International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Reevaluating Link Prediction and New Approaches and Best Practices

Dr. Ugranada Channabasava[1], Dr. Raghu Nandan R[2]

[1]*Department of Artificial Intelligence and Data Science, Global Academy of Technology,Bangalore*
uchannabasava@gat.ac.in
[2]*Department of Computer science and Engineering, Navkis College of Engineering, Hassan.*
raghu.siet@rediffmail.com

Abstract –

Link prediction (LP) is a crucial problem in network science and machine learning. However, existing evaluation methods often adopt a standardized approach, overlooking key factors that influence LP performance across different applications. In this study, we identify several important factors, including network type, problem type, geodesic distance between end nodes and its class distribution, the applicability of LP methods, class imbalance and its impact on early retrieval, and evaluation metrics. To address these challenges, we propose a rigorous experimental framework that accounts for these factors, ensuring a more controlled and comprehensive evaluation of LP methods. We conduct extensive experiments on real-world network datasets to analyze the interactions between these factors and LP performance. By testing a series of carefully designed hypotheses, we uncover valuable insights into how these elements influence LP outcomes. Based on our findings, we provide a set of best practices for evaluating LP methods, ensuring more reliable and meaningful comparisons. Our study highlights the need for a nuanced evaluation approach that goes beyond conventional setups, ultimately improving the robustness and applicability of LP techniques across diverse domains.

Key Words: Link Prediction (LP), Network Science, Evaluation Framework, Class Imbalance, Performance Analysis

1.INTRODUCTION

Complex networks depict intricate and significant relationships among entities and are widely observed across various domains, serving as models for real-world systems. Examples include social, biological, and information networks. Social networks represent interactions between individuals, biological networks illustrate associations between biological entities such as proteins, and information networks track data exchanges, like emails. In these networks, entities are referred to as nodes, while connections, interactions, or data exchanges between them are termed links. A graph-based data structure is commonly used to represent these networks, where nodes correspond to vertices and links are depicted as edges.

Link prediction (LP) is a fundamental challenge in network science, focusing on identifying missing, hidden, unobserved, or future links within a network. The specific nature of these predictions is influenced by the properties of the network and its intended application.

LP methods are generally divided into two categories: similarity-based and machine learning-based techniques. Similarity-based methods assign a score that reflects the likelihood of a connection between two nodes based on their structural proximity. These techniques are further classified as local or global. Local similarity methods utilize the triadic closure principle, which posits that two nodes without a direct link are more likely to establish one if they share at least one common neighbor.

2. Related Works

Nowell et al. [1] introduced a comprehensive list of both local and global similarity-based link prediction (LP) methods, which are widely used to estimate the likelihood of link formation between node pairs in a network. These similarity-based approaches work by computing a numerical score that quantifies the probability of a connection forming between two given nodes. Local similarity-based methods, in particular, rely on neighborhood structures and are designed to operate on node pairs that are precisely two hops apart. These methods leverage the principle that nodes sharing common neighbors are more likely to establish a direct link in the future.

Global similarity-based methods primarily rely on path-based metrics and can be applied to any pair of nodes within a network, regardless of their distance. These methods assess link likelihood by analyzing network topology beyond immediate neighborhoods, often considering multiple paths or structural patterns. In this paper, we incorporate most of the similarity-based methods proposed by Nowell et al. in our experiments, alongside an additional widely used local method, Resource Allocation [2], to provide a more comprehensive evaluation.

The state-of-the-art LP methods discussed above are primarily designed for simple, undirected, and homogeneous networks. However, several studies have explored the incorporation of various network attributes, such as node features, edge weights, and temporal information, to enhance link prediction accuracy in more complex and heterogeneous network structures [3].

The advanced LP methods outlined above are mainly designed for simple, undirected, and homogeneous networks. However, several studies have leveraged different network attributes, such as node characteristics, edge weights, and temporal dynamics, to improve link prediction in more complex and heterogeneous networks [4].

Few examples are: Temporal link prediction leverages the evolving patterns of link formation and maintenance over time in dynamic networks, such as social networks, where relationships continuously change. This approach considers factors like the frequency, recency, and persistence of past interactions to predict future connections. Similarly, link prediction in bipartite networks focuses on predicting links in networks composed of two distinct sets of nodes, such as user-product networks in recommendation systems or term-document networks in information retrieval. These methods account for the structural properties unique to bipartite graphs, ensuring accurate predictions tailored to their specific characteristics [5].

Temporal link prediction utilizes the evolving link dynamics in dynamic networks, such as social networks, by analyzing interaction patterns over time. Similarly, link prediction in bipartite networks focuses on connections between two distinct node sets, like user-product or term-document networks. These methods leverage structural properties and past interactions to improve predictive accuracy [6].

Kumar et al. provide a comprehensive survey on LP methods and assess their performance using standard metrics such as AUROC, AUPR, and Precision@K. Most existing studies on link prediction evaluation primarily address the challenge of class imbalance [7].

3. LINK PREDICTION METHODS

3.1 Similarity Based Methods

Similarity-based link prediction (LP) methods assign a score to a given pair of nodes, indicating the probability of link formation between them. These methods are categorized based on the geodesic distance between node pairs in the network into two types: local and global. Local methods are specifically applied to node pairs that are two hops apart, whereas global methods can be utilized for any node pair, regardless of their distance in the network. Since machine learning-based approaches also operate globally, we refer to local similarity-based methods as **local-sim** and global similarity-based methods as **global-sim** for clarity. Below, we outline some commonly used local and global similarity-based methods that are employed in this study.

- **Local methods**

Common Neighbors (CN) [1]: The Common Neighbors method (CN) measuresn the number of common neighbors or two-hop paths between node pairs. The CN score between nodes x and y can be expressed as:

$$CN(x,y)=|\Gamma(x)\cap\Gamma(y)|$$

Where Γ(x) and Γ(y) represents the set of neighbors of x and y respectively.

Jaccard's Coefficient (JC) [1, 49]: Jaccard's Coefficient (JC) extends the CN method by penalizing it for non-shared neighbors between the nodes. The JC score between nodes x and y can be expressed as:

$$JC(x,y)=|\Gamma(x)\cup\Gamma(y)||\Gamma(x)\cap\Gamma(y)|$$

Adamic Adar (AA) [1, 50]: Adamic Adar (AA) extends the CN method by penalizing each common neighbor by its degree logarithmically. The AA score between nodes x and y can be expressed as:

$$AA(x,y)=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\boldsymbol{log}\,|\,\Gamma(z)\,|}$$

Resource Allocation (RA) [17]: Unlike AA, Resource Allocation (RA) penalizes each common neighbor by its degree without logarithmic scaling. The RA score between nodes x and y can be expressed as:

$$RA(x,y)=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{|\,\Gamma(z)\,|}$$

Global methods

Preferential Attachment (PA) relies on the principle that nodes with higher degrees are more likely to acquire new connections. The PA score between nodes x and y can be expressed as:

$$PA(x,y)=|\Gamma(x)|\times|\Gamma(y)|$$

Katz Similarity (Katz): Katz similarity index (Katz) enumerates all the possible paths of different lengths between the node pairs, and takes sum over this collection, exponentially damping by path length. The Katz score between nodes x and y can be expressed as:

$$Katz(x,y)=\sum_{l=1}^{\infty}\beta^{l}\cdot|\,paths_{x,y}^{\langle l\rangle}\,|,$$

where β acts as decay factor to give exponentially higher weight to longer paths, and pathsx,y(l)| is the number of different paths of length l connecting the node pair.

*Hitting Time (HT) and Normalized HT (Norm-HT***)** Hitting Time (HT) leverages random walks on a network to quantify node similarity. The hitting time between nodes x and y is the expected number of steps it takes for a random walker starting at node x to reach node y for the first time. It quantifies how easily information or influence can spread between the nodes. It indicates easier information flow or shorter travel times between the nodes in the network, where smaller HT indicates better link formation likelihood. The scoring function can be expressed as:

$$HT(x,y)=-\sum_{t=1}^{\infty}t.P(T_{xy}=t)$$

Here, Txy is the random variable denoting the time it takes for a random walker to reach node y from x, and P(Txy = t) is the probability of this happening in t steps. HT (x, y) is quite small when the node y has a large stationary probability.

To mitigate this issue, the score is multiplied with y's stationary probability. We call this measure as normalized heating time (Norm-HT).

Commute Time (CT) and Normalized CT (Norm-CT) : Commute Time (CT) signifies the expected time for a random walker to travel from node x to y and back to x. It encapsulates the symmetric nature of node connectivity. The CT score between nodes x and y can be expressed as:

$$CT(x, y) = HT(x, y) + HT(y, x)$$

Like HT, we consider normalized commute time (Norm-CT) along with its un-normalized version.

3.2 Machine Learning Based Methods

We use two popular node embedding methods in this category, namely, Deepwalk [19], and Graphsage [21], towards LP. All of these methods learn a function $f : V \rightarrow R^d$, given a network graph G(V,E), where V and E are the set of vertices and edges respectively, which maps each node to a d dimensional latent space where d << |V |. We learn edge features following the method presented in Grover et al. [20] to get a link vector given a node pair, and apply logistic regression and random forest supervised learning technique to predict links. We refer this group of LP methods as learning. Below, we brief the aforementioned node embedding methods and the edge feature learning methods.

- **Deepwalk**

Deepwalk adapts Skip-gram [54] method of natural language processing to generate node embeddings. It solves the following optimization problem:

$$\underset{f}{\text{maximize}} \sum_{v_i \in V} \left[-\log Z_{v_i} + \sum_{n_i \in S^i} f(n_i) \cdot f(v_i) \right],$$

where Zvi = Pu∈V exp(f(u) ・ f(vi)), and Si are the set of nodes inside the context window of vi. To avoid the explosion of labels, precisely |V | numbers, Deepwalk uses Hierarchical Softmax [55, 56] with stochastic gradient descent (SGD) to approximate
the optimization.

- **GraphSAGE**

GraphSAGE is a graph neural network approach for scalable and inductive learning on large graphs. GraphSAGE leverages node attributes (e.g., node2vec embeddings) to learn embedding functions that generalize to unseen nodes during the training phase.
GraphSAGE does this by learning aggregator functions that can induce an embedding of a node by aggregating the attributes of its neighboring nodes, sampled from its direct connections. This aggregation process is executed k times for all nodes in the network, which way it learns k sets of weight matrices {Wk}. It uses four different aggregator techniques: mean, max-pooling, mean-pooling, and LSTM. GraphSAGE optimizes similar objective as Deepwalk and Node2vec, and approximates it with negative sampling, where given a node, its positive instances are sampled from nodes appearing in the chain of short random walks starting at the given node, and negatives are sampled from the degree distribution. It uses SGD as the optimization procedure.

DATASET
Facebook 1: It is a social network, built on the Facebook platform, focuses on a single user ("ego") and their connections with other users. Nodes represent Facebook users, and links represent friendships between them. We refer the dataset prepared from this network as fb.
The future LP datasets were prepared for the networks where the interaction available. The interactions can be directed, and there could be multiple interactions between a node pair.

4. RESULTS AND DISCUSSION

We applied the 10 LP methods presented in Section 3 on the FB datasets. We evaluated their performance based on the AUROC evaluation metric. Here we aim to understand whether prediction performance of various LP methods vary when the distance between the two nodes in test node pairs differ. For each dataset, we performed paired t-test to determine if the prediction performance differ in terms of their absolute values for the two cases: two-hop away test node pairs
Following table represents fb dataset with different methods.

Table 1 AUROC results

Dataset	Method type	P- Value
Missing two-hop	Local sim	0.199
	Global sim	0.577
	Learning	0.384
Missing three-hop	Local sim	0.216
	Global sim	0.505
	Learning	0.456

The table 1 presents P-values for different link prediction methods applied to datasets with missing links at two-hop and three-hop distances. It compares three method types: Local Similarity, Global Similarity, and Learning-based approaches. In both cases, Global Similarity shows the highest P-values (0.577 for two-hop and 0.505 for three-hop), indicating weaker statistical significance in distinguishing predicted links. Local Similarity has the lowest P-values (0.199 and 0.216), suggesting relatively better differentiation. The Learning-based approach falls in between, with moderate P-values (0.384 and 0.456). Overall, the results imply that Local Similarity methods may be more effective, while Global Similarity methods may struggle to provide significant differentiation.

5. CONCLUSIONS

This paper introduces an experimental framework designed to provide a structured and controlled environment for evaluating link prediction (LP) methods. The framework considers key factors such as prediction type, network

characteristics, method category, distance between end nodes, and evaluation metrics to ensure a rigorous assessment. Extensive experiments were conducted on real-world network datasets using existing LP techniques, followed by hypothesis-driven analyses to understand the influence of these factors on LP performance. Based on the insights gained, we offer a set of best practices for systematically evaluating LP methods.

ACKNOWLEDGEMENT

We express our sincere gratitude to Global Academy of Technology, Bangalore and Navkis College of Engineering, Hassan for their support and resources that contributed to this paper.
we would like to thank Visvesvaraya Technological University (VTU), GnanaSangama, Belagavi - 590018, Karnataka, India for the monetary support extended our work.

REFERENCES

1. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121

2. Zhou, T., L¨u, L., Zhang, Y.-C.: Predicting missing links via local information. The European Physical Journal B 71, 623–630 (2009)

3. Sett, N., Basu, S., Nandi, S., Singh, S.R.: Temporal link prediction in multirelational network. World Wide Web 21, 395–419 (2018)

4. Qin, M., Yeung, D.-Y.: Temporal link prediction: A unified framework, taxonomy, and review. ACM Computing Surveys 56(4), 1–40 (2023)

5. Kunegis, J., De Luca, E.W., Albayrak, S.: The link prediction problem in bipartite networks. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, pp. 380–389 (2010). Springer.

6. Ozer, S¸.D.I., Orman, G.K., Labatut, V.: Link prediction in bipartite networks. In: 28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES) (2024).

7. Kumar, A., Singh, S.S., Singh, K., Biswas, B.: Link prediction techniques, applications, and performance: A survey. Physica A: Statistical Mechanics and its Applications 553, 124289 (2020).