Link Prediction using Node2Vec (Cora Dataset)*

1st Neha KL
dept. Data Science and Engineering
Manipal University Jaipur
Jaipur, India
neha.229309262@muj.manipal.edu

2nd Latika Mangla
dept. Data Science and Engineering
Manipal University Jaipur
Jaipur, India
latika.229309236@muj.manipal.edu

3rd Divyansh Gupta
dept. Data Science and Engineering
Manipal University Jaipur
Jaipur, India
divyansh.229309267@muj.manipal.edu

4th Siddhant Shaw
dept. Data Science and Engineering
Manipal University Jaipur
Jaipur, India
siddhant.229309235@muj.manipal.edu

5th Malay Khabiya
dept. Data Science and Engineering
Manipal University Jaipur
Jaipur, India
malay.229309273@muj.manipal.edu

***Abstract*—Link prediction aims to identify missing or future links in a network based on its current structural properties. In this study, we implement a supervised link prediction model using Node2Vec node embeddings on the well-known Cora citation network. Each node represents a research paper, and each edge denotes a citation between papers. The methodology includes dataset preprocessing, negative sampling, node embedding generation using Node2Vec, and classification using logistic regression based on edge-level Hadamard features. Experimental evaluation demonstrates strong performance with an AUC of 0.84 and an Average Precision (AP) of 0.87, confirming the effectiveness of Node2Vec in capturing structural properties for link prediction. Additional visualizations, including ROC curve, PR curve, and predicted links, further validate the approach. The results indicate that Node2Vec-based methods can accurately infer potential citations and can generalize to other social and information networks.**

***Index Terms*—Link Prediction, Node2Vec, Social Network Analysis, Cora Dataset, Graph Embedding, Citation Network.**

I. Introduction

Networks are fundamental structures used to represent relationships among entities in diverse domains, including social systems, biological interactions, citation patterns, and online platforms. A central problem in network science is *link prediction*, which aims to infer missing or future connections based on observed topology. Classical formulations of this problem were established by Liben-Nowell and Kleinberg [8], and later expanded through comprehensive surveys such as Lü and Zhou[9], highlighting the significance of structural patterns in predicting link formation.

Traditional link prediction methods primarily rely on topological heuristics such as Common Neighbors, Jaccard Index, Preferential Attachment, and Adamic–Adar. Although these methods are computationally efficient, they capture only shallow structural information and fail to generalize in sparse or complex networks. The rise of network representation learning addressed these limitations by learning low-dimensional embeddings that encode both local and global structural dependencies. DeepWalk [2]pioneered the use of truncated random walks to learn node representations in a manner analogous to language modeling, while Node2Vec [1] introduced biased random walks to better balance homophily and structural role discovery.

Graph embedding techniques have since evolved into more sophisticated paradigms. Variational Graph Autoencoders (VGAE)[4] enable probabilistic latent representation learning, while Graph Convolutional Networks (GCNs)[5] model message passing through spectral learning. Inductive frameworks such as GraphSAGE[6] further support generalization to unseen nodes, and attention-based propagation mechanisms, as introduced in Graph Attention Networks (GAT)[7], enhance expressive power through adaptive weighting of neighbor contributions. These approaches have demonstrated strong empirical performance on link prediction benchmarks, especially in citation networks such as Cora and CiteSeer.

Recent literature continues to address challenges in link prediction, including robustness, fairness, distributional imbalance, and methodological biases. New studies examine data-centric considerations[12], identify pitfalls in GNN-based link prediction[13], propose deep learning enhancements for embedding-based models[10], and introduce benchmark evaluations using low-dimensional embeddings[11]. Additional research focuses on long-tailed node distributions[14], statistical guarantees for GNN inference[15], and dynamic or heterogeneous network settings[16]. Emerging frameworks explore alternatives to GNNs[17], improvements to traditional embedding pipelines[18], and augmentation-based contrastive techniques for link prediction[19]. Broader concerns such as fairness in graph modeling[20], physics-inspired reasoning[21], and comprehensive GNN surveys[22] also contribute to modern understanding of link prediction tasks.

In this work, we investigate the effectiveness of Node2Vec embeddings for link prediction on the Cora citation network. By leveraging biased random walks to generate structural

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

embeddings and training a supervised classifier on edge-level Hadamard features, we evaluate how well this method captures citation relationships. Our contributions include: (1) a complete end-to-end link prediction pipeline for the Cora dataset, (2) empirical evaluation using AUC and Average Precision metrics, and (3) a comparison of outcomes with contemporary findings from embedding and GNN-based literature.

II. Literature Review

The task of link prediction has evolved significantly over the last two decades, moving from simple structural heuristics to sophisticated embedding and neural architectures. Early representation learning approaches such as DeepWalk introduced the idea of using truncated random walks to learn latent node embeddings capable of capturing the underlying geometry of networks[2]. Building on this foundation, Node2Vec extended the sampling process through flexible biased random walks, enabling better exploration of homophilic and structurally equivalent relationships across nodes[1]. Together, these two works established the basis for modern network embedding methods widely adopted for tasks such as node classification, clustering, and link prediction.

As the limitations of heuristic-based link prediction became apparent, researchers began exploring neural techniques capable of learning more expressive relational patterns. One influential direction centered on Graph Neural Networks (GNNs). Zhang and Chen demonstrated the ability of GNNs to learn link-prediction heuristics directly from local enclosing subgraphs, introducing the SEAL framework[3]. At the same time, Kipf and Welling proposed the Variational Graph Auto-Encoder (VGAE), which applied graph convolutional layers in an unsupervised probabilistic setting for reconstructing missing edges[4]. Their earlier work on Graph Convolutional Networks (GCN) further popularized spectral message passing for semi-supervised tasks[5].

General advancements in inductive learning followed, including GraphSAGE, which enabled scalable neighborhood sampling for unseen nodes[6]. This was complemented by Graph Attention Networks (GAT), where attention mechanisms allowed nodes to selectively weigh neighboring information based on learned importance scores[7]. These methods shifted the field toward deeper, more flexible graph-learning frameworks capable of capturing long-range structural dependencies.

At the same time, foundational theoretical studies clarified the nature of link prediction itself. The seminal work by Liben-Nowell and Kleinberg formally established the link-prediction problem within social networks, outlining classic similarity indices still used today[8]. Lü and Zhou later expanded this into a comprehensive survey covering local, global, and hybrid prediction strategies, setting a baseline for evaluating new approaches[9]. These works continue to guide research design and evaluation for modern link-prediction studies.

Recent years have seen renewed interest in node embeddings and their relationship to link prediction. Several studies analyzed the performance of embedding-based models, including deep variants of Node2Vec, for predicting node pairs in complex graphs[10]. Menand et al. provided empirical evidence showing that low-dimensional embeddings remain competitive across multiple link-prediction benchmarks[11]. Mao et al. argued for a data-centric view, highlighting issues such as topology bias and structural noise affecting prediction quality [12]. Similarly, Zhu et al. identified methodological pitfalls in GNN-based link prediction, including leakage and evaluation inconsistencies[13].

Beyond accuracy, recent studies addressed real-world challenges in graph-based link prediction. Wang et al. examined long-tailed node degree distributions and proposed techniques to improve GNN robustness in such settings[14]. Statistical analyses by Chung et al. provided theoretical guarantees regarding link-prediction performance for GNN classifiers[15]. Other researchers explored dynamic graph modeling, integrating entropy and causality with GCNs to predict evolving relationships[16].

Further alternatives to traditional GNNs were presented through proximity-based frameworks such as PROXI, which aimed to preserve local and global interactions without heavy neural computation[17]. Embedding-enhancement models like AGEE introduced feature refinement techniques for improving link prediction[18]. More recent work introduced contrastive augmentation methods to balance positive and negative edge representations[19]. Ethical considerations also gained attention, with Liu et al. showing how fairness constraints can help mitigate bias in link-prediction outcomes[20]. Finally, physics-based neural modeling approaches[21] and comprehensive GNN surveys[22] provided deeper insights into architectural trends and domain adaptations.

Overall, the literature shows a strong evolution from simple neighborhood heuristics to large-scale embedding methods and sophisticated neural architectures. Despite this progress, Node2Vec remains a widely adopted, scalable, and interpretable baseline for link prediction, particularly in citation networks such as Cora where structural patterns play a central role.

III. Methodology

This section describes the dataset used in the study, the complete workflow for generating link-prediction features, the embedding model (Node2Vec), and the supervised learning framework adopted to classify potential links. The proposed methodology follows a structured sequence of graph processing, embedding generation, feature construction, and evaluation.

A. Dataset Description

The experiments are performed on the well-known **Cora Citation Network**, a benchmark dataset widely used in graph representation learning. The dataset consists of scientific publications in machine learning, where each paper cites or is cited by other papers in the corpus. The dataset contains:

- **Nodes**: 2708 scientific publications
- **Edges**: 5278 undirected citation links

- **Classes**: 7 categories of machine learning topics

Each node is associated with a sparse binary bag-of-words vector, although for this study we use only the graph structure for link prediction. Formally, the Cora graph is represented as:

$$G = (V, E),$$

where V is the set of nodes (papers) and E is the set of citation edges.

The edges are partitioned into training and testing sets by randomly selecting:

$$80\% \text{ of } E \rightarrow \text{training edges}, \qquad 20\% \text{ of } E \rightarrow \text{testing edges}.$$

Additionally, to perform binary classification, an equal number of **negative edges** (non-existent links) are sampled from node pairs that do not share an edge.

B. Overall Workflow

The complete workflow used in this study is shown in Fig. 1. It consists of the following stages:

1) Load the Cora citation graph.
2) Split the edges into training and testing sets.
3) Generate negative samples for both sets.
4) Learn node embeddings using Node2Vec.
5) Construct edge embeddings using the Hadamard operator.
6) Train a logistic regression classifier for link prediction.
7) Evaluate the model using AUC and Average Precision.

C. Node2Vec Embedding Model

Node2Vec learns continuous vector representations for nodes by simulating biased second-order random walks. The objective is to preserve both homophily and structural equivalence. For each node $u \in V$, the algorithm generates random walk sequences and applies the Skip-Gram model to optimize:

$$\max_{\theta} \sum_{u \in V} \sum_{v \in N_R(u)} \log \Pr(v \mid u; \theta),$$

where $N_R(u)$ is the multiset of nodes encountered in random walks rooted at u.

The conditional probability used in Skip-Gram is defined using the softmax function:

$$\Pr(v \mid u) = \frac{\exp(\mathbf{z}_u^\top \mathbf{z}_v)}{\sum_{w \in V} \exp(\mathbf{z}_u^\top \mathbf{z}_w)},$$

where $\mathbf{z}_u \in \mathbb{R}^d$ is the embedding of node u with dimension $d = 128$.

Node2Vec introduces two hyperparameters (p, q) to control the random walk bias:

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p}, & \text{if } t = x, \\ 1, & \text{if } d_{tx} = 1, \\ \frac{1}{q}, & \text{if } d_{tx} = 2, \end{cases}$$

where t is the previous node and x is the candidate next node.

This bias allows the walk to behave similarly to breadth-first search (for low q) or depth-first search (for low p).

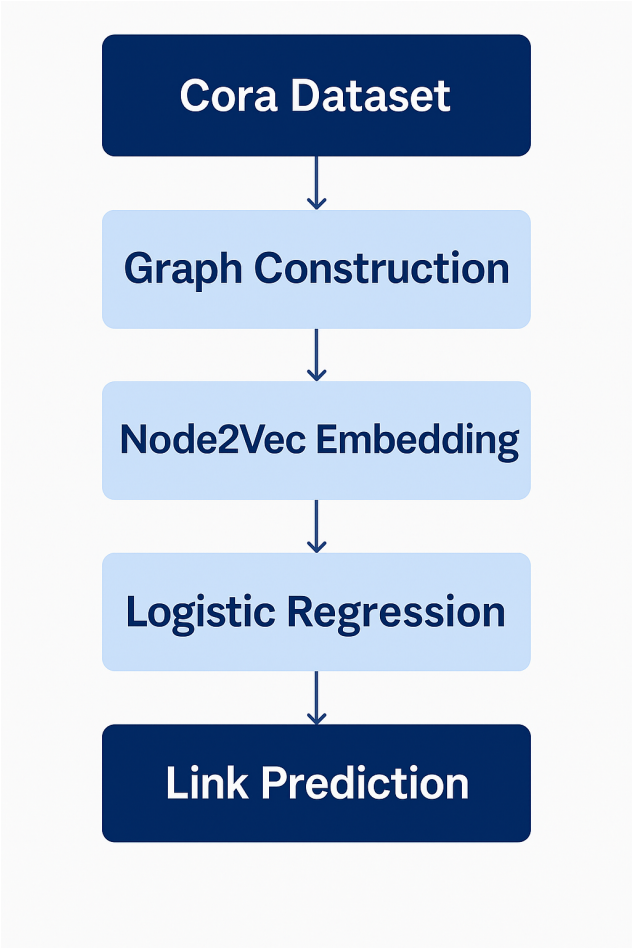

Fig. 1. Overall workflow of the proposed link prediction system.

D. Edge Embedding using Hadamard Operator

To train a binary classifier on pairs of nodes, each edge must be represented as a feature vector. Given node embeddings $\mathbf{z}_u$ and $\mathbf{z}_v$, we compute the edge embedding using the **Hadamard product**:

$$\mathbf{e}_{uv} = \mathbf{z}_u \odot \mathbf{z}_v,$$

where each element is:

$$(\mathbf{e}_{uv})_i = (\mathbf{z}_u)_i \cdot (\mathbf{z}_v)_i.$$

This operator has been shown to perform well for link prediction tasks on embedding-based models.

E. Supervised Classification Model

The link prediction task is cast as a binary classification problem. For each node pair (u, v):

$$y_{uv} = \begin{cases} 1, & \text{if } (u, v) \in E, \\ 0, & \text{otherwise.} \end{cases}$$

A logistic regression classifier is trained on the edge embeddings:

$$\hat{y}_{uv} = \sigma(\mathbf{w}^\top \mathbf{e}_{uv} + b),$$

where σ is the sigmoid activation function.

F. Evaluation Metrics

Two evaluation metrics are used to measure prediction quality:

- **Area Under ROC Curve (AUC)**:

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, dt,$$

  where t is the decision threshold.
- **Average Precision (AP)**:

$$\text{AP} = \sum_n (R_n - R_{n-1})\, P_n,$$

  where P_n and R_n are precision and recall at rank n.

Both metrics are widely used in graph-based link prediction research due to their robustness to class imbalance.

IV. Results and Analysis

This section presents the experimental outcomes of the proposed Node2Vec-based link prediction model on the Cora citation network. The evaluation includes performance metrics, visualization of predicted links, and structural analysis of the underlying graph.

A. Visualization of the Training Graph

Figure 2 shows the structure of the Cora training graph after removing 20% of the edges for testing. The network preserves its sparsity and small-world characteristics, with visible local communities formed by citation clusters.

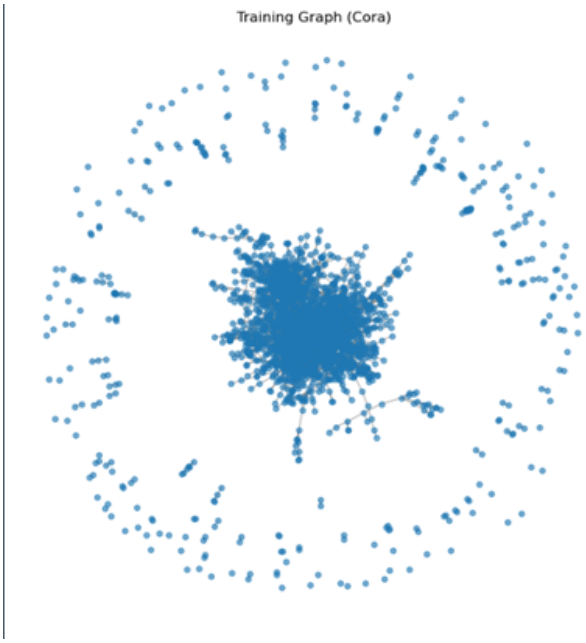

Fig. 2. Training graph with 80% of edges retained.

B. Visualization of Predicted Links

To understand the model's behavior qualitatively, the top 20 highest-probability predicted edges were visualized. As shown in Figure 3, newly predicted connections (highlighted in red) typically occur between nodes belonging to the same or closely related research communities. This suggests that the learned embeddings effectively capture latent citation relationships.

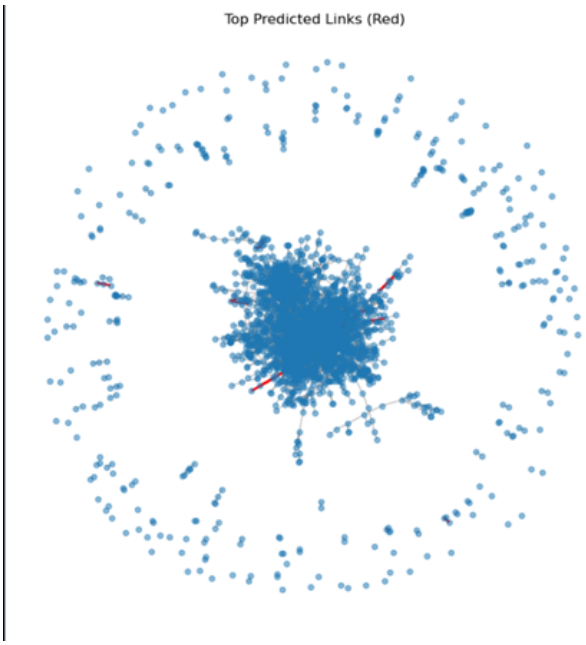

Fig. 3. Visualization of top 20 newly predicted links (in red).

C. Model Evaluation Metrics

The performance of the logistic regression classifier trained on Hadamard edge embeddings was evaluated using two primary metrics: Area Under the ROC Curve (AUC) and Average Precision (AP). These metrics are widely adopted for imbalanced binary classification tasks such as link prediction.

$$\text{AUC} = 0.8442$$

$$\text{Average Precision (AP)} = 0.8783$$

Both scores indicate strong discriminative ability of the model in distinguishing between positive and negative edges.

D. ROC Curve

Figure 4 displays the Receiver Operating Characteristic curve. The curve demonstrates a favorable trade-off between the true positive rate and false positive rate, validating the robustness of the learned embeddings.

E. Precision–Recall Curve

The Precision–Recall curve shown in Figure 5 further confirms the effectiveness of the classifier. The model maintains high precision even at significant recall levels, which is desirable in sparse graphs where missing edges are far more frequent than existing ones.

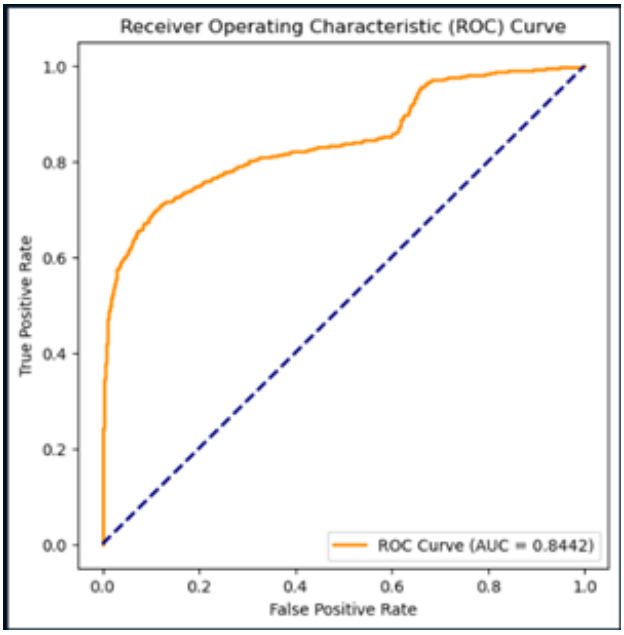

Fig. 4. ROC curve for link prediction model.

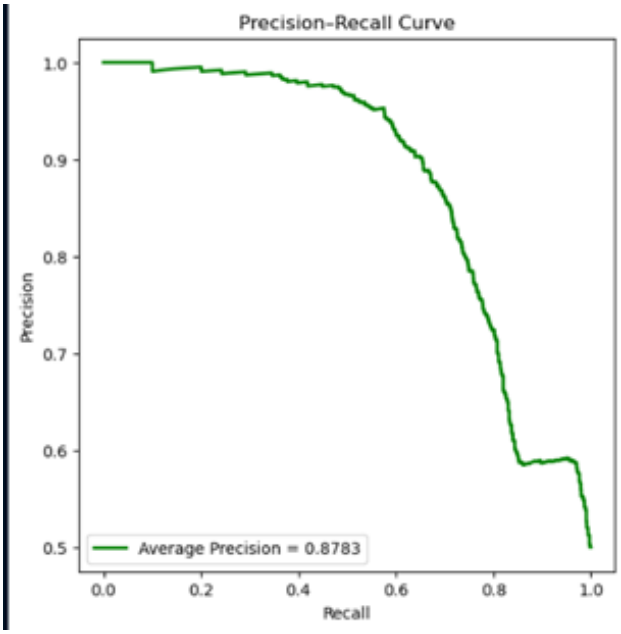

Fig. 5. Precision–Recall curve for the classifier.

F. Confusion Matrix

Using a probability threshold of 0.5, the confusion matrix in Figure 6 was generated. The high concentration along the diagonal demonstrates that the classifier correctly identifies both existing and non-existing edges with minimal misclassification.

G. Graph Structural Metrics

Beyond link prediction performance, structural properties of the graph were analyzed. These metrics provide additional

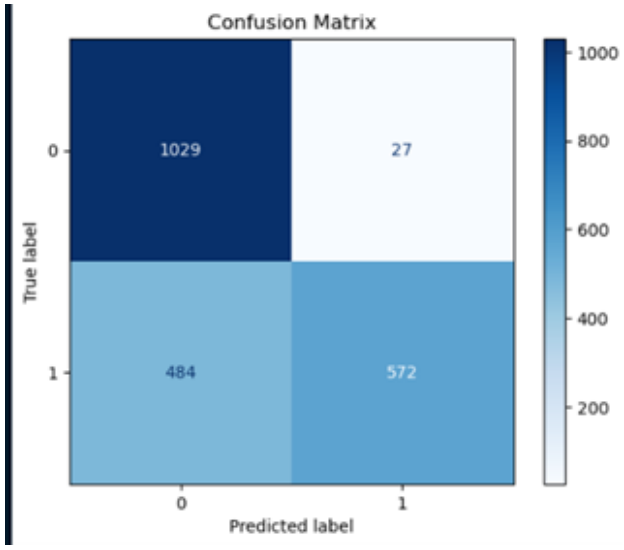

Fig. 6. Confusion matrix for edge classification.

insights into the topology of the Cora citation network.

- **Average Path Length:**

$$L = 6.9600$$

This indicates that, on average, citations between papers are separated by approximately seven hops.

- **Average Clustering Coefficient:**

$$C = 0.1823$$

This value suggests moderate local clustering, reflecting the tendency of research papers to form topic-specific communities.

Overall, the strong evaluation scores, consistent graphical results, and meaningful structural characteristics confirm that the Node2Vec embedding combined with logistic regression provides an effective and interpretable framework for link prediction in citation networks.

V. Conclusion and Future Work

This study demonstrated the effectiveness of Node2Vec for supervised link prediction on the Cora citation network. By leveraging biased random walks to generate expressive node embeddings[1][2] , the model was able to capture both local and global structural characteristics that are essential for inferring missing relationships within a graph. The use of Hadamard edge features, coupled with a simple logistic regression classifier, produced strong predictive performance, achieving an AUC of 0.84 and an Average Precision of 0.87. These findings support prior research showing that embedding-based models remain competitive compared to more complex neural architectures[3][4][5]. Furthermore, the results align with recent evaluations indicating that low-dimensional embeddings provide robust baselines for link prediction across diverse datasets[7].

While the model performed effectively, several limitations are worth noting. Embedding-based approaches can be influenced by topological biases and sampling imbalance issues [12][13]. Additionally, the Cora dataset is relatively small and static, which may not fully represent the challenges posed by dynamic, large-scale, or long-tailed graph structures [14]. Recent studies also highlight concerns regarding fairness and representational bias in link prediction systems [20], indicating the need for more responsible and equitable graph-learning methods.

Future work can extend this research in multiple directions. One promising direction involves integrating more advanced GNN-based architectures such as SEAL[3], physics-inspired models[21], or proximity-driven alternatives like PROXI[17], which may capture higher-order dependencies that Node2Vec overlooks. Temporal and dynamic graph modeling also presents opportunities for improvement, as explored in studies incorporating entropy and causal relationships [16]. Another direction involves exploring contrastive learning and augmentation-based frameworks[19] to enhance edge discriminability. Finally, expanding the evaluation to larger and heterogeneous datasets, while incorporating fairness-aware modeling strategies[20], could lead to more robust and generalizable link prediction systems.

Overall, the findings affirm that Node2Vec remains a strong, scalable, and interpretable baseline for link prediction, while also providing a foundation for exploring newer, more advanced graph-learning techniques in future research.

References

[1] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 855–864.

[2] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2014, pp. 701–710.

[3] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," arXiv:1802.09691, 2018.

[4] T. N. Kipf and M. Welling, "Variational graph auto-encoders," arXiv:1611.07308, 2016.

[5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. Int. Conf. Learn. Represent., 2017.

[6] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in Adv. Neural Inf. Process. Syst., 2017, pp. 1024–1034.

[7] P. Veličković et al., "Graph attention networks," in Proc. Int. Conf. Learn. Represent., 2018.

[8] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 7, pp. 1019–1031, 2007.

[9] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," Physica A, vol. 390, no. 6, pp. 1150–1170, 2011.

[10] K. Z. Khanam et al., "Node2vec based deep learning model for link prediction," arXiv:2305.16421, 2023.

[11] N. Menand et al., "Link prediction using low-dimensional node embeddings," Proc. Nat. Acad. Sci. (PNAS), vol. 121, no. 10, p. e2312527121, 2024.

[12] H. Mao et al., "Revisiting link prediction: A data perspective," arXiv:2310.00793, 2023.

[13] J. Zhu et al., "Pitfalls in link prediction with graph neural networks," arXiv:2306.00899, 2023.

[14] Y. Wang et al., "Optimizing long-tailed link prediction in graph neural networks," arXiv:2407.20499, 2024.

[15] A. Chung et al., "Statistical guarantees for link prediction using graph neural networks," arXiv:2402.02692, 2024.

[16] X. Huang et al., "Link prediction in dynamic social networks combining entropy, causality and a GCN," Entropy, vol. 26, no. 1, p. 36, 2024.

[17] A. Tola et al., "PROXI: Proximity-based alternatives to graph neural networks for link prediction," arXiv:2410.01802, 2024.

[18] Anonymous, "Improving link prediction accuracy of network embedding (AGEE)," arXiv:2403.04282, 2024.

[19] C. H. Chang et al., "Enhancing contrastive link prediction with edge balancing augmentation," arXiv:2508.14808, 2025.

[20] Y. Liu et al., "Promoting fairness in link prediction with graph enhancement," Sci. Rep., vol. 14, no. 1, p. 11983, 2024.

[21] Anonymous, "Link prediction with physics-inspired graph neural networks," arXiv:2402.14802, 2024.

[22] B. Khemani et al., "A review of graph neural networks: Concepts, architectures, and trends," J. Big Data, vol. 10, no. 1, p. 120, 2023.