Research of Motif-Based Similarity for Link Prediction Problem

CHAO LI[1,2], WEI WEI[3,4,5,6], XIANGNAN FENG[3,4,7], AND JIAOMIN LIU[1]
[1]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
[2]Department of Mathematics and Computer Science, Hengshui University, Hengshui 053000, China
[3]Key Laboratory of Mathematics, Informatics and Behavioral Semantics (LMIB), Beihang University, Beijing 100191, China
[4]School of Mathematics and Systems Science, Beihang University, Beijing 100191, China
[5]Peng Cheng Laboratory, Shenzhen 518066, China
[6]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China
[7]Max Planck Institute for Human Development, 14195 Berlin, Germany

Corresponding author: Wei Wei (weiw@buaa.edu.cn)

ABSTRACT The link prediction problem in the network concerns to predict the existence of links between node pairs, which is a research hotspot in different scenarios with network applications. Methods of predicting links based on network topology and structures provide a number of measurements to reveal the underlying relationship between two nodes. In this paper, a motif-based similarity index for link prediction is proposed to calculate the similarity score of two nodes concerning their local environment, which takes advantage of existing similarity definitions and the motifs. This motif-based similarity can be generalized to more complicated cases by considering different motifs and keeps the same level of computational complexity with the existing indexes. Experimental results on 9 public benchmark datasets and 1 randomly generated dataset show the effectiveness of our proposed index, and accuracies on several datasets are significantly improved. The performance of motif-based similarity suggests that considering typical motifs on networks could improve the precisions of link prediction tasks, and exploring specific structure characteristics on networks will point out an important and effective direction for more research with network methods applied.

INDEX TERMS Link prediction, similarity score, motif, high-order structure.

I. INTRODUCTION

Many social, biological and other systems could be described by complex networks. The nodes in the networks represent the individuals in the systems, and the edges represent the relationships between individuals. In recent years, complex networks have provided a number of effective methods to describe and analyze complex systems and are widely researched. Various behaviors on network are described and predicted on the basis of known network structural characteristics, and one of the most important problems is the link prediction problem, which has attracted more and more attention recently. The link prediction problem in the network refers to how to predict the possible existence of a link between two unconnected nodes through known information such as existing network nodes and structures [1]. The link

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

prediction problem includes both the prediction of unknown links and the prediction of future links. This problem has important meaning and significant values in both aspects of methodology and application.

Recently link prediction methods based on the network structure have received a lot of attention. Compared with the attribute information of nodes which requires extra data, the network structure information is easier to obtain and requires less extra effort to gather and collect. At the same time, this type of method has universal applicability to networks with similar structures, thereby avoiding the huge needs for computing cost in obtaining specific parameter combinations for different networks. Based on the characteristics of network structure, many similarity-defined methods based on network topological features have emerged, and the abilities of these indexes in predicting links in various scenarios like social cooperation networks have been analyzed [2]. Another type of link prediction method is based

on maximum likelihood estimation of the network structure. This method uses the hierarchical structure of the network for link prediction, and performs well on networks with clear hierarchical structure [3]. Existing similarity definitions such as *Common Neighbor* [4], *Jaccard* [5], *Adamic-Adar* [6], *Preferential Attachment* [7] and *Cannistrai-Alanis-Ravai* [8] take advantage of the local connection environments to measure the correlations of two nodes, and another index *Clustering Coefficient for Link Prediction* [9] uses the clustering coefficient to fulfill the link predictions. As a widely used graph learning method, *Node2vec* (n2v) provides another way to understand the structure and correlation among nodes in a global viewpoint [10]. These methods embody the local and global connection types, and have achieved good performance on the prediction tasks of some network datasets.

As an important network structure concept, motif is well studied in many scenarios of network science [11]. Motif can be viewed as basic/frequent building blocks of networks and can be treated as the molecules composing the complex systems. Generally, typical motifs could work as elemental components and are able to determine some global characteristics of large networks. Attentions on specific motifs and organizations promote deeper research and understanding of network-related tasks. Particularly, the classical clustering efficient is also a calculation related to 3-motifs (aiming at the triangle structure).

In this paper, inspired by the existing similarity measurements and the motif structures, a new similarity definition based on motifs will be proposed to calculate the correlations of target node pairs, and it is a more comprehensive index that can have more compatible applications by considering different motif types. Experiments on some real network data and randomly generalized one will be conducted to show the performance of our proposed similarity index. Our contributions are three-fold:

- The relationship of the existing similarities for link prediction with motifs is discussed, which provides a new angle to view the link prediction tasks by considering network structures.
- A motif-based similarity definition for link prediction is proposed to measure the correlations of targeted node pairs, and this definition can be generalized to more complicated cases by considering different motifs.
- Experimental results on public benchmark datasets and randomly generated dataset validate the effectiveness of our motif-based similarity, and the accuracy on some datasets shows significant improvement.

The remainder of this paper is organized as follows. Section II summarizes the related work on link prediction problems and motif studies. Section III introduces the backgrounds, definitions and some properties of our proposed similarity score. Section IV describes the experiments, a comparison with existing indexes, and analysis of the performance of our index. Finally, we conclude the paper and propose some future work in Section V.

II. RELATED WORK

This paper will focus on the link prediction problem and build a similarity measurement based on the network motif structure, and the related work of these two aspects will be reviewed in this section.

A. LINK PREDICTION RELATED

As a research area with extremely high theoretical and empirical values, link prediction has long attracted the attention of academia and industry. The ideas and methods are mainly based on Markov chain and machine learning. Sarukkai applied Markov chain to network link prediction and path analysis [12]. Popescul and Ungar proposed a regression model to predict the citation relationship of scientific literatures in the literature citation network [13]. Lin defined the similarity between nodes based on nodes attributes, which could be directly used for link prediction [14]. Although the application of external information such as node attributes can indeed obtain good prediction results, it is very difficult or even impossible to obtain this information in many cases. Furthermore, even if accurate information about node attributes could be obtained, identifying useful and useless information for network link prediction and which information is still a problem.

Among the many proposed methods, similarity-based approaches appear to be particularly vibrant, with new ideas and indicators keeping emerging and taking into account various features. Inspired by cosine similarity, G. Salton and M. McGill defined Salton Index to illustrate the link probability [15]. Ravasz *et al.* introduced Hub Promoted Index and Hub Depressed Index to acquire the promotion in link formation for two situations: links between hubs and high-degree nodes and links between hubs and low-degree nodes respectively [16]. Leicht *et al.* proposed Local Leicht-Holme-Newman index to define the similarity between a pair of nodes related to the amount of neighbors they share [17]. Zhou *et al.* defined Resources Allocation in 2009 based on the resource allocation process [18]. Most of these existing methods are based on the local information of graphs, which is relatively easier to collect but may not be effective in some cases, due to their limited representation abilities on graphs' local characteristics. Therefore, it is supposed to improve the accuracy of link prediction greatly if having a more throughout understanding of the graph.

Deep learning related work also acts as an important role in the link prediction problem, which always uses more global information of the targeted network data. The deep learning model E-LSTM-D for predicting end-to-end dynamic links can not only deal with long-term prediction problems, but also be suitable for networks of different scales [19]. The weighted link prediction model using learning automata strategy predicts the occurrence of each link according to the weight information of the current network [20]. Multi-level graph neural network based on the original method uses the multi-layer perceptron strategy to model the hierarchical structure and introduces the attention mechanism, which

can accurately predict on the noisy knowledge graph [21]. A series of stacking models are constructed by using the diversity, and the prediction factors are combined into a single algorithm and applied to the real world integrated network, which can achieve almost optimal prediction [22]. With the development of deep learning, future research should focus on learning appropriate heuristics for a given network, rather than using predefined heuristics [23]. The characteristics of interpretability, simplicity and scalability make link prediction widely used in practice. Node2vec is an algorithm for learning the continuous feature representation of nodes in a network, that is, learning the mapping from a node to a low dimensional feature space, and keeping the network neighborhood of nodes to the maximum extent [10], [24].

B. MOTIF STUDY RELATED
Motifs are simple building blocks of complex networks, which were first proposed by Milo *et al.* [11]. More formally, motifs are defined as patterns of interconnections, whose frequencies of occurring in real-world complex networks are significantly higher than those in random networks with same scales. Later Milo *et al.* have found crucial motifs structure in networks from biochemistry [25], [26], neurobiology [27], ecology [28], and engineering [7] fields.

One of the most important aspect of motifs is that they provide a new solution for defining universal classes of networks from the perspective of topological structure levels [11]. For example, motifs in the electronic sequential logic circuit networks analyzed from ISCAS89 benchmark set [29] are found to be exactly the same as those in the synaptic connection network among neurons in C. Elegans [27]. On the other hand, the motifs in the synaptic connection network responding to information transmission are different from those in the ecological food networks [28] that respond to energy transmission, which seems to indicate that information flow and energy flow have significantly different patterns and features. In particular (see Figure 1), three motifs in subfigure (**1**), (**2**), and (**3**) appear frequently in the electronic sequential logic circuit networks and synaptic connection network, while the motifs in subfigure (**3**) and (**4**) appear frequently in the ecological food networks.

In addition, we can capture key structure information by analyzing specific motifs, which is difficult to get through the single edges in networks. For example, using motifs to bring high-order local structures into the graph neural network (GNN) can enhance the capability of the GNN model compared to edge-based models and improve the performance of downstream tasks [30]. Benson *et al.* proposed that important hub cities can be found in the airport networks by using specific motifs rather than edges [31]. Zhao *et al.* proposed to merge high-level relationships into regular PageRank algorithm [32], which can significantly improve the performance of user ranking in social networks [33]. The motif is a highly active research topic and a lot more research combining real-world data could be expected in the future.
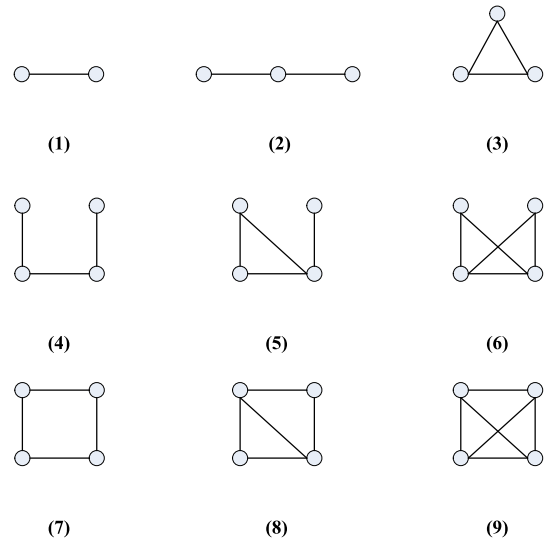

FIGURE 1. Motifs for undirected graph. (1) is a 2-motif, (2)-(3) are 3-motifs, and (4-9) are 4-motifs.

III. METHODS
In this section, a comparative analysis of the existing similarity measurements for link prediction is performed, and a new similarity index based on the high-order structure, in our case motifs, will be proposed to investigate the possibility of connecting two nodes on a given graph.

A. ANALYSIS OF EXISTING SIMILARITIES
Similarity indexes play an important role in predicting links and evaluating the possibility of adding new edges by setting scores for node pairs on a graph. Many similarity indexes are proposed to measure the score for each node pair, such as *Common Neighbor* (CN), *Jaccard* (JC), *Resource Allocation* (RA), *Preferential Attachment* (PA), *Cannistrai-Alanis-Ravai* (CAR):

$$CN(x, y) = |N(x) \cap N(y)|, \tag{1}$$

$$JC(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}, \tag{2}$$

$$PA(x, y) = |N(x)| \cdot |N(y)|, \tag{3}$$

$$RA(x, y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{deg(z)}, \tag{4}$$

$$CAR(x, y) = CN(x, y) \cdot |E(N(x) \cap N(y))|, \tag{5}$$

where $N(x)$ denotes the neighborhood of node x, $E(N(x) \cap N(y))$ denotes the edge set induced by the node set $N(x) \cap N(y)$, and $|N(x)|$ denotes the number of nodes in the set $N(x)$. These indexes mainly focus on common neighbors of a pair of nodes (x, y), and RA and CAR consider the edge density among the local environment of the node pair (x, y), in which $deg(z)$ reflects number of nodes connected to the common neighbor z and $E(N(x) \cap N(y))$ indicates the edge density among the common neighbors of node x and y.

There is another index *Clustering Coefficient for Link Prediction* (CCLP):

$$CCLP(x, y) = \sum_{z \in N(x) \cap N(y)} CC_z, \quad (6)$$

where CC_z is the local clustering coefficient of node z. The CAR and CCLP have better performance on many different datasets [9]. In the viewpoint of motif, the CAR and CCLP have strong correlations with 3-motifs (triangle motif). The definitions of CCLP can be rewritten as:

$$CCLP(x, y) = \sum_{z \in N(x) \cap N(y)} \frac{motif(z)}{deg(z) \cdot (deg(z) - 1)/2}, \quad (7)$$

where $motif(z)$ counts the 3-motif number related with node z. For the CAR index, any $(z_1, z_2) \in E(N(x) \cap N(y))$ is an edge between the common neighbors z_1, z_2 of x and y, and two 3-motifs are composed by x, z_1, z_2 and y, z_1, z_2. The CAR and CCLP indexes take into account the information of the common neighbors and perform better on the link prediction problem, as shown in Figure 2.
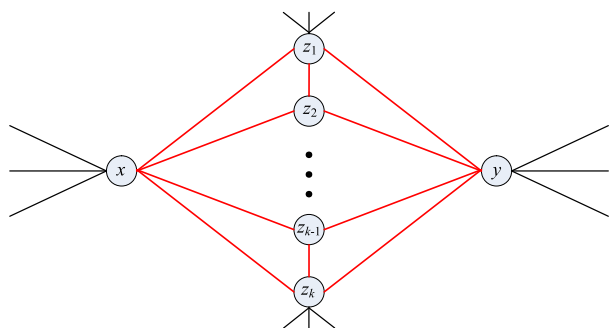
FIGURE 2. The local environment of the CAR and CCLP indexes. The red lines illustrate the 3-motifs related with the common neighbors. CCLP counts the clustering coefficient of the nodes $z_1, \cdots, z_k$, and CAR can induce 3-motifs by the edges between common neighbors, such as the red triangles in the figure.

B. MOTIF-BASED SIMILARITY MEASUREMENT

The CN, JC, PA and RA indexes all focus on the common neighbor number of the predicted edge (x, y). The structures (x, z_i, y) can be viewed as the motif in Figure 1 (**2**), and these structures can also be treated as underlying motifs in Figure 1 (**3**) if nodes x and y are connected. Then, the structures (x, z_i, y) are named as *underlying triangle motifs*. The CAR and CCLP indexes mainly reflect the influence of triangle motifs in the local neighborhood of the predicted edge.

Motivated by the existing similarities and link prediction requirements, the prediction of edge (x, y) should consider three parts: the neighborhood of node x, the neighborhood of node y and their intersection. Similarities for link prediction aim to measure the correlations between nodes x and y, and if the neighborhood intersection is relatively larger (compared to the neighborhood union), nodes x and y should have higher possibility to be connected, which is reflected by the definition of JA index concerning the node degrees.

To provide a high-order structure description for the correlation between x and y, we propose a *Motif-Based Similarity*, which considers the local environment based on the motifs:

$$MS(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cap N(y)| + motif(x) + motif(y)}. \quad (8)$$

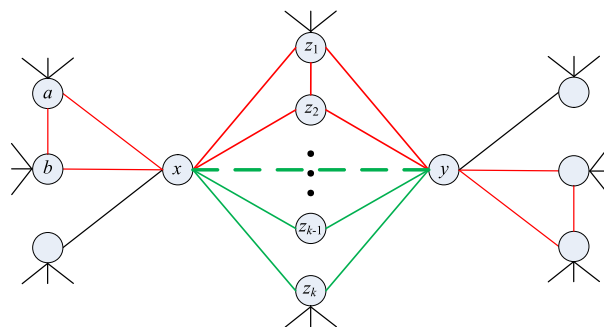A schematic graph is presented in Figure 3.

FIGURE 3. The schematic graph of the motif-based similarity. The common neighbors $z_1, \cdots, z_k$ will form underlying triangle motifs with nodes x and y, which is illustrated by the green triangles. The $motif(x)$ counts two kinds of motifs, one is induced by the common neighbor such as the red triangle motif (x, z_1, z_2), the other is induced other neighbors such as the red triangle motif (x, a, b). $motif(y)$ has similar cases as $motif(x)$.

In this research, the motif-based similarity aims at the triangle motif to measure the local correlations between nodes. The numerator denotes the new generated triangle motifs if node x and y are connected, and the denominator gives all the possible triangle motifs with edge (x, y) added. It is worth noting that here we require $motif(x) > 0$ and $motif(y) > 0$ to avoid trivial case that the MS index degenerates to 1.

C. GENERALIZATION OF THE MOTIF-BASED SIMILARITY

The definition of motif-based similarity can be easily generalized to more complicated structures. The motifs in this paper mainly focus on the triangle motif. When considering other high-order motifs, the *generalized Motif-Based Similarity* can be defined as:

$$MS_g(x, y) = \frac{|Umotifs(x, y)|}{|Umotifs(x, y)| + motif(x) + motif(y)}, \quad (9)$$

where $Umotifs(x, y)$ denotes the new generated motifs after x and y get connected and $motif(x)$ and $motif(y)$ are the numbers of existing motifs related with nodes x and y. Here, the motifs could be 4-motifs, 5-motifs or other high-order motifs, and we could focus on some specific motifs or a combination of some different motif types. Thus, the definition of motif-based similarity can fulfill more complicated task requirements, and taking advantage of the motif structures could promote the research of nodes correlation/similarity and the link prediction problem. In this paper we only consider the 3-motifs in all experiments, since for most small-scale networks 3-motifs are sufficient to reflect the typical structure characteristics, and it is still a challenge to identify and apply appropriate and exact high-order motifs accurately for large-scale networks.

D. COMPUTATIONAL COST OF THE MOTIF-BASED SIMILARITY

For the motif-based similarity, it only refers to the neighbors of node x and y. If node x and y have degree $deg(x)$ and $deg(y)$, calculating $N(x) \cap N(y)$ needs at most $deg(x) \times deg(y)$ basic steps and calculating $motif(x)$ and $motif(y)$ only needs $(deg^2(x) + deg^2(y))/2$ basic steps. So, calculating the motif-based similarity of node pair (x, y) requires $deg(x) \times deg(y) + (deg^2(x) + deg^2(y))/2$ computational cost. For a network with nodes number N and average degree $\langle k \rangle$, this time cost has an average $2\langle k \rangle^2$ value and the total time cost for evaluating the similarity scores of all the $N^2/2$ node pairs is $N^2\langle k \rangle^2$ in average.

The CAR index also requires to calculate the common neighbors $N(x) \cap N(y)$ which costs at most $deg(x) \times deg(y)$ basic steps, and counting $|E(N(x) \cap N(y))|$ will cost at most $min^2(deg(x), deg(y))/2$ basic steps, which leads to average $\frac{3}{2}\langle k \rangle^2$ time cost for each pair of nodes and $\frac{3}{4}N^2\langle k \rangle^2$ in total. The CCLP index requires to calculate the clustering coefficients of the common neighbors in $N(x) \cap N(y)$, which refers to the second-order neighbors of nodes x and y. On high-coefficient networks the number of second-order neighbors is generally larger than that of the first-order neighbors, which lead to a higher time cost for each node pair and the total. Compared to the CCLP and CAR similarity indexes, the motif-based similarity has time cost with the same level.

IV. EXPERIMENTS AND RESULTS

A. DATASET

In the experiments, 9 public benchmark datasets and 1 random generated dataset will be used to exhibit the effectiveness of the motif-based similarity. *Dolphins* dataset is an undirected social network of frequent associations among 62 dolphins in a community living off Doubtful Sound, New Zealand. Two well-established subcommunities are contained in this dataset [34]. *Protein1, protein2, protein3* are network-based datasets which represent the 3D structure of proteins. In addition, *protein3* is a dataset of an unconnected network and contains isolated nodes [35]. *NetScience* dataset is about the co-authorship network among scientists in the field of network science since 2006. This network is a one-mode projection from the bipartite graph of authors and their scientific publications. In this article, we select the largest maximal clique in the dataset as our experiment data [36]. *Jazz* dataset is an undirected social network of cooperations among 198 Jazz musicians [37]. The *football* network is a representation of the schedule of Division I games in season 2000. Nodes in the graph represent different colleges and edges represent regular-season games between the two teams they connect. 115 teams were separated into 12 conferences. In the schedule, games between members of the same conference went earlier than between members of different conferences. And intra-conference games are more frequent than inter-conference games, which indicates it is highly likely to emerge community structure in this dataset [38]. *Geom* values undirected network with 7343 vertices

and 11898 edges based on the co-work between two authors. Two authors are linked with an edge, if and only if they wrote a common work (paper, book, ...). The value of an edge is the number of common works [39]. *Ca-GrQC(ca)* is an undirected and unweighted graph which describes the collaboration network of Arxiv General Relativity [40]. In the *random geometry graph* model (RGG), nodes are placed in a unit cube randomly. Two nodes are linked if the distance between them is no more than the radius threshold value set in advance. We generate a dataset of graphs with 200 nodes and about 1000 edges.

In Table 1, some basic topological features such as number of nodes and edges, average degree, average shortest distance and clustering coefficient are presented.

TABLE 1. Basic topological features of Network datasets, including number of nodes and edges, average degree, average shortest distance and clustering coefficient (* means unconnected).

Nets	N	M	$\langle k \rangle$	$\langle d \rangle$	CC
jazz	198	2742	27.7	2.24	0.62
dolphins	62	159	5.13	3.36	0.26
football	115	613	10.66	2.51	0.4
netscience	379	914	4.82	6.04	0.74
protein1	95	213	4.48	6.28	0.4
protein2	53	123	4.64	3.79	0.41
protein3	99	212	4.28	*	0.36
RGG	200	1083	11.87	6.31	1.64
geom	7343	11898	3.24	5.31	0.73
ca	4158	13422	6.46	6.05	0.63

In the preprocessing of the 10 dataset, 90% edges are randomly selected from the original network datasets as training set, and the rest 10% are set as part of the testing set T. Then, we calculate the similarity indexes of all the non-existing edges (including the testing set), and then sort the indexes values of nodes pairs in descending order. A *top-L precision* is calculated as L_T/L, where L indicates the top-L node pairs under the sorted similarity index and L_T is the number of edges belonging to the testing set in the top-L node pairs.

B. RESULTS AND ANALYSIS

In Table 2, the Area Under Curves (AUCs) of the CAR, CCLP, CN, JC, PA, RA, n2v and motif-based similarity (MS) indexes are provided to show the efficiency of the proposed motif-based similarity. Except on the *jazz, geom* and *ca* datasets, the MS index has better performance than others and RA works better on these three datasets, but the MS index also has comparable results on them. On the datasets *dolphins, netscience, protein3, geom, ca* the MS index improves the AUC values larger than 2% than the existing methods. Compared with CAR, CCLP and n2v, the MS index outperforms them greatly on all the datasets except *jazz*.

In Table 3, the precision of methods CAR, CCLP, n2v and MS are provided to show the link prediction accuracy, which give the average ratio of L_T/L for $L = 20$. By the results it can be seen that the MS index outperforms other except only the *geom* dataset, for the relatively small networks the MS index can have better performance, but the CCLP also has
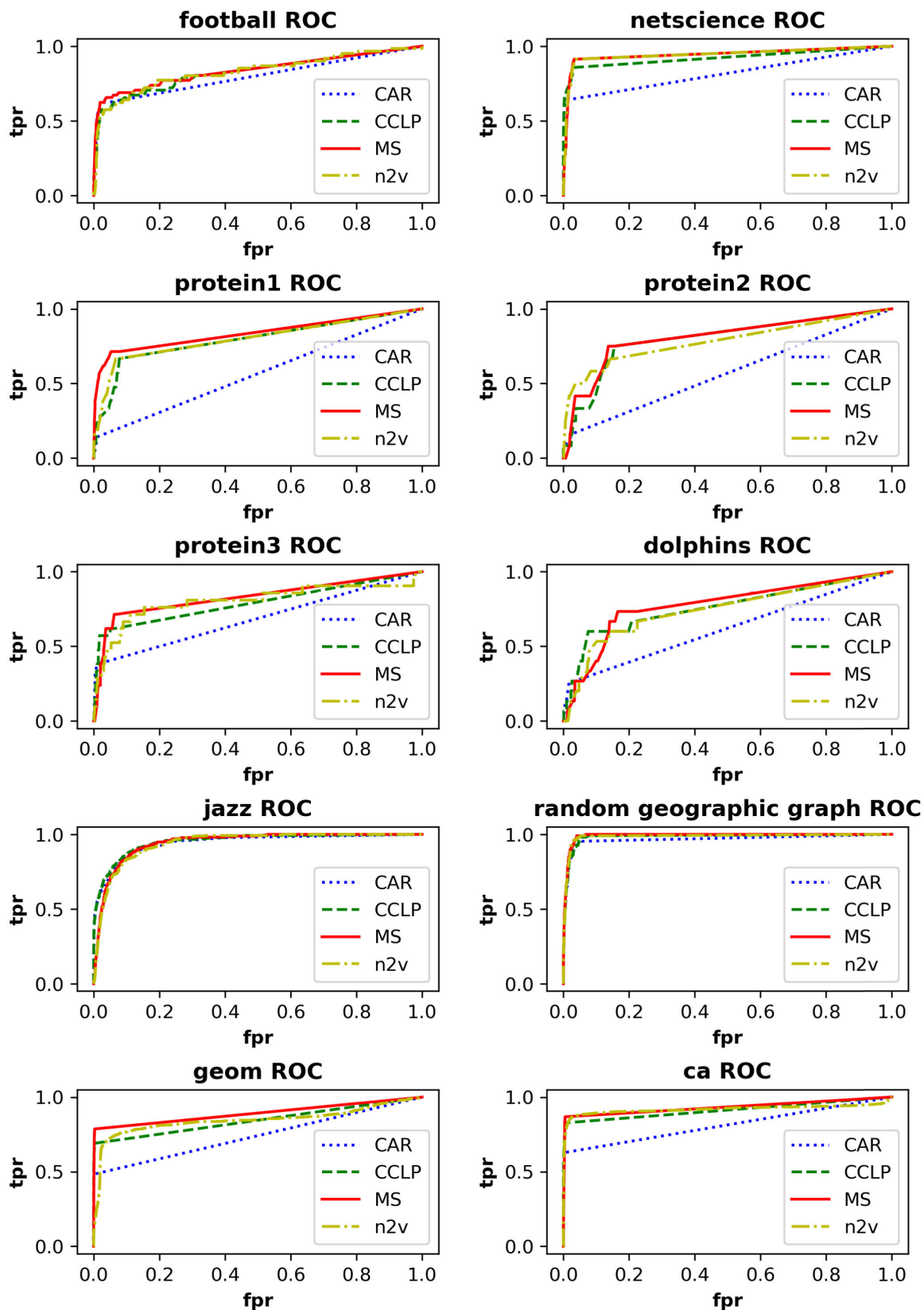
FIGURE 4. The ROC curves of index MS, n2v, CAR and CCLP. The vertical axis shows TPR (true positive rate) and the horizontal axis represents FPR (false positive rate). Each result is the average on 100 such experiments.
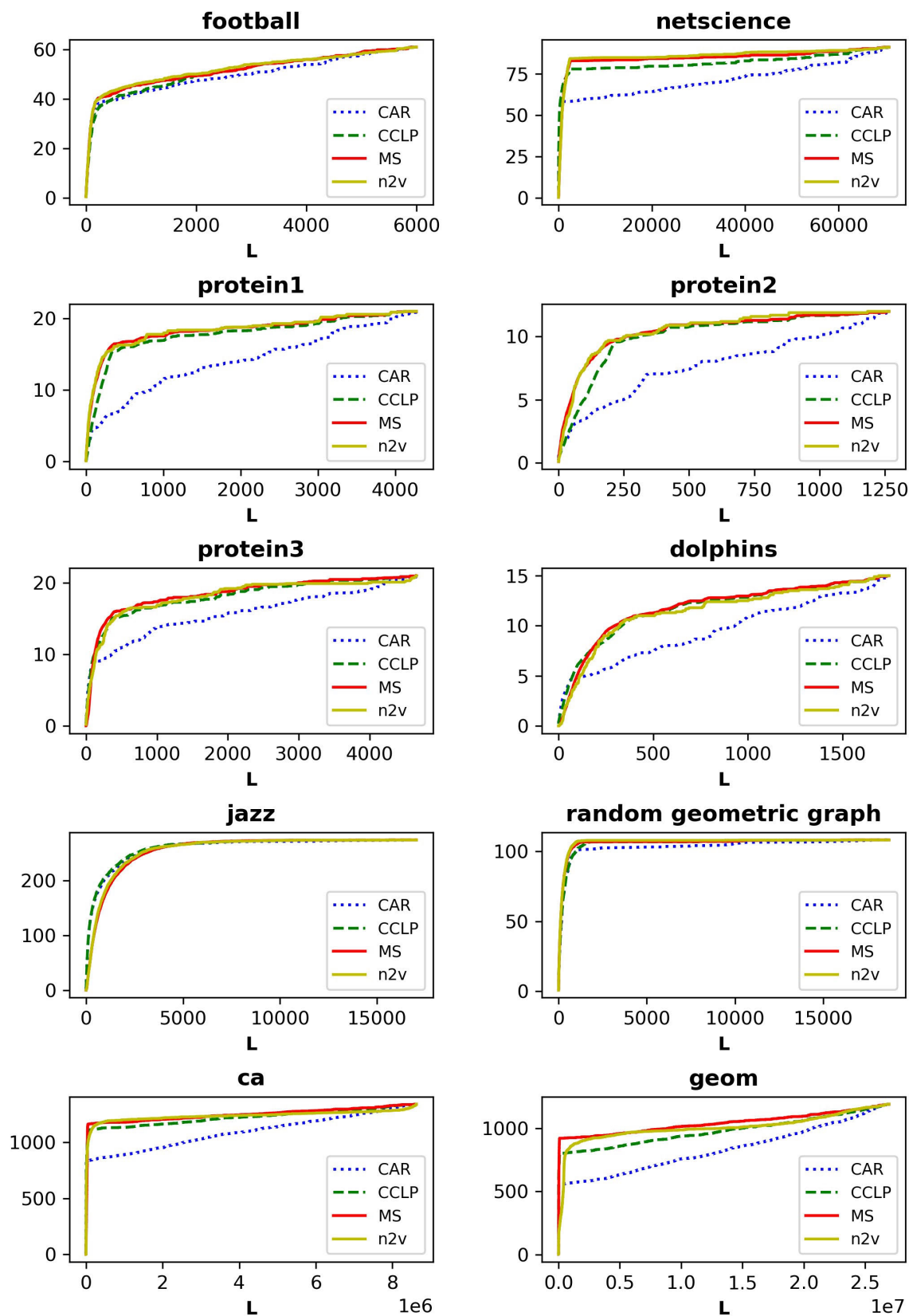
FIGURE 5. The results of prediction by the MS, n2v, CAR and CCLP indexes. The vertical coordinate is the cumulative number of predicted edges L_T, the horizontal coordinate is L. Each result is the average on 100 such experiments.

TABLE 2. Link prediction accuracy of compared similarity indexes estimated by AUC. Each result is the average on 100 such experiments.

AUC	CAR	CCLP	MS	CN	JC	PA	RA	n2v
jazz	0.954	0.959	0.941	0.955	0.96	0.771	**0.97**	0.939
dolphins	0.652	0.838	**0.865**	0.739	0.733	0.643	0.732	0.777
football	0.814	0.84	**0.851**	0.839	0.85	0.27	0.839	0.83
netscience	0.817	0.923	**0.947**	0.852	0.864	0.273	0.853	0.945
protein1	0.564	0.84	**0.872**	0.861	0.868	0.39	0.863	0.858
protein2	0.577	0.819	**0.854**	0.834	0.85	0.494	0.841	0.843
protein3	0.672	0.822	**0.851**	0.823	0.827	0.532	0.825	0.804
RGG	0.971	0.985	**0.989**	0.984	0.988	0.53	0.986	0.988
geom	0.731	0.835	0.886	0.891	0.891	0.765	**0.891**	0.846
ca	0.809	0.913	0.932	0.935	0.935	0.740	**0.936**	0.924

TABLE 3. Link prediction accuracy of compared similarity indexes estimated by precision. Each result is the average on 100 such experiments.

precision	MS	CCLP	CAR	n2v
jazz	**0.981**	0.919	0.938	0.944
dolphins	**0.113**	0.05	0.05	0.013
football	**0.431**	0.406	0.375	0.3
netscience	**0.713**	0.625	0.619	0.45
protein1	**0.2**	0.075	0.1	0.075
protein2	**0.219**	0.069	0.069	0.094
protein3	**0.175**	0.106	0.138	0.113
RGG	**0.819**	0.65	0.613	0.65
geom	0.72	**0.99**	0.97	0.69
ca	**0.99**	**0.99**	0.97	**0.99**

superiority on the large-scale networks *geom* and *ca*, which indicates more complicated high-order motifs should be considered to optimize the MS index. As the link prediction problem aims at predicting the potential connections among nodes, the False Negative Rate means little for the aim and the evaluation of recall for these methods is not provided here.

As the motif-based similarity mainly focuses on the triangle motif structures and has close relations with CAR and CCLP, in Figure 4 we illustrate the ROC (receiver operating characteristic) curves of MS, CAR and CCLP as local methods, together with n2V as global method for link prediction. AUC is the area under the ROC curve, and ROC reflects a similar effect as AUC. By the ROC, it can be seen that on *jazz* and *RGG*, the four methods work similarly; on the *dolphins*, *netscience*, *protein2* and *protein3*, MS, n2v and CCLP outperform in different FPR ranges; on the *football*, *protein1*, *geom*, *ca* datasets, MS presents obvious superiority compared to others.

The cumulative numbers of predicted edges L_T for increasing L is shown in Figure 5 for different datasets. The proposed index MS is fully dominant on the *geom* dataset compared to other indexes, and in data set *protein3* and *dolphins*, MS gets the more comparable cumulative number of predicted edges than others in a large range of L. Moreover, except for some initial L values, MS and n2v achieve the best performance in most cases, which is true for datasets such as *protein1*, *protein2*, *football*, *netscience* and *RGG*. For datasets *football* and *RGG*, when L exceeds around 20, MS and n2v exceed CCLP in performance and remain in the lead since then, and for datasets *protein3* and *dolphins*, this threshold may be a bit higher and around 200. While MS and n2v do not perform as well as CCLP in some cases, it remains its advantage over

CAR in all these datasets. For dataset *netscience* and *ca*, MS and n2v outperform for relatively large L values, which accounts for only a small proportion if we consider all the values of L. For the dataset *jazz*, the method MS and n2v have not gained better performance than CAR and CCLP for a long time but reached the same level of predicting ability as them in the end. The curves of cumulative numbers of predicted edges for MS and n2v are almost the same for all datasets except *geom*.

By the topological features of Network datasets from Table 1 and AUC results from Table 2, it could be observed that the motif-based index works more efficiently when the average degree of the network is not too large. For the *jazz*, its average degree $\langle k \rangle$ is 27.7 and it means that the local connection is very dense, which suggests that there will be too many triangle motifs in each node-pair neighborhoods. In this case, the statistical characteristics of the common neighbors $N(x) \cap N(y)$ is prominent to reflect the correlation and similarity of two targeted nodes x and y, so the CCLP and CAR indexes outperform the motif-based similarity. In the *RGG* and *football* dataset, the average degree is near around 10 and the MS, n2v, CAR and CCLP indexes have similar performance. When the average degree is around 5, motif-based similarity has significant performance on the rest 5 datasets. This phenomenon illustrates that the motif-based similarity is more suitable for the low-average-degree datasets. On the large-scale networks *geom* and *ca*, MS does not have the best performance and it is believed high-order motif may improve the results by using proper motifs.

The similarity indexes listed for link prediction in Table 2 can be classified into 4 classes: the one aiming at only the common neighbor information including CN, PA, CAR and CCLP, the one only considering the neighbors of the aimed node pair including PA, the indexes JC and MS which calculate both the neighborhood and common neighbor information, and the last one n2v which uses global information of the network. For the high-average-degree networks, the common neighbor information will play a major role in measuring the node-pair similarity for link prediction, and for a relatively low-average-degree network, the neighborhood and common neighbors should both be considered to calculate the similarity of link prediction. Also, the main difference of the JC and MS indexes is that the high-order structure is considered for our proposed motif-based similarity. High-order structure and more connection information considered in a network dataset will help understand the underlying node correlations and improve the efficiency in the link prediction problem.

V. CONCLUSION

Motif-based similarity for link prediction provides a new viewpoint to understand the local environments of target node-pair correlations. It takes the advantages of the existing indexes such as CAR and CCLP, and has a better performance by experiments on some datasets. Typical or designed motifs on a network play important roles in the link prediction problem. Sufficiently and effectively using proper

motif structures can help achieve deeper understanding of the underlying mechanism for link prediction problem. Considering the expendability of the motif-based similarity, extracting and locating the most related motifs will be an interesting direction to improve the efficiency, which will be further studied in our future work. Besides, combining the typical motif structures with other link prediction techniques (e.g., Graphic Neural Networks) would also be a valuable research direction.

The research direction of the link prediction problem has gradually shifted from methods that rely on node attributes to methods that use network structure information [2], and the accuracy has been significantly improved. Existing algorithms mainly describe the structural characteristics of a certain aspect of the network, and their predicting capabilities vary a lot in different networks. For more complex networks, such as weighted networks and directed networks, how to predict through structural information is still worthy of in-depth discussion [41]. This research not only helps to reveal the advantages and limitations of the link prediction problem itself, but also has significant practical values.

REFERENCES

[1] G. Lise and D. Christopher, "Link mining: A survey," *ACM SIGKDD Explor. Newslett.*, vol. 7, no. 2, pp. 3–12, 2005.

[2] 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, May 2007.

[3] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008.

[4] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *J. Math. Sociol.*, vol. 1, no. 1, pp. 49–80, Jan. 1971.

[5] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull. Soc. Vaudoise Sci. Nat.*, vol. 37, no. 1901, pp. 547–579, 1901.

[6] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Social Netw.*, vol. 25, no. 3, pp. 211–230, Jul. 2003.

[7] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[8] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks," *Sci. Rep.*, vol. 3, no. 1, pp. 1–14, Dec. 2013.

[9] Z. Wu, Y. Lin, J. Wang, and S. Gregory, "Link prediction with node clustering coefficient," *Phys. A, Stat. Mech. Appl.*, vol. 452, pp. 1–8, Jun. 2016.

[10] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 855–864.

[11] R. Milo, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002.

[12] R. R. Sarukkai, "Link prediction and path analysis using Markov chains," *Comput. Netw.*, vol. 33, nos. 1–6, pp. 377–386, Jun. 2000.

[13] A. Popescul and L. H. Ungar, "Statistical relational learning for link prediction," in *Proc. IJCAI Workshop Learn. Stat. Models Relational Data*. 2003, pp. 81–87.

[14] D. Lin, "An information-theoretic definition of similarity," in *Proc. ICML*, vol. 98, 1998, pp. 296–304.

[15] G. Salton, "The SMART and SIRE experimental retrieval systems," in *Introduction to Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983, p. 448.

[16] E. Ravasz, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, Aug. 2002.

[17] 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.

[18] 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.

[19] J. Chen, J. Zhang, X. Xu, C. Fu, D. Zhang, Q. Zhang, and Q. Xuan, "E-LSTM-D: A deep learning framework for dynamic network link prediction," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Aug. 22, 2019, doi: 10.1109/TSMC.2019.2932913.

[20] B. Moradabadi and M. R. Meybodi, "Link prediction in weighted social networks using learning automata," *Eng. Appl. Artif. Intell.*, vol. 70, pp. 16–24, Apr. 2018.

[21] Z. Wang, Z. Ren, C. He, P. Zhang, and Y. Hu, "Robust embedding with multi-level structures for link prediction," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5240–5246.

[22] A. Ghasemian, H. Hosseinmardi, A. Galstyan, E. M. Airoldi, and A. Clauset, "Stacking models for nearly optimal link prediction in complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 38, pp. 23393–23400, Sep. 2020.

[23] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," 2018, *arXiv:1802.09691*. [Online]. Available: https://arxiv.org/abs/1802.09691

[24] S. De Winter, T. Decuypere, S. Mitrovic, B. Baesens, and J. De Weerdt, "Combining temporal aspects of dynamic networks with Node2Vec for a more efficient dynamic link prediction," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 1234–1241.

[25] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genet.*, vol. 31, no. 1, pp. 64–68, May 2002.

[26] M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, and C. Lengieza, "YPD, PombePD and WormPD: Model organism volumes of the bioKnowledge Library, an integrated resource for protein information," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 75–79, 2001.

[27] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, "The structure of the nervous system of the nematode Caenorhabditis elegans," *Philos. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 314, no. 1165, pp. 1–340, 1986.

[28] R. J. Williams and N. D. Martinez, "Simple rules yield complex food webs," *Nature*, vol. 404, no. 6774, pp. 180–183, Mar. 2000.

[29] R. F. I. Cancho, C. Janssen, and R. V. Solé, "Topology of technology graphs: Small world patterns in electronic circuits," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 4, Sep. 2001, Art. no. 046119.

[30] X. Li, W. Wei, X. Feng, X. Liu, and Z. Zheng, "Representation learning of graphs using graph convolutional multilayer networks based on motifs," 2020, *arXiv:2007.15838*. [Online]. Available: https://arxiv.org/abs/2007.15838

[31] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, Jul. 2016.

[32] A. Langville and C. Meyer, "Deeper inside pagerank," *Internet Math.*, vol. 1, no. 3, pp. 335–380, Jan. 2004.

[33] H. Zhao, X. Xu, Y. Song, D. L. Lee, Z. Chen, and H. Gao, "Ranking users in social networks with motif-based pagerank," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 5, pp. 2179–2192, May 2020.

[34] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behav. Ecol. Sociobiol.*, vol. 54, no. 4, pp. 396–405, Sep. 2003.

[35] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004.

[36] 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.

[37] P. M. Gleiser and L. Danon, "Community structure in jazz," *Adv. Complex Syst.*, vol. 6, no. 4, pp. 565–573, Dec. 2003.

[38] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Apr. 2002.

[39] B. Jones, "Computational geometry database," FTP/HTTP, Feb. 2002.

[40] R. Rossi and N. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 29, no. 1, 2015, pp. 1–2.

[41] T. Murata and S. Moriyasu, "Link prediction of social networks based on weighted proximity measures," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Nov. 2007, pp. 85–88.

XIANGNAN FENG was born in 1993. He received the Ph.D. degree in mathematics from the School of Mathematical Sciences, Beihang University, Beijing, China, in 2021. He is currently a Postdoctoral Fellow with the Max Planck Institute for Human Development, Berlin, Germany. His research interests include complex networks, computing social science, and artificial intelligence.

CHAO LI was born in 1981. He received the master's degree in measurement technology and instrumentation from the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, China, in 2010. He is currently pursuing the Ph.D. degree with Yanshan University. He is an Associate Professor with the School of Mathematics and Computer, Hengshui University, Qinhuangdao, China. His research interests include artificial intelligence and complex networks.

WEI WEI was born in 1981. He received the Ph.D. degree in mathematics from the School of Mathematical Sciences, Peking University, Beijing, China, in 2009. He is currently an Associate Professor with the School of Mathematical Sciences, Beihang University, Beijing. His research interests include dynamical system and complexity, complex networks, and artificial intelligence.

JIAOMIN LIU was born in 1958. He received the Ph.D. degree from the Hebei University of Technology, Tianjin, China, in 1998. He is currently a Professor with the School of Information Science and Engineering, Yanshan University, Qinhuangdao, China. His research interests include computer intelligent control and multimedia fusion technology.

• • •