Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.37054

Link Prediction in Networked Data Using Structural Features and Machine Learning

Kaihua Li

College of Arts and Sciences, University of Oregon, Eugene, USA
kaihualimm@gmail.com

Abstract. Link prediction is a problem in network analysis that seeks to find missing or future links in a graph based on observed graph structures and available node information. This paper reviews the development of link prediction methods from traditional structural heuristics to network embeddings and graph neural networks. Common Neighbors and other similarity indices are traditional methods that are simple to use and interpret but rely heavily on manually designed structural features. Representation-learning methods, such as node2vec, reduce the requirement for feature engineering by learning low-dimensional node embeddings. More recently, graph neural network (GNN)-based methods have been developed to address issues such as neighbourhood overlap, multi-hop paths, graph incompleteness and long-tailed structural patterns in link prediction, including Neighborhood Overlap-aware Graph Neural Networks (Neo-GNNs), Neural Bellman-Ford Network (NBFNet), Common Neighbor Completion with Information Entropy (IECNC) and Long-Tailed Link Prediction (LTLP). Recently, temporal link prediction has also begun to attract some attention, and EAGLE has been developed to combine both recent temporal information and global structural patterns effectively. In short, the reviewed studies indicate that structural information remains important for link prediction in contemporary learning-based models. Future work will continue to improve the scalability, robustness, temporal adaptation and interpretability of link prediction models for large-scale, dynamic real-world networks.

Keywords: link prediction, networked data, structural features, graph neural networks, temporal link prediction

1. Introduction

Network data are frequently used to represent the connections among all kinds of objects in many real-life situations, such as social media, biological systems, citation networks and knowledge graphs. The components of a network are nodes and edges. Nodes are the entities, and edges are the connections among these entities. With the continuous increase in the amount of network data, a problem of finding missing links or potential connections in a graph has emerged in graph mining and network analysis. Link prediction estimates whether a future or missing connection is likely to exist based on the current network structure and available information. Link prediction is frequently applied in social recommendation, biological interaction prediction, fraud detection and knowledge

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.37054

graph completion. Recently, graph representation learning and graph neural networks have made significant improvements in link prediction performance and are now active research areas in graph machine learning. Classical structural heuristics estimate link likelihood based on local topology, while network embeddings learn compact representations and graph neural networks extract structural information through trainable architectures. Different families of methods have different uses for handcrafted features, different abilities to capture local or long-range structure, different computational costs, and different applications in static or dynamic graphs. The construction of negative samples and the separation of training, validation, and test links in evaluation also affect the results; different protocols can yield different performance estimates [1, 2]. Therefore, this paper will review how structural information is represented in classical and learning-based link prediction methods, how recent models address graph incompleteness, long-tail structural patterns and temporal evolution, and what limitations remain in scalability, generalisation and interpretability.

2. Basic concepts of networked data and link prediction

Structured datasets can illustrate the associations between different items within a practical environment more clearly. Such systems are also referred to as graphs, and they are denoted as G= (V, E), where V represents the collection of vertices and E denotes the collection of links. Vertices are objects like individuals, documents, molecules and goods, while edges are the interactions or bonds between them. Graphs might be organized as directed or undirected, labelled or unlabelled, homogeneous or heterogeneous, etc., depending on the specific use case. A standard representation of a graph is an adjacency matrix, and every entry reflects whether a link exists between the two specific vertices. Such graph structures serve as the basis for graph analytics and graph-based machine learning. Link prediction aims to determine whether a nonexistent or potential edge is probable to occur between two vertices given the architecture of the observed network and accessible vertex data. This endeavour is typically structured as a binary classification issue in real-world applications. Existing edges are treated as positive instances, while unobserved vertex pairs are selected as negative samples; nonetheless, some of these pairs may correspond to missing or future positive links. Three typical partitions of data for model generalisation evaluation are training subsets, validation subsets and test subsets. The area under the receiver operating characteristic curve (AUC), average precision (AP), Precision@K, Recall@K and Hits@K are frequently utilized assessment metrics that evaluate different aspects of the predictive performance of various algorithms.Link prediction may be examined within two distinct learning frameworks: transductive learning and inductive learning. Under the transductive paradigm, every node remains accessible for training, while only a portion of links is concealed during the prediction phase. Conversely, inductive learning aims to generalize the task to unseen nodes or entirely novel graphs, making it more appropriate for practical situations involving evolving networks. Lately, numerous researchers have started focusing on inductive link prediction due to the fact that many actual-world networks are dynamic and undergo temporal transformations. The efficacy of link prediction largely hinges upon the inherent topology of the network. Entities possessing numerous mutual neighbors, brief path lengths, or analogous local patterns are typically more probable to establish future associations. These aforementioned insights have motivated various structural heuristics, including Common Neighbors, Jaccard's Coefficient, Adamic-Adar, and Preferential Attachment. Lately, machine learning approaches, graph embedding methodologies and graph neural networks have been devised to autonomously extract meaningful embeddings from graph architectures, and their forecasting precision on vast and intricate networks has advanced considerably. Subsequent sections shall elaborate on these aforementioned techniques [1, 2].

DOI: 10.54254/2753-8818/2026.37054

3. Structural features and machine learning methods for link prediction

The first type of link prediction method uses the structure of a network. These ways assume that nodes with similar structures are more likely to be connected in the future. One of the first and most popular ways is Common Neighbors (CN), which predicts a link based on the number of shared neighbours of the two nodes. Other classic similarity measures, such as the Jaccard Coefficient, Adamic-Adar (AA) index, Resource Allocation (RA) index and Preferential Attachment (PA) score, have also been proposed later to improve the prediction accuracy under different network structures. Heuristic methods are relatively simple to compute and implement, and often perform well for small or sparse networks. However, because they are based on manually designed structural features, their ability to learn complex graph patterns in large-scale and heterogeneous networks is relatively weak [1, 3].

Machine learning methods have been used to address the problems of heuristic methods in link prediction. Instead of using a single similarity measure, these methods combine multiple structural features as input and train classifiers such as logistic regression, random forest, or gradient boosting to determine whether a link exists between two nodes. Most machine learning methods are more flexible and can consider various topological features and node attributes simultaneously. Nevertheless, they still require manual feature engineering, and their prediction performance heavily depends on the quality of the extracted features. As the size and complexity of graph datasets continue to increase, it is becoming more difficult to design good features manually [1, 2].

Network embedding techniques have been introduced to learn low-dimensional vector representations of nodes automatically. These methods do not manually construct features and maintain the graph structure by mapping nodes to a continuous embedding space. Node2Vec is one such well-known network embedding algorithm. Biased random walks are used to explore different neighbourhoods, and node representations that contain both local and global structural information are learned. The learned embeddings can be employed in the following tasks later, such as node classification and link prediction. Embedding-based methods have improved the representation ability and reduced the demand for handcrafted features compared with the old heuristic method. However, they generally treat graph structures as static and fail to fully use the rich relational information [1].

Recently, some advancements in Graph Neural Networks (GNNs) have shown improved link-prediction capabilities under various graph structures, such as Neighborhood Overlap-aware Graph Neural Networks (Neo-GNNs), Neural Bellman-Ford Network (NBFNet), Common Neighbor Completion with Information Entropy (IECNC), and Long-Tailed Link Prediction (LTLP). GNN-based models iteratively collect information from neighbouring nodes to learn expressive node representations directly from the graph structure. Many models have been proposed to address various problems. Neo-GNNs are graph neural networks that add neighbourhood overlap information to better use the structural patterns that are often ignored by traditional GNNs. NBFNet is a path-based framework that extends the Bellman-Ford algorithm to gather evidence from multiple paths for link prediction. Recently, IECNC has introduced a message-passing neural network and an information-entropy-based common-neighbor completion mechanism to address graph incompleteness. Recently, LTLP has been introduced to address the problem of long-tail link prediction and improve the accuracy of prediction for node pairs with few common neighbours. The above development indicates that, in recent years, GNN-based methods have led the way in link prediction research due to their ability to jointly learn structural information, node attributes and complex graph patterns [2, 4-7].

DOI: 10.54254/2753-8818/2026.37054

4. State-of-the-art studies of link prediction in networked data

Recently, with the development of graph representation learning and graph neural networks, link prediction research has been progressing rapidly. The new ones are better at handling difficult cases than the previous heuristic and embedding-based methods, such as sparse graphs, heterogeneous networks, dynamic graphs, and long-tail structural patterns. Several models have been proposed by researchers to improve the accuracy of prediction and enhance scalability and generalisation [1, 2].

A typical type is Neo-GNNs, which add neighbourhood overlap information to graph neural networks. Traditional GNN models generally use message passing to aggregate information from neighbouring nodes and may fail to consider structural similarities between node pairs. Neo-GNNs integrate neighbourhood overlap with graph representation learning to address the limitations of GNNs in capturing local structure more effectively, thereby improving their expressive power. According to the experiments reported in the original paper, Neo-GNNs have outperformed several baseline models on the benchmark datasets [4].

Another kind of development is Neural Bellman-Ford Networks (NBFNet), which is based on the path-based reasoning method of the original Bellman-Ford shortest-path algorithm. NBFNet considers all paths connecting the two target nodes and learns from them, rather than just neighbouring nodes. The design can address long-range dependencies through multiple paths. Therefore, NBFNet has performed well on multiple standard benchmark datasets [5].

Recently, some studies have also addressed specific problems in the application stage. Common Neighbor Completion with Information Entropy (IECNC) extends common-neighbor-based link prediction by adding a message-passing neural network and an information-entropy-based completion mechanism for uncertain links. The above method is to reduce the impact of graph incompleteness on common-neighbor information. At the same time, LTLP addresses a long-tail structural problem in which node pairs with fewer common neighbours generally have lower prediction accuracy. LTLP improves the structural representation of tail-node pairs to enhance prediction performance for these underrepresented pairs. These studies show that the current research trend is to improve the robustness and performance of models in real-world network environments [6, 7].

Temporal and dynamic link prediction have also been the subjects of research in addition to static graphs. Real-world networks, such as social media networks, communication systems and financial transaction networks, are constantly changing over time and need to be modelled with both time-dependent and structure-preserving characteristics. EAGLE is a relatively new temporal link prediction model that is both accurate and computationally light. EAGLE can model the changing structure of graphs well and is relatively efficient for dynamic network data. With the development of temporal graphs, future research in link prediction will probably focus more on adaptability, scalability and computational efficiency [2, 8, 9].

5. Limitations and prospects

Link prediction has achieved some results in the past ten years, but many problems still exist. Traditional heuristic methods are simple and computationally inexpensive, but they only use local structural information. As a result, they often have poor performance in large-scale or heterogeneous networks with more complex node relationships. Machine learning can increase the accuracy of the prediction by incorporating many features; however, it generally requires manual feature engineering and may not be suitable for all kinds of networks [1, 2].

DOI: 10.54254/2753-8818/2026.37054

Recently, graph embedding methods and graph neural networks have improved the performance of link prediction significantly. However, these models also have some deficiencies. Many GNN-based methods need large amounts of training data and computational resources, making them less suitable for very large graphs. Some models are also relatively simple; they assume that the network is static and does not change over time. Social networks and communication systems are continuously adding new nodes and edges in practice; therefore, models trained on static graphs may not perform well in dynamic environments [1, 2, 8, 9].

Another problem is the long-tail structure of link prediction. Although node degrees are often skewed in real-world networks, recent studies have shown that node pairs with fewer common neighbours are even harder to predict. Recently, some studies have started to address this problem, but there are still many deficiencies. Most of the current models are not very transparent. Although these models have achieved high predictive accuracy, it is often unknown why a particular link is predicted; therefore, they are not suitable for high-stakes applications that require transparent decision-making [7, 10].

Looking ahead, the following will be the direction of future research. First, building high-efficiency models that can handle large and continuously changing graphs will be required to some extent. Second, models should integrate structural information, node attributes, temporal information and data from multiple graphs to improve prediction accuracy. Recently, people have begun to be interested in interpreting what kind of problems these models are used to solve. With the rapid development of large language models and foundation models, graph learning may be combined with strong semantic representations in the future for link prediction research. Although many problems still exist, link prediction is expected to be used in recommendation systems, biological network analysis, knowledge graphs and other places in the future [1, 2, 8-10].

6. Conclusion

This paper reviews link prediction in networked data from the perspectives of structural heuristics, machine-learning classifiers, network embeddings and graph neural networks. The original indices are still practical because they are simple to compute and interpret, such as common neighbors, the Jaccard index, Adamic-Adar index, resource allocation and preferential attachment. However, their performance is very sensitive to the local topology and may be relatively low in sparse, heterogeneous, dynamic graphs or graphs with long-tail structural patterns. Machine-learning classifiers extend these heuristics by combining several structural and attribute-based features, but their effectiveness depends on careful feature engineering and negative-sample construction. Network embedding methods reduce manual feature design by learning low-dimensional node representations, and GNN-based methods model neighbourhood, paths and common-neighbor information more directly. Recently, some methods to deal with structural overlap, long-range reasoning, graph incompleteness, long-tailed structural patterns and temporal evolution have been proposed, such as Neo-GNNs, NBFNet, IECNC, LTLP and EAGLE. Not all of these ways work for every network. A reliable comparison should use consistent training, validation and test splits; explicitly perform negative-sampling; employ multiple ranking metrics; report computational cost; and evaluate on tail-node pairs or previously unseen nodes. Future work will enhance the scalability, temporal adaptability, robustness and interpretability of the model by incorporating structural information and node attributes/semantic representations. Link prediction is still a general graph learning problem, and whether the model, data representation and evaluation methods are suitable for the structure and development of the target network will determine its actual application performance. Practically, the studies should also report how the candidate links are sampled and how

DOI: 10.54254/2753-8818/2026.37054

the performance changes under sparse, imbalanced or temporally shifted graph conditions to support reproducible comparisons across datasets.

References

[1] Kapoor, P., Kaushal, S., Kumar, H., & Kanwar, K. (2024). A survey on feature extraction and learning techniques for link prediction in homogeneous and heterogeneous complex networks. *Artificial Intelligence Review, 57*, Article 348. https://doi.org/10.1007/s10462-024-10998-7

[2] Sun, C., Song, Y., Zhai, C., Tian, J., Yang, J., Rui, X., Zhang, J., Wang, Z., & Yu, P. S. (2026). A survey on GNN-based link prediction: Techniques, applications, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 16*(2), e70093. https://doi.org/10.1002/widm.70093

[3] Ahmad, I., Akhtar, M. U., Noor, S., & Shahnaz, A. (2020). Missing link prediction using common neighbor and centrality based parameterized algorithm. *Scientific Reports, 10*, Article 364. https://doi.org/10.1038/s41598-019-57304-y

[4] Yun, S., Kim, S., Lee, J., Kang, J., & Kim, H. J. (2021). Neo-GNNs: Neighborhood overlap-aware graph neural networks for link prediction. In *Advances in Neural Information Processing Systems (Vol. 34)*, pp. 13683-13694. Curran Associates, Inc.

[5] Zhu, Z., Zhang, Z., Xhonneux, L.-P., & Tang, J. (2021). Neural Bellman-Ford networks: A general graph neural network framework for link prediction. In *Advances in Neural Information Processing Systems (Vol. 34)*, pp. 29476-29490. Curran Associates, Inc.

[6] Zhou, Z., Wan, G., & Du, B. (2025). Common neighbor completion with information entropy for link prediction in social networks. *Data Science and Engineering, 10*, 40-53. https://doi.org/10.1007/s41019-024-00267-6

[7] Wang, Y., Wang, D., Liu, H., Hu, B., Yan, Y., Zhang, Q., & Zhang, Z. (2024). Optimizing long-tailed link prediction in graph neural networks through structure representation enhancement. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3222-3232). ACM. https://doi.org/10.1145/3637528.3671864

[8] Li, H., Xu, Y., Li, Y., Liu, H., Li, D., Zhang, C. J., Chen, L., & Li, Q. (2025). When speed meets accuracy: An efficient and effective graph model for temporal link prediction. *Proceedings of the VLDB Endowment, 18*(10), 3396-3405. https://doi.org/10.14778/3748191.3748203

[9] Mei, P., & Zhao, Y. H. (2024). Dynamic network link prediction with node representation learning from graph convolutional networks. *Scientific Reports, 14*, Article 538. https://doi.org/10.1038/s41598-023-50977-6

[10] Kaczyńska, P., Sienkiewicz, J., & Ślęzak, D. (2026). Accumulated local effects and graph neural networks for link prediction. *Scientific Reports, 16*, Article 8574. https://doi.org/10.1038/s41598-026-39000-w