

Received April 20, 2020, accepted May 4, 2020, date of publication May 8, 2020, date of current version May 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993455

# Quantifying the Effect of Community Structures for Link Prediction by Constructing Null Models

XIAO-KE XU<sup>1</sup>, (Member, IEEE), KE-KE SHANG<sup>1,2</sup>, AND JING XIAO<sup>1</sup>

<sup>1</sup>College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China

<sup>2</sup>Computational Communication Collaboratory, Nanjing University, Nanjing 210023, China

Corresponding author: Xiao-Ke Xu (xuxiaoke@foxmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773091, Grant 61603073, and Grant 61803047, in part by the LiaoNing Revitalization Talents Program under Grant XLYC1807106, and in part by the Program for the Outstanding Innovative Teams of Higher Learning Institutions of Liaoning under Grant LR2016070.

**ABSTRACT** The existing studies of link prediction mostly focused on using network topology properties to improve the accuracy of link prediction. More broadly, researches on the role of community structures for link prediction have recently received increasing attention. In this study, we propose a succinct algorithm that is built on community structures to improve the performance of link prediction, and it has been verified by both of synthetic benchmarks and real-world networks. More importantly, we introduce different null models to study the role of community structures on link prediction more carefully. Firstly, it is found that clearer community structures correspond to the higher performance of link prediction algorithms that are based on community information. Secondly, the role of links within a community and that between two communities are further distinguished. The edges within one community play a vital role for link prediction of the whole network, and conversely the edges between two communities have a minimal effect on that. At last, we reveal the relationship and dependence between this special meso-scale structure (community) and micro-scale structures of different orders (i.e., degree distribution, assortativity, and transitivity) for link prediction.

**INDEX TERMS** Null model, link prediction, community structure, null network, network topology.

## I. INTRODUCTION

In the field of complex networks, link prediction is a significant problem of predicting the unknown or fake links from a network which has uncertain structures [1], [2]. Link prediction algorithms can apply in identifying unknown interactions to reduce the cost of experiments [3]. For example, the vast majority of interactions among different proteins are unknown in biology [4]–[6], so traditionally we have to spend expensive cost to restore these unclear interactions.

Link prediction plays an essential role in various applications of complex networks. For example, link prediction can be utilized to analyze the evolution of social networks, such as using historical structures of an online social network to predict whether a pair of friends will comment on each other in future [7]–[10]. Link prediction also can measure the fitness of network models for real-life networks [11]. Furthermore, link prediction is helpful for designing recommendation algorithms [12], detecting spurious links [2], ana-

lyzing cascading failures [13]–[16], network reconstruction and node classification [17].

The existing link prediction algorithms can be divided into two categories: structural similarity algorithms in network domain [18] and network embedding algorithms in the field of machine learning [19]. Traditionally, link prediction methods in network domain tend to use micro-scale topology properties, such as link existence [11], link weights [20], common neighbors [21], node degree [17], [22], clustering coefficient [23], to predict the links. In recent years, the technique of network embedding has been widely applied in link prediction [19]. It aims to map network data into a low dimensional space in which the network neighborhood information is maximumly preserved. In the previous study, we systematically compared the two categories of prediction algorithms and studied the shortcomings of network embedding algorithms in short-path networks [24]. Our previous findings imply that the algorithms in network domain can effectively characterize node similarity by micro-scale and meso-scale (e.g., community) structure information, thus better performance can be achieved in the category of short-path networks compared with those of network embedding algorithms.

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao<sup>1</sup>.

Community structure is one of the most important meso-scale characteristics of complex networks, allowing us to unearth in-depth network information that may not be obtained from direct observation, e.g., functional and dynamic characteristics. Recently, researches pay more attention to link prediction algorithms which are based on community structure information [25]–[29]. In detail, researchers have proposed some link prediction algorithms which take advantage of a pair of nodes that locate in the same community, and found that the performance of traditional local link prediction algorithms can be improved [25], [26]. However, in previous studies, the algorithms of link prediction based on community structures are always used to predict unknown links whether a synthetic or real-life network has the characteristic of community structures or not. Obviously, the networks that do not have community characteristics will not experience any performance improvement using such kind of algorithms. Moreover, as far as we know, there are insufficient studies to systematically discover detailed effects of community structures on link prediction.

In this study, our two main contributions are as follows. Inspired by previous studies such as [26], we proposed a novel algorithm of link prediction based on Common Neighbors (CN) [30] and the known information of node neighbors in the same community [31], and we name that algorithm as Global Community Neighbors (GCN). It requires less limiting conditions than previous community-based link methods (i.e., Local Community Neighbors, LCN), so the method of GCN can significantly improve the performance of link prediction by fully exploring community information compared to previous studies [26]–[29]. Our findings have been verified by both of synthetic benchmarks and real-world networks.

More importantly, we introduce different null models [32]–[34] to study the contribution of community structures for link prediction more carefully. Firstly, it is found that clearer community structures correspond to the higher performance of link prediction algorithms that are based on community information. Secondly, comparing the original network and corresponding null models, we find that the edges within a community play a significant role for link prediction of the whole network. Conversely, the edges between two communities have a minimal effect on that. At last, our study also provides a framework can be utilized to distinguish different roles of micro-scale and meso-scale structures on link prediction in complex networks. In particular, it reveals the relationship and dependence between this special meso-scale structure (community) and micro-scale structures of different orders (i.e., degree distribution, assortativity, and transitivity) for link prediction.

## II. THE PROPOSED METHOD OF GLOBAL COMMUNITY NEIGHBORS

### A. COMMON NEIGHBORS (CN)

Traditionally, link prediction methods tend to use local topology properties. With this intuition, we try to modify several basic local structure similarity methods for link prediction:

Common Neighbors (CN), Jaccard Similarity, Resource Allocation (RA), and Leicht-Holme-Newman (LHN) [17]. CN is the basic and simplest form of these metrics, so we only present the CN-based method here.

The link prediction method of Common Neighbors (CN) is based on the node similarity of local information proposed by Liben-Nowell and Kleinberg [30]. First, we should select a pair of nodes ( $a$  and  $b$ ) that have common neighbors, and let  $\Gamma(a, b)$  denote the set of common neighbors of nodes  $a$  and  $b$ . Then, the score of the link between  $a$  and  $b$  is equal to the number of common neighbors between them:

$$CN(a, b) = |\Gamma(a, b)|. \quad (1)$$

In this case, a higher CN score means that it is more likely to have a relationship between  $a$  and  $b$ .

### B. LOCAL COMMUNITY NEIGHBORS (LCN)

Recently, some novel algorithms attempt to use the information that a pair of nodes and their common neighbors which locate in the same community [35]–[39] to improve the performance of traditional methods for link prediction [25], [26]. Based on CN, they introduce the number of common neighbors which locate in the same community to propose the method of Local Community Neighbors (LCN):

$$LCN(a, b) = CN(a, b) + \sum_{i \in \Gamma(a, b)} |C(i) \cap C(a) \cap C(b)|. \quad (2)$$

Here,  $C(a)$  and  $C(b)$  are the communities that nodes  $a$  and  $b$  belong to, respectively. In Equation 2, the first half part is the number of common neighbors for nodes  $a$  and  $b$ , and the second half part is the weight of community information. Because node  $i$  has to be the common friend of nodes  $a$  and  $b$ , we call it a local way of using community information. If a node  $i$  belongs to the community of node  $a$ , and node  $b$  also belongs to this community at the same time, the second half is going to add 1.

### C. GLOBAL COMMUNITY NEIGHBORS (GCN)

The LCN algorithm has the following disadvantages. Firstly, such kind of algorithms only utilize local community information but does not fully explore global community information. Secondly, whether a synthetic or real-life network has the characteristic of community structures, the LCN algorithms in previous studies are usually used to predict unknown links. Obviously, the networks that do not have community characteristics will not experience any performance improvement using such kind of algorithms. At last, as far as we know, there are insufficient studies to systematically discover detailed effects of community structures on link prediction.

In this study, we propose a novel algorithm of link prediction to take advantage of the number of common neighbors in the same community. The number of neighbors which locate in the same community is also introduced to present another method of using community information, we call it as the

method of Global Community Neighbors (GCN):

$$GCN(a, b) = CN(a, b) + \sum_{i \in \Gamma(a), j \in \Gamma(b)} |C(i) \cap C(j)|. \quad (3)$$

Here, node  $i$  is the neighbor of node  $a$  and  $C(i)$  denotes the community that node  $i$  belongs to. At the same time, node  $j$  is the neighbor of node  $b$  and  $C(j)$  denotes the community that node  $j$  belongs to. The main difference between GCN and LCN is that GCN does not require node  $i$  or  $j$  is a common neighbor of nodes  $a$  and  $b$ . Therefore, GCN requires less limiting conditions than LCN and other CN-based methods [25], [26], hence it contains more community structure information. Moreover, we introduce different null models to study the role of community structures on link prediction more carefully.

In the GCN algorithm, it is necessary to balance the effects of micro-scale (common neighbors) and meso-scale (common community neighbors) characteristics. To quantify the role of community structures for link prediction, we use  $\alpha$ GCN to weight community information. A new parameter  $\alpha$  is added to GCN to propose  $\alpha$ GCN:

$$\alpha GCN(a, b) = CN(a, b) + \alpha \sum_{i \in \Gamma(a), j \in \Gamma(b)} |C(i) \cap C(j)|. \quad (4)$$

Here we set the range of  $\alpha$  to be  $[-3, 3]$ .  $\alpha > 0$  means that the information of community structures has a positive effect for link prediction. On the contrary,  $\alpha < 0$  means that the information of community structures has a negative effect for link prediction. If  $|\alpha| < 1$ , the weight of community structure information is less than that of common neighbors; if  $|\alpha| > 1$ , community information will play a more prominent role than the number of common neighbors.

### III. THE EFFECT OF PREDICTION METHODS

#### A. EVALUATION METRICS FOR LINK PREDICTION

Referring to [9], [20], the standard link prediction problem can be described as follows. A part of the non-observed links construct the set  $\bar{E}$ , and the observed link set  $E$  is divided into two parts  $E^T$  and  $E^P$ , where  $E^T \cup E^P = E$  and  $E^T \cap E^P = \emptyset$ . All the links in  $E = E^P \cup E^T$  have been observed and known. Typically 90% of observed links belong to  $E^T$ , and  $E^P$  includes 10% of observed links. The division into  $E^T$  and  $E^P$  is arbitrary and will be used for the scoring purpose [20]. Links in  $E^T$  form a *training set* and are utilized to implement a link prediction score, the effectiveness of which will be evaluated over the *probe set*  $E^P$ .

#### 1) AUC

The so-called Area Under the Receiver Operating Characteristic Curve (AUC), originally applied to evaluate communication schemes, has been widely applied to measure prediction accuracy in a variety of settings [40]. Here, we use the metric of AUC as one of the measures for link prediction accuracy. Only the information of  $E^T$  is allowed to be used to compute the performance score, and we compare the prediction scores of  $m$  pairs of nodes from  $E^P$  and  $\bar{E}$  randomly. If there are  $m'$

**TABLE 1. The corresponding meanings of parameters in the LFR benchmark network.**

Index	Description	Value
$N$	the number of node	2000
$\langle k \rangle$	the average degree of node	20
$k_{max}$	the maximal node degree	40
$\mu$	the mixing parameter	0.5
$C_{max}$	the maximal community size	50
$C_{min}$	the minimal community size	40
$O_n$	the overlapping nodes for communities	0
$\lambda$	the power-law exponent for node degree	2
$\beta$	the power-law exponent for community size	1

times which the score measured from  $E^P$  is larger than that from  $\bar{E}$  and  $m''$  times which the two scores are equal, then,

$$AUC = (m' + 0.5m'')/m. \quad (5)$$

Generally, the range of AUC values is in  $[0.5, 1]$ . An AUC value close to 1 means that the corresponding method of link prediction is very efficient.

#### 2) PRECISION

Precision is also widely used to measure the performance of designing recommendation algorithms [41] and detecting spurious links [20] in various settings. Precision is defined as the ratio of selected observed links to the number of selected items. Here, we use Precision as the other measure for prediction accuracy. In this study 10% pairs of nodes from  $E^P$  and the same number of pairs of nodes from the nonexistent links are selected randomly. We select top  $L$  (200, about 1% pairs of nodes) pairs of nodes by their scores in the LFR network. In other cases, we select 1% of network size as the value of  $L$ . If there are  $m$  links from  $E^P$ , then,

$$Precision = m/L. \quad (6)$$

The values of Precision are in the range of  $[0, 1]$ . The closer the Precision value is to 1, the better the prediction performance of the prediction method is.

### B. EXPERIMENTAL RESULTS

The LFR network is one of the most commonly used standard benchmarks to test the performance of community detection algorithms [42]. It takes into account not only the distribution of node degree, but also the distribution of community size. In this study, to provide an impartial evaluation for the role of links within the same community and that between different communities, we ensure that the two kinds of links have a similar quantity. Moreover, our LFR network is undirected. Table 1 provides sufficient background details for the LFR benchmark network.

Except the above LFR network in Table 1, we also use a series of LFR benchmark networks whose  $\mu$  values change from 0.8 to 0.1 to test how community structures affect the performance of link prediction in Sect. IV-D. Furthermore,

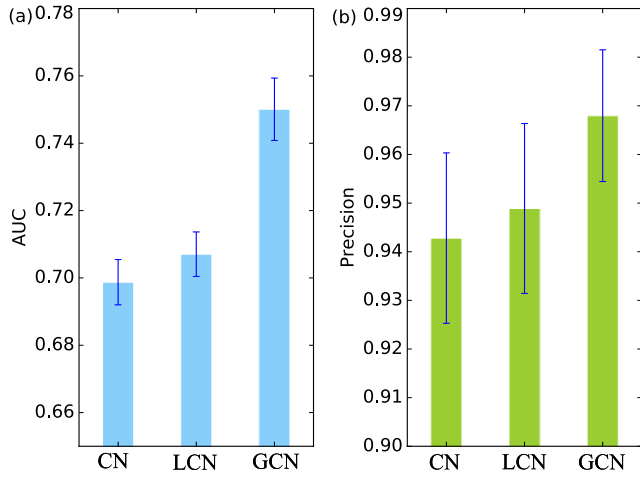


FIGURE 1. For AUC and Precision, the performance of CN, LCN and GCN algorithms in the LFR network.

we use eight real-life networks to verify that GCN can obtain higher performance than LCN and traditional CN-based algorithms in this section.

Firstly we try to select a suitable way to utilize the information of community structures for link prediction. As shown in Fig. 1, there is an obvious difference among the performance of CN, LCN and GCN. The results of LCN and GCN are both superior to that of CN. That is to say, community structures can improve the accuracy of link prediction. In detail, we can see that GCN achieves the most optimal performance for AUC. For Precision, GCN can also achieve the highest accuracy. In general, the performance of GCN stays ahead of those of CN and LCN. Our results indicate that the number of neighbors that locate in the same community can substantially alter the result of link prediction. Therefore, in the following section, we mainly take the advantage of community structure information by the GCN method.

For  $\alpha$ GCN, if  $\alpha > 0$ , community information plays a positive role for link prediction. On the contrary, if  $\alpha < 0$ , the information of community structures plays a negative role. We set the range of  $\alpha$  to be  $[-3, 3]$ . As shown in Fig. 2, we find that when  $\alpha > 0$ , community structures are always helpful to improve the performance. In contrast, the negative role ( $\alpha < 0$ ) can significantly reduce the prediction accuracy. In addition, at the cases of  $\alpha > 0$ , there is no obvious fluctuation for the algorithm performance when we vary the value of  $\alpha$ . Hence, we suggest that the specific value of the weight factor  $\alpha$  does not strongly affect the performance of link prediction, when it is greater than zero. In the following section, we set  $\alpha = 1$  and use GCN to inspect the effect of community structures for link prediction.

Apart from the synthetic benchmark networks (i.e., GN and LFR), eight well-studied real-world networks are also adopted to study the effect of community structures, including the networks of Adjectives and nouns [43], Zachary Karate Club [44], Sawmill communication [45], Bottlenecked Dolphins [46], Football [47], Polbooks [48], and

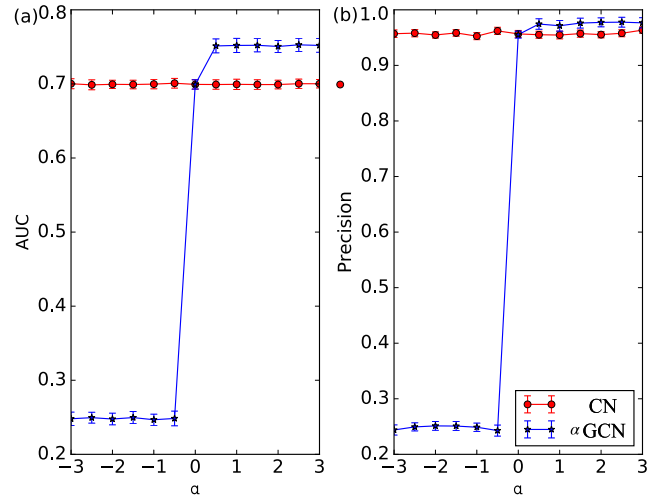


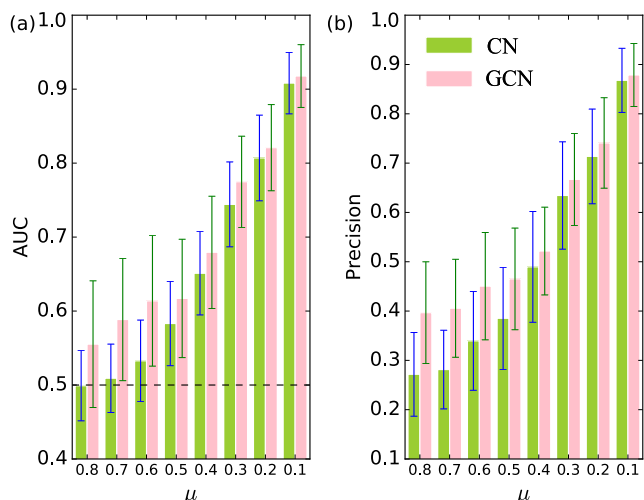
FIGURE 2. For AUC and Precision, the performance of CN and  $\alpha$ GCN algorithms varying with parameter  $\alpha$  in the LFR benchmark network.

TABLE 2. The statistics of eight real-world networks.  $N$  represents the number of nodes,  $M$  represents the number of edges,  $C_n$  is the number of communities,  $Q$  is the Modularity index obtained by Infomap,  $R_{in}$  is the ratio of the number of inner edges within each community to the total number of edges,  $\Delta AUC\%$  is the percentage of the performance improvement of GCN with CN for the value of AUC, and  $\Delta Precision\%$  is the percentage of the performance improvement of GCN with CN for the value of Precision.

Statistic	$N$	$M$	$C_n$	$Q$	$R_{in}$	$\Delta AUC\%$	$\Delta Precision\%$
Karate	34	78	3	0.402	0.821	16.33	<b>49.10</b>
Words	112	425	2	0.009	0.991	10.43	39.66
Sawmill	36	62	5	0.542	0.806	12.22	21.37
Dolphins	62	159	6	0.528	0.757	1.59	10.69
Football	115	613	12	0.601	0.690	1.93	5.14
Email	1133	5452	69	0.527	0.580	5.93	3.42
Polbooks	105	441	6	0.523	0.855	3.00	4.16
LesMis	77	254	9	0.546	0.729	1.77	3.17
LFR_5k	5000	50245	113	0.491	0.500	5.85	1.22
PowerGrid	4941	6594	498	0.813	0.816	<b>31.57</b>	42.40

Co-appearances of characters in Les Miserables [49]. The characteristics of these networks are summarized in Table 2.

As illustrated in Table 2, the method of GCN can significantly improve the performance of link prediction by fully exploring community information compared to previous studies. The new framework of link prediction by GCN is superior to the classical CN method. Experimental results on real-life networks show the highest performance improvement is 31.57% for AUC and 49.10% for Precision. Most Link prediction algorithms based on community information are scalable, because community size generally does not increase with the increase of network size. In order to prove the effectiveness of our algorithm in large-scale networks, two large-scale networks were added, namely the LFR network with 5000 nodes and 50245 edges, and the Power grid network with 4941 nodes and 6594 edges. The experimental results of the two networks show that the proposed method is robust and scalable.



**FIGURE 3.** For AUC and Precision, the performance of link prediction for a series of LFR benchmark networks. The values of  $\mu$  change from 0.8 to 0.1 with a span of  $-0.1$ .

#### IV. THE EFFECT OF COMMUNITY INTENSITY

##### A. EXPERIMENTAL RESULTS OF LFR NETWORKS

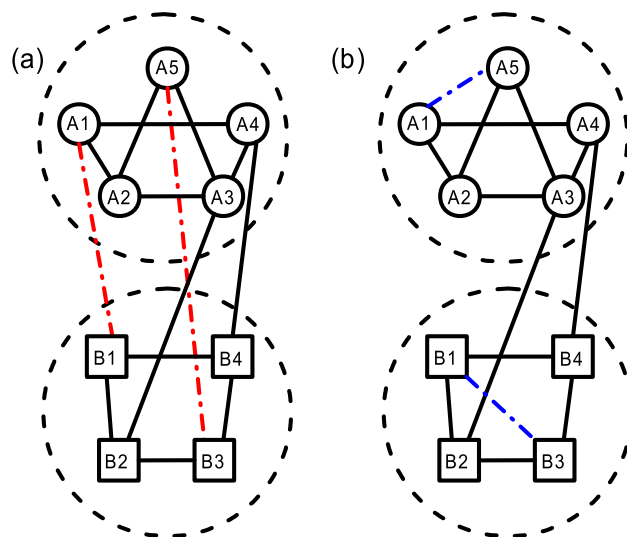
We have verified that the information of community structures can significantly improve the accuracy of link prediction algorithms. Therefore, the relationship between community structures and link prediction should be studied more carefully.

A lot of connections within communities mean that the intensity of community structures is high. In contrast, if there are few connections within each community, it means that the community intensity of a network is relatively low. By adjusting the parameter of  $\mu$ , the strength of community structures can be modified for a series of LFR networks. We conduct experiments on LFR networks in Fig. 3. In these networks, the values of  $\mu$  changing from 0.8 to 0.1 with a span of  $-0.1$  and a lower  $\mu$  value means clearer community structures. It is observed that the performance of GCN is always superior to that of CN, and both of them increase with the growing of community intensity in Figs. 3 (a) and (b). It is suggested that a higher community intensity is helpful to achieve the more outstanding accuracy for link prediction. In other words, clearer community structures (i.e., denser links in a community) correspond to the preferable accuracy of link prediction that is based on the information of community structures.

At the same time, with the increase of community intensity, the performance improvement of link prediction obtained by using community information becomes smaller and smaller as shown in Fig. 3. Although this phenomenon is very interesting and deserves further study, we cannot use these LFR networks to reveal the intrinsic mechanism, because when we vary the mixed parameter  $\mu$ , the whole topology structure of the corresponding LFR network also is modified.

##### B. NULL MODEL OF INCREASING COMMUNITY INTENSITY

The null models proposed in this section can enhance or weaken the community density of an original network, at the



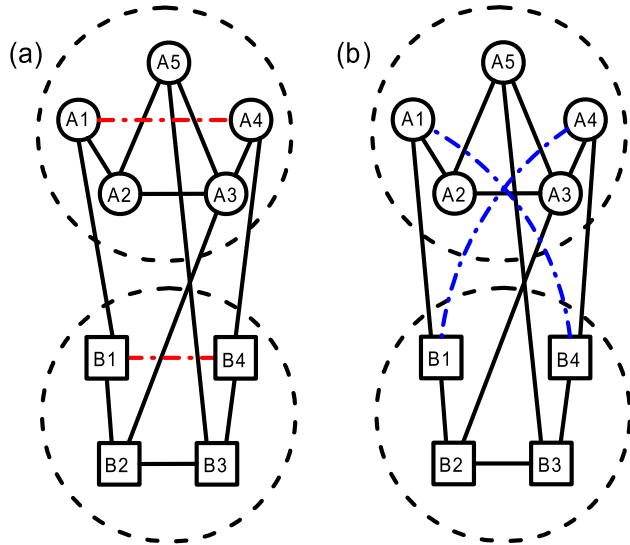
**FIGURE 4.** Constructing the null model of increasing community structures. (a) The original network, and (b) its corresponding null model of increasing community intensity.

same time maintaining its micro-scale structural characteristics. When the community structure of a network has been determined, we attempt to study the influence of its community intensity on link prediction. Traditionally, a series of benchmark networks can be constructed, but they lose the restriction for micro-scale structures of the original network [50]–[52]. Here, we introduce the first null model that is similar as Bagrow proposed in [53], which enhances community structure intensities. Meanwhile, in the case of varying network meso-scale characteristics (i.e., community structures), our null models can maximally maintain micro-structures of different orders, such as degree distribution, assortativity, and transitivity.

The process of constructing the null model of increasing community intensity is illustrated in Fig. 4. Firstly, we split a toy network into multiple communities by a classical community detection algorithm (e.g., Infomap). Secondly, we attempt to exchange a pair of external edges between two communities to be the edges within a community, in the case of keeping micro-scale structures (e.g., degree distribution) unchanged. For example, we can disconnect two edges A1-B1 and A5-B3. If two pairs of nodes A1-A5 and B1-B3 both are not connected, and then we connect them respectively. The result of the reconnection is illustrated in Fig. 4(b). The process of rewiring links should be repeated a lot of times until the null model meets the requirement of increasing community intensity.

##### C. NULL MODEL OF DECREASING COMMUNITY INTENSITY

To study the influence of decreasing community intensities on link prediction, we introduce the second null model, which can weaken community structures. The constructing process of reducing community structures is shown in Fig. 5. Firstly,



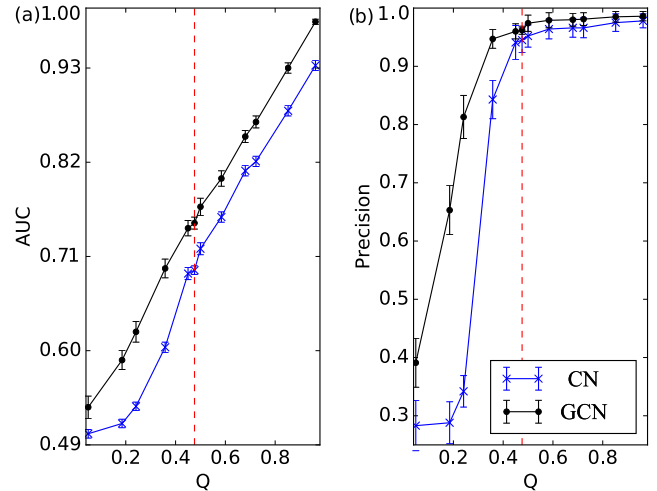
**FIGURE 5.** Constructing the null model of decreasing community intensity. (a) The original network, and (b) its null model of decreasing community intensity.

we divide the original network into multiple communities. Secondly, we try to exchange a pair of edges within a community to be external edges between two communities, at the same time keeping micro-scale structures unchanged. For example, we firstly disconnect two edges A1-A4 and B1-B4. If two pairs of nodes A1-B4 and A4-B1 are both not connected, and then we can connect them respectively. The reconnection result is displayed in Fig. 5(b). Repeating the process of rewiring links, we can obtain a series of null models with different levels of reducing community intensities. Actually, when the property of community structures of a network is weak, the kind of network can be utilized to test the robustness of community detection algorithms [33], so this type of null models has more extensive applications.

**D. EXPERIMENTAL RESULTS OF NULL MODELS**

The density of links in one community is far denser than that between two communities, corresponding to a high value of  $Q$ , which means the network has distinct community structures. Therefore, if we add the inner links of a community and at the same time remove the links between two communities,  $Q$  is going to increase. On the contrary, if we remove the links within a community and add the links between two communities, the value of  $Q$  will decrease. In this study, we use the two null models to increase and decrease community intensities (see Sect. IV-B and IV-C). As displayed in Fig. 6, we represent the prediction result of the original LFR network by the red dashed line, and we also show the prediction performance in the cases of increasing and decreasing  $Q$  values.

Obviously, the intensity of community structures for promoting the AUC and Precision performance is quite different. For the results of AUC, with the variation of  $Q$ , community information can improve the performance in a linear proportion [Fig. 6(a)]. However, the improvement of Precision



**FIGURE 6.** For AUC and Precision, the performance of CN and GCN algorithms in the increasing and decreasing  $Q$  null modes of the LFR benchmark network.

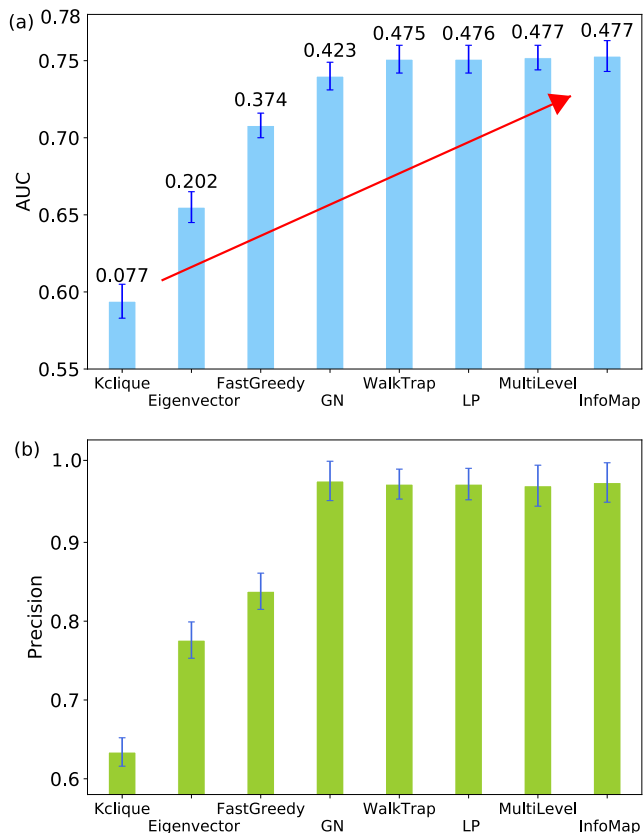
is nonlinear under different  $Q$  values [Fig. 6(b)]. When the values of  $Q$  are small ( $Q < 0.5$ ), the community structure information can greatly improve the accuracy of link prediction. On the contrary, in the cases of large  $Q$  values ( $Q > 0.5$ ), community information has little effect on the prediction performance. This is because that the value of Precision is high in itself based on CN when  $Q > 0.5$ , and it can not be greatly improved using community information.

**E. THE EFFECT OF COMMUNITY DETECTION ALGORITHMS**

Most of complex networks have community structures, and the problem of community detection is a hotspot topic in recent years [54]–[57]. If we adopt different community detection algorithms, the results of community detection for the same network may have very large variations, and they correspond to the distinct intensity of community structures. Because the detected community information can be fused in the methods of link prediction, it is significant to explore how community detection algorithms affect the performance of link prediction.

In this section, we examine the performance of eight kinds of classical community detection algorithms: MultiLevel is based on the multi-resolution version of modularity [58], FastGreedy is based on the greedy optimization [54], WalkTrap is based on random walks [59], GN is based on edge betweenness [47], Kclique is based on the relationship between nodes and subgraphs [60], InfoMap is based on random walk dynamics [61], Label Propagation (LP) is based on the propagation of labeled nodes [62] and the leading Eigenvector algorithm is based on the leading eigenvector of modularity matrix [63]. These diverse algorithms make it more accurate to analyze the effect of community intensity for link prediction in various complex networks.

As shown in Fig. 7(a), the methods of WalkTrap, LP, MultiLevel and Infomap can achieve higher performance for AUC. On the other side, Kclique, Eigenvector, FastGreedy and GN do not perform the desired results. And the similar



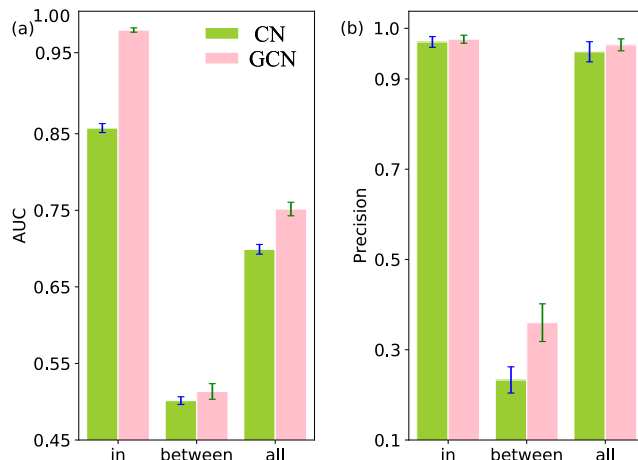
**FIGURE 7.** For AUC and Precision, the performance of eight kinds of community detection algorithms under GCN in the LFR network. We mark the value of  $Q$  on the top of each corresponding result bar of community detecting algorithms.

results for Precision are illustrated in Fig. 7(b). Furthermore, it is evident that the performance of AUC and Precision is proportional to the values of  $Q$  induced by detection algorithms. There is a large difference between the performance of the best community detection algorithm and the worst one. Hence, it suggests that the performance of link prediction is not related to the specific division obtained by community detecting algorithms, but depend on the values of  $Q$ . An excellent community detection algorithm can achieve a higher value of  $Q$ , which means that inner edges of each community in this partition are denser, and the edges between two communities are more sparse. The intensity of community structures induced by community detection methods can also affect the prediction performance, so a more accurate community detection can get higher performance for link prediction. Because the method of Infomap can achieve the largest value of  $Q$ , we utilize it to divide every benchmark network into communities for constructing null models.

## V. THE EFFECT OF LINK LOCATION

### A. PREDICTING LINKS IN A COMMUNITY AND BETWEEN TWO COMMUNITIES

To analyze how community structures affect link prediction on links in a community and links between two communi-



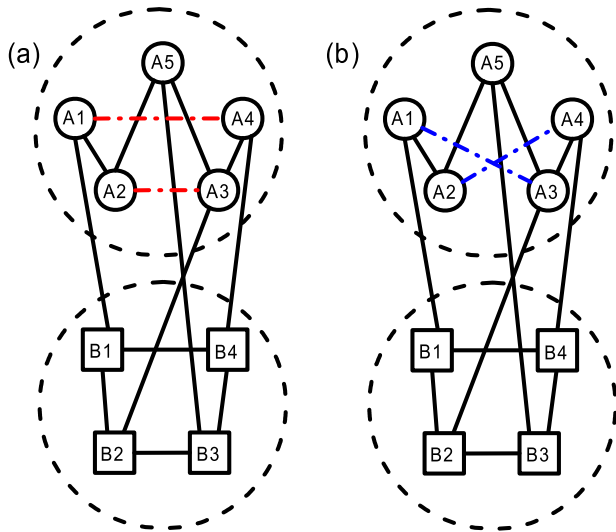
**FIGURE 8.** For AUC and Precision in the LFR network when  $\mu = 0.5$ , the performance of link prediction for all the links, the links within a community and the links between two communities under both CN and GCN algorithms.

ties more clearly, we examine the prediction results for the two kinds of links independently. As shown in Fig. 8, the performance of predicting the links within a community is much higher than that of the links between two communities. In general, the performance of predicting the links within a community is also superior to that of all the links in Fig. 8(a). In addition, the performance of GCN is higher than that of CN. In particular, for Precision in Fig. 8(b), the performance of predicting the links within a community is similar to that of all the links, but the performance of the links between two communities is very low. However, the results suggest that community structure information plays a significant role for predicting the links between two communities.

### B. NULL MODEL OF REWIRING EDGES WITHIN A COMMUNITY

The traditional null networks are based on random rewiring, which completely destroy community structures of the original network [64], [65]. The characteristic of community structures is that the density of inner edges is relatively denser than that of external edges between communities. In order to study the influence of inner edges on link prediction, we introduce a new null model, which only modifies the inner structure of each community but maintains the number and structure characteristics of the communities.

The construction process of the null model of random rewiring edges within a community is shown in Fig. 9. Firstly, we divide the original network into multiple communities using a classical community detection algorithm such as Infomap. Secondly, while keeping community structures (external edges between communities) and the whole degree distribution unchanged, we only exchange the inner edges of the communities. For example, we break two edges  $A1-A4$  and  $A2-A3$  in community-A. If the pairs of nodes  $A1-A3$  and  $A2-A4$  are not linked, we connect them. The result of the reconnection is shown in Fig. 9(b). The process of random



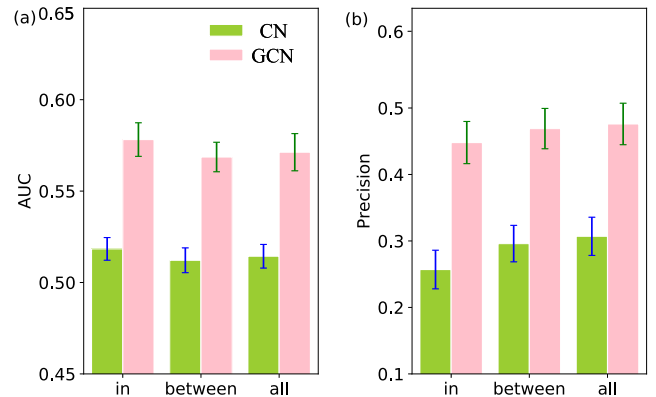
**FIGURE 9.** The process of constructing the null network of random rewiring edges within a community. (a) The original toy network, and (b) its corresponding null network of random rewiring edges within a community.

reconnection should be repeated many times until the null network is completely randomized. At last, we can get a special 1k null network that only destroys the inner topology of each community.

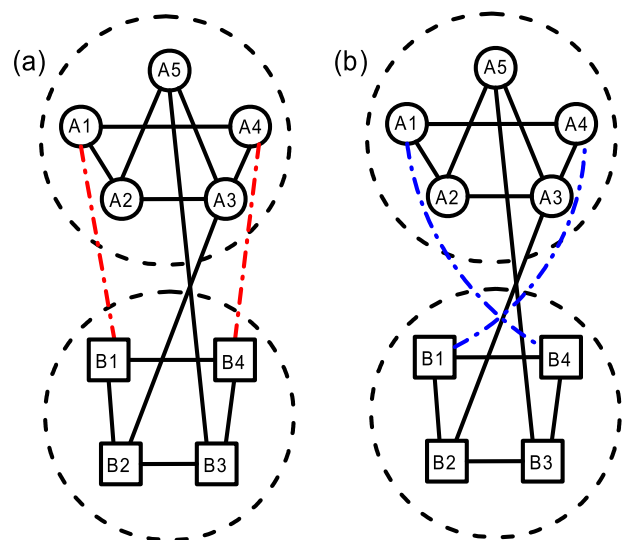
As shown in Fig. 10, for the null model of reconnecting edges within each community, the prediction performance of the links within a community is similar to that of the links between two communities, and that of all the links. In addition, compared with Fig. 8, the performance of predicting the links within a community decreases sharply for both AUC and Precision in Fig. 10. Hence, the result suggest that the original community structure strongly benefits predicting the links within one community. In other words, we confirm that community structures play the crucial role in predicting the links within a community, and these links also have a significant effect on link prediction of the whole network. We have to emphasize that even in this kind of case, the performance of GCN is always much better than that of CN in Fig. 10, which strongly demonstrates the effectiveness and robustness of the proposed algorithm in this study.

**C. NULL MODEL OF REWIRING EDGES BETWEEN COMMUNITIES**

To investigate the impact of external edges between community structures, we introduce another null model that manipulates only the links between two communities but maintains the number of communities and all the links within each community. The process of constructing the null network of random reconnecting edges between two communities is shown in Fig. 11. First, we use the classical community detection algorithm to divide the original network into multiple communities. Second, we only exchange the edges between two communities, keeping the topology within each community



**FIGURE 10.** For AUC and Precision in the null model of randomly rewiring edges within each community, the performance of link prediction for all the links, the links within a community and the links between two communities is achieved under both CN and GCN algorithms.

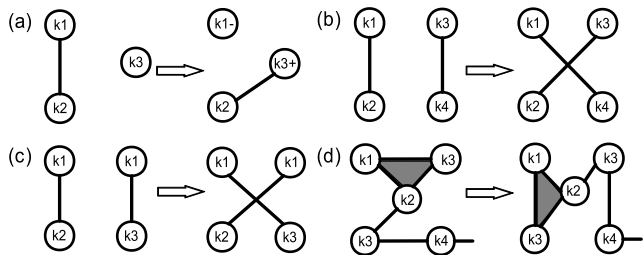


**FIGURE 11.** The process of constructing the null network of random rewiring edges between two communities. (a) The original toy network, and (b) its corresponding null network of random rewiring edges between two communities.

and the degree distribution of the entire network unchanged. For example, we break the two edges A1-B1 and A4-B4. If the pairs of edges A1-B4 and A4-B1 are not linked, we connect them and the reconnection result is shown in Fig. 11(b). The process of random reconnection should be repeated many times, and then we can get a special 1k null network that only destroys the external structures of all the communities. The two types of null models that are depicted in Figs. 9 and 11 can distinguish the role of edges within a community and that of edges between different communities.

For the null network of randomly reconnected edges between communities, the performance of predicting all kinds of the links are similar to that of the original LFR network in Fig. 8, so the relevant results are not shown here. This phenomenon means that the edges between communities





**FIGURE 12.** A toy network and its corresponding different order null models. (a) 0k null model, (b) 1k null model, (c) 2k null model, and (d) 3k null model.

have little impact on link prediction, and further verifies the significance of the links within a community.

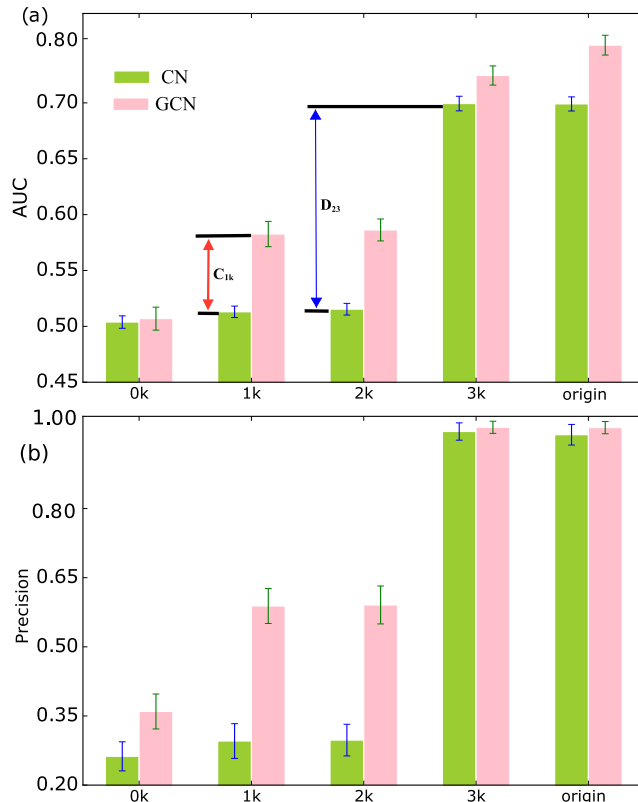
## VI. THE EFFECT OF MICRO-SCALE STRUCTURES

### A. 0k-3k NULL MODELS FOR MICRO-SCALE STRUCTURES

In general, compared with lower-order null models, the structure of a higher-order null model is closer to the topology of the original network [65]. The null models of all these orders are interconnected, i.e.,  $0K \supseteq 1K \supseteq 2K \supseteq 3K$ . Any higher-order null network encompasses the features of lower-order null networks [64]. Here, we briefly introduce the null models of different orders. 0k null model is the simplest but completely randomized version of the original network, which only has the same number of nodes and the same average node degree as the original network in Fig. 12(a).

Compared with 0k model, 1k-3k null models maintain more topology structure characteristics, so they have been widely used in previous studies. As shown in Fig. 12, we describe a toy model of 1k-3k order null models. All the null models are generated by the edge reconnection algorithm, which can randomize all the edges under certain constraints [66], [67]. 1k null model can maintain the degree sequence of the original network and it is depicted in Fig. 12(b). The original network and its correspond 2k null model have the same joint degree distribution, that is, the degree value of each link-end node is the same as that of the original network. The result of 2k null model is depicted in Fig. 12(c). There are two  $k_1$  in it, which mean that the two nodes have the same degree and the value is  $k_1$ . To keep the same joint degree distribution of the original network, the nodes of  $k_2$  and  $k_3$  in the corresponding 2k null network have to possess the same degree value ( $k_1$ ) for the nodes on the end of each link. Finally, the clustering coefficients of all the nodes in the original network and its 3k null model are the same, and at the same time the degree distribution and joint degree distribution are retained. The result of 3k null model is shown in Fig. 12(d).

It is important to note that not all the information of the original network will be randomized in the null models, so we can use them to explore the impact of different micro-scale and meso-scale network characteristics on link prediction. In each real-life network, as the order of the null model increases, the average number of shuffled edges in the cor-



**FIGURE 13.** For AUC and Precision, the performance of CN and GCN algorithms for different order null models of the LFR benchmark network when  $\mu = 0.5$ .

responding null models decreases. It means that the diversity of generated benchmarks is decreasing because fewer edges can be rewired, making the network structures similar to the original network. Therefore, 1k-3k null models can actually lead to different changes in the original real-life networks, providing diversified test networks for performance evaluation of link prediction algorithms.

### B. EXPERIMENTAL RESULTS

0k-3k null models can be utilized to verify the effect of different micro-scale attributes (such as degree distribution, assortativity, and transitivity) on link prediction. First, we use 0k-3k null models to randomize the LFR benchmark network. As shown in Fig. 13, the performance of GCN is always higher than that of CN. Here we define  $C_{1k}$  as the difference between CN and GCN for 1k null model in Fig. 13(a), which quantifies the improvement for the accuracy of link prediction after adding community information. Using this approach, we can detect the impact of community structures on link prediction performance in the null networks of different orders (i.e.,  $C_{0k}$ ,  $C_{1k}$ ,  $C_{2k}$ , and  $C_{3k}$ ).

The higher order null models correspond to the smaller structural variation of the original network. Meanwhile, if we increase the order of null models, the performance of link prediction will be more preferable, which means that the retention of more micro-scale characteristics can allow us

**TABLE 3.** The effect of micro-scale structures on link prediction. We quantify the effect of community structures on link predicting on different orders of null networks.  $C_{0k}$  is the performance difference between CN and GCN for 0k null model of the LFR benchmark network,  $C_{1k}$  is the performance difference for 1k null model,  $C_{2k}$  is the performance difference for 2k null model, and  $C_{3k}$  is the performance difference for 3k null model. At the same time, we also quantify the effect of different order micro-scale structures on link prediction.  $D_{01}$  is the performance difference of CN between 0k and 1k null models,  $D_{12}$  is the performance difference between 1k and 2k null models,  $D_{23}$  is the performance difference between 2k and 3k null models,  $D_{3o}$  is the performance difference between 3k null model and the original LFR network.

Statistic	$C_{0k}$	$C_{1k}$	$C_{2k}$	$C_{3k}$	$D_{01}$	$D_{12}$	$D_{23}$	$D_{3o}$
$\Delta$ AUC%	0.6	<b>13.6</b>	<b>13.8</b>	3.6	1.8	0.5	<b>35.7</b>	0.1
$\Delta$ Precision%	37.1	<b>99.1</b>	<b>98.5</b>	2.6	12.7	0.7	<b>224.9</b>	0.7

to improve the prediction performance. We attempt to measure the effect of different order microscopic characteristics on link prediction of a specific network. For instance,  $D_{23}$  represents the performance difference of CN between 2k and 3k null models. In this way, we can quantify the effect of different order micro-scale structures on link prediction ( $D_{01}$ ,  $D_{12}$ ,  $D_{23}$ , and  $D_{3o}$ ). Specifically, as shown in Fig. 13(b), the performance of 3k null model is similar to that of the LFR benchmark network. Furthermore, the link prediction accuracy grows sharply when the order of null model increases to 3k. Hence, for link prediction, maintaining the 3k micro-scale property (i.e., transitivity) can achieve the same function of the original network.

In brief, based on the results of Table 3, we can distinguish the different roles of micro-scale and meso-scale structures on link prediction, and reveal the relationship between them. Moreover, the effect of community structures on link prediction can be quantified for different orders of null networks. For example,  $C_{1k}$  is the performance difference between CN and GCN for 1k null model of the LFR benchmark network, and  $C_{2k}$  is the performance difference for 2k null model. In these two kinds of micro-scale null networks, the information of community structures has a strong positive effect, because it can greatly improve prediction performance as shown in Table 3.

At the same time, the results in Table 3 can quantify the effect of different order micro-scale structures on link prediction. For instance,  $D_{23}$  is the performance difference for CN between 2k and 3k null models, which shows that the transitivity characteristic is a significant structural feature that affects link prediction. Actually, this framework can also be used to distinguish different roles of micro-scale and meso-scale structures on other applications in complex networks, such as social recommendation [12], detecting spurious links [2] and node classification [17].

## VII. CONCLUSION

In summary, we proposed a succinct algorithm that is based on extended community common neighbors to improve the performance of link prediction. More importantly, we firstly

quantified the role of community structures for link prediction using the proposed null models. Comparing the original network to null models, our results suggest that the intensity of community structures plays an important role for link prediction. Especially, the higher modularity value leads to more clear community structures, hence it can give rise to a higher accuracy of link prediction. Even only detecting more structure information of the original network by an efficient community detection algorithm, can achieve a higher prediction performance. Our findings have been verified by both of synthetic benchmarks and real-life networks.

We also distinguished the role of links in the same community and that between two communities. Our study uncovers that community information plays a crucial role for predicting links within one community, and links within one community also have a strong impact on predicting all the links, especially on the links between two communities. Our study also provides a start point for detailed analysis for the correlation of community and micro-scale structures by constructing different kinds of null models. In particular, it reveals the relationship and the dependence between this special meso-scale structure (community) and micro-scale structures of different orders (i.e., degree distribution, assortativity, and transitivity) for link prediction. In future, we try to extend the proposed framework to signed networks [68], and will pay attention to applying the information obtained by fuzzy overlapping communities [69] for link prediction.

## REFERENCES

- [1] S. Redner, "Teasing out the missing links," *Nature*, vol. 453, no. 7191, pp. 47–48, May 2008.
- [2] A. Zeng and G. Cimini, "Removing spurious interactions in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 85, no. 3, Mar. 2012, Art. no. 036101.
- [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] H. Yu, "High-quality binary protein interaction map of the yeast interactome network," *Science*, vol. 322, no. 5898, pp. 104–110, Oct. 2008.
- [5] M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf, "Estimating the size of the human interactome," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 19, pp. 6959–6964, May 2008.
- [6] L. A. N. Amaral, "A truer measure of our ignorance," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 19, pp. 6795–6796, May 2008.
- [7] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Social Netw.*, vol. 25, no. 3, pp. 211–230, Jul. 2003.
- [8] Q.-M. Zhang, X.-K. Xu, Y.-X. Zhu, and T. Zhou, "Measuring multiple evolution mechanisms of complex networks," *Sci. Rep.*, vol. 5, no. 1, Sep. 2015, Art. no. 10350.
- [9] K.-K. Shang, M. Small, X.-K. Xu, and W.-S. Yan, "The role of direct links for link prediction in evolving networks," *EPL Europhysics Lett.*, vol. 117, no. 2, p. 28002, Jan. 2017.
- [10] Z. Wang, J. Liang, and R. Li, "Exploiting user-to-user topic inclusion degree for link prediction in social-information networks," *Expert Syst. Appl.*, vol. 108, pp. 143–158, Oct. 2018.
- [11] K.-K. Shang, W.-S. Yan, and M. Small, "Evolving networks—Using past structure to predict the future," *Phys. A, Stat. Mech. Appl.*, vol. 455, pp. 120–135, Aug. 2016.
- [12] Z.-K. Zhang, C. Liu, Y.-C. Zhang, and T. Zhou, "Solving the cold-start problem in recommender systems with social tags," *EPL Europhysics Lett.*, vol. 92, no. 2, p. 28002, Oct. 2010.
- [13] S. Hong, H. Yang, E. Zio, and N. Huang, "A novel dynamics model of fault propagation and equilibrium analysis in complex dynamical communication network," *Appl. Math. Comput.*, vol. 247, pp. 1021–1029, Nov. 2014.

- [14] S. Hong, B. Wang, X. Ma, J. Wang, and T. Zhao, "Failure cascade in interdependent network with traffic loads," *J. Phys. A, Math. Theor.*, vol. 48, no. 48, Dec. 2015, Art. no. 485101.
- [15] S. Hong, C. Lv, T. Zhao, B. Wang, J. Wang, and J. Zhu, "Cascading failure analysis and restoration strategy in an interdependent network," *J. Phys. A, Math. Theor.*, vol. 49, no. 19, May 2016, Art. no. 195101.
- [16] S. Hong, J. Zhu, L. A. Braunstein, T. Zhao, and Q. You, "Cascading failure and recovery of spatially interdependent networks," *J. Stat. Mech., Theory Exp.*, vol. 2017, no. 10, 2017, Art. no. 103208.
- [17] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Stat. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, Mar. 2011.
- [18] Z.-M. textcolor red Ren, A. Zeng, and Y.-C. Zhang, "Structure-oriented prediction in complex networks," *Phys. Rep.*, vol. 750, pp. 1–50, Aug. 2018.
- [19] H. Cai, V. W. Zheng, and K. C.-C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1616–1637, Sep. 2018.
- [20] L. Lü and T. Zhou, "Link prediction in weighted networks: The role of weak ties," *EPL Europhysics Lett.*, vol. 89, no. 1, p. 18001, Jan. 2010.
- [21] 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.
- [22] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.
- [23] Y. Liu, C. Zhao, X. Wang, Q. Huang, X. Zhang, and D. Yi, "The degree-related clustering coefficient and its application to link prediction," *Phys. A, Stat. Mech. Appl.*, vol. 454, pp. 24–33, Jul. 2016.
- [24] R.-M. Cao, S.-Y. Liu, and X.-K. Xu, "Network embedding for link prediction: The pitfall and improvement," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 29, no. 10, Oct. 2019, Art. no. 103102.
- [25] B. Yan and S. Gregory, "Finding missing edges in networks based on their community structure," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 85, no. 5, May 2012, Art. no. 056112.
- [26] 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, Dec. 2013, Art. no. 1613.
- [27] J. Ding, L. Jiao, J. Wu, Y. Hou, and Y. Qi, "Prediction of missing links based on multi-resolution community division," *Phys. A, Stat. Mech. Appl.*, vol. 417, pp. 76–85, Jan. 2015.
- [28] J. Ding, L. Jiao, J. Wu, and F. Liu, "Prediction of missing links based on community relevance and ruler inference," *Knowl.-Based Syst.*, vol. 98, pp. 200–215, Apr. 2016.
- [29] Z. Wang, Y. Wu, Q. Li, F. Jin, and W. Xiong, "Link prediction based on hyperbolic mapping with community structure for complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 450, pp. 609–623, May 2016.
- [30] 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.
- [31] F. Li, J. He, G. Huang, Y. Zhang, Y. Shi, and R. Zhou, "Node-coupling clustering approaches for link prediction," *Knowl.-Based Syst.*, vol. 89, pp. 669–680, Nov. 2015.
- [32] B. Liu, S. Xu, T. Li, J. Xiao, and X.-K. Xu, "Quantifying the effects of topology and weight for link prediction in weighted complex networks," *Entropy*, vol. 20, no. 5, p. 363, 2018.
- [33] W.-K. Cui, K.-K. Shang, Y.-J. Zhang, J. Xiao, and X.-K. Xu, "Constructing null networks for community detection in complex networks," *Eur. Phys. J. B*, vol. 91, no. 7, p. 145, Jul. 2018.
- [34] K.-K. Shang, M. Small, and W.-S. Yan, "Link direction for link prediction," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 767–776, Mar. 2017.
- [35] V. textcolor red Rebaza, J. Carlos, and A. Lopes, "Link Prediction in Complex Networks Based on Cluster Information," in *Proc. SBIA*, 2012, pp. 92–101.
- [36] H. A. Deylami and M. Asadpour, "Link prediction in social networks using hierarchical community detection," in *Proc. 7th Conf. Inf. Knowl. Technol. (IKT)*, May 2015, pp. 1–5.
- [37] H. Jeon and T. Kim, "Community-adaptive link prediction," in *Proc. Int. Conf. Data Mining, Commun. Inf. Technol. (DMCIT)*, 2017, pp. 1–5.
- [38] J. Wang, Y. Ma, M. Liu, H. Yuan, W. Shen and L. Li, "A vertex similarity index using community information to improve link prediction accuracy," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2017, pp. 158–163.
- [39] J. Wang, Y. Ma, M. Liu, and W. Shen, "Link prediction based on community information and its parallelization," *IEEE Access*, vol. 7, pp. 62633–62645, 2019.
- [40] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [41] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, Jan. 2004.
- [42] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, no. 4, Oct. 2008, Art. no. 046110.
- [43] 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.
- [44] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropol. Res.*, vol. 33, no. 4, pp. 452–473, Dec. 1977.
- [45] J. H. Michael and J. G. Massey, "Modeling the communication network in a sawmill," *Forest Products J.*, vol. 47, no. 9, pp. 25–30, 1997.
- [46] 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. Ecology Sociobiol.*, vol. 54, no. 4, pp. 396–405, Sep. 2003.
- [47] 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, Jun. 2002.
- [48] 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*, 2005, pp. 36–43.
- [49] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*. New York, NY, USA: ACM Press, 1993.
- [50] K. Judd, M. Small, and T. Stemler, "What exactly are the properties of scale-free and other networks?" *EPL Europhys. Lett.*, vol. 103, no. 5, p. 58004, Sep. 2013.
- [51] 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.
- [52] K.-K. Shang and X.-K. Xu, "Construction and application for null models of complex networks based on randomized algorithms," *J. Univ. Electron. Sci. Technol. China*, vol. 43, pp. 7–20, Oct. 2014.
- [53] J. P. Bagrow, "Evaluating local community methods in networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 5, 2008, Art. no. P05001.
- [54] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, Dec. 2004, Art. no. 066111.
- [55] R. Ghosh and K. Lerman, "Structure of heterogeneous networks," in *Proc. Int. Conf. Comput. Sci. Eng.*, 2009, pp. 98–105.
- [56] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.
- [57] Z. Zhao, C. Li, X. Zhang, F. Chiclana, and E. H. Viedma, "An incremental method to detect communities in dynamic evolving social networks," *Knowl.-Based Syst.*, vol. 163, pp. 404–415, Jan. 2019.
- [58] R. L. Breiger, S. A. Boorman, and P. Arabie, "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling," *J. Math. Psychol.*, vol. 12, no. 3, pp. 328–383, Aug. 1975.
- [59] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. Graph Algorithms Appl.*, vol. 10, no. 2, pp. 191–218, 2006.
- [60] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," 2005, *arXiv:physics/0506133*. [Online]. Available: <https://arxiv.org/abs/physics/0506133>
- [61] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 4, pp. 1118–1123, Jan. 2008.
- [62] J. Xie and B. K. Szymanski, "Towards linear time overlapping community detection in social networks," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, Berlin, Germany: Springer, 2012, pp. 25–36.
- [63] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006.
- [64] P. Mahadevan, C. Hubble, D. Krioukov, B. Huffaker, and A. Vahdat, "Orbis: Rescaling degree correlations to generate annotated Internet topologies," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 37, no. 4, pp. 325–336, Oct. 2007.

- [65] C. Orsini, M. M. Dankulov, P. Colomer-de-Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and D. Krioukov, "Quantifying randomness in real networks," *Nature Commun.*, vol. 6, no. 1, p. 8627, Dec. 2015.
- [66] S. Maslov, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, pp. 910–913, May 2002.
- [67] P. Holme, "Network reachability of real-world contact sequences," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 71, no. 4, Apr. 2005, Art. no. 046119.
- [68] C. Liu, J. Liu, and Z. Jiang, "A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2274–2287, Dec. 2014.
- [69] J. Su and T. C. Havens, "Quadratic program-based modularity maximization for fuzzy community detection in social networks," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 5, pp. 1356–1371, Oct. 2015.



**KE-KE SHANG** received the Ph.D. degree from the School of Marine Science and Technology, Northwestern Polytechnical University, in 2017. He is currently an Assistant Researcher with the Computational Communication Collaboratory, School of Journalism and Communication, Nanjing University. His current research interest includes network science.



data mining on complex networks.

**XIAO-KE XU** (Member, IEEE) received the Ph.D. degree from the College of Information and Communication Engineering, Dalian Maritime University, in 2008. He was a Postdoctoral Fellow with Hong Kong Polytechnic University and a Visiting Scholar with the City University of Hong Kong. He is currently a Professor with the College of Information and Communication Engineering, Dalian Minzu University. His current research interests include community detection, link predicting, and



**JING XIAO** received the Ph.D. degree in signal and information processing from Harbin Engineering University, Harbin, China, in 2012. She received postdoctoral training in automation from Harbin Engineering University, in 2018. She is currently a Lecturer of information and communication engineering with Dalian Minzu University, Dalian, China. Her current research interests include network community detection, swarm intelligence computing, and multi-objective optimization.

...