Inconsistency among evaluation metrics in link prediction

Yilin Bi[a,1], Xinshan Jiao[a,1], Yan-Li Lee[b], and Tao Zhou[a,*]

[a]CompleX Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[b]School of Computer and Software Engineering, Xihua University, Chengdu 610039, China
[1]Yilin Bi and Xinshan Jiao contributed equally to this work.
[*]To whom correspondence should be addressed: zhutou@ustc.edu

February 27, 2024

Abstract

Link prediction is a paradigmatic and challenging problem in network science, which aims to predict missing links, future links and temporal links based on known topology. Along with the increasing number of link prediction algorithms, a critical yet previously ignored risk is that the evaluation metrics for algorithm performance are usually chosen at will. This paper implements extensive experiments on hundreds of real networks and 25 well-known algorithms, revealing significant inconsistency among evaluation metrics, namely different metrics probably produce remarkably different rankings of algorithms. Therefore, we conclude that any single metric cannot comprehensively or credibly evaluate algorithm performance. Further analysis suggests the usage of at least two metrics: one is the area under the receiver operating characteristic curve (AUC), and the other is one of the following three candidates, say the area under the precision-recall curve (AUPR), the area under the precision curve (AUC-Precision), and the normalized discounted cumulative gain (NDCG). In addition, as we have proved the essential equivalence of threshold-dependent metrics, if in a link prediction task, some specific thresholds are meaningful, we can consider any one threshold-dependent metric with those thresholds. This work completes a missing part in the landscape of link prediction, and provides a starting point toward a well-accepted criterion or standard to select proper evaluation metrics for link prediction.

Keywords: link prediction, evaluation metrics, inconsistency

1 Introduction

Network is a powerful tool to represent many complex social, biological and technological systems, and network science is an increasingly active interdisciplinary research domain [1,2]. Link prediction is one of the most productive branches of network science, which aims at estimating likelihoods of missing links, future links and temporal links, based on known topology [3–9]. As many observed networks are incomplete or dynamically changing, link prediction can find direct applications in inferring missing or upcoming links, such as the inference of missing biological interactions [10–12], the online recommendation of friends and products [13,14], and the prediction of future scientific discoveries [15,16]. Link prediction can also be considered as a touchstone to evaluate network models [17–20], because a better understanding of network formation will in principle lead to a more accurate prediction algorithm. In addition, link prediction can be an important step in solving some challenging problems, like network reconstruction [21,22] and sparse training [23], or an essential reason for some impressive phenomena, like polarization [24] and information cocoons [25] in online social networks.

A huge number of link prediction algorithms have been proposed recently (see some selected representatives [26–42]), and an accompanying question is how to evaluate algorithm performance. A standard procedure is to divide the observed links into a training set and a probe set, and to train model parameters by using only the information contained in the training set. An algorithm's performance is then measured by the closeness between ground truth and the algorithm's prediction. Many

evaluation metrics have already been utilized to quantify the above-mentioned closeness, including the two very popular ones, namely AUC [43,44] and Balanced Precision (BP) [45], the one with increasing popularity, say AUPR [46], as well as some occasionally used ones, such as Precision [47], Recall [47], F1-measure [47], Matthews Correlation Coefficient (MCC) [48], NDCG [49,50], AUC-Precision [33], and so on.

Everyone should be immediately aware of the crucial role of evaluation metrics, however, discussions about how to choose metrics in link prediction are rare. Recently, a few scientists have conducted criticism on popular metrics. Yang, Lichtenwalter, and Chawla [51] argued that, when evaluating link prediction performance, the precision-recall curve might provide better accuracy than the ROC curve. Saito and Rehmsmeier [52] pointed out that AUC is inadequate to evaluate the performance of imbalanced classification problem, while link prediction is a typical imbalanced classification problem as most real-world networks are sparse [53]. Menand and Seshadhri claimed that neither AUC nor AUPR can well characterize the algorithm performance in link prediction for sparse networks, and proposed a vertex-centric measure [54]. To overcome the shortcoming of AUC for imbalanced classification, Muscoloni, and Cannistraci [55] designed a novel metric named the area under the magnified ROC curve (AUC-mROC), which assign remarkably high weights to top-ranked links. Zhou *et al.* [45,56] proposed a method to quantitatively measure the discriminating ability of any metric and showed that AUC, AUPR and NDCG have significantly higher discriminating abilities than other well-known metrics.

In despite of those studies on evaluation metrics, thus far, there is no criterion or standard to select evaluation metrics: some scientists are conditioned to follow popular metrics, while some others have their own niche preferences (see, for example, Table 1 of Ref. [45]). Such fact reminds us of a even more fundamental question, that is, whether those evaluation metrics provide statistically consistent rankings of algorithms. If the answer is YES, then we can breathe easy since it is not a big deal in choosing metrics, while if the answer is NO, we have to reexamine the related literature because an algorithm being superior according to some metrics may be at a disadvantage for other metrics, and even worse, researchers who are too eager to get their works published may only report beneficial results from some metrics but ignore negative results from other metrics. This is not a groundless worry, as a recent large-scale experimental study has shown that a winner for one metric may be a loser for another metric [57].

Unfortunately, we have not found any answers to the above question in the literature. In this paper, we intend to provide a direct answer by analyzing correlations between evaluation metrics based on 25 algorithms and hundreds of real-world networks. The answer is a clear NO, and further analysis arrives to four practical suggestions in the selection of metrics, which are presented in the last section of this paper. We believe those suggestions can be considered as a useful guide in choosing evaluation metrics, before a commonly recognized standard for metric selection that may appear in the future.

2 Results

Consider a simple network $G(V, E)$, where V is the set of nodes, E is the set of links, the directionalities and weights of links are ignored, and the multiple links or self links are not allowed. We assume that there are some missing links in the set of unobserved links $U - E$, where U is the universal set containning all $|V|(|V|-1)/2$ potential links. The task of link prediction is to find out those missing links. However, as we do not known whether a link in $U - E$ is a missing link or a nonexistent link, to evaluate the algorithm's accuracy, a standard procedure is to use part of the links in E to predict the other part. Practically, we randomly divide the set E into a training set E^T and a probe set E^P, use only information in E^T to predict missing links, and approximately treat E^P as positive samples (i.e., missing links) while $U - E$ as negative samples (i.e., nonexistent links). Intuitively speaking, an algorithm assigning higher likelihoods to positive samples and lower likelihoods to negative samples is considered to be a well-performed algorithm.

Consider the evaluation metrics M_1 and M_2, as well as a series of algorithms $A_1, A_2, \cdots, A_P$ (see figure 1 for an illustration for $P = 5$), for an arbitrary network G, M_1 and M_2 will respectively give each algorithm an evaluation score. According to those scores, we can obtain two rankings of algorithms by M_1 and M_2 (see figure 1A), and then get the correlation of M_1 and M_2 by measuring the correlation of the two rankings. Using the same procedure, we can consider a number of networks $G_1, G_2, \cdots, G_Q$ to calculate the average correlation between any two evaluation metrics (see figure 1B and figure 1C

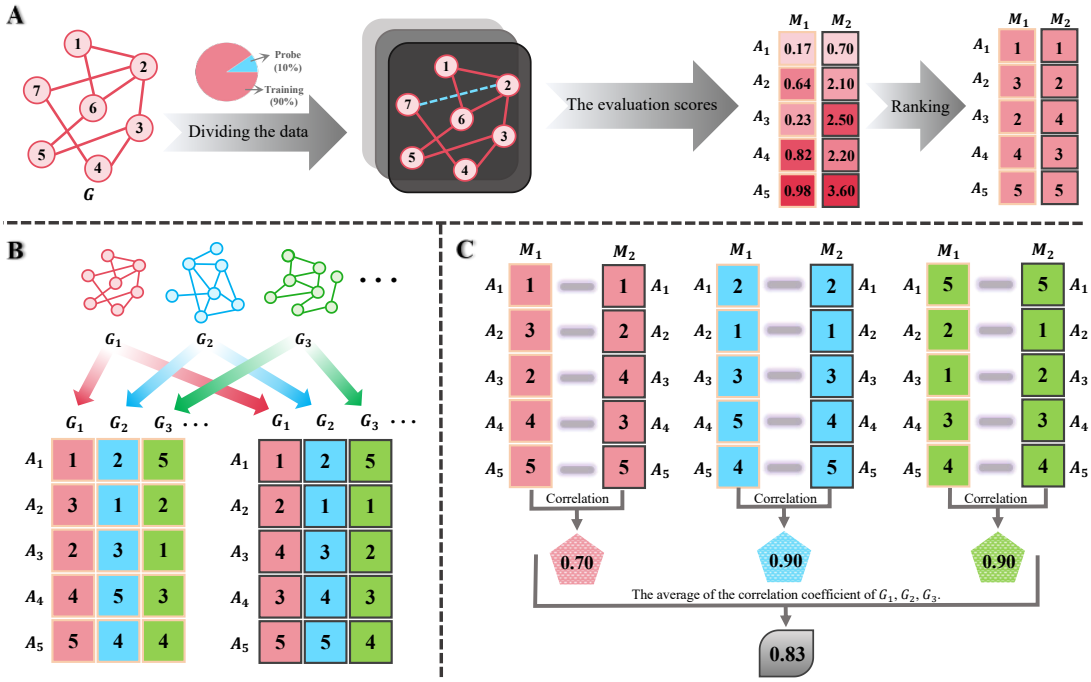

Figure 1: Schematic flowchart of the proposed method to measure the correlation between any two evaluation metrics M_1 and M_2. (A) Initially, the original network is divided into training and probe sets at a ratio, for example 9:1. Next, the evaluation scores of different algorithms $A_i (i = 1, 2, \ldots, P)$ (here we show an example for $P = 5$) are calculated by M_1 and M_2. The average scores can be obtained by multiple implementations with different random divisions of training and probe sets. Based on the average scores, we can get two rankings of algorithms corresponding to M_1 and M_2, respectively. (B) We select a large number of real-world networks $G_1, G_2, \cdots, G_Q$, and for each network G_i and each metric M_j, we can obtain a ranking of the P algorithms (here we show an example for $Q = 3$). (C) We calculate the correlation coefficient of M_1 and M_2 by applying some ranking correlation coefficients (e.g., the Spearman correlation coefficient [58, 59] and the Kendall's τ correlation coefficient [59, 60]) and averaging over the Q selected networks.

for an illustration for $Q = 3$). To reduce the possible fluctuations, we utilize up to $P = 25$ algorithms and up to $Q = 340$ real networks (see Materials and Methods for detailed information).

2.1 Inconsistency among Metrics

We first calculate the correlations between pairwise metrics by using the above-mentioned framework. The following 12 metrics are under consideration: Precision [61], Recall [61], Accuracy [62], Specificity [63], F1-measure [64], Youden Index [65], MCC [48], AUC [43], AUPR [46], AUC-Precision [33], NDCG [49], and AUC-mROC [55]. The first seven are threshold-dependent metrics while the last five are threshold-free. The threshold-dependent metrics depend on some threshold parameters, for example, the number of predicted links k (the top-k links with the highest likelihoods are considered as predicted links) or the threshold likelihood L_c (links with likelihoods larger than L_c are considered as predicted links). Detailed definitions of those metrics are presented in the Materials and Methods.

After obtaining the likelihoods of links in $U - E^T$, different kinds of thresholds (e.g., k and L_c) are essentially equivalent, as the function of any kind of thresholds is to cut all $|U - E^T|$ links into two parts according to their likelihoods. Therefore, we concentrate on the most intuitive and popular threshold k. For any fixed k, we have rigorously proved that all considered threshold-dependent metrics are equivalent, namely the rankings of algorithms by any two threshold-dependent metrics are exactly the same, provided they share the same k. The mathematical proof are shown in the Materials and Methods. As a consequence, we select Precision to represent all threshold-dependent metrics and only discuss Precision later. In addition, since BP is equivalent to Precision at $k = |E^P|$, we will no longer

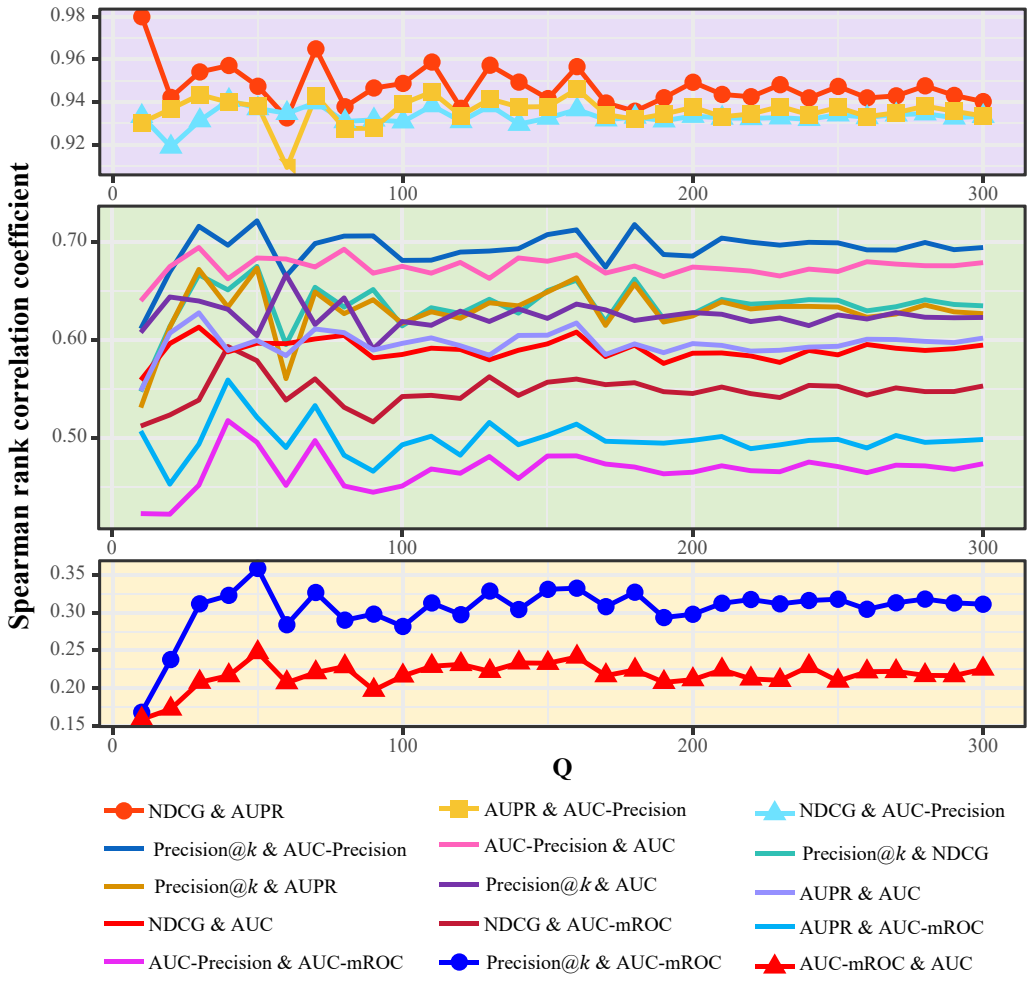

Figure 2: The trend of correlations between metrics as the increase of Q. For each Q, we implement 10 independent runs, where in each run we randomly select Q networks from the collection of 340 real networks. Here the threshold for Precision is set as $k = 0.1 \cdot |U - E^T|$.

specifically analyze BP.

To obtain the average correlations between metrics, we randomly select Q networks from a collection of 340 real networks in disparate fields, and apply the Spearman correlation coefficient to quantify the correlations. Figure 2 shows the correlations in the range $10 \leq Q \leq 300$, where each curve represents a pairwise correlation between two metrics. Obviously, as the increasing of Q, the correlation between any two metrics becomes stable. All 15 pairwise relations are divided into three categories: highly correlated (AUPR & NDCG, AUPR & AUC-Precision, AUC-Precision & NDCG), weakly correlated (Precision & AUC-mROC, AUC & AUC-mROC), and moderately correlated (others). If two metrics are consistent to each other, their correlation should be close to 1. Unfortunately, as shown in figure 2, except for the three highly correlated metric pairs, the correlations for the other 12 pairs are significantly less than 1, indicating that most pairwise metrics are inconsistent to each other. The results on Kendall's τ correlation coefficient, and the results for different dividing ratios of training and probe sets are essentially same to the results reported in figure 2 (see SI Appendix).

2.2 Quandary of Threshold-dependent Metrics

For threshold-dependent metrics, the choices of thresholds are highly relevant. Figure 3 reports how the correlations between Precision and the five threshold-free metrics change for different thresholds k. Except for very small k (see the insets of figure 3), all curves exhibit an overall decaying trend

as the increasing of k, while their decaying patterns are slightly different, namely the Precision-AUC correlation decays slowly and other four curves decay faster. Notice that, every threshold-dependent metric is designed to be meaningful at a relatively small threshold. In contrast, when k approaches its maximum $|U - E^T|$, the score of each metric only depends on the ratio of positive samples to negative samples, irrelevant to the algorithm performance. To summarize, the observed decaying trend results from the fact that Precision (and other threshold-dependent metrics) will become less informative for large k. Furthermore, one can infer that if a threshold-free metric puts higher weights to the top-ranked links, the correlation between Precision and this metric will decay faster as the increasing of k, because the ranking of links in non-top positions has less effect on this metric. This inference is in line with the observations in figure 3, say the correlation for AUC-mROC decays fastest as AUC-mROC assigns very high weights to top positions by applying logarithmic transformations to both coordinates, and the correlation for AUC decays most slowly since AUC considers the overall advantage of positive samples and is less sensitive to top-ranked links.

As indicated by figure 3, the value of threshold k largely impacts the ranking of algorithms, so how to determine k is still a puzzle needing to be solved. As Precision is originally designed to evaluate the early retrieval performance [47, 66], namely to measure the accuracy of a very few top-ranked predictions, k should be much smaller comparing with the total number of potential links $|U - E^T|$. At the same time, when k is very small, the correlations between Precision and some threshold-free metrics (i.e., AUPR, AUC-Precision and NDCG) are very high, all larger than 0.8. Hence Precision at a very small k provides less additional information to AUPR, AUC-Precision and NDCG. If we choose a large k, Precision itself will be less meaningful, though it seems to be more informative at the presence of some threshold-free metrics. Therefore, behind the observations in figure 3 is a quandary in determining the threshold: it should not be small, it should not be large, and it cannot be dug out from the data. In a word, we do not suggest the usage of threshold-dependent metrics if we do not have any clues to determine the threshold. In contrast, if some certain thresholds are meaningful for a specific problem, we can choose one threshold-dependent metric at these thresholds that best fits the practical requirement. For example, if in a e-commercial website, each user will be recommended eight produces (this task can be considered as link prediction in user-product bipartite networks), and the recommender system care most about the click rate, we can choose Precision at $k = 8$.

2.3 Correlation Graph Analysis

To obtain the stable correlations, for each pair of metrics under consideration, we implement 10 independent runs, and for each run we randomly select $Q = 300$ networks from the collection of 340 real networks. The average correlations over 300 networks and 10 runs are presented as a histogram in figure 4, ranked in a descending order. The corresponding correlation graph is shown in the top-right corner of figure 4, which is a complete graph (also called clique or fully connected network in the literature) with metrics being nodes and strengths of correlations being link weights. As we have already analyzed the threshold-dependent metrics in the above subsection, here we do not discuss them again but only draw the correlation graph with a specific case $\rho = 0.1$.

The most noticeable structure in the correlation graph is the purple triangle {AUPR, AUC-Precision, NDCG}, wherein all pairwise correlations are very high (with an average value 0.936). As a consequence, we suggest only choose one of these three metrics to avoid redundant computation. The correlations between AUC and the above three metrics are of moderate strength, so AUC is still informative even at the presence of some of the three metrics. Therefore, AUC should be considered as one metric for algorithm evaluation. Another conspicuous observation is that AUC-mROC is relatively weakly correlated with other metrics. Indeed, among the 15 pairwise correlations, the least five (i.e., those with the weakest correlations) are all related to AUC-mROC. On the one hand, it is a good point as AUC-mROC is information-rich even at the presence of other metrics, on the other hand, it is a dangerous signal as AUC-mROC will produce a ranking of algorithms probably largely different from all other widely and longly used metrics. This observation is closely related to the individualistic feature of AUC-mROC, namely it pays great attention to a few top-ranked predictions. Therefore, our suggestion is to use AUC-mROC in the scenario where only a limited number of predictions are relevant and the values of those predictions decay fast with their positions. For example, one may search for news, products and friends in some websites, and the search returns are of limited number (the results not appearing in the first page are rarely to be visited) and decaying weights (the click rate decays fast with rank). In such kind of scenarios, AUC-mROC could be relevant.

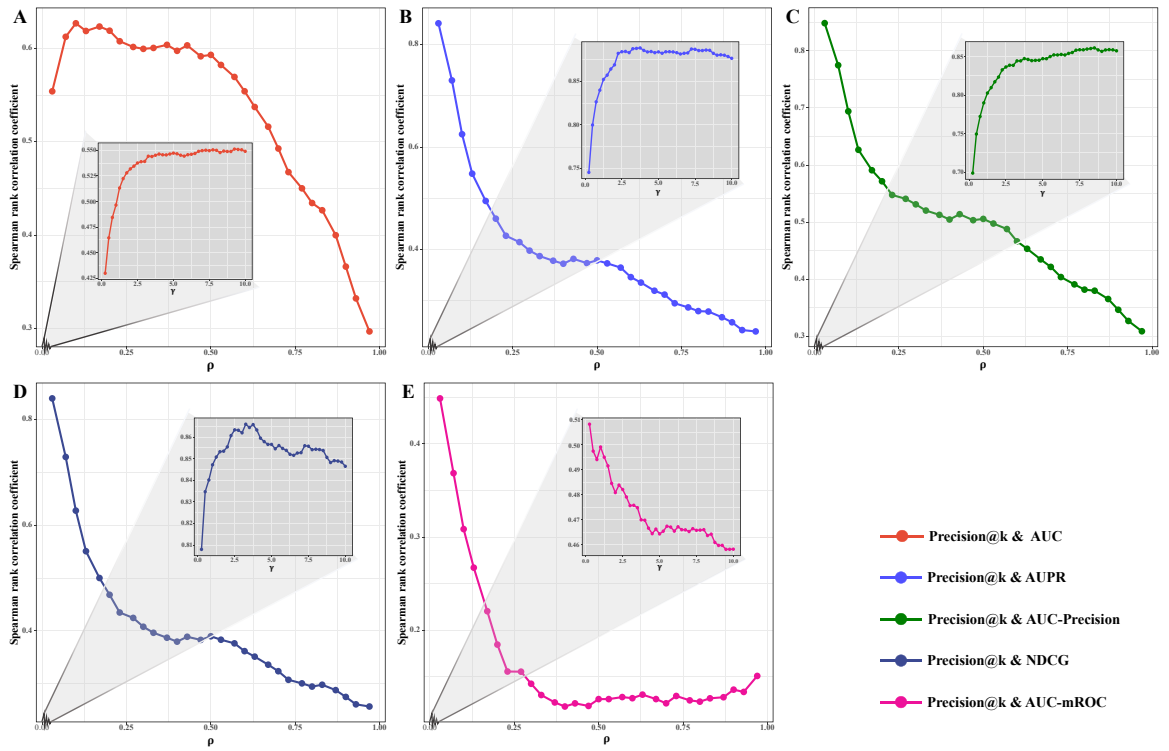

Figure 3: The change of correlations between Precision@k and threshold-free metrics for varying k. In the main plots, we set $k = \rho|U - E^T|$, and in the insets, we set $k = \gamma|E^P|$. The average Spearman rank correlation coefficients correspond to $Q = 300$. (A)-(E) respectively show the cases for AUC, AUPR, AUC-Precision, NDCG, and AUC-mROC.

3 Discussion

In this paper, we have implemented extensive experiments involving 340 real networks and 25 algorithms to analyze the consistency among 12 well-known evaluation metrics. As we have proved the essential equivalence of the seven threshold-dependent metrics, our analyses focus on one representative threshold-dependent metric, Precision, and the five threshold-free metrics. Here we emphasize three important observations from the experiments. Firstly, there exists significant and robust inconsistency among evaluation metrics, that is to say, different metrics may provide different rankings of algorithms. Secondly, the ranking of algorithms produced by a threshold-dependent algorithm is sensitive to the threshold k, and with the increasing of k, the correlation between any threshold-dependent metric and any of the five threshold-free metrics displays an overall decaying trend. The decaying speeds associated with different threshold-free metrics are different: the one with AUC-mROC is the fastest while the one with AUC is the slowest. Thirdly, all pairwise correlations within the set {AUPR, AUC-Precision, NDCG} are very high (with an average value 0.936), the correlations between AUC and the above three metrics are of moderate strengths, and the correlations between AUC-mROC and other metrics are weakest.

The above observations are robust to different settings. Firstly, the results are not sensitive to the ratio of training set to probe set. Here, we only report results with $|E^T| : |E^P| = 9 : 1$, while figure 5 (see SI Appendix) shows clearly that such ratio has negligible effect on the metrics' pairwise correlations. Secondly, regarding the ranking correlation coefficients, we consider another famous one, say the Kendall's τ correlation coefficient. As shown in figure 6 (see SI Appendix), the correlations measures by the Spearman rank correlation coefficient and the Kendall's τ correlation coefficient exhibit completely the same trend.

For any pair of evaluation metrics, in the calculation of their average correlation over a large number of networks, there are two different methods to aggregate the batch data. The first is to get the ranking of algorithms as well as the corresponding correlation for each network, and them to average those

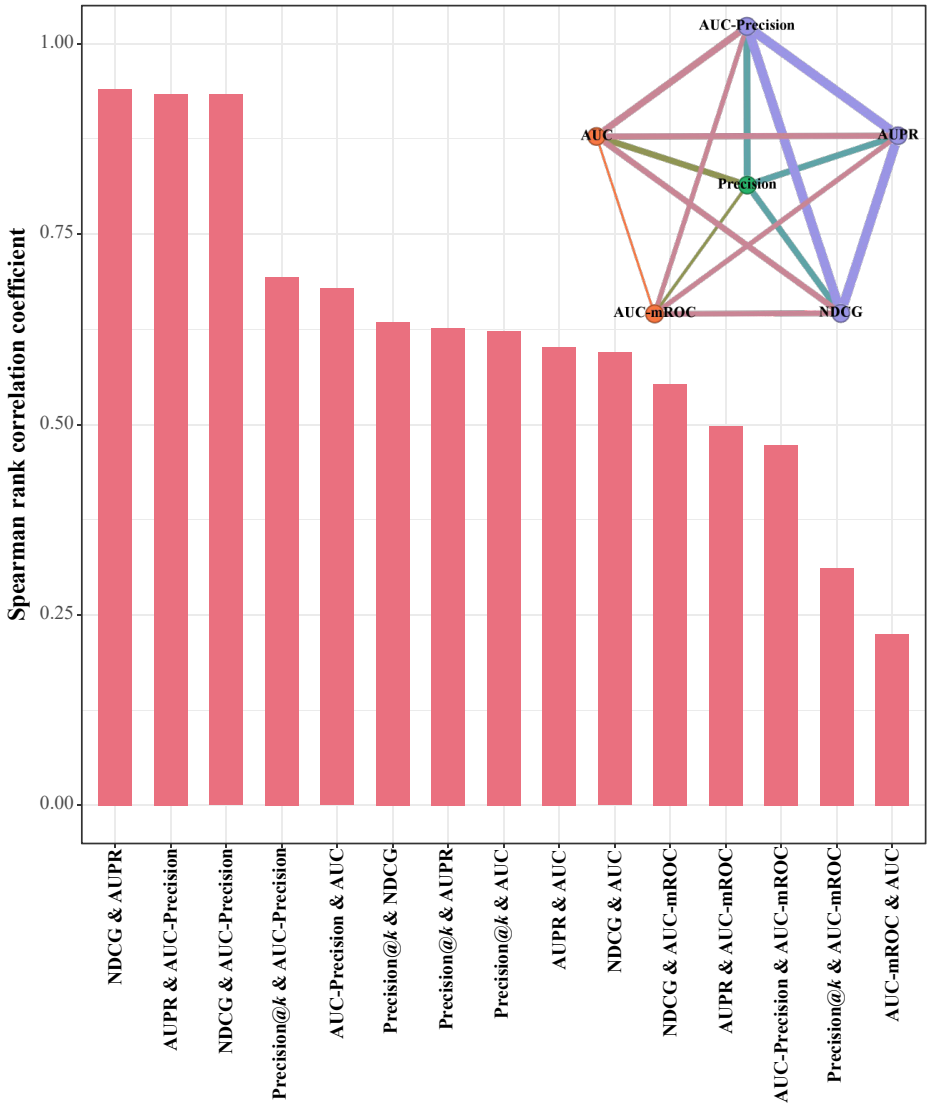

Figure 4: The Spearman rank correlation coefficients for all metric pairs, averaged over 10 independent runs and 300 selected networks in each run. For Precision, the threshold is set as $k = 0.1 \cdot |U - E^T|$. The top-right corner shows the corresponding correlation graph, with the thickness of each link representing the strength of correlation.

correlations over the Q selected networks. The second is to get the ranking of algorithms for each network at first, and then to obtain a mean rank of each algorithm by averaging its ranks over the Q selected networks, and lastly to calculate the correlation between the two vectors of mean ranks of algorithms produced by the two considered metrics. The first method is more intuitive and thus utilized in this paper, with its detailed procedure presented in figure 1. The schematic flowchart of the second method is shown in figure 7 (see SI Appendix, and only the different part from figure 1 is illustrated). Readers should be aware of that the two methods may result in different relationships between the average correlation and the number of selected networks Q. For example, if the average correlation of two metrics M_1 and M_2 will stabilize at about 0.5 for large Q based on the first method, it does not imply that the average correlation of M_1 and M_2 will converge in the large limit of Q or will approach to 0.5 if it converges. It is because there exists a rarely observed but mathematically possible situation that two metrics are indeed highly consistent to each other, but their correlation is not very high for a typical network because there is some unknown and unbiased noises that randomly

perturb the evaluation scores. Therefore, the correlation for each network is not very high and thus the average correlation obtained by the first method is not very high too. In contrast, the correlation between mean ranks can be very high, because when more and more networks are taken into account, the random effects due to unbiased noises tend to cancel each other out. To be more intuitive, we consider a toy model where two metrics, denoted as X and Y, are consistent to each other and both assign an evaluation score j to the jth algorithm. However, there exists some random noises that perturb the evaluation scores, say $x_{ij} = j + \sigma_{ij}$ and $y_{ij} = j + \eta_{ij}$ ($i = 1, 2, \ldots, Q$ and $j = 1, 2, \ldots, P$) for the ith network and jth algorithm, where the noises σ_{ij} and η_{ij} are independently generated from a uniform distribution $U(0, P)$. Figure 8A and figure 8B respectively shows the results obtained by the first and second methods, with $P = 100$ is fixed and Q is up to 500. One can observe that, based on the first method, the average correlation will converge to about 0.49 in the large limit of Q, while based on the second method, the correlation of mean ranks will approach to 1 for large Q, because the random effect vanishes then. As a consequence, we can confidently claim that some evaluation metrics are essentially inconsistent to each other only if both the first and second methods give similar and supportive results. As shown in figure 9 (see SI Appendix), the second method produces qualitatively the same and quantitatively very close results to the first method. Therefore, we can arrive a more believable conclusion that the observed inconsistency among evaluation metrics essentially underlies the definitions of metrics, which does not simply result from some external randomness and thus cannot be eliminated by any statistical skills.

After extensive experiments and analyses, we eventually arrive to four suggestions about how to select evaluation metrics in link prediction. (i) Despite recent debates, we still recommend AUC as one metric because it has moderate correlations to most metrics and thus can provide additional information to other metrics. (ii) One (and no more than one) of AUPR, AUC-Precision and NDCG should be chosen as a metric. (iii) If we don't have any clues to determine the threshold, it is better not to use threshold-dependent metrics, while if for a specific problem, some thresholds are meaningful, we can choose one (and no more than one) threshold-dependent metric with those thresholds. (iv) To use AUC-mROC in the scenario where only a limited number of predictions are relevant and the values of those predictions decay fast with their positions.

In general, during the early stages of a discipline's development, exploratory work tends to be more attractive than reflective work, hence the majority of scientists typically allocate their primary efforts to exploring new frontiers. However, once the discipline reaches a certain level of maturity, reflective work becomes essential; otherwise, the defects in the foundation underlying a taller and taller building will lead to greater losses. Link prediction is a young and niche branch of network science. Study in link prediction is very active, with thousands of algorithms being proposed in the past two decades. In comparison, reflective and critical studies are rare [45, 57, 67, 68]. Now is the time for us to reexamine the fundamentals of link prediction research, with a central problem being how to evaluate whether an algorithm is good or bad, or how to compare which of two algorithms performs better. The solution to this problem may be a kind of guideline that we need to follow in the later studies, just as the double-blind principle in medical experiments. Such guideline should clarify at least four issues. (i) **How to sample the probe set**? It is natural to use random sampling [3] and temporal sampling [67] to get probe set for missing link prediction and future link prediction, respectively. But there are still some technical details that need to be addressed. For example, how to deal with the situations if the removal of links lead to an unconnected network (this issue becomes more serious for higher-order link prediction [69]) or some nodes only appear in the probe set (i.e., all links associated with these nodes are allocated to the probe set). The current approaches are often reasonable yet ad hoc, and thus, we need to assess the extent to which these approaches affect the evaluation of algorithms. For certain specific purposes, there are some other sampling methods. For example, negative sampling method [70] that samples a set of non-existent links with comparable size to the probe set is proposed to manage the cases where the number of missing links is extremely smaller than the number of non-existent links (e.g., for very sparse networks or higher-order networks), and cold sampling method [71] that prefers to sample probe links with low-degree ends since in many practical applications to dig out potential interactions between unpopular nodes is more informative and valuable. These less common sampling methods may have subtle but yet unknown relations to algorithm evaluation, meaning that the appropriate methods and metrics for algorithm evaluation can differ under different sampling methods. Very recently, He et al. [72] tested 20 different sampling methods and found that different link prediction algorithms exhibit significant differences in accuracy

contingent upon the sampling methods. Therefore, the fairness, scope of application and potential impacts of sampling methods requires further analysis and validation. (ii) **How to determine the model parameters**? An undoubtable principle is any information contained in the probe set cannot be used to train the model parameters. However, in the literature, a commonly-used but incorrect method is to obtain the so-called optimal parameter(s) by comparing the prediction with the probe set. This is largely unfair to parameter-free algorithms, while algorithms that are prone to overfitting will get unjustifiable benefit. Accordingly, on the one hand, in the future studies, researchers should train their model parameters using only the information in E^T (e.g., by further dividing E^T into two parts), on the other hand, maybe more important, we have to reevaluate known algorithms in the above-mentioned fair way. We guess those algorithms that are highly sensitive to parameters and inherently prone to overfitting will exhibit decreased performance, while the relative performance of parameter-free algorithms or those with strong generalization capabilities tends to increase. (iii) **How to select proper evaluation metrics**? This is a difficult question to answer. While this paper does not provide a complete answer, it raises the value and urgency of the question. The four specific suggestions presented in this paper focus solely on maximizing the informational content of selected metrics without considering the rationality of these metrics themselves. The recommendations may be different if we consider different aspects. For example, if we intend to encourage the early retrieval ability, NDCG and AUC-mROC are good candidates [55], while if we emphasize on the discriminating ability, AUC and NDCG are superior [45, 56]. A fesiable and useful answer may be a combination of suggestions for general tasks and suggestions accounting for some special conditions, such as the data distributions (e.g., the extremely imbalanced learning) and network organization principles (e.g., the higher-order link prediction). (iv) **How many and how large networks we should use**? In most early studies, only a very few networks (usually of small sizes) are utilized to evaluate the algorithm performance. In comparison, the experiment reported by Ghasemian *et al.* [41] involves 550 real-world networks from diverse fields and of varying sizes. After that, experiments involving a huge number of real-world networks become more popular [68,73,74]. However, we still lack an analytical or statistical answer as to how many networks and of what size are necessary to obtain a reliable assessment of an algorithm's performance.

This paper only provides a tiny step towards the answer to the third question. However, we believe that the value of this paper is substantial; not only in its provision of four constructive suggestions to help researchers select proper metrics to quickly and accurately evaluate algorithm performance, but also, and perhaps more significantly, it compels us to reevaluate the validity of previously known results. In addition, as link prediction is a kind of binary classification problem, our perspectives and methods could be extended to the selection of evaluation metrics for classification.

4 Materials and Methods

4.1 Algorithms of Link Prediction

In this work, we consider 25 algorithms. Some are well-known and some are very recently proposed. Table 1 lists those algorithms, together with the corresponding references, where readers can find more details.

4.2 Evaluation Metrics

This subsection will introduce the 12 considered metrics, say Precision, Recall, Accuracy, Specificity, F1-measure, Youden, MCC, AUC, AUPR, AUC-Precision, NDCG, and AUC-mROC. The first seven are threshold-dependent metrics and the last five are threshold-free metrics.

Without loss of generality, we assume that each algorithm will assign a score s_{ij} to characterize the existence likelihood of any potential link $(i, j) \in U - E^T$, and all links in $U - E^T$ are ranked in a descending order of their scores. The threshold k cuts the set of potential links into two parts: the top-k ranked links are predicted missing links, while the others are predicted non-existent links. As link prediction is a binary classification problem, we can use the confusion matrix to formulate threshold-dependent metrics. In the confusion matrix, all samples are classified into four categories based on whether they are positive samples or negative samples, and whether they are correctly predicted. These four categories are: true positive (TP), where a positive sample is correctly predicted as positive; false

Table 1: List of link prediction algorithms considered in this paper, with abbreviations showing in the brackets.

Algorithms	References
Common Neighbor Index (CN)	[75]
Resource Allocation Index (RA)	[29]
Local Path Index (LP)	[29]
Adamic-Adar Index (AA)	[76]
Preferential Attachment Index (PA)	[77]
Jaccard Index	[78]
Average Commute Time (ACT)	[79]
Sim Index	[80]
Length Three (L3)	[39]
Adjacency Three (A3)	[39]
Katz Index	[81]
Liner Optimization (LO)	[82]
Salton Index	[66]
Sørenson Index	[83]
Hub Promoted Index (HPI)	[84]
Hub Depressed Index (HDI)	[3]
Local Random Walk (LRW)	[31]
Superposed Random Walk (SRW)	[31]
Leicht-Holme-Newman-1 Index (LHN-1)	[85]
Matrix Forest Index (MFI)	[86]
Local Naive Bayes based Adamic-Adar Index (LNBAA)	[87]
Local Naive Bayes based Resource Allocation Index (LNBRA)	[87]
Salton Cosine Similarity (S1)	[66]
Controlling the Leading Eigenvector (CLE)	[88]
Common neighbor and Centrality based Parameterized Algorithm (CCPA)	[89]

positive (FP), where a negative sample is incorrectly predicted as positive; true negative (TN), where a negative sample is correctly predicted as negative; and false negative (FN), where a positive sample is incorrectly predicted as negative.

Next we can describe the threshold-dependent metrics using the language of confusion matrix. **Precision** is proportion of true positives to all predicted positives [61]. **Recall** measures the ratio of true positives to the total number of positives [61]. **Accuracy** quantifies the proportion of correctly classified instances out of the total instances [62]. **Specificity** measures the ratio of true negatives to the total number of negatives [63]. **F1-measure** is the harmonic mean of Precision and Recall [64]. **Youden Index** is defined as the sum of Recall and Specificity minus 1, which captures the overall performance of a diagnostic test [65]. **MCC** takes into account the roles of all elements in the confusion matrix, which is particularly useful when dealing with imbalanced learning problem [48]. Accordingly, the mathematically formulas for these threshold-dependent metrics are as follows.

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{k}, \tag{1}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{|E^P|}, \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} = \frac{TP + TN}{|U - E^T|}, \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} = \frac{TN}{|U - E|}, \tag{4}$$

$$F1 = 2\left(\frac{1}{Precision} + \frac{1}{Recall}\right)^{-1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{5}$$

$$Youden = Recall + Specificity - 1, \tag{6}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{7}$$

AUC measures the ability of the model to discriminate between positive and negative classes across all possible threshold values. It delegates the are under the Receiver Operating Characteristic (ROC) curve [43]. The range of AUC is $[0, 1]$, where a higher value indicates better performance. As AUC is equavalent to the probability that a randomly selected positive sample (i.e., missing link) is scored higher than a randomly selected negative sample (i.e., non-existent link), we can obtain the approximation of AUC through directly comparing positive and negative samples. If we randomly compare n positive-negative pairs, and there are n_1 times the missing link having higher score and n_2 times the missing link and nonexistent link having the same score, then AUC is approximated as

$$AUC = \frac{n_1 + 0.5n_2}{n}. \tag{8}$$

AUC will approach 0.5 if the algorithm produce a random classification, and thus to what extent AUC exceeds 0.5 indicates how much better the algorithm performs than pure chance. **AUPR** measures the area under the precision-recall curve, which plots Precision (on the Y-axis) against Recall (on the X-axis) for different threshold values [46]. A higher AUPR value indicates better performance. If the positions of the $|E^P|$ missing links are $r_1 < r_2 < \cdots < r_{|E^P|}$ in the $|U - E^T|$ ranked links, then AUPR can be calculated as

$$AUPR = \frac{1}{2|E^P|}\left(\sum_{i=1}^{|E^P|}\frac{i}{r_i} + \sum_{i=1}^{|E^P|}\frac{i}{r_{i+1}-1}\right), \tag{9}$$

where $r_{|E^P|+1}$ is defined as $|U-E^T|+1$. Analogously, **AUC-Precision** is the area under the threshold-precision curve, which plots Precision (on the Y-axis) against the threshold k (on the X-axis) for different threshold values [33]. **NDCG** assigns larger weights to higher positions, normalized by the ideal discounted cumulative gain, as [49]

$$NDCG = \sum_{i=1}^{|E^P|}\frac{1}{\log_2(1+r_i)} \Big/ \sum_{l=1}^{|E^P|}\frac{1}{\log_2(1+l)}, \tag{10}$$

where the contribution of a missing link at position r is $\frac{1}{\log_2(1+r)}$. **AUC-mROC** is a variant of AUC by transforming both axes of the ROC curve using logarithmic scale [55]. It applies another transformation to ensure the a random classification also lies in the diagonal line. The finalized horizontal and vertical coordinates are defined as

$$mFPR = \log_J(1 + FP) \tag{11}$$

and

$$mTPR = mFPR + [\log_Z(1 + TP) - H] \cdot (1 - H)^{-1} \cdot (1 - mFPR), \tag{12}$$

where $J = 1 + |U - E|$, $Z = 1 + |E^P|$, and $H = \log_Z(1 + FP \cdot \frac{Z-1}{J-1})$. The AUC-mROC is the area under the above transformed curve.

4.3 Equivalence of Threshold-dependent Metrics

Theorem. *Given the threshold k, metrics in the set Ω={Precision, Recall, F1, Specificity, Youden, Accuracy, MCC} are equilevent to each other, namely any two metrics in Ω will give exactly the same rankings of algorithms.*

Proof. Consider two link prediction algorithms A_1 and A_2, and any two metrics $M_i, M_j \in \Omega$, denote $M_i(A_1)$ the evaluation score A_1 received from M_i, then the theorem can be unfolded to the following three propositions: (1) if $M_i(A_1) < M_i(A_2)$, then $M_j(A_1) < M_j(A_2)$; (2) if $M_i(A_1) > M_i(A_2)$, then $M_j(A_1) > M_j(A_2)$, (3) if $M_i(A_1) = M_i(A_2)$, then $M_j(A_1) = M_j(A_2)$. In subsequent proof, it is assumed that k is given as a constant.

Obviously, from proposition (1), we can deduce propositions (2) and (3): (1)⇒(2) can be obtained by exchanging A_1 and A_2, (1)⇒(3) can be proved by contradiction. Therefore, to prove the equivalence between two metrics M_i and M_j, we only need to show that for any two algorithms A_1 and A_2, if $M_i(A_1) < M_i(A_2)$, then $M_j(A_1) < M_j(A_2)$. For convenience, we use Precision as the central metric, and then prove the following six inequalities provided the condition $Precision(A_1) < Precision(A_2)$: (i) $Recall(A_1) < Recall(A_2)$; (ii) $F1(A_1) < F1(A_2)$; (iii) $Specificity(A_1) < Specificity(A_2)$; (iv) $Youden(A_1) < Youden(A_2)$; (v) $Accuracy(A_1) < Accuracy(A_2)$; (vi) $MCC(A_1) < MCC(A_2)$.

According to the definition in Eq. (1), as k is fixed, the condition $Precision(A_1) < Precision(A_2)$ is equivalent to $TP(A_1) < TP(A_2)$, so that a smart way to prove the above inequalities is expressing elements in the confusion matrix by TP and other constants. Using the following evident relationships

$$\begin{aligned} &FP + TP = k, \\ &FN + TP = |E^P|, \\ &TN + FP = |U - E|, \\ &TN + FN = |U - E^T| - k, \end{aligned} \tag{13}$$

we have

$$\begin{aligned} &FP = k - TP, \\ &FN = |E^P| - TP, \\ &TN = |U - E| - k + TP. \end{aligned} \tag{14}$$

Next, we prove the six inequalities one by one. Inequality (i) is evident as

$$Recall(A_1) = \frac{TP(A_1)}{|E^P|} < \frac{TP(A_2)}{|E^P|} = Recall(A_2). \tag{15}$$

If $Precision(A_1) < Precision(A_2)$ and $Recall(A_1) < Recall(A_2)$, it is very clear that the harmonic mean of $Precision(A_1)$ and $Recall(A_1)$ is also smaller than the harmonic mean of $Precision(A_2)$ and $Recall(A_2)$, say the inequality (ii) holds. Substituting Eq. (14) to Eq. (4), we have

$$Specificity(A_1) = \frac{|U - E| - k + TP(A_1)}{|U - E|} < \frac{|U - E| - k + TP(A_2)}{|U - E|} = Specificity(A_2), \tag{16}$$

namely the inequality (iii) holds. If $Precision(A_1) < Precision(A_2)$, we can deduce that $Recall(A_1) < Recall(A_2)$ and $Specificity(A_1) < Specificity(A_2)$ by inequality (i) and inequality (iii) respectively, hence $Youden(A_1) < Youden(A_2)$ according to the definition Eq. (6), namely the inequality (iv) holds. Combining Eq. (3) and Eq. (14), we have

$$Accuracy = \frac{|U - E| - k + 2TP}{|U - E^T|}, \tag{17}$$

so that

$$Accuracy(A_1) = \frac{|U - E| - k + 2TP(A_1)}{|U - E^T|} < \frac{|U - E| - k + 2TP(A_2)}{|U - E^T|} = Accuracy(A_2), \tag{18}$$

namely the inequality (v) holds. According to Eq. (14), the numerator of MCC is

$$TP \cdot TN - FP \cdot FN = TP(|U - E| - k + TP) - (k - TP)(|E^P| - TP) = (|U - E^T|TP - k|E^P|), \tag{19}$$

and the denominator of MCC can be expressed by Eq. (13), therefore

$$MCC = \frac{|U - E^T|TP - k|E^P|}{k|E^P||U - E|(|U - E^T| - k)}, \tag{20}$$

and thus

$$MCC(A_1) = \frac{|U - E^T|TP(A_1) - k|E^P|}{k|E^P||U - E|(|U - E^T| - k)} < \frac{|U - E^T|TP(A_2) - k|E^P|}{k|E^P||U - E|(|U - E^T| - k)} = MCC(A_2), \tag{21}$$

namely the inequality (vi) holds. □

4.4 Ranking Correlation Coefficients

This paper applies two classical coefficients to measure the correlation between two rankings, say Spearman rank correlation coefficient [58,59] and Kendall's τ correlation coefficient [59,60]. Denoting R_{iu} the rank of the uth algorithm by the ith metric, the Spearman rank correlation coefficient between two rankings produced by metrics i and j is

$$r_{ij} = \frac{\sum_{u=1}^{P}(R_{iu} - \bar{R}_i)(R_{ju} - \bar{R}_j)}{\sqrt{\sum_{u=1}^{P}(R_{iu} - \bar{R}_i)^2} \cdot \sqrt{\sum_{u=1}^{P}(R_{ju} - \bar{R}_j)^2}}, \tag{22}$$

where $P = 25$ is the number of algorithms under consideration and $\bar{R}_i$ is the average rank. Clearly, the Spearman rank correlation coefficient lies in the range $-1 \leq r_{ij} \leq 1$. The Kendall's τ measurs the strength of association of the cross tabulations. Considering two algorithms u and v ($1 \leq u, v \leq P$) and two metrics i and j, if $R_{iu} > R_{iv}$ and $R_{ju} > R_{jv}$, or $R_{iu} < R_{iv}$ and $R_{ju} < R_{jv}$, we say the pair (u, v) is concordant, if $R_{iu} > R_{iv}$ but $R_{ju} < R_{jv}$, or $R_{iu} < R_{iv}$ but $R_{ju} > R_{jv}$, we say the pair (u, v) is discordant, and if $R_{iu} = R_{iv}$ or $R_{ju} = R_{jv}$, we say the pair (u, v) is tied. Counting all $P(P-1)/2$ pairs, the Kendall's τ reads

$$\tau_{ij} = \frac{2(N_C - N_D)}{P(P-1)}, \tag{23}$$

where N_C is the number of concordant pairs, and N_D is the number of discordant pairs.

4.5 Data and Codes

The 340 real-world networks mainly come from the two public datasets (http://konect.cc/networks/ and https://networkrepository.com/networks.php). The details of the data and code are deposited in GitHub: https://github.com/98YiLin/IEMLP.git.

References

[1] Barabási A-L. 2016. Network science. *Cambridge University Press*.

[2] Newman MEJ. 2018. Networks. *Oxford University Press*.

[3] Lü L, Zhou T. 2011. Link prediction in complex networks: a survey. *Physica A*. 390(6):1150-1170.

[4] Wang P, Xu B, Wu Y, Zhou X. 2015. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*. 58(38):011101.

[5] Martínez V, Berzal F, Cubero J-C. 2016. A survey of link prediction in complex networks. *ACM Computing Surveys*. 49(4):69.

[6] Kumar A, Singh SS, Singh K, Biswas B. 2020. Link prediction techniques, applications, and performance: a survey. *Physica A*. 553:124289.

[7] Divakaran A, Mohan A. 2020. Temporal link prediction: a survey. *New Generation Computing*. 38:213-258.

[8] Zhou T. 2021. Progresses and challenges in link prediction. *iScience*. 24(11):103217.

[9] Chen C, Liu Y-Y. 2023. A survey on hyperlink prediction. *IEEE Transactions on Neural Networks and Learning Systems*. (in press).

[10] Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R. 2013. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology and Therapeutics*. 138(3):333-408.

[11] Ding H, Takigawa I, Mamitsuka H, Zhu S. 2014. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Briefings in Bioinformatics*. 15(5):734-747.

[12] Bi Y, Wang P. 2022. Exploring drought-responsive crucial genes in *Sorghum*. *iScience*. 25(11):105347.

[13] Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F. 2012. Friendship prediction and homophily in social media. *ACM Transactions on the Web*. 6(2):9.

[14] L*ü* L, Medo M, Yeung CH, Zhang Y-C, Zhang Z-K, Zhou T. 2012. Recommender systems. *Physics Reports*. 519(1):1-49.

[15] Nagarajan M, Wilkins AD, Bachman BJ, Novikov IB, Bao S, Haas PJ, Terrón-Díaz ME, Bhatia S, Adikesavan AK, Labrie JJ, Regenbogen S, Buchovecky CM, Pickering CR, Kato L, Lisewski AM, Lelescu A, Zhang H, Boyer S, Weber G, Chen Y, Donehower L, Spangler S, Lichtarge O. 2015. Predicting future scientific discoveries based on a networked analysis of the past literature. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press). pp. 2019-2028.

[16] Krenn M, Buffoni L, Coutinho B, Eppel S, Foster JG, Gritsevskiy A, Lee H, Lu Y, Moutinho JP, Sanjabi N, Sonthalia R, Tran NM, Valente F, Xie Y, Yu R, Kopp M. 2023. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence*. 5:1326-1335.

[17] Wang W-Q, Zhang Q-M, Zhou T. 2012. Evaluating network models: a likelihood analysis. *Europhysics Letters*. 98(2):28004.

[18] Zhang Q-M, Xu X-K, Zhu Y-X, Zhou T. 2015. Measuring multiple evolution mechanisms of complex networks. *Scientific Reports*. 5:10350.

[19] Vallès-Català T, Peixoto TP, Sales-Pardo M, Guimerà R. 2018. Consistencies and inconsistencies between model selection and link prediction in networks. *Physical Review E*. 97(6):062316.

[20] Ghasemian A, Hosseinmardi H, Clauset A. 2019. Evaluating overfit and underfit in models of network community structure. *IEEE Transactions on Knowledge and Data Engineering*. 32(9):1722-1735.

[21] Peixoto TP. 2018. Reconstructing networks with unknown and heterogeneous errors. *Physical Review X*. 8(4):041011.

[22] Squartini T, Caldarelli G, Cimini G, Gabrielli A, Garlaschelli D. 2018. Reconstruction methods for networks: the case of economic and financial systems. *Physics Reports*. 757:1-47.

[23] Zhang Y, Zhao J, Wu W, Muscoloni A, Cannistraci CV. 2022. Ultra-sparse network advantage in deep learning via cannistraci-hebb brain-inspired training with hyperbolic meta-deep community-layered epitopology. *Preprints*:2022070139.

[24] Santos FP, Lelkes Y, Levin SA. 2021. Link recommendation algorithms and dynamics of polarization in online social networks. *PNAS*. 118(50):e2102141118.

[25] Hou L, Pan X, Liu K, Yang Z, Liu J, Zhou T. 2023. Information cocoons in online navigation. *iScience*. 26(1):105893.

[26] Hasan MA, Chaoji V, Salem S, Zaki M. 2006. Link prediction using supervised learning. In *Proceedings of SDM06: Workshop on Link Analysis, Counter-Terrorism and Security* (SIAM Press). pp. 798-805.

[27] Liben-Nowell D, Kleinberg J. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*. 58(7):1019-1031.

[28] Clauset A, Moore C, Newman MEJ. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature*. 453:98-101.

[29] Zhou T, L*ü* L, Zhang Y-C. 2009. Predicting missing links via local information. *The European Physical Journal B*. 71:623-630.

[30] Guimerà R, Sales-Pardo M. 2009. Missing and spurious interactions and the reconstruction of complex networks. *PNAS*. 106(52):22073-22078.

[31] Liu W, Lü L. 2010. Link prediction based on local random walk. *Europhysics Letters*. 89(5):58007.

[32] Menon AK, Elkan C. 2011. Link prediction via matrix factorization. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer Press). pp.437-452.

[33] Cannistraci CV, Alanis-Lobato G, Ravasi T. 2013. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific Reports*. 3:1613.

[34] Lü L, Pan L, Zhou T, Zhang Y-C, Stanley HE. 2015. Toward link predictability of complex networks. *PNAS*. 112(8): 2325-2330.

[35] Pan L, Zhou T, Lü L, Hu C-K. 2016. Predicting missing links and identifying spurious links via likelihood analysis. *Scientific Reports*. 6:22955.

[36] Pech R, Hao D, Pan L, Cheng H, Zhou T. 2017. Link prediction via matrix completion. *Europhysics Letters*. 117(3):38002.

[37] Zhang M, Chen Y. 2018. Link prediction based on graph neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (NeurIPS Press). pp. 5171-5181.

[38] Benson AR, Abebe R, Schaub MT, Jadbabaie A, Kleinberg J. 2018. Simplicial closure and higher-order link prediction. *PNAS*. 115(48):E11221-E11230.

[39] Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, Bian W, Kim D-K, Kishore N, Hao T, Calderwood MA, Vidal M, Barabási A-L. 2019. Network-based prediction of protein interactions. *Nature Communications*. 10:1240.

[40] Kitsak M, Voitalov I, Krioukov D. 2020. Link prediction with hyperbolic geometry. *Physical Review Research*. 2(4):043113.

[41] Ghasemian A, Hosseinmardi H, Galstyan A, Airoldi EM, Clauset A. 2020. Stacking models for nearly optimal link prediction in complex networks. *PNAS*. 117(38):23393-23400.

[42] Wang H, Cui Z, Liu R, Fang L, Sha Y. 2023. A multi-type transferable method for missing link prediction in heterogeneous social networks. *IEEE Transactions on Knowledge and Data Engineering*. 35(11):10981-10991.

[43] Hanely JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 143(1):29-36.

[44] Bradley AP. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*. 30(7):1145-1159.

[45] Zhou T. 2023. Discriminating abilities of threshold-free evaluation metrics in link prediction. *Physica A*. 615(1):128529.

[46] Davis J, Goadrich M. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (ACM Press). pp. 233-240.

[47] Herlocker JL, Konstan JA, Terveen LG, Riedl JT. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*. 22(1):5-53.

[48] Mattews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*. 405(2):442-451.

[49] Jšrvelin K, Kekäläinen J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*. 20(4):422-446.

[50] Wang Y, Wang L, Li Y, He D, Chen W, Liu T-Y. 2013. A theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory* (COLT Press). pp. 25-54.

[51] Yang Y, Lichtenwalter RN, Chawla NV. 2015. Evaluating link prediction methods. *Knowledge and Information Systems*. 45:751-782.

[52] Saito T, Rehmsmeier M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 10(3):e0118432.

[53] Del Genio CI, Gross T, Bassler KE. 2011. All scale-free networks are sparse. *Physical Review Letters*. 107(17):178701.

[54] Menand Nicolas, C. Seshadhri. 2024. Link prediction using low-dimensional node embeddings: the measurement problem. *PNAS*. 121(8):e2312527121.

[55] Muscoloni A, Cannistraci CV. 2022. Early retrieval problem and link prediction evaluation via the area under the magnified ROC. Preprints: 2022090277.

[56] Jiao X, Wan S, Liu Q, Bi Y, Lee Y-L, Xu E, Hao D, Zhou T. 2024. Comparing discriminating abilities of evaluation metrics in link prediction. arXiv: 2401.03673.

[57] Muscoloni A, Cannistraci CV. 2023. "Stealing fire or stacking knowledge" by machine intelligence to model link prediction in complex networks. *iScience*. 26(1):105697.

[58] Spearman C. 1987. The proof and measurement of association between two things. *The American Journal of Psychology*. 100(3/4):441-471.

[59] Lü J, Wang P. 2020. Modeling and analysis of bio-molecular networks. *Springer Singapore Press*.

[60] Kendall MG. 1938. A new measure of rank correlation. Biometrika. 30(1/2):81-93.

[61] Buckland M, Gey F. 1994. The relationship between precision and recall. *Journal of the Association for Information Science and Technology*. 45(1):12-19.

[62] Swets JA. 1988. Measuring the accuracy of diagnostic systems. *Science*. 240(4857): 1285-1293.

[63] Jones KS. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. 28(1): 11-21.

[64] Sasaki Y. 2007. The truth of the F-measure. *Teach Tutor Mater*. 1: 1-5.

[65] Youden WJ. 1950. Index for rating diagnostic tests. *Cancer*. 3(1): 32-35.

[66] Salton G, McGill MJ. 1986. Introduction to modern information retrieval. *McGraw-Hill Book Company Press*.

[67] Lichtenwalter RN, Lussier JT, Chawla NV. 2010. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press). pp. 243-252.

[68] Mara AC, Lijffijt J, Bie T De. 2020. Benchmarking network embedding models for link prediction: are we making progress? In *Proceedings of the 7th IEEE International Conference on Data Science and Adavnced Analytics* (IEEE Press). pp. 138-147.

[69] Kumar T, Darwin K, Parthasarathy S, Ravindran B. 2020. HPRA: Hyperedge Prediction using Resource Allocation. In *Proceedings of the 12th ACM Conference on Web Science* (ACM Press). pp. 135-143.

[70] Kotnis B, Nastase V. 2017. Analysis of the Impact of Negative Sampling on Link Prediction in Knowledge Graphs. arXiv: 1708.06816.

[71] Zhu Y-X, Lü L, Zhang Q-M, Zhou T. 2012. Uncovering missing links with cold ends. *Physica A*. 391: 5769-5778.

[72] He Xie, Ghasemian Amir, Lee Eun, Schwarze Alice. 2024. Link prediction accuracy on real-world networks under non-uniform missing edge patterns. arXiv: 2401.15140.

[73] Muscoloni A, Michieli U, Zhang Y, Cannistraci CV. 2020. Adaptive Network Automata Modelling of Complex Networks. Preprint: 202012.0808.

[74] Zhou T, Lee Y-L, Wang G. 2021. Experimental analysis on 2-hop-based and 3-hop-based link prediction algorithms. *Physica A*. 564: 125532.

[75] Newman MEJ. 2001. Clustering and preferential attachment in growing networks. *Physical Review E*. 64(2):025102.

[76] Adamic LA, Adar E. 2003. Friends and neighbors on the web. *Social Networks*. 25(3):211-230.

[77] Barabási A-L, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. 2002. Evolution of the social network of scientific collaborations. *Physica A*. 311(3-4):590-614.

[78] Jaccard P. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*. 37:241-272.

[79] Liu W, Lü L. 2010. Link prediction based on local randomwalk. *Europhysics Letters*. 89(5):58007.

[80] Chen Y, Wang W, Liu J, Feng J, Gong X. 2020. Proteininterface complementarity and gene duplication improve link prediction of protein-protein interaction network. *Frontiers in Genetics*. 11:291.

[81] Katz L. 1953. A new status index derived from sociometric analysis. *Psychometrika*. 18:39-43.

[82] Pech R, Hao D, Lee Y-L, Yuan Y, Zhou T. 2019. Link prediction via linear optimization. *Physica A*. 528:121319.

[83] Sørensen T.A. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*. 5:1-34.

[84] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L. 2002. Hierarchical organization of modularity in metabolic networks. *Science*. 297(5586):1551-1555.

[85] Leicht EA, Holme P, Newman MEJ. 2006. Vertex similarity in networks. *Physical Review E*. 73(2):026120.

[86] Chebotarev PY, Shamis EA. 1997. A matrix-forest theorem and measuring relations in small social group. *Avtomatika i Telemekhanika*. 9:125-137.

[87] Liu Z, Zhang Q-M, Lü L, Zhou T. 2011. Link prediction in complex networks: a local naïve Bayes model. *Europhysics Letters*. 96(4): 48007.

[88] Lee Y-L, Dong Q, Zhou T. 2021. Link prediction via controlling the leading eigenvector. *Applied Mathematics and Computation*. 411: 126517.

[89] Ahmad I., Akhtar MU, Noor S, Shahnaz A. 2020. Missing link prediction using common neighbor and centrality based parameterized algorithm. *Scientific Reports*. 10:364.

5 Supplemental Information

5.1 Sensitivity Analysis

We first test the impacts of the ratio of the training set to the probe set. In addition to the commonly used ratio [3], say $|E^T| : |E^P| = 9 : 1$, we consider other ratios like 8:2, 7:3 and 6:4, which are also usually used in binary classification. As shown in figure 5, the change in ratio does not affect the correlations between metrics, suggesting the robustness of observations in the main text. We next check whether our results are sensitive to the choice of correlation measures by comparing the Spearman rank correlation coefficient with another well-known coefficient, say the Kendall's τ. As shown in figure 6, they show completely the same trend, indicating that our results are robust to the correlation coefficients.

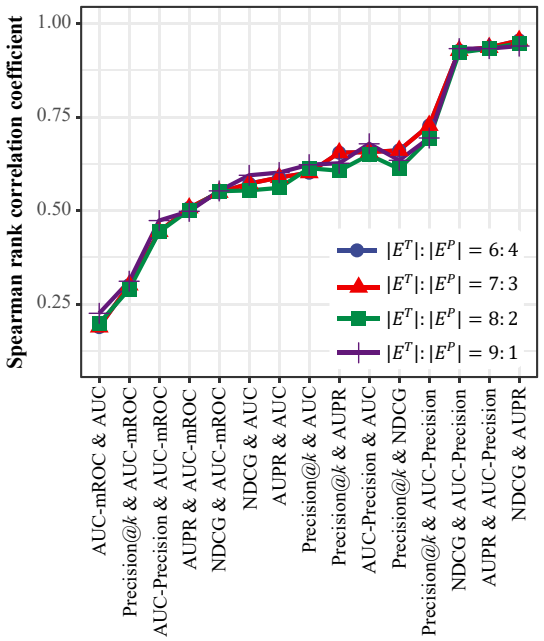

Figure 5: The average pairwise correlations over 300 randomly selected real networks for different splitting ratios of $|E^T|$ to $|E^P|$. The blue, red, green, and purple lines represent the results for $|E^T| : |E^P| = 6 : 4$, $|E^T| : |E^P| = 7 : 3$, $|E^T| : |E^P| = 8 : 2$, and $|E^T| : |E^P| = 9 : 1$, respectively.

5.2 The Alternative Method

Different from the method in figure 1, there is an alternative way to calculate the correlation between metrics based on a large number of real networks. The key point of this method is to first average the ranks of different algorithms over selected networks, and then calculate the Spearman rank correlation coefficient of the mean ranks. The different part of this method from the method applied in the main text is shown in figure 7. Figure 8 shows the results of the toy model introduced in the main text by using the first and second methods. For the first method (see figure 8A), when Q is large enough, the correlation coefficient between X and Y stabilizes at about 0.49. However, for the second method (see figure 8B), the correlation coefficient rapidly increases to 1 as Q increases. Figure 9 reports the results by using the second method for real networks, where the other settings are completely the same to those of figure 4.

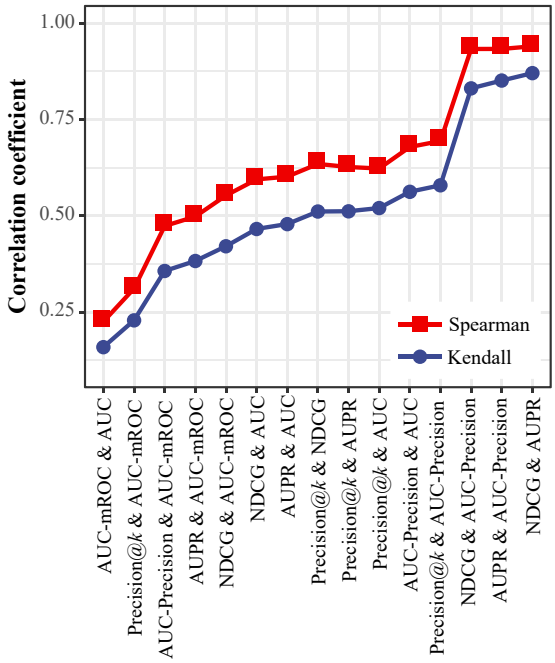

Figure 6: The average pairwise correlations over 300 randomly selected real networks, obtained by different correlation coefficients. The red and blue lines represent the results obtained using the Spearman rank correlation coefficient and the Kendall's τ, respectively.

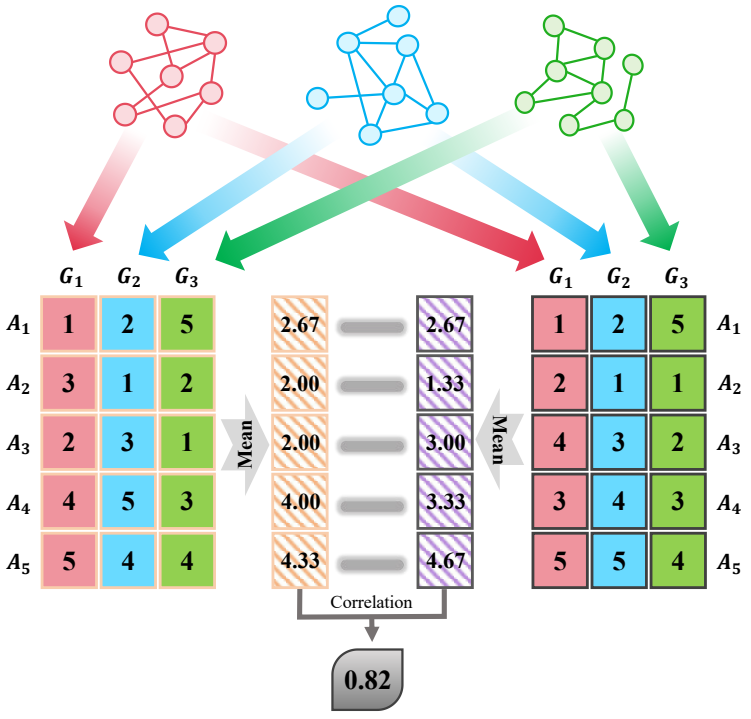

Figure 7: Schematic flowchart of an alternative averaging method to measure the correlation between any two evaluation metrics M_1 and M_2 in for five algorithms ($P = 5$). After obtaining the rankings of algorithms for the Q selected networks (here we show an example for $Q = 3$), we first calculate the mean ranks and then measure the correlation between two vectors of mean ranks.

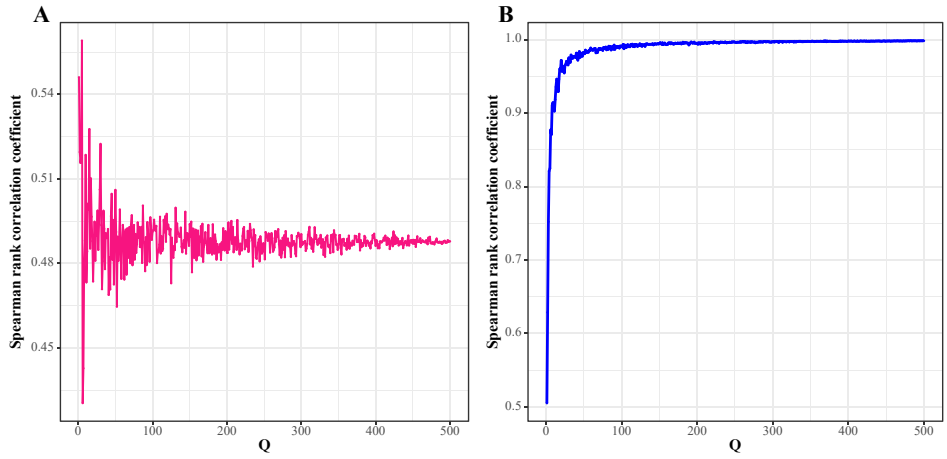

Figure 8: The Spearman rank correlation coefficients between X and Y as the increasing of Q for the toy model using (A) the first method and (B) the second method.

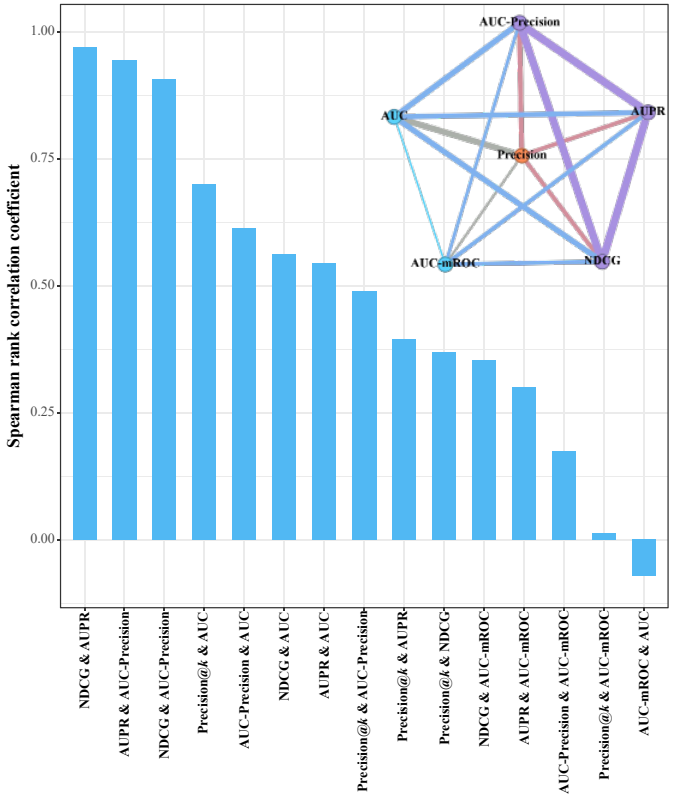

Figure 9: The Spearman rank correlation coefficients for all metric pairs, obtained by the method presented in figure 7, which are averaged over 10 independent runs and 300 selected networks in each run. For Precision, the threshold is set as $k = 0.1 \cdot |U - E^T|$. The top-right corner shows the corresponding correlation graph, with the thickness of each link representing the strength of correlation.