量指标之一是导通率 (conductance)。较小的导通率意味着社区 C_Q 具有较少的外部连接泄漏，同时保持较大的内部凝聚性，因此常被视为高质量局部社区的标志。围绕这一目标，Andersen 等人的一系列工作^[102-104] 证明了：在不访问整个图的情况下，仅通过查询节点附近的局部扩散，即可有效找到导通率较低的局部社区。后续的局部谱方法^[105] 进一步将谱方法局部化，使得围绕查询种子寻找低导通率集合成为传统社区搜索中的另一条重要路线。值得一提的是，局部扩散方法后来还被推广到高阶结构场景。MAPPR^[106] 通过在 motif 加权图上执行近似 PPR，将传统基于 conductance 的局部聚类推广到 motif conductance，从而能够搜索围绕查询种子的高阶局部社区。

为了衡量节点与查询集合 Q 之间的接近程度，这一类方法通常依赖某种扩散接近度 (diffusion proximity)。其中最经典的工具是 Personalized PageRank (PPR)^[24,107] 及 heat kernel diffusion^[108]。分数越大，意味着该节点越容易通过多条局部路径与查询节点产生关联，因此越有可能属于目标社区。在实际计算中，PPR 往往通过近似迭代获得。需要指出的是，这种接近度指标本质上都只是一个对所有节点 $v_i \in V$ 的连续评分，而不是直接给出离散社区 $\hat{C}_Q$ 。因此，经典方法通常还需要一个离散化步骤：先依据某种归一化规则对节点排序，例如按照 PPR 与节点度的归一化比值排序；再对排序后的前缀社区依次计算得到候选集合，最终选出导通率最小的那个候选集合作为输出社区^[102]。从这个意义上说，这一类方法通常由三个环节构成：局部扩散、节点排序以及基于社区选取。

这一路线的主要优势在于灵活性较强。由于它不强制社区满足唯一的结构模式，因此能够更自然地处理边界模糊、结构稀疏或存在重叠的真实社区。在许多场景下，查询相关群体未必表现为高三角闭合的稠密块，而更可能是围绕若干种子逐步衰减扩散形成的局部影响区域；扩散方法恰好适合描述这类现象。此外，连续评分也为“核心成员—边界成员”的渐进式刻画提供了自然工具。

但其局限同样不可忽视。首先，这类方法通常对参数较为敏感。对于 PPR 而言， α 决定了扩散在局部邻域中的停留程度： α 较大时，随机游走更频繁返回个性化向量

s , 结果通常更局部; α 较小时, 扩散更容易向外传播, 返回社区的尺度也可能更大。与此同时, 不同的 PPR 近似算法在实现效率与性能也有明显的分野。类似地, heat kernel 的时间参数、以及候选集合排序过程中所采用的归一化与停止准则, 也都会显著影响社区边界与结果规模^[109]。其次, 这类方法输出的社区更多是通过“评分 + 截断”得到的近似最优解, 因此虽然具有较强经验有效性, 但其结果解释通常不如 k -core 或 k -truss 那样直接对应一个明确的结构定义。最后, 当目标社区并不主要体现为低导通率局部区域, 而更多依赖于属性一致性、异质语义互补性或复杂高阶关系时, 单纯依赖拓扑扩散的方法也可能受到限制。

2.3.2 属性感知与学习驱动社区搜索

上述传统社区搜索方法大多建立在纯拓扑图之上, 即主要依据节点间的连接关系来定义和搜索社区。然而, 现实中的图数据通常并非仅包含拓扑结构。无论是社交网络中的用户画像、科研网络中的研究主题与机构信息, 还是电商网络中的商品类别、价格区间与交互行为, 节点和边往往都携带丰富的属性信息, 甚至还可能属于不同的语义类型。相应地, 用户对社区搜索的要求也不再局限于“结构紧密”, 而是进一步希望搜索结果在语义上具有一致性或任务相关性。换言之, 一个高质量的社区不仅应当在拓扑上相对内聚, 也应当在属性层面呈现出较强的同质性或协同性。

在这一背景下, 属性驱动社区搜索 (Attributed Community Search, ACS) 逐渐形成。与仅依赖拓扑的传统社区搜索不同, ACS 的目标是在给定查询集合 Q 的条件下, 返回一个同时满足结构约束与属性约束的社区 $\hat{C}_Q$ 。从建模角度看, 这类方法通常试图在如下两类目标之间取得平衡: 其一是结构可行性, 例如社区内部具有足够的连接强度、闭合性或鲁棒性; 其二是属性一致性, 即社区成员在关键词、类别、兴趣或语义表征上与查询意图保持较高匹配程度。

早期 ACS 方法大多沿用了传统社区搜索的基本框架, 只是在其中加入属性过滤或属性相似性约束。例如, ACQ^[110] 在 k -core 索引的基础上引入关键词约束, 并通过 CL-tree 索引将结构约束与属性检索结合起来; ATC^[111] 则将 k -truss 与属性相似性联合建模, 使得返回社区既满足较强的三角闭合结构, 又与查询属性保持较高一致性。这类方法的共同特点是: 属性信息虽然被纳入搜索过程, 但其作用通常仍然从属于既有的结构框架, 即作为附加约束、过滤条件或后验排序信号存在。

进一步的工作开始尝试在统一的目标函数中同时刻画结构质量与属性质量, 而不再仅仅将属性视为“结构搜索后的补充条件”。DMCS^[112] 是一个具有代表性的例子。该工作指出, 仅强调社区内部稠密性并不足以刻画高质量结果, 因为一个候选社

区即便内部连接较强,也可能与外部区域存在大量连接,从而导致边界模糊。因此,DMCS 在建模中同时考虑了内部密度与外部分离性,并允许属性信息以加权方式进入目标函数,从而在社区边界辨识方面取得了更好的效果。类似地,一些对传统硬约束进行软化的模型,如 k -plex^[113],也为属性社区搜索提供了更灵活的结构载体。这一阶段的发展表明,社区搜索的目标已经从单纯的“拓扑局部聚团”扩展为“结构—语义协同”的联合优化问题。

不过,无论是早期的属性过滤方法,还是后续将结构与属性联合建模的 ACS 方法,它们大多仍然依赖人工设计的社区度量或约束形式。例如,方法通常需要事先规定何种结构模式应被视为“紧密”,何种属性相似性函数应被视为“相关”,以及二者之间应如何加权折中。这种建模方式虽然具有较强可解释性,但也带来了明显局限:一方面,人工设计的度量往往难以适应复杂多变的真实社区形态;另一方面,当图同时包含拓扑、属性乃至异质语义关系时,不同信息源之间的耦合方式往往难以通过固定规则充分表达。也正因如此,社区搜索逐步走向学习驱动范式,尤其是基于图神经网络的方法。

学习驱动社区搜索的核心思想是:不再显式手工定义“什么是好社区”,而是通过数据学习一个与任务对齐的社区判别函数。从任务形式上看,这一过程可以理解围绕查询节点展开的节点二分类问题,也可以理解为对候选节点进行相关性排序的问题。与传统方法相比,学习范式的优势在于:结构模式、属性相关性乃至更高阶的语义依赖关系都不必被预先硬编码,而可以通过表示学习与任务监督共同确定。

根据训练与推理方式的不同,现有学习驱动社区搜索大致可以分为归纳式方法和直推式方法两类。

归纳式学习 归纳式方法遵循“离线训练、在线推理”的思路。它通常假设可以获得一批历史训练样本,即若干查询集合与其对应目标社区的配对数据,据此离线训练一个可泛化的模型;当新的查询到达时,直接调用预训练模型进行前向传播,输出节点得分并生成社区结果。QD-GNN^[20] 是这一方向的代表性工作。其核心思想是同时编码查询信息与图结构信息,并在图神经网络的消息传递过程中显式注入查询表示,使模型学习到“与查询相关”的节点表征,而非仅学习与图本身相关的通用表征。归纳式方法的主要优势在于在线响应快,适合需要高频查询的场景;但其前提是存在足够丰富且高质量的历史监督数据,而这在社区搜索任务中往往较难满足,因为真实用户查询通常高度个性化,且标准社区标签的获取成本较高。为缓解这一问题,COCLEP^[21] 引入了对比学习等策略,通过构造正负样本对来降低对显式标注的依赖,

从而支持少标签或弱监督场景下的训练。

直推式学习 与归纳式不同，直推式方法不预先训练一个全局可迁移模型，而是在查询到达后，围绕该查询即时构造候选子图 CS_Q ，并在该局部范围内训练或更新一个与当前查询绑定的模型 M_Q 。其出发点是：社区搜索本质上是一个高度个性化的局部任务，因此针对当前查询进行局部学习，往往比依赖全局共享模型更灵活。ICS-GNN^[22] 是该路线的典型代表。该方法构建了一个“候选子图抽取—局部模型训练—社区预测—用户反馈—模型更新”的迭代闭环：初始时利用查询节点及其邻域构造候选范围，在弱监督信号下训练轻量级 GNN；随后将预测结果呈现给用户，并通过用户提供的少量“是/否”反馈修正模型，再迭代更新社区边界。与归纳式方法相比，直推式方法不依赖大规模历史标注数据，且天然适合交互式社区搜索，因此在实际探索式分析场景中具有较强吸引力。

不过，直推式方法也引入了新的挑战。首先，即时训练或反复更新 GNN 会带来不可忽视的在线计算开销，这与社区搜索对快速响应的要求天然存在张力。其次，若模型初始推荐质量不高，则用户需要提供较多交互反馈才能逐步收敛到满意结果，从而带来额外的人机交互成本。围绕这些问题，近期研究开始探索预训练^[114-115]、元学习^[116] 等策略，希望在保持个性化适应能力的同时降低在线优化成本。但就目前而言，这些方向在社区搜索场景中的稳定收益、任务迁移能力以及与查询特定目标之间的对齐程度仍有待进一步验证，因此基于轻量局部建模的直推式框架仍然是一个相对务实且常用的技术选择。

2.3.3 交互式社区搜索

尽管直推式社区搜索方法为查询相关社区的个性化建模提供了新的可能，但其走向实用仍面临两个核心瓶颈：一是用户反馈所带来的标注成本，二是局部模型在线训练与更新所带来的计算成本。前者决定了人机交互是否高效，后者决定了系统能否满足在线社区搜索对低延迟响应的要求。因此，近年来交互式社区搜索的研究重点，逐渐从“如何训练一个更准确的模型”转向“如何在序贯在线交互过程中，以尽可能低的用户与计算代价持续提升社区质量”。

围绕标注成本问题，主动学习 (Active Learning, AL)^[117] 为交互式社区搜索提供了自然的技术路径。其基本思想是：在每一轮交互中，系统不再被动展示当前得分最高的一组候选节点，而是主动选择那些对社区边界最具判别价值的节点请求用户标注。通常 AL 综合考虑模型不确定性、节点对决策边界的影响程度，以及与已标注样

本的冗余性等因素^[118-119]。其核心目标并不是获取尽可能多的标签，而是在固定交互预算 $B \subset V$ 下，使每一次用户反馈都尽可能带来最大的社区质量增益。对于社区搜索而言，这一点尤为重要，因为用户的精力通常有限，而目标社区的修正往往高度依赖少量边界节点的精确信息。

然而，仅仅降低标注成本并不足以支撑实用化的交互式社区搜索。与传统离线学习不同，社区搜索中的交互过程具有显著的在线性和序贯性：查询 Q 到达后，系统需要在极短时间内完成候选社区生成、关键节点选择、用户反馈吸收以及模型更新，并将新的结果立即返回给用户。这意味着在线计算代价本身成为与模型精度同等重要的优化目标。尤其在直推式方法中，若每次交互都需要从头训练或大幅更新图神经网络，则计算开销很容易成为阻碍毫秒级或近实时响应的主要瓶颈。

因此，低延迟交互式社区搜索已经不再是单一算法设计问题，而更接近一个算法与系统协同优化的问题。其核心思路通常是将“精确但昂贵”的学习模块与“高效但相对粗糙”的传统图算法进行分层结合：首先，可利用 k -core、局部扩散或 PPR 近似等传统社区搜索技术^[103] 作为快速候选收缩器，在原图 G 中迅速定位与查询集合 Q 相关的局部区域，从而显著缩小后续学习模块需要处理的搜索空间；随后，仅在压缩后的候选子图上执行局部打分、主动查询和轻量更新。通过这种“粗筛选 + 精修正”的两阶段机制，系统可以在保持社区质量的同时，将在线优化问题限制在一个可控规模内。

从这一视角看，现代交互式社区搜索的关键不再只是设计某一种更强的社区评分函数或图神经网络结构，而是构建一个端到端的人机协同闭环：候选社区快速生成、关键节点主动选择、用户反馈实时吸收、局部模型轻量更新，以及结果持续修正与展示，必须在统一框架下协同工作。最新的交互式框架，如 COMET^[120]，正体现了这一发展趋势。此类方法通常不再将主动学习、模型更新和系统加速视为相互独立的模块，而是将其共同纳入面向在线交互场景的整体设计之中，从而实现“高效候选生成—主动反馈采样—轻量模型更新”的闭环优化。

2.3.4 总结

总体来看，社区搜索的发展经历了从拓扑驱动到属性感知、再到学习驱动与交互优化的逐步演进^[27]。传统方法依靠显式结构约束^[14-15,17,98] 或局部扩散机制^[103]，在查询效率、可解释性和结果稳定性方面具有坚实基础；而面向属性与异质信息的扩展，以及基于图神经网络的学习范式^[20-22]，则进一步提升了模型对复杂社区形态和查询语义的表达能力。与此同时，交互式社区搜索^[22,120]的兴起表明，社区搜索已不再只

是一个静态的局部子图识别问题，而正在转变为一个面向序贯在线场景的人机协同优化问题，其中社区质量、交互成本与系统时延需要被统一考虑。社区搜索研究仍有两个值得进一步推进的方向：其一，如何在弱监督甚至无监督条件下学习更具泛化能力的查询相关社区表示，从而减少对高质量标注和人工反馈的依赖；其二，如何将主动学习、轻量更新与系统级加速更紧密地结合起来，以支持大规模动态图与复杂异质图上的低延迟交互式社区搜索。这些问题的持续推进，也将推动社区搜索从研究型算法进一步走向真实复杂场景中的可部署系统。

2.4 本章小结

本章从图神经网络、社区检测与社区搜索三个层面回顾了相关研究进展。整体而言，图神经网络为图数据上的协同建模提供了统一基础；社区检测的发展展示了从经典结构划分到深度表示学习的范式迁移；而社区搜索则在查询驱动场景下进一步体现出局部性、个性化与在线性的要求，并逐步演进出属性感知、学习驱动和交互优化等研究方向。由此可见，现代社区搜索已不再是单纯依赖拓扑模式匹配的局部检索问题，而是一个需要同时统筹表示能力、查询效率和用户反馈机制的综合建模问题。基于这一认识，本文在后续章节中将面向交互式社区搜索场景，涉及一种兼顾轻量性、灵活性与实用性的直推式社区搜索框架。

第3章 方法论

本章介绍本文提出的交互式社区搜索框架。首先，对社区搜索及交互式社区搜索问题进行形式化定义；其次，给出 *COMET* 的整体框架；最后，详细说明各组成模块的设计与实现。

3.1 问题描述

在复杂图数据中，社区通常表现为内部连接紧密、外部连接相对稀疏的节点子集，是理解网络结构与功能的重要载体。传统的全局社区发现方法旨在对整个图进行一次性划分，从而识别所有潜在社区。然而，在许多实际应用中，用户往往并不关心图的全局组织形式，而更关注与某个特定实体、对象或兴趣相关的局部结构。由此，社区搜索问题应运而生。

设 $G(V, E)$ 表示一个无向图，其中 V 为节点集合， $E \subseteq V \times V$ 为边集合。图 G 的邻接矩阵记为 $A \in \{0, 1\}^{|V| \times |V|}$ ：若节点 v_i 与 v_j 之间存在边，则 $A_{ij} = 1$ ，否则 $A_{ij} = 0$ 。在此基础上，社区搜索可形式化定义如下。

定义 3.1 (社区搜索) 给定图 $G(V, E)$ 和查询节点集合 Q ，其中 Q 由正查询节点 Q^+ 和负查询节点 Q^- 构成。社区搜索旨在找到一个连通的社区 $\hat{C}_Q \subseteq V$ ，使其满足

$$Q^+ \subset \hat{C}_Q, \quad Q^- \cap \hat{C}_Q = \emptyset. \quad (3.1)$$

生成的社区 $\hat{C}_Q$ 应具有较强的内部凝聚性，即社区内部节点连接紧密，而与社区外部节点的连接相对较少。

从形式上看，上述定义对应于图上的一个局部优化问题：在满足查询约束的前提下，从图 G 中寻找一个高质量的节点子集。这里的“高质量”通常由某种社区质量度量来刻画，例如导通率 (conductance) 等^[121]。因此，在非重叠社区设定下，社区搜索往往可以理解为：在查询约束下，寻找使某一社区质量指标最优的局部社区。

然而，标准社区搜索定义隐含了一个较强的理想化假设，即初始查询 Q 已经能够充分表达用户意图，从而唯一确定一个令人满意的目标社区。现实场景中，这一假设往往并不成立。首先，真实网络中的社区普遍存在重叠现象，同一节点可能同时属

于多个语义不同但结构上都较为紧密的局部群体；其次，社区内部还可能包含层次性或更细粒度的局部子结构，使得仅依赖少量查询节点难以准确锁定用户真正关心的目标。

延续图 1.1 中的示例，假设某个“内存”节点同时属于两个紧密相关但语义不同的购买清单：一个面向工作站配置，另一个面向游戏主机配置。这两个局部子图都包含该内存节点，且都具有较高的社区质量，例如较低的导通率。在这种情况下，如果用户只以该内存节点作为初始查询，一个标准社区搜索算法由于缺乏额外上下文，可能返回任意一个局部社区，也可能返回一个混合了两类需求的更大超集。虽然这种结果在结构意义上可能是合理的，但并不一定符合用户的真实意图。

这说明，静态社区搜索的局限并不完全来自社区质量度量本身，而在于其默认用户意图已经被一次性、充分地编码到查询集合 Q 中。当初始查询信息不足时，算法难以区分“结构上合理”与“用户真正需要”的社区。尤其在很多社区搜索方法仅依赖极少量正样本节点的情况下，这一问题会更加突出。为了解决这一困难，用户需要能够在搜索过程中提供额外反馈。例如，当算法返回了偏向“工作站配置”的社区时，用户可以将其中某些不相关节点标记为负样本 Q^- ，或补充某些“游戏主机配置”中的关键节点作为正样本 Q^+ ，从而逐步修正搜索方向。

基于这一观察，交互式社区搜索将用户反馈显式纳入社区搜索过程，通过迭代交互不断校正目标社区，使生成结果更贴近用户真实偏好。其形式化定义如下。

定义 3.2 (交互式社区搜索) 交互式社区搜索是一个由算法与用户共同参与的迭代过程，主要包括以下三个步骤：

- **迭代**：在第 t 轮迭代中，算法基于当前查询集合生成社区 $\hat{C}_Q^t$ ，并给出若干待标记的候选节点；
- **交互**：用户对候选节点进行标记，或手动调整查询集合 Q ，从而更新正、负查询约束；
- **终止**：上述过程持续进行，直到输出社区 $\hat{C}_Q$ 与用户偏好充分一致。

因此，交互式社区搜索本质上是一个 human-in-the-loop 的社区搜索过程，其目标是在用户反馈的指导下逐步逼近目标社区 $\hat{C}_Q \subseteq V$ 。与仅允许用户被动审查搜索结果或仅依赖初始查询节点的方法不同，交互式社区搜索强调一种双向协同机制：模型不仅输出当前社区结果，还主动提供具有判别价值的候选节点供用户评估；而用户的反馈又反过来更新搜索约束并指导后续迭代。基于这一机制，社区搜索不再是一次性的

静态检索，而转化为一个面向用户真实意图的动态优化过程。

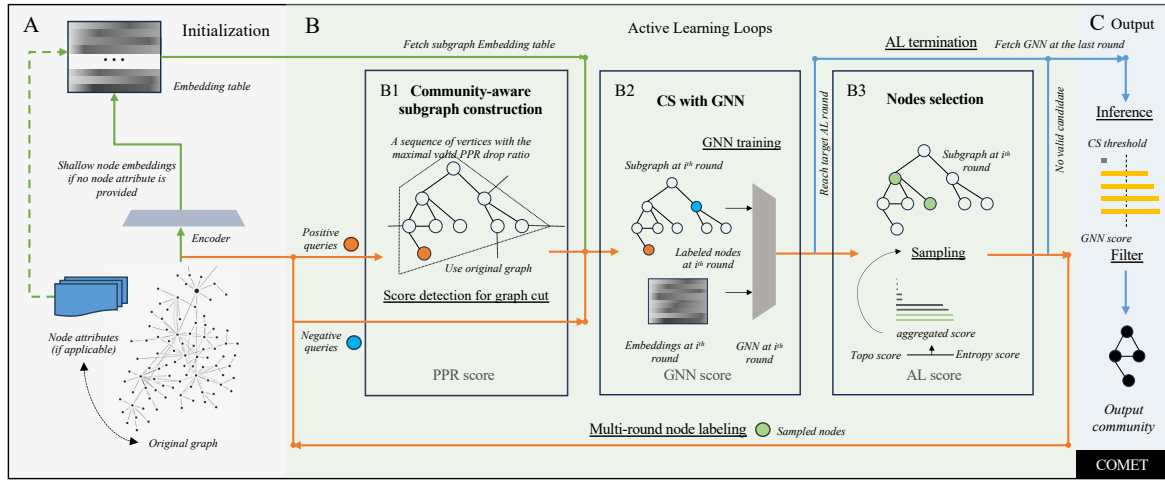


图 3.1 *COMET* 的整体框架

Figure 3.1 Proposed Interactive Community Search Framework: *COMET*

阶段 A 为初始化阶段（绿色箭头），阶段 B 为基于主动学习的交互式社区搜索阶段（橙色箭头），阶段 C 为输出阶段（蓝色箭头）。其中，阶段 B 由社区感知子图构建、基于 GNN 的社区搜索以及节点选择三个核心模块组成。

3.2 框架概述

COMET 的整体框架如图 3.1 所示，由三个阶段构成：初始化阶段、交互式社区搜索阶段以及输出阶段。初始化阶段（A）根据原始图及其可用属性生成节点表示，为后续局部学习提供输入特征；交互式社区搜索阶段（B）是框架的核心，其围绕当前查询集合 Q 迭代执行社区感知子图构建、基于 GNN 的社区搜索以及节点选择三个模块；输出阶段（C）则根据最终一轮模型得分与社区搜索阈值 τ ，生成最终输出社区 $\hat{C}_Q$ 。

其中，模块 B1 负责从原始图中提取与当前查询最相关的候选子图 CS'_Q ，以缩小后续搜索空间；模块 B2 在候选子图上训练局部 GNN 模型 M' ，并为节点输出社区得分 $p(v_i | M')$ ；模块 B3 则从当前候选节点中选择最具判别价值的节点供用户标记，并将反馈结果并入查询集合 Q ，从而驱动下一轮迭代。通过上述三个模块的协同，*COMET* 将局部社区建模、用户反馈吸收与多轮主动学习统一到同一个闭环框架中。

为实现实用的交互式社区搜索，*COMET* 的设计主要围绕两个核心目标展开：其一是在大规模图上实现低延迟的局部社区搜索，其二是在有限标注预算下，通过不确

定性驱动的主动学习尽快修正社区边界。前者要求框架能够在每轮迭代中快速收缩搜索空间，避免在全图范围内重复训练与推理；后者则要求框架能够主动选择最具判别价值的节点，以最小的标注代价持续提升社区搜索结果。基于这两个目标，*COMET* 分别在候选子图构建与节点选择两个环节进行了针对性设计。

3.2.1 面向低延迟的候选子图收缩

交互式社区搜索首先面临的是在线计算代价问题。与传统社区搜索不同，学习驱动的交互式方法需要在每一轮中重复执行候选子图构建、模型训练与结果更新；若这些操作直接在全图上进行，则其计算开销会随图规模迅速增长，从而难以满足交互式场景对低延迟响应的要求。尤其对于直推式方法而言，若每次查询均在大规模图上重新训练 GNN，则模型更新本身就可能成为限制系统实用性的主要瓶颈。

基于这一观察，*COMET* 采用“先收缩、后学习”的局部化策略。具体而言，*COMET* 利用近似 PPR 算法在原始图中快速估计与查询最相关的局部区域，并据此构造社区感知候选子图 CS'_Q 。这一过程能够有效排除大量与当前查询无关的节点，使后续 GNN 仅需在一个规模受控、与目标社区高度相关的局部子图上训练与推理。借助这一候选收缩机制，*COMET* 显著降低了每轮交互的计算负担，并使模型训练更加聚焦于潜在目标社区附近的关键结构区域。

3.2.2 基于不确定性采样的主动学习优化

除计算效率外，交互式社区搜索的另一核心挑战在于如何在有限交互预算下尽快逼近用户真正关心的目标社区。若系统仅被动返回当前预测结果，而不能进一步判断哪些节点最值得用户标注，则用户往往需要提供较多额外反馈，才能逐步修正社区边界。特别是在初始查询较少、社区存在重叠或语义歧义时，这种低效交互会显著削弱学习驱动方法的实际效果。

为此，*COMET* 在每一轮社区预测之后引入主动学习驱动节点选择的模块。该模块不再随机或被动地请求用户标注，而是优先从当前候选子图中选取最具判别价值的节点，以最小化反馈成本并最大化下一轮模型更新所带来的社区质量增益。通过这种方式，用户反馈不再只是对当前结果的事后修正，而成为持续驱动候选子图更新与局部模型改进的关键信号。由此，*COMET* 能够在有限轮次内逐步收敛到更贴合用户偏好的社区结果，并增强整个交互式社区搜索过程的有效性。

3.3 社区感知的子图构建

候选子图构建的目标,是在原始大图 $G(V, E)$ 中快速提取一个与当前查询高度相关的局部子图 CS'_Q , 使其既尽可能覆盖目标社区, 又尽可能排除大量无关节点。该步骤在 *COMET* 中具有两方面作用: 一方面, 它将后续 GNN 的训练与推理限制在一个规模受控的局部区域内, 从而显著降低在线计算开销; 另一方面, 它也为后续社区预测提供了更聚焦的候选空间, 有助于减弱全图噪声结构对局部社区搜索的干扰。

从社区搜索的视角看, 一个理想的候选子图应当满足两点: 其一, 包含查询节点及其最相关的局部邻域; 其二, 在结构上尽可能对应一个内部连接紧密、对外连接较少的局部区域。直接在图中搜索满足这一条件的最优子图通常代价高昂, 因为这等价于在指数规模的候选集合中寻找最优切分。因此, 实际方法通常依赖某种能够刻画“查询相关性”的局部启发式指标。本文采用 Personalized PageRank (PPR) 作为这一阶段的核心工具。

3.3.1 PPR 及其高效近似

PPR 已被广泛证明是构建查询相关局部子图的有效工具。其核心作用在于度量图中节点相对于给定查询的相关性: 分值越高, 说明该节点越可能位于查询节点附近, 或与查询所在社区具有更强的结构关联。基于这一性质, PPR 可以作为社区感知候选子图构建的基础排序信号。

定义 3.3 (Personalized PageRank, PPR) PPR 定义为:

$$\text{pr}(\alpha, s) = \alpha s + (1 - \alpha)W \text{pr}(\alpha, s) \quad (3.2)$$

此处, $\text{pr}(\alpha, s)$ 是 PageRank 向量, $W = A^\top D^{-1}$ 是归一化的转移矩阵, α 是瞬移概率 (teleport probability), s 是个性化向量。指示函数 $\chi_S(v)$ 通常用于 PPR 中的个性化向量 (指示节点 v 是否在查询集 S 中)。

在这一设定下, PPR 可以理解作为一种带重启的随机游走稳态分布: 游走过程从查询分布 s 出发, 每一步以概率 α 返回查询节点, 以概率 $1 - \alpha$ 沿图边继续传播。因此, PPR 自然兼顾了局部可达性与多路径支持, 能够较好地刻画“与查询社区相关”的节点区域。

幂迭代 最传统且最直接的近似方法是幂迭代 (Power iteration)。该方法从一个初始近似值 $\text{apr}^{(0)}(\alpha, s) = s$ 开始, 然后使用以下公式进行递归更新:

$$\text{apr}^{(n+1)}(\alpha, s) = \alpha s + (1 - \alpha)W \text{apr}^{(n)}(\alpha, s) \quad (3.3)$$

此方法在概念上等同于模拟一个从 s 出发、每步以 α 概率重启的截断随机游走 (truncated random walk)。其收敛性是保证的, 当给定一个 L_∞ 误差界限 ϵ , 即

$$\|\text{apr}^{(n+1)}(\alpha, s) - \text{apr}^{(n)}(\alpha, s)\|_\infty < \epsilon \quad (3.4)$$

该迭代方法的计算复杂度被证明为 $O\left(\frac{m}{\alpha} \log \frac{1}{\epsilon}\right)^{[107]}$ 。一个简单的证明思路就是因为 W 的特征值为 $\lambda_1 = 1, 1 \geq |\lambda_2| \geq \dots \geq |\lambda_n| \geq 0$, 而收敛速度由 $\frac{2}{\lambda_1}$ 控制。

然而, 幂迭代的根本缺陷在于其全局性: 无论 PPR 向量 $\text{pr}(\alpha, s)$ 的分布多么稀疏或局部化, 幂迭代的每一步都需要 $O(m)$ 的计算量, 即访问图中的所有边。这在 α 较小 (导致 PPR 分值高度集中在 s 的邻域内) 或图规模极大时, 造成了较大的计算浪费。

局部算法 为了克服全局迭代的低效, 研究界转向了“局部”算法, 其计算复杂度仅取决于解的稀疏度, 而与图的整体规模 n 或 m 无关。

Andersen et al. [102] 提出的 FORWARD PUSH 是这一领域的开创性工作。该方法不再维护一个完整的 n 维向量, 而是巧妙地维护两个稀疏向量: 一个“解”向量 $p = \text{apr}(\alpha, s)$ 和一个“残差”向量 r (待分配的概率质量)。

算法从 $p = 0$ 和 $r = s$ 开始, 并始终保持如下不变量:

$$\text{pr}(\alpha, s) = p + \text{pr}(\alpha, r) \quad (3.5)$$

即真实的 PPR 向量等于已求解部分加上残差部分的贡献, 这里的 p 一定是原 $\text{pr}(\alpha, s)$ 的下估计 (underestimate)。算法的核心是 (Push) 操作: 算法每次选择一个残差过大的节点 u , 即满足 $r(u) > \epsilon \cdot d(u)$ 的节点 (其中 $d(u)$ 是节点 u 的出度, ϵ 是设定的精度阈值), 并执行以下更新

1. 更新: 将 u 上的部分残差 ($\alpha \cdot r(u)$) 加入到解向量 $p(u)$ 中, 这代表随机游走在 u 处传送 (teleport) 并终止。
2. 扩散: 将其余残差 ($(1 - \alpha) \cdot r(u)$) 按照转移概率 (即根据 $d(u)$) 分配给 u 的所有出边邻居, 加入到它们各自的残差项中。

3. 清空 $r(u)$ 。

算法持续进行 push 操作，直到所有节点的残差都满足 $r(u) \leq \epsilon \cdot d(u)$ 。Andersen et al. [102] 证明，该算法在度归一化的情况下计算复杂度为 $O(1/(\alpha\epsilon))$ 。这一复杂度是 PPR 计算领域的一个里程碑，因为它完全独立于图的规模 n 和 m 。它表明，可以在仅与精度要求 (ϵ) 和传送概率 (α) 相关的恒定时间内，计算出近似值 $\text{apr}(\alpha, s)$ 的估计。

基于 Forward Push 的思想，后续研究进一步发展了多种变体，例如用于高效计算单目标 PPR (single-target PPR) 的反向推进算法 (Backward Push)^[108,122]，以及结合前向和反向搜索的各种混合算法，如这篇近期的综述^[123]所述。

3.3.2 PPR 作为子图提取的实践依据

PPR 之所以适合用于候选子图提取，不仅因为它能够为节点提供查询相关性排序，还因为经典结果表明：若图中存在与查询相关的低导通率局部社区，则基于 PPR 排序的 sweep 过程有机会恢复这一局部切分^[102-103]。因此，PPR 实际上为“从查询节点出发截取一个社区感知局部子图”提供了重要理论依据。

为便于讨论，给定社区 $C_Q \subseteq V$ ，其导通率定义如下。

定义 3.4 (图导通度 Conductance) 给定一个图 G 和一个社区 C_Q ， C_Q 的图导通度定义为：

$$\phi(C_Q) = \frac{|\partial(C_Q)|}{\min(\text{Vol}(C_Q), 2m - \text{Vol}(C_Q))} \quad (3.6)$$

其中 $|\partial(C_Q)|$ 是连接社区 C_Q 中的节点与 C_Q 之外节点的边数， $\text{Vol}(C_Q)$ 是 C_Q 的体积，计算为其节点的总度数。

经典的 PPR 局部聚类方法通常先根据度归一化的 PPR 分值对节点排序，再通过 sweep 在所有前缀集合中搜索低导通率切分。尽管这种方法具有较强理论保证^[102-103]，但在本文场景下，COMET 并不直接将 sweep-cut 作为最终候选子图构建策略，而是进一步采用一种更轻量的启发式截取方法。其原因在于：对交互式社区搜索而言，需要的并不是一个严格意义上的最优局部切分，而是一个能够高效覆盖目标社区、且规模适中的候选子图 CS'_Q 。

基于对 PPR 排序序列的观察，发现当节点按 PPR 分值降序排列时，目标社区内部节点通常形成一个相对平滑的高分前缀；而一旦越过社区边界，PPR 分值往往会呈现较明显的下降。因此，排序后的 PPR 序列中存在显著落差的位置，可以作为识

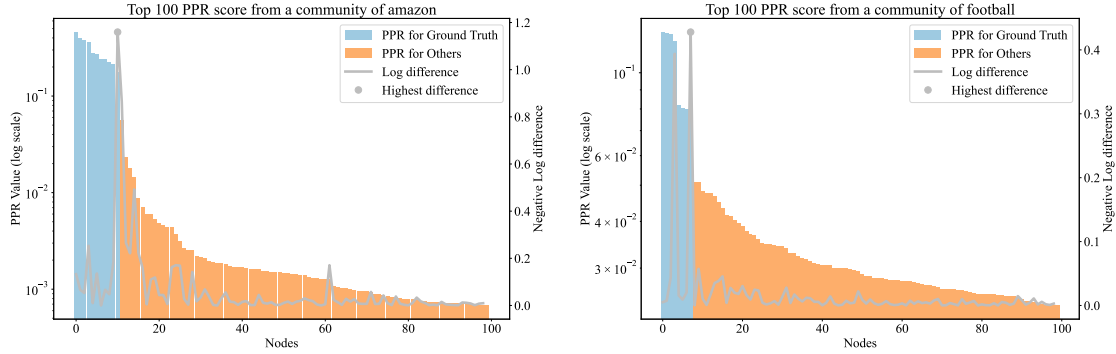


图 3.2 Amazon 数据集中的社区其对应的 PPR 与之负对数差异

Figure 3.2 PPR and Corresponding Negative Log Differences of a Community on Amazon Datasets

Amazon (左) 和 Football (右) 数据集上一个社区的 PPR 及相应的负对数差 PPR 代表了社区中每个节点的重要性, 而负对数差则突显了显著下降发生的位置, 指示了内聚子图的潜在边界。

算法 3.1 基于 PPR 的社区感知候选子图构建

Input: 无向图 $G(V, E)$, 正查询集 Q^+ , 瞬移概率 α , 收敛阈值 ϵ , 最大子图大小 $k_{\max}$, 最小子图大小 $k_{\min}$ 。

Output: 候选子图 CS_Q^t 。

- 1 使用收敛阈值 ϵ 计算 $\text{apr}(\alpha, \chi_{Q^+})$
- 2 $\text{apr}(\alpha, \chi_{Q^+}) \leftarrow$ 将 $\text{apr}(\alpha, \chi_{Q^+})$ 按降序排序
- 3 $\text{apr}(\alpha, \chi_{Q^+}) \leftarrow$ 删除尾随零和所有索引大于 $k_{\max}$ 的元素
- 4 $\text{apr}_r(\alpha, \chi_{Q^+}) \leftarrow \text{apr}(\alpha, \chi_{Q^+})$ 的负对数
- 5 $\text{apr}_r^{\text{FO}}(\alpha, \chi_{Q^+}) \leftarrow \text{apr}_r(\alpha, \chi_{Q^+})$ 的一阶差分
- 6 $\ell \leftarrow \text{apr}_r^{\text{FO}}(\alpha, \chi_{Q^+})$ 中大于 $k_{\min}$ 的最大元素的索引
- 7 $CS_Q^t \leftarrow \{v_0, v_1, \dots, v_{\ell-1}\}$ 来自 $\text{apr}(\alpha, \chi_{Q^+})$
- 8 **return** CS_Q^t

别候选子图边界的有效启发式信号。图 3.2 给出了 Amazon 和 Football 数据集上的两个示例, 其中前 100 个 PPR 分值及其负对数差分清楚展示了这种现象: 最大的下降点通常对应于社区内部节点与外部节点之间的过渡位置, 而较小的下降则可能反映更细粒度的局部划分。

基于这一观察, 本文设计了算法 3.1 来构建候选子图。算法首先计算正查询集合 Q^+ 的近似 PPR 向量, 并按分值降序排序 (删去尾随零与最大子图约束 $k_{\max}$); 随后, 对排序后的分值序列取负对数并计算一阶差分, 以突出潜在的边界点; 最后, 在满足最小子图规模约束 $k_{\min}$ 的前提下, 选取差分最大的索引位置 ℓ 作为截断点, 并将前 ℓ 个节点作为候选子图 CS_Q^t 。该方法避免了直接枚举和评估大量候选切分, 以极低的

额外代价利用 PPR 排序信息完成局部社区感知截取，因而更适合本文所关注的在线交互式社区搜索场景。

3.4 基于学习算法的在线社区搜索优化

在本文框架中，社区搜索被建模为一个查询依赖的节点判别问题。给定查询集合 $Q = Q^+ \cup Q^-$ ，目标是在图中识别与查询相关的目标社区，并将社区成员判定转化为对节点是否属于该社区的预测。与传统方法依赖手工设计的结构约束或相似度度量不同，本文采用参数化模型从图结构与节点属性中学习查询相关的判别边界，从而以数据驱动方式完成社区搜索。

3.4.1 学习驱动的社区搜索建模

从学习视角看，社区搜索可以形式化为一个二元分类问题。

定义 3.5 (基于学习算法的社区搜索优化) 给定图神经网络模型 M ，其为每个节点 v_i 输出一个社区得分 $p(v_i | M) \in [0, 1]$ ，用于刻画节点属于查询相关社区的概率。于是，社区成员判定可写为如下优化问题：

$$\min_M \mathbb{E}_{C_Q} [L_c(Q) | M(G)] \quad (3.7)$$

其中， C_Q 表示与查询 Q 对应的目标社区， $L_c(Q)$ 表示基于查询集合构造的分类损失。该模块采用二元交叉熵作为监督目标：

$$L_c(Q) = - \sum_{v_i \in Q} (y(v_i | Q) \log p(v_i | M) + (1 - y(v_i | Q)) \log(1 - p(v_i | M))) \quad (3.8)$$

其中， $y(v_i | Q) \in \{0, 1\}$ 表示节点 v_i 相对于当前查询集合 Q 的标签：若 $v_i \in Q^+$ ，则 $y(v_i | Q) = 1$ ；若 $v_i \in Q^-$ ，则 $y(v_i | Q) = 0$ 。

在这一建模下，GNN 的作用是学习一种能够刻画查询依赖结构特征的节点表示，使模型在潜在空间中区分属于目标社区与不属于目标社区的节点。与静态启发式方法相比，这种参数化建模方式不需要显式预设社区边界的具体形式，而是通过训练过程自动学习局部社区模式。

3.4.2 子图局部化建模与社区输出

尽管上述形式给出了学习驱动社区搜索的一般建模方式，但在大规模图上直接对完整图 G 进行训练通常代价过高，而且查询相关社区本身也往往只分布在一个局部区域内。为兼顾在线效率与局部判别能力，本文进一步将该优化过程局部化到候选子图 CS_Q^t 上。也就是说，在第 t 轮迭代中，模型仅在当前候选子图上训练与推理，而不在全图上进行社区判别。

相应地，局部化后的优化目标可写为

$$\min_{M^t} \mathbb{E}_{C_Q} [L_c(Q) | M^t(CS_Q^t)] \quad (3.9)$$

其中， M^t 表示在候选子图 CS_Q^t 上训练的局部 GNN 模型。训练完成后，模型为每个候选节点 $v_i \in CS_Q^t$ 输出社区得分 $p(v_i | M^t)$ ，并依据社区搜索阈值 τ 生成当前轮预测社区：

$$\hat{C}_Q^t = \{v_i | v_i \in CS_Q^t \wedge p(v_i | M^t) > \tau\} \quad (3.10)$$

因此，本文中的学习驱动社区搜索并不是在全图上一次性预测所有节点的社区归属，而是在查询相关的局部候选空间中，通过训练局部模型逐步逼近目标社区边界。这样的局部化建模具有两方面优势：其一，它显著降低了训练与推理时需要处理的图规模，使在线社区搜索在大图上成为可能；其二，由于模型仅在与查询高度相关的局部子图上学习，因此能够更集中地捕获局部结构特征，从而提升对查询相关社区的判别能力。

3.5 用于交互式社区搜索的节点选择

在交互式社区搜索中，系统不仅需要预测当前查询相关的社区结构，还需要在用户反馈的作用下逐步修正社区边界。因此，节点选择机制成为连接模型预测与用户交互的关键环节。其核心任务是在有限标注预算下，从当前候选子图中选择最具信息价值的节点提交给用户标记，并利用这些新增标签持续提升后续轮次的社区搜索质量。换言之，节点选择并非简单地扩充查询集合，而是在有限预算约束下，主动决定哪些节点最值得被询问，从而最大化每轮交互对社区边界修正的收益。

3.5.1 主动学习目标与可实现近似

在交互式社区搜索中，节点选择模块的目标是在有限标注预算下，为当前模型选择最具信息价值的节点，并通过用户反馈尽快提升后续社区预测质量。设第 t 轮的查

询集合为 Q ，候选子图为 CS_Q^t ，局部模型为 M^t 。若从未标记节点中选择一个预算集合 B^t 并获得其标签，则新的查询集合将更新为

$$Q' = Q \cup B^t \quad (3.11)$$

理想情况下，交互式社区搜索希望所选节点能够最大程度改善下一轮模型训练与社区预测结果。于是，节点选择的理想目标可写为

$$\arg \min_{B^t \subset V, |B^t|=k} \mathbb{E}_{C_Q} [L_c(Q') | M^{t+1}(CS_{Q'}^{t+1})] \quad (3.12)$$

其中， k 表示本轮标注预算， $CS_{Q'}^{t+1}$ 与 M^{t+1} 分别表示在加入新标注节点后下一轮重新构建的候选子图与重新训练的局部模型。

然而，上述目标在实际中无法直接求解。原因在于，节点集合 B^t 的加入不仅会改变查询集合本身，还会进一步影响下一轮候选子图 $CS_{Q'}^{t+1}$ 的拓扑结构以及局部模型 M^{t+1} 的训练结果。换言之，在当前轮进行节点选择时，模型并不知道不同候选节点将如何改变未来的搜索空间与模型状态，因此无法直接评估其对真实目标损失的精确影响。

基于这一点，本文采用一个可实现的近似策略：节点选择仅依赖当前轮已经可获得的信息，即当前候选子图 CS_Q^t 、当前查询集合 Q 以及当前局部模型 M^t 。

定义 3.6 (基于代理评分的局部节点选择目标) 设第 t 轮中的未标记节点集合为

$$U^t = \{v_i | v_i \in CS_Q^t \wedge v_i \notin Q\} \quad (3.13)$$

给定预算大小 k ，本文将节点选择的局部替代目标定义为

$$B^t = \arg \max_{B \subseteq U^t, |B|=k} \sum_{v_i \in B} \phi(v_i | CS_Q^t, M^t, Q) \quad (3.14)$$

其中， $\phi(v_i | CS_Q^t, M^t, Q)$ 表示节点 v_i 在当前轮中的代理评分。该评分由当前候选子图的结构信息、当前模型输出以及查询约束共同决定，用于近似衡量节点对下一轮社区搜索质量提升的贡献。

因此，本文的节点选择过程可视为对原始主动学习目标的两步松弛。第一层是空间上的局部化近似：节点选择不再在全图 V 上进行，而是被限制在当前轮候选子图 CS_Q^t 的未标记节点集合 U^t 内。这样做的原因在于，候选子图已经由前一阶段的社区

感知构建过程压缩为与当前查询最相关的局部搜索空间，因此在该局部范围内进行采样能够在保持查询相关性的同时显著降低选择与评估代价。第二层是单轮的贪心近似：*COMET* 不再直接优化多轮主动学习过程下的全局最优目标，而是仅利用当前轮可获得的信息选择一组在当前看来最有可能提升后续社区搜索质量的节点。换言之，本文将原始的跨轮次全局优化问题近似为当前轮上的局部贪心优化问题，从而将节点选择转化为当前候选子图上的主动学习采样问题。针对节点采样，本文首先介绍几类基础采样策略，并在此基础上给出本文采用的自适应聚合采样方法。

3.5.2 基础节点采样策略

在上述近似框架下，节点选择的关键在于如何定义代理评分函数 $\phi(v_i | CS_Q^t, M^t, Q)$ 。直观地说，一个理想的候选节点应当同时满足两类性质：其一，对当前模型而言具有较高的不确定性，即其标签一旦被确认，可能显著修正当前社区边界；其二，对当前候选子图而言具有较强的结构重要性，即该节点位于查询相关区域的关键位置。基于这一思路，基础节点采样策略大致可分为不确定性驱动方法与拓扑重要性驱动方法两类。

不确定性采样 (Uncertainty Sampling, US) 不确定性采样聚焦于当前模型预测最模糊的节点。其基本假设是：若一个节点的预测概率接近决策阈值，则该节点更可能位于当前模型尚未清晰分辨的社区边界附近，因此对后续训练最具信息价值。本文采用预测熵来刻画节点的不确定性，记作 $\phi_{\text{ent}}(v_i)$ 。对于任意节点 v_i ，其熵定义为

$$\phi_{\text{ent}}(v_i) = -p(v_i | M^t) \log p(v_i | M^t) - (1 - p(v_i | M^t)) \log(1 - p(v_i | M^t)) \quad (3.15)$$

熵越大，表示模型对该节点的预测越不确定，因此其被优先标注的价值越高。

小批量不确定性采样 (Mini-batched US) 在实际交互式社区搜索中，系统通常每轮向用户推荐多个节点，以减少频繁重训练带来的额外开销^[118-119]。因此，不确定性采样可自然扩展为小批量形式：在未标记节点集合 U^t 中，按照 $\phi_{\text{ent}}(v_i)$ 从高到低排序，并选取得分最高的前 k 个节点作为本轮推荐集合。该方法实现简单，但其缺点在于被选节点可能集中于局部相似区域，导致推荐结果缺乏多样性^[118]。

结合拓扑的采样 仅依赖不确定性得分可能忽略节点在候选子图中的结构作用。因此，这里进一步引入拓扑重要性指标，以衡量节点在查询相关局部区域中的关键程

度。一个直观选择是基于子图内部度定义拓扑得分：

$$\phi_{\text{topo-SD}}(v_i) = \frac{d_i^{\text{sub}}}{\max_{v_j \in CS_Q^t} d_j^{\text{sub}}} \quad (3.16)$$

其中， d_i^{sub} 表示节点 v_i 在候选子图 CS_Q^t 中的度数。若考虑原始图中的全局结构，也可相应使用 d_i^G 替代 d_i^{sub} 。

然而，在本文场景中，一个更合适的拓扑得分来自预先计算的近似 PPR。由于 $\text{apr}(\alpha, \chi_{Q^+})$ 同时刻画了节点与查询正样本的相关性以及其在局部子图中的结构位置，因此可更自然地反映“查询相关的结构重要性”。对应地，定义基于 PPR 的拓扑得分为

$$\phi_{\text{topo}}(v_i) = \frac{\text{apr}(v_i \mid \alpha, \chi_{Q^+})}{\max_{v_j \in CS_Q^t} \text{apr}(v_j \mid \alpha, \chi_{Q^+})} \quad (3.17)$$

在后续方法中，采用 ϕ_{topo} 作为拓扑重要性评分。

3.5.3 自适应聚合采样

单独依赖不确定性采样虽然能够优先选择当前模型最模糊的节点，但容易忽略节点在候选子图中的结构重要性；相反，单独依赖拓扑得分虽然能够强调查询相关区域中的关键节点，却可能无法充分反映当前模型最需要被修正的判别边界。为在这两类信息之间取得平衡，本文提出一种基于 Log-Sum-Exp (LSE) 的自适应聚合采样方法，将拓扑得分 ϕ_{topo} 与不确定性得分 ϕ_{ent} 在不同主动学习轮次中进行动态融合。

其基本思想是：在主动学习初期，查询集合较小，模型尚未建立稳定的社区边界，此时更应强调拓扑先验，以优先获取位于查询相关区域中的关键节点；而随着交互轮次推进，模型逐渐积累了更多标注信息，不确定性得分对边界修正的作用会更加突出。因此，本文通过轮次相关的动态加权方式，在拓扑信息与不确定性信息之间进行平衡。

定义 3.7 (基于 LSE 的自适应聚合采样) 设主动学习总轮次为 R ，当前轮次为 $t \in \{0, 1, \dots, R-1\}$ 。对于任意未标记节点 $v_i \in U^t$ ，其聚合得分定义为

$$\phi_{\text{aggregated}}(v_i) = \log \left\{ \exp \left[\frac{\phi_{\text{topo-norm}}(v_i)}{t+1} \right] + \exp \left[\frac{\phi_{\text{ent-norm}}(v_i)}{R-t} \right] \right\} \quad (3.18)$$

其中， $\phi_{\text{topo-norm}}(v_i)$ 与 $\phi_{\text{ent-norm}}(v_i)$ 分别表示经过归一化后的拓扑得分与不确定性得分。基于该得分，节点选择集合定义为

$$B^t = \arg \max_{B \subseteq U^t, |B|=k} \sum_{v_i \in B} \phi_{\text{aggregated}}(v_i) \quad (3.19)$$

由于 ϕ_{topo} 与 ϕ_{ent} 的数值尺度通常不同，直接相加可能导致某一项主导最终结果。因此，这里采用最小-最大归一化，将二者分别缩放到 $[0, 1]$ 区间：

$$\phi_{\text{topo-norm}} = \frac{\phi_{\text{topo}} - \min(\phi_{\text{topo}})}{\max(\phi_{\text{topo}}) - \min(\phi_{\text{topo}})} \quad (3.20)$$

$$\phi_{\text{ent-norm}} = \frac{\phi_{\text{ent}} - \min(\phi_{\text{ent}})}{\max(\phi_{\text{ent}}) - \min(\phi_{\text{ent}})} \quad (3.21)$$

在实现上，本模块采用贪婪排序策略来近似求解上述优化问题：首先为所有 $v_i \in U^t$ 计算聚合得分 $\phi_{\text{aggregated}}(v_i)$ ，再按得分降序选取前 k 个节点作为本轮推荐集合 B^t 。这样做的好处在于，框架可以随着主动学习轮次动态调整采样重心：在早期更强调与查询相关的结构关键节点，在后期则更强调当前模型最不确定、最值得进一步确认的边界节点。动态节点选择的整体过程如算法 3.2 所示。

算法 3.2 动态节点选择

Input: 当前轮次 t ，主动学习总轮次 R ，未标记节点 U ，样本大小 k

Output: 选中的节点 B^t

- 1 为 $v_i \in U$ 计算 $\phi_{\text{topo}}(v_i), \phi_{\text{ent}}(v_i)$
 - 2 $\phi_{\text{topo-norm}}(v_i), \phi_{\text{ent-norm}}(v_i) \leftarrow$ 归一化 $\phi_{\text{topo}}(v_i), \phi_{\text{ent}}(v_i)$
 - 3 **foreach** $v_i \in U$ **do**
 - 4 $\phi_{\text{aggregated}} \leftarrow \log \left\{ \exp \left[\frac{\phi_{\text{topo-norm}}(v_i)}{t+1} \right] + \exp \left[\frac{\phi_{\text{ent-norm}}(v_i)}{R-t} \right] \right\}$
 - 5 $B^t \leftarrow$ 从 U 中按 $\phi_{\text{aggregated}}$ 降序排列选出前 k 个节点
 - 6 **return** B^t
-

3.6 COMET 的整体流程

本节结合算法 3.3，说明 COMET 的整体执行流程。

在初始化阶段，若原图未提供节点属性，则首先基于图结构生成浅层节点表示，并将其作为后续模型的输入特征（算法第2行）。本文采用浅层嵌入方法（如 Node2vec^[124]）完成该步骤。这样做的目的，是为后续局部 GNN 提供统一的初始表征，从而使框架能够同时适用于有属性图与无属性图两种场景。

随后，框架进入主动学习驱动的迭代过程。在第 t 轮中，算法首先仅利用当前查询集合中的正样本节点，执行社区感知候选子图构建，得到当前轮候选子图 CS'_Q （算法第4行）。在此基础上，框架初始化局部 GNN 模型 M^t ，并在候选子图 CS'_Q 上进行训练（算法第6–8行），从而得到节点相对于当前查询的社区得分 $p(v_i | M^t)$ 。训练达到预设轮次 T 或满足提前停止条件后终止。随后，模型依据社区搜索阈值 τ 生成当前轮预测社区，对应算法第9行。

在进入节点选择模块之前，框架首先执行终止判定。具体而言，若当前候选子图 CS'_Q 与预测社区 $\hat{C}_Q$ 的相似度超过阈值 β_1 ，说明当前候选空间与模型输出已经较为一致；若当前正查询集合 Q^+ 与预测社区 $\hat{C}_Q$ 的相似度超过阈值 β_2 ，说明当前预测结果已较好覆盖用户显式确认的正样本约束；此外，当迭代轮次达到上限 R 时，算法同样终止（算法第10行）。这一终止机制用于判断当前搜索空间或当前查询约束是否已足以稳定刻画目标社区，从而避免不必要的后续训练与交互。

若未满足终止条件，框架则继续执行主动学习循环。首先，从当前候选子图中收集未标记节点集合，对应算法第12行；随后，根据未标记集合的规模动态调整当前轮采样数 k （算法第13行）。在此基础上，节点选择模块从 U 中选取最具判别价值的节点组成 B^t （算法第14行），并交由用户进行标注。获得新标签后，框架将 B^t 并入查询集合 Q ，从而进入下一轮候选子图构建与局部模型训练过程。通过这一机制，COMET 将社区预测、不确定性采样与用户反馈更新整合为一个持续迭代的闭环过程。当迭代终止时，框架返回最终输出社区 $\hat{C}_Q$ 。

因此，从整体上看，COMET 的执行逻辑可以概括为：先通过社区感知候选收缩将全图搜索问题局部化，再通过局部 GNN 完成当前轮社区估计，并借助主动学习不断引入高价值标注，从而在有限交互预算下逐步逼近目标社区。

3.7 本章小结

本章详细阐述了交互式社区搜索的形式化定义及其解决方案 COMET 框架。首先明确了社区搜索的局部优化目标与交互机制；随后，深入介绍了框架的三个核心组成部分：利用近似 PPR 实现高效剪枝的社区感知子图构建模块、基于图神经网络

算法 3.3 *COMET* 框架

Input: 图 $G(V, E)$, 查询集 Q , 特征矩阵 X (如果可用), 邻接矩阵 A , 训练周期 T , 主动学习轮次 R , 样本大小 k , 相似度阈值 β_1, β_2 , CS 阈值 τ

Output: 输出社区 $\hat{C}_Q$

```

1 if  $X$  不可用 then
2    $X \leftarrow$  从  $G(V, E)$  生成图嵌入
3 while 当前轮次  $t = 0$  to  $R$  do
4    $CS_Q^t \leftarrow$  基于  $Q$  的正样本运行算法 3.1
5   初始化 GNN 模型  $M^t$ 
6   for  $epoch = 1$  to  $T$  do
7      $p(v_i | M^t) \leftarrow M^t(CS_Q^t)$ 
8      $L_c(Q) \leftarrow -\sum_{i \in Q} (y(v_i | Q) \log(p(v_i | M^t)) + (1 - y(v_i | Q)) \log(1 - p(v_i | M^t)))$ 
9      $\hat{C}_Q = \{v_i | v_i \in CS_Q^t \wedge p(v_i | M^t) > \tau\}$ 
10    if  $\text{sim}(CS_Q^t, \hat{C}_Q) > \beta_1 \vee \text{sim}(CS_Q^t, Q^+) > \beta_2 \vee t = R$  then
11      break
12    未标记样本  $U = \{v_i | v_i \in CS_Q^t \wedge v_i \notin Q\}$ 
13    样本大小  $k \leftarrow \min(k, |U|)$ 
14     $B^t \leftarrow$  基于算法3.2 对  $U$  进行节点选择
15     $Q \leftarrow Q \cup B^t$ 
16 return 输出社区  $\hat{C}_Q$ 

```

的局部社区分类模型，以及融合图拓扑结构与模型不确定性的自适应节点选择策略。通过将高效的子图提取与人在回路的主动学习相结合，本章提出的方法论有效地解决了大规模图中社区搜索在效率与准确性之间的平衡问题，为后续的实验验证奠定了坚实的理论基础。

第4章 实验评估

本章通过系统实验评估 *COMET* 在真实世界图数据上的表现。实验的核心目标有二：一方面，验证 *COMET* 在不同规模、不同结构复杂度的数据集上是否能够同时兼顾社区搜索质量与在线交互效率；另一方面，分析框架内部各组件的作用及其在不同实现配置下的稳定性与适应性。为此，本文在八个真实世界数据集上开展实验，并围绕以下两个研究问题展开分析。

关键问题一 *COMET* 在不同规模与复杂度的数据集上，能否同时保持较好的准确性与效率？随着交互轮次增加，多轮交互式社区搜索能够带来多大程度的性能提升？

关键问题二 *COMET* 的各组成模块是否具有足够的鲁棒性，使其在候选子图构建方法、GNN 架构以及节点选择策略的不同配置下，仍能保持相对稳定的性能？这些不同配置又将如何影响框架的整体适应性？

为回答上述问题，下面首先介绍实验所使用的数据集与基本设置，随后分别从性能评估、消融实验与敏感性分析以及合成图模拟等方面展开分析。

4.1 数据集

为全面评估本文方法，本文选用了八个真实世界图数据集，并按照规模与结构特征将其划分为两类。第一类为四个规模较小的无属性社区网络，主要用于观察方法在中小规模图上的行为细节与交互收益；第二类为四个来自 SNAP 存储库^[125]的大规模无属性社区网络，主要用于检验方法在复杂图上的可扩展性与鲁棒性。

第一类数据集包括 Karate、Dolphins、Football 和 EU-Core。Karate 数据集^[126]来源于经典的空手道俱乐部社交网络研究；Dolphins 数据集^[127]描述了宽吻海豚之间的社会关系；Football 数据集^[126]表示单赛季美国大学橄榄球比赛网络；EU-Core 数据集^[128]则记录了一个欧洲研究机构内部的电子邮件通信关系。第二类数据集包括 Amazon、DBLP、Youtube 和 LiveJournal。它们分别对应产品共购网络、学术合著网络、在线视频社交网络和博客社交网络，图规模显著更大，也包含更复杂的社区结构。

这些数据集在图规模、社区数量、社区大小分布、局部密度、结构混合程度以及是否存在重叠社区等方面具有明显差异，因此能够较好覆盖本文关注的多种实验场

景。

表 4.1 数据集的描述性统计
Table 4.1 Descriptive Statistics of Datasets

指标 ¹	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
n	34	62	115	1,005	334,863	317,080	1,134,890	3,997,962
m	78	159	613	16,706	925,872	1,049,866	2,987,624	34,681,189
数量	2	2	12	42	1,255	4,961	2,250	4,151
大小	17.00 ± 1.00	31.00 ± 11.00	9.58 ± 2.22	27.66 ± 23.39	20.49 ± 24.78	22.57 ± 201.85	29.30 ± 87.90	31.97 ± 62.69
直径	3.0 ± 0.00	5.0 ± 0.00	1.75 ± 0.60	2.91 ± 1.20	3.42 ± 2.45	2.08 ± 1.81	3.59 ± 1.65	2.52 ± 1.30
连通分量	1.00 ± 0.00	1.00 ± 0.00	1.42 ± 0.86	3.69 ± 3.15	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
$\rho_{\text{local}}(C_Q)$	0.25 ± 0.02	0.18 ± 0.05	0.73 ± 0.25	0.45 ± 0.25	0.53 ± 0.27	0.72 ± 0.27	0.29 ± 0.19	0.76 ± 0.21
$\bar{d}^{\text{sub}}(C_Q)$	4.01 ± 0.12	4.77 ± 0.47	6.43 ± 2.31	9.15 ± 5.51	5.59 ± 1.37	5.33 ± 1.84	3.06 ± 1.51	23.32 ± 33.95
$\bar{d}^G(C_Q)$	4.60 ± 0.15	4.99 ± 0.39	10.57 ± 0.63	37.51 ± 22.19	5.92 ± 1.46	10.07 ± 5.12	63.98 ± 63.65	45.22 ± 61.48
$\mu(C_Q)$	0.10 ± 0.02	0.04 ± 0.01	0.40 ± 0.21	0.60 ± 0.19	0.05 ± 0.03	0.24 ± 0.15	0.72 ± 0.18	0.34 ± 0.19
$C_{\text{local}}(C_Q)$	0.69 ± 0.03	0.30 ± 0.06	0.75 ± 0.26	0.55 ± 0.27	0.71 ± 0.15	0.91 ± 0.10	0.33 ± 0.21	0.88 ± 0.10
重叠	✗	✗	✗	✗	✓	✓	✓	✓

1. 指标描述如下: 数量统计有效社区。大小是每个社区的平均节点数。直径是平均社区直径。连通分量是每个社区的平均连通分量数。 $\rho_{\text{local}}(C_Q)$ 是局部密度。 $\bar{d}^{\text{sub}}(C_Q)$, $\bar{d}^G(C_Q)$ 是基于社区 c 和图 G 度数的平均社区度。 $\mu(C_Q)$ 是社区混合水平。 $C_{\text{local}}(C_Q)$ 是聚类系数。重叠检查是否存在重叠社区。

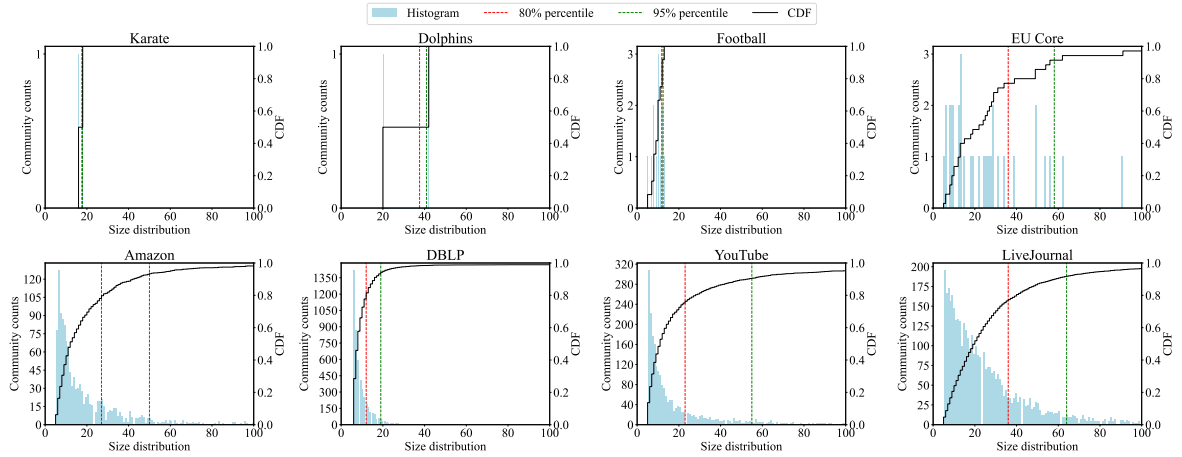


图 4.1 社区大小的分布
Figure 4.1 Distribution of Community Sizes

图 4.1 展示了各数据集中社区大小的分布。可以看到, 尽管八个数据集在图规模上差异显著, 但大多数社区仍然集中在相对较小的规模范围内, 平均社区大小大约为 30 个节点左右。这与一些大型图分类数据集形成对比, 在那些数据集中, 单个类别可以轻易超过数千个节点或更多。

这一现象对于本文任务具有重要意义: 一方面, 它说明真实社区搜索问题通常具有较强的局部性, 目标社区往往仅占据原图中的很小一部分; 另一方面, 它也表明, 在大规模图上直接进行全图级训练或搜索通常存在明显冗余, 因此采用候选子图收缩是合理且必要的。

为便于理解表 4.1 中的统计量，这里对其中涉及的几个社区结构指标作统一说明。给定查询 Q 对应的真实社区 $C_Q \subseteq V$ ，记其在原图 G 上诱导的边集为

$$E_{C_Q} = \{(u, v) \in E \mid u \in C_Q, v \in C_Q\} \quad (4.1)$$

局部密度定义为

$$\rho_{\text{local}}(C_Q) = \frac{2|E_{C_Q}|}{|C_Q|(|C_Q| - 1)} \quad (4.2)$$

用于衡量社区内部的稠密程度。基于社区子图和原图的平均度分别定义为

$$\bar{d}^{\text{sub}}(C_Q) = \frac{\sum_{v \in C_Q} d_v^{\text{sub}}}{|C_Q|}, \quad \bar{d}^G(C_Q) = \frac{\sum_{v \in C_Q} d_v^G}{|C_Q|} \quad (4.3)$$

其中， d_v^{sub} 和 d_v^G 分别表示节点 v 在社区内部和原图中的度。社区混合水平定义为

$$\mu(C_Q) = \frac{|\{(u, v) \in E \mid u \in C_Q, v \notin C_Q\}|}{|E_{C_Q}| + |\{(u, v) \in E \mid u \in C_Q, v \notin C_Q\}|} \quad (4.4)$$

用于刻画社区与外部节点之间的连接强度。局部聚类系数定义为

$$C_{\text{local}}(C_Q) = \frac{1}{|C_Q|} \sum_{v \in C_Q} \frac{2T_v}{d_v^{\text{sub}}(d_v^{\text{sub}} - 1)} \quad (4.5)$$

其中， T_v 表示社区 C_Q 中经过节点 v 的三角形数量。

表 4.1 进一步展示了数据集在局部密度、平均社区度、社区混合水平和聚类系数等方面的统计差异。这些统计量从不同角度反映了社区结构的复杂程度。例如，*Karate* 与 *Dolphins* 的局部密度相对较低，而 *DBLP* 与 *LiveJournal* 则呈现出更高的局部聚团性；*Youtube* 的社区混合水平较高，说明其社区边界更模糊、外部连接更复杂；相对地，*Amazon* 的混合水平较低，表明其社区划分更清晰。

4.2 实验设置

本节介绍实验中采用的比较方法、*COMET* 的实现细节以及实验运行环境。具体而言，本文将比较方法划分为传统社区搜索方法和基于学习的社区搜索方法两类，并在统一的数据预处理流程与查询设置下对各方法进行评估。

4.2.1 基线方法

为了全面评估 *COMET* 的有效性，本文选择了六种具有代表性的社区搜索方法作为基线，分别涵盖传统结构方法与近年的学习驱动方法。

传统方法包括 k -core、 k -truss 以及基于模块度的 Louvain 方法。这三类方法均属于文献中常见的结构驱动社区搜索模型，能够为本文方法提供具有代表性的非学习基准。其中， k -core 与 k -truss 方法分别基于对应的 k -内聚结构搜索包含查询节点的局部子图，并返回满足最高 k 值的社区结果；Louvain 方法则基于模块度优化，并在实验中采用分辨率参数 0.1。需要说明的是，这类传统方法通常只能处理单个正查询节点，因此在对应实验中每个查询仅包含一个正样本。

基于学习的方法包括 ICS-GNN、QD-GNN 和 COCLEP。三者中，ICS-GNN 属于直推式社区搜索方法，而 QD-GNN 与 COCLEP 属于归纳式方法。对于 ICS-GNN，本文总体遵循其原始默认设置，但为了与本文实验场景保持一致，将训练与测试所允许的最小社区规模从原文中的 30 个节点调整为 5 个节点。对于 QD-GNN，由于该方法面向属性图设计，而本文所使用的数据集不提供预定义属性，因此本文采用与 COMET 相同的 Node2vec 浅层嵌入作为其输入特征。考虑到 QD-GNN 在训练阶段可以访问完整社区信息，本文进一步采用三种设置考察其样本效率与分布偏移适应能力：E50、I50 和 E200。具体而言，E50/E200 表示将社区集合划分为两部分，使用前一部分中的 50/200 个训练样本进行训练，并在后一部分上采样测试查询；I50 表示训练样本数为 50，且训练与测试查询来自相同分布。对于 COCLEP，本文在小规模图 (Karate、Dolphins、Football 和 EU-Core) 上将分区数设置为 2，并将参数 α 与 λ 均设为 0.2；对于其余数据集，则采用原论文中的默认设置。根据 COCLEP 的设定，在无属性图场景下，其节点输入表征使用归一化核心度生成。

4.2.2 COMET 实现

数据预处理 本文过滤掉了所有少于五个节点的社区。对于来自 SNAP 的四个大型数据集，使用排名前 5000 的社区作为真实社区。此外，为了确保数据的完整性和唯一性，所有重复的社区都已从 SNAP 数据集中移除。所有数据集均在没有预定义特征矩阵 X 的情况下使用。

实现细节和超参数选择 在节点表示初始化阶段，本文基于邻接矩阵 A 使用 Node2vec 的默认配置为每个节点生成 128 维浅层嵌入。对于候选子图构建模块，近似 PPR 的瞬移概率设置为 0.05，收敛阈值设置为 10^{-4} ，候选子图大小的下界与上界分别设为 5 和 200。

在局部社区搜索模型方面，默认实现采用两层 GCN，其中隐藏维度为 128，隐藏层激活函数为 ReLU，输出层使用 sigmoid。训练过程中采用 Adam 优化器，学习率设

为 0.01，dropout 率设为 0.5，并设置最多 200 个 epoch 的训练上限与早停机制。社区搜索判别阈值 τ 固定为 0.5。

在交互式节点选择阶段，主动学习轮次上限设为 3，终止相似度阈值 (β_1, β_2) 分别设为 (0.9, 1.0)。所有实验默认采用 $\phi_{\text{aggregated}}$ 作为节点选择评分函数。在每一轮交互中，本文假设用户会对系统推荐的节点进行一致且正确的标注，并将其中至多 3 个具有最高聚合得分的新标记节点并入查询集合。为保证实验可复现性，所有实验默认随机种子均设为 42。

平台 *COMET* 基于 Python 3.10 和 PyTorch 1.12 实现。所有实验均运行在配备两颗 2.80GHz Intel(R) Xeon(R) Gold 6342 CPU、512GB DDR4 内存以及两块 GeForce RTX 4090 GPU 的服务器上。GNN 模型基于 PyTorch Geometric 实现。此外，本文采用 cdlib 构建并运行模块度 (Modularity) 相关的基线方法。

对于其他传统图算法与敏感性分析中的比较方法， k -core、 k -truss 以及 Panther 均基于 NetworkX 实现；METIS 采用连接至 METIS 5 的 Python 接口；SimRank 则使用 ExactSim^[129] 提供的实现。

4.2.3 评估指标

为了进行性能评估，本文采用三个常用的指标：归一化互信息 (NMI)、Jaccard 指数和 F1-score 对生成结果质量进行量化。

定义 4.1 (归一化互信息 (Normalized Mutual Information, NMI)) 归一化互信息用于衡量预测结果与真实标签之间的相互依赖关系，其定义如下：

$$\text{NMI} = \frac{2 \cdot I(CS_Q, C_Q)}{H(CS_Q) + H(C_Q)} \quad (4.6)$$

其中， $I(\cdot, \cdot)$ 表示互信息， $H(\cdot)$ 表示熵。NMI 的取值范围为 $[0, 1]$ ，值越大表示预测结果与真实标签越一致。

定义 4.2 (Jaccard 指数) Jaccard 指数反映预测集合与真实集合之间的重叠程度, 其定义如下:

$$\text{Jaccard} = \frac{|CS_Q \cap C_Q|}{|CS_Q \cup C_Q|} \quad (4.7)$$

该指标衡量两集合的相似性, 取值越接近 1 表示匹配越好。

定义 4.3 (F1-score) F1-score 综合考虑了精确率与召回率的平衡性, 其定义如下:

$$F_1 = \frac{2 \cdot |CS_Q \cap C_Q|}{|CS_Q| + |C_Q|} \quad (4.8)$$

F1-score 兼顾正确性与覆盖性, 是衡量整体识别性能的重要指标。在其他性能评估场景中, 本文仅使用 F1 作为核心评价标准。

4.2.4 社区和查询设置

在真实社区 (ground-truth communities) 的选择上, 本文剔除了节点数少于 5 的社区。对于规模较大的数据集, 如 Amazon、DBLP、Youtube 和 LiveJournal, 本文随机选取 200 个真实社区作为评估基准。对于真实社区数量较少的数据集, 本文将所有可用社区纳入评估, 并通过设置不同查询构造测试样本, 使每个数据集上的查询总数均为 200。

在查询设置方面, ICS-GNN 对每个查询使用社区内的 4 个正样本和社区外的 3 个负样本; QD-GNN 采用相同设置, 即 4 个正样本和 3 个负样本; COCLEP 则仅使用 4 个正样本。对于本文提出的 *COMET*, 默认设置 (Default) 与上述监督信息配置保持一致, 即每个查询包含 4 个正样本和 3 个负样本。此外, 在交互式设置 (Interactive) 下, 本文为 *COMET* 设置每轮最多 3 个节点的标注预算, 并最多进行 3 轮节点选择。

所有性能指标均在每个数据集的 200 个查询上取平均值, 并分别报告各基线方法与本文模型的实验结果。

4.3 有效性评估

针对**关键问题一**, 本文通过与多个代表性基线方法进行对比, 系统评估了 *COMET* 的整体有效性。在此基础上, 本文进一步引入多轮交互式社区搜索设置, 以考察 *COMET* 在迭代反馈场景下的性能提升及其交互机制的有效性。

表 4.2 代表性数据集上社区搜索的有效性评估
Table 4.2 Effectiveness Evaluation for Community Search on Representative Datasets

指标 ¹	方法	设置	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
NMI	k -core	-	2.59	0.73	2.51	0.24	44.00	5.29	0.30	1.53
	k -truss	-	8.90	7.50	76.04	0.34	72.21	33.44	2.34	33.35
	Modularity	-	17.80	11.68	57.56	6.90	57.58	55.64	10.49	64.95
	ICS-GNN	-	4.68	40.09	17.71	28.22	47.61	33.08	19.49	46.58
	QD-GNN	E50	100.00	100.00	64.72	13.73	o.o.m. ⁴	o.o.m.	o.o.m.	o.o.m.
		E200	100.00	100.00	69.45	19.75	o.o.m.	o.o.m.	o.o.m.	o.o.m.
		<i>I50³</i>	98.71	93.94	94.10	30.55	o.o.m.	o.o.m.	o.o.m.	o.o.m.
	COCLEP	E50	51.07	42.26	4.12	0.58	0.85	0.37	0.22	1.11
		E200	47.72	31.79	63.05	14.02	80.56	67.73	20.65	63.99
		<i>I50</i>	51.49	33.53	2.52	0.90	0.99	0.34	0.20	1.03
	COMET	默认	78.47	82.10	59.17	34.33	88.59	70.75	41.24	64.92
		交互式	91.82	100.00	78.75	37.59	92.86	79.61	56.24	72.81
Jaccard	k -core	-	38.08	38.99	8.27	2.97	41.36	5.11	0.23	1.39
	k -truss	-	43.88	36.54	78.57	2.93	67.08	31.53	1.92	30.19
	Modularity	-	25.01	16.11	62.30	8.21	50.79	49.60	8.10	58.29
	ICS-GNN	-	52.18	66.65	26.25	30.05	39.74	24.79	13.58	37.51
	QD-GNN	E50	0.00	0.00	70.78	16.91	o.o.m.	o.o.m.	o.o.m.	o.o.m.
		E200	0.00	0.00	73.40	22.21	o.o.m.	o.o.m.	o.o.m.	o.o.m.
		<i>I50</i>	99.45	93.94	95.37	33.13	o.o.m.	o.o.m.	o.o.m.	o.o.m.
	COCLEP	E50	78.26	61.94	11.78	1.42	0.54	0.26	0.16	0.67
		E200	76.60	57.34	69.10	15.24	76.73	61.91	15.07	56.40
		<i>I50</i>	77.48	64.45	10.01	2.96	0.63	0.22	0.15	0.62
	COMET	默认	91.32	92.29	62.65	32.17	85.87	64.02	33.18	57.39
		交互式	96.88	100.00	82.45	33.30	91.42	74.73	48.27	66.58
F1	k -core	-	53.97	54.35	15.21	5.66	45.95	5.45	0.34	1.63
	k -truss	-	60.68	51.44	83.49	5.60	75.77	34.90	2.80	35.24
	Modularity	-	38.71	26.68	72.16	13.05	61.56	60.18	12.73	68.65
	ICS-GNN	-	68.51	79.86	40.97	42.29	53.49	38.66	23.32	51.91
	QD-GNN	E50	0.00	0.00	79.95	27.90	o.o.m.	o.o.m.	o.o.m.	o.o.m.
		E200	0.00	0.00	81.51	34.44	o.o.m.	o.o.m.	o.o.m.	o.o.m.
		<i>I50</i>	99.70	96.71	97.49	46.65	o.o.m.	o.o.m.	o.o.m.	o.o.m.
	COCLEP	E50	86.26	76.06	20.69	2.75	1.06	0.51	0.32	1.31
		E200	85.14	72.44	78.69	23.85	83.35	72.18	24.03	68.12
		<i>I50</i>	85.56	77.70	17.90	5.58	1.23	0.44	0.30	1.22
	COMET	默认	95.45	95.91	70.42	45.06	91.03	75.52	46.92	69.54
		交互式	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76

1. 本章报告了基线方法和 *COMET* 框架的性能结果，所有指标均以百分比形式呈现。
2. 粗体数字表示最高值，而下划线数字表示次高值。
3. I50 设置因信息泄漏到训练集中而被排除在排名之外。所有归纳式 CS 的 I50 性能均以斜体标注。
4. “o.o.m.” 表示模型因超过 GPU 内存限制而无法运行。

4.3.1 性能评估

表 4.2 评估了八个数据集上的 CS 性能。交互式 *COMET* 优于大多数基线方法，而非交互式 *COMET* 与其他算法相比也具有竞争力。表 4.3 展示了经过三轮主动学习后 F1-score 的显著提高。

其他基线方法在特定数据集及其设置上表现出独特的行为。对于传统的 CS 方法， k -core 和 k -truss 在较小的数据集（如 Karate、Dolphins 和 Football）上表现良好。 k -truss 在关系密集的数据集（如 Football）中擅长生成高质量的社区。模块度方法具

有较少的严格约束，在较大数据集上很有效，并且通常优于 k -core 和 k -truss 方法。对于基于学习的 CS，QD-GNN 由于 GPU 内存溢出 (o.o.m.) 在大数据集上表现不佳。尽管在 NMI 得分上完全匹配（完美地生成了另一个社区而不是查询的社区），它在 I50/E50 比较中（尤其是在 Karate 和 Dolphins 上）难以处理分布偏移。增加训练样本大小带来的额外好处微乎其微。与 QD-GNN 相比，COCLEP 在 I50/E50 比较中对分布偏移保持鲁棒。基于 E50/E200 的比较，它受训练样本大小的影响更大。ICS-GNN 在所有数据集上表现尚可，在小数据集上表现良好，但在较大数据集上与其它方法相比表现不佳。

总体而言，*COMET* 在大多数数据集上以显著优势取得了最佳结果，并且其性能在交互模式下得到进一步增强。这突显了 *COMET* 在 CS 任务中的鲁棒性和适应性。

表 4.3 *COMET* 多轮 ICS 的有效性评估
Table 4.3 Effectiveness Evaluation of Multi-round ICS of *COMET*

AL ¹	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal	平均 ²
$\times$	95.45	95.91	70.42	45.06	91.03	75.52	46.92	69.54	-
R-1	96.96	97.02	64.91	45.64	92.44	76.31	53.46	73.63	1.32
R-2	98.39	98.28	77.20	46.35	94.43	80.02	58.37	75.45	3.52
R-3	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76	2.30
累计提升 ³	2.94	4.09	17.23	1.39	3.29	7.48	13.41	7.22	7.13

1. $\times$ 表示没有交互，R-1/2/3 表示有 1/2/3 轮主动学习（或来自真实情况的交互）。

2. 表中最后一列给出了相邻交互轮次之间的平均性能提升幅度。

3. 累计提升³表示相较于初始轮次，各数据集在多轮交互后的平均累计百分比提升。

4.3.2 效率评估

本节分析 *COMET* 的运行效率。本文首先将其与基于学习的基线方法进行对比，随后评估其核心组件的时间开销，从而全面考察 *COMET* 的效率表现。

效率的基线比较 在效率分析中，本文主要将 *COMET* 与基于学习的基线方法进行比较。这是因为本文关注的核心对象是学习式社区搜索方法，而传统社区搜索方法（如 k -内聚子图方法）的运行效率在很大程度上受实现平台、工程优化和并行化策略等因素影响。例如，采用 C++ 实现或并行加速可能显著改变其运行时间。因此，为保证比较的公平性与可解释性，本文不将传统方法纳入效率对比。

表 4.4 显示，*COMET* 在默认设置下是节省时间的。在交互模式下，*COMET* 的性能也与其它方法相当。使用最多 3 个交互轮次，*COMET* 花费的时间大约是默认设置的 3 倍，正如预期的那样。*COMET* 在默认设置上以显著优势优于 ICS-GNN，并且在交互模式下相当。QD-GNN 和 COCLEP 通常需要更多时间，因为训练数据随样本量

的增加而增加。

表 4.4 基于学习的社区搜索的效率评估

Table 4.4 Efficiency Evaluation of Learning-based Community Search

方法	设置	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
ICS-GNN	-	142.05	113.60	161.31	244.84	434.57	423.62	986.84	3428.92
QD-GNN	E50	279.64	290.33	277.62	264.49	o.o.m.	o.o.m.	o.o.m.	o.o.m.
	E200	1177.59	1072.12	1122.32	1194.02	o.o.m.	o.o.m.	o.o.m.	o.o.m.
COCLEP	E50	1126.13	930.54	1025.36	1103.80	1461.44	1440.35	2695.24	5747.50
	E200	4625.42	3849.48	4548.95	4391.42	4882.10	4855.20	5082.15	9114.06
COMET	默认	39.80	28.13	27.76	56.92	89.54	93.96	305.44	1781.04
	交互式	107.94	83.62	75.19	106.27	241.77	307.55	1023.28	5815.46

表 4.5 COMET 组件的效率评估

Table 4.5 Efficiency Evaluation of COMET Components

阶段	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
T_{PPR}	0.0037 ± 0.0001	0.0043 ± 0.0001	0.0038 ± 0.0020	0.0109 ± 0.0038	0.9039 ± 0.4156	1.5655 ± 0.3195	4.3086 ± 0.8430	27.9950 ± 10.2131
T_{GNN}	0.5145 ± 0.0465	0.3976 ± 0.0005	0.3523 ± 0.1959	0.5382 ± 0.2744	0.2965 ± 0.1852	0.4705 ± 0.2547	0.7937 ± 0.4275	1.0668 ± 0.5411
T_{NS}	0.0093 ± 0.0000	0.0092 ± 0.0001	0.0071 ± 0.0052	0.0048 ± 0.0021	0.0040 ± 0.0030	0.0054 ± 0.0024	0.0067 ± 0.0019	0.0066 ± 0.0036
T_{infer}	0.0085 ± 0.0001	0.0071 ± 0.0000	0.0054 ± 0.0026	0.0054 ± 0.0018	0.0045 ± 0.0022	0.0054 ± 0.0019	0.0074 ± 0.0027	0.0089 ± 0.0033
T_{total}	0.5397 ± 0.0464	0.4181 ± 0.0004	0.3760 ± 0.2029	0.5314 ± 0.2842	1.2089 ± 0.5657	1.5378 ± 0.5163	5.1164 ± 1.0706	29.0773 ± 10.5379

COMET 组件的效率评估 表 4.5 列出了 COMET 4 个关键组件的时间消耗：PPR 组件 T_{PPR} 、GNN 组件 T_{GNN} 、AL 组件 T_{NS} 和推理组件 T_{infer} 。 T_{PPR} 与图的大小成正比。由于 PPR 组件生成的子图大小一致， T_{GNN} 在所有数据集上都相似。 T_{NS} 和 T_{infer} 的时间消耗极小。总处理时间 T_{total} 表明，基于 PPR 的候选子图构建时间 T_{PPR} 极大地影响了 COMET 的效率，并随图的大小而扩展。

4.4 消融实验

对于**关键问题二**，为辨析 COMET 两个核心组件的功能边界与相互依赖性，本节分别对候选子图构建（B1）与“移除 GNN 核心（B2）进行消融。具体地，移除 B2 时，系统跳过所有基于标签的学习与推理，直接将 B1 产出的候选子图 CS'_Q 视为估计社区 $\hat{C}_Q$ ；移除 B1 时，B2 不再在一个与查询语义对齐的局部子图上学习，而是被动地在原始大图 G 上以极少标签进行直推式训练与推理。

表 4.6 GNN 和候选子图的消融实验
Table 4.6 Ablation on GNN backbone and candidate subgraph

组件	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
原始模型	98.39	100.00	89.87	45.44	94.32	83.04	60.30	76.75
移除 GNN 节点分类 $\Delta\%$	67.25 -31.14	68.74 -31.26	52.77 -37.10	33.23 -12.21	79.00 -15.32	59.84 -23.20	41.43 -18.87	66.19 -10.56
移除候选子图 $\Delta\%$	100.00 1.61	96.34 -3.66	85.28 -4.59	27.21 -18.23	0.18 -94.14	0.15 -82.89	0.01 -60.29	o.o.m. -

表 4.6 的结果显示二者缺一不可：其一，移除 B2 会在所有数据集上显著降分（如 *Football* -37.10%、*DBLP* -23.20%），说明在仅有 B1 的拓扑子图的情况下，仍需要一个利用正负标签刻画边界的学习器来完成优化，否则难以在复杂边界处维持高精度；其二，移除 B1 在小图上影响有限（*Karate* 仍为 100.00），但在大图上几乎无法得到正常的社区推荐（*Amazon/DBLP/Youtube* 的 F1 分别降至 0.18/0.15/0.01，*LiveJournal* 更出现 OOM）。这一退化并非仅是效率退步，而是学习问题本身被放大到不可解的症状：在大规模 G 上，在线直推式范式只能从极少的查询标签（少量 Q^+ 与少量 Q^- ）获得监督，信号在百万级节点的传递中被严重稀释，标签极度不平衡诱发决策边界的退化，从而使 F1 在大图上接近于零；反观小图， $|G|$ 与 $|CS'_Q|$ 近似，同样的标签密度足以维持有效学习，因而成为例外。

从机理上看，B1 主要发挥定域作用。具体而言，本文基于正查询节点集合 Q^+ 进行个性化 PageRank (PPR) 随机游走，得到节点重要性排序 $\text{apr}(\alpha, \chi_{Q^+})$ ，随后对该排序结果取负对数并计算一阶差分。通过在差分最大跃迁处进行截断，可得到候选搜索空间 CS'_Q （见图 3.2）。该截断点通常与真实社区的导通边界相一致，使得搜索空间在拓扑结构上与查询语义更好地对齐，并显著提高边界附近的类别区分度。因此，B1 能够将原本难以通过标签传递直接学习的大规模图压缩为规模适中、结构更清晰的可学习子图。

从效率角度看，PPR 构建时间 T_{PPR} 是总体延迟的主要来源。这主要是因为本文采用并行图计算库 *graph-tool* 实现 PPR，并使用了较为基础的迭代式计算方法。尽管如此，实验结果表明，PPR 构建时间随图规模增长仍表现出良好的扩展性。因此，该步骤虽然带来一定计算开销，但能够显著提升后续搜索空间的质量，是一种具有较高性价比的前置定域操作。

与之互补，B2 的作用在于定界：在 CS'_Q 上以二分类损失联合使用 Q^+ （内聚性）与 Q^- （排他性），借助图消息传递把负例的约束压到边界上。表 4.6 中原始方法与移除 GNN 的性能差即是这种子图精炼带来的净增益（如 *Amazon* 的 +15.32%）。值得注

意的是，主动学习阶段对高不确定性样本的偏好，使模型能够优先采集更具判别价值的负样本，从而进一步强化边界约束，并加速搜索空间与真实社区边界的对齐。

4.5 敏感性分析

对于**关键问题二**，本章对 *COMET* 框架的完整流程进行了五项关键的敏感性分析，包括 1) 候选子图构建、2) PPR 子图的参数选择、3) 子图表示、4) GNN 架构和 5) 节点选择策略的变体。

4.5.1 子图构建

第一部分基于子图方法。对于以查询为中心的子图构造方法，BFS 和 PPR 同样分别设置为返回 50 和 200 个节点；对于拓扑内聚子图方法，本文采用 k -core、 k -truss 和 METIS 作为代表；对于基于相似性的子图方法，本文采用 SimRank^[130] 和 Panther^[131]，并分别在 50 和 200 个节点规模下进行评估。

算法 4.1 基于 k -内聚子图的候选子图构建

Input: 无向图 $G(V, E)$ ，正查询集 Q^+ ， $k_{\min}$ ， $k_{\max}$

Output: 候选子图 CS_Q^t 。

```

1 if  $G_{sub}$  为空 then
2    $G_{sub}[k] \leftarrow$  预计算  $k \in [k_{\min}, k_{\max}]$  范围内的所有子图
3  $k \leftarrow -1$ 
4 forall  $k_{cur}$  从  $k_{\max} \rightarrow k_{\min}$  do
5   if  $G_{sub}[k_{cur}]$  为 None then
6     continue
7   forall  $subgraph \in G_{sub}[k_{cur}]$  do
8     if  $Q^+ \subseteq subgraph$  then
9        $CS_Q^t \leftarrow subgraph$ 
10       $k \leftarrow k_{cur}$ 
11      break
12   if  $k \neq -1$  then
13     break
14 if  $CS_Q^t$  为 None then
15   return “未找到合适的  $k$ -子图”
16 else
17   return  $CS_Q^t$ 

```

第一部分是预计算所有可能的子图（第 2 行），其基本直觉很简单：在指定范围

($k \in [k_{\min}, k_{\max}]$) 内的所有潜在 k -内聚子图仅依赖于图 G 的拓扑。因此, 这种方法通过避免在查询阶段进行冗余计算, 显著减少了框架的运行时间。预计算之后, 按 k 的降序枚举所有可能的 k -内聚子图。较大的 k 值对应于具有较高内聚性的子图, 使其更符合上下文中社区的定义。随后, 识别出所有正查询节点都包含在子图中的最大 k (第 8 行)。此步骤遵循了内聚子图 (CS) 的定义。如果遇到这样的子图, 算法立即终止循环并返回结果 (第 13 行)。在没有子图满足标准的情况下, 算法输出错误消息 (第 15 行)。否则, 候选子图 CS'_Q 作为输出返回 (第 17 行)。

对于 k -core 和 k -truss 分解, 本文采用自适应的 k 值来构建候选子图。具体而言, 对于给定查询, 本文在预设搜索范围内寻找能够生成可行子图的最大 k 值, 以获得尽可能紧致的拓扑内聚结构。对于 k -内聚子图方法, 需要预先指定搜索范围 $k_{\min}$ 与 $k_{\max}$ 。本文将 k -core 和 k -truss 的 $k_{\max}$ 均设置为 10; 对于 $k_{\min}$, 分别将 k -core 设置为 1、将 k -truss 设置为 3, 对应两类 k -内聚子图方法的最小可行取值。

对于 METIS, 分区数设置遵循 COCLEP^[21] 中的实验配置。由于单个分区未必包含所有正查询节点, 本文采用一种直接的重连策略以保证有效候选子图的存在。具体而言, 本文将所有包含至少一个正查询节点的分区进行合并, 并将合并后的结果作为对应查询的候选子图。相比之下, k -core 和 k -truss 分解本身与查询节点无关, 且缺乏类似的简单重连机制。因此, 尤其是在 k -truss 设置下, 这类方法可能无法生成同时包含所有正查询节点的合适候选子图。

对于诸如 SimRank^[130] 和 Panther^[131] 之类的基于相似性的方法, 这两种方法都需要为每个查询节点迭代聚合相似性得分。此过程将各个相似性结果相加, 并假设所有查询节点具有相等的权重。对于任何给定节点 v , 此节点与整个查询集之间的相似性可以表述为: $S(v, Q^+) = \sum_{q \in Q^+} w(q) \cdot s(v, q)$, $v \in G$, 在本文中 $w(q) = 1$, 而 $s(v, q)$ 是节点 v 和查询节点 q 之间的相似性得分。一旦单源评估完成, 就可以快速对结果进行排序以提取前 50/200 个节点。

表 4.7 概述了子图构建方法的变化。对于具有 50 和 200 个节点的以查询为中心的子图, BFS 和 PPR 产生相似的结果, 尤其是在 BFS-200 和 PPR-200 上。拓扑内聚子图表现出不同的性能: k -core 方法在小图上效果很好, 但在大图上不行, k -truss 方法由于其严格的约束, 适用性有限。METIS 在前六个图上表现一致, 尤其在较小的图上表现出色。对于基于相似性的子图, 两种方法在小图上表现良好, 但在大图上表现不佳。

表 4.7 *COMET* 的子图构建方法选择
Table 4.7 Selections of Subgraph construction methods of *COMET*

类型	设置	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
本文方法	PPR-自适应	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76
以查询为中心的 子图	BFS-50	100.00	95.31	66.75	56.81	93.56	67.79	38.62	68.40
	BFS-200	100.00	99.41	85.82	56.28	76.19	43.53	20.47	55.13
	PPR-50	100.00	92.29	78.30	52.75	94.17	64.65	39.32	73.30
	PPR-200	100.00	97.03	85.33	61.66	76.14	38.99	22.59	58.58
拓扑内聚子图	k -core	92.47	100.00	84.06	36.91	41.88	6.69	1.05	12.05
	k -truss	- ¹	-	87.27	-	-	-	-	-
	METIS	98.57	99.40	90.95	61.64	95.53	74.04	17.77	56.43
基于相似性的子 图	simrank-50	98.57	97.56	85.72	36.15	11.97	7.22	2.26	26.42
	simrank-200	97.04	93.05	84.45	17.89	81.63	48.72	21.20	51.11
	panther-50	98.39	93.65	86.60	28.89	73.54	30.48	15.67	41.02
	panther-200	100.00	94.24	73.55	63.27	TLE ²	TLE	TLE	TLE

1. 不适用 (Not applicable) 意味着该方法未能生成有效的子图以进行后续程序。
2. TLE (Time Limit Exceeded) 意味着该方法无法在 24 小时内生成所有有效结果。

表 4.8 基于 PPR 的子图构建的参数选择
Table 4.8 Parameter choices on PPR-based subgraph construction

参数	设置	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
α	0.5	98.39	97.50	80.50	56.91	93.73	75.17	45.18	72.01
	0.2	100.00	93.22	94.23	53.72	95.62	79.00	51.80	76.15
	0.1	88.46	83.33	94.06	50.43	93.77	81.34	57.38	76.90
	0.05	98.39	100.00	86.89	46.45	94.32	83.00	60.33	76.76
	0.02	98.39	98.78	87.00	35.19	94.53	83.11	58.75	75.33
ϵ	10^{-2}	100.00	100.00	87.24	44.41	94.25	82.90	59.27	76.36
	10^{-3}	98.39	97.56	85.82	44.32	94.92	82.78	59.41	76.77
	10^{-4}	98.39	100.00	86.89	46.45	94.32	83.00	60.33	76.76
	10^{-5}	98.39	98.78	86.84	45.07	94.37	82.87	60.30	76.78
$k_{\min}$	5	65.79	100.00	86.52	38.90	90.87	76.39	58.60	70.22
	10	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76
	15	98.39	100.00	86.15	48.88	95.55	78.17	55.27	76.90
	20	100.00	100.00	80.71	50.51	95.97	73.53	50.48	76.65
$k_{\max}$	50	98.39	81.64	92.76	42.87	91.57	85.07	59.89	72.69
	100	98.39	98.78	89.73	44.66	94.06	82.69	60.30	74.55
	200	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76
	500	100.00	99.40	85.60	44.76	94.32	83.04	60.33	76.81

4.5.2 基于 PPR 的子图构建的参数选择

表 4.8 显示了用于基于 PPR 的子图构建的参数选择 (α 、 ϵ 、 $k_{\min}$ 和 $k_{\max}$)。基线配置 ($\alpha = 0.05$ 、 $\epsilon = 10^{-4}$ 、 $k_{\min} = 10$ 、 $k_{\max} = 200$) 作为参考点, 这些参数的变化揭示了它们对输出社区质量的影响。该实验参考^[106]进行参数网格选择。

参数 α 极大地影响了不同数据集的生成质量。较大的 α 针对查询附近的节点, 这对于像 EU-Core 这样具有较少内聚结构的数据集有利。较小的 α 适用于像 Football 这样具有内聚结构的数据集。相反, ϵ 显示出有限的敏感性, 有助于在宽松的收敛阈值下实现一致的性能。对于大小参数, $k_{\max}$ 在超过 200 的阈值后影响甚微。较低的 $k_{\max}$

可以限制边界,在某些情况下会提高质量,但超过该值后,该参数的影响可以忽略不计。这支持将 $k_{\max} = 200$ 设置为以保持性能一致。 $k_{\min}$ 则存在不同趋势。在 EU-Core 数据集中,随着 $k_{\min}$ 的增长,指标也随之增加,这表明 EU-Core 需要比初始 PPR 边界建议的更多节点,但此设置会降低 Youtube 上的性能。

4.5.3 子图表示

在节点嵌入评估中,本文根据不同方法在捕捉子图结构信息方面的相关性与互补性,选取了六类代表性方法进行比较。

首先,本文引入可学习嵌入作为基础基线,用于考察在不显式引入拓扑属性的情况下,模型是否能够仅通过训练过程学习到有效的拓扑与语义表示。其次,本文采用 COCLEP^[21] 中提出的子图核心度作为一种适用于无节点属性图的结构表示,并将其与 PPR 表示进行对比。二者均能够在一定程度上刻画子图拓扑结构,因而与本文所研究的社区搜索问题具有较高相关性。第三,本文引入 ICS-GNN^[22] 中采用的高斯嵌入作为节点表示基线。

此外,本文还选取了两种基于节点摘要 (sketch) 的嵌入方法:其一是基于 NodeSketch^[132] 的节点表示方法;其二是 GraphSketch^[133] 中面向链接预测任务的简图式链接表示方法。为将后者适配至社区搜索任务,本文将其原有的链接预测模块替换为分类头,并基于该读出函数预测给定查询下各节点属于目标社区的概率。

表 4.9 显示了六种额外的子图表示: Learned embedding、normalized coreness^[21]、PPR、Gaussian^[22]、Nodesketch^[132] 以及用于节点分类的 ELPH^[133] 的修改版本。

表 4.9 子图表示的选择
Table 4.9 Selections of subgraph representations

方法	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
Node2vec	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76
Learned Embedding	90.74	100.00	78.97	41.74	94.03	81.59	59.19	76.02
Coreness	72.59	69.01	47.66	32.47	85.57	66.00	38.73	71.37
PPR	73.21	94.41	84.67	36.43	92.90	71.54	42.01	73.11
Gaussian	98.39	97.56	82.70	43.35	94.72	82.51	61.58	76.85
Nodesketch	88.46	68.74	55.82	32.68	92.34	71.25	51.55	58.23
ELPH-revised	92.82	95.45	74.66	38.65	88.46	75.11	51.83	73.91

根据结果, COMET 使用的 Node2vec 在大多数情况下优于大多数方法。Learned Embedding 方法在较小的图上表现稍差,但在较大的图上表现出有竞争力的性能,表明其在处理复杂结构时具有鲁棒性。依赖单值表示的核心度和 PPR 未能捕获足够的语义和拓扑信息,导致作为子图嵌入时性能下降。Gaussian 嵌入尽管是固有随机的,但产生了相对较强的结果,可能是由于为局部化提供了足够的可变性。NodeSketch 和

ELPH-修改版虽然旨在用于嵌入任务，但表现不如 Node2vec 和 Learned Embedding。它们的设计更适合链接预测而非节点分类。

4.5.4 GNN 架构

表 4.10 比较了 GNN 架构的三种变体：MLP、GraphSAGE^[32] 和 GAT^[33]。

MLP 缺乏节点聚合，在小图上表现不佳，但在大图上具有竞争力，因为子图嵌入可能封装了重要的拓扑信息，允许 MLP 在不关注邻居节点的情况下捕获更多的全局语义模式。对于 GraphSAGE，其简单性通过避免过拟合有助于在较大图上的性能，尽管它在较小图上表现不佳，而在这些图上 GCN 的完全聚合是有益的。GAT 在基准测试中表现不佳，可能是由于其结构复杂而训练资源有限。

表 4.10 GNN 架构选择
Table 4.10 GNN architecture choices

模型	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
GCN	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76
MLP	67.62	75.06	57.05	44.97	93.46	83.69	62.18	76.02
GraphSAGE	98.39	95.65	90.14	44.18	94.53	84.40	63.09	78.82
GAT	88.46	93.00	79.59	38.37	94.59	82.16	59.63	77.16

4.5.5 节点选择

在本部分中进一步分析节点选择机制中的两个关键因素，即 LSE 聚合得分的时间调整策略以及聚合器选择的变化，以考察二者对交互式社区搜索性能的影响。

表 4.11 基于 LSE 的聚合函数中的时间调整策略
Table 4.11 Time-adjust strategy in LSE-based aggregate function

时间调整	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
TRUE	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76
FALSE	100.00	100.00	86.47	41.85	93.40	82.15	58.52	75.70

时间调整策略 为考察上述时间调整项的作用，进一步构造不带时间调整的聚合形式，即移除分母中的轮次相关项，得到

$$\phi_{\text{aggregated-static}}(v_i) = \log\{\exp[\phi_{\text{topo-norm}}(v_i)] + \exp[\phi_{\text{ent-norm}}(v_i)]\} \quad (4.9)$$

表 4.11 展示了基于 LSE 聚合得分中时间调整策略的消融结果。总体来看，引入时间调整后，COMET 在八个数据集上的平均 F1 提升约为 1.1%。尽管在 Karate 数据

集上结果略有下降,但该数据集仅包含两个社区,规模较小且波动更易受个别查询影响;从整体结果看,时间调整策略仍然对基于 LSE 的聚合采样带来了稳定的正向贡献。

表 4.12 主动学习 (AL) 得分函数的选择
Table 4.12 Selections of AL score functions

配置	Karate	Dolphins	Football	EU-Core	Amazon	DBLP	Youtube	Livejournal
LSE	98.39	100.00	87.65	46.45	94.32	83.00	60.33	76.76
随机 熵 PPR	84.10	97.02	86.40	42.21	95.56	83.18	57.18	75.50
	98.39	99.40	87.83	41.29	94.23	82.64	57.60	75.69
	90.74	93.60	81.46	40.09	92.38	82.46	59.01	72.82
均匀 Beta	98.39	98.78	80.62	43.79	94.26	82.91	58.57	75.06
	98.39	100.00	89.38	40.94	94.10	82.19	56.75	74.94

采样函数 本部分共选择三种单标准节点选择策略和两种双标准节点选择策略进行比较。

首先,单标准策略包括基于 PPR、熵和随机选择的得分函数。其次,双标准策略包括使用 Beta 分布和均匀分布进行加权。Beta 分布因其在纳入时变效应方面的灵活性而在先前的研究中被广泛采用^[134-135],其公式为:

$$\phi_{\text{aggregated}}(v_i) = (1 - \beta_t) \phi_{\text{topo-norm}}(v_i) + \beta_t \phi_{\text{ent-norm}}(v_i) \quad (4.10)$$

默认为先前的研究, Beta 分布参数是从时间调整参数中采样的,初始值分别为 2:

$$\beta_t \sim \text{Beta}(a, b) \quad a = 2, b = 2 \cdot \frac{R - t + 1}{R + 1}, \quad t = 0, 1, 2, \dots, R - 1 \quad (4.11)$$

此处比较均匀分布(也称为等权重策略)作为参考。

表 4.12 展示了 AL 的五种额外得分函数配置。单标准选择使用随机、熵和 PPR 得分,而双标准则应用 PPR 和熵的均匀(uniform)和 beta 分布加权。*COMET* 的 LSE 配置在所有数据集中都很鲁棒,双标准方法显示出比单标准方法更好的一致性。

4.6 合成图模拟

为了探索边界条件并理解 *COMET* 在不同图密度和结构下的性能,使用社区检测中两种广泛采用的图生成模型进行模拟: Planted Partition Model^[136] (PP, 也称为 Stochastic Block Model, SBM^[59]) 和 LFR 基准图^[137]。

鉴于现有研究在合成图参数设置上尚未形成统一标准，为确保评估的公平性与可比性，本文在两类合成图模型上对 *COMET* 进行系统测试。具体而言，本文选取了一组能够平衡图稀疏性、密度以及结构复杂性的参数配置，用以覆盖不同难度的社区搜索场景。上述参数设置基于对八个真实数据集结构特征的经验分析确定。

4.6.1 模拟设置

LFR 基准图和 PP 图的设置基于所有八个数据集所使用的参数范围。为了更有效地展示性能比较，本文主要调节两个关键参数：混合水平 (μ) 和节点度。其中，LFR 图的节点度由 $\bar{d}^G(C_Q)$ 表示，PP 图的节点度由 $\bar{d}^{sub}(C_Q)$ 表示。首先，控制社区间边重叠程度的混合水平 μ 在 0.1 到 0.7 之间变化，与真实数据集中的观测范围保持一致。其次，对于 LFR 图，节点度由 $\bar{d}^G(C_Q)$ 控制，并分别设置为 10 和 50，以对应低密度和高密度图。对于 PP 图，节点度由社区内边概率 p_{in} 控制，同样设置为 0.1 和 0.5 以反映低密度和高密度配置。网络规模在所有合成图中保持一致，均固定为 1,000 个节点。对于 LFR 基准图中的大小和社区规模分布参数，本文遵循^[138]的设置，取 $\tau_1 = 3$ 和 $\tau_2 = 1.5$ 。对于 PP 图，本文将社区规模设置为 20 或 50 个节点，分别对应 50 个和 20 个社区。

在实验设置上，每种参数组合生成 10 张图，并在每张图上构造 20 个查询，因此每个参数网格共包含 200 个查询。与前述实验保持一致，每种设置均采用 3 轮主动学习交互，并使用相同的标注预算。

4.6.2 模拟图结果

表 4.13 和表 4.14 分别给出了 PP 图与 LFR 图上的模拟结果。总体来看，两类模拟图呈现出一致的趋势。首先，在绝大多数参数设置下，引入 GNN 都能够进一步提升社区搜索性能，说明即使候选子图已经将搜索空间压缩到局部区域，学习驱动的社区判别仍然能够有效修正社区边界。其次，随着混合水平 μ 的增加，社区与外部节点之间的连接增强，目标社区边界逐渐模糊，因此社区搜索性能整体上呈下降趋势。这一现象在 PP 图和 LFR 图中都较为明显。最后，图的局部稠密程度以及 PPR 的参数配置会进一步影响候选子图构建质量，从而影响最终结果。

对于 PP 图，可以观察到 p_{in} 、 $|C_Q|$ 与 α 之间存在明显的交互作用。当 $p_{in} = 0.1$ 时，图整体较为稀疏，不同设置之间的差异相对有限，但 α 的选择仍会影响模型在低混合与高混合区间中的稳定性。当 $p_{in} = 0.5$ 时，社区内部结构更加清晰，整体性能显著提升；此时，较小的 α 往往更适合规模较小且边界清晰的社区，而在规模较大、

表 4.13 PP 图上的模拟结果
Table 4.13 Simulation on Planted Partition Models

p_{in}	$ C_Q ^1$	α	GNN ²	$\mu^{0.1}$	$\mu^{0.2}$	$\mu^{0.3}$	$\mu^{0.4}$	$\mu^{0.5}$	$\mu^{0.6}$	$\mu^{0.7}$
0.1	20	0.5	✗	72.84 ± 2.96	67.00 ± 1.23	62.28 ± 2.35	59.93 ± 1.33	53.70 ± 1.45	45.90 ± 1.37	38.30 ± 1.14
			✓	76.63 ± 2.72	70.36 ± 1.84	65.47 ± 2.47	63.76 ± 1.02	60.86 ± 1.36	53.14 ± 1.07	45.23 ± 1.48
		0.05	✗	68.99 ± 4.22	64.04 ± 3.62	61.48 ± 3.90	57.57 ± 2.54	55.51 ± 2.33	51.22 ± 2.86	44.08 ± 3.38
			✓	73.86 ± 4.26	67.56 ± 3.36	65.29 ± 3.42	61.07 ± 2.25	59.24 ± 1.66	55.60 ± 2.06	49.02 ± 2.87
	50	0.5	✗	66.19 ± 2.55	63.04 ± 2.31	58.64 ± 3.89	54.08 ± 1.45	48.39 ± 1.44	42.26 ± 1.23	35.24 ± 0.95
			✓	68.28 ± 3.28	65.00 ± 2.39	60.96 ± 4.11	56.76 ± 1.76	50.98 ± 1.65	45.02 ± 1.05	37.83 ± 1.11
		0.05	✗	83.22 ± 3.38	62.03 ± 4.97	51.04 ± 3.35	42.56 ± 2.34	38.27 ± 0.96	34.54 ± 1.23	30.83 ± 1.31
			✓	84.41 ± 3.50	63.20 ± 5.31	51.88 ± 3.61	43.10 ± 2.45	38.73 ± 1.13	34.96 ± 1.29	31.48 ± 1.32
0.5	20	0.5	✗	91.56 ± 2.66	78.42 ± 1.80	66.09 ± 2.84	53.84 ± 4.09	42.62 ± 2.90	29.98 ± 1.14	25.71 ± 2.69
			✓	94.73 ± 1.88	83.21 ± 1.66	71.80 ± 2.36	59.30 ± 3.45	48.15 ± 2.52	35.66 ± 1.02	29.49 ± 2.67
		0.05	✗	97.20 ± 0.20	96.37 ± 0.43	94.24 ± 1.10	89.97 ± 1.89	84.24 ± 2.42	78.59 ± 1.57	74.51 ± 1.33
			✓	97.20 ± 0.20	96.40 ± 0.42	94.40 ± 1.09	90.22 ± 1.85	84.59 ± 2.55	79.09 ± 1.65	75.15 ± 1.40
	50	0.5	✗	90.72 ± 1.23	78.23 ± 1.13	61.95 ± 1.40	48.42 ± 1.21	79.87 ± 3.28	70.95 ± 2.58	60.48 ± 3.03
			✓	94.23 ± 0.92	83.67 ± 1.07	71.89 ± 1.27	60.76 ± 1.43	82.29 ± 2.77	72.32 ± 2.99	65.70 ± 2.96
		0.05	✗	98.99 ± 0.00	98.91 ± 0.07	97.61 ± 0.78	88.11 ± 2.96	69.38 ± 5.65	55.19 ± 2.51	44.57 ± 2.28
			✓	98.99 ± 0.00	98.94 ± 0.05	97.78 ± 0.75	88.57 ± 3.03	69.74 ± 5.72	55.62 ± 2.58	44.96 ± 2.42

1. $|C_Q|$ 表示目标社区的大小。

2. ✓ 表示启用 GNN, ✗ 表示仅使用候选子图。

表 4.14 LFR 上的模拟结果
Table 4.14 Simulation on LFR graphs

$\bar{d}^G(C_Q)$	α	GNN	$\mu^{0.1}$	$\mu^{0.2}$	$\mu^{0.3}$	$\mu^{0.4}$	$\mu^{0.5}$	$\mu^{0.6}$	$\mu^{0.7}$
10	0.5	✗	64.82 ± 4.10	58.15 ± 2.15	47.71 ± 2.97	37.35 ± 2.95	31.53 ± 1.90	25.76 ± 2.40	21.33 ± 1.68
		✓	80.95 ± 3.79	70.18 ± 3.33	57.19 ± 2.37	46.07 ± 2.33	38.76 ± 2.11	32.98 ± 3.37	28.66 ± 2.56
	0.05	✗	59.99 ± 3.96	58.83 ± 4.60	52.78 ± 7.00	42.65 ± 4.51	36.14 ± 5.07	31.83 ± 2.89	33.03 ± 2.92
		✓	75.94 ± 3.99	70.01 ± 5.33	66.23 ± 8.06	61.56 ± 3.36	52.72 ± 5.01	47.66 ± 6.28	49.94 ± 6.49
50	0.5	✗	94.80 ± 2.24	92.83 ± 4.57	75.08 ± 7.75	56.07 ± 6.42	39.68 ± 2.97	27.29 ± 4.99	16.74 ± 2.76
		✓	96.03 ± 1.75	95.65 ± 2.68	75.46 ± 7.92	57.08 ± 6.68	40.38 ± 3.25	28.20 ± 5.03	18.31 ± 2.82
	0.05	✗	99.16 ± 0.04	79.59 ± 6.11	45.80 ± 5.29	32.43 ± 5.38	24.20 ± 2.92	17.68 ± 1.41	15.57 ± 1.88
		✓	99.16 ± 0.04	79.70 ± 6.08	46.01 ± 5.31	32.88 ± 5.30	25.34 ± 3.25	18.86 ± 1.52	16.34 ± 1.85

外部混合程度更高的情形下，较大的 α 反而更有利于维持候选子图对查询局部邻域的聚焦。特别地，表 4.15 显示了在 $p_{in} = 0.5$, $|C_Q| = 50$, $\alpha = 0.5$ 的设置下，性能并未随 μ 单调下降，而是在 $\mu = 0.5$ 后出现阶段性回升。这表明当混合水平进入更高区间时，候选子图构建机制本身可能发生变化，使模型从“边界较松但覆盖更广”的搜索模式转向“范围更保守但精度更高”的局部搜索模式。

对于 LFR 图，整体规律与 PP 图一致，但表现出更明显的结构稀疏性效应。随着 μ 增加，社区搜索性能持续下降；同时，在低平均度设置 $\bar{d}^G(C_Q) = 10$ 下，GNN 带来的增益通常更加显著，说明在稀疏图中，学习模块对局部社区边界的修正作用更为

表 4.15 PP 图在 $p_{\text{in}} = 0.5$, $|C_Q| = 50$ 设置下的详细结果
Table 4.15 Detailed Results on Planted Partition Models with $p_{\text{in}} = 0.5$ and $|C_Q| = 50$

Metric	GNN	$\mu^{0.1}$	$\mu^{0.2}$	$\mu^{0.3}$	$\mu^{0.4}$	$\mu^{0.5}$	$\mu^{0.6}$	$\mu^{0.7}$
NMI	✗	80.57 ± 1.99	63.13 ± 1.68	44.72 ± 1.45	31.72 ± 1.03	66.55 ± 3.91	55.12 ± 2.79	43.16 ± 2.68
	✓	86.75 ± 1.72	69.97 ± 1.67	55.13 ± 1.57	43.24 ± 1.37	69.03 ± 3.49	56.78 ± 3.33	48.87 ± 2.42
Jaccard	✗	83.20 ± 2.01	64.98 ± 1.67	45.53 ± 1.57	32.09 ± 1.09	69.55 ± 4.00	57.07 ± 2.79	44.36 ± 3.05
	✓	89.23 ± 1.62	72.30 ± 1.67	56.49 ± 1.63	43.80 ± 1.47	72.04 ± 3.60	58.78 ± 3.36	50.21 ± 2.94
F1	✗	90.72 ± 1.23	78.23 ± 1.13	61.95 ± 1.40	48.42 ± 1.21	79.87 ± 3.28	70.95 ± 2.58	60.48 ± 3.03
	✓	94.23 ± 0.92	83.67 ± 1.07	71.89 ± 1.27	60.76 ± 1.43	82.29 ± 2.77	72.32 ± 2.99	65.70 ± 2.96
Precision	✗	83.25 ± 2.03	65.22 ± 1.72	45.68 ± 1.61	32.13 ± 1.11	73.14 ± 4.48	74.48 ± 3.93	71.14 ± 4.87
	✓	89.28 ± 1.64	72.56 ± 1.71	56.66 ± 1.65	43.85 ± 1.49	75.66 ± 3.99	76.32 ± 3.96	77.69 ± 3.95
Recall	✗	99.94 ± 0.08	99.58 ± 0.12	99.58 ± 0.15	99.78 ± 0.11	95.79 ± 1.58	81.00 ± 4.44	70.13 ± 5.23
	✓	99.94 ± 0.08	99.58 ± 0.12	99.58 ± 0.15	99.78 ± 0.11	95.79 ± 1.58	81.00 ± 4.44	70.11 ± 5.21
$ CS_Q $	✗	60.33 ± 1.59	78.76 ± 1.78	113.78 ± 3.77	158.00 ± 5.06	80.07 ± 8.76	67.02 ± 8.01	65.78 ± 7.62
	✓	56.16 ± 1.07	69.53 ± 1.49	89.47 ± 2.46	115.27 ± 4.12	71.15 ± 5.84	63.88 ± 7.34	54.07 ± 5.89

关键。当 $\bar{d}^G(C_Q) = 50$ 时，图的整体连通性增强，因此在较低混合水平下通常能够取得更高性能，但当 μ 进一步增大时，社区边界同样会迅速退化。总体而言，PP 图与 LFR 图上的结果共同表明，*COMET* 对不同图结构条件具有较好的适应性，而 GNN 模块在弱结构信号、边界模糊或候选子图质量受限的场景中尤为重要。

4.7 本章小结

本章通过在八个具有不同拓扑特征的真实世界数据集及两类合成图上的广泛实验，全面评估了 *COMET* 的有效性、效率及鲁棒性。实验结果表明，*COMET* 不仅在各项关键评价指标上显著优于现有的传统与基于学习的基线方法，而且通过多轮交互机制有效提升了社区搜索的精度与用户意图的对齐度。消融实验与敏感性分析深入揭示了基于 PPR 的候选子图构建与 GNN 节点分类模块的互补优势，证实了模型在不同参数配置与组件选择下的稳定性。此外，合成数据集上的模拟测试进一步验证了该框架在面对不同网络密度与噪声水平时的良好适应能力，确立了 *COMET* 作为一种高效、鲁棒的交互式社区搜索解决方案的优越性。

第 5 章 COMET 扩展与案例分析

5.1 有关有向图扩展的可行性分析

将 *COMET* 扩展到有向图设置, 对于建模信息传播、推荐关系以及具有显式方向含义的交互网络具有重要意义。在许多实际应用中, 边的方向往往蕴含了不对称的交互语义, 例如邮件网络中的“发件人 $\rightarrow$ 收件人”、引用网络中的“被引文献 $\rightarrow$ 引用文献”、以及关注网络中的“被关注者 $\rightarrow$ 关注者”。因此, 简单地将有向图无向化虽然在工程实现上较为方便, 但可能导致关键信息的丢失, 从而影响社区结构的判定与解释。本节从框架组件、随机游走/PPR 建模、GNN 消息传递以及社区定义与性质四个方面, 系统讨论 *COMET* 在有向图上的扩展及其潜在影响。

有向图设置下的图嵌入模块 在无向图情形下, *COMET* 采用 Node2vec^[124] 等随机游走式图嵌入方法来捕捉节点的局部与全局结构信息。对于有向图 $G = (V, E)$, 随机游走策略需要显式地考虑由出度矩阵决定的转移概率。若直接沿用无向图中的对称转移矩阵, 则会忽略出入度的差异, 削弱方向性在表示学习中的作用。幸运的是, Node2vec 本身支持在有向图上定义非对称的转移概率: 随机游走从节点 u 走向 v 的概率由有向边 (u, v) 是否存在以及预设的回溯/探索参数共同决定。在强连通分量内部, 随机游走能较好地刻画信息流动的主干路径; 但对于仅弱连通甚至存在大量汇点 (sink) 或源点 (source) 的图结构, 部分节点之间可能在有向意义下不可达, 从而导致对应嵌入向量落在彼此差异较大的子空间中。

有向图上的 PPR 与子图构建 PPR 本身就是在有向图上提出的局部重要性度量, 其核心思想是: 从查询节点集 Q 出发, 以概率 $1 - \alpha$ 沿有向边进行随机游走, 以概率 α 回到 Q , 从而得到一类偏向 Q 的稳态分布。对于有向图, 转移矩阵通常定义为

$$W = A^T D_{\text{out}}^{-1} \quad (5.1)$$

其中 A 为有向图的邻接矩阵, D_{out} 为出度对角矩阵。与无向情形不同, 有向图上的 W 不再对称, 稳态分布的存在与唯一性需要依赖图的强连通性或通过加入随机跳转机制来保障。实际网络中通常不存在完美的强连通结构, 大量节点可能仅单向可达或构成“悬挂”结构, 因此 PPR 的实现需要综合考虑以下几个方面:

- **强连通性与可达性。**若查询集合 Q 所在的子图与其它区域仅通过单向边相连, 则从 Q 出发的随机游走可能在有限步数后完全无法回到目标社区, 导致 PPR 得分在目标社区内部被稀释。通过适当增大回跳概率 α , 可以在一定程度上缓解这一问题, 但也会降低 PPR 对局部结构边界的敏感性。
- **近似算法的收敛性。**幂迭代与 push 等近似算法在有向情形下依然适用, 但需要针对 D_{out} 的长尾分布进行数值稳定性处理。特别是在出度高度偏斜的邮件或关注网络中, 少数“超级枢纽”节点会吸收大量流量, 若不做截断或归一化处理, 可能导致候选子图被这些枢纽主导, 而忽略弱连通但语义上紧密的局部结构。
- **方向性对子图边界的影响。**在无向图中, *COMET* 通过对 PPR 得分排序并检测数值的急剧下降点来确定子图边界; 在有向图设置下, 这一边界更应被解读为: 从查询出发的稳态流量在何处发生显著衰减。这意味着某些入度很高但从 Q 出发难以到达的节点, 即便在无向拓扑上靠近 Q , 也会得到较低的 PPR 得分, 从而被排除在候选子图之外。这样的行为更加符合以查询为源点的影响范围这一直观定义。

基于上述考量, 有向 *COMET* 在子图构建阶段保持了 PPR 框架不变, 但在参数选择与数值实现上更强调对出度分布与可达性的鲁棒性, 因此对于特定的数据集需要针对优化。

有向图上的 GNN 消息传递 *COMET* 在无向图场景中采用基于 GCN 的 GNN 模块, 通过对称归一化的邻接矩阵实现各层节点特征的聚合。在有向图中, 若直接采用 $(A+A^T)$ 的对称化邻接矩阵, 则会在一定程度上丢失方向信息; 但如果完全沿有向邻接矩阵 A 聚合, 则又可能忽略入边与出边在语义上的差异。因此, 有向 *COMET* 的 GNN 设计需要在保留方向性与控制模型复杂度之间作出权衡。

一种简洁可行的方案是引入双通道的消息传递: 分别基于入邻居和出邻居构造两类聚合算子, 例如

$$H^{(l+1)} = \sigma\left(\tilde{D}_{\text{in}}^{-1}A^T H^{(l)} W_{\text{in}}^{(l)} + \tilde{D}_{\text{out}}^{-1}A H^{(l)} W_{\text{out}}^{(l)}\right), \quad (5.2)$$

其中 $\tilde{D}_{\text{in}}$ 和 $\tilde{D}_{\text{out}}$ 分别为入度与出度的归一化矩阵, $W_{\text{in}}^{(l)}$ 与 $W_{\text{out}}^{(l)}$ 为不同方向的变换参数。该结构在参数量上对标准 GCN 仅增加常数倍, 但允许模型在学习社区判定边界时区分接收信息的节点 (高入度) 与传播信息的节点 (高出度)。在实验中, 本文也进一步比较了三种设计的前两个: 完全无向化 (只使用 $A + A^T$)、仅使用出边聚合

(因为由于所选数据集只有一类属于有向图, 双通道聚合的泛用性无法能够得到很好的验证)。后续研究可以对双通道的消息传递进行进一步探讨。

需要指出的是, 有向 GNN 的引入同样会影响主动学习阶段的节点选择策略。由于预测熵不仅取决于节点在嵌入空间中的位置, 也受到其在有向拓扑中所处流入/流出瓶颈位置的影响, 部分关键的桥接节点在有向图上可能更容易被高熵得分所识别, 从而被优先推荐给用户进行标注。

有向图下社区定义与性质的讨论 在无向图情形下, *COMET* 隐含采用了一种密集子图 + 局部随机游走保留率高的社区概念; 即一个好的社区应当在结构上内聚、在 PPR 意义上对查询节点具有高保留概率。在有向图中, 若直接沿用“强连通分量”作为社区的前提条件, 会带来过于严格的约束: 许多实际网络中, 一个功能上紧密的群体 (例如一组持续向外部群体输出信息的专家) 在有向图上可能表现为具有明确层次结构的“半流形”结构, 而非完美的强连通图。如果强行要求社区内部任意两个节点均有有向路径相互到达, 则这类具有明显层级与流向特征的社区会被排除在外。

因此, 在有向 *COMET* 中, 本文采用更为宽松且与 PPR 一致的社区定义: 相对于查询节点集 Q , 一个有向社区应当满足以下两点: (1) 在有向随机游走中, 绝大多数从 Q 出发的流量被限制在该子图内部, 即 PPR 质量在该子图上的累积占比高; (2) 子图内部的有效边密度较高, 且在方向上呈现相对一致或稳定的流动模式 (例如以若干“核心节点”为源, 向外围节点扩散)。在这种定义下, 强连通性可以视为一种足够条件, 而非必要条件; *COMET* 的子图构建与 GNN 判定过程则共同在这一宽松定义下搜索与查询最匹配的流量社区。

这一社区定义的改变也会影响结果的解释方式: 对于同一查询, 有向图上的社区往往具有更明确的信息源和信息汇, 社区内部可能出现明显的源—汇结构或链式结构; 而无向图版本则倾向于给出几何形态更加圆润的密集簇。在实际应用中, 前者更适合解释方向性强的行为模式 (如邮件发送、关注关系), 后者则更适合刻画对称交互关系 (如共同购买、双向好友)。

对 *COMET* 修正的实际表现 为了更具体地考察有向扩展对 *COMET* 行为及结果的影响, 本章选取 EU-Core 邮件网络作为代表性数据集进行实证分析。该数据集来源于某欧洲研究机构内部的电子邮件往来记录, 原始形式为有向图: 每条边 (u, v) 表示邮箱 u 向 v 至少发送过一封邮件。与第表4.1 中对 EU-Core 的基本统计描述一致, 该网络整体规模适中, 既包含少量高度活跃的“枢纽”邮箱, 也包含大量出入度较低的

表 5.1 基于 EU-Core 数据集将 COMET 扩展到有向图
Table 5.1 Extension of COMET to directed graph based on EU-Core dataset

设置	EU-Core 有向	使用无向嵌入	使用无向 PPR	使用无向 GNN
默认	50.01 $\pm$ 3.46	50.19 $\pm$ 3.54	49.55 $\pm$ 3.74	48.65 $\pm$ 2.53
交互式	52.33 $\pm$ 4.34	52.32 $\pm$ 4.28	52.66 $\pm$ 2.37	49.73 $\pm$ 3.06

普通员工邮箱，并呈现典型的分层结构。

在实验设置上，本章遵循上一章的基本配置。除此之外，该实验将 200 次查询分成每批 20 个查询（共 10 批）来进行对比。实验主要对比下列两种主版本：其一是论文主体中使用的无向基线，即在构建图嵌入、PPR 子图以及 GNN 消息传递时均采用对称化的无向图；其二是完全有向版本，即三个组件均基于有向邻接矩阵实现。如表 5.1 所示，在相同的初始查询规模和标注预算的条件下，有向版本的 COMET 在 EU-Core 上取得了 52.33 $\pm$ 4.34 的 F1，而无向版本的 F1 为 56.91。二者之间存在约 4 个百分点的差距，但并未出现显著退化，说明在该数据集上引入方向信息带来的优化空间与建模难度大致相当。

为了进一步区分不同组件的作用，本章在 EU-Core 上额外考察了若干混合配置，即将某一组件替换为其无向版本，而保持其余组件仍按有向设定工作。具体而言，分别考察：(i) 仅在图嵌入阶段使用基于无向图训练的 Node2vec 向量，而 PPR 与 GNN 均基于有向结构；(ii) 仅在子图构建阶段使用无向图 PPR 得分，而嵌入与 GNN 仍保持有向；(iii) 仅在 GNN 中采用无向化的消息传递结构，而嵌入与 PPR 继续利用有向拓扑。实验结果表明，当仅替换 Node2vec 嵌入或 PPR 得分时，整体 F1 的变化幅度较小，与完全有向版本的差距通常在实验方差范围内。而使用无向 GNN 的性能有一定损失，可能是由于基于无向图的 GNN 与 embedding 的不匹配导致的。但整体而言，在 EU-Core 上，以有向或无向方式建模局部相似性与局部重要性对最终社区边界的影响是有限的：邮件往来本身具有一定的互惠性，使得无向近似仍能捕捉到相近的结构模式。最后，平均 F1 指标体现不出长尾数据的性能（因为本身长尾分布的数据会被大量普通查询所掩盖），因此后续的工作可以从长尾数据分布的角度讨论有向图拓扑诱导的框架设计。

5.2 案例研究

这里展示两个相关案例，第一个为本文所使用的数据集上的交互式社区搜索示例，第二个则是考察 *COMET* 在其他领域（比如这里的虚假评论检测）上的应用潜力。

5.2.1 在 Karate 数据集上的交互式社区搜索

首先，本小节在经典的 Karate 空手道俱乐部数据集上进行了一个交互式社区搜索案例研究，用于直观展示 *COMET* 在小规模网络上的工作机理。该数据集包含 34 个节点与 78 条边，刻画了俱乐部成员之间的社交关系，并天然划分为两个社区，每个社区约 17 个成员，是社区搜索领域中广泛使用的示例之一。在这样一个规模可视化、结构清晰的小图上，可以将整个交互过程细致地画出来，从而更好地解释模型如何在多轮人机交互中逐步逼近真实社区。

在本案例中刻意采用了极为克制的标注设置：初始仅给出 1 个正查询节点和 1 个负查询节点，这两个节点分别位于两个真实社区中，起到锚点的作用；随后，在每一轮交互中，仅保留 1 个节点的标注预算，即模型每轮最多只能向用户询问一个额外节点的标签。尽管这一设置在实践中较为严格，但实验结果表明，模型在首轮交互后便已能够识别出基本的社区结构，说明即使在极低标注预算下，该交互机制仍能较好地满足此类社区搜索任务的需求。

如图 5.1 所示，该案例一共展示了三张子图，分别对应初始 + 两轮交互的完整过程：图 5.1a 给出了基于初始 1 个正、1 个负查询节点训练得到的第一轮社区搜索结果；图 5.1b 展示在第一轮主动标注之后，模型更新得到的第二轮社区；图 5.1c 则对应在第二轮主动标注之后得到的第三次社区搜索结果，也是 *COMET* 自动判定收敛、结束搜索的时刻。三张子图中：紫色节点表示正查询节点，棕色节点表示负查询节点；蓝色节点表示真阳性节点，红色节点表示真阴性节点；橙色节点表示假阳性节点，绿色节点表示假阴性节点。

具体来看，在图 5.1a 第一轮搜索中，*COMET* 首先基于 PPR 得分从全图中自适应抽取候选子图；在 Karate 这样的小规模图上，候选子图实际上覆盖了几乎全部节点。随后，GNN 以 1 正 1 负为监督信号，在候选子图内完成一次训练，输出每个节点属于当前社区的概率分数，并通过阈值化得到一个初步社区。从图中可以看到，初始社区已大致围绕正查询节点形成一个连通子图，但在社区边界附近仍存在少量假阴性节点：一些位于两个真实社区桥接位置的节点（主要是个别真实社区成员）尚未被完全覆盖，说明当前决策边界仍有优化空间。

为了提高后续轮次的效率, *COMET* 通过主动学习模块, 从当前候选子图中计算每个未标注节点的综合不确定性评分。该评分一方面使用 GNN 输出的预测熵刻画模型对该节点判断的不确定程度, 另一方面结合 PPR 得分衡量该节点在查询相关局部结构中的拓扑重要性, 并通过 $\log\text{-sum-exp}$ 形式进行聚合和时间加权, 从而在探索(选结构关键节点)与利用(选决策边界附近节点)之间取得平衡。根据这一评分, *COMET* 在第一轮结束时挑选出得分最高的一个未标注节点推荐给用户, 消耗本轮 1 个标记预算。用户反馈该节点的真实标签后, 模型将其与原有 1 正 1 负一并作为已标注集合, 重新进行一次 PPR 候选子图筛选并使用 GNN 进行一次训练。

图 5.1b 展示了在加入这一额外标注之后, 第二轮社区搜索的结果。可以观察到, 相比图 5.1a, 社区的整体形状更加接近真实划分。接下来, *COMET* 再次在更新后的模型输出上运行主动学习模块, 依据新的预测熵与 PPR 得分重新计算所有未标注节点的综合不确定性评分, 并再推荐一个节点给用户标注。凭借这一第二轮反馈, 模型再一次完成候选字体更新、参数更新与社区推断, 得到图 5.1c 所示的第三次搜索结果。此时, 可以看到被预测为社区成员的节点集合已经与 Karate 数据集的真实社区吻合(实际上在文献中也存在不同划分版本的边界节点上)。在这种情形下, *COMET* 根据预先设定的停止条件, 自动判定搜索已收敛并终止交互过程。

值得强调的是, 这一案例本身是一个相对简单的场景: 图规模很小、真实社区接近均匀划分、且查询节点选择理想。因此, 本案例可以将其视为 *COMET* 的演示: 在极低标注预算(初始 2 个标签 + 2 轮各 1 个标签, 共 4 个用户标注)的前提下, 主动学习策略能够驱动模型在三次社区生成中逼近真实答案, 直观展示了框架在交互效率与搜索质量之间取得的良好平衡。相比之下, 在真实的大规模社区搜索中, 候选子图往往包含更多节点, 社区结构也更加复杂、多层次, 用户的查询可能包含多个正负节点且存在噪声, 这些因素都会使交互过程远比 Karate 示例复杂。

5.2.2 在 Amazon-CN 数据集上的虚假评论检测

其次, 本章同时也在 Amazon-CN 数据集^[139]上开展了虚假评论检测的案例研究。该数据集由 Wu et al. [139] 基于 Amazon China 平台构建: 他们首先根据商品的 ASIN 在站内随机抽取评论数不少于 15 条的商品, 覆盖图书、电子产品、美妆、教育等多个品类; 随后收集所有曾点评这些商品的用户, 并在此基础上构建评论者网络 (reviewer network)。在公开的图数据中, 节点对应评论者, 边为作者按照共同评论行为等规则定义的用户-用户关系, 该网络被视作一张无向图, 用于后续的刷评检测建模。该数据集最终确认了 985 名高置信度刷评用户 (总用户数量: 8374)。

鉴于异常社区在该数据集中的分布具有显著异构性, 该实验首先需要在评论者关系网络上应用 *Louvain* 算法进行社区划分, 并移除节点数少于 5 的小社区, 仅保留规模较大的社区作为后续搜索的目标社区。随后, 在目标社区中选取 4 个已知刷评账号作为正查询节点、3 个正常用户作为负查询节点构成初始查询集, 在没有额外标记预算的约束下运行 *COMET*。图 5.2 展示了一个代表性的生成社区 (为便于可视化, 省略了大多数真阴性节点), 可以看到 *COMET* 能够在高度噪声和标签稀缺的现实场景下有效聚集潜在刷评账号, 体现了其在在线虚假评论检测等实际应用中的潜力。

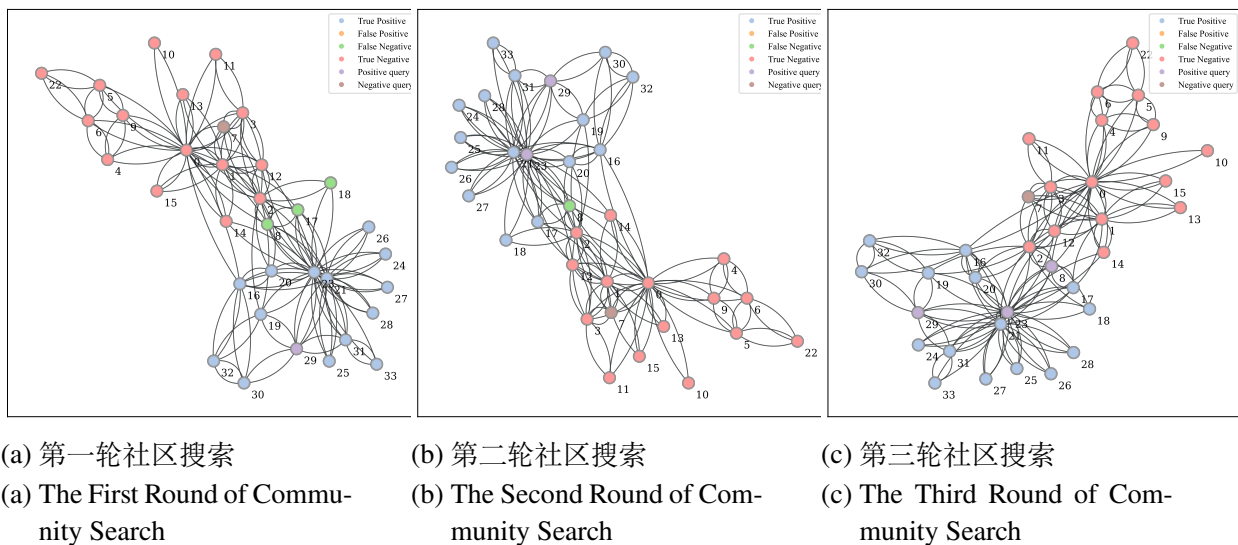


图 5.1 Karate 数据集中的社区搜索
Figure 5.1 Community Search in the Karate Club Dataset

5.3 本章小结

本章重点探讨了 *COMET* 的可扩展性与实际应用能力。首先, 本章论证了将框架从无向图迁移至有向图的可行性, 详细阐述了图嵌入、PPR 子图构建及 GNN 聚合机制在有向场景下的适配方案。在 EU-Core 数据集上的实验表明, 有向版本的 *COMET* 依然能够保持具有竞争力的性能, 且对组件的无向化近似表现出一定的鲁棒性。其次, 通过 Karate 数据集上的全过程可视化, 本章清晰展示了“人机交互—模型更新—主动推荐”这一闭环机制如何驱动社区搜索质量的逐步提升。最后, 在 Amazon-CN 数据集上的应用案例进一步证实了该框架在虚假评论检测等现实世界异常挖掘任务中的有效性与泛化潜力。

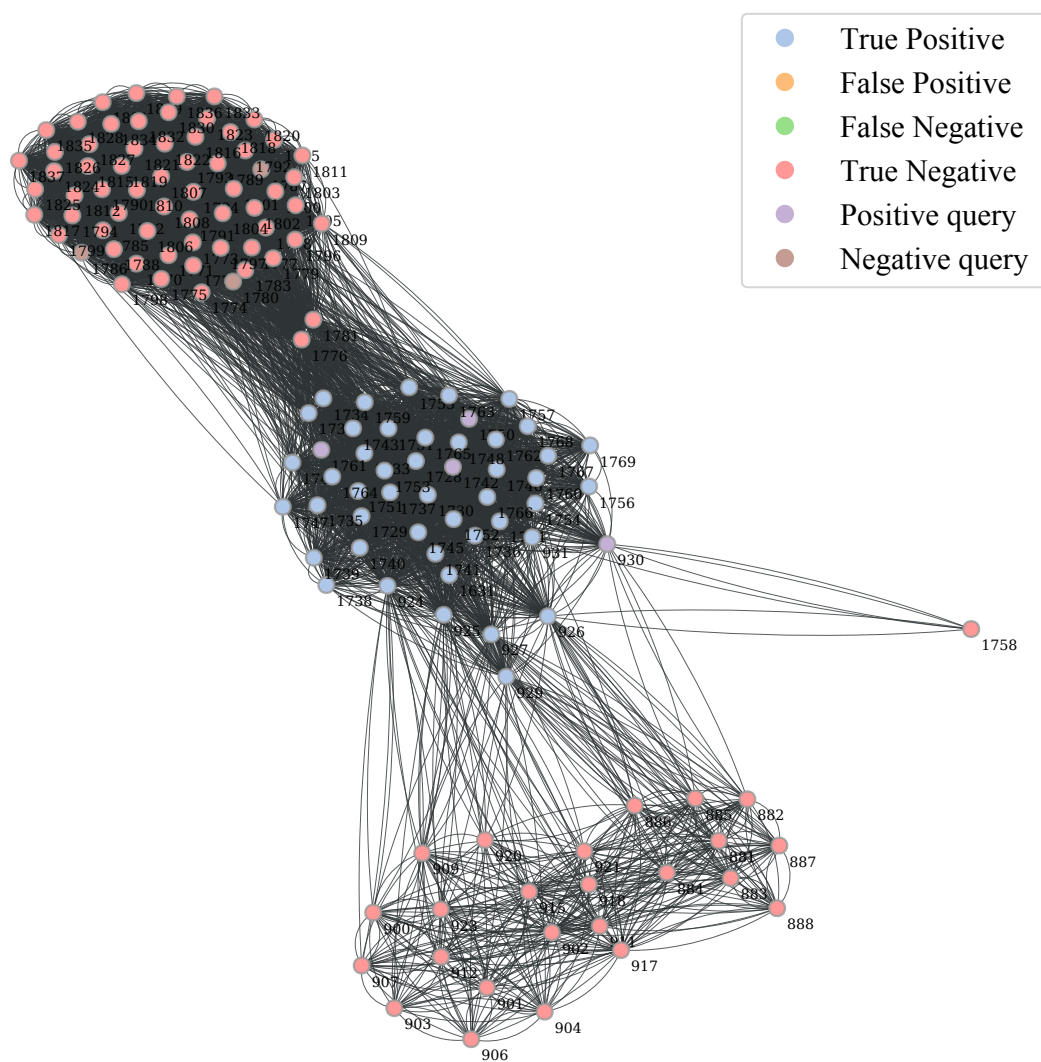


图 5.2 用于欺诈检测的社区搜索
Figure 5.2 Community Search for Fraud Detection

第 6 章 全文结论

6.1 主要结论

COMET 框架的核心贡献在于将社区感知的子图挖掘与主动学习驱动的交互统一纳入一个端到端的流程。该模型首先基于 PPR 提取社区感知的候选子图，该策略利用 PPR 向量中的“急剧下降”点来识别社区的“天然边界”，从而在兼顾查询语义与拓扑结构的同时，显著压缩搜索空间。

随后，*COMET* 以多轮的方式，在候选子图上利用图神经网络预测社区归属。为了解决交互效率低下的问题，本文创新性地引入了主动学习阶段。该阶段设计了一个双重标准 (dual-criteria) 的节点选择目标，主动向用户推荐那些模型信息最丰富（基于 PPR 拓扑重要性）且最不确定（基于 GNN 预测熵）的节点进行标记。

本文在八个真实世界的基准数据集上进行了广泛的实验。结果充分表明，与现有的基线方法相比，*COMET* 在鲁棒性、效率和可扩展性方面均表现出显著优势。通过利用这种动态和交互式识别查询相关社区的能力，*COMET* 为用户提供了一个更高效、更准确的图数据探索工具，有助于加快其下游应用的决策过程。

6.2 应用价值

本文提出的 *COMET* 框架以及交互式社区搜索技术，在多个关键领域具有广泛的应用前景和实践价值：

社交网络分析 在复杂的社交网络中，社区往往对应着兴趣小组或社会圈子。*COMET* 框架可以帮助用户或运营者即时识别以特定兴趣或活动为中心的社区。具体应用包括：通过社交网络进行有影响力的主题搜索^[140]、检测谣言传播的“小团体”^[141]、通过关键节点检测^[142] 或节点影响最大化^[143-144] 来识别受影响的社区。此外，鉴于机器人对社交网络用户行为的重大影响^[145-146]，本框架可用于根据特定查询探索机器人社区及其影响。

欺诈检测与金融风控 在金融交易网络或电子商务评论网络中，欺诈团伙往往形成异常紧密的“小团体”^[52,147]。交互式社区搜索 (ICS) 可以从已知的可疑账户（种子节点）出发，快速识别相关的可疑活动集群，使欺诈检测团队能集中精力处理高风险区

域。*COMET* 所采用的主动学习策略尤为关键，它能迭代地挖掘并标记那些“潜在但高度不确定”的边界节点，从而提高欺诈团伙识别的准确性和速度，减少潜在损失。

隐私保护的社区生成 在社区生成过程中确保隐私至关重要^[148-149]。*COMET* 所采用的 PPR 子图构建阶段，天然具有隐私保护的潜力。因为它首先过滤掉了大量与查询不相关的节点，后续的 GNN 分析和主动学习仅在高度相关的候选子图上进行，这意味着模型仅需获取与查询强相关的节点属性，从而避免了对全图敏感数据的非必要访问，有助于实现更安全、更保密的分析。

6.3 未来研究展望

尽管 *COMET* 框架在实验中取得了良好效果，但通过 ICS 揭示查询相关的社区仍是一个充满挑战的领域。目前还没有一个统一的模型可以无需特定调整就能直接处理所有类型的图，因为每个图都包含其自己独特的语义和动态，因此本文选择一种高效剪枝与人在回路的方式对这个问题建模。但预训练模型带来的对世界知识的理解一定能对这个领域有新的启发。结合近期的研究趋势，本文认为社区搜索的未来研究可以沿着以下极具前景的方向深度发展：

融合 Agentic LLM 的智能交互范式 传统的交互式搜索（包括 *COMET*）主要依赖于“是/否”的主动学习反馈。然而，2024-2025 年 Agentic LLM（即具有代理能力的大型语言模型）的兴起，基于智能体的检索与推荐系统开始变得流行^[150]，这为构建更智能的交互系统提供了全新思路。未来的研究可以探索将 *COMET* 作为一个“工具”或“能力”，被一个大型语言模型代理所调用。用户可以通过自然语言提出模糊的探索性请求（例如：“帮我分析一下，最近讨论我们产品的开发者中，哪些人是高度相关的，但又不属于我们已知的合作伙伴？”）。LLM 代理负责理解意图，并自动将自然语言查询分解为 *COMET* 所需的 $Q+$ （正查询节点）和 $Q-$ （负查询节点），实现真正的对话式、意图驱动在社区搜索。

面向多模态与异构数据的 GraphRAG 融合 *COMET* 主要在结构化的图上运行。但未来的数据本质上是多模态的（Multimodal）^[151]。一个关键趋势是 GraphRAG^[152]，它首先从非结构化文本（如公司财报、技术文档）中构建知识图谱，然后应用社区检测来生成主题摘要。未来的工作可以将 *COMET* 的交互性与 GraphRAG 相结合。例如，

系统可以先从海量文档中构建一个大型异构图^[153-154]，当用户对某个主题进行 RAG 查询时，*COMET* 不再是进行一次性的静态社区划分，而是允许用户交互式地探索和修正这个从文本中生成的“知识社群”，从而获得更精确、更可信的上下文 (Context)，最终显著提升 RAG 的答案质量。

协同图异常检测与主动防御 *COMET* 的应用价值中提到了欺诈检测。近期的研究热点是利用 GNN 进行图异常检测 (GAD)^[155]。特别是在 2025 年的研究中，如何处理标签极其有限的欺诈检测已成为关键挑战^[156]。这为 *COMET* 这样的主动学习框架提供了绝佳的切入点。未来的研究不应局限于被动的欺诈社区搜索，而应探索主动防御：系统可以利用 *COMET* 的不确定性采样模块，自动识别那些高度可疑但不易被自动模型察觉的边界账户，并主动将其推送给人类专家进行标注。这种人机协同的主动挖掘模式，对于发现新型、隐蔽的欺诈团伙具有不可估量的价值。

基于图基础模型 (Graph Foundation Model) 的统一框架 *COMET* 与许多 GNN 模型一样，依赖于在特定任务和数据集上的即时训练。然而，图学习领域正面临基准和评估协议碎片化的问题^[157-159]。未来的趋势是构建大规模的图基础模型 (Graph Foundation Models) 或使用大规模预训练^[160]。*COMET* 的未来演进可以考虑将其 GNN 模块替换为一个在海量图数据（如 OGB 等大型基准）与其他模态数据 continue pretraining 的图 Transformer。*COMET* 的交互式主动学习则转变为一种高效的 Parameter-Efficient Fine-Tuning (PEFT) 策略^[161]，用于在极少的用户反馈下，快速将强大的图基础模型适配到特定用户的个性化社区搜索任务上，从而在实现 SOTA 性能的同时，极大提升模型的泛化能力和部署效率。

6.4 总结

社区搜索作为“带查询约束的社区发现”问题，近年来取得了长足的发展。研究范式从早期基于 k-core、k-truss 等刚性结构约束的算法，逐步演进到融合图嵌入和深度学习的数据驱动方法。*COMET* 框架的探索证明，将高效的拓扑先验（如 PPR）与灵活的深度学习（GNN）以及人在回路（Active Learning）相结合，是平衡社区搜索效率、准确性与个性化需求的关键路径。

面对真实世界中日益增长的数据规模和异质性，挑战依然存在。有理由相信，随着图机器学习、大型语言模型和多模态技术的持续演进和跨领域融合，社区搜索技术



将在社交网络分析、推荐系统、金融反欺诈和知识图谱探索等诸多领域发挥愈加重要的作用，为揭示复杂网络的组织结构和动力学机制提供更深入的洞见。

参考文献

- [1] LESKOVEC J, HUTTENLOCHER D, KLEINBERG J. Predicting positive and negative links in online social networks[C]// Proceedings of the 19th international conference on World wide web. 2010: 641-650.
- [2] FAN W, MA Y, LI Q, et al. Graph neural networks for social recommendation[C]// The world wide web conference. 2019: 417-426.
- [3] WANG D, LIN J, CUI P, et al. A semi-supervised graph attentive network for financial fraud detection[C]// 2019 IEEE International Conference on Data Mining (ICDM). 2019: 598-607.
- [4] ZENG M, ZHANG F, WU F X, et al. Protein-protein interaction site prediction through combining local and global features with deep neural networks[J]. Bioinformatics, 2020, 36(4): 1114-1120.
- [5] HUANG Z, ZENG D D, CHEN H. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems[J]. Management science, 2007, 53(7): 1146-1164.
- [6] CHEN T, TSOURAKAKIS C. AntiBenford Subgraphs: Unsupervised Anomaly Detection in Financial Networks[C]// KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington DC, USA, 2022: 2762-2770.
- [7] MOGHADDASS R, GUAN Y. Optimal Frameworks for Detecting Anomalies in Sensor-Intensive Heterogeneous Networks[J]. INFORMS Journal on Computing, 2022, 34(5): 2583-2610.
- [8] RASCHID L, SAYYADI H, HRISTIDIS V. Learning to Rank in Entity Relationship Graphs[J]. INFORMS Journal on Computing, 2019, 31(4): 671-688.
- [9] TIMONINA-FARKAS A, SEIFERT R W. Information Retrieval Under Network Uncertainty: Robust Internet Ranking[J]. Operations Research, 2023, 71(6): 2328-2351.
- [10] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs[J]. The Bell system technical journal, 1970, 49(2): 291-307.
- [11] GIRVAN M, NEWMAN M E. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.
- [12] 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.
- [13] SOZIO M, GIONIS A. The community-search problem and how to plan a successful cocktail party [C]// KDD '10: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 939-948.
- [14] CUI W, XIAO Y, WANG H, et al. Local search of communities in large graphs[C]// SIGMOD '14: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data.

- Snowbird, Utah, USA, 2014: 991-1002.
- [15] HUANG X, CHENG H, QIN L, et al. Querying k-truss community in large and dynamic graphs [C]//SIGMOD '14: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. Snowbird, Utah, USA, 2014: 1311-1322.
 - [16] HUANG X, LAKSHMANAN L V S, YU J X, et al. Approximate closest community search in networks[J]. Proc. VLDB Endow., 2015, 9(4): 276-287.
 - [17] AKBAS E, ZHAO P. Truss-based community search: a truss-equivalence based indexing approach[J]. Proc. VLDB Endow., 2017, 10(11): 1298-1309.
 - [18] LUO F, WANG J Z, PROMISLOW E. Exploring local community structures in large networks [J]. Web Intelligence and Agent Systems: An International Journal, 2008, 6(4): 387-400.
 - [19] KLOUMANN I M, KLEINBERG J M. Community membership identification from small seed sets[C]//KDD '14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York, USA, 2014: 1366-1375.
 - [20] JIANG Y, RONG Y, CHENG H, et al. Query driven-graph neural networks for community search: from non-attributed, attributed, to interactive attributed[J]. Proc. VLDB Endow., 2022, 15(6): 1243-1255.
 - [21] LI L, LUO S, ZHAO Y, et al. COCLEP: Contrastive Learning-based Semi-Supervised Community Search[J]. IEEE 39th ICDE, 2023.
 - [22] GAO J, CHEN J, LI Z, et al. ICS-GNN: lightweight interactive community search via graph neural network[J]. Proc. VLDB Endow., 2021, 14(6): 1006-1018.
 - [23] KARYPIS G, KUMAR V. METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices[J]. 1997.
 - [24] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: Bringing order to the web [J]. 1999.
 - [25] FORTUNATO S. Community detection in graphs[J]. Physics reports, 2010, 486(3-5): 75-174.
 - [26] NEWMAN M E, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113.
 - [27] FANG Y, HUANG X, QIN L, et al. A survey of community search over big graphs[J]. The VLDB Journal, 2020, 29: 353-392.
 - [28] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2020, 32(1): 4-24.
 - [29] ZHOU J, CUI G, HU S, et al. Graph neural networks: A review of methods and applications[J]. AI open, 2020, 1: 57-81.
 - [30] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv preprint arXiv:1312.6203, 2013.
 - [31] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks

- [C]//5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. 2017.
- [32] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[J]. Advances in neural information processing systems, 2017, 30.
- [33] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph Attention Networks[C]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. 2018.
- [34] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for quantum chemistry[C]//International conference on machine learning. 2017: 1263-1272.
- [35] XU K, HU W, LESKOVEC J, et al. How Powerful are Graph Neural Networks?[C]//7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. 2019.
- [36] LI Q, HAN Z, WU X M. Deeper insights into graph convolutional networks for semi-supervised learning[C]//Proceedings of the AAAI conference on artificial intelligence: vol. 32: 1. 2018.
- [37] LI G, MULLER M, THABET A, et al. DeepGCNs: Can GCNs go as deep as CNNs?[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9267-9276.
- [38] RONG Y, HUANG W, XU T, et al. Dropedge: Towards deep graph convolutional networks on node classification[J]. arXiv preprint arXiv:1907.10903, 2019.
- [39] CHEN J, ZHU J, SONG L. Stochastic training of graph convolutional networks with variance reduction[J]. arXiv preprint arXiv:1710.10568, 2017.
- [40] ZENG H, ZHOU H, SRIVASTAVA A, et al. Graphsaint: Graph sampling based inductive learning method[J]. arXiv preprint arXiv:1907.04931, 2019.
- [41] CHIANG W L, LIU X, SI S, et al. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 257-266.
- [42] KARYPIS G, KUMAR V. Multilevelk-way partitioning scheme for irregular graphs[J]. Journal of Parallel and Distributed computing, 1998, 48(1): 96-129.
- [43] KIPF T N, WELING M. Variational graph auto-encoders[J]. arXiv preprint arXiv:1611.07308, 2016.
- [44] VELIČKOVIĆ P, FEDUS W, HAMILTON W L, et al. Deep graph infomax[J]. arXiv preprint arXiv:1809.10341, 2018.
- [45] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. 2020: 1597-1607.
- [46] ZHU Y, XU Y, YU F, et al. Deep graph contrastive representation learning[J]. arXiv preprint arXiv:2006.04131, 2020.
- [47] ZHU Y, XU Y, YU F, et al. Graph contrastive learning with adaptive augmentation[C]//

- Proceedings of the web conference 2021. 2021: 2069-2080.
- [48] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//European semantic web conference. 2018: 593-607.
- [49] WANG X, JI H, SHI C, et al. Heterogeneous graph attention network[C]//The world wide web conference. 2019: 2022-2032.
- [50] SEO Y, DEFFERRARD M, VANDERGHEYNST P, et al. Structured sequence modeling with graph convolutional recurrent networks[C]//International conference on neural information processing. 2018: 362-373.
- [51] ROSSI E, CHAMBERLAIN B, FRASCA F, et al. Temporal graph networks for deep learning on dynamic graphs[J]. arXiv preprint arXiv:2006.10637, 2020.
- [52] DOU Y, LIU Z, SUN L, et al. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters[C]//Proceedings of the 29th ACM international conference on information & knowledge management. 2020: 315-324.
- [53] HU A L, CHAN K C. Utilizing both topological and attribute information for protein complex identification in PPI networks[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2013, 10(3): 780-792.
- [54] XIE J, KELLEY S, SZYMANSKI B K. Overlapping community detection in networks: The state-of-the-art and comparative study[J]. Acm computing surveys (csur), 2013, 45(4): 1-35.
- [55] JAVED M A, YOUNIS M S, LATIF S, et al. Community detection in networks: A multidisciplinary review[J]. Journal of Network and Computer Applications, 2018, 108: 87-111.
- [56] SU X, XUE S, LIU F, et al. A comprehensive survey on community detection with deep learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.
- [57] MCPHERSON M, SMITH-LOVIN L, COOK J M. Birds of a feather: Homophily in social networks[J]. Annual review of sociology, 2001, 27(1): 415-444.
- [58] AMINI A A, CHEN A, BICKEL P J, et al. Pseudo-likelihood methods for community detection in large sparse networks[J]. The Annals of Statistics, 2013, 41(4): 2097-2122.
- [59] HOLLAND P W, LASKEY K B, LEINHARDT S. Stochastic blockmodels: First steps[J]. Social networks, 1983, 5(2): 109-137.
- [60] KARRER B, NEWMAN M E. Stochastic blockmodels and community structure in networks[J]. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, 2011, 83(1): 016107.
- [61] AIROLDI E M, BLEI D, FIENBERG S, et al. Mixed membership stochastic blockmodels[J]. Advances in neural information processing systems, 2008, 21.
- [62] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, 2007, 76(3): 036106.
- [63] ROSVALL M, BERGSTROM C T. Maps of random walks on complex networks reveal commu-

- nity structure[J]. Proceedings of the national academy of sciences, 2008, 105(4): 1118-1123.
- [64] NEWMAN M E. Modularity and community structure in networks[J]. Proceedings of the national academy of sciences, 2006, 103(23): 8577-8582.
- [65] LIU F, XUE S, WU J, et al. Deep learning for community detection: progress, challenges and opportunities[J]. arXiv preprint arXiv:2005.08225, 2020.
- [66] JIN D, LIU Z, LI W, et al. Graph convolutional networks meet markov random fields: Semi-supervised community detection in attribute networks[C]//Proceedings of the AAAI conference on artificial intelligence: vol. 33: 01. 2019: 152-159.
- [67] PARK C, KIM D, HAN J, et al. Unsupervised attributed multiplex network embedding[C]//Proceedings of the AAAI conference on artificial intelligence: vol. 34: 04. 2020: 5371-5378.
- [68] YANG L, CAO X, HE D, et al. Modularity based community detection with deep learning[C]//IJCAI'16: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, New York, USA: AAAI Press, 2016: 2252-2258.
- [69] ZHANG X, LIU H, LI Q, et al. Attributed graph clustering via adaptive graph convolution[C]//IJCAI'19: Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: AAAI Press, 2019: 4327-4333.
- [70] JING B, PARK C, TONG H. Hdmi: High-order deep multiplex infomax[C]//Proceedings of the web conference 2021. 2021: 2414-2424.
- [71] GAO H, PEI J, HUANG H. Progan: Network embedding via proximity generative adversarial network[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 1308-1316.
- [72] LI J, WANG X, DENG K, et al. Most influential community search over large social networks[C]//2017 IEEE 33rd international conference on data engineering (ICDE). 2017: 871-882.
- [73] SPERLÍ G. A deep learning based community detection approach[C]//Proceedings of the 34th ACM/SIGAPP symposium on applied computing. 2019: 1107-1110.
- [74] DU L, LU Z, WANG Y, et al. Galaxy network embedding: a hierarchical community structure preserving approach.[C]//IJCAI. 2018: 2079-2085.
- [75] FU X, ZHANG J, MENG Z, et al. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding[C]//Proceedings of the web conference 2020. 2020: 2331-2341.
- [76] WANG Z, WANG C, GAO C, et al. An evolutionary autoencoder for dynamic community detection[J]. Science China Information Sciences, 2020, 63(11): 212205.
- [77] CHEN Z, LI L, BRUNA J. Supervised Community Detection with Line Graph Neural Networks [C]//7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. 2019.
- [78] SHCHUR O, GÜNNEMANN S. Overlapping community detection with graph neural networks [J]. arXiv preprint arXiv:1909.12201, 2019.

- [79] HU R, PAN S, LONG G, et al. Going deep: Graph convolutional ladder-shape networks[C]// Proceedings of the AAAI conference on artificial intelligence: vol. 34: 03. 2020: 2838-2845.
- [80] LEVIE R, MONTI F, BRESSON X, et al. Cayleynets: Graph convolutional neural networks with complex rational spectral filters[J]. IEEE Transactions on Signal Processing, 2018, 67(1): 97-109.
- [81] CUI G, ZHOU J, YANG C, et al. Adaptive graph encoder for attributed graph embedding[C]// Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020: 976-985.
- [82] ZHANG X, LIU H, WU X M, et al. Spectral embedding network for attributed graph clustering [J]. Neural Networks, 2021, 142: 388-396.
- [83] ZHANG T, XIONG Y, ZHANG J, et al. CommDGI: community detection oriented deep graph infomax[C]// Proceedings of the 29th ACM international conference on information & knowledge management. 2020: 1843-1852.
- [84] WANG X, LIU N, HAN H, et al. Self-supervised heterogeneous graph neural network with co-contrastive learning[C]// Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. 2021: 1726-1736.
- [85] LUO L, FANG Y, CAO X, et al. Detecting communities from heterogeneous graphs: A context path-based graph neural network model[C]// Proceedings of the 30th ACM international conference on information & knowledge management. 2021: 1170-1180.
- [86] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
- [87] JIA Y, ZHANG Q, ZHANG W, et al. Communitygan: Community detection with generative adversarial nets[C]// The world wide web conference. 2019: 784-794.
- [88] ZHANG Y, XIONG Y, YE Y, et al. SEAL: Learning heuristics for community detection with generative adversarial networks[C]// Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020: 1103-1113.
- [89] SHI M, TANG Y, ZHU X, et al. Multi-class imbalanced graph convolutional network learning[C]// Proceedings of the twenty-ninth international joint conference on artificial intelligence (IJCAI-20). 2020.
- [90] WANG J, CAO J, LI W, et al. CANE: community-aware network embedding via adversarial training[J]. Knowledge and Information Systems, 2021, 63(2): 411-438.
- [91] CAVALLARI S, ZHENG V W, CAI H, et al. Learning community embedding with community detection and node embedding on graphs[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 377-386.
- [92] LI T, LEI L, BHATTACHARYYA S, et al. Hierarchical community detection by recursive partitioning[J]. Journal of the American Statistical Association, 2022, 117(538): 951-968.
- [93] HUANG X, CHEN D, REN T, et al. A survey of community detection methods in multilayer

- networks[J]. Data Mining and Knowledge Discovery, 2021, 35(1): 1-45.
- [94] FAN S, WANG X, SHI C, et al. One2multi graph autoencoder for multi-view graph clustering[C] // proceedings of the web conference 2020. 2020: 3070-3076.
- [95] LI R, ZHANG C, HU Q, et al. Flexible multi-view representation learning for subspace clustering. [C] // Ijcai: vol. 2019. 2019: 2916-2922.
- [96] ZHU J, YAN Y, ZHAO L, et al. Beyond homophily in graph neural networks: Current limitations and effective designs[J]. Advances in neural information processing systems, 2020, 33: 7793-7804.
- [97] ROSSETTI G, CAZABET R. Community discovery in dynamic networks: a survey[J]. ACM computing surveys (CSUR), 2018, 51(2): 1-37.
- [98] KHAOUID W, BARSKY M, SRINIVASAN V, et al. K-core decomposition of large networks on a single PC[J]. Proc. VLDB Endow., 2015, 9(1): 13-23.
- [99] TSOURAKAKIS C. The k-clique densest subgraph problem[C] // Proceedings of the 24th international conference on world wide web. 2015: 1122-1132.
- [100] AKIBA T, IWATA Y, YOSHIDA Y. Linear-time enumeration of maximal k-edge-connected subgraphs in large networks by random contraction[C] // Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 909-918.
- [101] CHANG L, LIN X, QIN L, et al. Index-based Optimal Algorithms for Computing Steiner Components with Maximum Connectivity[C] // SIGMOD '15: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne, Victoria, Australia, 2015: 459-474.
- [102] ANDERSEN R, CHUNG F, LANG K. Local graph partitioning using pagerank vectors[C] // 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). 2006: 475-486.
- [103] ANDERSEN R, CHUNG F. Detecting sharp drops in PageRank and a simplified local partitioning algorithm[C] // International Conference on Theory and Applications of Models of Computation. 2007: 1-12.
- [104] ANDERSEN R, CHUNG F, LANG K. Local Partitioning for Directed Graphs Using PageRank[C] // BONATO A, CHUNG F R K. Algorithms and Models for the Web-Graph. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 166-178.
- [105] SPIELMAN D A, TENG S H. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning[J]. SIAM Journal on computing, 2013, 42(1): 1-26.
- [106] YIN H, BENSON A R, LESKOVEC J, et al. Local higher-order graph clustering[C] // Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017: 555-564.
- [107] HAVELIWALA T H, KAMVAR S D. The second eigenvalue of the Google matrix[R]. Citeseer, 2003.

- [108] FOGARAS D, RÁCZ B, CSALOGÁNY K, et al. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments[J]. *Internet Mathematics*, 2005, 2(3): 333-358.
- [109] YANG M, WANG H, WEI Z, et al. Efficient Algorithms for Personalized PageRank Computation: A Survey[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2024, 36(9): 4582-4602.
- [110] FANG Y, CHENG R, LUO S, et al. Effective community search for large attributed graphs[J]. *Proc. VLDB Endow.*, 2016, 9(12): 1233-1244.
- [111] HUANG X, LAKSHMANAN L V S. Attribute-driven community search[J]. *Proc. VLDB Endow.*, 2017, 10(9): 949-960.
- [112] KIM J, LUO S, CONG G, et al. DMCS: Density Modularity based Community Search[C] // *SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data*. Philadelphia, PA, USA, 2022: 889-903.
- [113] WANG Y, JIAN X, YANG Z, et al. Query optimal k-plex based community in graphs[J]. *Data Science and Engineering*, 2017, 2: 257-273.
- [114] WANG J, WANG K, LIN X, et al. Efficient Unsupervised Community Search with Pre-trained Graph Transformer[J]. *arXiv preprint arXiv:2403.18869*, 2024.
- [115] WANG J, WANG K, LIN X, et al. Neural Attributed Community Search at Billion Scale[J]. *Proceedings of the ACM on Management of Data*, 2024, 1(4): 1-25.
- [116] FANG S, ZHAO K, LI G, et al. Community search: a meta-learning approach[C] // *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. 2023: 2358-2371.
- [117] REN P, XIAO Y, CHANG X, et al. A survey of deep active learning[J]. *ACM computing surveys (CSUR)*, 2021, 54(9): 1-40.
- [118] ASH J T, ZHANG C, KRISHNAMURTHY A, et al. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds[C] // *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 2020.
- [119] SENER O, SAVARESE S. Active Learning for Convolutional Neural Networks: A Core-Set Approach[C] // *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. 2018.
- [120] ZHOU J, WANG K, WANG J, et al. COMET: An interactive framework for efficient and effective community search via active learning[J]. *INFORMS Journal on Computing*, 2025.
- [121] YANG J, LESKOVEC J. Defining and evaluating network communities based on ground-truth [C] // *Proceedings of the ACM SIGKDD workshop on mining data semantics*. 2012: 1-8.
- [122] BERKHIN P. Bookmark-coloring algorithm for personalized pagerank computing[J]. *Internet Mathematics*, 2006, 3(1): 41-62.
- [123] YANG M, WANG H, WEI Z, et al. Efficient algorithms for personalized pagerank computation: A survey[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2024, 36(9): 4582-4602.
- [124] GROVER A, LESKOVEC J. node2vec: Scalable Feature Learning for Networks[C] // *KDD '16*:

- Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, 2016: 855-864.
- [125] LESKOVEC J, KREVL A. SNAP Datasets: Stanford Large Network Dataset Collection[Z]. <http://snap.stanford.edu/data>. 2014.
- [126] ZACHARY W W. An information flow model for conflict and fission in small groups[J]. Journal of anthropological research, 1977, 33(4): 452-473.
- [127] LUSSEAU D, SCHNEIDER K, BOISSEAU O J, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait?[J]. Behavioral Ecology and Sociobiology, 2003, 54: 396-405.
- [128] LESKOVEC J, KLEINBERG J, FALOUTSOS C. Graph evolution: Densification and shrinking diameters[J]. ACM transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 2-es.
- [129] WANG H, WEI Z, YUAN Y, et al. Exact Single-Source SimRank Computation on Large Graphs [C]//SIGMOD ' 20: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. Portland, OR, USA, 2020: 653-663.
- [130] JE H G, WIDOM J. Simrank: a measure of structural-context similarity[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002: 538-543.
- [131] ZHANG J, TANG J, MA C, et al. Panther: Fast top-k similarity search on large networks[C]// Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015: 1445-1454.
- [132] YANG D, ROSSO P, LI B, et al. Nodesketch: Highly-efficient graph embeddings via recursive sketching[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 1162-1172.
- [133] CHAMBERLAIN B P, SHIROBOKOV S, ROSSI E, et al. Graph neural networks for link prediction with subgraph sketching[J]. arXiv preprint arXiv:2209.15486, 2022.
- [134] CAI H, ZHENG V W, CHANG K C C. Active learning for graph embedding[J]. arXiv preprint arXiv:1705.05085, 2017.
- [135] ZHANG Y, LEASE M, WALLACE B. Active discriminative text representation learning[C]// Proceedings of the AAAI conference on artificial intelligence: vol. 31: 1. 2017.
- [136] CONDON A, KARP R M. Algorithms for graph partitioning on the planted partition model[J]. Random Structures & Algorithms, 2001, 18(2): 116-140.
- [137] LANCICHINETTI A, FORTUNATO S, RADICCHI F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, 2008, 78(4): 046110.
- [138] YANG Z, ALGESHEIMER R, TESSONE C J. A comparative analysis of community detection algorithms on artificial networks[J]. Scientific reports, 2016, 6(1): 30750.

- [139] WU Z, LIU G, WU J, et al. Are neighbors alike? A semisupervised probabilistic collaborative learning model for online review spammers detection[J]. *Information Systems Research*, 2024, 35(4): 1565-1585.
- [140] LI J, LIU C, YU J X, et al. Personalized Influential Topic Search via Social Network Summarization[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(7): 1820-1834.
- [141] HE Z, LI C, ZHOU F, et al. Rumor Detection on Social Media with Event Augmentations[C] // SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event, Canada, 2021: 2020-2024.
- [142] ZHOU Y, LI J, HAO J K, et al. Detecting critical nodes in sparse graphs via “reduce-solve-combine” memetic search[J]. *INFORMS Journal on Computing*, 2024, 36(1): 39-60.
- [143] GÜNNEÇ D, RAGHAVAN S, ZHANG R. Least-cost influence maximization on social networks [J]. *INFORMS Journal on Computing*, 2020, 32(2): 289-302.
- [144] RAGHAVAN S, ZHANG R. Influence maximization with latency requirements on social networks [J]. *INFORMS Journal on Computing*, 2022, 34(2): 710-728.
- [145] Des MESNARDS N G, HUNTER D S, el HJOUJI Z, et al. Detecting bots and assessing their impact in social networks[J]. *Operations research*, 2022, 70(1): 1-22.
- [146] FENG S, TAN Z, WAN H, et al. Twibot-22: Towards graph-based twitter bot detection[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 35254-35269.
- [147] VAN VLASSELAER V, ELIASSI-RAD T, AKOGLU L, et al. Gotcha! network-based fraud detection for social security fraud[J]. *Management Science*, 2017, 63(9): 3090-3110.
- [148] ZHENG X, CAI Z, LUO G, et al. Privacy-preserved community discovery in online social networks[J]. *Future Generation Computer Systems*, 2019, 93: 1002-1009.
- [149] GUO T, PENG S, LI Y, et al. Community-based social recommendation under local differential privacy protection[J]. *Information Sciences*, 2023, 639: 119002.
- [150] ZHANG Y, QIAO S, ZHANG J, et al. A survey of large language model empowered agents for recommendation and search: Towards next-generation information retrieval[J]. *arXiv preprint arXiv:2503.05659*, 2025.
- [151] PENG C, HE J, XIA F. Learning on multimodal graphs: A survey[J]. *arXiv preprint arXiv:2402.05322*, 2024.
- [152] EDGE D, TRINH H, CHENG N, et al. From local to global: A graph rag approach to query-focused summarization[J]. *arXiv preprint arXiv:2404.16130*, 2024.
- [153] LIU J, YANG C, LU Z, et al. Graph foundation models: Concepts, opportunities and challenges [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [154] ZHAO Z, LUO X, CHEN M, et al. A Survey of Knowledge Graph Construction Using Machine Learning.[J]. *CMES-Computer Modeling in Engineering & Sciences*, 2024, 139(1).
- [155] QIAO H, TONG H, AN B, et al. Deep graph anomaly detection: A survey and new perspectives

- [J]. IEEE Transactions on Knowledge and Data Engineering, 2025.
- [156] SONG Y, WEI Y, LU Y, et al. Mitigating Message Imbalance in Fraud Detection with Dual-View Graph Representation Learning[J]. arXiv preprint arXiv:2507.06469, 2025.
- [157] HU W, FEY M, ZITNIK M, et al. Open graph benchmark: Datasets for machine learning on graphs[J]. Advances in neural information processing systems, 2020, 33: 22118-22133.
- [158] DWIVEDI V P, JOSHI C K, LUU A T, et al. Benchmarking graph neural networks[J]. Journal of Machine Learning Research, 2023, 24(43): 1-48.
- [159] BECHLER-SPEICHER M, FINKELSHTEIN B, FRASCA F, et al. Position: Graph learning will lose relevance due to poor benchmarks[J]. arXiv preprint arXiv:2502.14546, 2025.
- [160] WANG Z, LIU Z, MA T, et al. Graph Foundation Models: A Comprehensive Survey[J]. arXiv preprint arXiv:2505.15116, 2025.
- [161] LI S, HAN X, BAI J. Adaptergnn: Parameter-efficient fine-tuning improves generalization in gnns [C]//Proceedings of the AAAI conference on artificial intelligence: vol. 38: 12. 2024: 13600-13608.

致 谢

时光荏苒，硕士阶段的学习与科研旅程即将落下帷幕。回首两年多来的点滴，我深感自己在科研道路上的不断成长，离不开诸多师长、同学与亲友的支持与陪伴。在此，谨向所有给予我关心与帮助的人致以最诚挚的感谢。

首先，我要衷心感谢我的导师王凯老师。在论文选题、研究思路、实验设计以及论文撰写与修改的过程中，王老师始终给予我耐心细致的指导。老师严谨的治学态度、认真负责的工作作风和深厚的学术素养，使我受益良多。正是在王老师的悉心指导下，我逐步加深了对科研工作的理解，也提升了独立思考和解决问题的能力。

同时，我诚挚感谢上海交通大学安泰经济与管理学院数据与商务智能系为我提供的良好科研环境与浓厚学习氛围。感谢一路同行的各位同学，在学术讨论与技术交流中给予我的启发与帮助；也特别感谢身边的好友孟令凯、朱睿诚、周天缘、陈梦实、赵敬业、丁航、张春贤、魏君陶、林一晨、周宸宇等，在学习与生活中给予我的支持与陪伴。在我开展第一篇以 UTD 期刊为目标的研究工作过程中，王建伟师兄给予了宝贵的经验分享与建议，马里兰大学张昆鹏老师也提供了无私的指导与支持，使我受益匪浅。这些经历与帮助将成为我今后科研道路上弥足珍贵的财富。

因对人工智能领域的热爱，我曾有幸在上海人工智能实验室实习，并在此过程中得到付杰老师的悉心指导。付老师对我进入人工智能研究领域起到了重要的引领作用，也使我更加坚定了继续深耕这一方向的决心。此后，我非常荣幸获得陆超超、王翔、孙友邦三位老师的指导，前往复旦大学，并在上海人工智能实验室联合培养，继续攻读计算机科学博士学位。未来，我希望能在大模型后训练、自进化智能体系统、安全可信等方向持续探索，力争有所突破，不断拓展对智能边界的理解。

最后，衷心感谢我的家人，感谢他们始终如一的理解与支持。无论是面对科研的挫折，还是生活的挑战，家人始终是我最强有力的后盾。他们无私的爱与鼓励，是我能够坚持到最后的动力源泉。

硕士阶段的学习虽已告一段落，但求索之路永无止境。AGI 的曙光已然可见，愿未来的我能在这条探索之路上持续前行，为中国的人工智能事业贡献绵薄之力。

学术论文和科研成果目录

学术论文

- [1] Zhou J, Wang K, Wang J, et al. COMET: An interactive framework for efficient and effective community search via active learning[J]. INFORMS Journal on Computing, 2025.

个人简历

基本情况

周家纬，2001 年 08 月生于上海。

教育背景

- 2023 年 09 月至 2026 年 06 月，上海交通大学，硕士研究生，管理科学与工程专业
- 2019 年 09 月至 2023 年 06 月，华东师范大学，本科，数据科学与大数据技术专业

研究兴趣

大模型后训练、自进化智能体系统、图数据挖掘、大模型可解释性。

联系方式

- 地址：上海市华山路 1954 号，200030
- E-mail: davidzjw@sjtu.edu.cn

AN INTERACTIVE COMMUNITY SEARCH FRAMEWORK VIA GRAPH ACTIVE LEARNING

Community search on graph-structured data has become a crucial research problem due to its applications in social network analysis, fraud detection, and bioinformatics. Unlike traditional community detection, which seeks a global and static partitioning of an entire graph, community search is an online, query-driven process that aims to find a single cohesive community relevant to user-provided query nodes in real time. To enhance flexibility and robustness, recent developments have introduced interactive community search, which incorporates a human-in-the-loop learning paradigm. This enables iterative refinement of search results through user feedback, allowing the model to better align with human intent by leveraging labeled examples such as positive and negative nodes. However, despite its promise, interactive community search still faces substantial challenges when deployed on large and heterogeneous graphs.

The first major challenge is the lack of a robust and adaptive subgraph extraction mechanism. Performing online graph neural network inference on billion-scale graphs is computationally infeasible, requiring all practical systems to first extract a smaller, query-relevant subgraph. Existing strategies often rely on fixed-size exploration such as breadth-first search or pre-partitioned subgraphs, both of which fail to adapt to the diversity of real-world communities. This often leads to under-coverage, where relevant nodes are missed, or over-coverage, where irrelevant noise is included. The second challenge lies in inefficient interaction strategies. Because user annotation is costly, interactive frameworks must make every query count. Yet most current methods employ passive interaction schemes, requiring users to manually identify and correct errors without algorithmic guidance. This absence of an active learning mechanism results in slow convergence and excessive annotation effort, limiting scalability and usability in practical settings.

To overcome these challenges, the thesis proposes *COMET*, an interactive community search framework built upon graph active learning. *COMET* integrates adaptive subgraph construction with a model-guided, feedback-driven interaction process. The framework begins with a personalized PageRank-based subgraph mining module that automatically deter-

mines the natural boundary of a query-relevant community. By analyzing the distribution of the personalized PageRank vector, *COMET* adaptively identifies the point where semantic relevance sharply drops, yielding a subgraph that captures both structural cohesion and semantic focus. Within this subgraph, a graph neural network is trained transductively using the user's feedback, treating community search as a series of binary node classification tasks. The network propagates labels from user-provided examples through the subgraph, estimating the probability of membership for each node. To enhance interaction efficiency, *COMET* incorporates an active learning module that recommends the most informative nodes for user labeling. This recommendation is governed by a hybrid scoring function that fuses topological importance derived from personalized PageRank and model uncertainty derived from the entropy of GNN predictions. The scores are dynamically combined using a log-sum-exp formulation with a time-adjusted balance between exploration and exploitation, ensuring that early rounds favor structurally significant nodes and later rounds focus on uncertain boundary cases.

Extensive experiments on eight real-world datasets, spanning small networks such as Karate and Dolphins to massive graphs like Amazon, DBLP, and LiveJournal, demonstrate the strong performance of *COMET*. Compared with traditional and learning-based baselines, *COMET* achieves consistently higher accuracy across standard metrics; its interactive version achieves top-tier F1-scores such as 98.39% on Karate, 100.00% on Dolphins, and 94.32% on Amazon. Its interactive version shows particularly rapid improvement with minimal user effort, reaching substantial performance gains within only a few annotation rounds, such as an F1-score increase of 17.23% on Football and 13.41% on Youtube after three rounds. *COMET* also exhibits remarkable computational efficiency, outperforming existing learning-based approaches in runtime; for instance, it runs approximately 2x faster than ICS-GNN and 5x faster than COCLEP on the LiveJournal dataset. Ablation and sensitivity studies confirm the complementary value of its components: removing either the adaptive subgraphing (causing an F1-score drop of 94.14% on Amazon) or the GNN-based classification (an F1-score drop of 15.32% on Amazon) significantly reduces performance, underscoring the importance of both. The framework's adaptive PPR-based subgraphing outperforms fixed-size methods (e.g., 60.33% F1 on Youtube vs. 20.47% for BFS-200), and its hybrid active learning strategy yields more robust information gain (76.76% F1 on LiveJournal) than single-criterion

(75.69% for entropy-only) or static weighting approaches.

Beyond empirical evaluation, the thesis extends *COMET* to new contexts. The framework can be applied to directed graphs with minimal adjustment, as demonstrated on communication networks such as the EU-Core email dataset. Case studies further highlight its interpretability and practical relevance. On the Karate network, *COMET* reveals community structures that align with known ground truth, while in a real-world fake review detection scenario on Amazon data, it successfully isolates fraudulent clusters guided by limited human feedback. These examples show how *COMET* effectively balances automation with interpretability, supporting domain experts in exploring complex relational data.

In conclusion, *COMET* represents a unified and efficient solution for interactive community search, systematically addressing the two core bottlenecks of robustness and interaction efficiency. By integrating adaptive personalized PageRank-based subgraph construction with an active learning mechanism that dynamically optimizes user engagement, *COMET* achieves state-of-the-art performance across multiple benchmarks. Its scalability and transparency make it well suited for real-time, human-guided exploration of large-scale graphs, offering valuable tools for applications in social network analysis, financial fraud detection, and other domains requiring interpretable, user-centered graph intelligence.

上海交通大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：周家伟

日期：2026年5月15日

上海交通大学 学位论文使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于 ☒ 公开论文

☐ 内部论文，☐ 1 年/☐ 2 年/☐ 3 年 解密后适用本授权书。

☐ 秘密论文，____ 年（不超过 10 年）解密后适用本授权书。

☐ 机密论文，____ 年（不超过 20 年）解密后适用本授权书。

（请在以上方框内打“√”）

学位论文作者签名：周家伟

日期：2026年5月15日

指导教师签名：王仁

日期：2026年5月18日



上海交通大学硕士学位论文答辩决议书

姓名	周家纬	学号	123120910040	所在学科	管理科学与工程(120100)
指导教师	王凯	答辩日期	2026-05-13	答辩地点	A407
论文题目	基于主动学习框架的交互式图社区搜索研究				
投票表决结果: 3 / 3 / 3 (同意票数/实到委员数/应到委员数)					
答辩结论: ☒ 通过 ☐ 未通过					
评语和决议: 论文针对现有的学习型交互式社区搜索方法在应对大规模图上的复杂查询任务时面临的挑战,提出一种面向大规模图的交互式社区搜索框架,提高了社区搜索的准确性和效率。 论文思路清晰、结构合理、研究方法科学,表明作者掌握了本门学科基础理论,拥有一定的专业知识,具有独立从事科学研究的能力。该生在答辩中能理解和回答老师的提问。经答辩委员会无记名投票一致通过其学位论文答辩,并建议授予周家纬同学管理科学与工程硕士学位。 2026年5月13日					
答辩委员会 工作职务	姓名	职称	工作单位	签名	
主席	徐丽群	教授	安泰经济与管理学院	徐丽群	
委员	徐海峰	副教授	安泰经济与管理学院	徐海峰	
委员	江浦平	副教授	安泰经济与管理学院	江浦平	
秘书	闫飞扬		安泰经济与管理学院	闫飞扬	