

基于 SPMA 框架的数据集结构对链接预测性能影响机制研究

姜小波 邓亚东 邱 崧

(华南理工大学电子与信息学院 广州 510641)

摘要 针对知识图谱链接预测任务中数据去冗余操作(剔除逆向与对称关系三元组)导致的模型性能显著下降及优化困难问题,本研究提出结构扰动与多视角分析框架 SPMA(Structural Perturbation and Multi-view Analysis Framework),系统探究数据结构特征对模型性能的作用机制。以 FB15k-237 为基准数据集,该框架构建“结构特征-性能指标-作用机制”三维分析体系,整合结构扰动生成、量化评估及多视角解释方法,形成可迁移的通用分析范式。研究首先开发基于启发式广度优先搜索的连通子图采样算法,构建结构特征差异化子图集合。通过敏感性实验发现,关系类别分布不均衡构成关键结构性瓶颈,其负向影响在不同采样场景下呈现显著相关性。进一步通过收敛动态分析、梯度分布解析和嵌入空间可视化等多维度方法,揭示不同关系类别对优化强度、梯度贡献度及空间结构需求存在显著差异,这种内在冲突导致模型难以达成多目标优化平衡。TransE、ComplEx、SEGNN 等主流模型上的系统性验证表明:性能制约主要源于数据结构固有特征而非模型架构限制。本研究建立的 SPMA 框架为知识图谱数据质量评估、结构优化及模型训练策略设计提供了新的理论工具。

关键词 知识图谱嵌入; 链接预测; 机制研究; 关系类别分布; LIME 模型

中图分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.00520

A Study on How Dataset Structure Influences Link Prediction Performance Using the SPMA Framework

JIANG Xiao-Bo DENG Ya-Dong QIU Song

(School of Electronics and Information, South China University of Technology, Guangzhou 510641)

Abstract The removal of redundant triples (specifically inverse and symmetric relations) in knowledge graph link prediction tasks, commonly adopted for data leakage mitigation, often leads to a significant degradation in model performance and presents considerable challenges for optimization. This study proposes the Structural Perturbation and Multi-view Analysis Framework (SPMA) to systematically investigate the underlying mechanisms by which data structural features influence model performance. Using FB15k-237 as a benchmark dataset, SPMA establishes a three-dimensional analysis system encompassing “structural features-performance metrics-influence mechanisms”. It integrates structural perturbation generation, quantitative evaluation, and multi-view explanation methods, forming a transferable and generalizable analytical paradigm. Firstly, this research develops a connected subgraph sampling algorithm based on heuristic Breadth-First Search (BFS) to construct subgraph sets with

收稿日期: 2025-05-25; 在线发布日期: 2025-11-05。本课题得到国家自然科学基金(U1801262)和广东省科技项目基金(2019B010154003)资助。

姜小波, 博士, 副教授, 主要研究领域为自然语言处理(信息抽取、自然语言生成、金融领域情感分析)、AI芯片设计、差错控制码设计。E-mail: jiangxb@scut.edu.cn。邓亚东(通信作者), 硕士研究生, 主要研究领域为深度学习、自然语言处理。E-mail: 202321011978@mail.scut.edu.cn。邱崧, 本科生, 主要研究领域为深度学习、自然语言处理。

differentiated structural features. Through sensitivity experiments, we discover that the imbalanced distribution of relationship categories constitutes a critical structural bottleneck. Its negative impact exhibits a significant correlation across different sampling scenarios. Furthermore, multi-dimensional analyses, including convergence dynamics, gradient distribution analysis, and embedding space visualization, reveal that different relation categories demand distinct optimization intensities, gradient contributions, and spatial structural requirements. This inherent conflict hinders models from achieving a balanced multi-objective optimization. Systematic validation on mainstream models such as TransE, ComplEx, and SEGNN consistently indicates that performance constraints primarily stem from the inherent structural features of the data rather than limitations of the model architectures. The SPMA framework established in this study provides a novel theoretical tool for knowledge graph data quality assessment, structural optimization, and the design of model training strategies.

Key words knowledge graph embedding; link prediction; mechanism study; relationship categories; LIME model

1 引言

现代人工智能系统的演进高度依赖于高质量数据集的支撑。数据集不仅决定模型的学习能力与预测准确性，还推动着跨学科研究范式的革新^[1-6]。数据集设计中一个重要的趋势是消除冗余三元组(如逆关系和对称关系)来减少信息泄露，从而推动模型从“记忆模式”转向真正的“推理能力”。

FB15k-237 是这一趋势的典型代表^[7]，来源于 FB15k 数据集，但其内部存在大量冗余关系如逆向关系和对称关系导致信息泄露，使得模型在测试集上能够通过简单的模式匹配或反向推理来预测缺失链接，而非真正学习到复杂的语义关联，进而导致模型性能虚高。FB15k-237 通过去除其中冗余关系避免了数据泄露，但其严格的数据构建标准也带来了显著的性能下降的问题。研究表明，在同等模型设定和训练规模下，FB15k-237 的平均倒数排名(MRR)最高仅达到 0.42，远低于同类型数据集的 FB15k 的 0.86^[8]以及 WN18RR 的 0.74^[9]。因此，FB15k-237 已逐渐成为衡量模型实际推理能力的挑

战性基准^[10-12]。

然而，传统优化策略未能提高 FB15k-237 上的链接预测性能。以往研究普遍认为，提升性能需优化复杂实体关系建模、增强语义表达能力或调整模型结构与超参数等。例如，早期的基于翻译的模型 TransE 被 TransD、TransR 扩展^[13]，以更好地捕捉复杂的关系模式。同样，DistMult、TuckER、InterTris 等^[14-17]双线性模型通过矩阵和张量分解来捕捉语义关联；QuatCNNE_x 和 HittER^[18-19]等神经网络方法分别采用了卷积和基于 Transformer 的架构来处理复杂信息。此外，由于图神经网络(GNN)对图结构的强大处理能力，近年来使用图神经网络进行链接预测的研究也取得了显著进展。这些研究不仅探索了如 NBFNet^[20]等路径查找算法，还包括 SEGNN^[21]等复杂的 GNN 架构，它们能够有效地聚合邻居信息以预测缺失链接。除了模型架构的创新，研究人员还探索了各种超参数调整 and 不同种类的损失函数以进一步优化性能^[22-24]。然而，这些在传统基准测试中表现良好的策略，在 FB15k-237 上却未能实现预期的增益。图 1 描绘了不同知识图谱

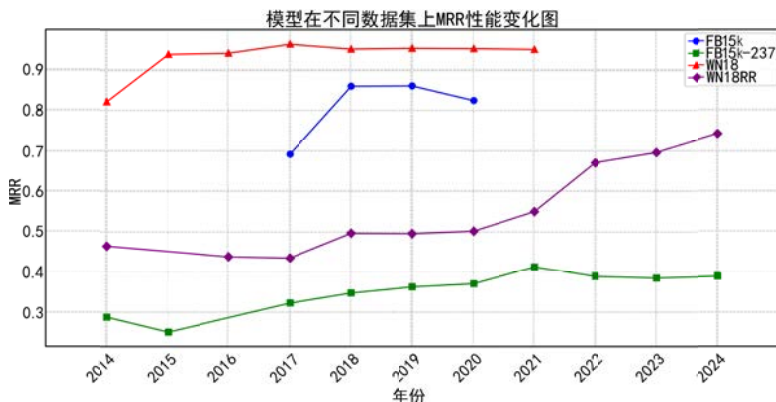


图 1 不同数据集上模型在各年份的性能峰值变化曲线

基准数据集上, 链接预测模型在各年份所能达到的峰值性能变化曲线, 显示在更具结构约束的数据集下, 传统的建模方法面临显著瓶颈。

这一现象的根本原因并非完全源于模型能力的不足, 而在于 FB15k-237 的结构特征影响性能的机制尚未被揭示。FB15k-237 虽未改变实体或关系本身的信息, 但其构造过程显著改变了数据的结构分布, 使得模型在训练与预测中面临潜在的结构适应性挑战。为系统揭示这一问题背后的机制, 本文提出一种结构扰动与多视角机制分析框架 SPMA (Structural Perturbation and Multi-view Analysis Framework), 用于研究数据集结构特征影响模型性能的机制。该框架由结构扰动构造、性能影响量化及机制路径剖析三部分组成, 具备良好的可迁移性与通用性。具体而言, SPMA 首先设计了一种基于启发式广度优先搜索的连通子图采样方法, 用于构建具有多样结构特征子图集合; 随后在多个主流 KGE 模型(如 TransE、CompLex、SEGNN)上进行相关性与 Sobol 敏感性分析, 发现关系类别分布的不均衡性是主要性能瓶颈因素。本文进一步使用上述流程在 WN18RR 数据集上进行了普适性实验, 实验结果发现 WN18RR 虽然也通过移除冗余关系来避免信息泄露, 但其性能瓶颈却呈现出不同的结构来源, 图密度与全局聚类系数等宏观连接特性对模型有较大影响。这种在不同数据集上性能瓶颈的结构差异, 恰恰凸显了现有研究缺乏一个通用的分析框架来系统性地诊断和揭示特定数据集的内在结构性挑战。本文的研究重点为性能更低的 FB15k-237, 为了探究关系类别在 FB15k-237 上对模型性能影响的机制, SPMA 从训练过程的三个视角深入剖析关系类别分布对性能的影响路径: (1) 从收敛视角, 当前训练过程被高频关系类别主导, 而低频关系类别面临梯度稀疏、学习停滞, 关系类别的分布不均衡会使整体训练过程出现“主关系类别欠拟合、次关系类别过拟合”的结构不平衡; (2) 从梯度视角, 不同关系类别在训练中的梯度占比与其数据量相关, 通过改变占主导的 n - n 类别的占比, 发现性能与在训练时的梯度占比之间呈现正相关, 这表明关系类别的分布不均会使影响梯度的分配进而影响模型在不同关系类别上的性能; (3) 从嵌入向量视角, 借助 LIME 可解释方法分析不同关系类别对实体与关系表征学习的作用, 发现关系

分布的不均衡会导致嵌入空间中不同类别关系的表征偏移, 从而削弱模型对多关系类别的统一表达能力。三视角下的实验结果都验证了关系类别分布不均衡对模型性能的显著影响, 且该影响主要源自数据集结构本身, 而非模型架构所限。

综上所述, 本文不仅系统揭示了 FB15k-237 中结构特征影响性能的作用机制, 更提出了可迁移、通用的 SPMA 框架。该框架量化了结构因素的性能影响强度, 明确了其机制路径, 为提升模型在结构复杂数据集中的泛化能力提供了理论依据, 也为未来的数据集设计与结构优化研究提供了分析工具与范式支持。

本文结构安排如下: 第 1 节介绍研究背景与问题定义; 第 2 节综述现有 KGE 模型与相关结构性研究; 第 3 节详述所提出方法, 包括子图采样、结构指标计算与解释性分析方法; 第 4 节展示实验设置与结果; 第 5 节总结全文并讨论未来工作方向。

2 相关工作

2.1 KGE 模型

知识图谱补全(KGC)旨在识别实体之间缺失的交互关系, 从而解决知识图谱不完整的问题。知识图谱嵌入(KGE)模型是解决 KGC 任务的主流方法之一, 其核心思想是将知识图谱中的实体和关系映射为低维连续向量, 并通过定义一个打分函数来衡量三元组的合理性^[25]。通常情况下, KGE 模型遵循四个步骤^[26]: (1) 随机初始化: 对实体向量和关系向量进行随机初始化; (2) 打分函数: 定义一个打分函数来衡量三元组的合理性; (3) 交互机制: 设计交互机制来对实体和关系的交互进行建模, 并计算三元组的匹配得分; (4) 训练策略: 通过负采样、正则化等策略最大化对三元组的置信度来训练 KGE 模型。

基于上述流程, 研究人员提出了多种模型。最早的是基于翻译的 TransE 模型, 其优点在于简单且高效, 但在表示复杂关系时能力有限。该局限性推动了后续研究在进一步扩展知识图谱嵌入的建模能力方面做出尝试。例如, 为了克服 TransE 在处理复杂关系类别方面的不足, 研究者提出了 TransH 和 TransR^[13]。TransH 引入了特定于关系的超平面, 使同一个实体在不同的关系下能拥有不同的投影

表示, 从而更好地处理复杂关系。TransR 则进一步提出将实体和关系嵌入到不同的向量空间中, 通过投影矩阵将实体投影到特定于关系的空间内, 解决了实体与关系共享同一空间可能导致的表达不足问题。随后, RotatE^[15]在复数向量空间中把每个关系定义为一个旋转操作, 将关系表示为从头实体到尾实体的旋转。由于这一设计, RotatE 能够有效建模对称、反对称以及多对多等关系, 从而弥补了 TransE 在处理此类关系类别时的不足。

除了上述基于翻译的模型之外, 还有采用三线形乘积形式的打分函数来衡量实体与关系间语义相似度的语义模型。例如, DistMult 使用双线性模型来捕捉实体与关系之间的交互, 其打分函数采用三线形乘积形式。这样的设计使得 DistMult 在对称关系的建模上表现良好, 但难以准确刻画反对称关系。ComplEx 则将实体和关系的嵌入表示从实数空间扩展到复数空间, 从而能够更好地表示如循环关系、 n 对 n 等更复杂的关系类别。此外, 一些模型借助深度学习技术来捕捉知识图谱中更复杂的模式^[27]。例如 ConvE 基于卷积神经网络, 能够捕捉实体与关系嵌入之间的局部特征交互, 从而有效学习实体与关系的多样组合; 同样地, CapsE 将胶囊网络引入知识图谱嵌入框架, 以捕获实体与关系之间更精细的特征交互, 在对复杂关系的建模上展现出更佳的表现。

尽管研究人员为提升 KGE 模型在链接预测任务上的性能做出了诸多努力, 但在诸如本研究所关注数据集这样的复杂数据集上, 当前的实验结果似乎已经遭遇瓶颈。这表明仅通过优化模型本身的结构和参数, 提升效果已十分有限。为此, 亟需从数据集结构特征的角度, 对模型的内在机制进行深入分析, 并探索导致性能瓶颈的潜在原因, 从而为后续模型改进提供方向和启示。

2.2 机制研究

近年来, 知识图谱嵌入(KGE)模型的研究取得了显著进展, 但在影响模型性能的关键因素方面仍存在研究空白。一些现有研究通过优化模型设计、调整超参数以及处理数据分布问题来提高嵌入质量。Oliver 等人^[28]利用 Sobol 敏感性分析来评估不同超参数对嵌入质量方差的影响, 旨在找出哪些超参数可以在不显著影响嵌入质量的情况下被排除于搜索范围之外。Zhang 等人^[12]提出了加权知识图谱

嵌入(WeightE)方法, 用以验证长尾分布对模型性能的影响; Akrami 等人^[29]则揭示了数据泄漏问题可能导致对嵌入模型性能的高估。在 KGE 模型可解释性方面, Zhang 等人^[30]提出了基于路径的异质链接预测 GNN 解释方法(PaGE-Link), 用以解决 GNN 模型在链接预测任务中缺乏可解释性的问题。PaGE-Link 能够生成可解释路径并拥有可扩展的模型能力。Ma 等人^[31]则提出了 KGExplainer, 用以解决现有 KGE 模型解释方法在推理信息不足时的局限性。

在去除冗余关系后的 FB15k-237 数据集中, 其结构特征会显著影响嵌入模型的学习过程和性能。然而, 这一方面尚未得到深入研究。因此本研究将探讨其结构特征对嵌入性能的影响机制。

3 方法

3.1 子图采样方法

我们提出了一种连通子图采样方法, 以构建不同结构特征子图^[32]。该方法基于一种启发式策略, 优先选择高连接度节点, 确保采样子图具有良好的连通性和代表性, 避免采样到孤立或稀疏区域。同时利用广度优先搜索(BFS)^[33]进行采样子图, 采样以一个明确的起始节点为起点, BFS 算法的性质保证了所有通过扩展加入到 V_s 的节点, 都必然是通过一条或多条边与先前的已采样节点相连。这意味着, 从起始节点 v_o 到任何一个被采样的节点 $v \in V_s$, 都存在一条由采样子图中的边构成的路径。由于整个采样过程是一个从中心向外扩展、不断添加相邻节点的过程, 且只有与已采样节点直接相连的节点才会被加入, 因此当采样节点数量达到目标后, 最终形成的子图 G_s 中的所有节点都将是相互连通的。

具体方法如下:

对于给定知识图 $G=(V, E)$, 节点数为 $|V|$ 。采样目标是一个子图 $G_s=(V_s, E_s)$, 其中子图的节点数 V_s 满足:

$$|V_s| = \lceil r \cdot |V| \rceil \quad (1)$$

其中, r 代表采样比率, 它决定了要采样的节点数量。该比率是从 $[r_{\min}, r_{\max}]$ 范围内的均匀分布中随机选择的。我们设置采样比率 r 的范围为 $[0.5, 1.0]$, 即子图节点数控制在总节点数的 50% 到 100% 之间。该范围的设定主要基于两点考虑: 其一, 保证子图的

结构代表性。过低的采样率(如 $r < 0.5$)可能导致生成的子图过于稀疏或碎片化,其结构特征会严重偏离原始图,0.5的下限确保了采样子图足以保留原图关键的宏观结构信息。其二,实现充分的结构扰动,[0.5, 1.0]的区间提供了足够大的子图规模变化,能够使关系类别分布、图密度等结构指标产生显著的数值变动,这对于后续进行鲁棒的相关性分析和敏感性分析至关重要。

为了提高子图采样的连通率,优先从高连接度节点开始采样。具体做法为,定义节点 $v \in V$ 的度数为:

$$\deg(v) = |\{u \in V : (u, v) \in E \text{ 或 } (v, u) \in E\}| \quad (2)$$

节点 v 的度数是其直接邻居的数量。

将节点按度数降序排列后,选择前 k 个节点形成候选集,研究取 k 为50,即选择度数排名前50的节点作为候选集:

$$S_k = \{v_1, v_2, \dots, v_k\} \quad (3)$$

其中,

$$\deg(v_1) \geq \deg(v_2) \geq \dots \geq \deg(v_k) \quad (4)$$

从候选集中随机选择一个节点 v_0 作为采样起始点,初始化采样节点集 V_s 和待扩展队列 Q :

$$V_s = \{v_0\}, Q = [v_0] \quad (5)$$

之后,从队列 Q 中取出当前节点 u ,遍历 u 的所有邻居节点:

$$N(u) = \{v \in V : (u, v) \in E \text{ 或 } (v, u) \in E\} \quad (6)$$

将尚未采样的邻居节点加入采样集 V_s 和队列 Q 。即

$$V_s = V_s \cup (N(u) \setminus V_s), Q = Q \cup (N(u) \setminus V_s) \quad (7)$$

重复以上步骤直到满足

$$|V_s| \geq |r \cdot |V|| \quad (8)$$

通过广度优先搜索(BFS)机制,确保了采样的所有节点都与起始节点连通,从而保证了采样的子图 G_s 是连通的,更完整地保留知识图的结构特性,有助于研究知识图的局部结构特性。同时为确保每次采样都能达到目标节点数量,我们设定了最大重试次数(100次),以应对某些极端图结构下难以达到采样比例的情况。

3.2 基尼指数

在本研究中,我们采用基尼指数^[34]来量化关系类别、关系类型以及节点度分布中的不平衡程度。基尼系数是一种广泛使用的评估分布不均衡程度的指标。基尼系数最初是为了衡量资源分配和经济

不平衡而开发的,但它也适用于描述其他领域中的不平衡情况,例如知识图谱中的数据分布。基尼系数在数学上定义为任意两个数据点之间平均绝对差值与数据均值之比,其公式表达如下:

$$G = \frac{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2\bar{x}} \quad (9)$$

其中, n 表示样本总数, x_i 为样本中第 i 个数据点, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 表示样本均值。较大的基尼指数 G 表示分布越不均衡。

3.3 LIME模型

LIME(局部可解释模型无关解释模型)^[35]是一种通用的模型解释方法。旨在通过局部线性模型来近似复杂模型在局部区域内的行为。从而揭示黑盒模型的决策过程。其基本原理是,在输入空间的特定区域内,复杂模型可以被简单的线性模型所局部近似。LIME的主要目标是评估输入特征的重要性,并量化它们对模型输出的贡献。该方法基于以下原理:

假设黑盒模型为

$$f: \mathbb{R}^d \rightarrow \mathbb{R} \quad (10)$$

其中,输入为 $x \in \mathbb{R}^d$,输出为预测值 $f(x)$ 。

LIME通过以下过程生成解释:首先,在输入样本 x 的邻域内生成一组扰动样本 $\{x'_i\}_{i=1}^n$ 。随后,这些扰动样本由黑盒模型 f 进行预测,得到相应的预测值 $\{f(x'_i)\}_{i=1}^n$ 。接下来,定义一个权重函数 $\pi_x(x')$,通常采用高斯核函数:

$$\pi_x(x') = \exp\left(-\frac{\|x - x'\|^2}{\sigma^2}\right) \quad (11)$$

它表示扰动样本 x' 相当于原始输入 x 的权重,其中, σ 为带宽参数。

最后,在扰动样本空间中,通过加权最小二乘法拟合一个简单的线性模型 g ,以最小化以下加权误差函数:

$$g(z') = \beta_0 + \sum_{j=1}^d \beta_j z'_j \quad (12)$$

$$L(f, g, \pi_x) = \sum_{i=1}^n \pi_x(x'_i) \cdot (f(x'_i) - g(z'_i))^2 + \Omega(g) \quad (13)$$

其中, L 为加权平方误差,用以衡量局部线性模型 g 对黑盒模型 f 的拟合程度; $\Omega(g)$ 表示正则化项,用于控制线性解释模型 g 的复杂度(例如通过 $L1$

正则限制特征维度)。

拟合得到的线性模型参数 $\{\beta_j\}_{j=1}^d$ 即为特征的重要性权重, 反映了每个特征在样本 x 附近对模型预测 $f(x)$ 的贡献, 可以用来解释特征对模型的影响。

在本文中, LIME 被用于解释不同嵌入维度对模型得分的影响。对于输入三元组 (h, r, t) , 其嵌入特征 $x = [h, r, t]$ 定义了模型 f 的打分函数, 即 KGE 模型(例如 TransE)的预测得分:

$$f(x) = -\|h + r - t\|_1 \quad (14)$$

通过优化目标函数 $L(f, g, \pi_x g)$, LIME 导出了嵌入特征 $x = [h, r, t]$ 的线性模型参数 $\{\beta_j\}_{j=1}^d$, 并将特征的重要性计算为:

$$I_j = |\beta_j| \quad (15)$$

其中, I_j 表示特征 j 的重要性。照嵌入类型(头实体、关系、尾实体)对重要性进行分类, 可以统计各嵌入类型对得分的总贡献:

$$\begin{aligned} I_{\text{head}} &= \sum_{j \in \text{head}} |\beta_j|, \\ I_{\text{relation}} &= \sum_{j \in \text{relation}} |\beta_j|, \\ I_{\text{tail}} &= \sum_{j \in \text{tail}} |\beta_j| \end{aligned} \quad (16)$$

由此, 通过 LIME 模型可以获得头实体、尾实体以及关系嵌入的特征重要性。

4 实验

在本节中, 本文基于所提出的结构扰动与多视角机制分析框架(Structural Perturbation and Multi-view Analysis, SPMA), 系统分析了 FB15k-237 数据集中多个关键结构特征对链接预测模型性能的影响路径。该框架集成了结构扰动构造、量化分析和机制解释三个环节, 旨在揭示结构因素在模型训练与推理中的具体作用方式。

4.1 实验设置

在链接预测任务中, 不同模型在多个常用基准数据集上的性能表现存在显著差异。已有研究表明, 无论是传统模型还是近年来提出的先进方法, 在 FB15k-237 上的性能普遍较低, 且以往在其他数据集上表现良好的方法在该数据集上的性能提升十分有限。这一现象表明, FB15k-237 的结构特性可能对模型性能具有显著影响。

为深入分析该问题, 本文借助 SPMA 框架, 选取了三个具有代表性的 KGE 模型: TransE、ComplEx 和 SEGNN, 分别代表基于翻译的模型、基于张量分解的模型、基于图神经网络的模型, 使用开源工具库 OpenKE^[36]进行实验, 评估各结构特征对模型性能的具体影响机制。性能评价指标选用平均倒数秩(MRR), 用于衡量正确实体在预测排名中的整体表现。

4.2 子图采样与相关性分析

本节作为 SPMA 框架的结构扰动模块, 本文提出一种基于高连接度节点启发式策略与广度优先搜索(BFS)结合的连通子图采样方法, 对 FB15k-237 数据集进行结构扰动, 本文总共进行了 60 次独立的子图采样实验, 每次采样生成一个具备不同结构属性的子图集合。这些子图在保持实体和关系基本信息不变的前提下, 在结构特征上与原始图存在差异。围绕结构-性能关系, 本文重点分析以下六类结构特征:

- (1) 关系类别分布指数: 衡量关系类别分布的平衡性, 使用基尼系数进行计算。
- (2) 关系类型分布指数: 评估不同关系类型对应的三元组数量的不平衡性, 使用基尼系数进行计算。
- (3) 度分布指数: 评估节点连接性分布的不均衡程度, 使用基尼系数进行计算。
- (4) 强连通分量: 统计图中强连通子图的数量。
- (5) 图密度: 表示图中实际边数与最大可能边数的比例。
- (6) 全局聚类系数: 衡量图中三角形(即节点之间形成的三元组)所占比例。

为了研究不同结构特征与模型性能之间的相关性, 本文将子图进行重新训练和测试后统计模型在各个子图上的性能(MRR), 得到结果后绘制每个结构特征与性能之间的散点图, 结果如图 2 所示。

通过散点图可以发现: (1) 关系类别分布指数与性能相关性比较明显, 呈现较强的负相关, 即关系类别分布越不均衡, 性能越差。不均衡的关系类别分布会导致某些关系类别的训练样本严重不足, 模型难以有效学习这些关系的语义模式, 从而影响模型的泛化能力。(2) 其次是强连通分量与性能, 呈现负相关, 强连通分量的增多表明图的整体连通性降低, 模型可能缺乏跨子图的语义信息整合能力, 影响推理效果, 导致模型的泛化能力下降。(3)

除此之外,度分布指数与图密度都与性能有一定程度的正相关,其相关程度虽不及关系类别分布指数,但仍具有一定研究价值。(4)对关系类型分布指数与全局聚类系数,其分布与性能之间的相关性

为了进一步验证本框架的普适性,我们同样在 WN18RR 数据集上进行了实验。详细的散点图分析结果见附录 A1,图 A-1。结果清晰地表明,与 FB15k-237 不同,在 WN18RR 上对模型性能影响最大的是图密度与全局聚类系数,它们与性能呈现出显著的正相关性,说明图的宏观连接和局部聚类特性成为影响模型学习和推理性能的关键因素。这初步证明了 SPMA 框架能够有效识别出不同数据集的特有结构瓶颈。

不大,这与以往研究得出关系类型分布导致的长尾效应影响模型性能的结论不完全相同。从不同模型的散点图中可以发现,不同模型上散点图的趋势大致相同,这说明影响是由数据集本身的结构导致。

为了进一步验证相关性,本文使用了皮尔逊相关系数^[37]和斯皮尔曼等级相关系数^[38]来量化各结构特性与模型性能之间的关系,其中皮尔逊相关系数和斯皮尔曼等级相关系数常用于检测数据集中的变量之间的线性关系,结果如表 1 所示。可以观察到关系类别分布指数与性能之间的相关性最为显著。这一发现表明了数据集 FB15k-237 数据结构中关系类别分布的重要性,表明更均衡的关系类别分布对应于更优的子图性能。需要强调的是,通过分

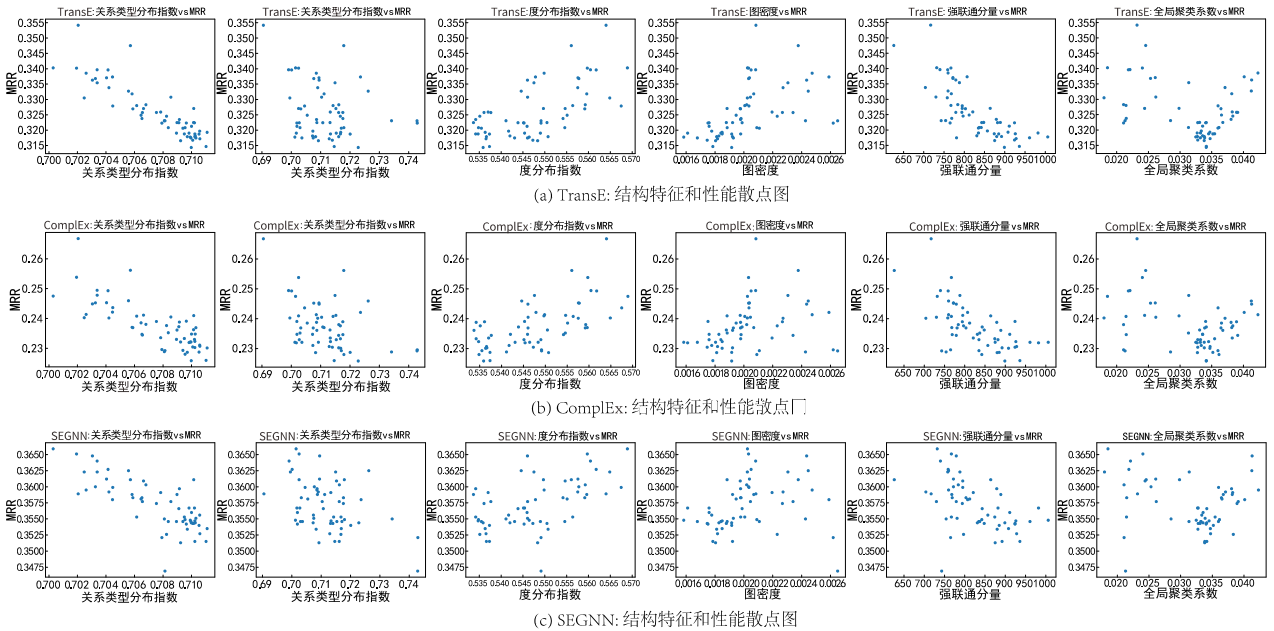


图 2 FB15k-237 结构特征与模型性能散点图

表 1 相关性分析结果

结构特征	方法	FB15k-237			WN18RR		
		TransE	ComplEx	SEGNN	TransE	ComplEx	SEGNN
关系类别分布指数	Pearson	-0.8795	-0.7829	-0.7673	0.0479	0.0585	0.0723
	Spearman	-0.8777	-0.7575	-0.7438	0.0326	0.1248	0.0799
关系类型分布指数	Pearson	-0.2514	-0.4274	-0.4295	0.3499	0.3555	0.3248
	Spearman	-0.1671	-0.3774	-0.3420	0.4462	0.5023	0.4666
度分布指数	Pearson	0.6386	0.6292	0.5391	0.2795	0.2521	0.2954
	Spearman	0.6157	0.5975	0.5102	0.2892	0.4033	0.3902
图密度	Pearson	0.5118	0.2959	0.2165	0.7171	0.8048	0.7656
	Spearman	0.6921	0.4271	0.4276	0.7412	0.8108	0.8069
强连通分量	Pearson	-0.7601	-0.6364	-0.4995	-0.0570	-0.2511	-0.1394
	Spearman	-0.7980	-0.6143	-0.5654	-0.3568	-0.2169	-0.2695
全局聚类系数	Pearson	-0.3376	-0.3283	-0.2039	0.7609	0.7957	0.7595
	Spearman	-0.1613	-0.1417	-0.1189	0.7155	0.7925	0.7745

别在 FB15k-237 以及 WN18RR 上对不同模型进行相关性实验，我们在不同模型上得到了相同的结果，说明实验结果不是仅限于某个模型。

通过在 FB15k-237 与 WN18RR 上的实验，可以观察到不同数据集影响模型性能的结构特征不同，这更加说明本文所提出的针对数据集结构特征影响模型性能的机理分析框架的重要性，不同数据集上的实验也证明了本文提出的框架的普适性。

4.3 不同结构特征的敏感性分析

在 SPMA 框架的量化分析模块中，本文进一步使用 Sobol 敏感性分析^[39]对结构特征的影响强度进行定量建模。Sobol 敏感性分析的核心思想是基于输入变量的变动，评估每个输入变量(或其交互作用)对输出结果的贡献程度，通过分解输出的方差来量

化输入变量的影响。本文通过子图采样得到具有不同结构特征子图从而改变输入变量以评估各个结构特征对性能的影响程度。同时，考虑到不同结构特征之间的交互性，通过 Sobol 敏感性分析的二阶指数得到不同结构特征之间的交互作用热力图。为了进行 Sobol 敏感性分析，本文使用了所有通过子图采样获得的 FB15k-237 子图样本(60 个)所计算的结构特征及其对应的模型性能数据作为输入。

结果如图 3 所示，其中 S1 代表一阶敏感性指数，它表示某个输入变量单独对输出结果的影响，衡量了该变量的变化对输出的直接贡献，S1 值越大，说明此变量的贡献越大。ST 代表总敏感性指数，表示该输入变量及其与其他输入变量的交互作用对输出结果的总贡献。

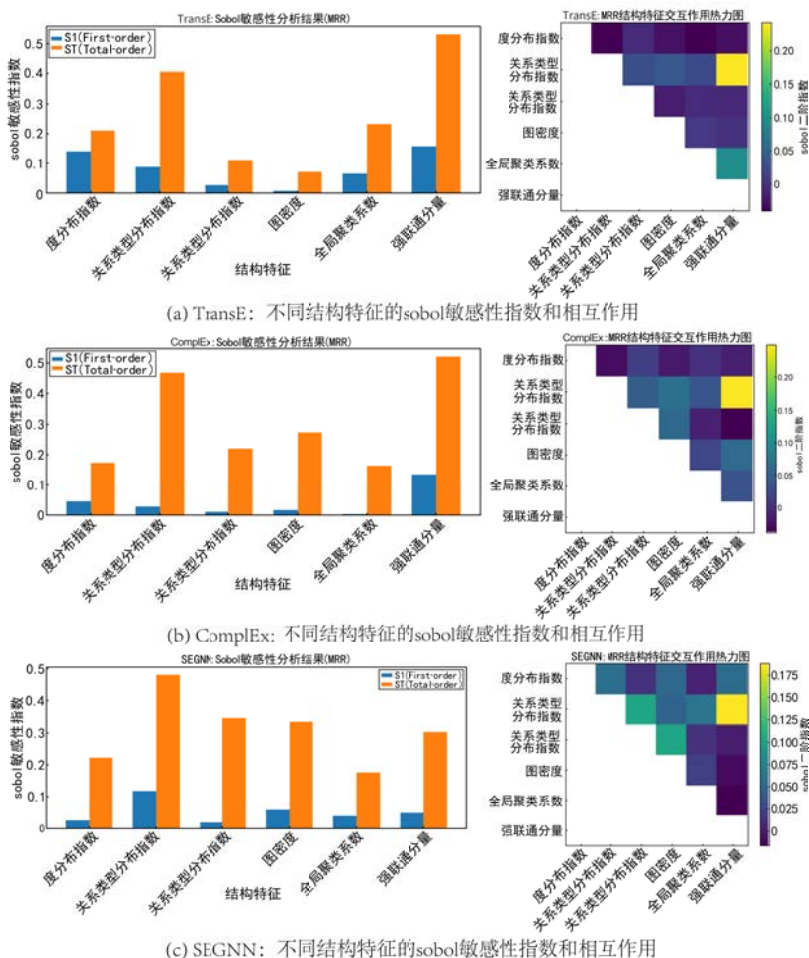


图 3 不同模型 Sobol 敏感性及其相互作用分析结果图

从结果图中可以看出，关系类别分布指数与强联通分量对不同模型的均有较大影响。结合之前的相关性分析，我们可以得出关系类别的分布程度不仅与模型性能的相关性较强，且对模型性能也有较

大影响。从二阶热力图中可以观察到，关系类别分布指数不仅自身对链接预测性能有重要影响，而且与其他结构特性之间的交互作用也会对模型的预测结果产生影响，相互影响较大的是关系类别分布

指数和强联通分量,这更加证明了在数据集中关系类别的分布是影响模型性能的关键因素。值得注意的是,之前有研究认为关系类型的分布不均衡导致的长尾分布现象会对模型性能的影响较大,然而我们通过研究发现在FB15k-237上其对模型影响远不如关系类别分布与强联通分量。

我们同样对WN18RR进行了Sobol敏感性分析(详细图表见附录A1,图A-2)。分析结果进一步确认,影响WN18RR性能的关键因素是图密度与全局聚类系数,这与FB15k-237中关系类别分布指数占据主导地位的情况形成了鲜明对比。

4.4 关系类别分布影响模型性能的机制研究

通过对两个数据集的对比分析,我们提出的SPMA框架已成功揭示出它们各自独特的结构瓶颈。鉴于FB15k-237的性能问题是本领域一个长期存在且极具挑战性的难题,本文后续章节将聚焦于FB15k-237的核心瓶颈——‘关系类别分布不均衡’,对其作用机制进行深入的多视角剖析,以期为

解决这一特定挑战提供理论依据和实践启示。SPMA框架设置了多视角解释模块,从训练收敛、梯度分布和嵌入向量三个维度探讨其机制路径。由于本文所使用的三个模型上均观察到了相似的性能分布趋势,为了保持图表的简洁性和清晰性,我们仅选取了最具代表性的基线模型TransE的结果进行后续机制研究。

4.4.1 关系类别分布影响模型性能训练收敛视角的机制研究

首先,本文对FB15K237数据集中的数据组成中各个关系类别的分布进行了统计,并以TransE模型为例,测试并统计了其在各个类别上的性能表现,结果如图4所示。从中可以看出,FB15K237数据集中的关系类别的分布极不均衡,其中 $n-n$ 类别占比超过70%,而其他关系类别数量较少。更为显著的是,模型在 $n-n$ 类别上的MRR预测性能表现不佳,这表示性能较低子图中 $n-n$ 关系类别的占比可能更大。

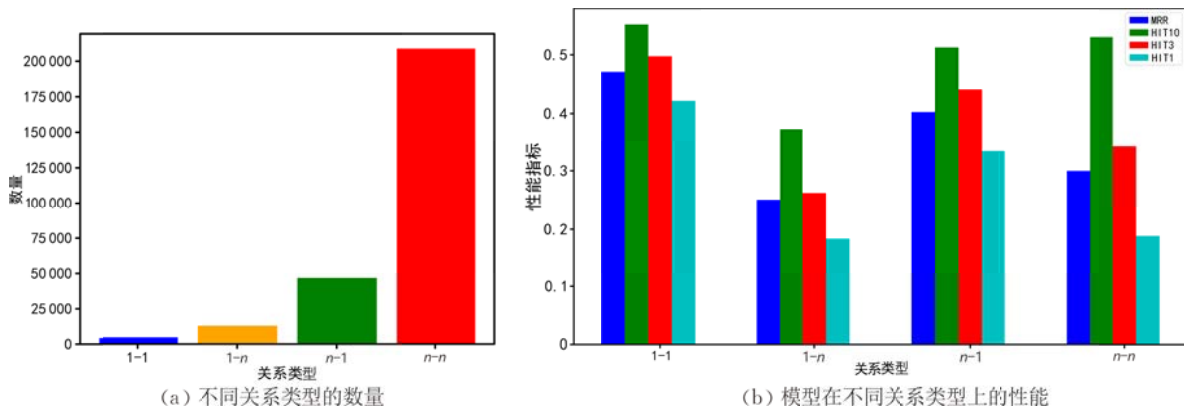


图4 不同关系类型的数量及性能分布图

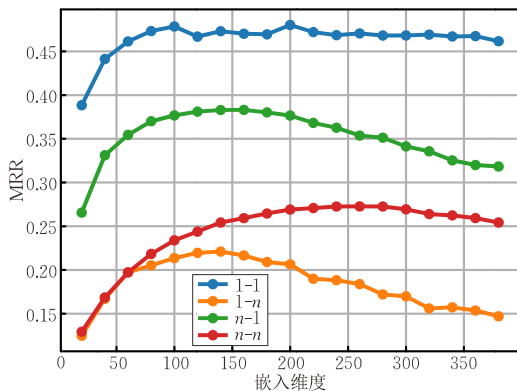
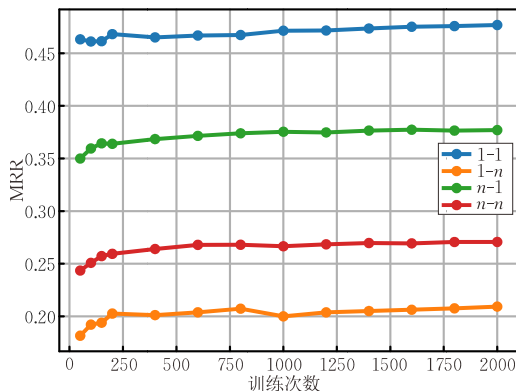
接下来,为了从收敛过程的角度进一步验证关系类别分布对模型性能的影响机制,本文重点考察不同关系类别在训练过程中对模型参数设置(如训练轮数、嵌入维度等)的敏感性差异。这有助于揭示为何某些关系类别在统一超参数下难以达到理想表现。为此本文分别设置不同训练时间、嵌入维度测试不同关系类别的性能并绘制了变化曲线,结果如图5所示。

通过以上结果获得了以下关键结论:(1)训练时间对性能的影响:不同关系类别在训练时间维度上表现出显著差异。以 $n-n$ 类别为代表的高频关系,由于包含更多训练样本,其性能在早期阶段提升较慢,需要更长的训练时间才能逐步收敛。然而,即

便经过更长训练,其最终性能仍低于其他关系类别,说明训练样本丰富并不必然带来性能提升。(2)嵌入维度对性能的影响:在不同嵌入维度设置下,复杂关系类别普遍需要更高的维度以捕捉更丰富的语义模式,模型才能实现有效学习;而对于1-1或 $n-1$ 等结构相对简单、语义关系明确的类别,较低的维度已足以获得较好性能。值得注意的是,在高维嵌入下,这些简单关系类别反而容易出现过拟合现象,原因在于其样本规模较小,难以支撑复杂模型带来的参数冗余。从以上收敛视角来看,不同关系类别的分布不均会导致在相同超参数下,难以同时收敛达到最优性能,因此关系类别的分布不均会影响模型性能。

综上，本文在收敛视角下揭示了关系类别结构对训练过程与模型表现的深层影响，强调了模型训

练策略与结构分布之间的匹配性。这一发现为结构自适应的知识图谱表示学习提供了理论依据。



(a) 不同关系类型性能随训练次数的MRR变化曲线

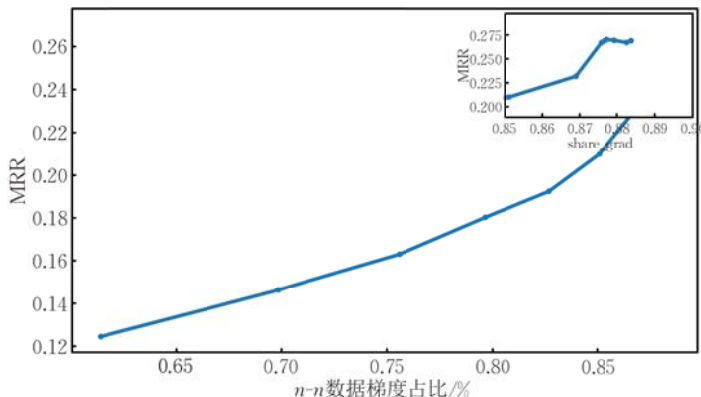
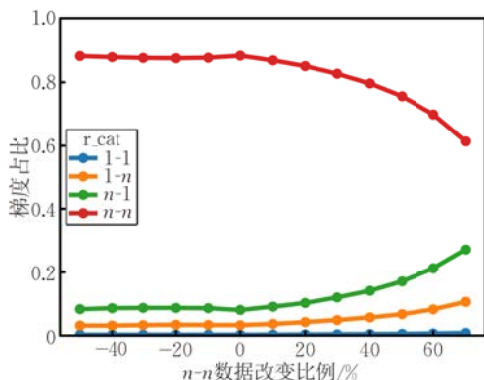
(b) 不同关系类型性能随嵌入维度的MRR变化曲线

图5 不同关系类别性能随训练次数、嵌入维度变化曲线

4.4.2 关系类别分布影响模型性能梯度视角的机制研究

在前一节收敛视角的分析中，已明确 n-n 类别在 FB15k-237 中占据主导地位，且该类别往往对应模型性能的薄弱环节。为了进一步揭示其影响机制，本文从训练过程中的“梯度信号”角度出发，研究 n-n 类别如何通过主导参数更新进而影响模型学习能力。具体来说本文设计了以下实验：通过按一定比例删减或增加 n-n 类别的数据(包括删减

10%到 70%，以及通过数据增强方法增加 n-n 类别的数据 10%到 50%)，构建一系列不同数量 n-n 类别的训练数据集。随后，在模型训练过程中，记录每一关系类别的梯度，并计算其梯度占比(某关系类别占总梯度的比例)。随后提取出测试集中 n-n 类别的数据以测量模型在该类别上的性能变化。实验结果如图 6 所示，显示了对 n-n 类别进行处理后在训练时的不同关系类别梯度所占比重变化，以及在 n-n 类别上的性能与训练时梯度占比的关联。



(a) 不同关系类型随 n-n 类别数据变化梯度占比曲线

(b) n-n 类型性能随梯度占比变化曲线

图6 n-n 类别梯度占比、性能变化曲线

实验结果表明：不同关系类别在训练时的梯度占比与其数据量相关。且在对 n-n 样本进行处理时，随着删减比例上升，n-n 的梯度占比显著下降，同时模型在该类别上的性能也明显下降；但在对 n-n 类别进行数据增强时，尽管数据量显著增加，但若模型训练过程中 n-n 的梯度占比未同步上升，其预测性能提升极为有限。这说明数据量本身并非直接决定模型表现的因素，而是通过主导训练过程中的梯

度分布，间接塑造模型的表示能力。换言之，模型在训练过程中对不同关系类别的“关注程度”——即其在参数更新中分配的梯度权重，是影响最终预测性能的核心路径。因此关系类别的分布不均会影响梯度的分配进而影响模型在不同关系类别上的性能。未来研究可考虑设计基于梯度占比的自适应训练策略如构建动态损失函数，按训练过程中实时监测的梯度占比调整不同关系类别的训练权重。

4.4.3 关系类别分布影响模型性能嵌入向量视角的机制研究

为了进一步揭示关系类别分布对模型内部表示机制的影响, 本文从嵌入向量的重要性出发, 探究不同关系类别如何影响实体与关系的表示学习过程。为此本文采用 LIME(Local Interpretable Model-agnostic Explanations) 方法对模型的预测结果进行解释, 分析给定三元组产生特定预测得分, 并识别三元组的哪些部分(头实体嵌入、关系嵌入、尾实体嵌入)对该预测贡献最大。对待解释的三元组, LIME 会在其对应的特征向量的周围生成大量的局部扰动样本。这些扰动是通过在原始特征向量上添加噪声或进行小幅度的修改来创建的。LIME 利用

这些扰动样本及其对应的预测得分, 通过加权线性回归拟合一个简单的局部线性模型。距离原始样本越近的扰动样本, 其权重越大。最终聚合这些单个维度层面的贡献, 统计出头实体嵌入、关系嵌入和尾实体嵌入这三大类对模型预测的影响频率和总贡献, 从而直观地展示在特定场景下, 哪个部分的嵌入对模型性能起到了关键作用。为确保样本的代表性, 我们对每个类别内的得分进行排序, 将得分排名前 10% 的三元组定义为“较高”得分样本, 排名后 10% 的定义为“较低”得分样本。通过 LIME 模型统计实体嵌入向量和关系嵌入向量对评分函数的重要性, 结果如图 7 所示。

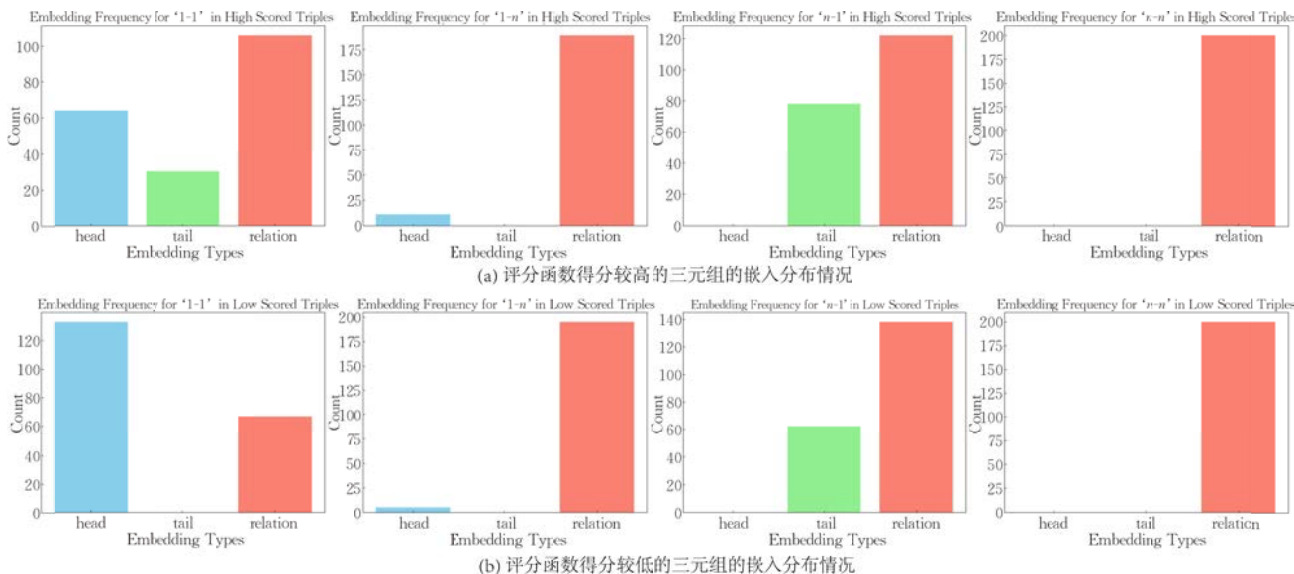


图 7 评分函数下代表性三元组嵌入向量分布

基于 LIME 模型的分析结果发现不同关系类别对模型嵌入向量的影响存在显著差异, 但关系嵌入向量在不同关系类别中的作用都尤为重要, 这表明关系嵌入向量是捕捉知识图谱中实体间相互关系的核心, 因此在模型设计中应特别关注关系嵌入向量的优化。此外不同关系类别对嵌入向量的影响具体如下: 对 1-1 关系类别, 得分较高的三元组对头尾实体的嵌入重要性相近, 表明在 1-1 关系类别中, 实体嵌入的需求较为平衡; 而 1-n 关系类别头实体嵌入的影响较大, n-1 关系类别尾实体嵌入的影响较大, 这表明, 模型在处理 1-n 或 n-1 关系类别时, 分别应该更注重头尾实体的嵌入; 而 n-n 关系类别则更依赖关系嵌入的重要性, 这意味着, 对 n-n 关系类别, 捕捉复杂的双向交互对于模型的嵌入表示至关重要。

本节通过 LIME 模型揭示了不同关系类别在训练过程中对嵌入向量的依赖模式差异, 发现不同关系类别对头尾实体以及关系嵌入的依赖程度不同, 关系分布的不均衡进一步导致嵌入空间中不同类别关系的表征偏移, 从而削弱模型对多关系模式的统一表达能力。这一发现提示未来的模型应根据关系类别的特性采用差异化的嵌入策略。

以上实验结果从多个角度验证了 SPMA 框架的有效性。通过结构扰动、敏感性量化与机制解释三重分析, 本文系统揭示了关系类别分布如何通过训练动态、梯度传递与表征学习影响模型性能。依据上述机制分析的结果, 未来研究可着重解决模型在不同关系类别上的差异, 针对不同关系类别对梯度贡献的差异, 设计能够动态调整不同关系类型训练权重的损失函数, 或针对嵌入空间中关系表征的

冲突问题, 探索能够更鲁棒、更具区分度地学习复杂关系类型嵌入的方法, 例如引入更高级的关系投影、多模型或基于对比学习的策略。

5 总结

本研究针对 FB15k-237 基准数据集为缓解数据泄露带来的性能下降和优化困难问题, 提出结构扰动与多视角分析框架 SPMA, 揭示数据集结构与模型性能的深层耦合机制。通过构建包含六类结构变量的可控扰动空间, 结合 Sobol 敏感性分析与跨模型验证, 首次定量证实关系类别分布不均是制约模型性能的核心结构瓶颈。从三个视角构建机制解释体系: (1) 从收敛视角揭示了关系类别分布不均会导致其难以同时收敛达到最优性能; (2) 梯度视角显示关系类别的分布不均会影响不同类别的梯度占比进而影响模型性能; (3) 基于 LIME 视角的可解释性分析发现关系分布的不均衡会导致嵌入空间中不同类别关系的表征偏移, 从而削弱模型对多关系模式的统一表达能力。在不同模型上实验 (TransE/ComplEx/SEGN) 验证了性能瓶颈的结构固有性, 这一发现为模型在 FB15k-237 上的优化提供了明确的结构切入点。综上所述, 本文不仅系统揭示了关系类别结构在知识图谱链接预测中的关键作用, 还构建了一个可推广的结构扰动与多视角分析框架, 为分析与改善其他结构复杂或表现受限数据集提供了通用的分析范式。

未来我们将致力于将 SPMA 框架推广应用于更广泛的知识图谱数据集, 发现并分析其他潜在结构性瓶颈。其次, SPMA 框架的计算开销是其应用于大规模知识图谱时需要关注的问题, 探索优化框架的计算效率将成为我们未来的工作。最后, 通过三个视角下的机理分析我们得出了关系类别影响模型性能的机理, 未来一个关键拓展方向是基于上述机理, 设计并验证针对不同关系类别特性的优化策略, 以缓解已有结构瓶颈的影响。

参考文献

- [1] Aguirre F, Sebastian A, Gallo M, et al. Hardware implementation of memristor-based artificial neural networks. *Nature Communications*, 2024, 15(1): 1-40
- [2] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//*Proceedings of the International Conference on Machine Learning*. Virtual, 2021: 8748-8763
- [3] Li H, Zhang Y, Koto F, et al. CMMLU: Measuring massive multitask language understanding in Chinese//*Proceedings of the Findings of the Association for Computational Linguistics* 2024. Bangkok, Thailand, 2024: 11260-11285
- [4] Suzgun M, Scales N, Schärli N, et al. Challenging BIG-Bench tasks and whether chain-of-thought can solve them//*Proceedings of the Findings of the Association for Computational Linguistics* 2023. Toronto, Canada, 2023: 13003-13051
- [5] Boudin M, Diallo G, Drancé M, et al. The OREGANO knowledge graph for computational drug repurposing. *Scientific Data*, 2023, 10(1): 871
- [6] Weth F R, Hoggarth G B, Weth A F, et al. Unlocking hidden potential: Advancements, approaches, and obstacles in repurposing drugs for cancer therapy. *British Journal of Cancer*, 2024, 130(5): 703-715
- [7] Toutanova K, Chen D. Observed versus latent features for knowledge base and text inference//*Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality*. Beijing, China, 2015: 57-66
- [8] Zhang Y, Yao Q, Dai W, et al. AutoSF: Searching scoring functions for knowledge graph embedding//*Proceedings of the 36th Institute of Electrical and Electronics Engineers International Conference on Data Engineering*. Virtual, 2020: 433-444
- [9] Li Q, Zhong Y, Qin Y. MoCoKGC: Momentum contrast entity encoding for knowledge graph completion//*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, USA, 2024: 14940-14952
- [10] Cao J, Fang J, Meng Z, et al. Knowledge graph embedding: A survey from the perspective of representation spaces. *Association for Computing Machinery Computing Surveys*, 2024, 56(6): 1-42
- [11] Zhang W, Chen J, Li J, et al. Knowledge graph reasoning with logics and embeddings: Survey and perspective//*Proceedings of the 2024 Institute of Electrical and Electronics Engineers International Conference on Knowledge Graph*. Nanjing, China, 2024: 492-499
- [12] Zhang Z, Guan Z, Zhang F, et al. Weighted knowledge graph embedding//*Proceedings of the 46th International Association for Computing Machinery Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*. Taipei, China, 2023: 867-877
- [13] Ruffinelli D, Broscheit S, Gemulla R. You can teach an old dog new tricks! on training knowledge graph embeddings//*Proceedings of the International Conference on Learning Representations*. Virtual, 2020: 1-12
- [14] Ahmed S F, Alam M S B, Hassan M, et al. Deep learning modelling techniques: Current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 2023, 56(11): 13521-13617
- [15] Chen X, Jia S, Xiang Y. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 2020, 141: 112948
- [16] Broscheit S, Ruffinelli D, Kochsiek A, et al. LibKGE-A knowledge graph embedding library for reproducible research//*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Virtual, 2020: 165-174
- [17] Zhang Yi, Meng Xiao-Feng. InterTris: Tri-interaction representation learning for domain knowledge graph. *Chinese Journal of Computers*, 2021, 44(8): 1535-1548 (in Chinese)
(张祎, 孟小峰. InterTris: 三元交互的领域知识图谱表示学习. *计算机学报*, 2021, 44(8): 1535-1548)
- [18] Xiong Wei, Chen Hao, Su Hong-Yu. QuatCNNE: A knowledge graph embedding model for multi-relational patterns. *Chinese Journal of Computers*, 2025, 48(1): 124-135 (in Chinese)
(熊伟, 陈浩, 苏鸿宇. QuatCNNE: 一种面向多关系模式的知识图谱嵌入模型. *计算机学报*, 2025, 48(1): 124-135)

[19] Chen S, Liu X, Gao J, et al. HittER: Hierarchical transformers for knowledge graph embeddings//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 10395-10407

[20] Zhu Z, Zhang Z, Xhonneux L P, et al. Neural bellman-ford networks: A general graph neural network framework for link prediction. Advances in Neural Information Processing Systems, 2021, 34: 29476-29490

[21] Li R, Cao Y, Zhu Q, et al. How does knowledge graph embedding extrapolate to unseen data: A semantic evidence view//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2022, 36(5): 5781-5791

[22] Zhang Z, Zhuang F, Zhu H, et al. Towards robust knowledge graph embedding via multi-task reinforcement learning. Institute of Electrical and Electronics Engineers Transactions on Knowledge and Data Engineering, 2023, 35(4): 4321-4334

[23] Wang P, Agarwal K, Ham C, et al. Self-supervised learning of contextual embeddings for link prediction in heterogeneous networks//Proceedings of the Web Conference 2021. Ljubljana, Slovenia, 2021: 2946-2957

[24] Liu Peng-Kai, Wang Xin, Liu Bao-Zhu, et al. DB4Trans: A database-inside knowledge graph embedding model training engine. Chinese Journal of Computers, 2022, 45(9): 1969-1982 (in Chinese) (柳鹏凯, 王鑫, 刘宝珠等. DB4Trans: 数据库内置知识图谱嵌入模型训练引擎. 计算机学报, 2022, 45(9): 1969-1982)

[25] Niu Guang-Lin, Lin Zhen. A survey on knowledge graph representation learning modeling relational properties. Computer Science, 2024, 51(9): 182-195 (in Chinese) (牛广林, 蔺震. 面向关系特性建模的知识图谱表示学习研究综述. 计算机科学, 2024, 51(9): 182-195)

[26] Yang H, Zhang L, Su F, et al. What affects the performance of models? sensitivity analysis of knowledge graph embedding//Proceedings of the International Conference on Database Systems for Advanced Applications. Virtual, 2022: 698-713

[27] Ye Z, Kumar Y J, Sing G O, et al. A comprehensive survey of graph neural networks for knowledge graphs. Institute of Electrical and Electronics Engineers Access, 2022, 10: 75729-75741

[28] Lloyd O, Liu Y, Gaunt T R. Assessing the effects of hyperparameters on knowledge graph embedding quality. Journal of Big Data, 2023, 10(1): 1-15

[29] Akrami F, Saeef M S, Zhang Q, et al. Realistic re-evaluation of knowledge graph completion methods: An experimental study//Proceedings of the 2020 Association for Computing Machinery SIGMOD International Conference on Management of Data. Virtual, 2020: 1995-2010

[30] Zhang S, Zhang J, Song X, et al. PaGE-Link: Path-based graph neural network explanation for heterogeneous link prediction//Proceedings of the Association for Computing Machinery Web Conference 2023. Austin, USA, 2023: 3784-3793

[31] Ma T, Tao W, Li M, et al. KGExplainer: Towards exploring connected subgraph explanations for knowledge graph completion. Ithaca: Cornell University, Technical Report: arXiv:2404.03893, 2024

[32] Bai Y, Zhang B, Xu N, et al. Vision-based navigation and guidance for agricultural autonomous vehicles and robots: A review. Computers and Electronics in Agriculture, 2023, 205: 107584

[33] Debowska A, Boduszek D, Ochman M, et al. Brain Fog Scale (BFS): Scale development and validation. Personality and Individual Differences, 2024, 216: 112427

[34] Martin A J F, Conway T M. Using the Gini index to quantify urban green inequality: A systematic review and recommended reporting standards. Landscape and Urban Planning, 2025, 254: 105231

[35] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?" explaining the predictions of any classifier//Proceedings of the 22nd Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1135-1144

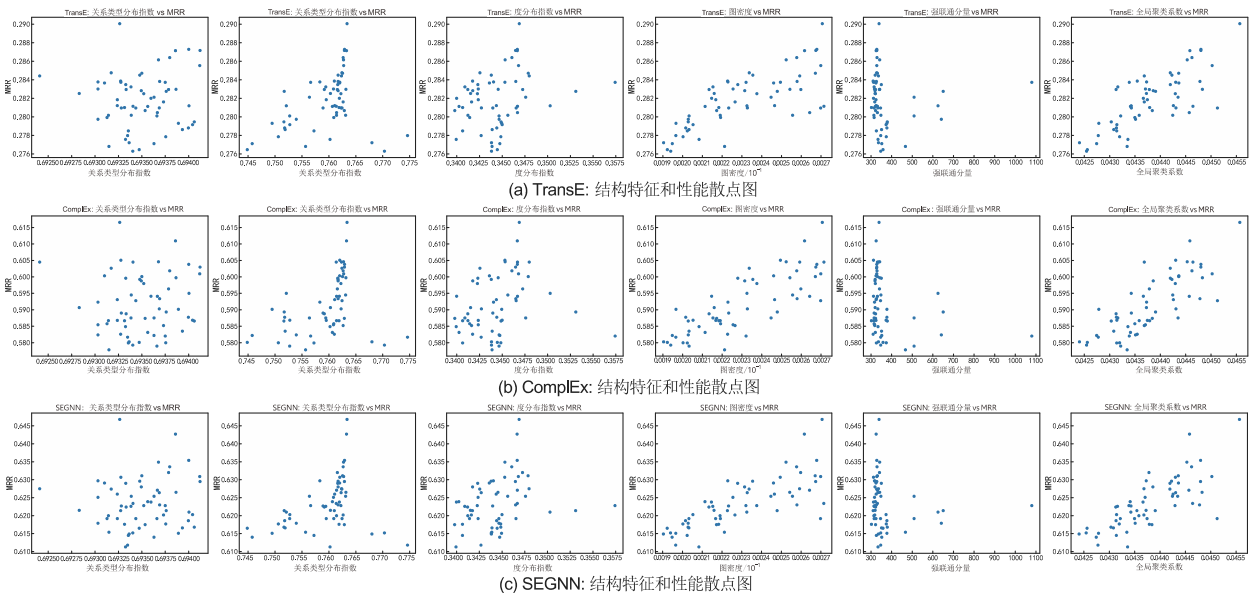
[36] Han X, Cao S, Lv X, et al. OpenKE: An open toolkit for knowledge embedding//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels, Belgium, 2018: 139-144

[37] Li Z, Yang Y, Li L, et al. A weighted Pearson correlation coefficient based multi-fault comprehensive diagnosis for battery circuits. Journal of Energy Storage, 2023, 60: 106584

[38] Ali Abd Al-Hameed K. Spearman's correlation coefficient in statistical analysis. International Journal of Nonlinear Analysis and Applications, 2022, 13(1): 3249-3255

[39] Renardy M, Joslyn L R, Millar J A, et al. To Sobol or not to Sobol? The effects of sampling schemes in systems biology applications. Mathematical Biosciences, 2021, 337: 108593

附录A.



图A-1 WN18RR结构特征与模型性能散点图

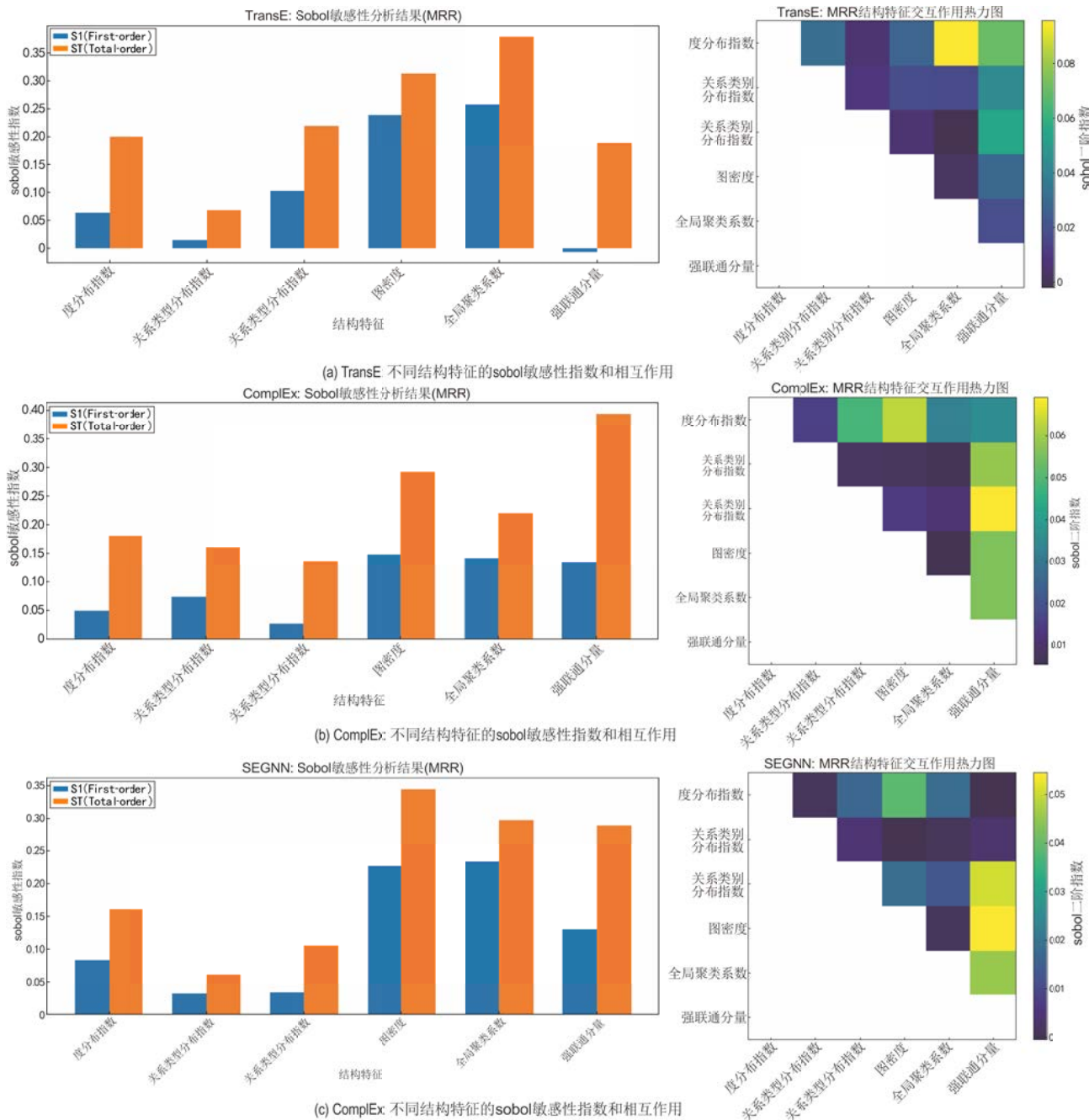


图 A-2 WN18RR 敏感性分析结果



JIANG Xiao-Bo, Ph.D., associate professor. His main research areas include natural language processing (information extraction, natural language generation, sentiment analysis in the financial domain), AI chip design,

and error control code design.

DENG Ya-Dong, M.S. candidate. His main research areas are deep learning, natural language processing.

QIU Song, undergraduate student. His main research areas are deep learning, natural language processing.

Background

This study is part of a broader project dedicated to advancing knowledge graph technologies, particularly in improving link prediction. Significant progress has been made

in this field, with many state-of-the-art models exceeding human-level performance on multiple benchmark datasets. However, models still face challenges on datasets such as

FB15k-237—where data leakage has been mitigated—and their performance has plateaued in recent years. The shortcomings of prior methods on FB15k-237 prompt us to investigate the structural characteristics of the dataset itself.

This paper proposes a general framework named SPMA (Structural Perturbation and Multi-view Analysis) to systematically uncover the mechanisms by which dataset structural features affect the performance of knowledge graph link prediction models. A subgraph sampling method is first introduced to perform structural perturbation, generating subsets with diverse structural features. Correlation and sensitivity analyses are then conducted to identify the key structural factor impacting model performance—namely, the imbalance in the distribution of relationship categories. Building on this, the study further performs a multi-view path analysis from three perspectives to elucidate how this structural feature influences model behavior. The findings offer theoretical insights for improving model performance on

FB15k-237 and provide guidance for enhancing the modeling of complex relation types, optimizing resource allocation strategies, and refining embedding representations in future research.

Furthermore, this project confirms the significant impact of relationship category distributions on link prediction and delves into their theoretical underpinnings. It thereby lays a foundation for enhancing how models represent complex relations, allocate resources, and optimize embedding methods. Our study also supplies a general analytical framework for modeling complex dataset structures and designing efficient embedding mechanisms, potentially driving further advancement in knowledge graph link prediction. This research was supported by the National Natural Science Foundation of China (Grant No. U1801262) and the Guangdong Provincial Science and Technology Program (Grant No. 2019B010154003).