IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 June 2022, accepted 25 July 2022, date of publication 1 August 2022, date of current version 4 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195222

RESEARCH ARTICLE

Is Megamerger Better?–Based on the Link Prediction Perspective

YANG WANG[1,2] AND JIFA WANG[1]
[1]Management School, Shenyang University of Technology, Shenyang 110870, China
[2]Architectural and Civil Engineering College, Shenyang University, Shenyang 110044, China
Corresponding author: Jifa Wang (team_jfwang@sut.edu.cn)

ABSTRACT Link prediction provides insight into the evolutionary mechanisms of complex networks by predicting missing edges. Existing research has proposed many similarity algorithms based on local information, and some link prediction algorithms typically perform better in different networks. It is generally believed that a megamerger is beneficial. In the perspective of link prediction, merging the good-performing algorithms brings higher prediction accuracy. And the more times the experiment is executed, the higher the accuracy of link prediction. Therefore, this research proposes a new link prediction algorithm based on the theory of megamerger in management and the concept of partnership, and uses ten actual complex networks for experiments to test the above two hypotheses. The experimental results show that megamerger is not applicable to the link prediction algorithm. In addition, there is no positive correlation between the increasing the quantity of experiments and improving the accuracy of the experiments, so the above two hypotheses are rejected. Hence, this research presumes that megamerger of the comprehensive information of the network, such as the resource flow between nodes, the degree of common neighbor nodes, and partnership of nodes, does not improve the accuracy of link prediction. For a refined network with a small number of nodes and a short average path length, it is recommended that the quantity of experiments be set to only ten can achieve the required accuracy of link prediction.

INDEX TERMS Link prediction, megamerger, complex network.

I. INTRODUCTION

In recent years, with the rapid development of network science, link prediction has been closely related to the structure and evolution of the network [1]. According to different research objects, the nodes of a complex network can be people, knowledge, proteins, or industries. The edges of the network represent the flow of information, technology, material, and resources. Link prediction is a prediction method that uses already existing nodes and edges to predict future links of a complex network. [2]. The higher the accuracy of link prediction, the clearer the guidance for the future development of the network, providing a theoretical and practical basis for the accuracy and science of decision making.

Megamerger [3] and partnership [4] are common theories in management. They are essential in inter-firm cooperation [5]. The introduction of megamerger and partnership in management science to the study of link prediction is one of the innovations of this research. From conceptual fit, relationships between nodes are linked through edges, and the connected edges of different network nodes can all be regarded as a broad partnership, although they represent different meanings. From the perspective of the network structure, two nodes that do not have an edge between them, but have a common partner, are obviously more likely to generate links in the future than nodes that do not have a common partner.

The incomplete or partially observable nature of complex network structures makes theoretical, and engineering studies of real networks face serious challenges [6]. Link prediction research has a wide range of practical applications as well as significant theoretical research implications, especially in promoting and contributing to the theoretical aspects of some related fields. Describing the similarity of nodes in a network

The associate editor coordinating the review of this manuscript and approving it for publication was Yichuan Jiang.

is one of the essential theoretical problems in link prediction [7]. There are many ways to measure similarity. Only by quickly and accurately evaluating whether a certain definition of similarity can describe the relationship between nodes of a given network can we further study the influence of network features on the selection of similarity indicators [8].

Research based on similarity link prediction can promote and reference link prediction itself and the establishment and improvement of the theoretical basis of complex network research. The study of link prediction based on similarity can be divided into two aspects: on the one hand, node similarity. Paths between nodes provide basic similarity features, and link prediction similarity is calculated based on factor information on relative paths [9]. From a statistical point of view, the importance of a node is the probability of attracting other nodes to connect with it, and the derivative value of a node pair is the probability of mutual attraction between nodes [10].

On the other hand, network structure similarity. According to the structure of the network, a similarity score is performed based on its topological characteristics, and the structural characteristics of the network are investigated by the clustering coefficient [11]. Link prediction can predict the undiscovered edges in complex networks to find the central node in the network and thus detect the community structure in complex networks [12]. New community detection metric is proposed based on the principle that internal links are more predictable than external links [13]. Community structures are also effective for link prediction [14]. Link prediction is applied to the management field, and link prediction algorithms for tree-like networks [15] and long-circle-like networks are studied [16].The set of structural features is extracted from the network, and node pairs are represented as vectors. The connection probability of node pairs is measured according to the distance between them and the local mean vectors of positive and negative nearest neighbors [17]. The similarity-based link prediction algorithm extracts the influencing factors of the topology of the target node pair. Still, it ignores the information about the influencing factors between the target node pair and its partners.

In this research, ten real networks are used to test two hypotheses: the first is that megamerger is better based on the link prediction perspective. The second hypothesis is that the more times the experiments, the greater the experimental accuracy. A new link prediction algorithm is constructed based on the theory of megamerger and the principle of partnership to verify whether the accuracy of the new algorithm is higher than the classical link prediction algorithms. The sensitivity of the link prediction algorithm to the quantity of experiments and the proportion of the test set is analyzed, leading to recommendations for the general practicability of link prediction experiments. This research may provide new ideas for applying link prediction algorithms in conjunction with management science. Based on the experimental results, the recommended number of experiments for link prediction experiments is proposed to guide the application of link prediction experiments in practice.

II. LINK PREDICTION ALGORITHM

A. THE BASIC PARAMETERS

1) THE BASIC COMPONENTS OF A COMPLEX NETWORK

An undirected, unweighted complex network G= (V, B), where V represents the set of points and B represents the set of edges. Let G contain n nodes and m edges, thus $V = (v_1, v_2, \cdots v_n)$, $B = (b_1, b_2, \cdots b_m)$. The degree of node $k\,(v_i)$ is defined as the number of other nodes connected to node v_i, $i \in \{1, 2, \ldots, n\}$.

2) MEGAMERGER AND PARTNERSHIP BETWEEN NODES

Megamerger refers to the cooperative relationship between enterprises, industries, or people, which belongs to the concept of management [18]. A megamerger is an acquisition or merger of two existing companies. Once completed, the two companies will likely maintain a significant market share in their respective industries. The concept of megamerger has been introduced in other research areas in recent years [19]. Megamerger is a way that the companies, banks, or other institutions to achieve more significant benefit [20]. Partnership is commonly applied in enterprises [21]. Partnership is a strategic cooperative relationship established on the premise of obtaining common interests and goals [22]. In a complex network, the connection between nodes represents the transfer of energy, information, materials, or knowledge within the network, and the cooperative partnership within the network is established between nodes through the edge. Conversely, network links can be predicted by the common partnership between network nodes. In this research, we study whether there is a megamerger of link prediction algorithms based on the idea of management science? That is, does the new algorithm constructed after the combination of several algorithms with high prediction accuracy still have good accuracy performance? Does its accuracy surpass the classical algorithm before megamerger?

To better describe the relationship of nodes in complex networks, partner relationships between nodes of complex networks are defined in this research. The meaning of cooperative partnership varies in different complex networks. For example, a partnership in industrial networks is a supply relationship established between supply chains and manufacturers. In the sales network, partnership is the buying and selling relationship. In the social network, partnership is follows or establishes friendships on social media. In the knowledge network, partnership is the co-writing of papers between scholars. The context of this research is a macroscopic complex network, which can be an industrial network, sales network, social network, knowledge network, and so forth.

Definition 1: The meaning of megamerger in this research is that the new algorithm obtained by the combination of two

link prediction algorithms with high prediction accuracy has higher accuracy.

Definition 2: Partnership is a personalized business relationship based on mutual trust, benefit, and risk-sharing, and win-win cooperation between enterprises, industries, or people. After establishing partnership, the performance of enterprises, industries, or people has been improved.

Definition 3: The node v_z is connected to both v_x and v_y, and a cooperative relationship exists between nodes, promoting the development of v_x and v_y, which is defined by v_z as a common partner of v_x and v_y, denoted as M_{xy}. The number of partners of v_x and v_y is $S_{xy} = \sum M_{xy}$.

Definition 4: Partnership between nodes in a complex network is defined as the existence of at least one partner between nodes, that is, nodes with common partners constitute a partnership. If at least one partner M_{xy} exists between v_x and v_y, v_x and v_y form a partnership. The set of all partners of node v_x is $\tau\ (v_x)$, and the set of all partners of node v_y is $\tau\left(v_y\right)$. Partnership not only represents the topological relationship between nodes but also describes the internal cooperative relationship between nodes.

3) AVERAGE PATH LENGTH L

The average path length represents the intermediate links required for the association between nodes. The larger L is, the more intermediate links are needed to establish the connection between nodes. The formula for calculating the average path length L of the network is shown in formula (1).

$$L = \frac{2}{n\,(n-1)} \sum\nolimits_{i \neq j} d\left(v_i, v_j\right), \quad i, j \in \{1, 2, \ldots, n\} \tag{1}$$

where n is the total number of nodes in the network, and $d\left(v_i, v_j\right)$ is the number of edges on the shortest path of v_i and v_j. The shortest path between two nodes is determined by the traversal algorithm in Ref. [23].

4) CLUSTERING COEFFICIENT

Suppose node v_i is joined to k_i nodes to form B_i edges, which has at most $k_i\,(k_i-1)/2$ edges. The clustering coefficient refers to the ratio of the actual number of connected edges B_i of node v_i to the sum of the possible number of edges. Its calculation is shown in formula (2).

$$C_i = \frac{2B_i}{k_i\,(k_i-1)} \tag{2}$$

B. LINK PREDICTION ALGORITHM BASED ON MEGAMERGER AND COMMON PARTNERSHIP

1) THE BASIC ASSUMPTIONS

The algorithm design is based on an undirected and powerless complex network. Assuming that any two nodes in the network have at least one partner, the possibility of future connection of two nodes is predicted only based on the points and edges of the complex network, without considering the problem of finite rationality of nodes as persons or the existence of human decisions. The greater the degree of common partners, the smaller the similarity of the two nodes. The more similar the prediction nodes are, the higher their scores.

2) ALGORITHM DESIGN

Undirected and powerless complex network $G{=}\ (V, B)$, the network has *N(N-1)/2* node pairs in total. The purpose of link prediction is to explore the possibility of a link between two nodes that have not yet been connected, which is represented by the score (hereinafter referred to as S). After scoring all pairs of unlinked edges the nodes are ranked in order of their scores, and the two nodes ranked first are considered to have a higher probability of being linked in the future.

Newman [24] firstly studied the role of triadic closure in the field of a complex network. In this research, we refer to Newman's ideas to understand common neighbor nodes from a management perspective. Nodes v_x and v_y are not directly connected, and their common partner is M_{xy}. The smaller $k\left(M_{xy}\right)$ is, the fewer connected edges of M_{xy} are, which means that their partnership is relatively concentrated. Therefore, the partner has the greater linking ability; that is to say, the nodes v_x and v_y connected to M_{xy} are more likely to be linked. For example, in a research collaboration network, a famous scholar M_{xy} in a popular research field publishes many papers, and it is improbable that any of its partners v_x and v_y will collaborate to publish papers. Conversely, it is more likely that $\tilde{v_x}$ and $\tilde{v_y}$, partners of $\tilde{M_{xy}}$ in a cold research field, will collaborate because it is highly likely that $\tilde{v_x}$ and $\tilde{v_y}$ are teachers or classmates of $\tilde{M_{xy}}$. In social networks, for example, it is improbable that fans of famous blogger M_{xy}'s fans v_x and v_y are less likely to follow each other. It is more likely that fans of unfamous blogger $\tilde{M_{xy}}$ will follow each other, as it is very reasonable that they will be introduced to follow $\tilde{M_{xy}}$ by their friends.

The computer scientist David Liben-Nowell [25] cited Newman's paper and proposed the well-known common neighbor index (hereinafter referred to as CN). Zhou *et al.* [26] compared nine well-known local similarity link prediction indexes on six real networks. The results showed that the simplest approach, the common neighbor algorithm, had the best overall performance, followed by the Adamic-Adar index (hereinafter referred to as AA) [27]. Based on the network resource allocation process, a new similarity measure was proposed by them and demonstrated to have higher prediction accuracy than ordinary neighborhoods, which was the RA index. According to the characteristics of partnership, and megamerger theory in management, a new link prediction algorithm based on megamerger is proposed (hereinafter referred to as MA). MA integrates the information on resource flows between nodes, the degree of common neighbors, and partnership between nodes. The future link probability score by MA of nodes v_x and v_y with a common partner M_{xy} is shown in (3).

$$S_{xy}^{MA} = \sum\nolimits_{M_{xy} \in \tau(v_x) \cap \tau(v_y))} \frac{1}{k(M_{xy})} + \frac{1}{log\left(1 + k(M_{xy})\right)} \tag{3}$$

where S_{xy}^{MA} represents the MA similarity score of nodes v_x and v_y that share a common partner M_{xy}; $k(M_{xy})$ is the degree of node M_{xy}, and obviously, $k(v_z) \geq 2$.

III. LINK PREDICTION EXPERIMENTS

A. HYPOTHESIS FORMULATION

The most straightforward design idea of the link prediction algorithm that can be analyzed through existing studies is based on the similarity of two nodes because the properties of two nodes are easily available in the network. However, link prediction only based on node information cannot fully consider the local characteristics of the network, which will lead to the loss of information in the process of link prediction. It is urgent to propose a link prediction algorithm that comprehensively integrates the information on resource flows between nodes, the degree of common neighbors, and partnership between nodes. In the case of link prediction, it is conjectured that megamerger is better. Therefore, a new link prediction algorithm based on megamerger and common partnership is designed to verify the first hypothesis. That is, whether there is a megamerger of link prediction algorithms based on the idea of management science?

In the case of the quantity of experiments, it is conjectured that more times of experiments bring higher accuracy of link prediction. Few scholars have analyzed the impact of the number of link prediction experiments and the proportion of the test set on the accuracy of the experiments, which results in poor generalizability or inefficiency of the experimental conclusions. Sensitivity analysis experiments of the ten link prediction algorithms to the proportion of the test set and quantity of experiments are conducted using ten real network datasets to test the second hypothesis. The purpose of this research is to test these two conjectures. (1) Hypothesis 1: megamerger is better. (2) Hypothesis 2: more times experiments bring higher accuracy.

B. DATA ACQUISITION AND NETWORK CONSTRUCTION

The experimental data in this research come from ten real complex networks. Nematode neural network (C. elegans) [27], in which points are nematode neurons and edges are synapses. Political blogs is a network of hyperlinks to political blogs in the United States [28]. Yeast protein interaction network, in which nodes are proteins and edges are protein interactions [29]. Router network, nodes of the network are routers, and the edge is the data exchange between routers [30]. USAir is an air network in the United States, with airports as nodes and routes as sides [31]. Jazz is a collaborative network of jazz musicians, and a network of partnership between jazz musicians [32]. FWFW is a food chain network of 128 species in the rainy season in Florida Bay, with the sides of the network indicating predation relationships [33]. FWMW is a mangrove estuarine wet season food chain network containing 97 species of organisms [34]. NS is a collaborative network of scientists, where the nodes of the network represent scientists and the edges represent collaborative relationships between scientists [35]. Metabolic is the network of C. elegans' metabolic [36]. Table 1 shows the statistical characteristics of experimental data.

TABLE 1. Network characteristics of experimental data.

Data Set	Number of Nodes	Number of Edges	Average Degree	Average Shortest Path	Aggregation Coefficient
C. elegans	297	2,148	14.47	2.46	0.308
Political blogs	1,222	19,021	27.36	2.74	0.36
Yeast	2,617	11,855	9.85	5.1	0.388
Router	5,022	6,258	2.49	6.45	0.033
USAir	332	2,126	12.81	2.74	0.749
Jazz	198	2,742	27.7	2.235	0.618
FWFW	128	2,075	32.422	1.776	0.335
FWMW	97	1446	29.814	1.693	0.468
NS	379	914	4.82	6.04	0.798
Metabolic	453	2,052	8.94	2.664	0.647

C. EVALUATION METRIC

There are three main ways to verify the accuracy of link prediction algorithms: AUC (Area Under Curve), precision, and ranking score. AUC focuses on the overall performance of the algorithm, precision focuses on whether the nodes in front of the prediction score are predicted accurately, and ranking score focuses on the prediction score ranking. In this research, a more comprehensive AUC index is more suitable to verify the accuracy of the algorithm. The reference [15] gave a detailed procedure for calculating AUC. AUC is the area under the Receiver Operating Characteristic curve (ROC), which can be simply understood as an edge chosen randomly in the test set [37]. Its score value is higher than that of a non-existent edge randomly selected. Its calculation formula is shown in (4). Where n denotes the number of extraction experiments, n' indicates that the fraction of the test set is bigger than the fraction of non-existent edges, and n'' indicates that the fraction of the test set is equal to the fraction of nonexistent edges.

$$AUC = \frac{n' + 0.5n''}{n} \tag{4}$$

D. EXPERIMENT DESIGN

Data set S is randomly divided into training set X and testing set C, test set ratio $\beta = \left|\frac{C}{X+C}\right|$, where $X \cap C = \emptyset, X \cup C = S$. Only the information in the training set can be used to estimate the unknown edges. The experimental steps are shown in Figure 1.

Step1: network input and construction: construct the adjacency matrix based on the network data set. Step2: data set division: randomly divide the test set and train set according to the ratio of the test set. Step3: calculate the AUC of ten link prediction algorithms. Verify hypothesis 1. Step4: sensitivity analysis of test set ratio: the quantity of experiments is unchanged, and step1-3 are repeated with different test set ratios. Step5: quantity of experiments in terms of sensitivity analysis: the proportion of the test set is kept constant, and

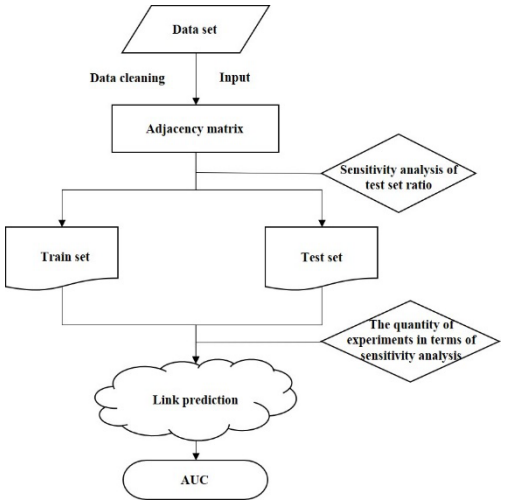

FIGURE 1. Experimental procedure.

TABLE 2. Classical link prediction algorithm.

Classical Link Prediction Algorithm	Calculation Formula
CN [25]	$S_{xy}^{CN} = \lvert\tau(v_x) \cap \tau(v_y)\rvert$
AA [27]	$S_{xy}^{AA} = \frac{1}{log k_z}$
RA [26]	$S_{xy}^{RA} = \frac{1}{k_z}$
Salton [38]	$S_{xy}^{Sa} = \frac{\lvert\tau(v_x) \cap \tau(v_y)\rvert}{\sqrt{k_x k_y}}$
Jaccard [39]	$S_{xy}^{Ja} = \frac{\lvert\tau(v_x) \cap \tau(v_y)\rvert}{\lvert\tau(v_x) \cap \tau(v_y)\rvert}$
Sørensen [40]	$S_{xy}^{So} = \frac{2 \times \lvert\tau(v_x) \cap \tau(v_y)\rvert}{k_x + k_y}$
HPI [41]	$S_{xy}^{HPI} = \frac{\lvert\tau(v_x) \cap \tau(v_y)\rvert}{min\{k_x, k_y\}}$
HDI [42]	$S_{xy}^{HDI} = \frac{\lvert\tau(v_x) \cap \tau(v_y)\rvert}{max\{k_x, k_y\}}$
LHN [43]	$S_{xy}^{LHN} = \frac{\lvert\tau(v_x) \cap \tau(v_y)\rvert}{k_x k_y}$

step1-3 is repeated under the different quantities of experiments. Test hypothesis 2.

E. EXPERIMENTAL RESULTS AND DISCUSSION

1) HYPOTHESIS 1: MEGAMERGER IS BETTER

The classical link prediction algorithm formulas are shown in Table 2.

The higher the AUC value is, the better the prediction effect is. Set quantity of experiments N=100 and the test set ratio β=0.1 to carry out link prediction experiments on ten real networks, and the AUC values of prediction accuracy obtained are shown in Table 3.

Among the ten networks, MA has the highest AUC in 3 networks, which are Router, FWMW, and Metabolic. The AUC value of MA is only 0.031% higher than AA in Router. The accuracy of the MA improved the most in the FWMW network, by 0.042% over RA, and megamerger is not reflected in the link prediction algorithm. The lowest MA accuracy improvement in the Metabolic network, with an improvement of only 0.021% compared to RA. All ten link prediction algorithms have the highest prediction accuracy in the NS network, with an average AUC of 99.24%, as the NS network has the smallest number of edges and the largest aggregation coefficient. The AUC of ten link prediction algorithms in the Router network is generally low. Router contains 5,022 nodes but only 6,258 links. The average degree of Router is the smallest among the ten networks at only 2.49, and the average shortest path is 6.45. The Router network has 12,607,731 node pairs, but its edge connection rate is only 0.05%.

TABLE 3. AUC values of link prediction algorithm for ten real networks.

	Salton	Jaccard	Sørensen	HPI	HDI	LHN	CN	AA	RA	MA
C. elegans	0.7987	0.7922	0.7917	0.8054	0.7814	0.7249	0.8481	0.8657	**0.8703**	0.8697
Political blogs	0.8795	0.8777	0.8769	0.8558	0.8738	0.7632	0.9238	0.9275	**0.9287**	0.9286
Router	0.6503	0.6508	0.6503	0.6506	0.6508	0.6509	**0.6514**	**0.6514**	0.6512	**0.6516**
Yeast	0.9149	0.9153	0.9152	0.9141	0.9152	0.9111	0.9163	0.9168	**0.9177**	0.9168
USAir	0.9263	0.9152	0.9154	0.8832	0.9091	0.7774	0.9549	0.9665	**0.9725**	0.9552
Jazz	0.9667	0.9620	0.9623	0.9482	0.9531	0.9028	0.9565	0.9636	**0.9724**	0.9717
FWFW	0.5291	0.5278	0.5278	0.5307	0.5234	0.4003	0.6096	0.6112	**0.6112**	0.6144
FWMW	0.6189	0.6254	0.6252	0.5692	0.6368	0.4090	0.7066	0.7108	0.7133	**0.7136**
NS	0.9925	0.9925	0.9923	0.9924	0.9924	0.9920	0.9924	**0.9927**	0.9926	0.9926
Metabolic	0.8145	0.7782	0.7775	0.9148	0.7653	0.7412	0.9241	0.9563	0.9609	**0.9611**

Focus on the analysis of CN, AA, RA, and MA that perform well in Table 3, and reserve only two decimal values of AUC, as shown in Table 4. Router, with the highest number of nodes, and NS, with the lowest number of nodes, are the two datasets with the lowest AUC values for MA. The AUC of MA are 0.65 and 0.71, respectively, indicating that MA accuracy is low in networks with a large and small number of nodes. MA has no obvious advantages, and the principle of megamerger cannot be realized in link prediction algorithm; that is to say, megamerger is not always better. In the other eight networks, although the AUC values of the four algorithms are different, there is no significant advantage or disadvantage. Therefore, the first question proposed in this research can be answered, that is, the megamerger is not stronger. In link prediction, the megamerger of several algorithms with better prediction accuracy to form a new link prediction algorithm cannot surpass the prediction accuracy of the traditional algorithm. Hence, the first hypothesis is false.

2) HYPOTHESIS 2: MORE TIMES EXPERIMENTS BRING HIGHER ACCURACY

To verify the second hypothesis, sensitivity analysis experiments of ten link prediction algorithms were carried out on test set ratio β and quantity of experiments N.

TABLE 4. AUC values of for link prediction algorithm that perform well.

Data set Name	CN	AA	RA	MA
C. elegans	0.85	**0.87**	**0.87**	**0.87**
Political blogs	0.92	**0.93**	**0.93**	**0.93**
Router	**0.65**	**0.65**	**0.65**	**0.65**
Yeast	**0.92**	**0.92**	**0.92**	**0.92**
USAir	0.95	**0.97**	**0.97**	0.96
Jazz	0.96	0.96	**0.97**	0.97
FWFW	0.61	0.61	**0.62**	0.61
FWMW	0.71	0.71	0.71	**0.71**
NS	0.99	**0.99**	0.99	0.99
Metabolic	0.92	0.96	0.96	**0.96**

TABLE 5. Applicability of data sets to different proportions of test sets.

β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
C. elegans	√	√	√	√	√	√	√	√	×
Political blogs	√	√	√	√	√	√	√	√	√
Router	√	×	×	×	×	×	×	×	×
Yeast	√	√	√	√	√	√	√	×	×
USAir	√	√	√	√	√	√	√	√	×
Jazz	√	√	√	√	√	√	√	√	√
FWFW	√	√	√	√	√	√	√	√	√
FWMW	√	√	√	√	√	√	√	√	√
NS	√	√	√	√	√	×	×	×	×
Metabolic	√	√	√	√	√	√	√	×	×

a: SENSITIVITY ANALYSIS OF TEST SET RATIO

Quantity of experiments N = 20, test set ratio β = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and the link prediction experiment was carried out on ten data sets respectively. It was found that not all data sets were suitable for each β value. The applicability of the data sets is shown in Table 5. Where √ means the dataset is applicable and × indicates that the data set is unsuitable. The reason why the data set is not applicable is that the continuity of the training set cannot be guaranteed under those test set ratios.

According to Table 5, all ten datasets can meet the applicability at β =0.1. The Political blogs dataset is applicable at β ={0.1,0.2, ⋯, 0.9} because its number of edges is the largest among the ten networks so that it can meet the proportional allocation of various test sets. A combined analysis of Table 5 and Table 1 shows that Political blogs, Jazz, FWFW, and FWMW rank in the top four of the ten datasets in terms of average degree, much larger than the other datasets. Political blogs, Jazz, FWFW, and FWMW are applicable to all nine values of β. The Router is applicable only at β =0.1. The Router network has the largest node number, the biggest average shortest path length, the smallest average degree, and aggregation coefficient, and the highest sensitivity to β. Correspondingly, the FWMW network has the least number of nodes and the smallest average shortest path length, which is effective at β = {0.1,0.2, ⋯, 0.9}. C. elegans is not applicable only when β =0.9 and can be predicted under other β values, indicating its low sensitivity to β. When β =0.9, only Political blogs, Jazz, FWFW, and FWMW are applicable, so setting the test set ratio to 0.9 in link prediction experiments is not recommended. The sensitivity of ten link prediction algorithms' accuracy to β in ten real network datasets is shown in Figure 2.

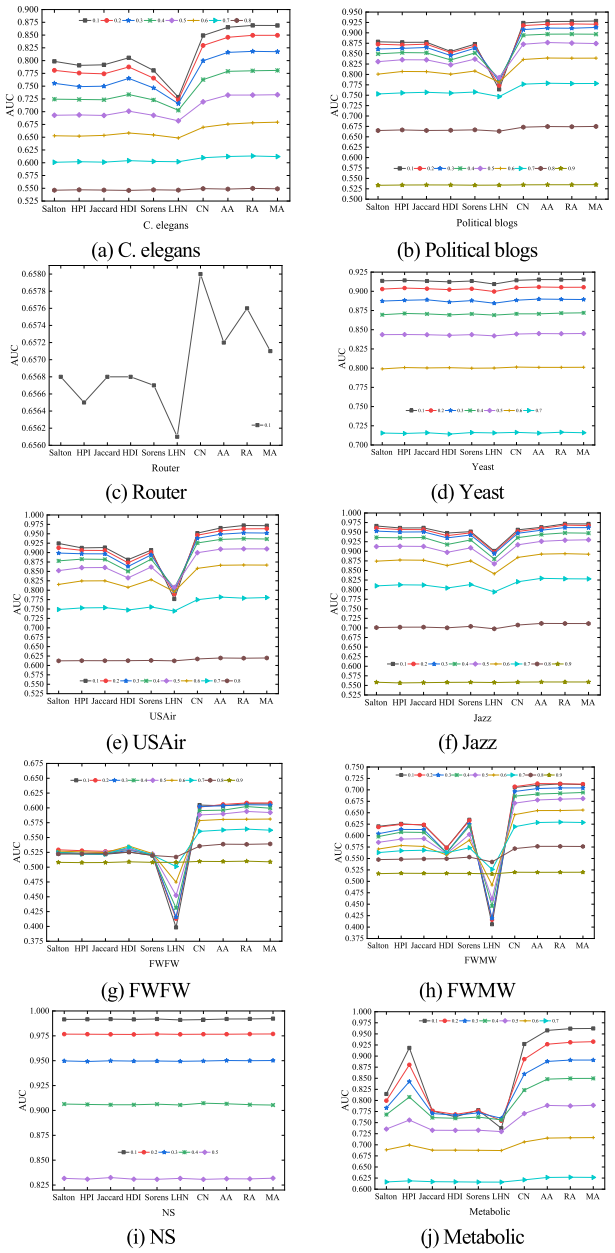

FIGURE 2. Sensitivity of ten link prediction algorithms' accuracy to β in ten real network datasets.

As shown from Figure (a), when $\beta \leq 0.4$, the AUC value of the C. elegans network is large, and the prediction accuracy is good, and in this range, the LHN algorithm performs significantly lower. When $\beta > 0.4$, the AUC tends to decline significantly, with the maximum decreasing rate as high as 37%., and the accuracy of each algorithm in this range does not vary to a great extent. Figure (b) shows that the AUC of the Political blogs network is generally good by all link prediction algorithms when $\beta \leq 0.6$, but the AUC of the LHN algorithm is still the lowest. When $\beta > 0.6$, the maximum accuracy drop is 42%. Figure (d) shows that the accuracies of different link prediction algorithms in the Yeast network are not significantly different, but the accuracy of MA when $\beta = 0.1$ is 22% higher than that of $\beta = 0.7$. According to the analysis of Figure (e), (f), (g), (h), and (j), the accuracy of the LHN algorithm is the lowest in all these ten datasets, and the AUC distribution trend of the ten link prediction algorithms does not change significantly with the change of β. In the Jazz dataset, the AUC of MA is 74% higher when $\beta = 0.1$ than when $\beta = 0.9$. As shown in Fig. (i), there is no significant fluctuation in the prediction accuracy of the ten link prediction algorithms for the NS network. When $\beta = 0.1$, the accuracy of all ten link prediction algorithms is maximum, the AUC values all reach above 99%, and the prediction accuracy of all ten link prediction algorithms for NS is very high. NS networks have the lowest number of edges and the largest aggregation coefficient, and the experimental results show that the link prediction algorithms have higher prediction accuracy on such characteristic networks. The combined analysis shows that the ten link prediction algorithms with $\beta = 0.1$ have the highest prediction accuracy for ten networks.

b: QUANTITY OF EXPERIMENTS IN TERMS OF SENSITIVITY ANALYSIS

According to the above analysis, when $\beta = 0.1$, all the ten data sets can meet the applicability, and the accuracy of each algorithm is the highest. Therefore, the test sets ratio was set to 0.1 during the sensitivity analysis of the quantity of experiments. Set experiment times $N = 10, 20, 30, 40, 50, 100, 200, 500, 1{,}000, 2{,}000, 10{,}000$ and carry out link prediction experiments on ten data sets respectively. The sensitivity of each network to N is shown in Figure 3.

According to the Figures (a), (b), (e), (f), (g), (h), (j) in Fig. 3, C. elegans, Political blogs, USAir, Jazz, FWFW, FWMW, and Metabolic networks are not sensitive to the change of N. The accuracy of each algorithm does not fluctuate with the change of the quantity of experiments. LHN algorithm generally performs poorly in ten data sets. The AUC of MA decreases clearly when the USAir network conducts 100 times experiments but performs well when N is set to other values. For $N = 10$, the three networks C. elegans, Jazz, and FWMW have the highest accuracy of the ten link prediction algorithms. C. elegans dataset has the highest accuracy at ten times, and the quantity of experiments should not exceed 50. When the quantity of experiments is greater than ten, AUC is relatively balanced. The AUC

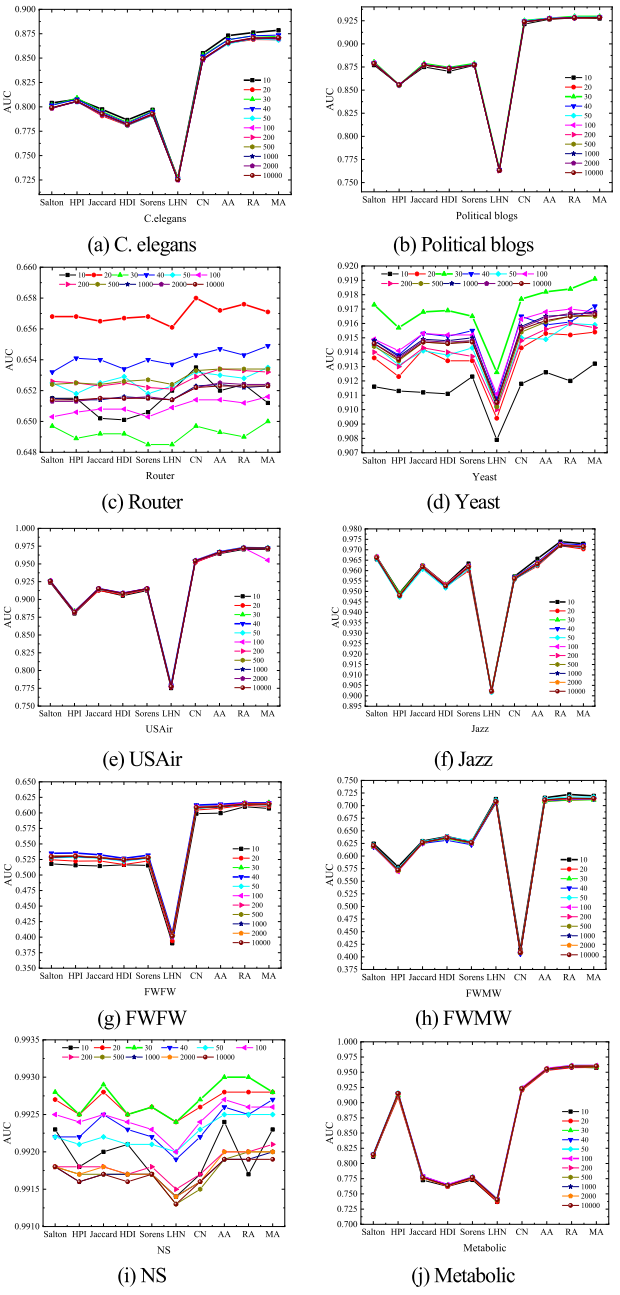

FIGURE 3. Sensitivity of ten link prediction algorithms' accuracy to N in ten real network datasets.

value reaches its peak when the quantity of experiments of the Political blogs data set is about 30. When the quantity of experiments is greater than 30, the prediction accuracy decreases obviously, and the quantity of experiments should not exceed 30. When conducting link prediction experiments on data like Political blogs database with a moderate number of nodes, a large number of sides, and large average degree, the quantity of experiments should be set to 30, and too many or too few experiments should not achieve good prediction results.

Figures (c), (d), and (i) appear that Router, Yeast, and NS networks have high sensitivity to N, but the sensitivity is only

about 1%. For the Router network, the AUC performs best when $N = 20$, whereas the AUC values of all algorithms are generally low when $N = 30$. As to the Yeast, the AUC shows the best performance when $N = 30$, whereas AUC values of all algorithms are generally low when $N = 10$. When $N = 500$, 1,000, 2,000, 10,000, the lines in Figure (i) almost overlap and the values are extremely similar. When $N = 10{,}000$, the computation time increases, but the algorithms have the lowest accuracy, indicating that for NS, networks $N = 30$ is sufficient to meet the experimental accuracy requirements.

A comprehensive analysis of Figure 3 shows no significant linear enhancement relationship between the quantity of experiments and the accuracy of link prediction algorithms when ten link prediction algorithms have experimented on ten real data sets. The prediction accuracy of the ten link prediction algorithms is highest when times of experiments of Political blogs with the highest number of edges are taken as 30 in the ten real network datasets. For the Router network with the largest number of nodes, the maximum accuracy of link prediction can be achieved by taking 20 experiments, which improves the efficiency of link prediction experiments. There is no need to increase the number of experiments to 100 or even 10,000, which vastly reduces the computational speed. The number of experiments for the FWMW network with the smallest number of nodes is taken as 10, and the number of experiments for the NS network with the smallest number of edges is taken as 30. Based on the experimental results, it is clear that more times experiments do not lead to higher link prediction accuracy for either large or small networks. Therefore, hypothesis 2 is rejected.

F. SUGGESTIONS OF LINK PREDICTION EXPERIMENTS

In previous studies, the quantity of link prediction experiments is usually taken to be 100 times, or larger [15], [44], which is that the scholars are guided by the inherent thought hypothesis 2. This research verifies that hypothesis 2 is wrong through experiments. According to the analysis in the previous section, the relationship between the characteristics of the data set, the test set ratio, and the quantity of experiments are summarized. Thus, the link prediction experiment suggestions are given as shown in Table 6.

TABLE 6. Suggestions for link prediction experiments.

Data set Name	Data Set Features	The Quantity of Experiments	Test Sets Ratio
C. elegans	Small network, the number of nodes is small, the average shortest path is short.	10	0.1
Political blogs	Medium network, the number of nodes is moderate, the number of edges is large, the average degree is large.	30	0.1
Router	Large network, more nodes, larger average path length, average degree and aggregation coefficient.	20	0.1
Yeast	Medium and large network.	30	0.1
USAir	Small network, with smaller average path length. The aggregation coefficient is large and has the characteristics of small world network.	40	0.1
Jazz	Small networks, with the large average degree and aggregation coefficient.	10	0.1
FWFW	Small networks, with the largest average degree among ten data sets.	40	0.1
FWMW	Small networks, with the minimum number of nodes and average shortest path in ten data sets.	10	0.1
NS	Small networks, with the minimum number of edges and the maximum aggregation coefficient in ten data sets.	30	0.1
Metabolic	Small networks, with the second largest aggregation coefficient among ten data sets.	50	0.1

According to the experimental results, the suggested the quantity of experiments for different types of networks is given, which is helpful to improve the experimental efficiency of link prediction, especially for huge networks with a large number of nodes and edges. The conclusions of this study can distinctly reduce the experimental time and have some recommended significance for guiding the experimental design of link prediction.

1) For refined networks with fewer nodes and shorter average path length, good prediction results can be achieved without many experiments, so it is suggested to set the experiment number to about ten times. 2) For a large network with many nodes and a large average path length, the larger quantity of experiments, the longer the experiment time will inevitably be. However, it is found through experiments that the quantity of experiments has no significant impact on the prediction accuracy. Therefore, it is suggested to set the quantity of experiments as 20-40 times to meet the accuracy requirement and effectively improve the experimental efficiency. 3) For networks with not a large number of nodes, the quantity of experiments can be set to about 40-50 times because such networks will not occupy a long experiment time even if they have a large quantity of experiments, and the accuracy is also good within 40-50 times of experiments. 4) When doing the link prediction experiment, the test set ratio is recommended to be 0.1.

IV. CONCLUSION AND DISCUSSION

The purpose of this research is to test two hypotheses: first, megamerger is better. That is to say, link prediction algorithms with high prediction accuracy combine to produce a new link prediction algorithm with higher accuracy. The other hypothesis is that the higher quantity of experiments, the higher the prediction accuracy. This research is a new attempt to apply management theory to complex network link prediction by experimentally verifying that people's inherent thinking is not always true. The recommended number of experiments for link prediction experiments is given according to the experimental results, which may be beneficial for scholars to conduct link prediction experiments scientifically and efficiently.

The experimental results showed that megamerger was inappropriate for the link prediction algorithms, and MA has no significant advantage over the classical link prediction algorithm. Although MA integrated the information on resource flows between nodes, the degree of common neighbors, and partnership between nodes, the AUC of MA was always basically equal to the original link prediction algorithm. The first hypothesis was rejected. Therefore, we infer that the more information about the network used in the link prediction algorithm doesn't deserve more accuracy.

Promising experimental results were achieved with a smaller quantity of experiments. Using ten real network datasets, a sensitivity analysis of the quantity of experiments done on the ten link prediction algorithms found that the AUC obtained from 10,000 times of experiments was not the highest but rather the highest link prediction accuracy of each algorithm when the quantity of experiments was set to 10-40, so the second hypothesis was rejected. This result is quite different from the usual practice of setting the quantity of experiments to 100 in previous studies. Hence, we suggest that ten times of experiments can meet the accuracy requirement for a refined network with a small number of nodes and a short average path length. For huge networks with a larger number of nodes and a larger average path length, 20-40 times of experiments are sufficient to obtain good link prediction accuracy, significantly reducing the prediction time and thus improving prediction efficiency.

In addition, only link prediction algorithms based on local similarity are investigated in this study, and megamerger of other types of link prediction algorithms will be further explored in future research. Experiments related to other network structures with special structures such as tree-like networks are also the subject of subsequent research.

ACKNOWLEDGMENT

The authors would like to thank the editors and the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] X. Chen, B. Kang, J. Lijffijt, and T. D. De Bie, "ALPINE: Active link prediction using network embedding," *Appl. Sci.*, vol. 11, no. 11, p. 5043, May 2021, doi: 10.3390/app11115043.

[2] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Statist. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011, doi: 10.1016/j.physa.2010.11.027.

[3] S. Dzuba and D. Krylov, "Cluster analysis of financial strategies of companies," *Mathematics*, vol. 9, no. 24, p. 3192, Dec. 2021, doi: 10.3390/math9243192.

[4] J. W. Wiśniewski, E. Sokołowska, J. Wu, and A. Dziadkiewicz, "Evolutionary game analysis of the partners' behavior in the rural E-payment market of China," *Risks*, vol. 9, no. 12, p. 220, Dec. 2021, doi: 10.3390/risks9120220.

[5] S. Rama Murthy, K. Roll, and A. Colin-Jones, "Ending business-non-profit partnerships: The spinout of social enterprises," *Scandin. J. Manage.*, vol. 37, no. 1, Mar. 2021, Art. no. 101136, doi: 10.1016/j.scaman.2020.101136.

[6] M. Coşkun and M. Koyutürk, "Node similarity-based graph convolution for link prediction in biological networks," *Bioinformatics*, vol. 37, no. 23, pp. 4501–4508, Dec. 2021, doi: 10.1093/bioinformatics/btab464.

[7] Y. Pan, D. Li, J.-G. Liu, and J.-Z. Liang, "Detecting community structure in complex networks via node similarity," *Phys. A, Stat. Mech. Appl.*, vol. 389, no. 14, pp. 2849–2857, 2010, doi: 10.1016/j.physa.2010.03.006.

[8] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 4, Oct. 2009, Art. no. 046122, doi: 10.1103/PhysRevE.80.046122.

[9] S. Li, J. Huang, J. Liu, T. Huang, and H. Chen, "Relative-path-based algorithm for link prediction on complex networks using a basic similarity factor," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 30, no. 1, Jan. 2020, Art. no. 013104, doi: 10.1063/1.5094448.

[10] X. Wu, J. Wu, Y. Li, and Q. Zhang, "Link prediction of time-evolving network based on node ranking," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105740, doi: 10.1016/j.knosys.2020.105740.

[11] S. Rafiee, C. Salavati, and A. Abdollahpouri, "CNDP: Link prediction based on common neighbors degree penalization," *Phys. A, Stat. Mech. Appl.*, vol. 539, Feb. 2020, Art. no. 122950, doi: 10.1016/j.physa.2019.122950.

[12] H. Jiang, Z. Liu, C. Liu, Y. Su, and X. Zhang, "Community detection in complex networks with an ambiguous structure using central node based link prediction," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105626, doi: 10.1016/j.knosys.2020.105626.

[13] K.-K. Shang, M. Small, Y. Wang, D. Yin, and S. Li, "A novel metric for community detection," *Europhys. Lett.*, vol. 129, no. 6, p. 68002, Apr. 2020, doi: 10.1209/0295-5075/129/68002.

[14] X.-K. Xu, K.-K. Shang, and J. Xiao, "Quantifying the effect of community structures for link prediction by constructing null models," *IEEE Access*, vol. 8, pp. 89269–89280, 2020, doi: 10.1109/ACCESS.2020.2993455.

[15] K. Shang, T. Li, M. Small, D. Burton, and Y. Wang, "Link prediction for tree-like networks," *Chaos*, vol. 29, no. 6, Jul. 2019, Art. no. 061103, doi: 10.1063/1.5107440.

[16] K.-K. Shang and M. Small, "Link prediction for long-circle-like networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 105, no. 2, p. 24311, Feb. 2022, doi: 10.1103/PhysRevE.105.024311.

[17] L. Li, H. Wang, S. Fang, N. Shan, and X. Chen, "A supervised similarity measure for link prediction based on KNN," *Int. J. Mod. Phys. C*, vol. 32, no. 9, Sep. 2021, Art. no. 2150112, doi: 10.1142/S0129183121501126.

[18] H. Fraisse, J. Hombert, and M. Lé, "The competitive effect of a bank megamerger on credit supply," *J. Banking Finance*, vol. 93, pp. 151–161, Aug. 2018, doi: 10.1016/j.jbankfin.2018.06.011.

[19] D. Huang, Z. Li, G. Zeng, C. Zhou, W. Xue, X. Gong, X. Yan, S. Chen, W. Wang, and M. Cheng, "Megamerger in photocatalytic field: 2D g-C_3N_4 nanosheets serve as support of 0D nanomaterials for improving photocatalytic performance," *Appl. Catal. B, Environ.*, vol. 240, pp. 153–173, Jan. 2019, doi: 10.1016/j.apcatb.2018.08.071.

[20] E. Devos, S. Krishnamurthy, and R. Narayanan, "Efficiency and market power gains in bank megamergers: Evidence from value line forecasts," *Financial Manage.*, vol. 45, no. 4, pp. 1011–1039, Dec. 2016, doi: 10.1111/fima.12134.

[21] T. W. Guinnane and J. Schneebacher, "Enterprise form: Theory and history," *Explor. Econ. Hist.*, vol. 76, Apr. 2020, Art. no. 101331, doi: 10.1016/j.eeh.2020.101331.

[22] V. L. Herzog, "Trust building on corporate collaborative project teams," *Project Manage. J.*, vol. 32, no. 1, 2001, Art. no. 1, doi: 10.1177/875697280103200105.

[23] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, Dec. 2001, doi: 10.1080/0022250X.2001.9990249.
[24] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, Jul. 2001, Art. no. 025102, doi: 10.1103/PhysRevE.64.025102.
[25] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.
[26] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, Oct. 2009, doi: 10.1140/epjb/e2009-00335-8.
[27] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Soc. Netw.*, vol. 25, no. 3, pp. 211–230, 2003, doi: 10.1016/S0378-8733(03)00009-1.
[28] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in *Proc. 3rd Int. Workshop Link Discovery (LinkKDD)*, 2005, pp. 36–43.
[29] C. Von Mering, R. Krause, and B. Snel, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, pp. 399–403, May 2002, doi: 10.1038/nature750.
[30] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, "Measuring ISP topologies with RocketFuel," *IEEE/ACM Trans. Netw.*, vol. 12, no. 1, pp. 2–16, Feb. 2004, doi: 10.1109/TNET.2003.822655.
[31] B. Vladimir and M. Andrej. (2006). *Pajek Datasets*. [Online]. Available: http://vlado.fmf.uni-lj.si/pub/networks/data/
[32] P. M. Gleiser and L. Danon, "Community structure in jazz," *Adv. Complex Syst.*, vol. 6, no. 4, pp. 565–573, Dec. 2003, doi: 10.1142/S0219525903001067.
[33] R. E. Ulanowicz, C. Bondavalli, and M. S. Egnotovich, "Network analysis of trophic dynamics in South Florida ecosystem, FY 97: The Florida bay ecosystem," Chesapeake Biol. Lab., Solomons, MD, USA, Tech. Rep., 1988. Accessed: Jun. 13, 2004. [Online]. Available: http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/foodweb.htm
[34] D. Baird, J. Luczkovich, and R. R. Christian, "Assessment of spatial and temporal variability in ecosystem attributes of the St Marks National Wildlife Refuge, Apalachee Bay, Florida," *Estuarine, Coastal Shelf Sci.*, vol. 47, no. 3, pp. 329–349, Sep. 1998, doi: 10.1006/ecss.1998.0360.
[35] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 3, Sep. 2006, Art. no. 036104, doi: 10.1103/PhysRevE.74.036104.
[36] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 72, no. 2, Aug. 2005, Art. no. 027104, doi: 10.1103/PhysRevE.72.027104.
[37] K.-K. Shang, M. Small, and W.-S. Yan, "Fitness networks for real world systems via modified preferential attachment," *Phys. A, Stat. Mech. Appl.*, vol. 474, pp. 49–60, May 2017, doi: 10.1016/j.physa.2017.01.066.
[38] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1986.
[39] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et du Jura," *Bull. del la Société Vaudoise des Sci. Naturelles*, vol. 37, no. 142, pp. 547–579, 1901, doi: 10.5169/seals-266450.
[40] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons," *Biol. Skrifter*, vol. 5, no. 4, pp. 1–34, 1948.
[41] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002, doi: 10.1126/science.1073374.
[42] L. Lü and T. Zhou, "Link prediction in weighted networks: The role of weak ties," *Europhys. Lett.*, vol. 89, p. 18001, Jan. 2010, doi: 10.1209/0295-5075/89/18001.
[43] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 73, no. 2, Feb. 2006, Art. no. 026120, doi: 10.1103/PhysRevE.73.026120.
[44] F. Aziz, H. Gul, I. Muhammad, and I. Uddin, "Link prediction using node information on local paths," *Phys. A, Stat. Mech. Appl.*, vol. 557, Nov. 2020, Art. no. 124980, doi: 10.1016/j.physa.2020.124980.

YANG WANG received the master's degree in management from the Harbin Institute of Technology, Harbin, China. She is currently pursuing the Ph.D. degree with the Management School, Shenyang University of Technology. She is also a Lecturer with the Architectural and Civil Engineering College, Shenyang University. Her research interests include link prediction and topology identification of complex networks.

JIFA WANG received the Ph.D. degree in management from the Harbin Institute of Technology, Harbin, China. He is currently a Professor with the Management School, Shenyang University of Technology. His research interests include complex system analysis and intelligent decisions, science and technology finance, and the development of small and medium-sized enterprises.