

庞可涵

求职方向：算法研究员 | 数据治理 / 机器学习 / 大模型

☎ +86 132 6451 8388 ✉ pangkehan@buaa.edu.cn 🏠 KehanPang.github.io

🎓 教育经历

北京航空航天大学 计算机学院 软件工程 博士研究生 2022.09 – 2027.07(预计)

- 导师：樊文飞 院士、陆平 副教授
- 研究方向：图数据质量、图数据挖掘、图模型可靠性

北京邮电大学 计算机学院 计算机科学与技术 工学学士 2018.09 – 2022.07

- GPA: 3.7/4.0 Rank: 33/396 (Top 8.3%)

🔧 工作经历

美团 核心本地商业 M17 算法研究实习生 2024.09 – 2025.01

- 参与基座大语言模型预训练网页数据的清洗与质量评估，并负责本地 7B 模型的部署与全量微调；在网页数据清洗任务上，微调后模型的 ROUGE-L 达到 GPT-4 基线的约 96%，较传统方法提升约 6%

深圳计算科学研究院 基础研究部 实习研究员 2023.06 – 2024.01

- 参与图数据治理与增强、图神经网络可解释性、图计算系统，以及大规模数据场景下大语言模型的低资源适配等方向的研究。期间成功产出多篇学术论文，发表于 TODS、ICDE 等 CCF A 类期刊与会议

📚 科研经历

本人的研究主要围绕以数据为中心的人工智能展开，关注图学习中的数据质量、知识发现与模型可靠性问题，具体工作包括 LLM 辅助的特征抽取与排序、图数据质量增强、图数据挖掘及 GNN 可解释性等

《Explaining GNN Negatives Globally and Locally》

- Kehan Pang, Wenfei Fan, Min Xie, and Dandan Lin. (ICDE, CCF A) 2026
- 针对 GNNs 解释方法中全局与局部建模不统一的问题，本人在该工作提出了基于规则学习的统一解释框架，将负预测解释建模为结构与特征联合刻画的问题；设计强化学习驱动的规则挖掘方法及基于 LLM 的谓词排序策略，在多个真实数据集上显著优于现有解释方法

《Imputing Sparse and Noisy Labels for GNNs》

- Wenfei Fan, Kehan Pang and Chao Tian. (ICDE, CCF A) 2025
- 针对图数据中标签稀疏、噪声及非结构化信息难以利用的问题，本人在该工作中设计了面向 GNN 的数据增强与标签治理方法，利用 LLM 从非结构化、半结构化属性中抽取任务相关特征，结合 Revised-1-WL 生成结构感知伪标签，并通过软标签推断与标签清洗提升下游分类性能

《Making It Tractable to Detect and Correct Errors in Graphs》

- Wenfei Fan, Kehan Pang, Ping Lu and Chao Tian. (TODS, CCF A) 2024
- 为解决静态图上的关联规则在时序图上语义存在欠缺的问题，本人在该工作提出了一类对时序图上的节点进行实时性排序的模型，以及将静态图上的关联规则适配至动态图的方法，显著增强了在数据动态更新场景下图关联规则对过时数据、错误数据检测的性能

《OpenMind: Query-Time Diversification via Logical Rules》

- Wenfei Fan, Kehan Pang and Dandan Lin. (VLDB, CCF A, 已录用) 2027
- 为解决现有推荐系统易形成信息茧房的问题，本人在该工作中提出了一种查询时观点多样化框架，并设计了面向 CTR 和 Top-k 推荐的联合推理与剪枝算法，在无需重训练原推荐模型的前提下，保证了大规模交互图上观点多样化推荐的准确性与运行效率

《ComCE: Community-Aware Counterfactual Explanations for Graph Neural Network》

- Kehan Pang, Dandan Lin. (Under Review)
- 为解决现有 GNNs 反事实解释方法在可解释性与效率上的不足，本工作提出了一种基于社区结构的反事实解释框架，提升了解释的语义一致性与计算效率的同时从理论上证明了最小反事实解释问题的 NP-hard 性，并设计了结合层次化搜索与强化学习的两阶段优化方法，在多个真实图数据集上取得了优于现有方法的效果

《Efficient Mixture of Experts based on Large Language Models for Low-Resource Data Preprocessing》

- Mengyi Yan, Yaoshu Wang, Kehan Pang, Min Xie and Jianxin Li. (SIGKDD, CCF A) 2024
- 为解决 MoE 模型在融合多专家时可能带来潜在的信息覆盖、信息冗余问题，本人在该论文中从理论上证明了数据增强时 Meta-path 的选取具有 Church-Rosser 性质，并负责设计 RAG 模块的信息瓶颈网络架构，提升了 MoE 模型相较单专家在下游任务上的性能

《Linking Entities across Relations and Graphs》

- Wenfei Fan, Ping Lu, Kehan Pang, Ruochun Jin and Wenyuan Yu. (TODS, CCF A) 2024
- 为解决异源、异构数据集下实体对齐的语义不一致问题，本人在该工作中提出了一种实现跨异构图与关系表的实体对齐算法、应对数据量增长的数据分区索引算法、增量实体对齐算法，保证了系统在兆级数据下的运行效率

♥ 获奖、专利与学术服务

北京市大学生数学竞赛（理工类甲组）一等奖	2019第30届、2020第31届
全国大学生数学竞赛（非数学类）一等奖	2019第11届、2020第12届
北京邮电大学大学生创新创业训练计划市级项目奖	2020年9月
美国大学生数学建模竞赛H奖	2020年2月
北京邮电大学本科奖学金	2019-2021年
北京航空航天大学计算机学院学业奖学金	2022-2025年
ICDE External Reviewer	2024-2026年
一种图数据增强方法、装置、设备及介质（中国发明专利申请）	2025年7月

⚙ 研究与技术专长

- 学术科研：具备独立推进研究课题并完成高水平学术论文撰写的能力，具有跨团队科研协作经验。
- 数据治理：长期从事图数据、关系数据及半结构化数据治理研究，熟悉异构数据对齐、错误检测与修复、标签清洗与增强、特征抽取及增量数据处理。
- 图计算与机器学习：长期从事图关联规则、图数据挖掘、图神经网络及其可解释性研究，具备面向大规模图数据的规则发现、索引构建、增量计算、搜索剪枝经验。
- 大语言模型：具备大语言模型部署、LoRA微调、全参数微调及下游任务适配经验，熟悉RAG、信息检索、语义表示与模型评测方法，能够将大语言模型应用于数据清洗、特征抽取、规则构造与复杂数据分析任务。
- 系统实现与实验能力：熟练使用 C++/Python 进行算法与深度学习系统开发，具备 Linux 环境下大规模实验、性能分析、并行计算与工程调试经验；熟练使用大模型与 AI 编程工具辅助研发，支持方案设计、代码实现、调试分析及快速原型迭代。
- 语言能力：英语 CET-6，具备英文论文、技术文档及研究报告的阅读与撰写能力，具有多次国际学术会议参会、交流经历。