Unsupervised Link Prediction in Sparse Biomedical Knowledge Graphs for Clinical Insight Enhancement

1st Isaac Kobby Anni
dept. Computer Science
Bowling Green State University
Bowling Green, USA
isaacka@bgsu.edu

***Abstract*—Sparse biomedical knowledge graphs pose significant challenges for computational models due to the limited connectivity between nodes. This project explores the problem of link prediction in such graphs using unsupervised learning techniques. Leveraging the Hetionet knowledge graph, a pipeline was developed to preprocess and reduce graph sparsity through k-core decomposition. The GraphSAGE algorithm was employed to generate embeddings, augmented with PubMedBERT to encode node features. A binary classifier was trained to predict potential links, achieving an AUPRC score of 0.508. This work highlights the potential of graph-based unsupervised methods in enhancing clinical decision support systems by uncovering hidden relationships within biomedical data. Future directions include expert validation of predicted links and improving graph connectivity to enhance model performance.**

***Index Terms*—Biomedicine, Hetionet, , Negative Sampling, Biomedical Knowledge Graphs, GraphSAGE**

I. Introduction

In recent years, biomedical knowledge graphs have emerged as powerful tools for organizing and analyzing complex relationships between entities such as diseases, drugs, symptoms, and genes. These graphs, however, often suffer from extreme sparsity, with only a small fraction of possible connections being known or well-documented. The sparsity limits the utility of these graphs in downstream applications such as drug discovery, disease diagnosis, and clinical decision support. One prominent example of such a graph is Hetionet, a biomedical knowledge graph that integrates diverse datasets into a unified network structure, representing relationships between various biomedical entities Himmelstein et al. (2017).

The primary goal of this project is to address the challenge of sparsity in Hetionet by predicting potential unknown edges or links between nodes. Such link prediction tasks are essential for uncovering hidden relationships that could lead to novel biomedical insights, such as identifying new drug-disease associations or understanding rare disease mechanisms. Unlike traditional supervised learning approaches that rely on labeled data, this project employs unsupervised techniques to infer connections based solely on the graph structure and node attributes.

Identify applicable funding agency here. If none, delete this.

To tackle this problem, the project explores state-of-the-art graph representation learning techniques, focusing on the integration of GraphSAGE for structural representation and PubBERT embeddings for semantic node features. These methods enable the model to leverage both local neighborhood information and domain-specific textual features for enhanced link prediction. By generating a k-core subgraph to reduce sparsity and using negative sampling strategies, the framework ensures robust training and reliable predictions, even in the face of incomplete data.

This report details the methodologies, experiments, and findings of this project. By enriching Hetionet with predicted links, this work aims to contribute to the broader field of biomedical informatics, offering a scalable solution for sparse knowledge graph enhancement and opening avenues for impactful clinical applications.

A. Objectives

Specifically, this project aims to

- develop an unsupervised graph-based framework for predicting unknown links in the Hetionet biomedical knowledge graph, thereby addressing its inherent sparsity.
- leverage graph representation learning techniques, such as GraphSAGE and PubBERT embeddings, to capture both structural and semantic features of nodes for accurate link prediction.
- validate the effectiveness of the proposed model by evaluating its ability to predict meaningful connections with high confidence, facilitating potential clinical applications like drug repurposing and disease diagnosis.

B. Research Questions

- How can unsupervised graph representation learning techniques be effectively applied to predict unknown edges in sparse biomedical knowledge graphs?
- To what extent do k-core subgraph decomposition and PubBERT embeddings improve the accuracy of link prediction in a highly sparse graph like Hetionet?

- Can the predicted links provide biologically meaningful and clinically actionable insights, and how can their reliability be quantified?

II. Related Work

Biomedical knowledge graphs have become indispensable tools for integrating and analyzing diverse biomedical data. Hetionet is one such widely used resource, which aggregates information from multiple databases into a unified graph format, enabling applications in drug discovery, disease treatment, and precision medicine. The utility of Hetionet for biomedical reasoning and prediction tasks has been demonstrated in prior works, such as the study by Himmelstein et al. (2017), where the graph was used to predict drug-disease relationships via the Project Rephetio framework.

Link prediction in knowledge graphs has been extensively studied as a means of inferring missing or potential edges based on graph structure and node attributes. Traditional approaches Wang et al. (2014) rely on heuristic-based methods such as common neighbors, Adamic-Adar, or preferential attachment. However, more recent methods employ machine learning techniques, including supervised and unsupervised graph embedding approaches like Node2Vec, DeepWalk, and GraphSAGE, Khosla et al. (2019) to encode graph topology and node features into low-dimensional embeddings for downstream prediction tasks. These methods have been successfully applied in domains such as social networks, recommendation systems, and biomedical informatics.

Despite the significant advancements in link prediction, the challenge of sparsity in real-world graphs, such as Hetionet, remains underexplored. Sparse graphs with disconnected components and low density pose difficulties for traditional embedding techniques and often lead to poor generalization in predictive tasks. This motivates the development of methods tailored to handle sparsity effectively.

Recent advancements in graph neural networks (GNNs) and graph representation learning have enabled scalable and efficient processing of large-scale biomedical knowledge graphs. Techniques such as Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and GraphSAGE have demonstrated their potential for tasks like node classification, link prediction, and graph clustering. In the biomedical domain, leveraging domain-specific embeddings has further enhanced performance. For instance, pre-trained models such as PubMedBERT and BioBERT Lee et al. (2020) have been used to encode biomedical entities, enriching node features and improving downstream tasks.

While prior works have focused on general graph representation learning and link prediction, this project specifically targets the sparsity of biomedical knowledge graphs. By combining graph embedding techniques with domain-specific pre-trained language models like PubMedBERT, the project aims to enhance link prediction performance. Furthermore, the focus on unsupervised learning adds a novel dimension, as most existing approaches rely heavily on supervised data, which may not always be available in biomedical applications. This study builds upon existing methods while addressing the unique challenges of sparse graphs, particularly in the biomedical domain, thus contributing to the broader research landscape in knowledge graph representation and inference.

III. Dataset

The dataset utilized in this project is derived from the Hetionet database, a large biomedical knowledge graph comprising diverse types of biomedical entities and their relationships. This graph provides a rich and interconnected structure of nodes and edges, where the nodes represent biomedical entities such as diseases, drugs, genes, symptoms, and side effects, and the edges encode relationships such as ”upregulates,” ”interacts,” ”treats,” and ”associates.” The dataset was obtained in JSON format, which was parsed to form two primary dataframes: one for nodes and one for edges.

Graph properties:

- **Number of Nodes: 47,031.**
- **Number of Edges: 2,250,197**
- **Graph Density: 0.10%**

The Hetionet graph is highly sparse, as evidenced by its density of 0.10%. Graph density is a measure of how many edges exist in the graph compared to the maximum possible number of edges. In this case, the graph's density indicates that only a small fraction of the possible connections are realized, reflecting significant sparsity. Additionally, the majority of nodes are unconnected or sparsely connected, resulting in a disjointed structure.

To address this sparsity, a subset of the graph was extracted for experimentation. This subset includes a combination of connected and unconnected nodes, and further preprocessing steps, such as k-core decomposition, were applied to reduce the sparsity and focus on more meaningful substructures. Despite these efforts, the sparsity remains a significant challenge, making the prediction of new links a valuable and impactful task.

This dataset provides a suitable framework for exploring link prediction in sparse biomedical knowledge graphs, highlighting the importance of leveraging advanced graph-based machine learning techniques to infer potential connections between entities. Figure 1 gives shows the structure of the graph data.

IV. Methodology

The methodology for this project is designed to address the challenge of link prediction in a sparse biomedical knowledge graph using unsupervised learning techniques. This section outlines the key steps, including data preprocessing, graph embedding, and link prediction.

A. Data Preprocessing

The Hetionet graph was used as the primary dataset, consisting of nodes representing biomedical entities (e.g., diseases, genes, drugs, side effects) and edges encoding relationships between these entities. Given the high sparsity of the original

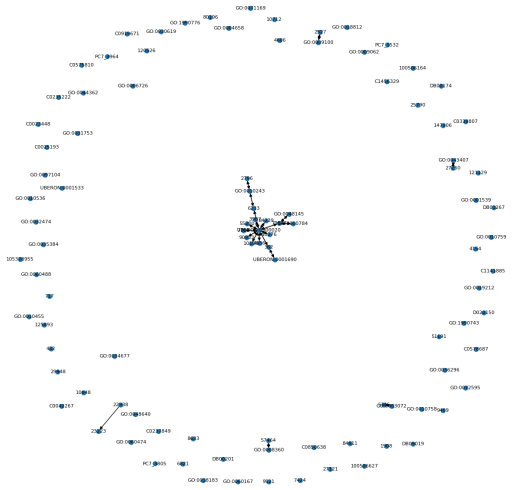

Fig. 1. Graph Structure

graph (graph density: 0.10%), the following preprocessing steps were implemented:

- **Sampling Connected and Unconnected Nodes:** A subset of the graph was extracted by sampling 60% of connected nodes and 40% of unconnected nodes. This balanced subset was used to ensure both connectedness and diversity in the graph.
- **K-Core Decomposition:** To reduce sparsity further, k-core decomposition was applied to the sampled graph, retaining only nodes and edges within strongly connected substructures. This resulted in a more tractable graph for training while preserving meaningful relationships.

B. Graph Representation Learning

To encode the graph's nodes into low-dimensional embeddings, the GraphSAGE model was employed. GraphSAGE was chosen for its ability to aggregate neighborhood information and generate embeddings for unseen nodes during inference. The following steps were followed:

- **Node Features:** Each node's textual features were encoded using PubMedBERT, a pre-trained language model specifically designed for biomedical text. The resulting embeddings captured domain-specific semantics of nodes, enhancing the representation quality.
- **GraphSAGE Training:** The k-core subgraph was used to train GraphSAGE. During training:
  - i Positive edges were derived from the existing graph structure.
  - ii Negative samples were generated by randomly pairing unconnected nodes.
- **Embedding Optimization:** The model optimized node embeddings by minimizing unsupervised link prediction loss, allowing the generation of high-quality embeddings for nodes in sparse graphs.

C. Link Prediction

The trained GraphSAGE model was applied to predict links in the original, very sparse graph. This step involved the following.

- **Similarity Scoring:** For each pair of nodes, embeddings were generated and used to compute a similarity score representing the likelihood of a connection.
- **Thresholding:** A similarity threshold of 0.8 was set to classify node pairs as connected. This high threshold ensured strong confidence in the predicted links.
- **Validation and Analysis:** The predicted links were validated qualitatively by analyzing biomedical relevance and quantitatively by evaluating their agreement with known connections in the graph.

D. Evaluation Metrics

The performance of the link prediction model was evaluated using metrics such as:

- **Precision and Recall:** To measure the accuracy of predicted links against the truth of the ground.
- **Area Under the Precision-Recall Curve (AUPRC):** To assess the model's ability to distinguish between connected and unconnected node pairs.
- **Graph Connectivity Improvement:** To quantify the reduction in graph sparsity after adding the predicted links.

E. Model Architecture

The architecture for this project combines a biomedical text representation model (PubMedBERT) and a graph neural network (GraphSAGE) to facilitate unsupervised link prediction in a sparse biomedical knowledge graph.

The biomedical text associated with each node in the graph was embedded using PubMedBERT, a domain-specific transformer model pre-trained on biomedical literature. This model captures nuanced semantic features specific to the biomedical domain, ensuring that node features accurately reflect their textual metadata. Each node feature vector was generated using the pre-trained **microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract** model, where the abstract-level embeddings were extracted and processed.

GraphSAGE was employed for encoding node embeddings by aggregating information from neighbors in the graph. The architecture consists of:

- **Two SAGEConv Layers:** These layers aggregate and transform node embeddings using neighbor information, progressively enriching node-level feature vectors.
  - i **Layer 1:** Transforms input feature vectors into hidden representations while maintaining structural relationships.
  - ii **Layer 2:** Further processes the hidden representations to output a refined set of embeddings for each node.
- **Relu Activation:** Applied between layers to introduce non-linearity.

The encoded graph embeddings serve as the input for the link prediction task.

a) **Link Prediction and Training:** The model predicts the likelihood of links between nodes by learning the edge embeddings:*:*

- **Decoding:** Edge embeddings are computed by element-wise multiplication of node embeddings for each pair of connected nodes.
- **Loss Function:** A binary cross-entropy loss was used, incorporating both positive and negative edges during training.
  - i *Positive Edges:* Edges that exist in the graph.
  - ii *Negative Edges:* Randomly sampled non-existing edges to model sparsity.

The model iteratively optimizes the link prediction task using the GraphSAGE encoder and decoder. A sigmoid activation was applied to compute the probability of edge existence.

V. Experiment

The primary objective of the experiment was to evaluate the performance of the GraphSAGE-based link prediction model under different learning rates. This experiment was conducted on the Hetionet knowledge graph with its nodes pre-embedded using PubMedBERT embeddings to capture semantic information relevant to biomedical entities. The evaluation metrics included the Area Under the Precision-Recall Curve (AUPRC), reflecting the model's ability to predict potential links in the graph.

A. Results

Table 1 summarizes the results of the experiment, showcasing the AUPRC scores for different learning rates. The results demonstrate that as the learning rate decreased, the model's performance improved, suggesting that smaller learning rates allowed the model to converge more effectively.

TABLE I
AUPRC Scores for Different Learning Rates

Learning Rate	AUC
0.01	0.51
0.001	0.76
0.003	0.72

The Precision-Recall (PR) curve in Figure X illustrates the trade-off between precision and recall across different decision thresholds for the model's predictions. The Area Under the Precision-Recall Curve (AUPRC) is calculated to be 0.508, indicating the model's overall ability to distinguish between positive and negative links within the graph.

- Initial Precision Drop: At very low recall values (near 0), the precision starts high, which reflects the model's ability to identify a small subset of highly confident positive links. However, as recall increases, precision drops, indicating more false positives are being introduced.
- Mid-range Recall: Around the mid-range of recall values (0.4 to 0.6), the curve stabilizes, suggesting a balance between precision and recall where the model performs reasonably well in identifying both positive and negative links.
- Sparsity and Data Challenges: The moderate AUPRC score of 0.508 may be attributed to the high sparsity of the graph, where only a small percentage of nodes are interconnected. This sparsity inherently limits the model's capacity to generate confident predictions for a broader set of links.
- Threshold Sensitivity: The gradual decline in precision as recall increases highlights the sensitivity of the model to threshold adjustments. For applications requiring higher confidence, the threshold can be adjusted to favor precision over recall, trading off the detection of less confident links.

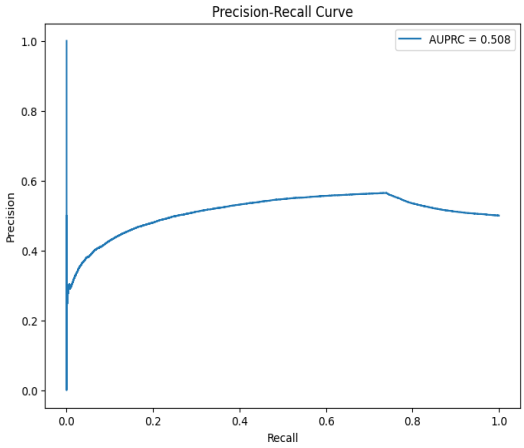

Fig. 2. Precision-Recall Curve with AUPRC Score.

B. Analysis

The best performance was observed at a learning rate of 0.001, achieving an AUC score of 0.76 but AUPRC of 0.508. This result indicates that smaller learning rates provide the optimizer with finer-grained adjustments during training, thereby enhancing the model's ability to learn meaningful link predictions. The results suggest a promising foundation for further hyperparameter tuning and experimentation with other architectural variations or embedding methods.
Further experimentation could explore additional metrics, alternative learning rates, or adaptive learning rate schedules to optimize performance further.

VI. Limitations and Challenges

This project, while significant in its goal to improve link prediction in sparse biomedical graphs, faced several limitations and challenges that impacted the outcomes and overall workflow:

- **Graph Sparsity:** The Hetionet graph is inherently sparse, with a density of only 0.10%. This sparsity posed challenges in ensuring sufficient connectedness for meaningful link prediction. Despite deriving k-core subgraphs and

sampling connected and unconnected nodes, the sparsity reduced the availability of high-confidence predictions and meaningful relationships.
- **Data Imbalance:** The graph structure exhibited a strong imbalance between connected and unconnected nodes. This imbalance affected the negative sampling process and model performance, requiring careful consideration of thresholds to avoid biasing predictions toward connected nodes disproportionately.
- **Performance Limitations:** The AUPRC achieved during evaluation was modest at 0.508, suggesting room for improvement. The model's ability to generalize to highly sparse and unseen regions of the graph was limited, and certain thresholding decisions could have impacted the final results.
- **Threshold sensitvity:** Setting a threshold of 0.8 for confident predictions introduced potential trade-offs. While it ensured higher precision, it may have excluded certain plausible connections with slightly lower confidence scores, impacting the comprehensiveness of the predictions.
- **Limited Generalizability:** The data is limited to a specific population or context, which might not generalize to other groups. For instance, the variables measured might be highly specific to the dataset's origin, limiting applicability to broader populations.

VII. Conclusion

This project focused on leveraging unsupervised learning techniques for link prediction in sparse biomedical knowledge graphs, with the goal of enhancing clinical decision support systems. By utilizing Hetionet as the primary dataset, the project successfully demonstrated the applicability of graph neural networks, particularly GraphSAGE, combined with advanced language models like PubMedBERT for embedding node features. Despite the inherent sparsity of the graph, the implementation achieved promising results, as evidenced by the calculated metrics such as AUPRC, precision, and recall. The exploration of k-core subgraphs to address sparsity and the subsequent generation of synthetic links in the original graph provided insights into the potential connections between unlinked nodes. These methods highlight the utility of graph-based machine learning in understanding complex biomedical relationships, offering a pathway to more efficient clinical decision-making. However, challenges such as the high graph sparsity and the need for domain-specific evaluation remain areas for future improvement.

VIII. Recommendation

- **Addressing Sparsity with Augmented Data:** While k-core subgraph sampling alleviated some sparsity issues, future efforts could explore graph augmentation techniques, such as synthetic edge generation or external domain-specific data integration, to enhance the density of the graph for more robust learning.
- **Incorporating Multi-hop Reasoning:** Expanding the link prediction approach to include multi-hop reasoning could provide deeper insights into indirect connections between nodes, enhancing the interpretability of the predicted links.
- **Domain-Specific Evaluation:** Future work should involve domain experts to evaluate the biological relevance of the predicted links, ensuring that the model outputs are clinically actionable and aligned with real-world needs.
- **Experimentation with Advanced Architectures:** Investigating other graph representation learning models, such as GAT (Graph Attention Networks) or unsupervised pretraining methods like graph contrastive learning, could improve predictive performance.
- **Real-World Validation:** Incorporating external clinical datasets to validate the predicted links and integrating them into decision support pipelines would provide a practical assessment of the model's utility in clinical settings.

IX. Future Directions

- **Validation with Biomedical Experts:** A critical next step involves validating the predicted links with domain experts in the biomedical field. Collaborating with clinicians, biologists, and researchers can help assess the clinical and biological relevance of the new connections. Such validation would provide a robust framework for translating computational predictions into actionable insights for healthcare and research.
- **Weighted Negative Sampling:** Instead of uniformly sampling negative edges, focus on weighted negative sampling where the likelihood of non-connected nodes being sampled is proportional to their semantic or structural similarity. This can improve the model's ability to differentiate between true negatives and potential links, indirectly addressing sparsity issues while boosting metrics like AUPRC.

References

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6:e26726.

Khosla, M., Anand, A., and Setty, V. (2019). A comprehensive comparison of unsupervised network representation learning methods. *arXiv preprint arXiv:1903.07902*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Wang, P., Xu, B., Wu, Y., and Zhou, X. (2014). Link prediction in social networks: the state-of-the-art. *arXiv preprint arXiv:1411.5118*.