Advancing Latent Relationship Inference Across Documents and Knowledge Graphs Using Graph Neural Networks for Explainable and Secure AI Systems

Tirth Kanani

Abstract

The integration of Knowledge Graphs (KGs) and Graph Neural Networks (GNNs) offers a promising approach to enhance the inference of both direct and latent relationships in complex datasets. Current AI systems, especially in critical sectors like healthcare and intelligence analysis, struggle to accurately identify hidden relationships within and across fragmented documents, limiting their ability to uncover unique perspectives and biases that shape decision-making. This research proposes novel GNN extensions to infer multi-hop latent relationships across multiple documents and KGs, introducing bias-aware attention mechanisms and scalable architectures. By incorporating Explainable AI (XAI) principles, the project ensures transparent and interpretable inferences, while secure, local deployment strategies protect sensitive data. These advancements aim to deliver fair, trustworthy, and privacy-preserving AI systems, with applications in bias detection for search engines, secure knowledge discovery, and beyond.

1 Background review

The integration of **Knowledge Graphs (KGs)** with **Graph Neural Networks (GNNs)** presents a promising solution for relationship inference, but current systems still face significant challenges in inferring **latent** or **indirect relationships**. As discussed by Liu et al. [1] and Li et al. [2], these challenges stem from the difficulty of capturing complex, multi-hop dependencies in large-scale, unstructured data environments. Current GNN models primarily focus on direct relationships, which limits their ability to extract deeper insights from fragmented data [3]. GNNs, when combined with KGs, provide structural advantages for information propagation, but they still struggle with indirect relationships, as shown by Pan et al. [4].

Graph Convolutional Networks (GCNs), widely used for relationship inference, update node representations by aggregating features from neighboring nodes, but they suffer from the inability to model **long-range dependencies**, as noted by Li et al. [2] and Gilmer et al. [5]. This limitation, rooted in their reliance on local message passing, reduces the model's capability to infer relationships between nodes separated by multiple hops, a critical flaw for cross-document analysis. Research into **multi-hop GCNs** by Chen et al. [6] seeks to overcome this by propagating information over multiple layers, but this approach remains computationally expensive and struggles to scale for large, interconnected KGs, as confirmed by Morris et al. [7] on higher-order GNNs. Similarly, **Graph Attention Networks (GATs)** introduce attention mechanisms to weigh neighbor importance differently, yet their focus on immediate neighbors limits multi-hop reasoning [8].

Subgraph-guided approaches to GNNs, such as those described by Chen et al. [6] and Li et al. [2], attempt to mitigate these issues by improving the model's ability to capture latent, complex relationships within KGs. However, these methods are still in their early stages, and the extraction of indirect relationships remains a major challenge in practice. Even advanced relationship prediction models like **GraphSAGE**, which have shown strong performance on benchmarks such as FB15K-237 and NELL-995, struggle to efficiently handle the sheer scale and complexity of indirect relationships in large KGs [2].

Explainable AI (XAI) is another critical aspect missing from current GNN systems. Although attention-based mechanisms in models like GATs provide some level of interpretability, as highlighted by Zhan et al. [9], they fall short for capturing and explaining **latent relationships** across multi-hop paths [10]. XAI remains underdeveloped in the context of multi-hop reasoning, with tools like GNNExplainer [10] offering

local insights but lacking scalability for large KGs [3]. As KG complexity grows, interpretable predictions become harder to achieve, particularly for hidden connections vital to bias detection [11].

In addition to the limitations in relationship inference and explainability, there is also an increasing demand for **secure, local deployment strategies**. Traditional cloud-based GNN systems introduce significant privacy risks, particularly when dealing with sensitive data, as noted by Zhan et al. [9] and Wu et al. [3]. While some advancements have been made in deploying GNNs on-premise, ensuring data security without compromising the model's inference performance remains a critical issue that has yet to be fully addressed in the literature [2].

Furthermore, the growing intersection of **Natural Language Processing (NLP)** and KGs has highlighted the need for more advanced GNN models capable of extracting latent relationships from unstructured text. Approaches such as **NLP4KGC**, as demonstrated by Vakaj et al. [11], attempt to bridge this gap by automating the construction of KGs from text. However, they still fall short when it comes to inferring complex, indirect relationships from this structured knowledge [12].

Finally, research into **relationship prediction** continues to demonstrate the limitations of current GNN-based systems in capturing indirect and missing relationships within KGs [2]. While progress has been made in specific applications, the broader challenge of efficiently identifying hidden relationships in large, fragmented datasets remains. As emphasized by Vakaj et al. [11] and Liu et al. [12], this represents a critical gap in the field, particularly for applications where uncovering latent relationships can significantly enhance AI-driven insights.

The integration of **large language models (LLMs)** with KGs has improved the accuracy of intelligent question-answering systems, but even these advancements have been limited by the inability to fully exploit latent relationships in large-scale KGs [9, 13, 14]. Despite these improvements, the challenge of scaling GNN-based systems for better inference of indirect relationships and their explainability remains, as emphasized by Wang et al. [15]. This is where the proposed PhD research aims to make a critical contribution.

2 Proposed Research

2.1 Research Question

How can Graph Neural Networks (GNNs) be advanced with novel bias-aware attention mechanisms to infer multi-hop latent relationships across multiple documents and Knowledge Graphs, identifying unique perspectives or biases, while incorporating Explainable AI (XAI) principles for transparent decision-making and ensuring scalable, secure, local deployment in privacy-sensitive domains?

2.2 Significance of the Research Question

The proposed research question is significant because it addresses critical gaps in GNNs and KGs by introducing the first framework to jointly model intra- and inter-document latent relationships with bias-aware attention, advancing beyond single-KG inference (Liu et al. [1]; Li et al. [12]). This capability is vital for uncovering hidden perspectives and biases in fragmented datasets—e.g., news corpora or web-scale document collections—aligning with goals like improving search relevance and fairness in content analysis. Current GNNs (e.g., GCNs, GATs) struggle with multi-hop reasoning across documents, a gap this work bridges with scalable architectures (Chen et al. [6]).

Detecting **perspectives** and **biases** is increasingly critical in domains like crime investigation, healthcare, and web search, where subtle biases in latent relationships can skew outcomes. This research offers tools to mitigate such biases, enhancing fairness in AI-driven insights—a priority for privacy-sensitive and equitable systems. Incorporating **XAI** ensures transparency (Zhou et al. [3]), while **secure, local deployment** leverages federated learning and differential privacy to protect sensitive data, aligning with privacy-preserving AI trends (Zhan et al. [9]). These advancements promise scalable, trustworthy solutions for real-world applications, such as bias-aware search engines or secure knowledge discovery.

Additionally, the detection of **perspectives** and **biases** in documents is a growing area of concern, particularly in domains like crime investigation, intelligence analysis, finance, and healthcare, where biased

information can significantly impact decision-making. Current AI systems lack the capability to effectively identify and interpret these biases, especially when they are subtly embedded in the latent relationships between entities across different documents. By extending GNN architectures to capture multi-hop latent relationships across documents, this research aims to fill this gap and provide tools for better understanding and mitigating bias.

Incorporating **Explainable AI (XAI)** principles is crucial for ensuring that the AI systems' decisions are transparent and interpretable. While attention-based models like **GATs** offer some level of explanation, recent works highlight the need for deeper integration of XAI principles to improve interpretability, especially when dealing with complex, cross-document relationships (Zhou et al. [3]; Zhan et al. [9]). This research will explore new avenues of XAI integration to ensure that AI models can explain their decisions effectively to end-users, facilitating trust and accountability.

Lastly, the emphasis on **secure, local deployment** is critical in privacy-sensitive environments to mitigate risks such as data leakage and breaches, which are common in cloud-based solutions. Security concerns surrounding AI systems that rely on cloud infrastructure have been a focal point of recent research, particularly when dealing with sensitive documents that may contain personal or proprietary information. This work will focus on developing secure, on-premise deployment strategies for GNNs, allowing AI systems to operate locally without compromising performance or risking data privacy (Zhou et al. [3]; Zhan et al. [9]).

In conclusion, this research aims to develop a holistic solution that tackles the challenges of latent relationship inference across multiple documents, explainability, and secure deployment, advancing the capabilities of GNNs and ensuring their applicability in real-world, sensitive applications where understanding perspectives and biases is essential.

2.3 Mathematical Formulation

The goal of this research is to extend **Graph Neural Networks (GNNs)** for inferring **multi-hop latent relationships** across **multiple documents** represented as interconnected **Knowledge Graphs (KGs)**, while integrating **Explainable AI (XAI)** principles and ensuring **secure, local deployment** in privacy-sensitive domains.

2.3.1 Multi-Document Knowledge Graph Construction

Each document $D^{(m)}$ is converted into a Knowledge Graph $G^{(m)} = (V^{(m)}, E^{(m)})$, where $V^{(m)}$ represents entities extracted from the document, and $E^{(m)}$ represents relationships between these entities. Entities may include authors, organizations, and topics, with features capturing semantic information.

An **integrated Knowledge Graph** $G = (V, E)$ is formed by merging individual graphs $G^{(m)}$ from multiple documents, creating inter-document edges based on shared entities or inferred relationships. This allows the model to capture latent relationships across documents.

2.3.2 Graph Representation

The integrated KG $G = (V, E)$ includes:

- Nodes $v_i \in V$ with feature vectors $h_i \in \mathbb{R}^d$.

- Edges $e_{ij} \in E$ representing relationships, which may include **edge features** f_{ij} capturing relationship types and contextual information.

2.3.3 Graph Neural Network Extensions

To capture latent relationships across multiple documents, we propose novel GNN extensions that uniquely separate intra- and inter-document relationships, enhancing multi-hop inference beyond existing models (e.g., GraphSAGE [2]). This involves modifying GCN and GAT architectures with bias-aware attention mechanisms, optimized for scalability on large KGs.

Graph Convolutional Networks (GCNs):

In the traditional GCN, each node updates its representation by aggregating information from its immediate neighbors within a single graph. To extend this to multiple documents, we modify the update rule to include both intra-document neighbors (within the same document) and inter-document neighbors (across different documents).

The updated node representation at layer $l+1$ is given by:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \frac{1}{\sqrt{d_i d_j}} A_{ij} W^{(l)} h_j^{(l)} + \sum_{k \in M(i)} \frac{1}{\sqrt{d_i d_k}} A_{ik} W^{(l)} h_k^{(l)}\right)$$

Where:

- $h_i^{(l+1)}$ is the updated feature vector of node v_i at layer $l+1$.
- σ is the activation function (e.g., ReLU).
- $N(i)$ denotes the set of intra-document neighbours of node v_i, i.e., nodes within the same document connected to v_i.
- $M(i)$ denotes the set of inter-document neighbours of node v_i, i.e., nodes in other documents connected to v_i.
- d_i is the degree of node v_i (number of edges connected to v_i), including both intra-document and inter-document edges.
- A_{ij} is the element of the adjacency matrix corresponding to the edge between nodes v_i and v_j. $A_{ij} = 1$ if an edge exists, 0 otherwise.
- $W^{(l)}$ is the weight matrix at layer l.
- $h_j^{(l)}$ and $h_k^{(l)}$ are the feature vectors of nodes v_j and v_k at layer l.

The first summation aggregates information from intra-document neighbours, while the second summation aggregates information from inter-document neighbours. The normalization term $\frac{1}{\sqrt{d_i d_j}}$ ensures that the contributions from each neighbour are appropriately scaled, preventing nodes with high degrees from dominating the aggregation.

Graph Attention Networks (GATs):

In GATs, attention mechanisms are used to assign different importance weights to neighbouring nodes during aggregation. We extend the GAT model to separately compute attention coefficients for intra-document and inter-document edges, allowing the model to learn the relative importance of relationships within and across documents.

The updated node representation at layer $l+1$ is given by:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)} + \sum_{k \in M(i)} \beta_{ik}^{(l)} W^{(l)} h_k^{(l)}\right)$$

Where:

- $\alpha_{ij}^{(l)}$ is the attention coefficient between node v_i and its intra-document neighbour v_j at layer l.
- $\beta_{ik}^{(l)}$ is the attention coefficient between node v_i and its inter-document neighbour v_k at layer l.
- $W^{(l)}$ is the shared weight matrix at layer l.

- Other terms are as previously defined.

The attention coefficients are computed using a shared attention mechanism but are applied separately for intra-document and inter-document edges.

For intra-document edges:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_j^{(l)}]\right)\right)}{\sum_{j' \in N(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_{j'}^{(l)}]\right)\right)}$$

For inter-document edges:

$$\beta_{ik}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{b}^T[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_k^{(l)}]\right)\right)}{\sum_{k' \in M(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{b}^T[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_{k'}^{(l)}]\right)\right)}$$

Where:

- $\mathbf{a}$ and $\mathbf{b}$ are learnable attention vectors for intra-document and inter-document edges, respectively.
- $[\cdot \,||\, \cdot]$ denotes the concatenation of two vectors.
- LeakyReLU is the activation function applied to introduce non-linearity.
- The denominators ensure that the attention coefficients sum to 1 over the respective sets of neighbours.

By computing separate attention coefficients for intra-document and inter-document edges, the model can learn to weigh the importance of relationships differently depending on whether they are within the same document or across different documents. This is crucial for detecting perspectives and biases, as inter-document relationships may reveal contrasting or reinforcing viewpoints.

Computation of Attention Coefficients Using Edge Features:

To capture the importance of relationships in detecting perspectives and biases, the attention coefficients incorporate edge features f_{ij} or f_{ik}, which may include information such as the type of relationship, the sentiment of the connection, or any domain-specific attributes.

The modified attention mechanism includes edge features:

For intra-document edges:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_j^{(l)} \,||\, f_{ij}]\right)\right)}{\sum_{j' \in N(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_{j'}^{(l)} \,||\, f_{ij'}]\right)\right)}$$

For inter-document edges:

$$\beta_{ik}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{b}^T[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_k^{(l)} \,||\, f_{ik}]\right)\right)}{\sum_{k' \in M(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{b}^T[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_{k'}^{(l)} \,||\, f_{ik'}]\right)\right)}$$

Including edge features allows the attention mechanism to consider not only the node features but also the nature of the relationships between nodes, which is essential for understanding how different connections contribute to the detection of perspectives and biases.

2.3.4 Modeling Perspectives and Biases

We define a **perspective function** $P(v_i)$ that maps node features to a latent space representing perspectives or biases:

$$P(v_i) = \phi(h_i)$$

- ϕ: A learnable function (e.g., neural network) that captures latent perspectives.
- The similarity of $P(v_i)$ across nodes indicates shared perspectives or biases.

2.3.5 Loss Function for Perspective and Bias Detection

We formulate a loss function that encourages the model to cluster nodes with similar perspectives and distinguish those with differing viewpoints:

$$L = L_{\text{prediction}} + \lambda L_{\text{perspective}}$$

- $L_{\text{prediction}}$: Loss for link prediction or classification tasks (e.g., cross-entropy loss).
- $L_{\text{perspective}}$: Regularization term that penalizes dissimilar perspectives for nodes known to share biases, defined as:

$$L_{\text{perspective}} = \sum_{(i,j)\in S} ||P(v_i) - P(v_j)||^2 - \sum_{(i,k)\in D} ||P(v_i) - P(v_k)||^2$$

- S: Pairs of nodes expected to share perspectives (e.g., documents from the same author).
- D: Pairs of nodes expected to have differing perspectives.
- λ: Hyperparameter balancing the two loss components.

2.3.6 Explainability Mechanisms

To ensure **explainability**, we enhance attention mechanisms through the following methods:

- **Attention Visualization**: Visualizing attention coefficients α_{ij} and β_{ik} to illustrate which relationships influence the model's decisions.
- **Feature Importance**: Using methods like **Integrated Gradients** to assess the contribution of node features to the output.
- **Post-Hoc Explanation**: Employing techniques such as **LIME** and **SHAP** adapted for graphs to provide local explanations.

2.3.7 Security and Privacy Considerations

For **secure, local deployment**, we incorporate:

- **Differential Privacy**: Applying mechanisms to ensure that the model's output does not compromise individual data points.
- **Federated Learning**: Enabling model training across multiple local nodes without sharing raw data.
- **Secure Aggregation**: Using cryptographic techniques to aggregate model updates securely.

2.4 Measurement and Evaluation

The success of the proposed research will be evaluated based on its ability to infer multi-hop latent relationships across documents, detect perspectives and biases, ensure explainability, and maintain secure deployment. The following metrics will be used:

2.4.1 Measuring Latent Relationship Inference Across Documents

- **Cross-Document Link Prediction Metrics**: Utilise **MRR** and **Hits@K** (e.g., Hits@10) on datasets like WikiNews or Twitter threads to evaluate inter-document relationship accuracy.
- **Perspective Detection Accuracy**: Measure precision/recall on labeled datasets (e.g., SemEval bias corpora) to assess bias identification.
- **Diversity Measures**: Use entropy-based metrics to quantify perspective variety across documents.

2.4.2 Measuring Explainability

- **Comprehensibility Scores**: Evaluate how easily domain experts can understand the explanations provided by the model.
- **Explanation Fidelity**: Measure the alignment between the model's explanations and its actual decision-making process.
- **User Studies**: Conduct surveys with domain experts to assess the usefulness of explanations in real-world tasks.

2.4.3 Measuring Security and Privacy

- **Privacy Loss (ϵ-Differential Privacy)**: Quantify the privacy guarantee provided by the model.
- **Resistance to Adversarial Attacks**: Evaluate the model's robustness against attempts to extract sensitive information.
- **Performance Overhead**: Measure the impact of security measures on model performance (e.g., inference time, accuracy).

2.4.4 Measuring Scalability

- **Scalability Metrics**: Track model performance as the number of documents and size of the integrated KG increase.
- **Resource Utilization**: Monitor memory usage and computational load during training and inference.
- **Convergence Rates**: Evaluate how quickly the model converges during training with increasing data size.

By systematically addressing these aspects, the research will ensure that the proposed models are practical, scalable, secure, and interpretable for real-world applications, particularly in detecting perspectives and biases across multiple documents.

3 Plan for Months 1-12 of the PhD

The first year will establish the research foundation via three phases:

- **Months 1-4 (Literature and Framework)**: Review GNNs, XAI, and bias detection; identify datasets (e.g., news articles); refine multi-hop inference framework.
- **Months 5-8 (Model Development)**: Build multi-document KGs; prototype GCN/GAT extensions with bias-aware attention.
- **Months 9-12 (Experiments)**: Test models on curated datasets; integrate initial XAI methods; define evaluation metrics (e.g., MRR, fidelity).

Expected outcomes: A refined framework, initial prototypes, and preliminary results by month 12, setting the stage for scalable, explainable AI advancements.

References

[1] X. Liu, Y. Su, and B. Xu, "The application of graph neural network in natural language processing and computer vision," in *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, 2021, pp. 708–714.

[2] J. Li, W. Xu, Y. Jin, and X. Wei, "Applying of graph neural network in relationship prediction in knowledge graph reasoning," in *2021 IEEE 23rd Int Conf on High Performance Computing Communications; 7th Int Conf on Data Science Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud Big Data Systems Application (HPCC/DSS/SmartCity/DependSys)*, 2021, pp. 2206–2210.

[3] J. Zhou, "The research and construction of the ai-based knowledge graph in multi-dimensional data," in *2023 International Conference on Computer Engineering and Distance Learning (CEDL)*, 2023, pp. 1–6.

[4] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3580–3599, 2024.

[5] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *International Conference on Machine Learning*, 2017. [Online]. Available: https://arxiv.org/abs/1704.01212

[6] Y. Chen, L. Wu, and M. J. Zaki, "Toward subgraph-guided knowledge graph question generation with graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 12 706–12 717, 2024.

[7] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and leman go neural: Higher-order graph neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4602–4609, 2019.

[8] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[9] L. Zhan and C. Huang, "Research on computer natural language processing intelligent question answering system based on knowledge graph," in *Proceedings of the 2024 International Conference on Machine Intelligence and Digital Applications*, ser. MIDA '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 70–74. [Online]. Available: https://doi.org/10.1145/3662739.3664744

[10] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Online]. Available: https://arxiv.org/abs/1903.03894

[11] E. Vakaj, S. Tiwari, N. Mihindukulasooriya, F. Ortiz-Rodríguez, and R. Mcgranaghan, "Nlp4kgc: Natural language processing for knowledge graph construction," in *Companion Proceedings of the ACM Web Conference 2023*, ser. WWW '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, p. 1111. [Online]. Available: https://doi.org/10.1145/3543873.3589746

[12] D. Liu, Y. Zhang, and Z. Li, "A survey of graph neural network methods for relation extraction," in *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 10, 2022, pp. 2209–2223.

[13] M. Guo, Y. Chen, J. Xu, and Y. Zhang, "Dynamic knowledge integration for natural language inference," in *2022 4th International Conference on Natural Language Processing (ICNLP)*, 2022, pp. 360–364.

[14] Q. Wu and Y. Wang, "Research on intelligent question-answering systems based on large language models and knowledge graphs," in *2023 16th International Symposium on Computational Intelligence and Design (ISCID)*, 2023, pp. 161–164.

[15] N. Wang, H. Yilahun, and A. Hamdulla, "Research on the construction and knowledge representation learning based on multi-modal knowledge graphs," in *Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering*, ser. ICAICE '23. New York, NY, USA: Association for Computing Machinery, 2024, p. 715–720. [Online]. Available: https://doi.org/10.1145/3652628.3652747