

Generative Graph Neural Networks for Link Prediction

Xingping Xian^a, Tao Wu^{a,*}, Xiaoke Ma^b, Shaojie Qiao^c, Yabin Shao^d, Chao Wang^e, Lin Yuan^a, Yu Wu^a

^aSchool of Cybersecurity and Information Law, Chongqing University of Posts and Telecommunications, Chongqing, China.

^bSchool of Computer Science and Technology, XiDian University, XiAn, China.

^cSchool of Software Engineering, Chengdu University of Information Technology, Chengdu, China.

^dSchool of Science, Chongqing University of Posts and Telecommunications, Chongqing, China.

^eSchool of Computer and Information Science, Chongqing Normal University, Chongqing, China.

Abstract

Inferring missing links or detecting spurious ones based on observed graphs, known as link prediction, is a long-standing challenge in graph data analysis. With the recent advances in deep learning, graph neural networks have been used for link prediction and have achieved state-of-the-art performance. Nevertheless, existing methods developed for this purpose are typically discriminative, computing features of local subgraphs around two neighboring nodes and predicting potential links between them from the perspective of subgraph classification. In this formalism, the selection of enclosing subgraphs and heuristic structural features for subgraph classification significantly affects the performance of the methods. To overcome this limitation, this paper proposes a novel and radically different link prediction algorithm based on the network reconstruction theory, called GraphLP. Instead of sampling positive and negative links and heuristically computing the features of their enclosing subgraphs, GraphLP utilizes the feature learning ability of deep-learning models to automatically extract the structural patterns of graphs for link prediction under the assumption that real-world graphs are not locally isolated. Moreover, GraphLP explores high-order connectivity patterns to utilize the hierarchical organizational structures of graphs for link prediction. Our experimental results on all common benchmark datasets from different applications demonstrate that the proposed method consistently outperforms other state-of-the-art methods. Unlike the discriminative neural network models used for link prediction, GraphLP is generative, which provides a new paradigm for neural-network-based link prediction. The code is available at <https://github.com/star4455/GraphLP>.

Keywords: Graph Machine Learning, Graph Neural Networks, Link Prediction, Structural Patterns, Network Reconstruction.

1. Introduction

Graphs provide an elegant representation for characterizing entities and their interrelations in complex systems. Given that real-world graphs can usually only be partially observed and are often noisy, link prediction aimed at inferring missing and spurious links based on observed graphs is a paradigmatic and fundamental problem across many scientific domains, including knowledge graph completion [52], experimental design in biological networks [6], fake account detection in online social networks [24], and product recommendation on e-commerce websites [31].

To address the link prediction problem, numerous heuristic methods have been proposed, including local indices such as Common Neighbors (CN) [2], and Resource Allocation (RA) [61], global indices such as Katz [25], and SimRank [23], and quasi-local indices such as the Local Path Index (LP) [61]. However, heuristic methods have a strong assumption on when two nodes are likely to be linked in real-world graphs and lack universal applicability to diverse areas [8]. Subsequently, statistical learning-based algorithms have been proposed to obtain groundbreaking results, such as maximum likelihood-based hierarchical structure model [13], stochastic block model [21], matrix factorization-based link prediction method [37], Linear Optimization (LO) link prediction method [38], and Low Frobenius norm-based Link Prediction (LFLP) method [53]. With the proposal of network representation learning, various network em-

bedding algorithms have been put forth so that the likelihood of a non-observed links can be estimated based on the proximity of nodes in low-dimensional vector space, including LINE [44], Node2Vec [20], and DNGR [10].

Recently, driven by the dramatic advances in deep learning techniques, neural networks have gradually been used to solve the link prediction problem. [56] trained a fully-connected neural network on the enclosing subgraphs of target links for link prediction, wherein a Weisfeiler-Lehman (WL) algorithm-based graph labeling mechanism was proposed to encode subgraphs. Based on the enclosing subgraphs extracted around links, [57] trained a Graph Neural Network (GNN) for link prediction to achieve a performance comparable to that of heuristic methods. Along this line of research, [36] encoded subgraphs into random-walk transition probabilities and then computed features using these probabilities to classify positive and negative links. Although these subgraph classification-based methods have achieved state-of-the-art link prediction performance, the prediction results are found to be considerably affected by the extraction process of the k -hop enclosing subgraphs and the graph structure features for them. For example, in representation learning on graphs [55], the range of enclosing subgraphs strongly depends on the graph structure, and the effective range should be different for subgraphs with varying properties.

Typically, from the perspective of subgraph classification, link prediction methods treat subgraphs in real-world graphs independently and equivalently. That is, the global structural information of real-world graphs is totally neglected during this process. However, extensive empirical analyses indicate that real-world graphs are not locally isolated but globally relevant

*Corresponding author.

Email addresses: xxp0213@gmail.com (Xingping Xian), wutaoadeny@gmail.com (Tao Wu)

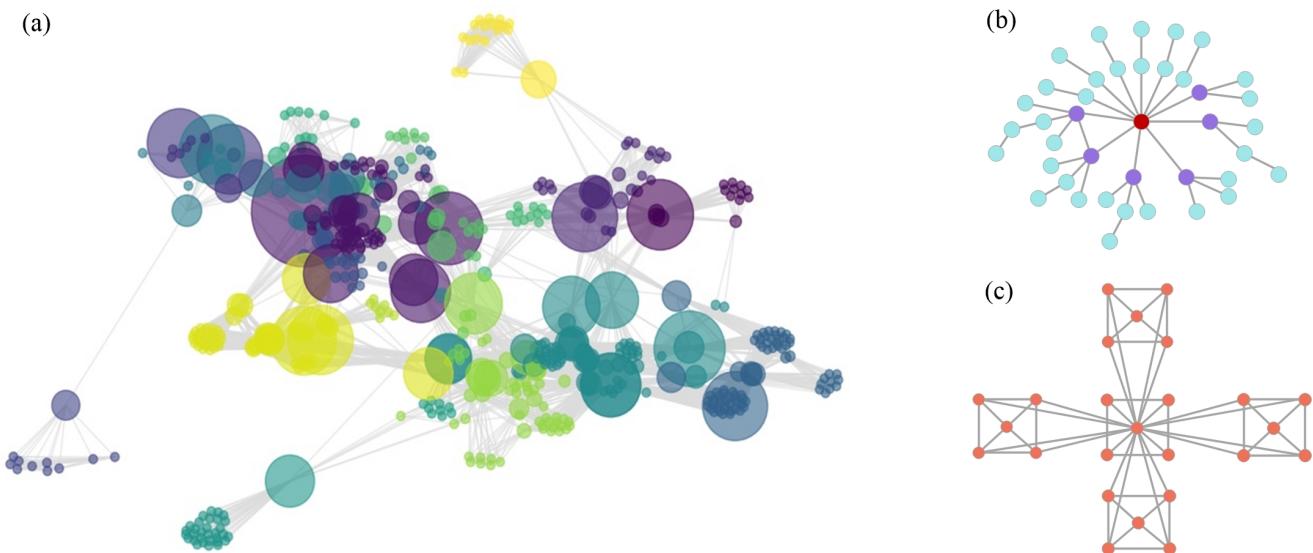


Figure 1: An illustrative example depicting the global and high-order organizations of real-world graphs. (a) Gene network for *C. elegans* [12]. (b) Representative hierarchical star-like structure [45]. (c) Representative hierarchical modular organization [39]. (b) and (c) depict the representative structural patterns of real-world graphs such as (a).

[35, 51]; here, nodes and edges naturally portray different structural roles and contribute differently to the global organization of real-world graphs [53, 54]. Moreover, subgraph classification-based link prediction methods assume that real-world graphs exhibit low-order connectivity patterns and can be captured at the level of individual nodes and edges. However, empirical studies have discovered that real-world graphs exhibit high-order organizations at the level of small subgraphs, which are recursively grouped into a hierarchical structure [7, 3]. An illustrative example of the global and high-order organization in real-world graphs is depicted in Figure 1. Hence, two challenges need to be addressed for link prediction: (i) how to learn good representation preserving both local and global graph structural features? and (ii) how to characterize and utilize hierarchical structure patterns?

To address these challenges, instead of predicting potential links through subgraph classification, this study designs a novel generative and multi-order GNN for link prediction, called GraphLP. Evidently, real-world graphs share some global properties, such as low-rank and sparsity, that can be used to provide guidance for graph learning. Hence, motivated by the network reconstruction theory [21], GraphLP defines a self-representation model-based collaborative inference operation to refine the observed graphs globally, which assumes that the original graph can be reconstructed utilizing the correlation between subgraph patterns. Assuming that the paths between a pair of nodes provide evidence for the existence of potential links, GraphLP extracts the local structural information via a high-order connectivity operation on the observed graphs. Thus, every neural network layer obtains the connectivity of node pairs within two-hop neighborhood, and a neural network with multiple connectivity layers captures the degree of connectivity between node pairs with various path lengths. Meanwhile, the weighted adjacency matrices generated by the connectivity operation in every neural network layer reflect the multi-order connectivity pattern in the graphs. Further, the hierarchical organizational structure of real-world graphs is explored by applying a collaborative inference operation. The contributions of this study can be summarized as follows:

- **Generative framework.** Rather than subgraph classification based discriminative schemes, a novel network reconstruction-based generative GNN is proposed for link prediction, which provides a new paradigm for the application of neural networks in link prediction problem.
- **End-to-end learning.** Instead of designing heuristic graph structural features for subgraph representation, local and global structural patterns are extracted and fused in an end-to-end fashion for link prediction.
- **Algorithm.** A novel collaborative inference operation and high-order connectivity computation mechanism are developed to characterize the structural patterns in real-world graphs at different scales.
- **Experiment.** Extensive experiments on real-world datasets from different areas reveal that the proposed method, GraphLP, achieves promising performance and consistently outperforms other state-of-the-art methods.

Paper Organization. The rest of this work is organized as follows. Section 2 discusses related studies. Section 3 presents the problem definitions and describes the preliminaries. Section 4 describes the proposed method. Section 5 presents the experimental results, and finally, Section 6 presents the conclusion and discussion.

2. Related Work

GNNs and link prediction task have been extensively investigated in recent years. A brief review of related studies is provided in this section.

2.1. Graph Neural Networks

Owing to their potential in modeling the complex structures of non-Euclidean graphs, GNNs have achieved state-of-the-art performance on almost all graph-based tasks, such as node classification, graph classification, link prediction. Based on different theories and perspectives, a plethora of different GNNs have

been proposed over the years. Generally, GNNs can be divided into two categories: spectral-based and spatial-based methods. Of these, spectral-based GNNs are types of GNNs that design graph convolution operators in the spectral domain using Fourier transform. The involved convolution operation is defined as follows:

$$f_1 * f_2 = \mathbf{U}[(\mathbf{U}^T f_1) \odot (\mathbf{U}^T f_2)], \quad (1)$$

where \odot denotes an element-wise product. The spectral filter is defined as $\mathbf{g} = \mathbf{U}^T f_1$, and the node signal \mathbf{X} can be processed as follows:

$$\mathbf{Z} = \mathbf{U}[\mathbf{g}(\Lambda) \odot (\mathbf{U}^T \mathbf{X})] = \mathbf{U}\mathbf{g}(\Lambda)\mathbf{U}^T \mathbf{X}. \quad (2)$$

where \mathbf{U} denotes a matrix of eigenvectors of the normalized Laplacian graph $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} = \mathbf{U}\Lambda\mathbf{U}^T$ [11]. Assuming that feature representation of node should be affected only by its k -hop neighborhood, [16] proposed a Chebyshev polynomial based k -localized convolution and developed a convolutional neural network, ChebNet, which eliminated the need to compute the eigenvectors of the Laplacian. Subsequently, [50] simplified the Chebyshev polynomial filter using its first-order approximation and proposed the popular spectral-based method called Graph Convolutional Networks (GCNs). Notably, spatial-based GNNs define graph convolution operator based on graph topology wherein the feature vectors of node's neighbors are aggregated via a permutation-invariant function. Specifically, [22] proposed a GraphSAGE approach that sampled fixed size neighborhood nodes and used max pooling, mean pooling, and LSTM pooling scheme to aggregate neighbor information. Considering the different weights of node's neighbors, [46] proposed a Graph Attention Network (GAT) algorithm to calculate attention coefficient and then aggregated the neighborhood information. Other related models include PATCHY-SAN [34], DCNN [4], and further details on GNNs can be found in the review [60].

2.2. Neural Networks based Link Prediction

Following heuristic methods, matrix completion-based methods and network embedding-based methods, neural networks have been gradually applied to link prediction problem and have achieved state-of-the-art results. Specifically, [56] proposed a link prediction method called Weisfeiler-Lehman Neural Machine (WLNM), which labeled nodes using the Weisfeiler-Lehman algorithm and encoded subgraphs to construct a feed-forward neural network-based classification model. Next, from the perspective of subgraph classification, [57] proposed a novel GNN-based link prediction framework, SEAL, to learn subgraph structures and node features from local enclosing subgraphs. Along this line, to directly leverage the topology features of local subgraphs, [36] proposed a new random-walk-based pooling scheme, WalkPool, and built features for subgraph classification. Moreover, [18] proposed a neural network-based link prediction method with only one-hop neighborhood information, which demonstrated almost equivalent performance to the WLNM and SEAL. Instead of subgraph classification, [8] converted the original graph into a corresponding line graph and solved the node classification problem for link prediction. To perform link prediction for general directed or undirected complex networks, [48] represented the adjacency matrices of networks as binary images and developed a generative adversarial networks (GANs)-based method. In addition, because existing GNN-based methods do not scale appropriately to large graphs, [30] extracted sparse enclosing subgraphs based on multiple random walks and presented a scalable link prediction solution,

called ScaLed. To reduce the time required to determine the distances between two nodes, [27] defined an anchor-based distance and proposed a new distance-enhanced GNN method for link prediction.

Among all existing methods for link prediction, the work closest to the one considered in this study is the GANs-based method [48]. However, this method predicts potential links via image processing within the GANs framework, whereas the proposed method conducts link prediction via GNNs-based network reconstruction.

2.3. Network Structure Analysis

Real-world graphs, also known as complex networks, are abstract representation of complex systems and have been extensively studied in the field of network science. Consequently, numerous studies have revealed that complex networks exhibit rich and diverse connectivity patterns. [32] augmented that the organization of real networks usually embodies both regularities and irregularities, where the former can be modeled and decides the extent to which the formation of a network can be explained. Notably, link predictability reflects the structural regularities in real-world networks and denotes the inherent difficulty of link prediction. [53] proposed a self-representation network model-based method, called NetSRE, for measuring and regulating link predictability of networks. [54] proposed a deep linear coding-based link prediction adversarial attack method by disturbing the underlying structural pattern of networks, which proved that links play global structural roles in network organization. Moreover, [7] suggested that high-order connectivity patterns are essential for understanding the fundamental structures of networks and developed a framework that identified clusters of network motifs. [41] claimed that hierarchical structure plays an important role in complex systems. To prove the existence of hierarchical organization, an unsupervised method for extracting the hierarchical organization of complex networks was introduced and validated.

Although real-world graphs exhibit various structural patterns, most existing neural networks-based link prediction methods simply assume that they are flattened and locally isolated, and these methods judge the existence of links explicitly based only on local enclosing subgraphs. With the exception of local structural features, this study focuses on integrating global and hierarchical structural patterns into neural networks for link prediction.

3. Problem Definition and Preliminaries

3.1. Problem Definition

Notations. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote an undirected and unweighted graph, where $\mathcal{V} = \{v_1, \dots, v_N\}$ denotes the set of nodes and $\mathcal{E} = \{e_1, \dots, e_M\}$ denotes the set of edges. The adjacency matrix of graph \mathcal{G} is denoted as $\mathbf{A} \in \{0, 1\}^{N \times N}$, where $\mathbf{A}_{ij} = 1$ if nodes i and j are connected and $\mathbf{A}_{ij} = 0$ otherwise. Each edge e can be represented as a node pair (u, v) , where $u, v \in \mathcal{V}$. Let $\mathcal{N}(u)$ denote the neighbors of node u , $\mathcal{N}(u) = \{v | (u, v) \in \mathcal{E}\}$.

Link Prediction. Given an observed graph $\mathcal{G}_o = (\mathcal{V}, \mathcal{E}_o)$ that corresponds to the original graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, link prediction aims to infer the presence or absence of an edge between a pair of target nodes based on \mathcal{G}_o , thereby generating a recovered graph \mathcal{G}^* to approximate the original graph \mathcal{G} . In particular, the prediction problem involves identifying a function that generates

a likelihood score for a pair of nodes $(u, v) \notin \mathcal{E}$ to infer the missing link (u, v) , or to produce a likelihood score for an existing edge $(u, v) \in \mathcal{E}$ to identify spurious links. Thus, the link prediction problem can be formulated as $s_{uv} = f(u, v, \mathbf{A}|\theta)$, where θ denotes the parameter of link prediction model. In this work, \mathcal{E}_m and \mathcal{E}_s denote the identified missing and spurious links, respectively.

Note that data augmentation is a set of techniques that increases the amount and diversity of data by creating reasonable virtual data points from existing data, such that better machine learning models can be constructed based on them. According to [59], this study considers graph data augmentation and adopts a random mapping mechanism to produce augmented graph set \mathcal{D} based on the observed graph $\mathcal{G}_o = (\mathcal{V}, \mathcal{E}_o)$. Specifically, the set of all possible edges in the graph \mathcal{G}_o is denoted as Ω , the existing edge set is denoted as \mathcal{E}_o , and the non-existing edge set is denoted as $\mathcal{E}_{non} = \Omega - \mathcal{E}_o$. Thus, the candidate sets for random mapping are defined as follows:

$$\mathcal{E}_{del}^c = \mathcal{E}_o, \quad \mathcal{E}_{add}^c = \mathcal{E}_{non}. \quad (3)$$

Thereafter, samples are randomly produced from the candidate sets to obtain the edge sets \mathcal{E}_{del} and \mathcal{E}_{add} . Finally, a new augmented graph is generated by modifying the graph \mathcal{G}_o based on \mathcal{E}_{del} and \mathcal{E}_{add} :

$$\mathcal{G}' = (\mathcal{V}, (\mathcal{E} \cup \mathcal{E}_{add}) \setminus \mathcal{E}_{del}). \quad (4)$$

Each input graph can be viewed as an instance for link prediction, owing to the generative learning scheme of the models considered in this work. Thus, the dataset containing a series of augmented graphs can be denoted as $\mathcal{D} = \{\mathcal{G}_i | i = 1, \dots, t\}$ and split to yield disjoint training and validation sets. These can be denoted as \mathcal{D}_{train} and \mathcal{D}_{val} respectively, wherein the missing and spurious links of the validation set are guaranteed not to appear in the training set. The observed graph \mathcal{G}_o used to generate the augmented graphs is defined as test set \mathcal{D}_{test} .

3.2. Graph Convolutional Networks

GCNs are a class of neural networks designed to generalize traditional convolution operator for non-euclidean graph-structured data. In essence, GCNs aim to learn new feature representations of nodes in graphs by exploiting their structural information. Let adjacency matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$ denote the structural information of the graph \mathcal{G} , and $\mathbf{X} \in \mathbb{R}^{N \times F}$ denote the feature matrix of all graph nodes. Mathematically, using the output of the l -th layer as the input for the next layer, each neural network layer can be formulated as a nonlinear function:

$$\mathbf{H}^{(l+1)} = f(\mathbf{H}^{(l)}, \mathbf{A}) \quad (5)$$

where $\mathbf{H}^{(l)}$ corresponds to the feature matrix of the l -th layer, and $\mathbf{H}^{(0)} = \mathbf{X}$ is the input feature matrix of the first layer. Specific GCNs models differ only in the manner in which the nonlinear function $f(\cdot)$ is instantiated. A simple example of $f(\cdot)$ is as follows:

$$f(\mathbf{H}^{(l)}, \mathbf{A}) = \sigma(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \quad (6)$$

where $\sigma(\cdot)$ denotes a nonlinear activation function, such as a Rectified Linear Unit (ReLU), and $\mathbf{W}^{(l)}$ denotes a trainable weight matrix for the l -th layer. With this propagation rule, the neighbour's features are aggregated to represent each node at every layer, and the features become increasingly abstract by stacking layers on top of each other. However, there exist two limitations: the propagation rule simply aggregates the features

Table 1: Notations and meanings.

Notations	Descriptions
\mathcal{G}	Original graph
\mathcal{G}_o	Observed graph
\mathbf{A}	Adjacency matrix of graph
\mathcal{E}_m	Missing links
\mathcal{E}_s	Spurious links
\mathcal{D}	Dataset that contains augmented graphs
$\mathbf{H}^{(l)}$	Feature matrix of l -th neural network layer
$\mathbf{W}^{(l)}$	Trainable weight matrix for the l -th layer
$\ \cdot\ _0$	ℓ_0 -norm
$\ \cdot\ _{2,1}$	$\ell_{2,1}$ -norm

of neighboring nodes but not the node itself, and the multiplication with \mathbf{A} expected to change the scale of the feature vectors. That is, the nodes with a high degree will have a larger value, and the nodes with a low degree may have smaller values. To address the problems, a new propagation function, $f(\cdot)$, is presented as follows:

$$f(\mathbf{H}^{(l)}, \mathbf{A}) = \sigma(\hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(l)} \mathbf{W}^{(l)}) \quad (7)$$

where $\hat{\mathbf{A}}$ is obtained by adding an identity matrix \mathbf{I} to the adjacency matrix $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\hat{\mathbf{D}}$ denotes the diagonal node degree matrix of $\hat{\mathbf{A}}$, and $\hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}}$ denotes symmetric normalization.

3.3. Low-rank and Sparse Modeling

Traditionally, Principal component analysis (PCA) was proposed to determine a low-dimensional representation of data while retaining as much information as possible. However, the PCA is particularly effective when dealing with Gaussian noise, which is independent and identically distributed with respect to the original data. Hence, the Robust Principal Component Analysis (RPCA) [9] has been proposed to eliminate the effect of erratic noise (outliers). PCA and RPCA methods implicitly assume that the underlying data structure is a single low-rank subspace; however, real-world data may be drawn from a union of multiple subspaces, and therefore, modeling may be inaccurate. To this end, Low-Rank Representation (LRR) [28] has been proposed.

Considering the correlation between the connectivity patterns of nodes in real-world graphs, the adjacency matrix of the graphs should be low-rank. In other words, the rows or columns of the adjacency matrix must not be linearly independent. Thus, assuming that hidden non-zero entries representing missing links can be recovered according to the adjacency matrix, [37] proposed an RPCA-based link prediction method, which is formulated as the following optimization problem:

$$\min_{\mathbf{X}^*, \mathbf{E}} \text{rank}(\mathbf{X}^*) + \gamma \|\mathbf{E}\|_0 \text{ s.t., } \mathbf{A} = \mathbf{X}^* + \mathbf{E} \quad (8)$$

where $\text{rank}(\mathbf{X}^*)$ denotes the rank of matrix \mathbf{X}^* , $\|\cdot\|_0$ is the ℓ_0 -norm, and γ denotes the balancing parameter. The method searches for \mathbf{X}^* with a low rank as low as possible and \mathbf{E} as sparse as possible from \mathbf{A} . Moreover, by representing a network structure with as few representative subgraphs as possible, [53] proposed an LRR-based link prediction method, wherein networks could be modeled via a low-rank and sparse represen-

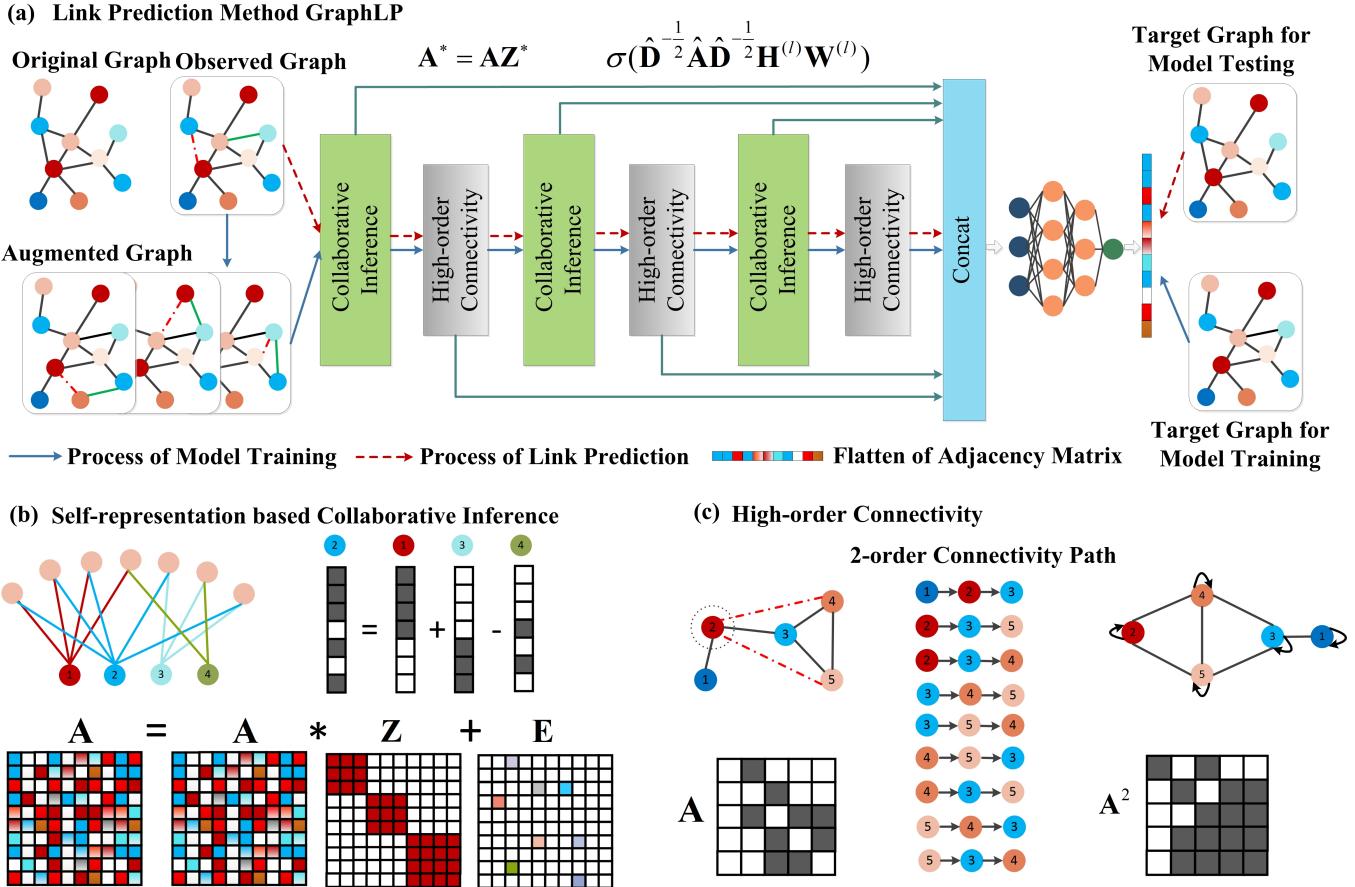


Figure 2: Demonstration of our link prediction method, GraphLP. (a) Link prediction method, GraphLP. The original graph is perturbed using a random mapping mechanism to obtain the observed graph; after this, the observed graph is further perturbed to generate augmented graphs. These augmented graphs are fed into GraphLP to learn the model using the observed graph as the label. Subsequently, the learned model is used to infer the original graph based on the observed graph. (b) Self-representation-based collaborative inference. Based on the structural regularity of graphs, the original graph can be reconstructed by utilizing the correlation between subgraph patterns. (c) Example of high-order connectivity. In addition to the 1-hop neighborhood, multi-hop connectivity influences the existence of links. The right graph represents the two-hop connectivity of the graph on the left, and the red dotted lines in the left graph provide an example of the two-hop connectivity path of node 2.

tation, as follows:

$$\min_{Z, E} \text{rank}(Z) + \alpha \|Z\|_0 + \beta \|E\|_{2,1} \text{ s.t., } A = AZ + E \quad (9)$$

where Z denotes the representation matrix reflecting the organization principle of the network, and $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$ -norm.

The notations used in this study are listed in Table 1.

4. The Proposed Method

This section presents the proposed link prediction method, GraphLP. As depicted in Figure 2, the framework of GraphLP consists of three main components:

- Collaborative inference operation. There exist certain similarities between the connection patterns of individuals in a complex system such that the perturbed structure of real-world graphs can be recovered globally based on the correlation between subgraph patterns (Section 4.1).
- High-order connectivity computation. The existence of a link between any two target nodes is intended to be primarily determined by the connectivity degree between nodes,

i.e., the number of paths and their length. Thus, the likelihood of a link can be estimated locally by computing the connectivity (Section 4.2).

- Pattern fusion operation. In addition to the first-order adjacency matrix, the connection patterns of nodes in the high-order adjacency matrix are also considered to be correlated, and the high-order connectivity can be reconstructed based on the collaborative inference. Thus, the graph topology can be estimated by fusing the k -order ($k \geq 1$) adjacency matrix (Section 4.3).

4.1. Collaborative Inference Operation

[32] suggested that link formation in real-world graphs is usually driven by both regular and irregular factors, and the former can be explained based on the mixture of multiple mechanisms, such as homophily, triadic closure, preferential attachment. Meanwhile, assuming that high-dimensional data are a mixture of simple data and are drawn from a union of multiple low-dimensional linear subspaces, the LRR has been proposed to represent the data $A = [a_1, a_2, \dots, a_N]$ as a linear combination of the basis in a "dictionary" $D = [d_1, d_2, \dots, d_M]$:

$$\min_Z \text{rank}(Z) \text{ s.t., } A = DZ, \quad (10)$$

Thus, the optimal representation matrix \mathbf{Z}^* uncovers the underlying subspaces in the data. By using each subspace to model a homogeneous subset of the data, multiple subspaces in LRR can capture heterogeneous structures within the data. Therefore, considering the above ideas, the regular structure of real-world graphs can be described appropriately by the LRR model, wherein the generation mechanisms of graph organization essentially corresponds to subspaces and the low rankness constraint captures the global correlation in graphs. Meanwhile, based on the generation mechanisms of graph organization, individual nodes may have similar connection patterns, and substructures that follow the same generation mechanism can be represented by each other, as depicted in Figure 2(b). Therefore, by using the adjacency matrix \mathbf{A} as the dictionary, the real-world graph can be represented by itself, as follows:

$$\min_{\mathbf{Z}} \text{rank}(\mathbf{Z}) \text{ s.t., } \mathbf{A} = \mathbf{AZ}. \quad (11)$$

In addition to their regular structure, real-world graphs also contain irregular components. Thus, we let matrix \mathbf{E} denote such irregular connections; then, the proposed self-representation model can be modified as $\mathbf{A} = \mathbf{AZ} + \mathbf{E}$. According to the LRR, data are considered to be "sample specific", and the ℓ_{21} -norm is adopted to constrain the matrix \mathbf{E} , i.e., $\|\mathbf{E}\|_{2,1}$. However, although the proposed method can be used to model real-world graphs, the low-rank model and ℓ_{21} -norm constraints are usually solved using Alternating direction method (ADM), which requires a large number of iterations and has high complexity. Therefore, a reasonable strategy is to relax the constraints with Frobenius norm:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{\mathcal{F}}^2 + \lambda \|\mathbf{A} - \mathbf{AZ}\|_{\mathcal{F}}^2 \text{ s.t., } \mathbf{A} = \mathbf{AZ} + \mathbf{E} \quad (12)$$

Let $\mathcal{L} = \|\mathbf{Z}\|_{\mathcal{F}}^2 + \lambda \|\mathbf{A} - \mathbf{AZ}\|_{\mathcal{F}}^2$ denote the partial derivative of \mathcal{L} with respect to \mathbf{Z} is $\partial\mathcal{L}/\partial\mathbf{Z} = 2\mathbf{Z} + \lambda(-2\mathbf{A}^T\mathbf{A} + 2\mathbf{A}^T\mathbf{AZ})$. By setting $\partial\mathcal{L}/\partial\mathbf{Z} = 0$, the optimal representation \mathbf{Z}^* can be obtained as follows:

$$\mathbf{Z}^* = \lambda(\lambda\mathbf{A}^T\mathbf{A} + \mathbf{I})^{-1}\mathbf{A}^T\mathbf{A}. \quad (13)$$

where \mathbf{I} denotes the identity matrix. Thus, in the case that the clean data is sufficient enough to represent the graph's structural patterns and the irregular connections are properly characterized, the structure perturbations can be inferred using \mathbf{AZ}^* . Hence, the collaborative inference operation is defined as follows:

$$\mathcal{C}\mathcal{G}(\mathbf{A}) = \lambda\mathbf{A}(\lambda\mathbf{A}^T\mathbf{A} + \mathbf{I})^{-1}\mathbf{A}^T\mathbf{A} \quad (14)$$

4.2. High-order Connectivity Computation

According to local similarity indices for link prediction, the more the number of paths two nodes possess, the greater the similarity between them. Specifically, two nodes with a high mutual connectivity are more likely to generate a link between them. Thus, n -hop-based ($n \geq 2$) paths must be explored to characterize the local structural features for link prediction. Using a deep learning framework, the n -hop computation can be decomposed into two-hop operations on each neural layer. Hence, a high-order connectivity computation calculates the two hop connectivity of graph nodes in each layer, and the mutual connectivity of two nodes can be estimated by stacking the high-order connectivity computation mechanism. Assuming that the integer powers of the adjacency matrix characterizes the mutual connectivity of graph nodes, that is, $[\mathbf{A}^n]_{ij}$ denotes the number of paths

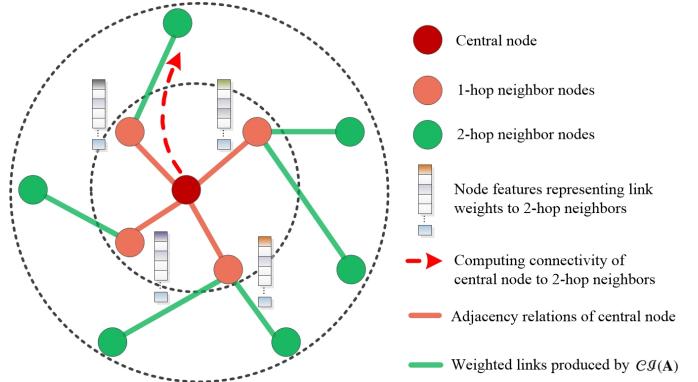


Figure 3: Illustration of high-order connectivity computation.

with length n connecting nodes i and j , the high-order connectivity computation in each neural layer can be defined based on the idea of the second power of adjacency matrix \mathbf{A} . From the perspective of graph convolution networks, high-order connectivity computation can be defined as

$$\mathcal{HCC}\mathcal{A}(\mathbf{A}) = \hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathcal{C}\mathcal{G}(\mathbf{A}), \quad (15)$$

where the weighted adjacency matrix generated by the proposed collaborative inference operation is viewed as the features of graph nodes. Figure 3 illustrates a high-order connectivity computation. As presented in Equation (15), the global and local structural features can be captured for link prediction at the level of individual nodes and edges. Thus, the nonlinear propagation function can be defined as follows:

$$\mathbf{H}^{(l+1)} = \hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathcal{C}\mathcal{G}(\mathbf{H}^{(l)})\mathbf{W}^{(l)}. \quad (16)$$

Thus, the hierarchical structure of real-world graphs can be characterized by executing the nonlinear propagation function iteratively, in which $\mathcal{HCC}\mathcal{A}(\mathbf{H}^{(l)}) = \hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathcal{C}\mathcal{G}(\mathbf{H}^{(l)})$ represents the high-order connectivity of graph nodes, as depicted in Figure 2(c), and $\mathcal{C}\mathcal{G}(\mathbf{H}^{(l+1)})$ denotes the collaborative inference.

4.3. Pattern Fusion Operation

To estimate the likelihood of potential links, the output of the $(l-1)$ -th layer, i.e., $\mathbf{H}^{(l)}$, is fed as the input of the l -th layer. Based on $\mathcal{C}\mathcal{G}(\mathbf{H}^{(l)})$ and $\mathcal{HCC}\mathcal{A}(\mathbf{H}^{(l)})$, the shallow layers extract the low-order global and local structure features, while the deep layers extract the high-order global and local structure features. Meanwhile, the effective range that the local structure features drawn from increases as the model depth increases. Therefore, the structure features in different range at various order, i.e., $\mathcal{HCC}\mathcal{A}(\mathbf{H}^{(l)})$ and $\mathcal{C}\mathcal{G}(\mathbf{H}^{(l)})$, $0 \leq l \leq L$, all contribute to the inference of potential links, although the exact extent of their contribution depends on the graph data.

To overcome the issues mentioned above, in addition to being used as the inputs of the next layer, the outputs of neural network layers are mapped to skip a block of several layers based on residual connections, as illustrated in Figure 2(a). Next, all outputs are concatenated and used as the input of a two-layer Multi-layer Perceptron (MLP), which is defined as:

$$\mathbf{O} = \text{MLP}(\text{concat}(\mathcal{C}\mathcal{G}(\mathbf{H}^{(l)}), \mathcal{HCC}\mathcal{A}(\mathbf{H}^{(l)}))), 0 \leq l \leq L. \quad (17)$$

where \mathbf{O} is a vector containing the probabilities of links between all possible node pairs, and missing and spurious links can be inferred based on it.

4.4. Model Training

To train the proposed model, augmented graphs generated based on the observed graph are used as training data, and the adjacency matrix of the observed graph is flatten as its labels \mathbf{Y} , where \mathbf{Y}_{i*N+j} denotes the existence of the link between nodes i and j . Correspondingly, \mathbf{O} represents the prediction results obtained by the proposed model for all possible links. Here, the Binary Cross-Entropy (BCE) is used as the loss function:

$$\mathcal{L} = -\frac{1}{N^2} \sum_{i=1}^{N^2} \mathbf{Y}_i \log(\mathbf{O}_i) + (1 - \mathbf{Y}_i) \log(1 - \mathbf{O}_i). \quad (18)$$

The learned model is then deployed on the observed graph to reconstruct the original graph. The training process of GraphLP is outlined in Algorithm 1.

Algorithm 1 Training Process of GraphLP

Input: Training set \mathcal{D}_{train} , validation set \mathcal{D}_{val} , and test set \mathcal{D}_{test} , number of neural network layers L .

Output: The well-trained model GraphLP.

```

1: while not convergence do
2:   for  $0 \leq l \leq L$  do
3:     Conduct collaborative inference operation using (14);
4:     Compute high-order connectivity using (16);
5:   end for
6:   Fuse the outputs based on MLP using (17);
7:   Update the model by minimizing the loss function (18);
8: end while
```

4.5. Model Analysis

(1) Generalized local similarity indices. The high-order connectivity computation $\mathcal{HCC}\mathcal{A}(\mathbf{H}^{(l)})$ in every neural network layer is essentially the second power of the adjacency matrix, and it obtains the connectivity of node pairs within two-hop neighborhood. As the model depth increases, the connectivity of node pairs in a wider range is considered. Thus, GraphLP can degenerate to $S_{ij} = \mathbf{A}^2 + \alpha\mathbf{A}^3 + \beta\mathbf{A}^4 + \dots$ when collaborative inference and deep learning mechanism are abolished.

(2) Connection to WalkPool. WalkPool [36] first generates node representations based on GNN and encodes them into edge weights of the extracted enclosing subgraphs; following this, it uses the edge weights to compute the transition probabilities of random walk. Next, the method calculates a list of features based on the transition probabilities to classify the subgraphs. However, for an enclosing subgraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, its variants $\mathcal{G}^+ = (\mathcal{V}, \mathcal{E} \cup \{i, j\})$ and $\mathcal{G}^- = (\mathcal{V}, \mathcal{E} \setminus \{i, j\})$ are used as positive and negative samples, respectively. In essence, this method discriminates only those subgraphs that differ by a single edge and is not suitable for practical link prediction scenarios. In contrast, GraphLP can predict any potential links based on graph structure features.

(3) Connection to LFLP. The LFLP [53] constructs an adjacency matrix based on a self-representation model and then combines it with the observed network to identify missing and spurious links. The collaborative inference operation $\mathcal{C}\mathcal{I}(\mathbf{H}^{(l)})$ of our work is similar to that in the LFLP with respect to modeling the global structure of graphs; however, the difference is that only low-order global structural features are considered in LFLP, whereas multi-order global and local structural features are characterized based on the deep-learning framework in GraphLP.

5. Experiments

Further, extensive experiments are conducted on real-world graphs to evaluate the performance of the proposed method GraphLP: (1) Compare GraphLP with state-of-the-art methods; (2) Compare GraphLP with traditional baseline methods; (3) Model architecture analysis; (4) Model sensitivity analysis. Here, Area Under Curve (AUC) and Average Precision (AP) are adopted to evaluate the performance of the methods. Furthermore, Precision is used to verify the superiority of GraphLP over traditional link prediction methods. Based on the link prediction results \mathbf{O} , the scores are sorted in descending and ascending orders, and following this, their top-L links are taken as the predicted missing and spurious links. Note that Precision is defined by calculating the ratio of accurately discovered links to the total number of links in the probe set:

$$\text{Precision} = \mathcal{T}/\mathcal{R} \quad (19)$$

where \mathcal{T} is the number of accurately identified links, and \mathcal{R} is the total number of links in the probe set.

5.1. Experimental Settings

5.1.1. Experimental Datasets

Herein, seven widely used graph datasets are used for link prediction. (1) USAir [40]. This is the transportation network of the United States, including 332 airports as nodes and 2,126 airlines as edges, connecting the United States worldwide. The average node degree is 12.81. (2) C.elegans [49]. This is a neural network of *C. elegans*, with 297 neurons representing nodes and 2,148 synaptic connections representing edges. The average node degree is 14.46. (3) PB [1]. This dataset is a network of hyperlinks between weblogs on US political blogs, with 1,222 blogs on US politics as nodes and 16,714 hyperlinks between blogs as edges. The average node degree is 27.36. (4) NS [33]. This is an undirected co-authorship network with 1,589 nodes and 2,742 edges, where the nodes denote the scientists engaged in network science research, and the edges denote two scientists have co-authored a publication. The average node degree is 3.45. (5) Yeast [47]. This represents a protein-protein interaction network formed in yeast with 2,375 proteins as nodes and 11,693 protein-protein interactions as edges. The average node degree is 9.85. (6) E.coli [58]. This is a pairwise reaction network of metabolites with 1,805 nodes and 14,660 edges. The average node degree is 12.55. (7) Router [43]. It is a snapshot of the Internet structure at the level of autonomous systems, with 5,022 nodes and 6,258 edges, in which the nodes represent routers and the edges represent the data transmission between routers. The average node degree is 2.49. The properties of the datasets are listed in Table 2.

To extensively validate the performance of the proposed method, 90% and 50% of the links of the original graph are selected randomly to first construct the observed graphs. Thereafter, based on the observed graph \mathcal{G}_o , 10% nonexisting links are added randomly as spurious links, and 10% existing links are removed randomly as missing links, denoted as \mathcal{E}_{del} and \mathcal{E}_{add} respectively, to generate the augmented graph set $\mathcal{D} = \{\mathcal{G}_i | i = 1, \dots, t\}$. Following this, 90% and 10% graphs are randomly selected from \mathcal{D} as the training and validation set, respectively, and the observed graph \mathcal{G}_o is used as the test set.

Table 2: Summary of the datasets. ACC is the average clustering coefficient, and AD is the average node degree.

Dataset	USAir	NS	PB	Yeast	C.ele	Router	E.coli
Node	332	1589	1222	2375	297	5022	1805
Edges	2126	2742	16714	11693	2148	6258	14660
ACC	0.625	0.638	0.320	0.306	0.292	0.012	0.516
AD	12.81	3.45	27.36	9.85	14.46	2.49	12.55

5.1.2. Comparison Methods

The proposed method was compared with six state-of-the-art deep learning-based link prediction methods, including:

(1) Weisfeiler-Lehman graph kernel (WLK) [42] is a fast feature extraction scheme based on the WL test for graph isomorphism, which maps the original graph to a graph sequence and adds the pair-wise similarities between the graphs.

(2) Weifeiler-Lehmam Neural Machine (WLNM) [56] is a subgraph classification-based link prediction method that leverage deep learning to automatically learn topological features from enclosing subgraphs.

(3) Node2Vec [20] is a network embedding method that encodes proximity information into low-dimensional vectors. The node features and low-dimensional vectors are then fed into the MLP for link prediction.

(4) LINE [44] learns network embeddings that preserve the first-order and second-order proximity, and the resulting low-dimensional vectors are used for link prediction.

(5) SEAL [57] extracts the enclosing subgraphs of positive and negative links and marks different roles of their nodes. The method then trains a GNN based on the node information matrix to classify subgraphs for link prediction.

(6) WalkPool (WP) [36] is a subgraph classification-based link prediction method. It encodes node feature and graph topology into the transition probabilities of random walk, and following this, a list of features is computed to classify subgraphs.

5.1.3. Parameter Settings

GraphLP is implemented on a Pytorch platform with a NVIDIA GeForce RTX GPU and optimized using Adam optimizer. All models are implemented using Python 3.6. The learning rate is set to 0.0012 for the NS dataset and 0.0005 for the other graphs. For all the datasets, the weight decay is set to 0.0. The number of epochs on the E.coli and Yeast dataset is 300, whereas it was 200 on the other datasets. Dropout is applied to the MLP, and the dropout rate is set to 0.5 on Router and 0.2 on the others. The trade-off parameter λ is set to 0.13, and the number of neural network layers in the GraphLP is set to three. The detailed hyperparameter settings for the model are listed in Table 3.

Table 3: Hyperparameter setting for the proposed method.

Name	Value
optimizer	Adam
loss function	binary_cross_entropy
learning rate	NS=0.0012, others=0.0005
weight decay	0.0
epochs	Ecoli, Yeast=300, others=200
dropout	Router=0.5, others=0.2
λ	0.13
number of network layers	3

5.2. Experimental Result

For 90% of the observed links, the results about the AUC and AP with standard deviations are presented in Table 4 and 5, which indicate that GraphLP significantly outperforms other state-of-the-art algorithms in terms of both AUC and AP, with exception of the NS and Router datasets. The results demonstrate that the learning of local and global graph structure entirely characterizes the underlying structural patterns; thus, the missing links and spurious links can be better identified. Table 4 indicates that GraphLP significantly improves the AUC on the PB, C.ele, and Router datasets, with approximately 4%, 7%, and 3% performance improvement, respectively, compared to the WP algorithm. In addition, the proposed method still performs better than other state-of-the-art methods on the USAir, Yeast and NS datasets. Moreover, the results for the AP presented in Table 5 also indicate that GraphLP outperforms state-of-the-art methods on most of datasets, and GraphLP achieves a maximum performance enhancement of approximately 9% compared to the best performing graph neural network method WP.

For 50% of the observed links, the results also demonstrate that the proposed model achieves remarkable performance compared to the methods, as described in Table 6 and 7. The results illustrate that, as the amount of structure perturbation increases, GraphLP can still appropriately learn the real graph structure, thus recovering the original graph effectively. Therefore, the values of the AUC and AP decreased to a lower extent. Furthermore, by comparing Table 4 with Table 6 and Table 5 with Table 7, we can infer that the AUC and AP values drop faster for the other state-of-the-art methods than those for GraphLP, which demonstrates that GraphLP can better capture the underlying structural patterns to demonstrate better performance.

5.3. Compared with Traditional Link Prediction Methods

To further verify the proposed method, the precision of GraphLP and traditional link prediction methods are calculated based on the following datasets: (1) Macaque [15], cortical networks of the macaque monkey; (2) Mangwet [5], the food web in Mangrove Estuary during the wet season; (3) Jazz [19], a collaboration network of jazz musicians; (4) Metabolic [17], a metabolic network of C.elegans; (5) USAir, (6) C.ele, (7) E.coli and (8) Yeast. Here, six representative traditional link prediction methods are selected for comparison.

(1) The CN [29] metric is among the most widely used methods for link prediction problem. It assumes that two nodes will be more easily connected if they share more common neighbors;

(2) The RA [61] metric is inspired by the physical processes involved in resource allocation, which suppresses the contribution of high-degree common neighbors;

(3) The LP [61] index measures the structural similarity of node pairs within three-hops;

(4) The Non-negative Matrix Factorization (NMF) [26] model is used for structure prediction by learning the latent features of real-world graphs;

(5) The Robust Principal Component Analysis (RPCA) [37] represents a real-world graph based on the sparsity and low rank property of its adjacency matrix and infers potential links based on matrix completion.

(6) The LFLP [53] uses a self-representation model to reconstruct the original graph based on a few representative subgraphs.

The results of missing link prediction with respect to Precision are shown in Table 8. For each network, the bold number

Table 4: Prediction measured by AUC (90% observed links). **Bold** numbers are the best results of all methods.

Data	USAir	NS	PB	Yeast	C.ele	Router	E.coli
WLK	96.63 \pm 0.73	98.57 \pm 0.51	93.83 \pm 0.59	95.86 \pm 0.54	89.72 \pm 1.67	87.42 \pm 2.08	96.94 \pm 0.29
WLN M	95.95 \pm 1.10	98.61 \pm 0.49	93.49 \pm 0.47	95.62 \pm 0.52	86.18 \pm 1.72	94.41 \pm 0.88	97.21 \pm 0.27
Node2Vec	91.44 \pm 1.78	91.52 \pm 1.28	85.79 \pm 0.78	93.67 \pm 0.46	84.11 \pm 1.27	65.46 \pm 0.86	90.82 \pm 1.49
LINE	81.47 \pm 10.71	80.63 \pm 1.90	76.95 \pm 2.76	87.45 \pm 3.33	69.21 \pm 3.14	67.15 \pm 2.10	82.38 \pm 2.19
SEAL	97.09 \pm 0.7	98.85 \pm 0.41	95.01 \pm 0.34	97.91 \pm 0.52	90.30 \pm 1.35	96.38 \pm 1.45	97.64 \pm 0.22
WP	98.68 \pm 0.48	98.95 \pm 0.41	95.60 \pm 0.37	98.37 \pm 0.25	95.79 \pm 1.09	97.27 \pm 0.28	98.58 \pm 0.19
GraphLP	99.26 \pm 1.01	99.64 \pm 0.98	99.73 \pm 0.25	99.41 \pm 0.15	99.90 \pm 0.14	99.02 \pm 0.19	99.23 \pm 0.23

Table 5: Prediction measured by AP (90% observed links). **Bold** numbers are the best results of all methods.

Data	USAir	NS	PB	Yeast	C.ele	Router	E.coli
WLK	96.82 \pm 0.84	98.79 \pm 0.40	93.34 \pm 0.89	96.82 \pm 0.35	88.96 \pm 2.06	86.59 \pm 2.23	97.25 \pm 0.42
WLN M	95.95 \pm 1.13	98.81 \pm 0.49	92.69 \pm 0.64	96.40 \pm 0.38	85.08 \pm 2.05	93.53 \pm 1.09	97.50 \pm 0.23
Node2Vec	89.71 \pm 2.97	94.28 \pm 0.91	84.79 \pm 1.03	94.90 \pm 0.38	83.12 \pm 1.90	68.66 \pm 1.49	90.87 \pm 1.48
LINE	97.70 \pm 11.76	85.17 \pm 1.65	78.82 \pm 2.71	90.55 \pm 2.39	67.51 \pm 2.72	71.92 \pm 1.53	86.45 \pm 1.82
SEAL	97.13 \pm 0.80	99.06 \pm 0.37	94.55 \pm 0.43	98.33 \pm 0.37	89.48 \pm 1.85	96.23 \pm 1.71	98.03 \pm 0.20
WP	98.66 \pm 0.55	99.09 \pm 0.29	95.28 \pm 0.41	98.64 \pm 0.28	91.53 \pm 1.33	97.20 \pm 0.38	98.79 \pm 0.21
GraphLP	99.91 \pm 1.03	98.94 \pm 0.96	98.32 \pm 1.43	98.74 \pm 0.16	99.41 \pm 0.42	79.30 \pm 0.19	98.96 \pm 0.19

Table 6: Prediction measured by AUC (50% observed links). **Bold** numbers are the best results of all methods.

Data	USAir	NS	PB	Yeast	C.ele	Router	E.coli
WLK	91.93 \pm 0.71	87.27 \pm 1.71	92.54 \pm 0.33	91.15 \pm 0.35	83.29 \pm 0.89	71.25 \pm 4.37	92.38 \pm 0.46
WLN M	91.42 \pm 0.95	87.61 \pm 1.63	90.93 \pm 0.23	92.22 \pm 0.32	75.72 \pm 1.33	86.10 \pm 0.52	92.81 \pm 0.30
Node2Vec	84.63 \pm 1.58	80.29 \pm 1.20	79.29 \pm 0.67	90.18 \pm 0.17	75.53 \pm 1.23	62.45 \pm 0.81	84.73 \pm 0.81
LINE	72.51 \pm 12.19	65.96 \pm 1.60	75.53 \pm 1.78	79.44 \pm 7.90	59.46 \pm 7.08	62.43 \pm 3.10	74.50 \pm 11.10
SEAL	93.36 \pm 0.67	90.88 \pm 1.18	93.79 \pm 0.25	93.90 \pm 0.54	82.33 \pm 2.31	86.64 \pm 1.58	94.18 \pm 0.41
WP	95.50 \pm 0.74	90.97 \pm 0.96	94.57 \pm 0.16	95.00 \pm 0.21	87.62 \pm 1.39	88.13 \pm 0.61	95.33 \pm 0.30
GraphLP	98.97 \pm 0.15	97.08 \pm 0.14	98.19 \pm 0.10	98.74 \pm 0.25	97.96 \pm 0.14	98.10 \pm 0.15	98.05 \pm 0.14

Table 7: Prediction measured by AP (50% observed links). **Bold** numbers are the best results of all methods.

Data	USAir	NS	PB	Yeast	C.ele	Router	E.coli
WLK	93.34 \pm 0.51	89.97 \pm 1.02	92.34 \pm 0.34	93.55 \pm 0.46	83.20 \pm 0.90	75.49 \pm 3.43	94.51 \pm 0.32
WLN M	92.54 \pm 0.81	90.10 \pm 1.11	91.01 \pm 0.20	93.93 \pm 0.20	76.12 \pm 1.08	86.12 \pm 0.68	94.47 \pm 0.21
Node2Vec	82.51 \pm 2.08	86.01 \pm 0.87	77.21 \pm 0.97	92.45 \pm 0.23	72.91 \pm 1.74	66.77 \pm 0.57	85.41 \pm 0.94
LINE	71.75 \pm 11.85	71.53 \pm 0.97	78.72 \pm 1.24	83.06 \pm 9.70	60.71 \pm 6.26	64.87 \pm 6.76	75.98 \pm 14.45
SEAL	94.15 \pm 0.54	92.21 \pm 0.97	93.42 \pm 0.19	95.32 \pm 0.38	81.99 \pm 2.18	87.79 \pm 1.71	95.67 \pm 0.24
WP	95.87 \pm 0.74	92.33 \pm 0.76	94.22 \pm 0.27	96.15 \pm 0.13	86.25 \pm 1.42	89.17 \pm 0.55	96.36 \pm 0.34
GraphLP	97.96 \pm 0.09	93.08 \pm 0.08	96.27 \pm 0.10	97.27 \pm 0.09	95.89 \pm 0.11	79.23 \pm 0.14	96.48 \pm 0.13

Table 8: The precision (90% observed links) of missing links prediction. **Bold** numbers are the best results of all methods.

Data	Macaque	Mangwet	Jazz	Metabolic	USAir	C.ele	E.coli	Yeast
RA	0.5099	0.1292	0.5547	0.2451	0.4443	0.093	0.4857	0.2609
CN	0.5695	0.125	0.5292	0.1127	0.3764	0.091	0.4399	0.1334
LP	0.5483	0.1319	0.5109	0.1275	0.3821	0.089	0.4837	0.1454
NMF	0.7316	0.4398	0.5309	0.2315	0.3981	0.1270	0.5013	0.3812
RPCA	0.7421	0.5421	0.6138	0.1842	0.3596	0.098	0.3418	0.5359
LFLP	0.7605	0.5572	0.5956	0.3241	0.4545	0.2010	0.5007	0.60
GraphLP	0.7881	0.7986	0.8212	0.7030	0.8208	0.8224	0.7169	0.7374

Table 9: The precision (90% observed links) of spurious links prediction. **Bold** numbers are the best results of all methods.

Data	Macaque	Mangwet	Jazz	Metabolic	USAir	C.ele	E.coli	Yeast
RA	0.5490	0.1380	0.5410	0.140	0.2650	0.2790	0.4171	0.1110
CN	0.5710	0.2880	0.5690	0.1670	0.2480	0.2458	0.3222	0.073
LP	0.5939	0.3280	0.7016	0.6911	0.6271	0.4780	0.6089	0.4585
NMF	0.8090	0.5660	0.6510	0.2430	0.4820	0.4333	0.4662	0.2330
RPCA	0.810	0.5180	0.5920	0.074	0.443	0.2609	0.3795	0.4250
LFLP	0.818	0.583	0.663	0.2210	0.5970	0.4390	0.4246	0.5680
GraphLP	0.9073	0.8750	0.9197	0.8812	0.9057	0.9533	0.8452	0.6210

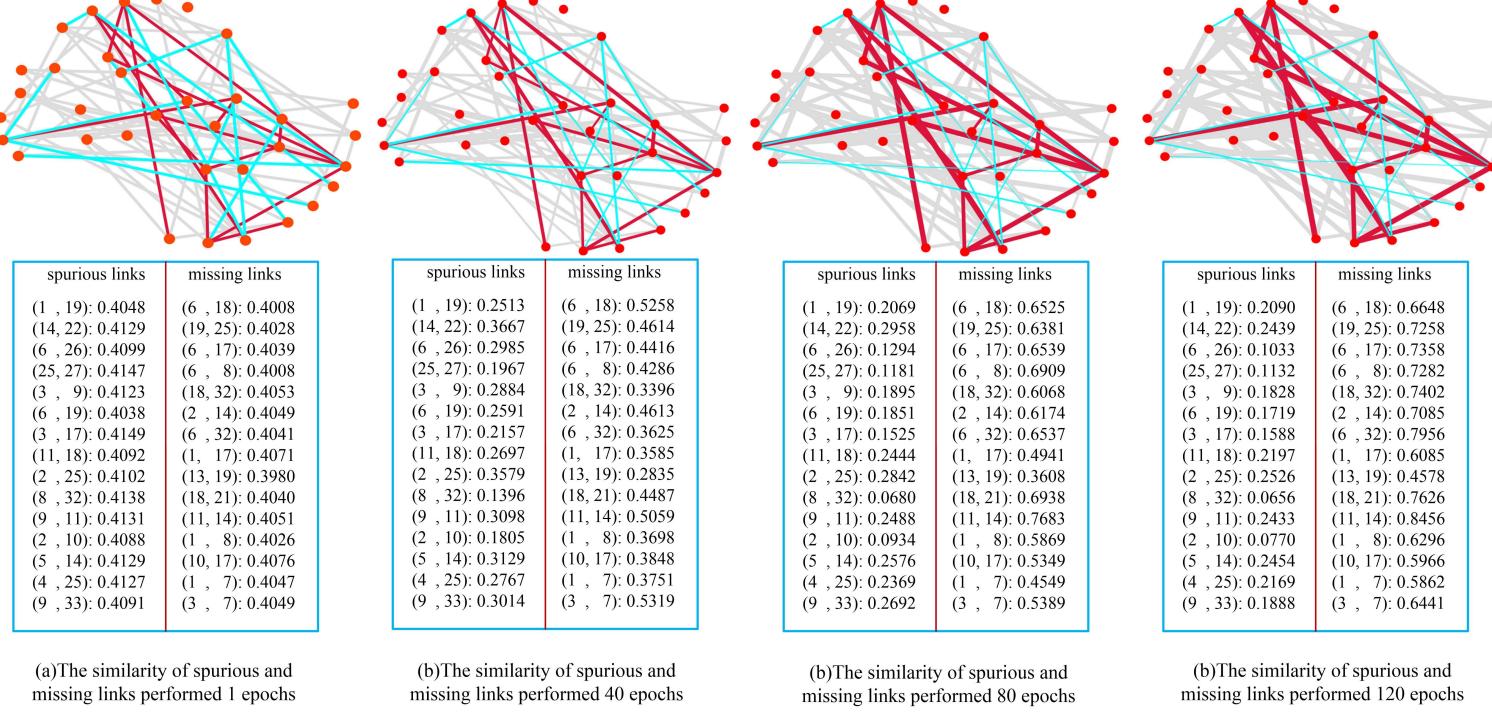


Figure 4: The topology visualization of Club dataset. The experiment performs 10% link perturbation, i.e. 10% spurious links are added and 10% missing links are deleted.

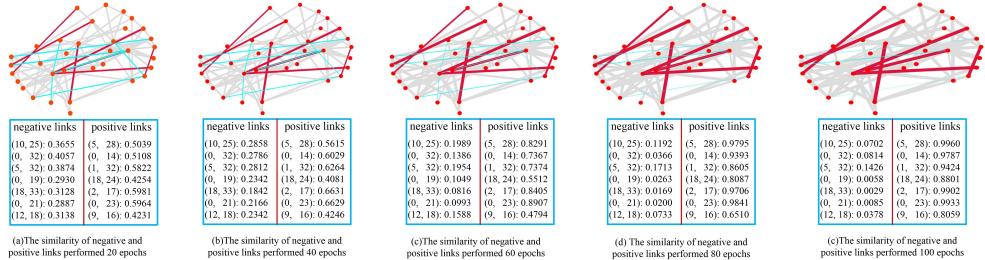


Figure 5: The topology visualization of Club dataset. The experiment performs 20% link perturbation, i.e. 20% spurious links are added and 20% missing links are deleted.

in the corresponding column indicates the highest accuracy. The results presented in Table 8 demonstrate that the proposed model GraphLP model performs the best among the methods. Furthermore, the link prediction accuracy of the proposed model is far higher than that of the other methods, which can be at least three times higher than that of the best-performing method. For spurious links prediction, the results measured by Precision are listed in Table 9. For all networks, GraphLP performs the best among the methods and is remarkably better than the second best algorithm. The results presented in Table 8 and 9 demonstrate that GraphLP has stronger ability to learn structural features, and can recover the structure of the original network more accurately.

Based on Table 2, it can be observed that the Precision of our proposed model performs best, despite the large differences between the ACC and AD across all the datasets; thus indicates that the proposed model performs well for heterogeneous graph structures.

5.4. Recovered Graph Visualization

To verify the effectiveness of the proposed model of missing and spurious links inference, the topology of the recovered graphs in the model training process on Club dataset is visually compared, as depicted in Figure 4 and 5. The top half depicts the topology of the graphs, wherein the red links denote the missing

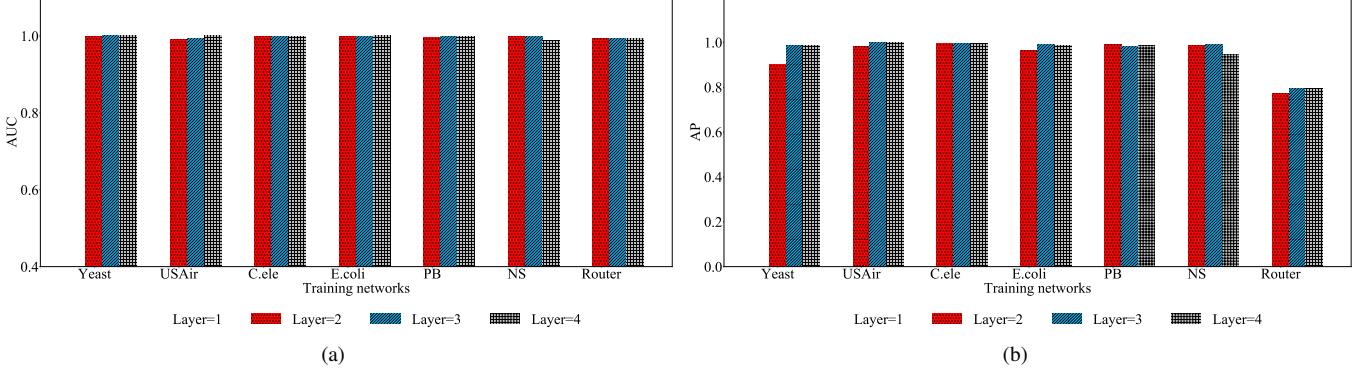


Figure 6: The performance of link prediction under various model depth. (a) AUC of link prediction method GraphLP with different layer number. (b) AP of link prediction method GraphLP with different layer number.

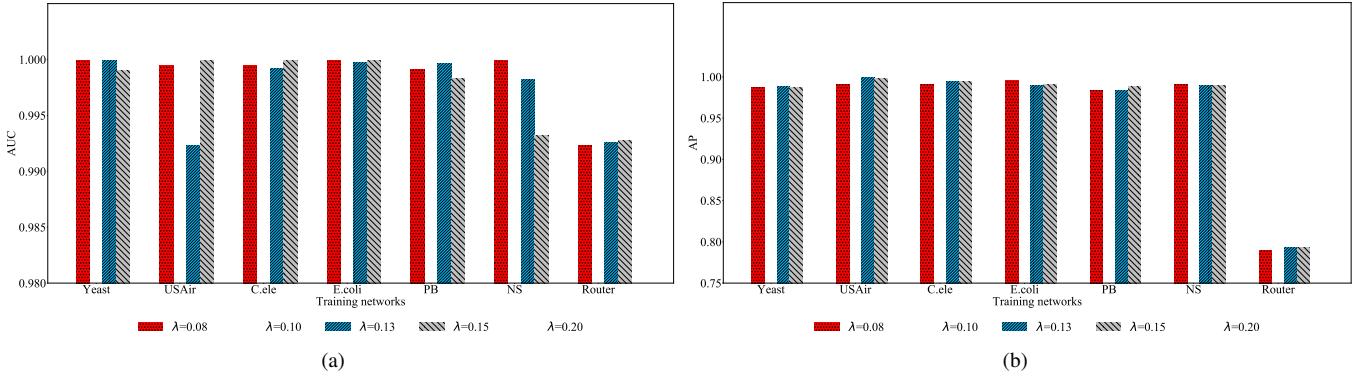


Figure 7: The performance of link prediction under different values of λ . (a) AUC of link prediction method GraphLP under different values of λ . (b) AP of link prediction method GraphLP under different values of λ .

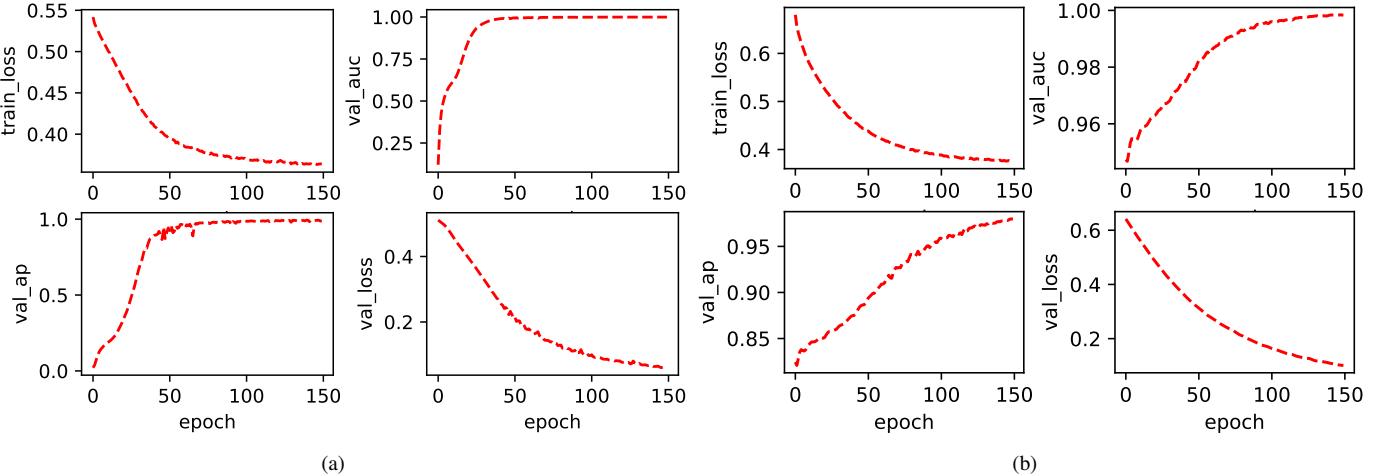


Figure 8: Visualization of model convergence. (a) Convergence of USAir dataset. (b) Convergence of C.ele dataset.

links, the blue links denote the spurious links, and the gray links denote the original links. The bottom half depicts the likelihood scores of missing and spurious links. Based on the results, it can be concluded that with an increase in epoch times, the likelihood scores of missing links increases gradually, and the likelihood scores of spurious links decline gradually. The widths of the lines indicate the following process: when the number of epochs reaches 100, the likelihood scores of missing links approach 1.0, and the likelihood scores of spurious links approach 0.0. This proves that the proposed GraphLP model can distinguish between missing links and spurious links and infer them

effectively. Moreover, to further prove the effectiveness of the proposed model, the topology of the recovered graphs in the model training process is visualized when 20% of the links of Club are perturbed, as depicted in Figure 5. Compared to Figure 4, we determined that the likelihood of missing and spurious links is weakened with an increase in the structure perturbation ratio. However, with an increase in the training epochs, the model is still able to distinguish and infer the missing and spurious links according to the structural patterns. For instance, when the training epoch reaches 140, the likelihood scores of missing links are greater than 0.5, and those of spurious links

are less than 0.3, indicating that the proposed model can still predict missing and spurious links with high accuracy.

5.5. Impact of Model Depth

Next, the performance of GraphLP is explored at various model depths. As depicted in Figure 6, the performance of the model in terms of the AUC and AP trained with one-layer neural network is poor; however, its performance improves significantly with an increase in the number of layers. In particular, when the model depth is equal to two, a significant performance improvement is noted. The primary reason for this is that the model with two-layers multi-order global and local structural features is integrated adaptively based on the MLP component, which considerably improves the performance of the model. Subsequently, as the layer number increases, a slight improvement in model performance is still noted. When the number of layers is four, the accuracy of GraphLP declines significantly on NS and fluctuates on other datasets. A possible reason for this is that the model with four layers becomes more complex, thereby requiring more training iterations or an appropriate learning rate [14]. In general, the performance of the proposed model is optimal when the depth is three, and a deep architecture is necessary.

5.6. Impact of Trade-off Parameter

To examine the sensitivity of the proposed model to the trade-off parameter, the AUC and AP values of link prediction methods with different λ are presented in Figure 7. Based on the results, it can be concluded that the performance of the proposed model is not sensitive to λ for most datasets. In Figure 7(a), for the USAir and NS datasets, the AUC value varies significantly under different λ , but the performance is still better than other of the other algorithms. In Figure 7(b), the AP value remains stable for different λ values, indicating that the proposed model is insensitive to different λ . Overall, our proposed algorithm exhibited satisfactory performance on most datasets with various λ .

5.7. The Convergence Analysis

Generally, GraphLP converges to optimal values after approximately 200 epochs on most datasets. In particular, Figure 8 plots the learning curves of GraphLP on the USAir and C.ele datasets, including the training loss, validation AUC, validation AP, and validation loss. The results indicate that the AUC and AP values increase rapidly with the decrease in training loss and validation loss, and these values converge to the optimal value when the validation loss approaches a minimum value. Additionally, we discover that validation loss is lower than training loss, and the difference between them remains relatively stable. A possible reason for this is that the dropout manipulation is only applied to the training process.

6. Conclusion

This paper aims to reconstruct graph structure to improve the performance of link prediction. In particular, unlike existing subgraph-classification-based discriminative methods, this work achieves the aforementioned objective by developing a generative GNN, namely GraphLP, which considered both global and local structure features and hierarchical structural patterns. Concurrently, a novel collaborative inference operation and high-order connectivity computation mechanism are developed. We

also present an analysis about the relation between GraphLP and other classical link prediction methods. Extensive experimental results demonstrate the superiority of the proposed method over other state-of-the-art models and traditional baseline methods. This could be a fruitful avenue for future research aimed at addressing graph learning tasks.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 62106030, 61802039, 62272066; Chongqing Municipal Postdoctoral Science Foundation under Grant No. csc2021jcyj-bsh0176; Chongqing Municipal Natural Science Foundation under Grant No. csc2020jcyj-msxmX0804; the Chongqing Research Program of Basic Research and Frontier Technology under Grant No. csc2021jcyj-msxmX0530.

References

- [1] Robert Ackland et al. Mapping the us political blogosphere: Are conservative bloggers more prominent? In *BlogTalk Downunder 2005 Conference, Sydney*. BlogTalk Downunder 2005 Conference, Sydney, 2005.
- [2] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.
- [3] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *nature*, 466(7307):761–764, 2010.
- [4] James Atwood and Don Towsley. Diffusion-convolutional neural networks. In *Advances in neural information processing systems*, pages 1993–2001, 2016.
- [5] Daniel Baird, J Luczkovich, and Robert R Christian. Assessment of spatial and temporal variability in ecosystem attributes of the st marks national wildlife refuge, apalachee bay, florida. *Estuarine, Coastal and Shelf Science*, 47(3):329–349, 1998.
- [6] Baruch Barzel and Albert-László Barabási. Network link prediction by global silencing of indirect correlations. *Nature biotechnology*, 31(8):720–725, 2013.
- [7] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.
- [8] Lei Cai, Jundong Li, Jie Wang, and Shuiwang Ji. Line graph neural networks for link prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [9] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.
- [10] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- [11] Zhiqian Chen, Fanglan Chen, Lei Zhang, Taoran Ji, Kaiqun Fu, Liang Zhao, Feng Chen, Lingfei Wu, Charu Aggarwal, and Chang-Tien Lu. Bridging the gap between spatial and spectral domains: A survey on graph neural networks. *arXiv preprint arXiv:2002.11867*, 2020.
- [12] Ara Cho, Junha Shin, Sohyun Hwang, Chanyoung Kim, Hongseok Shim, Hyojin Kim, Hanhae Kim, and Insuk Lee. Wormnet v3: a network-assisted hypothesis-generating server for *caenorhabditis elegans*. *Nucleic acids research*, 42(W1):76–82, 2014.
- [13] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.
- [14] Weilin Cong, Morteza Ramezani, and Mehrdad Mahdavi. On provable benefits of depth in training graph convolutional networks. *Advances in Neural Information Processing Systems*, 34:9936–9949, 2021.
- [15] Lucianoda F Costa, Marcus Kaiser, and Claus C Hilgetag. Predicting the connectivity of primate cortical networks from topological and spatial node properties. *BMC systems biology*, 1(1):1–17, 2007.
- [16] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3844–3852, 2016.
- [17] Jordi Duch and Alex Arenas. Community detection in complex networks using extremal optimization. *Physical review E*, 72(2):027104, 2005.
- [18] Zihong Fang, Shaolin Tan, Yaonan Wang, and Jinhua Lu. Elementary subgraph features for link prediction with neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

- [19] Pablo M Gleiser and Leon Danon. Community structure in jazz. *Advances in complex systems*, 6(04):565–573, 2003.
- [20] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.
- [21] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.
- [22] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- [23] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, 2002.
- [24] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. Detecting suspicious following behavior in multimillion-node social networks. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 305–306, 2014.
- [25] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.
- [26] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.
- [27] Boning Li, Yingce Xia, Shufang Xie, Lijun Wu, and Tao Qin. Distance-enhanced graph neural network for link prediction. In *ICML 2021 Workshop on Computational Biology*, 2021.
- [28] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012.
- [29] Francois Lorrain and Harrison C White. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80, 1971.
- [30] Paul Louis, Shweta Ann Jacob, and Amirali Salehi-Abari. Sampling enclosing subgraphs for link prediction. *arXiv preprint arXiv:2206.12004*, 2022.
- [31] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. Recommender systems. *Physics reports*, 519(1):1–49, 2012.
- [32] Linyuan Lü, Liming Pan, Tao Zhou, Yi-Cheng Zhang, and H Eugene Stanley. Toward link predictability of complex networks. *Proceedings of the National Academy of Sciences*, 112(8):2325–2330, 2015.
- [33] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.
- [34] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023. PMLR, 2016.
- [35] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814–818, 2005.
- [36] Liming Pan, Cheng Shi, and Ivan Dokmanić. Neural link prediction with walk pooling. In *International Conference on Learning Representations*, pages 5171–5181, 2021.
- [37] Ratha Pech, Dong Hao, Liming Pan, Hong Cheng, and Tao Zhou. Link prediction via matrix completion. *EPL (Europhysics Letters)*, 117(3):38002, 2017.
- [38] Ratha Pech, Dong Hao, Yan-Li Lee, Ye Yuan, and Tao Zhou. Link prediction via linear optimization. *Physica A: Statistical Mechanics and its Applications*, 528:121319, 2019.
- [39] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical review E*, 67(2):026112, 2003.
- [40] Ryan Rossi and Nesreen Ahmed. The network data repository with interactive graph analytics and visualization. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.
- [41] Marta Sales-Pardo, Roger Guimera, André A Moreira, and Luís A Nunes Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39):15224–15229, 2007.
- [42] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.
- [43] Neil Spring, Ratul Mahajan, and David Wetherall. Measuring isp topologies with rocketfuel. *ACM SIGCOMM Computer Communication Review*, 32(4):133–145, 2002.
- [44] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077, 2015.
- [45] Ala Trusina, Sergei Maslov, Petter Minnhagen, and Kim Sneppen. Hierarchy measures in complex networks. *Physical review letters*, 92(17):178702, 2004.
- [46] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [47] Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.
- [48] Xu-Wen Wang, Yize Chen, and Yang-Yu Liu. Link prediction through deep generative model. *iScience*, 23(10):101626, 2020.
- [49] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *nature*, 393(6684):440–442, 1998.
- [50] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016.
- [51] Tao Wu, Yuxiao Guo, Leiting Chen, and Yanbing Liu. Integrated structure investigation in complex networks by label propagation. *Physica A: Statistical Mechanics and its Applications*, 448:68–80, 2016.
- [52] Tao Wu, Hongyu Ma, Chao Wang, Shaojie Qiao, Liang Zhang, and Shui Yu. Heterogeneous representation learning and matching for few-shot relation prediction. *Pattern Recognition*, page 108830, 2022.
- [53] Xingping Xian, Tao Wu, Shaojie Qiao, Xi-Zhao Wang, Wei Wang, and Yanbing Liu. Netsre: Link predictability measuring and regulating. *Knowledge-Based Systems*, 196:105800, 2020.
- [54] Xingping Xian, Tao Wu, Shaojie Qiao, Wei Wang, Chao Wang, Yanbing Liu, and Guangxia Xu. Deepec: Adversarial attacks against graph structure prediction models. *Neurocomputing*, 437:168–185, 2021.
- [55] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*, pages 5453–5462. PMLR, 2018.
- [56] Muhan Zhang and Yixin Chen. Weisfeiler-lehman neural machine for link prediction. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 575–583, 2017.
- [57] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5171–5181, 2018.
- [58] Muhan Zhang, Zhicheng Cui, Shali Jiang, and Yixin Chen. Beyond link prediction: Predicting hyperlinks in adjacency space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [59] Jiajun Zhou, Jie Shen, and Qi Xuan. Data augmentation for graph classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2341–2344, 2020.
- [60] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.
- [61] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, 2009.