Hindawi
Complexity
Volume 2020, Article ID 7307058, 17 pages
https://doi.org/10.1155/2020/7307058

Research Article

W-Index: An Index for Evaluating Link Prediction considering Only the Role of Wins

Yun Yuan, Jingwei Wang, Yunlong Ma, and Min Liu

School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China

Correspondence should be addressed to Yunlong Ma; evanma@tongji.edu.cn and Min Liu; lmin@tongji.edu.cn

Received 2 January 2020; Revised 15 September 2020; Accepted 5 November 2020; Published 8 December 2020

Academic Editor: Hens Chittaranjan

With the emergence of numerous link prediction methods, how to accurately evaluate them and select the appropriate one has become a key problem that cannot be ignored. Since AUC was first used for link prediction evaluation in 2008, it is arguably the most preferred metric because it well balances the role of wins (the testing link has a higher score than the unobserved link) and the role of draws (they have the same score). However, in many cases, AUC does not show enough discrimination when evaluating link prediction methods, especially those based on local similarity. Hence, we propose a new metric, called W-index, which considers only the effect of wins rather than draws. Our extensive experiments on various networks show that the W-index makes the accuracy scores of link prediction methods more distinguishable, and it can not only widen the local gap of these methods but also enlarge their global distance. We further show the reliability of the W-index by ranking change analysis and correlation analysis. In particular, some community-based approaches, which have been deemed effective, do not show any advantages after our reevaluation. Our results suggest that the W-index is a promising metric for link prediction evaluation, capable of offering convincing discrimination.

1. Introduction

Link prediction is one of the most fundamental problems of complex networks, which aims to infer the network link formation process by predicting missed or future relationships based on currently observed links [1]. Being able to effectively and efficiently predict unobserved links will allow us to mine future interactions among members in the social network [2, 3], conduct successful recommendation in user-item bipartite networks [4, 5], provide guidance for the planning process of infrastructure systems [6], optimize asset allocation in the stock market [7], discover drug-target interactions or identify targeting drugs for new target-candidate proteins [8, 9], and save costs and reduce operational risks for manufacturing companies [10, 11]. Up to the present, various methods have been proposed for link prediction [12–16], most of which can be classified with heuristic-based approaches and learning-based approaches [17]. However, there are many long-standing challenges in the evaluation of link prediction methods.

Many quantitative evaluation metrics used in link prediction are adopted from binary classification tasks [18]. They can be divided into two broad categories: fixed-threshold metrics and threshold curves [19, 20]. As a typical fixed-threshold metric, the precision [21] is used commonly in link prediction literature studies. However, like other fixed-threshold metrics, it suffers from the limitation that it is difficult to select an appropriate threshold in the score space. In other words, the precision only focuses on the L links with the top ranks or the highest scores. Thus, the accuracy of link prediction varies according to the choice of L. Besides, Clauset et al. [22] presented that using the precision to evaluate prediction algorithms has a significant disadvantage. That means the precision may be high, considering the top L links with the highest scores, whereas an algorithm's overall performance is unsatisfactory because some missing connections are much easier to predict than others. For instance, if a network has a heavy-tailed degree distribution, the chances are excellent that two high-degree vertices have a missing connection, and such a connection can be easily predicted.

Due to these problems with fixed-threshold metrics, it is recommended to use threshold curves as an alternative [19, 23]. Threshold curves are especially suitable for class

imbalance tasks and hence are used increasingly commonly in link prediction evaluation [24–26]. AUC, the area under the receiver operating characteristic (ROC) curve [27], is a standard metric in the evaluation of link prediction. AUC evaluates a method's performance according to the whole list of all unobserved links and testing links. It is equivalent to the probability of a randomly selected testing link appearing above a randomly selected unobserved link in the score space. With the use of the AUC metric, the severe class imbalance problem in link prediction is largely mitigated, i.e., the number of testing links is much smaller than the number of unobserved links.

Although AUC is a good measure for link prediction evaluation [28], it has certain shortcomings. In the literature of data mining and machine learning [29–31], an important drawback is that the AUC measure ignores the scores of instances and only considers their ranking order. The result is that it is unreliable because the difference between scores of instances is ignored. Another consequence is the AUC can remain unchanged when the predicted scores change, as long as the ranking of instances remains the same. More importantly, it has been found that when using the AUC for link prediction evaluation, the discrimination between some methods is poor, especially for those neighborhood-based predictors, such as common neighbors (CN), Jaccard, and Sørensen (Sor) [1, 32]. Here, we argue poor discrimination is an inherent defect of the AUC because draws are considered in the calculation of the AUC. According to the AUC index, if a test link and an unobserved link are given the same score, they are in a draw, with a score of 0.5, but draws are, of course, less important than wins and losses for evaluation, and draws mean undiscriminating. Therefore, an intuitive idea is that we should pay more attention to wins and losses rather than draws. Even if the difference in scores is small, the method that assigns different scores to a testing link and an unobserved link is better than giving them the same score.

Briefly speaking, we deem it is the wins rather than the draws that count. The purpose of this paper is to present a new metric, called W-index, which only cares about who wins more between the testing link and the unobserved link, not about how many times they draw, to obtain discriminative evaluation of link prediction methods.

The rest of this paper is organized as follows. In Section 2, we briefly review commonly used evaluation metrics of link prediction and detail the proposed W-index metric. Link prediction problem and various predictors are illustrated in Section 3. Section 4 presents experimental results on six real-world networks followed by the discussion of the W-index in Section 5. Finally, we conclude the work in Section 6.

2. W-Index

In this section, we first introduce two widely used evaluation metrics, i.e., the precision and AUC. Then, we present a novel evaluation metric named W-index which only considers the number of wins that the testing link has a higher score than the unobserved link.

2.1. Precision. Given the ranking of all unobserved links and testing links, the precision is defined as the ratio of relevant links selected to the number of links selected. That is to say, if we take the top-L links as the predicted ones, among which m links are right (i.e., there are m links in the testing set E^{Test}), then the precision value is given by

$$\text{precision} = \frac{m}{L}. \tag{1}$$

Higher precision means higher prediction accuracy.

2.2. AUC. Given the rank of all unobserved links and testing links, the AUC value can be viewed as the probability that a randomly chosen testing link (a link in E^{Test}) is given a higher score than a randomly chosen unobserved link (a link in E^{Non}). Considering the computational complexity of large-scale networks, we usually implement sampling experiments to calculate this value. If, among n-times independent experiments, there are n' times the testing link having a higher score than the unobserved link and n'' times they have the same score, the AUC value is given by

$$\text{AUC} = \frac{n' + 0.5n''}{n}. \tag{2}$$

Higher precision means higher prediction accuracy. If all scores are generated from independent and identical distributions, then the AUC value should be approximately 0.5. Therefore, the degree to which the AUC value exceeds 0.5 indicates how much better an algorithm performs than a pure chance. Both the precision and the AUC metric are considered in most recent studies due to their different focus. If two link prediction methods have the same AUC score, the one with a higher precision score is considered better.

2.3. W-Index. In link prediction evaluation, draws are common. We consider draws as a byproduct of the "state degradation" problem, which was found by Lu et al. [32, 33], in many link prediction algorithms, the aforementioned neighborhood-based predictors in particular. The "degeneracy of the states" problem is that two node pairs are of high probability to be assigned the same similarity scores. This is because the state of the structural information is finite and is less distinguishable, especially for the unobserved links with little information, so the unobserved links are highly likely to obtain the same similarity score. As for some testing links connected by two low-degree nodes, they also have little information. Thus, the difference between scores of some testing links and the unobserved link is not obvious, and draws are not uncommon. Taking the Metabolic network [34] as an example, there are more than 10^5 unobserved node pairs, 62% of which are assigned score 0 by the CN algorithm. For all unobserved node pairs having scores higher than 0, 68% score 1, and 21% score 2. And the testing links with scores of 0, 1, and 2 account for 3%, 12%, and 28% of the total testing links, respectively. Therefore, when comparing the score of a randomly selected testing link with that of a randomly selected unobserved link, the possibility

of a draw exceeds 7%. Actually, in the experiments implemented in Section 4, there are more than 750 draws out of 10,000 independent experiments. Thereby, for the evaluation result, draws matter.

Here, we consider the side effects of draws in Figure 1 on three cases. In the following analysis, it is supposed that the A predictor outperforms the B predictor under the AUC metric. Case 1 shows that both the number of wins and the number of draws of the A predictor far exceed the corresponding number of the B predictor. In this case, ignoring the draws may widen or narrow the performance difference between the two predictors, depending on the difference in the number of draws between the two predictors.

In case 2, the number of draws of the B predictor is much larger than that of the A predictor, but the number of wins of B is less than that of A. If we do not reward 0.5 points for a draw, it yields a better distinction between the two predictors. Therefore, a side effect of the draws is that it narrows the difference in accuracy between the two predictors, making the predictors less distinguishable. In most applications, finding out a predictor that performs remarkably better than others is nontrivial. Taking protein-protein interaction networks as an example, they can reduce the experimental costs and speed the pace of uncovering the truth [35, 36]. Thus, a good link prediction evaluation metric needs to be not only effective but also discriminative.

To make matters worse, draws can be misleading, as shown in case 3. Predictor A with more draws is considered better than predictor B with clear wins and losses. Assuming in one case, using predictor A, the number of times that the similarity score of the testing link is higher than, equal to, and lower than the unobserved link makes up 88%, 6%, and 6% of the total times, respectively. The corresponding three proportions in the B predictor are 90%, 0.8%, and 9.2%, respectively. Then, the AUC score of predictor A is 0.91, and that of predictor B is 0.90, which means that A is better than B. However, predictor B wins more times than predictor A. It is doubtful whether predictor A is really better than predictor B.

To alleviate the two side effects of the draws, we propose a new metric, called W-index, for evaluating link prediction, which depends only on the number of wins, regardless of the number of draws. It is defined as

$$\text{W} - \text{index} = \frac{n'}{n}. \tag{3}$$

Apparently, the value of the W-index ranges from 0 to 1. The new scoring criteria may reduce the impact of the "degeneracy of the states" and make the accuracy scores more discriminative.

3. Link Prediction Method

In this section, we mainly depict the basic definitions and related concepts about the link prediction problem and then introduce ten link prediction predictors employed later.

3.1. Definitions. Consider an undirected network $G(V, E)$, where V is the set of nodes and E is the set of links. Multiple links and self-connections are not allowed. Consider the universal set, denoted by U, containing all $(|V|(|V|-1))/2$ possible links between the pair of nodes in V, where $|V|$ denotes the number of nodes in V. The basic task of link prediction is to find out the missing links (or the links that will appear in the future) in the unobserved set $U - E$.

Generally, we do not know which links are the missing or future links; otherwise, we do not need to do prediction. Therefore, to test the algorithm's accuracy, the observed links, E, are randomly divided into two parts: the training set, E^{Train}, is treated as known information, while the testing set, E^{Test}, is used for testing, and no information in this set is allowed to be used for prediction. The set of unobserved links is E^{Non}, which is equal to the set $U - E$. The set of links to be validated is E^P. Clearly, $E^{\text{Train}} \cup E^{\text{Test}} = E$ and $E^{\text{Train}} \cap E^{\text{Test}} = \varnothing$ and $E^{\text{Test}} \cup E^{\text{Non}} = E^P$ and $E^{\text{Test}} \cap E^{\text{Non}} = \varnothing$.

Each link in the set $U - E^{\text{Train}}$, say $(x, y) \in (x, y) \in U - E^{\text{Train}}$, where $x, y \in V$ are a pair of disconnected vertices, is assigned a score $S_{x,y}$ to quantify its existence likelihood by one link prediction method, i.e., the score $S_{x,y}$ measures the similarity between nodes x and y. Thus, the likelihood connected with nodes x and y is increasing as the score raises and vice versa.

3.2. Benchmark Prediction Algorithm. The simplest framework of link prediction algorithms is the similarity-based algorithm. Due to the high computational complexity of global similarity predictors [32], we only select predictors based on local and quasi-local similarity in our experiments. The definitions of these predictors are shown in Table 1 in detail.

3.2.1. Local Similarity Predictors. Here, we consider three classical predictors based on local information: common neighbors (CN), Leicht–Holme–Newman (LHN) index [37], and resource allocation (RA) [38]. Let $k(x)$ be the degree of node x, $\Gamma(x)$ be the neighbor set of node x, $|\cdot|$ be the cardinality of the set, and $\Lambda_{x,y} = \Gamma(x) \cap \Gamma(y)$ be the set of common neighbors of the pair of unconnected nodes (x, y).

Considering that every common neighbor contributes differently to the connection likelihood, some predictors based on community information are proposed. Consequently, we must apply a clustering scheme to the graph before computing these predictors. Here, we select five local similarity predictors based on community information predictors: WIC [39], intracommunity-based resource allocation (ICRA) [40], and the other three are the W-forms of CN, LHN, and RA [41], that is to say, they are CN-W, LNH-W, and RA-W. Note that x^c is vertex $x \in V$ belonging to a cluster with label C, $\Lambda^W_{x,y} = \left\{z \in \Lambda_{x,y} | x^C, y^C, z^C\right\}$ is the set of within-cluster (W) common neighbors, $\Lambda^{\text{IC}}_{x,y} = \Lambda_{x,y} / \Lambda^W_{x,y}$ is the set of intercluster (IC) common neighbors, and δ is a small value constant close to zero.

3.2.2. Quasi-Local Similarity Predictors. Quasi-local similarity predictors take consideration of local paths that

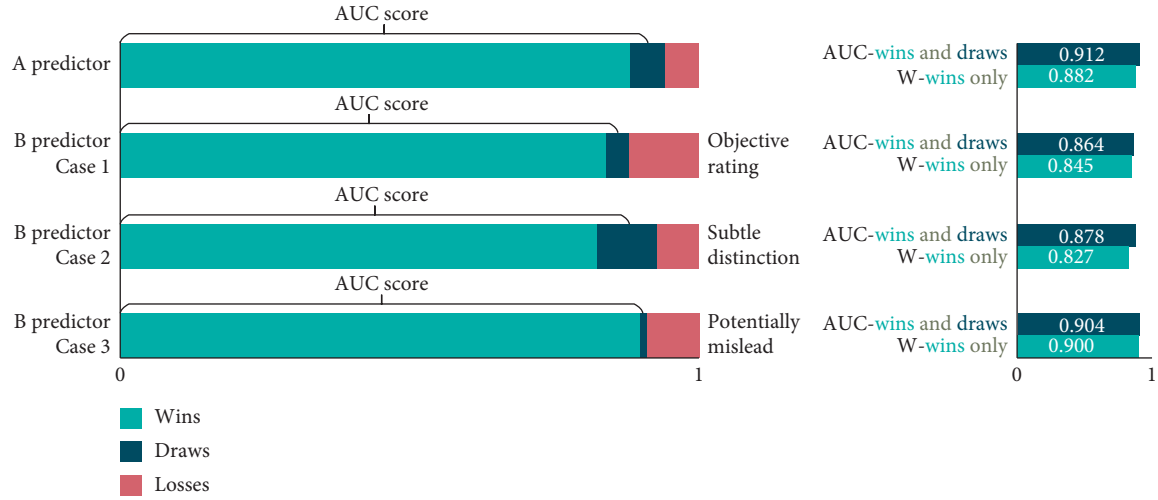

Figure 1: Two side effects of the draws for the evaluation of link prediction methods.

Table 1: Similarity-based algorithm.

Predictors	Equation
CN	$S_{x,y} = \lvert\Lambda_{x,y}\rvert$
LHN	$S_{x,y} = (\lvert\Lambda_{x,y}\rvert)/(k(x)k(y))$
RA	$S_{x,y} = \sum_{z\in\Lambda_{x,y}} (1/(k(z)))$
CN-W	$S_{x,y} = \lvert\Lambda^{W}_{x,y}\rvert$
LHN-W	$S_{x,y} = (\lvert\Lambda^{W}_{x,y}\rvert)/(k(x)k(y)$
RA-W	$S_{x,y} = \sum_{z\in\Lambda^{W}_{x,y}} (1/k(z))$
WIC	$S_{x,y} = \lvert\Lambda^{W}_{x,y}\rvert/(\lvert\Lambda^{IC}_{x,y}\rvert + \delta)$
ICRA	$S_{x,y} = ((\lvert\Lambda^{W}_{x,y}\rvert/\lvert\Lambda_{x,y}\rvert)) + \sum_{z\in\Lambda_{x,y}} ((\max\{k(x), k(y)\})/k(z)$
LP	$S = A^2 + \alpha A^3$
LRW	$S_{x,y}(t) = q_x\pi_{xy}(t) + q_y\pi_{yx}(t)$

provide a little bit more information than local ones. Here, we consider two predictors: local path (LP) [33, 38] and local random walk (LRW) [41]. Note that A is the adjacency matrix, α is a free parameter, $\pi_{xy}(t+1) = P^T\pi_{xy}(t)$, where $\pi_{xy}(0) = e_x$ and P is the transition matrix with $P_{xy} = 1/k_x$ if x and y are connected, and $P_{xy} = 0$, otherwise, and q is the initial configuration function; here, we apply a simple form determined by the node degree, namely, $q_x = (k_x/2|E|)$.

4. Experiments and Results

In this section, we implement experiments to validate the proposed W-index metric on six real-world networks from different fields. Firstly, we explore the prediction accuracy of each predictor under AUC and W-index. Next, we verify the accuracy and stability of the W-index through ranking change analysis and typical correlation analysis. Thirdly, we analyze the local and global distances between predictor performance. We further rethink the selection of link prediction methods from the perspective of W-index evaluation. Finally, we figure out the impact of the network structure and training set length on the performance of these predictors under W-index and AUC.

4.1. Datasets. We consider six representative real-world networks from typical domains of network science, including collaboration, transportation, biological, web, and social networks.

Note that all the similarity predictors considered here will give score 0 to a pair of nodes located in two disconnected components. Therefore, we do not consider those isolated nodes, and all the networks above are strong-connected. The detailed structural features of these real-world networks are shown in Table 2, where $|E|$ is the number of edges, k is the average degree of the network, p is the average short path length of the network, C is the clustering coefficient of the network, and r is the degree of the assortativity coefficient of the network.

(1) Jazz: a collaboration network between jazz musicians. Each node is a jazz musician, and an edge denotes that two musicians have played together in a band [42].

(2) USAir: a network of flights among the commercial airports in the United States [43].

(3) Metabolic: a metabolic network of the nematode *C. elegans* [35].

Table 2: Structural features and basic information of six real-world networks.

Networks	$\|V\|$	$\|E\|$	k	p	C	r
Jazz	198	2742	27.70	2.24	0.62	0.02
USAir	332	2126	12.81	2.74	0.63	−0.21
Metabolic	453	2025	8.94	2.66	0.65	−0.22
PB	1222	16714	27.36	2.74	0.32	−0.22
Tvshow	3892	17239	8.86	6.28	0.37	0.56
Yeast	6008	156945	52.25	2.54	0.17	−0.08

(4) PB: a network of the US political blogs during the 2004 US election. A node represents a blog, and an edge represents a hyperlink between two blogs. The original edges are directed; here, we treat them as undirected ones [44].

(5) Tvshow: a social network of Facebook pages about television shows. Nodes represent the pages, and edges are mutual likes among them [45].

(6) Yeast: a biology network of thousands of interactions between proteins [46].

4.2. Experimental Settings. To evaluate the validity of our evaluation metric, the set of observed edges, E, is randomly divided into two parts in each division: the training set E^{Train} and the testing set E^{Test}; $|E^{\text{Train}}|$: $|E^{\text{Test}}| = 9:1$. For each network, we use the training set to generate communities using the Louvain community detection algorithm [47]. Then, we use the W-index, AUC, and the precision metric to measure the prediction accuracy of these predictors mentioned in Table 2. In our experiments, we keep $n = 10000$ times independent experiments in the W-index and AUC and $L = E^{\text{Test}}$ in the precision metric for all networks. Each value is obtained by averaging over 20 implementations with independently random divisions of the training set and the testing set. Note that we choose $\alpha = 0.01$ for LP according to Lu and Zhou [32] and $t = 3$ for LRW due to the limitation of the running time.

4.3. Prediction Accuracy. In Table 3, we present the prediction accuracy results measured by AUC and W-index on the six networks. The entries corresponding to the highest accuracy for each network are marked in bold. Then, we calculate the difference in accuracy for each predictor under AUC and W-index, which is shown in Figure 2. Overall speaking, Table 3 and Figure 2 reveal that using the W-index metric does affect the performance of these predictors, and the accuracy scores of all predictors are reduced. However, the accuracy of the same predictor in different networks varies greatly. We find that the sparser the network is, the lower the accuracy evaluated by the W-index, which shows that the performance of a predictor is largely influenced by the network structure. For example, both average degree and clustering coefficient of Tvshow are extremely small, so the accuracy scores of all predictors are extremely low, while the Jazz network is the opposite.

4.4. Reliability of the W-Index. Whether the W-index can maintain reliability in different contexts is an important issue. We exploit two common methods based on empirical data for comparative analysis to verify the accuracy and stability of the W-index. Firstly, Figure 2 shows that, under the W-index metric, the performance ranking of these predictors in each network is almost the same as their ranking under the AUC metric. This means that if AUC is considered to be able to effectively evaluate the pros and cons of predictors, W-index has the same ability. Next, we compare the correlation coefficient between W-index and precision (see Figure 3) with that of AUC and precision on six networks. From Table 4, it can be seen that the correlation coefficient between W-index and precision is closely similar to that of AUC and precision in all six networks. Moreover, if we do not consider the CN-W, LHN-W, RA-W, and WIC predictors due to their poor performance, the correlation coefficient between W-index and precision is almost the same as that of AUC and precision (see Table 5). The consistency of correlation coefficient changes in different contexts indicates that the W-index has a similar evaluation effect as AUC. The above two comparative analyses can demonstrate the reliability of the W-index.

4.5. Performance of the W-Index. In link prediction, we argue evaluation serves two purposes. One is to quantify the performance of algorithms, which is called absolute evaluation. The other is to quantify the extent to which one predictor is better than another, which is called relative evaluation. Here, we use the prediction accuracy as the absolute evaluation score, e.g., a score of 1 measured by the AUC means a perfect predictor, and a score of 0.5 measured by the AUC means the predictor is not better than a pure chance. Besides, we use prediction accuracy differences between two predictors as the relative evaluation score, e.g., under the AUC metric, the accuracy of RA is 0.97 and that of CN is 0.95, and then the relative evaluation score is 0.02.

Generally, using the W-index metric, there is a larger separability between evaluated methods on these six real-world networks than using the AUC metric, which is caused by the following two reasons. Firstly, when measured by the W-index, the absolute evaluation score of the highest accuracy for each network is still high enough. In other words, there is no significant decrease in the highest score measured by the W-index compared with the AUC. Specifically, the difference of the highest accuracy does not exceed 0.01 on four of six networks. Secondly, when measured by the W-index, the relative evaluation score of the highest accuracies for each network becomes higher. That is to say, the prediction accuracy differences between different predictors measured by the W-index are larger than those measured by the AUC. For example, the prediction accuracy difference between RA and CN is 0.018 on Jazz under the AUC, while it is 0.024 under the W-index. Therefore, the W-index metric can better distinguish the performance of these link prediction methods.

Furthermore, we discuss the total distance between the predictor performance, which is given by

Table 3: The prediction accuracy on six networks.

		CN	CN-W	LHN	LHN-W	RA	RA-W	WIC	ICRA	LP	LRW
Jazz	AUC	0.953	0.826	0.902	0.796	**0.971**	0.829	0.809	0.956	0.945	0.912
	W	0.946	0.708	0.901	0.680	**0.970**	0.712	0.690	0.956	0.945	0.912
USAir	AUC	0.933	0.778	0.770	0.741	**0.951**	0.783	0.768	0.950	0.923	0.908
	W	0.909	0.604	0.755	0.570	**0.937**	0.613	0.594	0.936	0.919	0.902
Metabolic	AUC	0.920	0.748	0.739	0.731	**0.956**	0.753	0.744	**0.956**	0.917	0.869
	W	0.881	0.532	0.730	0.520	0.946	0.542	0.528	**0.948**	0.915	0.868
PB	AUC	0.917	0.893	0.762	0.770	0.922	0.895	0.890	0.927	0.931	**0.937**
	W	0.890	0.843	0.745	0.725	0.904	0.851	0.840	0.910	0.930	**0.936**
Tvshow	AUC	0.905	0.892	0.904	0.891	0.907	0.892	0.892	0.906	**0.952**	0.946
	W	0.813	0.786	0.814	0.786	0.815	0.786	0.785	0.815	**0.912**	0.899
Yeast	AUC	0.883	0.673	0.706	0.662	0.892	0.673	0.669	0.894	**0.907**	0.906
	W	0.843	0.383	0.682	0.374	0.868	0.386	0.379	0.871	**0.907**	0.906

$$d = \sum_{1 \le i < j \le |N|} \text{dist}\left(x_i, x_j\right), \tag{4}$$

where $|N|$ is the number of predictors analyzed and $\text{dist}(\cdot,\cdot)$ is one distance metric. We choose two standard distance metrics here: Manhattan distance and Euclidean distance. The total distance between the predictor performance measured by the W-index is consistently much larger than that measured by the AUC on all datasets, as shown in Figure 4. According to the above discussion, the W-index metric explicitly encourages a larger separability between evaluated methods.

4.6. Reevaluation of Link Prediction Methods. Since the W-index encourages discriminative evaluation of link prediction methods, it offers another perspective to observe the performance of the predictors. Specifically, we use the evaluation scores of CN and LP for comparison. Compared with LP, which uses high-order path information, CN only considers its second-order path. As a result, its state is generally too limited to tell the differences between testing links and unobserved links, and it is easier to draw. Therefore, one natural guess is that the performance of the CN is likely not as accurate as LP. However, we can find in Figure 2 that, under the AUC metric, CN is better than LP on Jazz and Metabolic. And although it is inferior to LP in the other four networks, the gap is acceptable. This is somewhat different from our assumption, but under the W-index, except for the similar performance on Jazz, the CN is always worse than the LP, and the difference is much larger. It is particularly noteworthy that, under the two evaluation metrics, the performance of CN and LP on Metabolic and USAir is reversed. This shows that LP can indeed alleviate the problem of "degeneracy of the states." By reducing the number of draws to gain more wins or losses, LP provides a more fine-grained score than CN.

From our definition of W-index in Section 2, the difference in prediction accuracy lies in draws. Based on our assumption that wins are much more convincing than draws for evaluating the superiority of a method, the smaller the difference, the more reliable the method. We can find in Figure 2 that LP has the smallest difference in five of six networks followed by LRW and ICRA. On the contrary, WIC, CN-W, LHN-W, and RA-W have the biggest difference among all six networks. Meanwhile, from the results of Table 3, the average prediction accuracy differences between local similarity measures and their W-forms measured by the W-index on these six networks are 3.76 times as those measured by the AUC. Besides, the average prediction accuracy difference between CN and LP, CN and LRW, and CN and ICRA measured by the W-index is 2.55, 2.41, and 2.21 times as that measured by the AUC. For example, the average prediction accuracy difference between CN and CN-W is 0.127 and 0.238 on Jazz measured by AUC and W-index, respectively; therefore, the difference measured by the W-index is 1.869 times than that measured by the AUC.

Hence, by using W-index, we can explicitly figure out that LP, LRW, and ICRA have superior overall prediction performance, whereas CN-W, LHN-W, RA-W, and WIC perform poorly. As we expected, under the W-index metric, quasi-local similarity predictors show more evident advantages over local similarity predictors. To our surprise, community information does not necessarily improve the accuracy of link prediction. For example, the W-form of CN does not show excellent performance as in the previous study, but there is no significant difference in the performance of ICRA, which indicates that the way of introducing community information has a great impact on the performance of the predictor. More intuitively, we show the statistical distributions of the performance of all predictors measured by AUC and W-index on six networks in Figure 5. Evidently, under the W-index, the performance between the methods is easier to distinguish. The performance scores of outstanding methods, such as LP and LRW, will have a larger mean and smaller variance, which remarkably outperform others.

4.7. Varying Network Structure. As mentioned in Section 4.3, the performance of a predictor is largely affected by the network structure, such as network size (node number), network density, average shortest path length, average centrality, clustering coefficient, and network diameter. To clarify the relationship between the W-index and the network structure, we conduct experiments to investigate this in extensive networks with different structures.

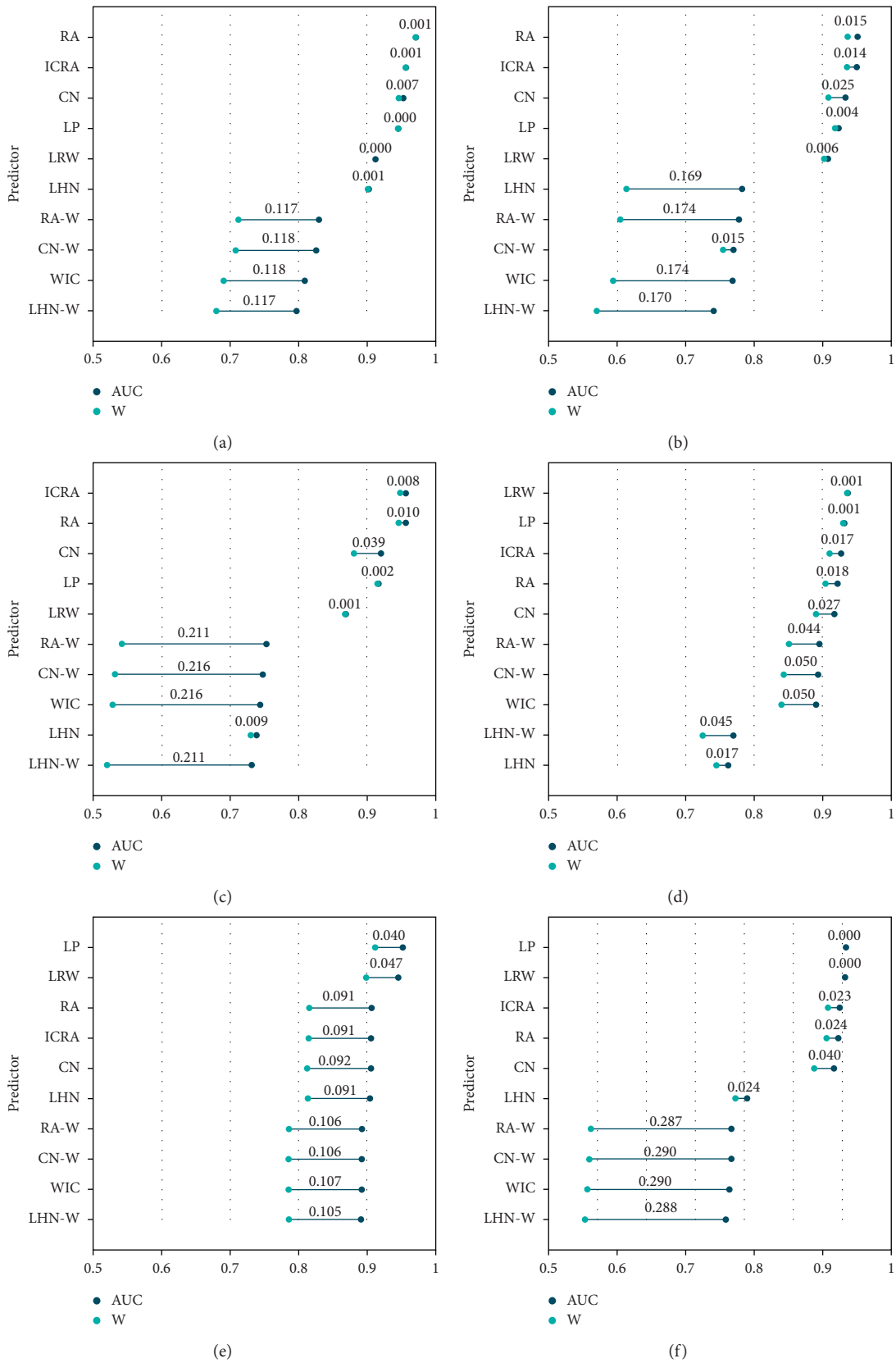

Figure 2: Difference in accuracy for each predictor under AUC and W-index metrics. (a) Jazz accuracy change W-AUC. (b) USAir accuracy change W-AUC. (c) Metabolic accuracy change W-AUC. (d) PB accuracy change W-AUC. (e) Tvshow accuracy change W-AUC. (f) Yeast accuracy change W-AUC.

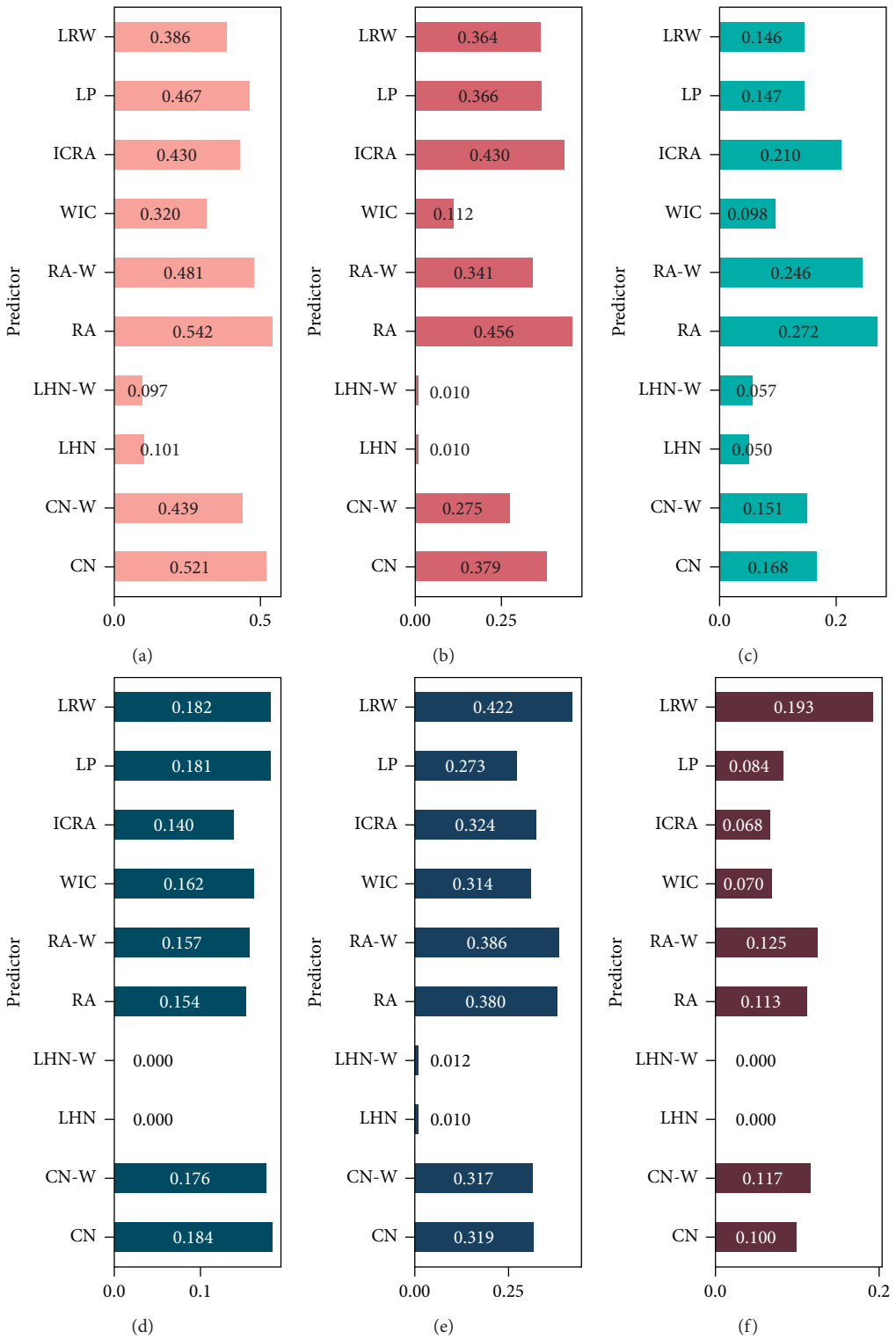

Figure 3: Accuracy of ten predictors quantified by precision on six networks. (a) Jazz. (b) USAir. (c) Metabolic. (d) PB. (e) Tvshow. (f) Yeast.

Table 4: The correlation coefficient between W-index score and precision score and that of AUC score and precision score on six networks.

Networks	Jazz	USAir	Metabolic	PB	Tvshow	Yeast
W-precision	0.373	0.412	0.695	0.880	0.236	0.295
AUC-precision	0.491	0.601	0.845	0.954	0.253	0.435

Table 5: The correlation coefficient between W-index score and precision score and that of AUC score and precision score on six networks without considering the CN-W, LHN-W, RA-W, and WIC predictors.

Networks	Jazz	USAir	Metabolic	PB	Tvshow	Yeast
W-precision	0.840	0.995	0.898	0.956	−0.295	0.774
AUC-precision	0.853	0.995	0.902	0.967	−0.311	0.748

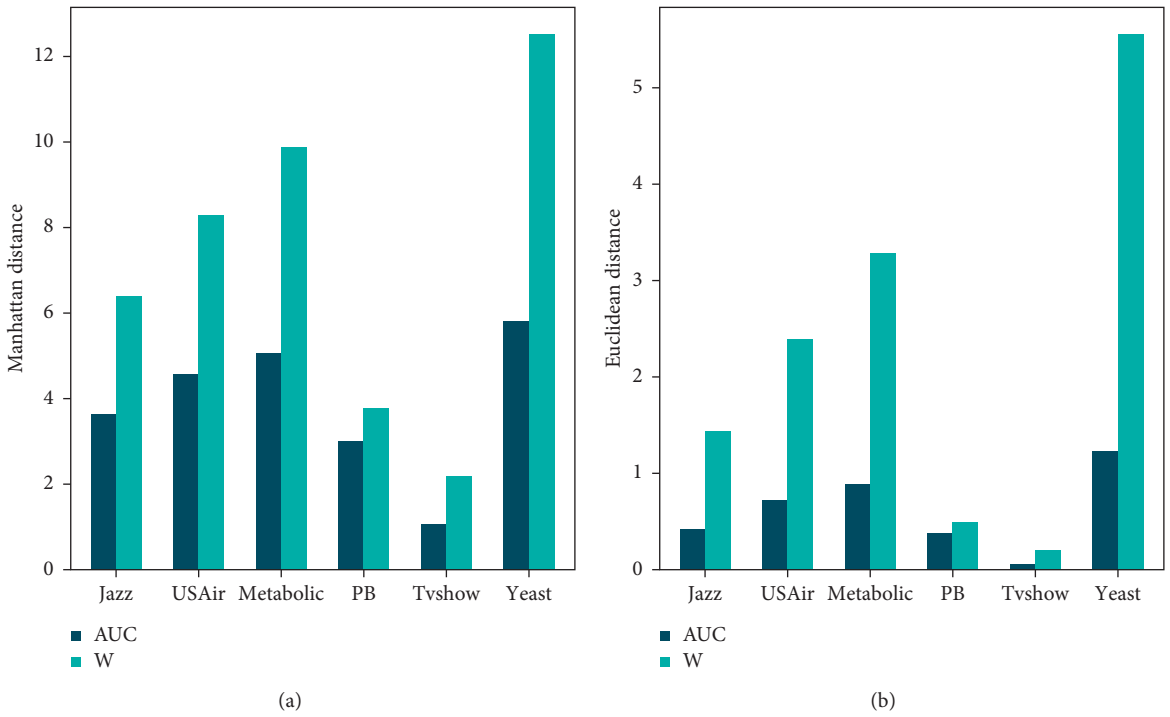

Figure 4: Total distance between the measure performance calculated by (a) Manhattan distance and (b) Euclidean distance.

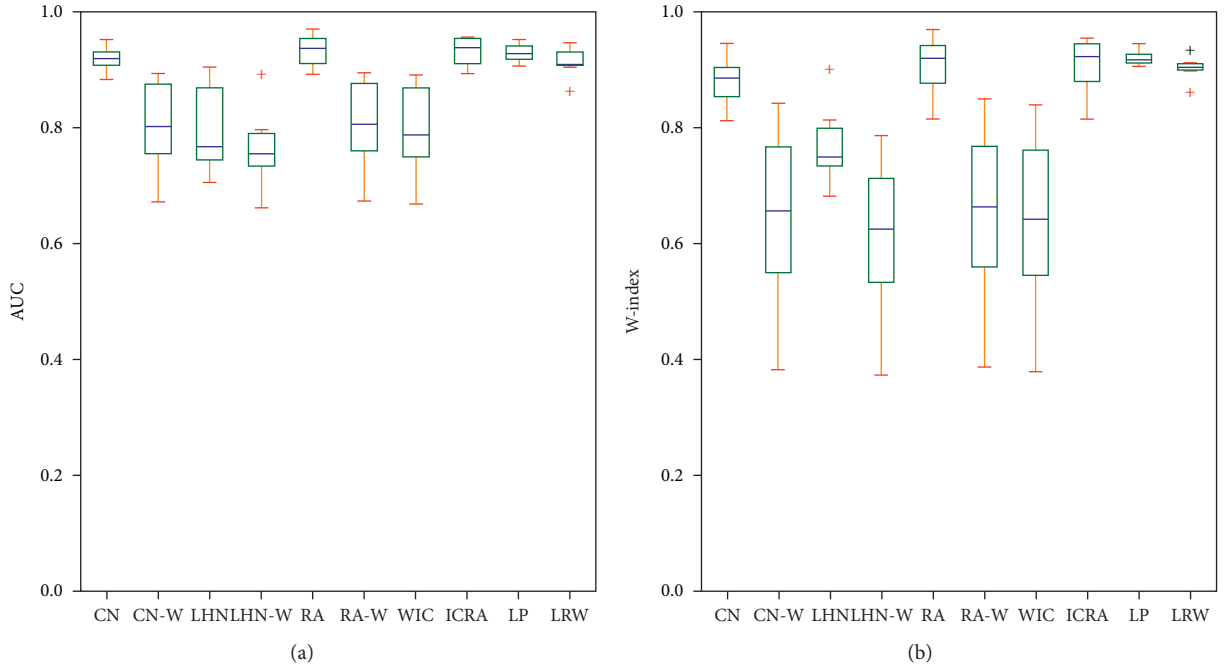

Figure 5: Statistical distributions of the performance of all measures measured by (a) AUC and (b) W-index. The line inside each box indicates the median of the prediction accuracies on six networks.

The Watts–Strogatz (WS) small-world network is a common network model, often used to describe real-world social networks. To vary the network structure, a series of WS small-world graphs are constructed as follows. First, a ring over n nodes is created, and then each node in the ring is joined to its k-nearest neighbors. Next, a series of rewirings are performed: for each edge (u, v) in the underlying "n-ring with k-nearest neighbors" with probability p, replace it with a new edge (u, w) with uniformly random choice of existing node w.

Table 6 shows the network structure of six groups of WS small-world networks used in the following experiments, namely, WS-Group1 to WS-Group6. Since network characteristics are closely related, a change in one characteristic will cause another change accordingly. For example, as the network size becomes smaller, the average shortest path of the network will inevitably become smaller. Therefore, we only list the fluctuation range of the independent variable of each group of networks and ignore the change of the dependent variable, which is marked by "—." Note that $|V|$ is the number of nodes, $|E|$ is the number of edges, k is the average degree of the network, d is the density of the network, p is the average shortest path length of the network, B is the average betweenness centrality of the network, C is the average clustering coefficient of the network, and D is the diameter of the network.

4.7.1. Network Size. We constructed ten networks with an average degree of 10, but the number of nodes varies from 100 to 1000, as shown in WS-Group1 in Table 6. Figure 6 shows the performance of each predictor for different networks in WS-Group1.

It can be seen that, under the AUC metric, as the number of nodes increases, the performance of community-based local similarity predictors increases significantly at first and then grows slightly. Figure 6 shows that the turning point is 600 nodes. The performance of other predictors rebounds after a moderate decline and eventually fluctuates within a narrow range after network size exceeds 400. Similarly, under W-index, the performance of community-based local similarity predictors surges first and then goes up steadily after the network size exceeds 600. Meanwhile, the trend of other local similarity predictors is consistent with that under the AUC metric, but the change range is greater. Exceptionally, the performance of quasi-local similarity predictors gradually goes down as the network size increases.

4.7.2. Network Density. Network density describes the portion of all potential connections in a network that are actual connections. We constructed eight networks with 1000 nodes, and the density of networks varies from 0.002 to 0.05, as shown in WS-Group2 in Table 6. As can be seen in Figure 7, when the density is extremely small, the performance of the predictors has a huge leap as the density increases whether measured by the AUC metric or the W-index. However, after the network density reaches a certain value, the predictors' performance shows different trends as the network density continues to go up. Specifically, under the AUC metric, the performance of community-based local similarity predictors drops slightly, while the performance of other predictors remains the same. Under the W-index, the trend of community-based local similarity predictors is consistent with that under the AUC metric, but the change range is greater. Unlike the previous trend, the performance of other predictors, especially quasi-local similarity predictors, is gradually increasing.

4.7.3. Average Shortest Path Length. The shortest path between two nodes is defined as the path with the minimal length. The average shortest path length of a network is defined as the mean of the shortest path of all pairs of nodes. We constructed ten networks with 1000 nodes and 5000 edges. The average shortest path length of these networks varies from 3.26 to 50.45, as shown in WS-Group3 in Table 6.

Figure 8 shows that no matter whether it is measured by the AUC metric or W-index, the performance of these predictors will quickly reach its peak as the average shortest path length increases. Since then, the performance of these predictors is in a stable state. The only difference is that, under the W-index, the performance changing range of these predictors is greater than that under the AUC metric.

4.7.4. Centrality. Centrality expresses the degree to which a node is at the centre of the entire network, which can help identify the vital nodes. There are several common centrality algorithms; here, we take betweenness centrality as an example. Betweenness centrality measures the fraction of the shortest paths passing through a node. We constructed ten networks with 2000 nodes and 10000 edges. The average betweenness centrality of these networks varies from 0.001 to 0.008, as shown in WS-Group4 in Table 6. It can be revealed from Figure 9 that when measured by the AUC metric and W-index, the performance of the predictors shows a rapid, exponential increase as the average betweenness centrality increases followed by a level off. Again, the performance changing range under the W-index is greater than that under the AUC metric.

4.7.5. Clustering Coefficient. Clustering coefficient is a coefficient used to describe the degree of clustering between the nodes of a graph. Specifically, it is the degree to which the adjacent nodes of a node are connected to each other. The average clustering coefficient of a graph is the arithmetic average of the local clustering coefficient values of all nodes, which measures the agglomeration degree of a graph on the whole. We constructed ten networks with 1000 nodes and 5000 edges. The average clustering coefficient of these networks varies from 0.09 to 0.67, as shown in WS-Group5 in Table 6. Figure 10 shows that, under AUC and W-index, the performance of predictors increases in proportion to the average clustering coefficient of the networks, but after the average clustering coefficient is over a certain value, the growth rate decreases. These mean that the average

Table 6: Network structure of six groups of WS small-world networks.

Networks	$\|V\|$	$\|E\|$	k	d	p	B	C	D
WS-Group1	100–1000	500–5000	10	—	—	—	—	—
WS-Group2	1000	1000–25000	—	0.002–0.05	—	—	—	—
WS-Group3	1000	5000	—	—	3.26–50.45	—	—	—
WS-Group4	2000	10000	—	—	—	0.001–0.008	—	—
WS-Group5	1000	5000	—	—	—	—	0.09–0.67	—
WS-Group6	1000	5000	—	—	—	—	—	5–100

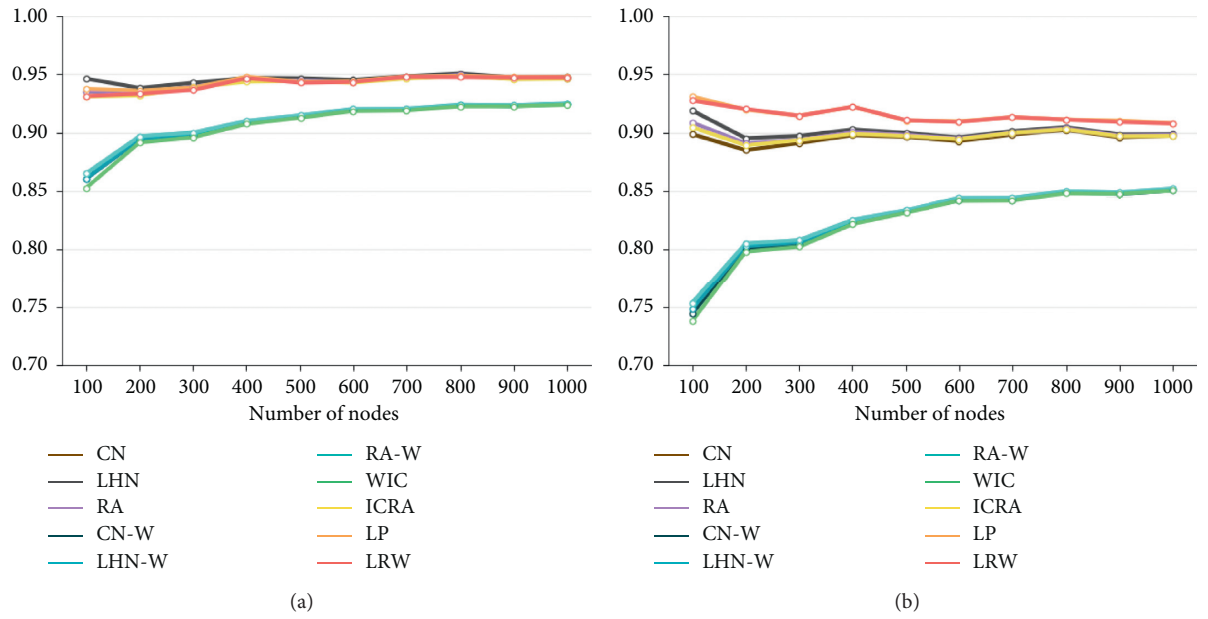

Figure 6: Performance of each predictor under AUC and W-index in WS-Group1. (a) Accuracy measured by AUC. (b) Accuracy measured by W-index.

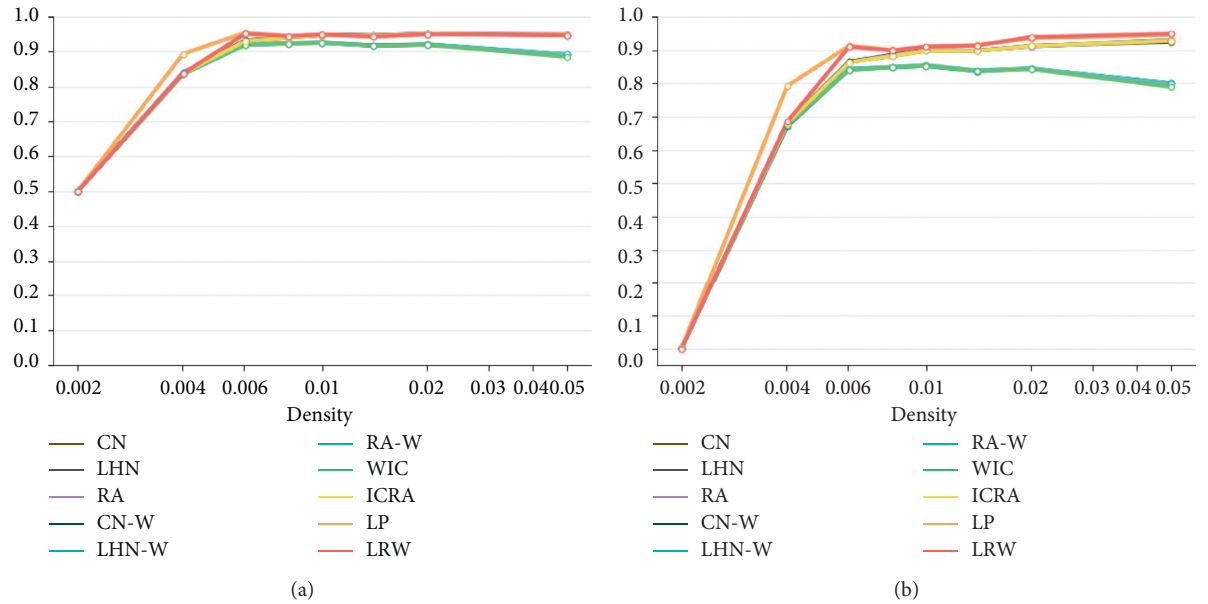

Figure 7: Performance of each predictor under AUC and W-index in WS-Group2. (a) Accuracy measured by AUC. (b) Accuracy measured by W-index.

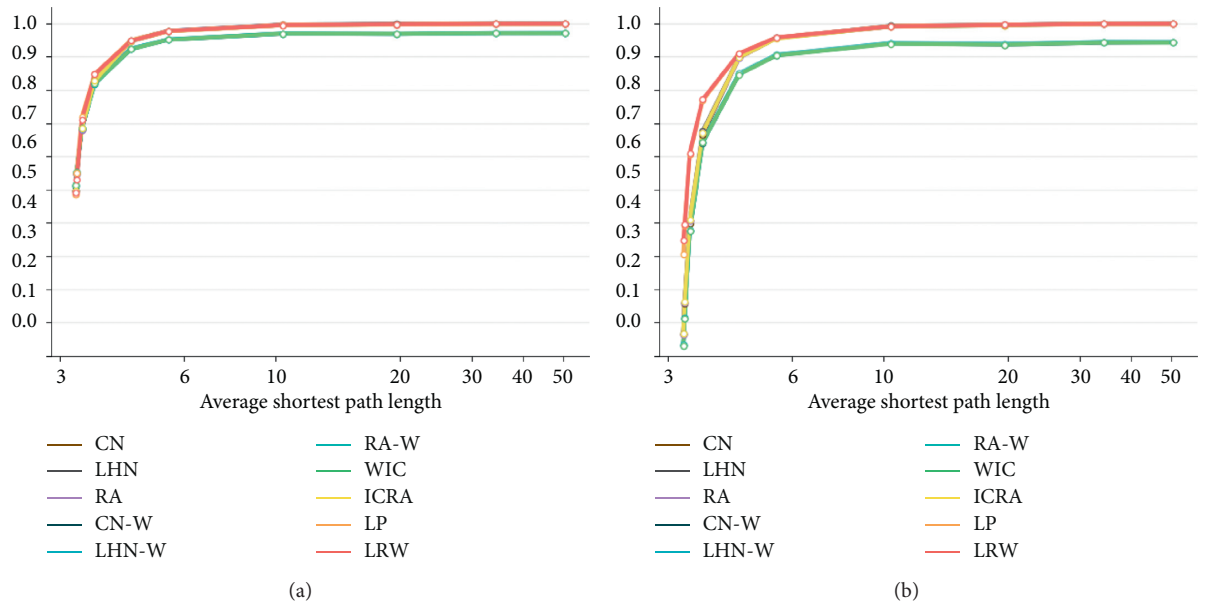

Figure 8: Performance of each predictor under AUC and W-index in WS-Group3. (a) Accuracy measured by AUC. (b) Accuracy measured by W-index.

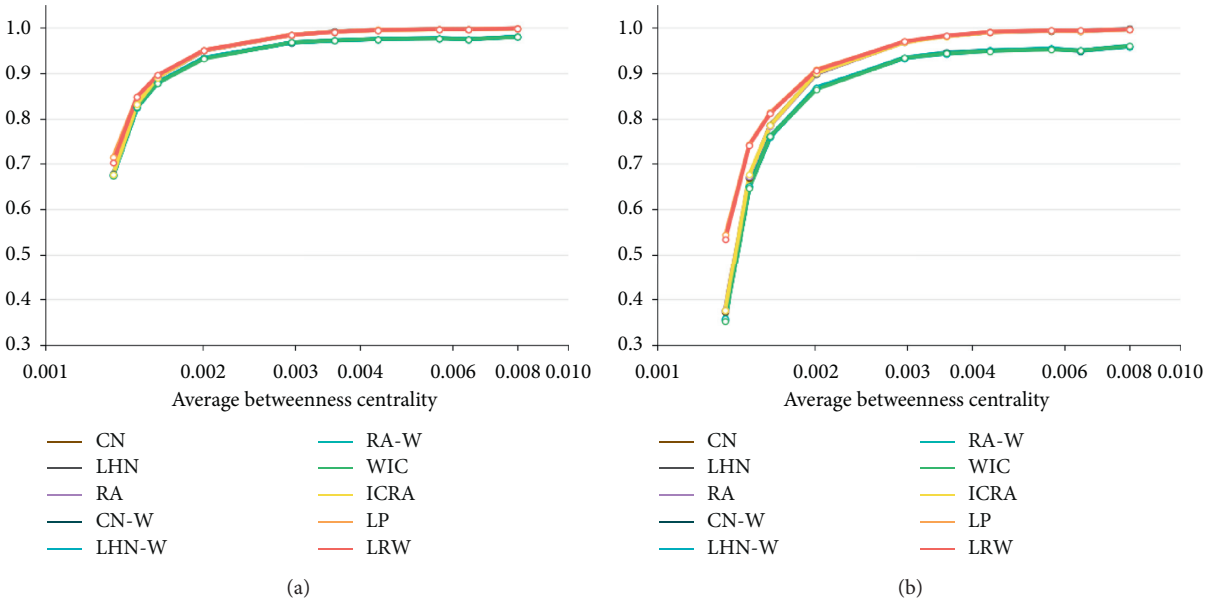

Figure 9: Performance of each predictor under AUC and W-index in WS-Group4. (a) Accuracy measured by AUC. (b) Accuracy measured by W-index.

clustering coefficient of networks has an incredibly important impact on the performance of these predictors.

4.7.6. Diameter. The diameter of a graph is defined as the maximum distance between all pairs of nodes. We constructed ten networks with 1000 nodes and 5000 edges. The diameter of these networks varies from 5 to 100, as shown in WS-Group6 in Table 6. Figure 11 shows that, in the initial range, whether measured by AUC or W-index, with the increase of the diameter of the network, the performance of predictors takes off, and after that point, the growth is negligible. Besides, the performance changing range is greater under the W-index than that under the AUC metric.

4.7.7. Summary. It can be seen from the above analysis that in addition to the number of nodes, other network structures have a greater impact on the performance of the predictors. Since these network structures are highly correlated, for example, the larger the network diameter, the larger the network clustering coefficient and the larger the edge

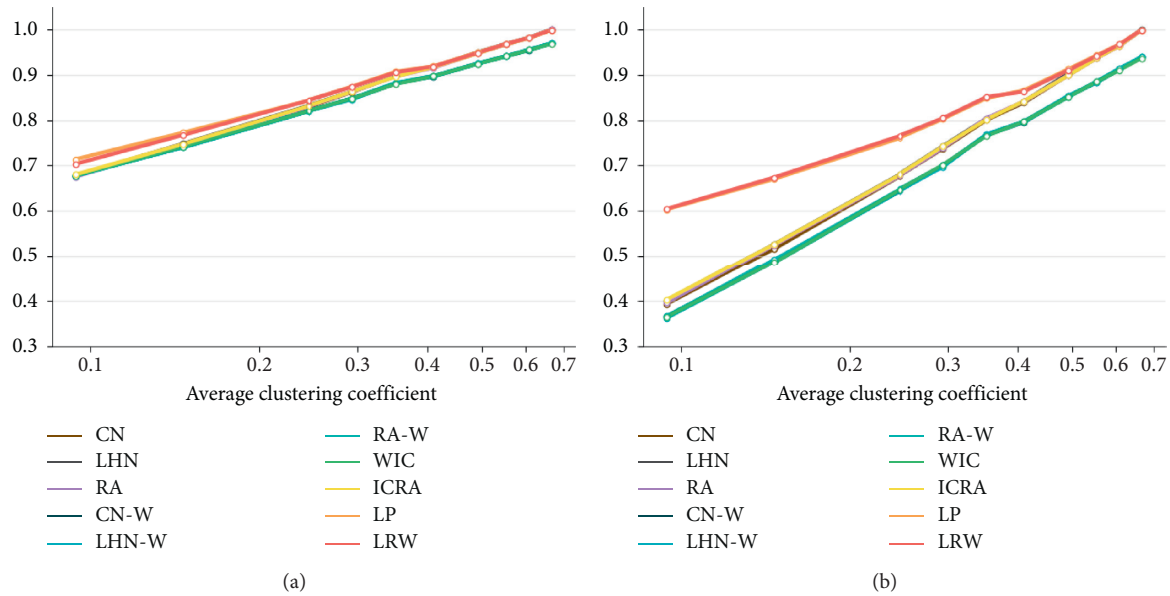

Figure 10: Performance of each predictor under AUC and W-index in WS-Group5. (a) Accuracy measured by AUC. (b) Accuracy measured by W-index.

betweenness, so the experimental results on these networks look very similar, showing a trend of gradual improvement in the performance. Of course, due to the characteristics of the W-index itself, its range of change will be larger than the AUC. Besides, although it is not very obvious, it can be seen that only three lines can be clearly seen in these 10 curves. This shows that these predictors are almost divided into three categories, namely, local similarity predictors (including ICRA), local similarity predictors based on community information predictors (except ICRA), and quasi-local similarity predictors. And the gap under the W-index is larger than the gap under the AUC, which is similar to our previous conclusion.

4.8. Varying Proportion of the Testing Set. As we all know, in machine learning and deep learning communities, it is important to reasonably divide the training set and testing set, and the same is true for the link prediction problem. Here, we carry out experiments to study the performance of predictors under different ratios of the training set when using W-index and AUC for evaluation. The datasets used in the experiments are the same as those in Section 4.1, and the experimental settings are the same as those in Section 4.2, except the ratio of the training set E^{Train} to the testing set E^{Test} which is not fixed at 9 : 1. We evaluate the performance of each predictor in the six networks when the ratio of the training set $E^{\text{Train}}$to the set of links E ranges from 50% to 90%, as shown in Figure 12. The value in the figure is the difference in accuracy for each predictor under AUC and W-index.

It can be seen that, as the proportion of the training set increases, the change in the predictor performance is related to the network structure. For example, in Tvshow, the performance of all predictors has been significantly improved, and the growth rate is decreasing as the proportion of the training set increases. The same happens in Metabolic and Yeast, except for the LHN predictor. However, in USAir, under both AUC and W-index metrics, the performance of all predictors fluctuates as the proportion of the training set increases, reaching the highest peak when the proportion of the training set is 0.7. Apart from the LHN predictor, so does Jazz. The PB network combines the above two situations. The performance of the LHN, LHN-W, LRW, and LP predictors oscillates as the proportion of the training set changes, while the performance of the remaining predictors improves as the proportion of the training set increases, and the growth rate is gradually decreasing. Besides, the improvement in the predictor performance under the W-index is greater than that under the AUC evaluation.

Through the above observations, it is not difficult to figure out that the predictors have adaptability to the network; whether the predictor performance can be improved as the lengths of the training set increase is also related to the nature of the predictor. For example, in most cases, the change of the LHN index is different from other local similarity predictors. Besides, the performance of quasi-local similarity predictors is much less sensitive to the sampling ratio than that of local similarity predictors.

5. Discussion

5.1. Properties of the W-Index. The properties of the W-index are compared with those of the AUC in the following:

(1) As mentioned in Section 2, the range of the W-index is 0 to 1, while the value of the AUC ranges from 0.5 to 1. This means that the limit on the values that the W-index can take is greater than the AUC. However, the AUC score of a link prediction method may also be lower than 0.5 [48], which indicates that the approach fails to predict the missing links and

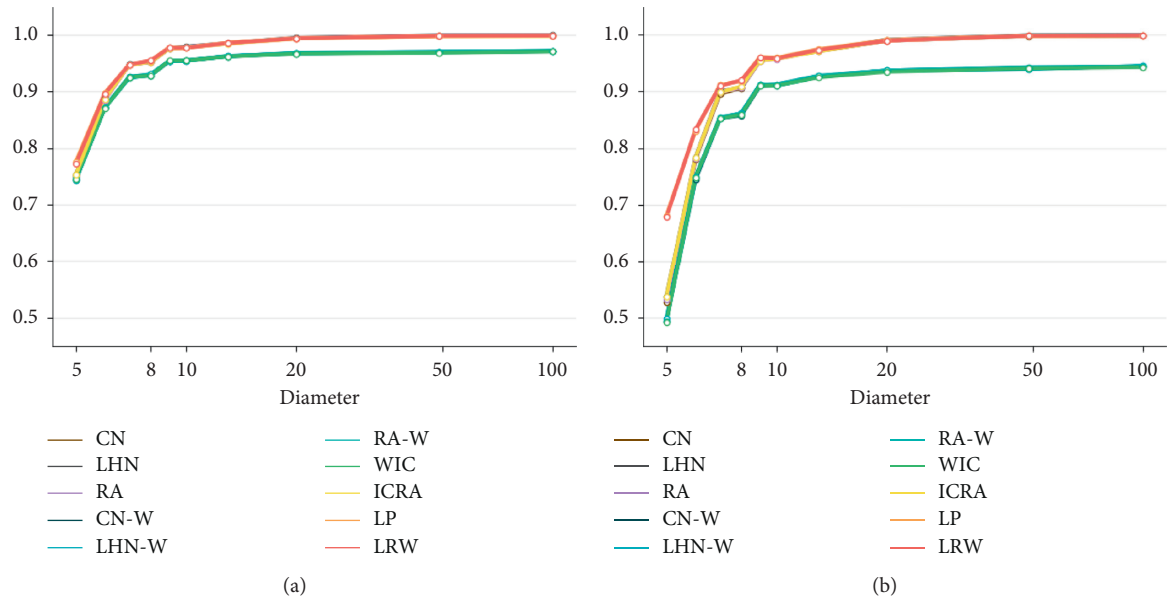

Figure 11: Performance of each predictor under AUC and W-index in WS-Group6. (a) Accuracy measured by AUC. (b) Accuracy measured by W-index.

cannot explain the evolution mechanism of the network.

(2) Random prediction (pure chance) is the benchmark for judging the pros and cons of link prediction methods. However, the W-index score of random prediction is no longer a fixed value, like 0.5 points scored by the AUC. Specifically, the score of random prediction varies with the way where a random score is generated. For example, if all testing links and unobserved links are given the same score, the W-index score of random prediction is 0. Alternatively, if the distribution of random scores is normal, the W-index score of random prediction is close to 0.5, or if the random scores are generated from a discrete uniform distribution, the W-index score of random prediction is a positive number less than 0.5. Briefly, the W-index score of random prediction ranges from 0 to 0.5.

(3) 0.5 points is still a benchmark score in the W-index, but its meaning has changed. Under the AUC metric, 0.5 points means the same number of wins and losses, that is, giving scores by pure chance. This means that the performance of a link predictor is better than chance only when its AUC is greater than 0.5. Unlike this, under the W-index, the score of pure chance is no longer a fixed value, but its upper limit is 0.5 points. Therefore, a link prediction method with a score higher than 0.5 must be better than the pure chance. Besides, when the number of wins is equal to the sum of the draws and losses, the W-index value is 0.5 points. Obviously, a competitive approach should have more wins than draws and losses. Taken together, only a link prediction method with score above 0.5 is effective. The greater the degree exceeds 0.5, the better the algorithm performs.

5.2. General Interpretation of the W-Index. As a matter of fact, the proposed W-index is a special case of cost-sensitive evaluation metrics. In undirected and unweighted graphs, we can regard link prediction as a binary classification problem where all pairs of nodes are divided into observed edges and unobserved edges. The classification result includes three cases. The first is the correct classification, that is, the testing link having a higher score than the unobserved link or, namely, win. The second is the wrong classification result, namely, loss, which can be further refined into the following two. One is treating observed edges as unobserved edges, named *L1*, and the other is treating unobserved edges as observed edges, named *L2*. The third is it cannot be classified, that is, the testing link having the same score as the unobserved link, or, namely, draw. Among them, the latter two classification results which are *L1*, *L2* and draw will bring costs.

Taking protein-protein interaction networks as an example, we have to perform lots of expensive and time-consuming experiments to discover unknown interactions. In *L1*, we miss the experiment we were supposed to perform and are unable to obtain discoveries. In *L2*, we do useless experiments and could not get findings either. In the draw, we need to perform all experiments, but this takes a lot of time and money.

Under the W-index, we have the same penalties for these three costs, that is, we get 0 points in draws and losses. However, in different contexts, different classification results often bring different costs. Furthermore, we can focus on the cases where the cost is high and use the total cost of classification results as the evaluation criterion. For example, we give −10 points in *L1*. In this way, although a link prediction method with the highest accuracy under the AUC metric may be abandoned, it has important practical significance in applications.

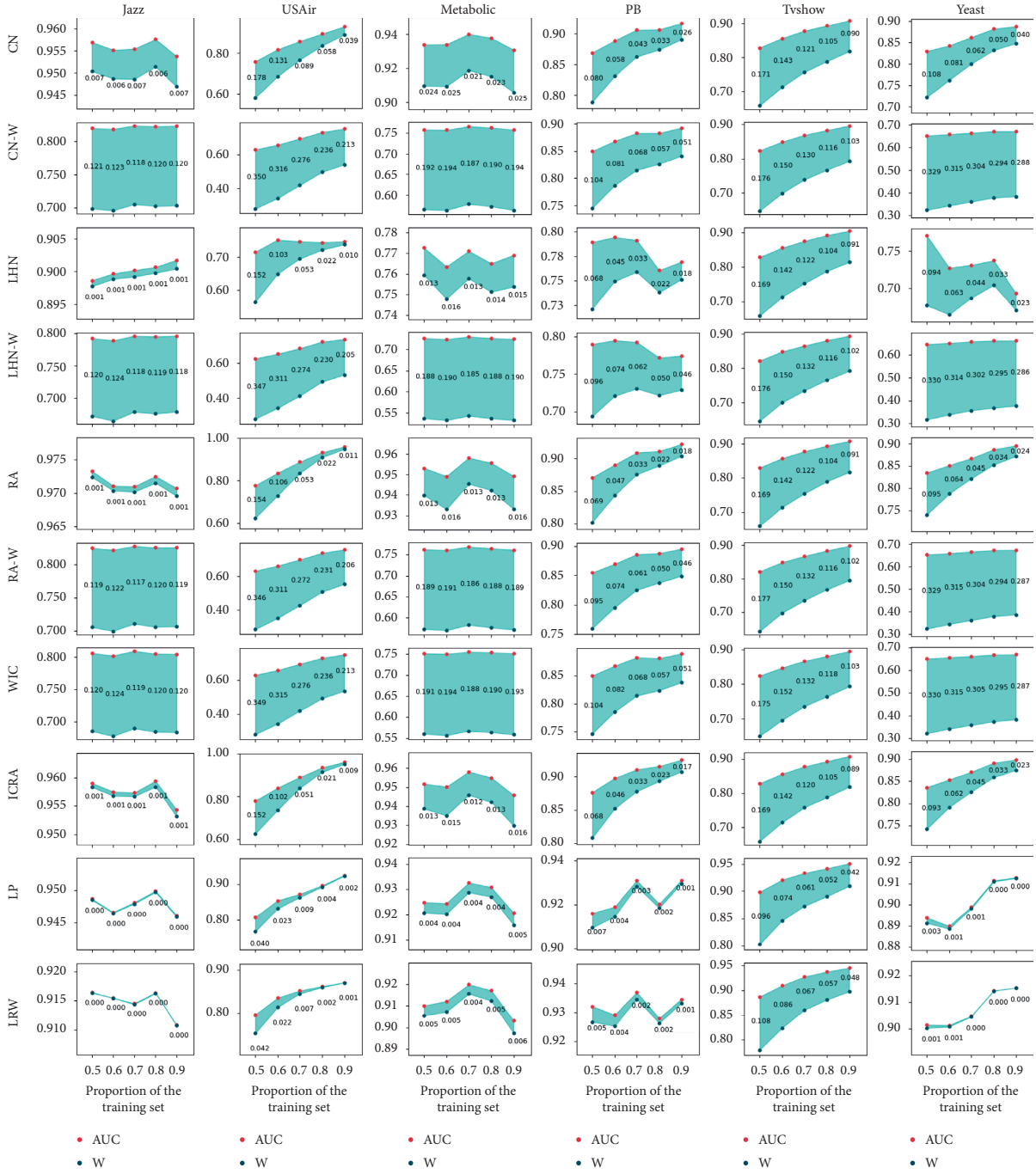

Figure 12: Performance of each predictor in the six networks under different proportions of the training set.

6. Conclusions

In this paper, we discuss two side effects of draws in the AUC and propose the W-index, which only cares about wins, to obtain discriminative evaluation of link prediction methods. A series of tools have been introduced for measuring the reliability and the performance of this new metric in this paper. Based on empirical data, two methods, i.e., ranking change and correlation analysis, are applied for comparative analysis to verify the reliability of the W-index. To evaluate the performance of the W-index, we utilize local and global distances to measure the differences between link prediction methods. Moreover, the impact of the network structure and training set length on the performance of predictors is clarified under W-index and AUC. These tools may shed light on the study of new evaluation metrics.

The main observations from our experiments on various networks are summarized as follows. Firstly, W-index is able to effectively evaluate the performance of predictors compared to AUC, which is supported by the following arguments. The performance ranking of link prediction methods does not change significantly when using W-index and

AUC, respectively, and the results of correlation analysis of W-precision and AUC-precision (see Tables 4 and 5) are also highly consistent. The next important observation is that using W-index instead of AUC to evaluate the performance of link predictors, the differences between these methods are more evident, both locally and globally. Then, our results show the use of community information does not necessarily improve the performance of a predictor. Whether it works depends on how it is utilized. For instance, the performance of W-form algorithms (i.e., CN-W, LHN-W, RA-W, and WIC) is not good, while ICRA always performs well. In addition, the performance of a link predictor is affected by the network structure, and our results in Section 4.7 show that the average clustering coefficient of networks is the main factor.

Finally, we would like to remind readers that, in 1995, the football league increased the reward for a win from two to three points, and the main objective of this rule change is to encourage more exciting and attractive matches. Thereafter, empirical data proved that the introduction of the three-point system reduced the number of draws in football matches and produced a more correct ranking of the teams [49]. In addition, the total number of goals scored per game also increased. Inspired by this, we hope that our work will contribute to further research and exploration of more competitive link prediction methods.

Data Availability

The networks used in this study are available from http://konect.uni-koblenz.de/networks/, http://networkrepository.com/, http://snap.stanford.edu/data/, http://vlado.fmf.uni-lj.si/pub/networks/data/, and https://icon.colorado.edu/#!/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Yun Yuan and Jingwei Wang contributed equally to this work.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (Grant no. 2019YFB1704700), the National Natural Science Foundation of China (Grant nos. 61573257, 71690234, 61873191, and 61973237), and the Science and Technology Commission of Shanghai Municipality (Grant nos. 19JG0500700 and 20JG0500200).

References

[1] V. Martinez, F. Berzal, and J. C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys*, vol. 49, no. 4, 2017.

[2] J. Tang, S. Chang, C. Aggarwal, and H. Liu, "Negative link prediction in social media," in *Proceedings of the ACM International Conference on WSDM*, pp. 87–96, Shanghai, China, February 2015.

[3] W. Yuan, J. Pang, D. Guan, Y. Tian, A. Al-Dhelaan, and M. Al-Dhelaan, "Sign prediction on unlabeled social networks using branch and bound optimized transfer learning," *Complexity*, vol. 2019, Article ID 4906903, 11 pages, 2019.

[4] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD'15*, pp. 785–794, Sydney, Australia, August 2015.

[5] Z. Sha, Y. Huang, J. Sophia Fu et al., "A network-based approach to modeling and predicting product coconsideration relations," *Complexity*, vol. 2018, Article ID 2753638, 14 pages, 2018.

[6] L. Mei, D. Tang, T. Wang, W. Du, Y. Xia, and X. Cao, "Predicting the evolution process of infrastructure networks with an NSIPA link prediction method," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 11, pp. 1895–1899, 2019.

[7] T. T. P. Souza and T. Aste, "Predicting future stock market structure by combining social and financial network information," *Physica A: Statistical Mechanics and its Applications*, vol. 535, 2019.

[8] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Drug-target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2013.

[9] Y. Luo, X. Zhao, J. Zhou et al., "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Communications*, vol. 8, no. 1, 2017.

[10] S. Mao and F. Xiao, "A novel method for forecasting Construction Cost Index based on complex network," *Physica A: Statistical Mechanics and its Applications*, vol. 527, 2019.

[11] A. Brintrup, P. Wichmann, P. Woodall, D. McFarlane, E. Nicks, and W. Krechel, "Predicting hidden links in supply networks," *Complexity*, vol. 2018, Article ID 9104387, 12 pages, 2018.

[12] J. Wang, Q.-M. Zhang, and T. Zhou, "Tag-aware link prediction algorithm in complex networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 523, pp. 105–111, 2019.

[13] S. Bai, L. Li, J. Cheng, S. Xu, and X. Chen, "Predicting missing links based on a new triangle structure," *Complexity*, vol. 2018, Article ID 7312603, 11 pages, 2018.

[14] L. Li, S. Bai, M. Leng, L. Wang, and X. Chen, "Finding missing links in complex networks: a multiple-attribute decision-making method," *Complexity*, vol. 2018, Article ID 3579758, 16 pages, 2018.

[15] M.-Y. Zhou, H. Liao, W.-M. Xiong, X.-Y. Wu, and Z.-W. Wei, "Connecting patterns inspire link prediction in complex networks," *Complexity*, vol. 2017, Article ID 8581365, 12 pages, 2017.

[16] S. N. Chowdhury, D. Ghosh, and C. Hens, "Effect of repulsive links on frustration in attractively coupled networks," *Physical Review E*, vol. 101, no. 2, 2020.

[17] S. Haghani and M. R. Keyvanpour, "A systemic analysis of link prediction in social network," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1961–1995, 2019.

[18] R. R. Junuthula, K. S. Xu, and V. K. Devabhaktuni, "Evaluating link prediction accuracy in dynamic networks with added and removed edges," in *Proceedings of the IEEE International Conferences on BDCloud*, pp. 377–384, Atlanta, GA, USA, September 2016.

[19] Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, "Evaluating link prediction methods," *Knowledge and Information Systems*, vol. 45, no. 3, pp. 751–782, 2015.

[20] R. Lichtenwalter and N. V. Chawla, "Link prediction: fair and effective evaluation," in *Proceedings of the ACM International Conference on ASONAM*, pp. 376–383, Istanbul Turkey, August 2012.

[21] J. Herlocker, J. Konstan, L. Terveen, and T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, pp. 5–53, 2004.

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

[23] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the ACM International Conference Proceeding Series*, pp. 233–240, Ontario, Canada, October 2006.

[24] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 322–331, Omaha, NE, USA, October 2007.

[25] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining-KDD'10*, pp. 243–252, Washington, DC, USA, July 2010.

[26] D. Davis, R. Lichtenwalter, and N. V. Chawla, "Multi-relational link prediction in heterogeneous information networks," in *Proceedings of the ACM International Conference on ASONAM*, pp. 281–288, Kaoshiung, Taiwan, July 2011.

[27] 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.

[28] H. Jin and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[29] S. Wu, P. Flach, and C. Ferri, "An improved model selection heuristic for AUC," in *Machine Learning: ECML 2007*, pp. 478–489, Springer, Berlin, Germany, 2007.

[30] C. Ferri, P. Flach, J. Hernández-Orallo, and A. Senad, "Modifying ROC curves to incorporate predicted probabilities," in *Proceedings of the Second Workshop on ROC Analysis in Machine Learning*, pp. 33–40, Bonn, Germany, September 2005.

[31] T. Calders and S. Jaroszewicz, "Efficient AUC optimization for classification," in *Knowledge Discovery in Databases: PKDD 2007*, pp. 42–53, Springer, Berlin, Germany, 2007.

[32] L. Y. Lu and T. Zhou, "Link prediction in complex networks: a survey," *Physica A*, vol. 390, no. 6, pp. 1150–1170, 2011.

[33] L. Lu, C. H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, p. 046122, 2009.

[34] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Physical Review E*, vol. 72, p. 027104, 2005.

[35] M. Zanin, D. Papo, P. A. Sousa et al., "Combining complex networks and data mining: why and how," *Physics Reports-Review Section of Physics Letters*, vol. 635, pp. 1–44, 2016.

[36] S. Redner, "Teasing out the missing links," *Nature*, vol. 453, no. 7191, pp. 47-48, 2008.

[37] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Physical Review E*, vol. 73, no. 2, 2006.

[38] T. Zhou, L. Lu, and Y.-C. Zhang, "Predicting missing links via local information," *European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.

[39] J. C. Valverde-Rebaza and A. de Andrade Lopes, "Link prediction in complex networks based on community information," in *Advances in Artificial Intelligence-SBIA 2012*, pp. 92–101, Springer, Berlin, Germany, 2012.

[40] 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 *Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 158–163, Banff, Canada, October 2017.

[41] W. Liu and L. Lue, "Link prediction based on local random walk," *EPL (Europhysics Letters)*, vol. 89, no. 5, 2010.

[42] P. M. Gleiser and L. Danon, "Community structure in Jazz," *Advances in Complex Systems*, vol. 6, no. 4, pp. 565–573, 2003.

[43] http://vlado.fmf.uni-lj.si/pub/networks/data/mix/USAir97.net.

[44] L. A. Adamic, "The political blogosphere and the 2004 U.S. Election: divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, New York, NY, USA, August 2005.

[45] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, "GEMSEC: graph embedding with self clustering," in *Proceedings of the ACM International Conference on ASONAM*, pp. 65–72, Vancouver, Canada, September 2019.

[46] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, TX, USA, February 2015, http://networkrepository.com.

[47] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics-Theory and Experiment*, 2008.

[48] K.-k. Shang, T.-c. Li, M. Small, D. Burton, and Y. Wang, "Link prediction for tree-like networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 6, p. 061103, 2019.

[49] A. Dilger and H. Geyer, "Are three points for a win really better than two? theoretical and empirical evidence for German Soccer," *SSRN Electronic Journal*, 2008.