Physica A 610 (2023) 128428

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Discussion

Link prediction based on the mutual information with high-order clustering structure of nodes in complex networks

Yabing Yao [a,*], Tianyu Cheng [a], Xiaoqiang Li [a], Yangyang He [a], Fan Yang [b], Tongfeng Li [c], Zeguang Liu [a], Zhipeng Xu [a]

[a] School of computer and communication, Lanzhou University of Technology, Lanzhou 730050, China
[b] School of computer science and technology, Guangxi University of Science and Technology, Liuzhou 545006, China
[c] Computer College, Qinghai Normal University, Xining 810016, China

ARTICLE INFO

Article history:
Received 21 September 2022
Received in revised form 6 December 2022
Available online 27 December 2022

Dataset link: https://github.com/yabingyao/MHOC4LinkPrediction.git

Keywords:
Complex networks
Higher-order clustering coefficient
Information entropy
Link prediction

ABSTRACT

In complex networks, link prediction can forecast missing links and identify spurious interactions has wide applications in the real world. Although the high-order structure plays a vital role in network evolution, its effect on link prediction is not always taken into consideration in traditional prediction algorithms. In this paper, we come up with a novel link prediction approach based on Mutual information of the High-Order Clustering structure (MHOC). The MHOC approach integrates the effects of multiple higher-order structures of nodes and quantifies the different contributions of common neighbors with the aid of the diverse high-order clustering coefficients of nodes based on information entropy for predicting missing links. Experimental results demonstrate that in the real world network, the high-order clustering patterns of nodes can improve link prediction accuracy significantly.

1. Introduction

The real-world complicated systems, such as online social websites and transportation system, can be represented as complex networks which characterize the topological structure and dynamic behavior of systems [1]. Generally, the system entities are regarded as nodes and the interactions or associations among entities are abstracted as links in complex networks. With the rapid development of network science and big data technology, complex network has been a powerful theoretical and practical tools to resolve the different problems in various fields [2,3]. As one of hot topics in complex networks, link prediction refers to predict the missing links and future possible links with the aid of observed network data [4,5]. It involves a significant in comprehending the evolution mechanism of network and mining the key factors that affect the network dynamic [6–8]. Furthermore, link prediction still has widely practical value in various fields. In protein–protein networks [9], the future interactions between proteins can be predicted to reduce the experiment cost [10]. In social networks, such as Twitter and Facebook, the potential friend relationship can be explored so as to increase the stickiness of users [11]. And moreover, link prediction can be used to knowledge graph completion [12] and e-commerce networks so as to automatically recommend products to customers [13].

In order to solve link prediction problem, many structures similarity-based approaches that exploit the topology structure information of nodes in networks have been proposed, such as Common Neighbor (CN) [4], Adamic–Adar (AA) [14] and Resource Allocation (RA) [15] et al. Generally, the approaches based on CN only consider the number of

* Corresponding author.
E-mail addresses: yaoyabing@lut.edu.cn (Y. Yao), cty18029@lut.edu.com (T. Cheng).

https://doi.org/10.1016/j.physa.2022.128428
0378-4371/

common neighbors and suffer from low prediction accuracy. With the in-depth exploration of network structure, the local clustering coefficient of one node or link that describes the connection probability of forming closed triangles between its neighbors is introduced into link prediction. Clustering coefficient of nodes is positively correlated with the probability of forming links between nodes, such as Clustering Coefficient based Link Prediction (CCLP) [16] and Asymmetric Link Clustering (ALC) [17] et al. In networks, in addition to triangle closure structures, closed sub-graphs with quadrilateral or even higher-order patterns are more possibly to form links between neighboring nodes, i.e., high-order clustering structures. Recently, high-order clustering coefficient has been proposed to measure the closure likelihood of high-order network cliques [18]. Since the high-order clustering coefficient of nodes can be regarded as the high-order substructure connection likelihood of its neighbors from higher viewpoint, for one common neighbor of a pair nodes, whether its high-order clustering coefficient can be helpful for link prediction is an interesting and valuable problem.

Generally, the traditional clustering coefficient is only applied to measure the closure probability of a simple wedge structure, which essentially represents the triangle clustering degree of nodes but cannot denote the clustering features of higher-order structures in the network. The high-order clustering coefficient can weigh the closure likelihood of high-order cliques structure of networks. Motivated by this, we propose a high-order clustering structure link prediction method based on mutual information (MHOC), which fuses the traditional clustering coefficient and the high-order structure as a whole in networks. MHOC method not only considers the high-order clustering coefficients, but also integrates different high-order clustering coefficients as a conditional probability to measure the contributions of common neighbors for node pairs. In addition, the mutual information is employed to distinguish the contributions of different common neighbors and quantifies the contribution of clique structures between neighbors to link formation. Experiment demonstrates that our proposed algorithm is superior to the benchmark algorithms, and the high-order clustering coefficient can significantly improve the prediction accuracy.

A brief introduction about the current link prediction methods is given in Section 2. Section 3 elaborates our high-order link prediction MHOC framework. Section 4 presents the popular baseline algorithms and evaluation metrics. Section 5 shows the experimental results. Section 6 summarizes this article.

2. Related work

Generally, link prediction algorithms can be divided into three categories [19]: similarity-based methods, maximum likelihood methods and probability model methods.

2.1. Similarity-based algorithms

Similarity-based methods, as implied by the name, assume that the higher the similarity between two nodes, the higher the existence possibility of a link [20]. Generally speaking, the similarity score is derived from the attribute or structural information of nodes. Due to the privacy protection issues, the attribute of nodes is difficult to obtain. Structural similarity methods are divided into three categories in light of the structural information used: local similarity indices, global similarity indices and quasi-local similarity indices.

The similarity score is calculated by the number of common neighbor nodes and its degree in the local similarity method. Common neighbors (CN) [4] is the simplest structure similarity-based methods. It deems that the more common neighbors between two unconnected nodes, the more likely they will be connected. However, CN only pays attention to the number of common neighbors, with the result that it hard to distinguish the contribution of different common neighbors. As variants of CN method, AA [14] and RA [15] penalize nodes with large degrees to differentiate the contributions of common neighbor nodes with different degrees. In addition, node centrality is also used to distinguish the contributions of common neighbors. PA [21] is proposed based on this property. Although the method based on common neighbors is effective, it does not excavate deeper into the structural information in the network. MI [22] uses information entropy to quantify the contribution of the clustering coefficients. Kumar et al. [23] extends the notion of clustering information of the CCLP index and extract clustering information of level-2 common neighbors of the seed node pair and computes the similarity score based on this information. In order to explore the relationship between link prediction and topological structure, Huang et al. [24] use the predictive value of clustering coefficient measure to extend the standard clustering coefficient measure and capture the trend of high-order clustering. Based on naive Bayes theory, Liu et al. [25] propose the Local Naive Bayes-based Common Neighbors (LNBCN) method that takes into account different functions of common neighbors. Liu et al. [26] propose a new degree correlation clustering coefficient to estimate the clustering ability of nodes, and design a Degree-related Clustering ability Path (DCP) index for link prediction. Wu et al. [16] extract the information of the link structure of the triangle prediction chain by calculating the node clustering coefficient, and the clustering information of nodes and links is introduced in [27], which achieves good results in large and medium-sized data sets. Zhou et al. [28,29] indicate that the introduction of local community paradigm and Hebbian learning rule could considerably improve the performance of routine local similarity indices. CCLP [16] considers the clustering structure information of common neighbor nodes. It believes that the higher the clustering coefficient of the common neighbor nodes of the predicted node pair, the greater the possibility of connecting the two nodes. In addition to common neighbors, the paths between node pairs also promote the formation of links. CAR-based Resource Allocation (CAR) [30] holds that if the common neighbor of two nodes has a series of strong internal links, the probability of establishing links between

two nodes will become greater. This premise enables us to obtain more important nodes with internal links with other neighbors. Wang et al. [31] propose Degree-related and Link Clustering coefficient (DLC) to better describe the common neighbors in different areas. Samira et al. [32] introduce a new link prediction measurement method Common Neighbors Degree (CNDP) with clustering coefficient as a structural characteristic of networks, which achieves higher prediction accuracy with lower complexity. NSI [33] quantifies the clustering coefficient of common neighbors and the contribution of links between neighbor nodes to the target node pair through information entropy. In general, local similarity indices often have low computational complexity resulting from the limited structural topology(i.e., common neighbors), but their prediction accuracy is also limited.

Global similarity indices compute similarity score ground on the entire network topological information i.e., the influence of paths of different lengths on the pairs of nodes. Katz [34] index calculates the effect of all different path lengths on similarity. Random walks with restart (RWR) [35] is a direct application of PageRank [36] algorithm. It assumes that the random walk particle has a certain probability of returning to its initial position at each step. Average Commute Time (ACT) [37,38] deems the average commuting time of two nodes is shorter, the two nodes are more similar. The global similarity method considers all topology structure, therefore the prediction performance is the best. But the computational complexity is too high to suitable for large-scale networks.

Quasi-local approaches use additional topological information as global indices, it strikes a trade-off between time complexity and prediction accuracy. The local path index (LP) [39] considers the common neighbors and also calculates the contribution of the 2-order neighbors. Local random walk (LRW) [40] is also a kind of random walk algorithm, which only limits the number of steps to walk. Path Entropy (PE) [41] uses information entropy to quantify the contribution of different order paths to the target node pair. Quasi-local similarity method balances the characteristics of local and global method. Therefore, it has both high prediction accuracy and low computational complexity.

2.2. Maximum likelihood methods

The core idea of the prediction method based on likelihood estimation is divided into two steps. Firstly, the likelihood value of the network is calculated according to the generation and organization of the network structure and the links that have been observed at present. Then the possibility of each pair of unconnected nodes to generate connection links is calculated according to the maximization of the network likelihood. This method is initially applied in the network structure with obvious hierarchy [42]. R. Guimera et al. [43] propose a general mathematical and computational framework for dealing with data reliability problems in complex networks for missing and pseudo-interactions in the observation of noisy networks. Pan et al. [44] transform the network organization mechanism into a link prediction algorithm, append the link to the conditional probability of the observation network to evaluate the non-observation link, which has excellent performance in detecting missing links and identifying false links in biological networks and social networks. Stochastic block model (SBM) [45,46] is one of the most common network models, where the nodes are divided into groups, and the probability that two nodes are connected depends only on the group they belong to. Although the methods based on maximum likelihood estimation have high computational complexity, they are not only applied to link prediction, but also bring us profound insights about network structure.

2.3. Probabilistic models

The core of probabilistic model for link prediction is to build a model containing a set of adjustable parameters, and then use optimization strategy to find the optimal parameter values, so that the obtained model can better reproduce the structure and relationship characteristics of the real network. The connection probability of two nodes without links is equal to the conditional probability of generating links under this set of optimal parameters. The probability models used in link prediction include: Probability Relationship Model (PRM) [47], Probability Entity Relationship Model (PERM) [48] and Random Relationship Model (SRM) [49]. PRM defines the joint probability distribution of relational data attributes and puts forward the relational attribute pattern. PERM is a model based on the entity relationship model, which considers the relationship between entities as important as the entity. The difference between PRM and PERM is that they express different databases in different ways. The former is based on relational models, while the latter is based on entity relationship model. The key concept of SRM is the random entity process caused by the interaction of multiple entity Gaussian processes.

In recent years, with the popularization of machine learning, many representation learning methods have been proposed for link prediction. These methods map one node in network to a relatively low-dimensional vector which can be used to represent the similarity between nodes base on the vector distance. Deep walk [50] and Node2vec [51] are pioneers in using graph embedding for link prediction. Topological deep network embedding [52] also uses deep learning graph embedding method for link prediction.

3. Motivation and method

Generally, the uncertainty of link formation between nodes always decreases with the increase of helpful structural feature information. High-order structure can capture more structure information and it plays an important role in network evolution, hence we introduce the high-order clique clustering structure of networks into link prediction.

3.1. Link prediction framework based on mutual information of topology structure

According to the definition of information theory [53], it measures the uncertainty of an event [54,55]. The event with high information entropy indicates the large occurrence uncertainty of this event, therefore its occurrence probability is small. Given an event Q and P(Q) denotes the occurrence probability of this event, its uncertainty (i.e., information entropy) I(Q) [56] can be defined as:

$$I(Q) = -\log(P(Q)) \tag{1}$$

based on Eq. (1), the conditional self-information [57] can be written as:

$$I(X \mid Y) = -\log(P(X \mid Y)) \tag{2}$$

where $P(X \mid Y)$ is the conditional probability that the event X happens given that Y has already happened.

Given a couple of nodes (x, y) in a network, the existence of a link between nodes x and y can be regarded as an event L^1_{xy} (i.e., there is a link between x and y) with respect to the link prediction problem. Therefore, the existence likelihood of the event L^1_{xy} depends on its occurrence uncertainty, i.e., the entropy of L^1_{xy}. The high entropy of L^1_{xy} means the large connection uncertainty between node pair (x, y), which leads to the small connection probability for this node pair, otherwise.

Many topological structures (such as common neighbor, hierarchical structure, community structure, etc.) between two nodes can be used to improve link prediction performance. For example, node pairs with more common neighbors tend to form links more often than those with less common neighbors. Moreover, two nodes within same community prefer to connect compared with those within different community. From the viewpoint of information entropy, these topological structures can be employed to eliminate the connection uncertainty with respect to two unconnected nodes to a large extent. The more useful information provided by the topology structure of network, the more helpful for eliminating connection uncertainty, and the more likely two nodes will connect. Accordingly, given a topological structure set Ω and node pair (x, y), its similarity based on the information entropy can be defined as [22,33]:

$$S^{\Omega}_{xy} = -I(L^1_{xy} \mid \Omega) \tag{3}$$

where $I(L^1_{xy} \mid \Omega)$ is conditional self-information, which represents the uncertainty of event that. If the uncertainty is greater, the similarity between the couple of nodes (x, y) is smaller, and the probability of them being connected is lower. Given a feature element ω in set Ω (i.e., $\omega \in \Omega$) and assuming that elements in set Ω are independent of each other, based on the mutual information theory, $I(L^1_{xy} \mid \Omega)$ can be defined as:

$$\begin{aligned} I(L^1_{xy} \mid \Omega) &= I(L^1_{xy}) - I(L^1_{xy}; \Omega) \\ &= I(L^1_{xy}) - \sum_{\omega \in \Omega} I(L^1_{xy}; \omega) \\ &= I(L^1_{xy}) - \sum_{\omega \in \Omega} (I(L^1_{xy}) - I(L^1_{xy} \mid \omega)) \end{aligned} \tag{4}$$

combining Eqs. (3) and (4), given a pair of nodes (x, y), the similarity between node pair (x, y) based on mutual information theory can be defined as [22,33]:

$$\begin{aligned} S^{\Omega}_{xy} &= -I(L^1_{xy} \mid \Omega) \\ &= \sum_{\omega \in \Omega} (I(L^1_{xy}) - I(L^1_{xy} \mid z)) - I(L^1_{xy}) \end{aligned} \tag{5}$$

Eq. (5) indicates that each feature ω can contribute towards eliminating the connection uncertainty for the node pair (x, y). The more valuable information provided by the feature ω, the more contribution for nodes x and y. In addition to the common neighbors between node pairs, we consider simultaneously the high-order clustering characteristics of each common neighbor to construct our prediction model based on the mutual information. Its specifics process is shown in following section.

3.2. High-order clustering properties of nodes

Several existing structure-based link prediction methods [58] adopted the local clustering coefficients of common neighbor nodes between node pairs to measure the node contribution. Given a node pair (x, y) and one of its common neighbors ω, the fundamental principle of such methods is that the local clustering coefficient evaluates the probability of forming a closed triangle between neighbor nodes with node ω as a core. The greater the clustering coefficient of node ω, the higher the likelihood of forming one link between (x, y) which contributes from node ω. However, the link prediction methods with local clustering coefficients only capture the contribution of the lowest-order aggregation properties of common neighbors to link formation, and does not fully utilize the higher-order aggregation properties of nodes. As shown in Fig. 1, with respect to node pair (x, y), it has two common neighbors m and n which own the same number of neighbors and also have the same local (or the lowest order) clustering coefficients. However, it is clear that the neighbors

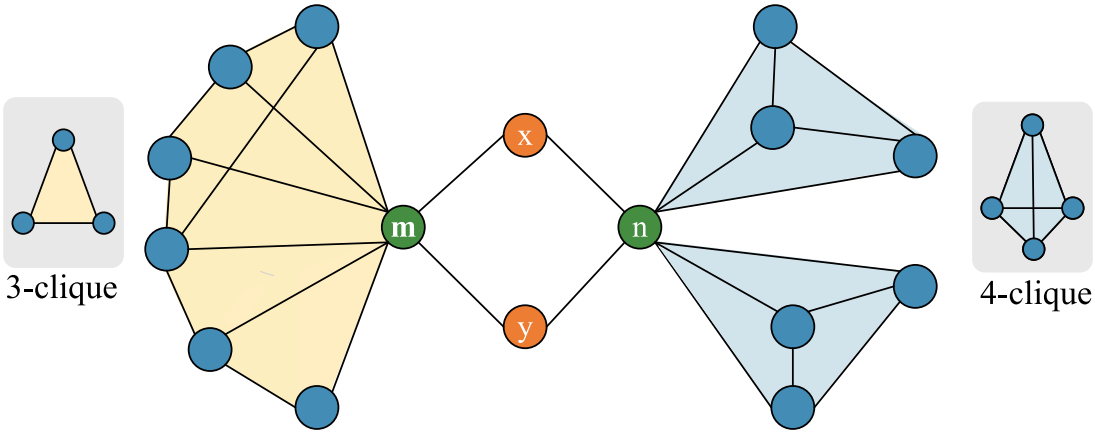

Fig. 1. An illustration example of calculating the MHOC. For a predicted target node pair (x, y), it has common neighbors m and n, which own the same number of neighbors and have the same local clustering as well. There are six 3-clique around node m, meaning that 3-clique is more likely to occur around node m. In addition to forming the same number of triangle closure structures, its neighbors also appear two four-node closure cliques with respect to the node n. This means that not only 3-clique is more likely to form around node n, but also 4-clique.

around node n are more aggregated than node m so as to form a novel high-order clustering structure [18], which may contribute to the formation of link between nodes x and y more heavily. Therefore, we naturally introduce this high-order clustering structure into link prediction problem to improve the prediction accuracy. Next, we take a deeper look at this structure from the perspective of clustering coefficients.

The local clustering coefficient measures the probability of forming triangular closure structure between neighboring nodes centered on a node θ. The higher the local clustering coefficient, the higher the degree of closure among neighbors and it is defined as:

$$C(\theta) = \frac{3\triangle}{\triangle + \wedge} = \frac{2M_\theta}{k_\theta * (k_\theta - 1)} \tag{6}$$

In Eq. (6), the numerator is denoted three times the number of closed triangles including node θ and the denominator is denoted the number of all possible triangles containing node θ. M_θ denotes the number of links between the k_θ neighborhoods of node θ. Closed triangle structure can be treated as a fully connected sub-graph and belongs to a kind of the lowest order clique composed by three nodes. In fact, the formation of closed triangles also can be regarded as the closing process of wedge-shaped structures, therefore the higher-order cluster structure can be naturally extended as the closing procedure of higher-order clique as well. According to the closing process of clique structure, local clustering coefficient (the 2-order clustering coefficient) can be further defined as [18]:

$$C_2(\theta) = \frac{2 \mid K_3(\theta) \mid}{\mid W_2(\theta) \mid} \tag{7}$$

where $K_3(\theta)$ is the set of 3-order clique containing node θ and $W_2(\theta)$ is the set of wedges centered on node θ. The meaning of Eq. (7) is the probability of forming a closed triangle between neighbors of θ. The larger the 2-order clustering coefficient, the higher the degree of closure. The 2-order clustering coefficient measures the formation process of the 3-order clique, the 3-order clustering coefficient can be regarded as a measure of the formation process of the 4-order clique [18], and so forth. The 3-order clustering coefficient measures the probability that the neighboring nodes form the 4-order clique with node θ as a core. It can be defined as:

$$C_3(\theta) = \frac{3 \mid K_4(\theta) \mid}{\mid W_3(\theta) \mid} \tag{8}$$

Furthermore, the 4-order clustering coefficient is defined as follows:

$$C_4(\theta) = \frac{4 \mid K_5(\theta) \mid}{\mid W_4(\theta) \mid} \tag{9}$$

The l-order clustering coefficient can be defined as:

$$C_l(\theta) = \frac{l \mid K_{l+1}(\theta) \mid}{\mid W_l(\theta) \mid} \tag{10}$$

where $K_{l+1}(\theta)$ is the set of (l+1)-cliques containing θ and $W_l(\theta)$ is the set of l-wedges with center θ (if $W_l(\theta)$ = 0, we say that $C_l(\theta)$ is undefined).

The high-order clustering coefficient characterizes the degree of high-order aggregation of neighboring nodes around the node. Combining with the topological structure mutual information theory in Section 3.1 and considering the high-order clustering characteristics of common neighbor nodes in Section 3.2, we propose a link prediction method based on the high-order clustering mutual information of common neighbors.

3.3. Link prediction method ground on mutual information of common neighbor high-order clustering

For an unconnected node pair (x, y) without any prior structure, the occurrence probability of event L^1_{xy} can be calculated as

$$P(L^1_{xy}) = 1 - P(L^0_{xy}) = 1 - \prod_{j=1}^{k_y} \frac{(M - k_x) - j + 1}{M - j + 1} = 1 - \frac{C^{k_y}_{M-k_x}}{C^{k_y}_M} \tag{11}$$

where $P(L^0_{xy})$ is the unconnected probability between x and y, k_x and k_y are degrees of nodes x and y respectively. M is the total number of links in network. $\frac{C^{k_y}_{M-k_x}}{C^{k_y}_M}$ represents the probability that no link is formed between x and y. According to the definition of information entropy, we have the entropy of L^1_{xy} as:

$$I(L^1_{xy}) = -\log(P(L^1_{xy})) = -\log(1 - \frac{C^{k_y}_{M-k_x}}{C^{k_y}_M}) \tag{12}$$

Given a node pair (x, y) with common neighbor set Γ_{xy} and z is one of its comm neighbors (i.e., $z \in \Gamma_{xy}$), let $p_l(z)$ represent the probability that there is a link between x and y in consideration of the l-order clustering characteristics of common neighbor z. $\overline{p_l(z)}$ denotes the probability of the neighbors around node z does not form connected links. For instance, $p_2(z)$ represents the connection probability considering the 2-order clustering characteristic of node z (i.e., local clustering coefficient). Since the high-order clustering coefficient $C_l(z)$ characterizes the connection probability of neighbors around node z under the l-order clustering structure, therefore $p_l(z) = C_l(z)$ and $\overline{p_l(z)} = 1 - C_l(z)$. We assume that the different order clustering coefficient of node z, i.e., $\{C_2(z), C_3(z) \ldots, C_l(z)\}$, are independent to each other. According to probability theory, the connection possibility between nodes x and y that simultaneously considers the l-order clustering structure can be expressed as

$$\begin{aligned} P(L^1_{xy} \mid z) &= 1 - \overline{p_2(z)} * \overline{p_3(z)} * \cdots * \overline{p_l(z)} \\ &= 1 - (1 - p_2(z)) * (1 - p_3(z)) * \cdots * (1 - p_l(z)) \\ &= 1 - (1 - C_2(z)) * (1 - C_3(z)) * \cdots * (1 - C_l(z)) \\ &= 1 - \prod_{j=2}^{l} (1 - C_j(z)) \end{aligned} \tag{13}$$

Eq. (13) indicates that at least one order clustering coefficient of node z will contribute to the link formation if a link between nodes x and y is exists. The conditional entropy $I(L^1_{xy} \mid z)$ with consideration of the l-order clustering coefficient of common neighbor z between nodes x and y is expressed as

$$\begin{aligned} I(L^1_{xy} \mid z) &= -\log(P(L^1_{xy} \mid z)) \\ &= -\log(1 - \prod_{j=2}^{l} (1 - C_j(z))) \end{aligned} \tag{14}$$

According to the definition of Eq. (3), given the node pair (x, y) and its common neighbor set Γ_{xy} (i.e., $\Omega = \Gamma_{xy}$), the similarity considering the l-order clustering characteristic of all common neighbors is defined as follows

$$\begin{aligned} S^{\Omega}_{xy} &= -I(L^1_{xy} \mid \Gamma_{xy}) \\ &= \sum_{z \in \Gamma_{xy}} (I(L^1_{xy}) - I(L^1_{xy} \mid z)) - I(L^1_{xy}) \\ &= \sum_{z \in \Gamma_{xy}} (-\log(1 - \frac{C^{k_y}_{M-k_x}}{C^{k_y}_M}) - (-\log(1 - \prod_{j=2}^{l} (1 - C_j(z))))) - (-\log(1 - \frac{C^{k_y}_{M-k_x}}{C^{k_y}_M})) \\ &= \sum_{z \in \Gamma_{xy}} (\log(1 - \prod_{j=2}^{l} (1 - C_j(z))) - \log(1 - \frac{C^{k_y}_{M-k_x}}{C^{k_y}_M})) + \log(1 - \frac{C^{k_y}_{M-k_x}}{C^{k_y}_M}) \end{aligned} \tag{15}$$

When $l = 2$, this method can degenerate to MI index [22] which only considers the contribution of the 2-order clustering coefficient (i.e., the traditional local clustering coefficient) of the common neighbors. Due to the computational complexity,

we merely take notice of the contribution of the 2-order, 3-order and 4-order clustering coefficient (i.e., $l \leq 4$), therefore the MHOC index is defined as

$$\begin{aligned} S_{xy}^{MHOC} &= -I(L_{xy}^1 \mid \Gamma_{xy}) \\ &= \sum_{z\in\Gamma_{xy}} (I(L_{xy}^1) - I(L_{xy}^1 \mid z)) - I(L_{xy}^1) \\ &= \sum_{z\in\Gamma_{xy}} (-\log(1-\frac{C_{M-k_x}^{k_y}}{C_M^{k_y}}) - (-\log(1-\prod_{j=2}^{4}(1-C_j(z))))) - (-\log(1-\frac{C_{M-k_x}^{k_y}}{C_M^{k_y}})) \\ &= \sum_{z\in\Gamma_{xy}} (\log(1-\prod_{j=2}^{4}(1-C_j(z))) - \log(1-\frac{C_{M-k_x}^{k_y}}{C_M^{k_y}})) + \log(1-\frac{C_{M-k_x}^{k_y}}{C_M^{k_y}}) \\ &= \sum_{z\in\Gamma_{xy}} (\log(1-(1-C_2(z))*(1-C_3(z))*(1-C_4(z))) - \log(1-\frac{C_{M-k_x}^{k_y}}{C_M^{k_y}})) + \log(1-\frac{C_{M-k_x}^{k_y}}{C_M^{k_y}}) \end{aligned} \tag{16}$$

In order to further understand MHOC model, we take Fig. 1 as an example to illustrate the specific calculation procedure. For convenience of expression, $MHOC_o^2$ indicates the 2-order clustering coefficient is only used and $MHOC_f^{23}$ represents the fusion usage of 2-order and 3-order clustering coefficients. For nodes m and n, they all have 8 neighbors (i.e., $k_m = k_n = 8$) and the same number of closed triangles (i.e., $M_m = M_n = 6$). The process of calculating $MHOC_o^2$ is as follows. Using Eq. (6), we obtained $C_2(m) = C_2(n) = \frac{2\times 6}{8\times(8-1)} = \frac{3}{14}$. According to Eq. (12), we have $I(L_{xy}^1) = -log(1-\frac{C_{28-2}^2}{C_{28}^2}) = 0.8532$. Using Eq. (13), because $C_2(m) = C_2(n)$, $P(L_{xy}^1 \mid m) = P(L_{xy}^1 \mid n) = 1-(1-C_2(m)) = C_2(m) = \frac{3}{14}$. Then $I(L_{xy}^1 \mid m) = I(L_{xy}^1 \mid n) = -log\{P(L_{xy}^1 \mid n)\} = -log(\frac{3}{14}) = 0.6690$. Therefore, $MHOC_o^2 = \sum_{z\in\Gamma_{xy}}(I(L_{xy}^1) - I(L_{xy}^1 \mid z)) - I(L_{xy}^1) = (I(L_{xy}^1) - I(L_{xy}^1 \mid m)) + (I(L_{xy}^1) - I(L_{xy}^1 \mid n)) - L_{xy}^1 = 0.2281$ (i.e., MI). The process of calculating $MHOC_f^{23}$ is as follows. Using Eq. (8), we have $C_3(m) = 0$ and $C_3(n) = \frac{1}{6}$. Then, using Eq. (13), we can obtained $P(L_{xy}^1 \mid m) = 1-(1-C_2(m))*(1-C_3(m)) = \frac{3}{14}$, $P(L_{xy}^1 \mid n) = 1-(1-C_2(n))*(1-C_3(n)) = \frac{29}{84}$. Further, $I(L_{xy}^1 \mid m) = -log(\frac{1}{7}) = 0.6690$ and $I(L_{xy}^1 \mid n) = -log(\frac{11}{35}) = 0.4619$. Finally, $MHOC_f^{23} = \sum_{z\in\Gamma_{xy}}(I(L_{xy}^1) - I(L_{xy}^1 \mid z)) - I(L_{xy}^1) = (I(L_{xy}^1) - I(L_{xy}^1 \mid m)) + (I(L_{xy}^1) - I(L_{xy}^1 \mid n)) - L_{xy}^1 = 0.4352$. Obviously, $MHOC_f^{23} > MHOC_o^2$. After using the third-order clustering coefficient, the prediction accuracy has been significantly improved.

4. Benchmarks and evaluations metrics

In this paper, MHOC is compared with ten typical methods, including CN [4], AA [14], RA [15], CAR [59], PA [21], CCLP [16], LNBCN [25], DLC [31], CNDP [32]. In addition, to explore the performance of MHOC, we also compare the MI [22] method that employs the mutual information of local clustering coefficient. The area under the receiver operation characteristic curve (AUC) [60] and the area under the precision–recall curve (AUPR) [61] are used for the evaluation of prediction performance.

4.1. Baseline methods for comparison

(1) Common neighbor (CN) [4]. For a pair of nodes, it is more likely to be existed a link if it owns more common neighbors.

$$S_{xy}^{CN} = \mid \Gamma_x \cap \Gamma_y \mid \tag{17}$$

where Γ_x and Γ_y are the neighbor node sets of nodes x and y respectively.

(2) Adamic–Adar (AA) [14]. It improves the simple count of common neighbors by giving more weight to nodes with lower degrees.

$$S_{xy}^{AA} = \sum_{z\in\Gamma_{xy}} \frac{1}{log(\mid\Gamma_z\mid)} \tag{18}$$

where Γ_{xy} is the common neighbors set of nodes x and y and $\mid\Gamma_z\mid$ represents the degree of node z.

(3) Resource Allocation (RA) [15]. If a pair of nodes x and y are not adjacent, node x can send resources to node y through its neighbor nodes. In this process, each neighbor node of x can obtain a unit of resources from node x, and then equally distribute them to its neighbors. The number of resources obtained by the final node y from its neighbors can be regarded as the similarity between node x and node y.

$$S_{xy}^{RA} = \sum_{z\in\Gamma_{xy}} \frac{1}{\mid\Gamma_z\mid} \tag{19}$$

Both the RA and AA methods end up in the same style, summing the weights over all common neighbor nodes. AA assigns the weight of each node to the inverse of the logarithm of the node's degree value, whereas RA assigns the inverse of the degree value.

(4) Cannistrai Alanis Ravai (CAR) [59]. This method takes into account the links between neighbors on the basis of common neighbors.

$$S_{xy}^{CAR} = CN(x,y) * \sum_{z \in \Gamma_{xy}} \frac{\mid \Gamma(z) \mid}{2} \tag{20}$$

where $\Gamma(z)$ is the intersection of the set of neighbors of node z with the set of common neighbors of nodes x and y. In other words, it is the search for a structure in the common neighbors of nodes x and y that has a role in the formation of a link between x and y. This structure is the concatenated link formed between the common neighbors of x and y.

(5) Preferential Attachment (PA) [21]. It supposes that the connection probability is proportional to the product of node degree.

$$S_{xy}^{PA} = \mid \Gamma_x \mid * \mid \Gamma_y \mid \tag{21}$$

where $\mid \Gamma_x \mid$ is the degree of the node x.

(6) Clustering Coefficient based Link Prediction (CCLP) [16]. CCLP index estimates the contributions of common neighbors with local clustering coefficient.

$$\begin{aligned} S_{xy}^{CCLP} &= \sum_{z \in \Gamma_{xy}} C(z) \\ C(z) &= \frac{t(z)}{\mid \Gamma_z \mid * (\mid \Gamma_z \mid -1)} \end{aligned} \tag{22}$$

where $t(z)$ is the total triangles passing through node z and $\mid \Gamma_z \mid$ is the degree of the node z.

(7) Mutual Information (MI) [22]. It uses mutual information theory to quantify the contribution of common neighborhoods.

$$\begin{aligned} S_{xy}^{MI} &= -I(L_{xy}^1 \mid \Gamma_{xy}) \\ &= \sum_{z \in \Gamma_{xy}} (I(L_{xy}^1; z) - I(L_{xy}^1)) \end{aligned} \tag{23}$$

where $I(L_{xy}^1; z)$ is calculated as follows:

$$I(L_{xy}^1; z) \approx \frac{1}{\mid \Gamma(z)(\mid \Gamma(z) \mid -1)} \sum_{m,n \in \Gamma_z, m \neq n} (I(L_{mn}^1) + \log \frac{\triangle}{\triangle + \wedge}) \tag{24}$$

(8) Local Naive Bayes-based Common Neighbors (LNBCN) [25]. This index is proposed based on the naive Bayes theory and considers different functions of different common neighbors.

$$S_{xy}^{LNBCN} = \sum_{z \in \Gamma_{xy}} [\log(\frac{C(z)}{1 - C(z)}) + log(\frac{1-\rho}{\rho})] \tag{25}$$

where ρ is the network density and $\rho = \mid E \mid \setminus \mid U \mid$, its value is the ratio of the number of links $\mid E \mid$ to the number of all possible links $\mid U \mid$, $C(z)$ is the clustering coefficient of the node z.

(9) Degree-related and Link Clustering coefficient (DLC) [31]. This index is defined by combining the asymmetry link clustering coefficient with the degree-related clustering coefficient.

$$S_{xy}^{DLC} = w_x(\sum_{z \in \Gamma_{xy}} DLC_{x,z}) + w_y(\sum_{z \in \Gamma_{xy}} DLC_{y,z}) \tag{26}$$

where the calculation of w_x, w_y and $DLC_{x,z}$ are shown below:

$$\begin{aligned} w_x &= \frac{\mid \Gamma_x \mid}{max(\mid \Gamma_x \mid, \mid \Gamma_y \mid)} \\ w_y &= \frac{\mid \Gamma_y \mid}{max(\mid \Gamma_x \mid, \mid \Gamma_y \mid)} \end{aligned} \tag{27}$$

$$DLC_{x,z} = \mid \Gamma_x \cap \Gamma_z \mid \cdot \overline{C(z)} \cdot \mid \Gamma_z \mid^{(r-1)} \tag{28}$$

where r is the assortative coefficient [62] of the network, $\overline{C(z)}$ is the average of the clustering coefficients of all nodes with degree equaling to the degree Γ_z of the node z, $\mid \Gamma_x \mid$ and $\mid \Gamma_y \mid$ represent the degree of x and y, respectively.

Table 1
The topological information of the real-world network. The number of nodes in the network is N and M demonstrates the number of links. The average degree is indicated $\langle k \rangle$. The average distance is represented $\langle d \rangle$. Clustering coefficient is denoted C_2. C_3 and C_4 are the 3-order and 4-order clustering coefficients, respectively. *Field* is the domain of the network. In order to preserve network connectivity, we extract the maximal connected subgraph of the original network for some networks.

DataSet	N	M	$\langle k \rangle$	$\langle d \rangle$	C_2	C_3	C_4	*Field*
Baydry	128	2106	32.90	1.77	0.34	0.17	0.11	Foodweb
Maayan	183	2434	27.00	2.63	0.33	0.29	0.20	Foodweb
Smagri	1024	4916	9.60	2.98	0.34	0.15	0.07	CoAuthor
Email	1133	5451	9.62	3.06	0.25	0.14	0.10	Email
Blogs	1222	16714	27.35	2.74	0.36	0.22	0.16	Politics
Google	1299	2773	4.27	6.47	0.60	0.62	0.77	Society
Ucsocial	1893	13835	35.67	6.81	0.06	0.04	0.01	Society
Yeast	2224	6609	5.94	5.09	0.20	0.36	0.38	Biology
Lastfm	7624	27804	13.46	3.50	0.38	0.18	0.14	Society

(10) Common Neighbors Degree Penalization (CNDP) [32]. CNDP is defined by combining CN, AA, RA methods with clustering coefficients and taking into account connections between common neighbors.

$$S_{xy}^{CNDP} = \sum_{z \in \Gamma_{xy}} \mid T_z \mid (\mid \Gamma_z \mid)^{-\beta C} \tag{29}$$

where z is a common neighbor for two nodes x and y, $\mid T_z \mid$ is the number of links between z's neighbors. $\mid \Gamma_z \mid$ is the degree of the node z, C is the average clustering coefficient, and β is a constant value.

4.2. Evaluation metrics

Given an undirected and unweighted network $G(V, E)$, where V is the set of nodes and E is the set of links in this network. All possible links is represented by set U. Self-loop and multi-links are not allowed. The non-existing link set is represented by $U \setminus E$. In the process of experiment, E is divided into two parts: the training set E^T and the test set E^P. Obviously, $E^P \cup E^T = E$ and $E^P \cap E^T = \phi$. In order to measure the prediction performance, we employ AUC and AUPR metrics [63,64].

(1) AUC [60]

AUC is the area under the receiver operating characteristic (ROC) curve. It represents the probability that a randomly selected link in the test network (i.e., link in E^P) will score higher than a randomly selected link that does not exist in the training network (i.e., link in $U - E$). The AUC value can be calculated as

$$AUC = \frac{n' + 0.5 \times n''}{n} \tag{30}$$

where n represents times of independent comparisons. n' shows that the score of the link from E^P is higher than the link from $U - E$ times, and n'' expresses times having the same scores.

(2) AUPR [61]

The link prediction problem is regarded as a binary classification task [65]. Therefore, most of the evaluation indexes of any binary classification task can be used for link predictive evaluation. The evaluation of a binary classification task having two classes can be represented as a confusion matrix [66]. In this confusion matrix, True Positive (TP) means positive data item predicted as positive; True Negative (TN) means negative data item predicted as negative; False Positive (FP) means negative data item predicted as positive; False Negative (FN) means positive data item predicted as negative. Based on the confusion matrix, the values of Precision and Recall metrics can be derived as follows [66]:

$$Precision = \frac{TP}{TP + FP} \tag{31}$$

$$Recall = \frac{TP}{TP + FN} \tag{32}$$

The Precision–Recall (PR) curve is made up of Precision on the Y-axis and Recall on the X-axis. In other words, the PR curve contains TP/(TP+FN) on the Y-axis and TP/(TP+FP) on the X-axis. AUPR is the average of precision across all recall values, and does not depend on the number of negative samples. Therefore, AUPR is the preferred evaluation index for imbalance problems that mainly focus on a few categories.

Table 2
Comparison of prediction accuracy of 11 indicators on 9 real-world networks under AUC measure. By performing 9:1 independent random segmentation on the training set, and achieving an average of 100 times. The best indicators with the best performance are marked in bold and the second-best results are marked with an underline on each network. Among them, $MHOC^*$ is the best result of all methods using high-order information. $MHOC_f^{23}$ is the result of both using 2-order and 3-order information.

DataSet	CN	LNBCN	AA	RA	CAR	PA	CCLP	MI	DLC	CNDP	$MHOC_f^{23}$	$MHOC^*$
Baydry	0.6060	0.6061	0.6070	0.6091	0.6216	0.7308	0.6331	0.5031	0.7392	0.7223	**0.7673**	**0.7673**
Maayan	0.6444	0.6446	0.6443	0.6444	0.6610	0.7223	0.6814	0.5377	0.6821	0.6731	**0.7754**	**0.7754**
Smagri	0.8480	0.8482	0.8579	0.8579	0.7172	0.8627	0.8560	0.8906	0.8878	0.8528	0.8981	**0.8983**
Email	0.8569	0.8568	0.8589	0.8584	0.7034	0.8046	0.8543	0.8696	0.8722	0.8674	**0.8832**	**0.8832**
Blogs	0.9239	0.9239	0.9272	0.9285	0.8960	0.9092	0.9263	0.9255	0.9256	0.9211	**0.9316**	**0.9316**
Google	0.9568	0.9567	0.9503	0.9503	0.8544	0.8737	0.9220	0.9438	0.9413	0.9534	**0.9628**	**0.9628**
Ucsocial	0.7819	0.7820	0.7866	0.7873	0.6413	0.9161	0.7894	0.9181	0.9175	0.8482	**0.9190**	**0.9190**
Yeast	0.9145	0.9143	0.9151	0.9155	0.8460	0.8641	0.9095	0.9276	0.9349	0.8765	0.9356	**0.9365**
Lastfm	0.8756	0.8757	0.8761	0.8761	0.7421	0.8586	0.8679	0.9139	0.9177	0.9170	**0.9187**	**0.9187**

5. Experiment result

5.1. Datasets and experimental setup

In this work, MHOC was made predictions on nine networks from six different fields. Table 1 gives details of the topological characteristics of these networks. (1) Baydry [67]: a food chain ecological network based on foodweb. (2) Maayan [67]: a foodweb network. (3) SmaGri [67]: the network constitute of citations to Small and Griffith and Descendants. (4) Email [67]: a network of Alex Arenas's email. (5) Blogs [68]: a network of the US political blogs. (6) Google [67]: a web-google social network. (7) Ucsocial [67]: an opsahl-ucsocial social network. (8) Yeast [59]: a protein–protein interaction network. (9) Lastfm [69]: a society network. All these networks can be downloaded from the websites of http://konect.uni-koblenz.de/, http://networkrepository.com/networks.php and http://archive.ics.uci.edu/ml/datasets/LastFM+Asia+Social+Network.

In order to explore the predictive performance of MHOC, we carry out 100 times independent experiments on nine real networks from three aspects: (1) It is verified whether the high-order cluster structure information in the network can improve the prediction accuracy and which high-order structure is finally selected; (2) We compare MHOC prediction performance with other methods (i.e., CN, LNBCN, AA, RA, PA, CAR, CCLP, MI, DLC and CNDP); (3) To verify the prediction performance stability of MHOC method.

5.2. The result of prediction with different clustering structures

From the definition of MHOC algorithm given above, we know that the MHOC algorithm uses information entropy to quantify the contribution of different higher-order clustering coefficients of common neighbor nodes. In this experiment, the ratio of train set is set at 90% and the detailed results are presented in Figs. 2 and 3. In these two figures, $MHOC_o^2$, $MHOC_o^3$ and $MHOC_o^4$ respectively indicate that the 2-order clustering coefficient, the 3-order clustering coefficient and the 4-order clustering coefficient are only considered. $MHOC_f^{23}$ represent the fusion use of 2-order, 3-order and $MHOC_f^{234}$ represent the fusion use of 2-order, 3-order and 4-order clustering coefficients simultaneously. From Figs. 2 and 3, we can see the use of high-order clustering structure brings significant improvement to the prediction performance. In particular in Fig. 2, on the Baydry network, the prediction performance of AUC was improved by 28% compared to MI ($MHOC_o^2$) after using the high-order clustering structure ($MHOC_f^{23}$) and 25% improvement in AUC results over traditional methods (MI) on the Maayan network. In Fig. 3, we can see that the prediction performance of high order cluster structure ($MHOC_f^{234}$) is better than that of MI on almost all networks.

To sum up, the use of high-order clustering structures in link prediction can significantly improve the prediction accuracy. The result shows that the more higher-order clustering coefficients are used in the MHOC algorithm, the better prediction results is. However, with the increase of the high-order clustering coefficients, the computational complexity also increase with it. In order to solve this problem, we choose $MHOC_f^{23}$ to balance the accuracy and efficiency, i.e., fusing both 2-order and 3-order clustering coefficients. It ensures the improvement of prediction accuracy while having a low computational complexity.

5.3. Comparative analysis of MHOC and benchmark methods

In this section, we focus on the comparison of the $MHOC_f^{23}$ with benchmark algorithms. We set the train set random sampling ratio to 0.9. The experimental results are shown in Tables 2 and 3. In these two tables, we put the best results of the prediction algorithm in bold and the second-best results are marked with an underline on each network. Table 2 shows that experimental results of *AUC* between $MHOC_f^{23}$ and other algorithms in nine networks after independently sampling and dividing the dataset 100 times. $MHOC_f^{23}$ has the best performer on the seven networks and the second best

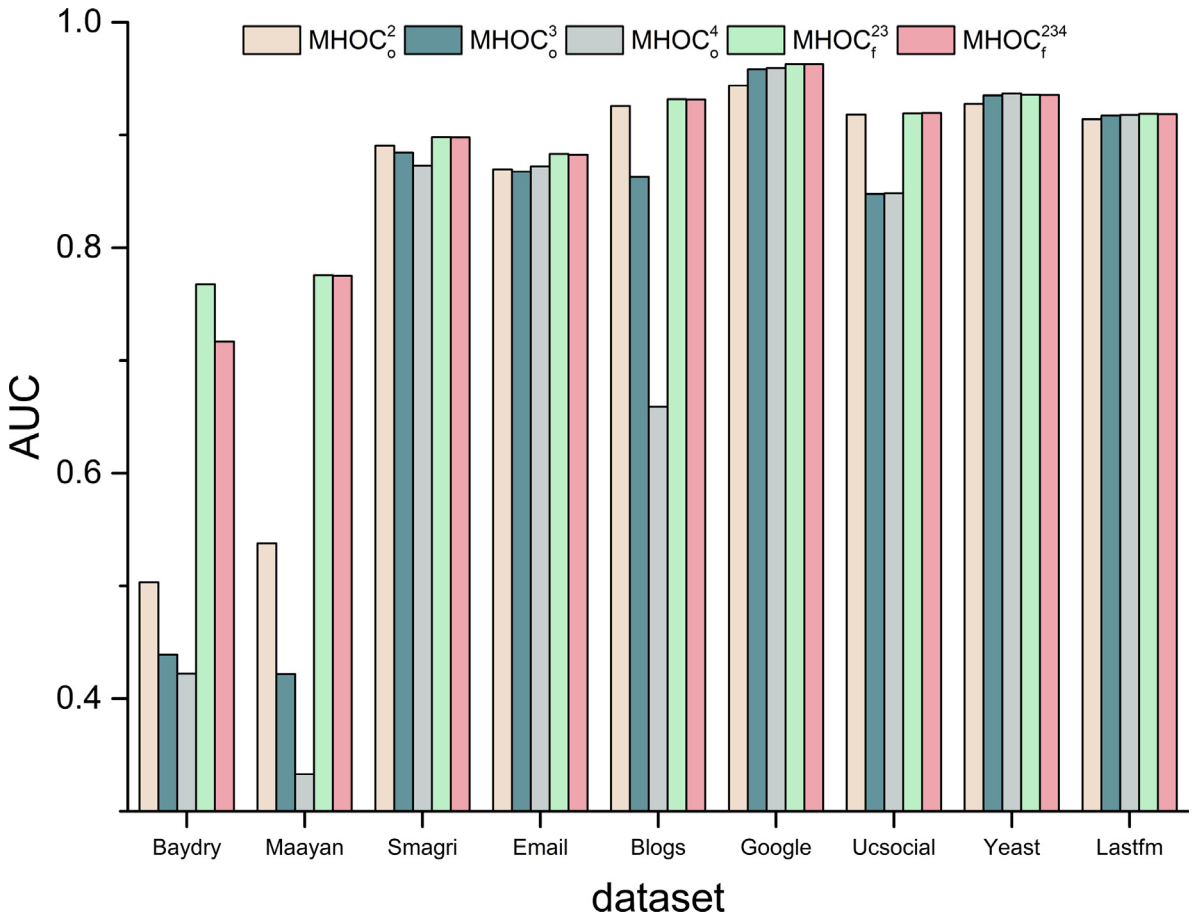

Fig. 2. Influence of different high-order cluster structures on AUC. The horizontal axis is the different networks and the vertical axis is the AUC value of the algorithm on the network. $MHOC_o$ indicates that only one kind of high-order clustering coefficient structure is considered, while $MHOC_f$ means that more than two kinds of high-order clustering coefficient structures are fused during the prediction process of 90% training set.

Table 3
Comparison of prediction accuracy of 11 indicators on 9 real-world networks under AUPR measure. By performing 9:1 independent random segmentation on the training set, and achieving an average of 100 times. The best indicators with the best performance are marked in bold and the second-best results are marked with an underline on each network. Among them, $MHOC^*$ is the best result of all methods using high-order information. MHOC$_f^{23}$ is the result of both using 2-order and 3-order information.

DataSet	CN	LNBCN	AA	RA	CAR	PA	CCLP	MI	DLC	CNDP	MHOC$_f^{23}$	$MHOC^*$
Baydry	0.4762	0.4801	0.4542	0.1015	0.4871	0.5067	0.4761	0.2162	0.4614	0.4531	**0.5103**	**0.5103**
Maayan	0.3937	0.3939	0.3785	0.2414	0.3845	0.1876	0.3945	0.1451	0.3914	0.3876	**0.4054**	**0.4054**
Smagri	0.3872	0.3875	0.3659	0.3684	0.3568	0.1059	0.3779	0.3588	0.3907	0.3822	**0.3995**	**0.3995**
Email	0.4767	0.4971	0.4531	0.4782	0.4219	0.0947	0.4852	0.4571	0.4952	0.4152	0.5036	**0.5058**
Blogs	0.0869	0.0859	0.0838	0.0747	0.0820	0.0489	0.0815	0.0830	0.0871	0.0842	**0.0937**	**0.0937**
Google	0.7602	0.7622	0.7597	0.7712	0.7314	0.3435	0.7752	0.7739	0.8438	0.7723	**0.8708**	**0.8708**
Ucsocial	0.3718	0.3714	0.3474	0.3424	0.3744	0.3642	0.3596	0.3852	0.3856	0.3594	**0.3966**	**0.3966**
Yeast	0.0422	0.0430	0.0419	0.0237	0.0421	0.0017	0.0435	0.0067	0.0432	0.0420	**0.0436**	**0.0436**
Lastfm	0.4937	0.4946	0.4812	0.3747	0.5871	**0.6171**	0.5594	0.4174	0.4996	0.4871	0.5962	0.5962

on the Smagri and Yeast networks. The $MHOC^*$ method performs best on the Smagri network (i.e., $MHOC_f^{234}$) and the Yeast network (i.e., $MHOC_o^4$), respectively.

Table 3 shows that the prediction performance of 11 indicators on 9 real-world networks in AUPR metric. We can find $MHOC_f^{23}$ has best predictive performance on eight real-world networks and the second predictive performance on Lastfm networks. For Lastfm network, PA index has the best predictive performance. We can analyze the reason from the network itself. From the topological characteristics of this network, the average degree of the Lastfm network is relatively

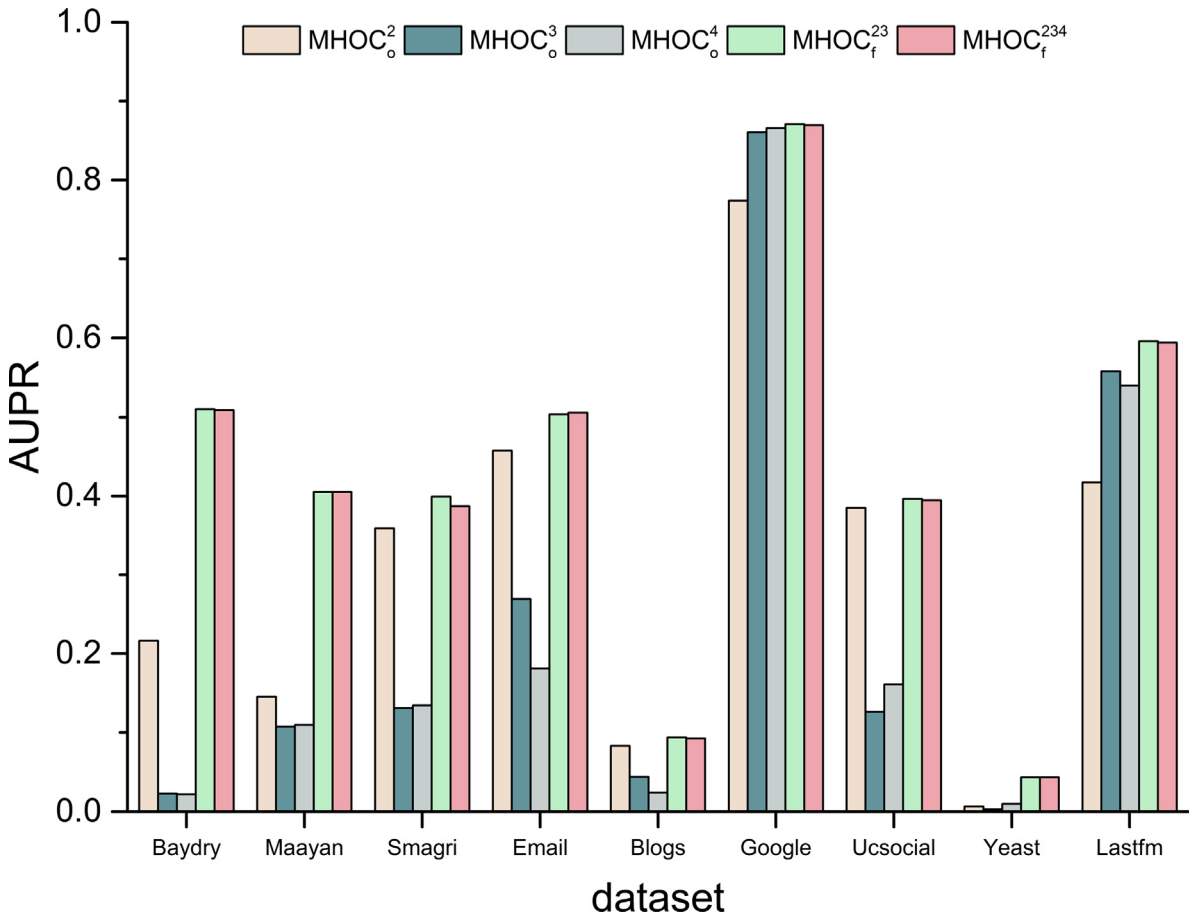

Fig. 3. Influence of different high-order clustering structures on AUPR. The horizontal axis is the different networks and the vertical axis is the AUPR value of the algorithm on the network. $MHOC_o$ indicates that only one kind of high-order clustering coefficient structure is considered, while $MHOC_f$ means that more than two kinds of high-order clustering coefficient structures are Fused during the prediction process of 90% training set.

high, which is also in line with the definition of the PA algorithm. The node with the highest degree will connect first, and the "rich people" phenomenon will appear in network.

From Tables 2 and 3, it is not difficult to find that $MHOC^*$ method usually performs better, and they have the approximate prediction performance in all most cases with respect to $MHOC_f^{23}$ and $MHOC^*$. Considering the calculation time and accuracy, we specify uniformly $MHOC_f^{23}$ as $MHOC$ in the following subsection for convenience of presentation.

5.4. Algorithm stability verification

To verify the stability of MHOC, we use MHOC to represent $MHOC_f^{23}$ and perform experiment with different training set ratios. The AUC and AUPR results of the models are shown in Figs. 4 and 5.

With the change of different training ratio, we compare the prediction result of our method with other baseline under AUC metric in Fig. 4. In all networks, the prediction results increase with the proportion of the training set in most cases. On the contrary, in Figs. 4(a) and 4(b), the performance of MI index decreases as the proportion of the training set increases. It is estimated that the network structure required by the MI index is destroyed in the process of link extraction. The results show that the performance of the MHOC method has different degrees of improvement compared with the baseline method under different test set partition ratios. It is obvious that our MHOC method has better performance than all other baseline methods under different percentages of training set in almost networks. But in the Baydry network, when the proportion of training set is lower at the beginning, PA index is obviously better than MHOC method. This is due to the fact that a low training ratio indicates that more links are being removed from networks, which results in a significant amount of high-order network structures to be destroyed. This causes a decline in the high-order clustering coefficient, which in turn leads to a decline in the performance of the MHOC index.

Fig. 5 shows the result of AUPR. From the figure, we observe that the MHOC performs best on Baydry, Email and Blogs networks, and outperforms the other methods on almost all split training set percentages. However, when the

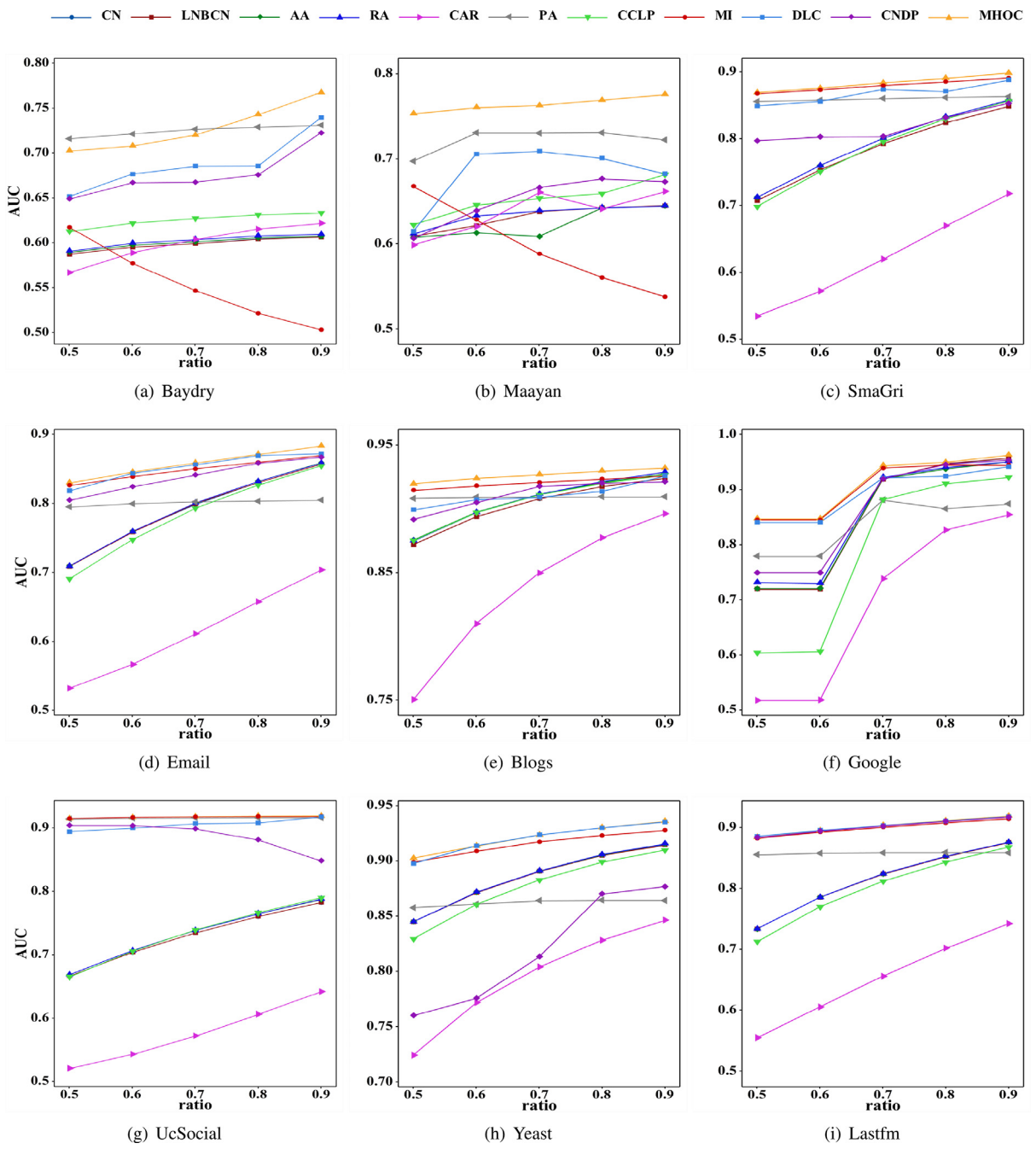

Fig. 4. The prediction performance of 11 indices measured by AUC in nine networks with different training set partition ratio. Each point represents the average value over 100 independent implementations. X-axis is training set ratio from 0.5 to 0.9, Y-axis is the AUC value.

training set partition ratio is low, the performance of MHOC method will significantly change on SmaGri, Google and Ucsocial networks. As mentioned before, when too many links are removed, the high-order structure in the network is largely destroyed, which leads to the decrease of the influence of the high-order clustering coefficient and causes the performance of MHOC to suffer. In addition, the performance of PA method on SmaGri, Email, Blogs and Google networks has a large gap with other methods, which indicates that PA is a very unstable method. In Fig. 5 we are able to observe that our proposed MHOC method maintains a better performance on all networks, which indicates that the MHOC method is stable and reliable. Based on the previous experimental results, the prediction performance of the MHOC method always performs well on different networks, whether it is AUC or AUPR evaluation metric.

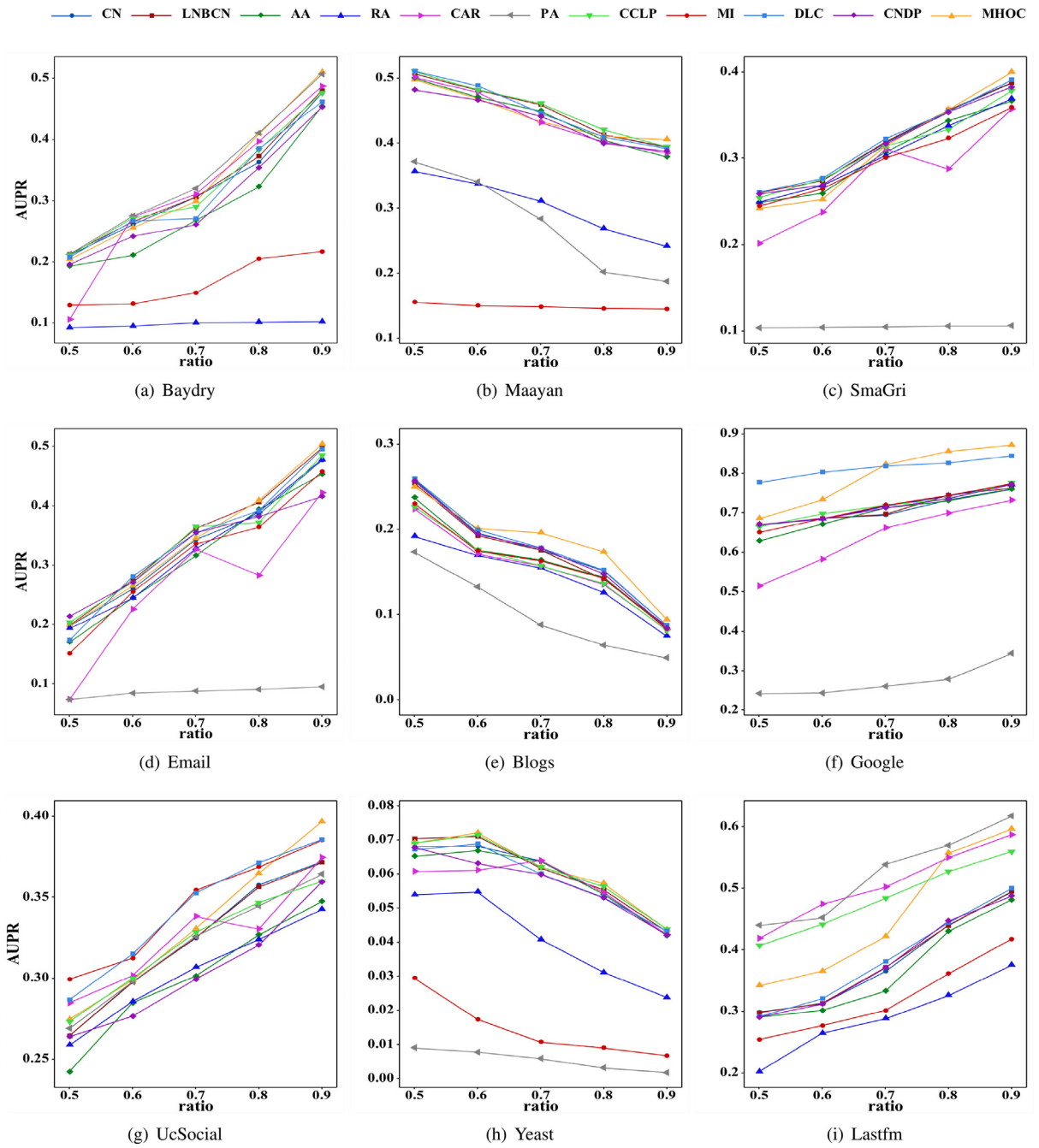

Fig. 5. The prediction performance of 11 indices measured by AUPR in nine networks with different training set partition ratio. Each point represents the average value over 100 independent implementations. X-axis is training set ratio from 0.5 to 0.9, Y-axis is the AUPR value.

5.5. Complexity analysis

Given an undirected network $G = (V, E)$, V is the set of all nodes in the network and E is the set of all the links in the network. Let $N = \mid V \mid$ means the number of the nodes in the network, $M = \mid E \mid$ means the number of the links in the network and $\langle k \rangle$ denotes the average degrees of network G. According to the graph theory, $N\langle k \rangle = 2M$. When calculating MHOC index, for each one node, the computational complexity for calculating the l-order local clustering

Table 4
Time costs of all compared methods on 9 networks (in s). The results are average of 100 independent runs.

DataSet	CN	LNBCN	AA	RA	CAR	PA	CCLP	MI	DLC	CNDP	MHOC
Baydry	0.003	0.008	0.003	0.003	0.004	0.005	0.006	0.005	0.011	0.007	0.007
Maayan	0.003	0.011	0.003	0.003	0.005	0.006	0.007	0.006	0.011	0.008	0.007
Smagri	0.054	0.106	0.053	0.062	0.088	0.064	0.149	0.106	0.194	0.067	0.131
Email	0.061	0.114	0.065	0.066	0.851	0.071	0.182	0.124	0.137	0.092	0.142
Blogs	0.102	0.207	0.113	0.114	0.142	0.122	0.216	0.209	0.291	0.149	0.274
Google	0.057	0.098	0.054	0.063	0.091	0.062	0.166	0.117	0.155	0.076	0.106
Ucsocial	0.112	0.315	0.117	0.126	0.144	0.132	0.302	0.202	0.283	0.146	0.279
Yeast	0.191	0.441	0.198	0.213	0.269	0.230	0.247	0.382	0.702	0.235	0.641
Lastfm	6.735	13.262	8.244	8.415	9.714	8.104	8.514	7.148	17.141	11.054	18.422

coefficients is $O(l\alpha^{l-2}M)$, where α is the arboricity of the graph network [18]. Accordingly, the time cost of the second-order clustering coefficient, the third-order clustering coefficient and the fourth-order clustering coefficient are $O(2M) = O(N\langle k\rangle)$, $O(3\alpha M) = O(1.5\alpha N\langle k\rangle) \approx O(\alpha N\langle k\rangle)$ and $O(4\alpha^2 M) = O(2\alpha^2 N\langle k\rangle) \approx O(\alpha^2 N\langle k\rangle)$, respectively. Therefore, the time cost of computing the clustering coefficients of anyone node in the network is $O(N\langle k\rangle + \alpha N\langle k\rangle + \alpha^2 N\langle k\rangle) \approx O(\alpha^2 N\langle k\rangle)$, and the time cost to calculate the higher-order clustering coefficients of all nodes in the network is $O(\alpha^2 N^2\langle k\rangle)$. After obtaining the clustering coefficient for each node, we next use the common neighbors between the nodes for MHOC, and its time complexity is $O(N\langle k\rangle^2)$ [39]. Therefore, the overall time complexity of computing the similarity score of all unconnected node pairs is $O(N\langle k\rangle^2 + \alpha^2 N^2\langle k\rangle) \approx O(\alpha^2 N^2\langle k\rangle)$.

All experiments are implemented in Julia programming language and perform on Windows10 with 2 processors and 32 GB RAM. Table 4 lists the practical running time of the 11 indicators over 9 networks for 100 runs.

6. Conclusion

The evolution of complex networks is often accompanied by the emergence of high-order structures, and the high-order clustering coefficient can measure the closed likelihood of high-order clique structures in the network. In this paper, we explore the effect of high-order clustering coefficient and propose the MHOC link prediction method that fuses the different order clustering coefficient with the aids of mutual information theory. By using AUC and AUPR evaluation metrics on nine real-world networks, we find that the prediction performance of MHOC method is affected when the proportion of split training set is low. Nevertheless, MHOC shows excellent performance and reliable stability on all networks. In addition, we compared the effect of using different order clustering coefficients on the prediction performance, and found that the fusion of order 2 and 3 has better prediction performance than the methods.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yabing Yao reports financial support was provided by the National Natural Science Foundation of China (No. 62062010). Yabing Yao reports financial support was provided by the Longyuan Youth Innovation and Entrepreneurship Talents Team Project of Gansu (No. 2021LQTD24). Yabing Yao reports financial support was provided by the Science and Technology Planning Project of Guangxi (No. AD19245101). Yabing Yao reports financial support was provided by the Higher Education Innovation Fund project of Gansu (No. 2022A-022). Yabing Yao reports financial support was provided by the Science and Technology Project of Lanzhou (No. 2018-4-56).

Data availability

Our datasets and codes are available on https://github.com/yabingyao/MHOC4LinkPrediction.git.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (No. 62062010), the Longyuan Youth Innovation and Entrepreneurship Talents Team Project of Gansu (No. 2021LQTD24), the Science and Technology Planning Project of Guangxi (No. AD19245101), the Higher Education Innovation Fund project of Gansu (No. 2022A-022), the Science and Technology Project of Lanzhou (No. 2018-4-56). We thank the support from Higher-Order Network Reading Group supported by the Save 2050 Programme jointly sponsored by Swarma Club and X-Order.

References

[1] Y. Yao, R. Zhang, F. Yang, J. Tang, Y. Yuan, R. Hu, Link prediction in complex networks based on the interactions among paths, Physica A 510 (2018) 52–67.
[2] C.-C. Chu, H.H.-C. Iu, Complex networks theory for modern smart grid applications: A survey, IEEE J. Emerg. Sel. Top. Circuits Syst. 7 (2) (2017) 177–191.
[3] G.A. Pagani, M. Aiello, The power grid as a complex network: A survey, Physica A 392 (11) (2013) 2688–2700.
[4] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031.
[5] N.N. Daud, S.H. Ab Hamid, M. Saadoon, F. Sahran, N.B. Anuar, Applications of link prediction in social networks: A review, J. Netw. Comput. Appl. 166 (2020) 102716.
[6] Q.-M. Zhang, L. Lü, W.-Q. Wang, Yu-Xiao, T. Zhou, Potential theory for directed networks, PLoS One 8 (2) (2013) e55437.
[7] S. Pulipati, R. Somula, B.R. Parvathala, Nature inspired link prediction and community detection algorithms for social networks: A survey, Int. J. Syst. Assur. Eng. Manag. (2021).
[8] A. Kumar, S.S. Singh, K. Singh, B. Biswas, Link prediction techniques, applications, and performance: A survey, Physica A 553 (2020) 124289.
[9] V. Martínez, C. Cano, A. Blanco, ProphNet: A generic prioritization method through propagation of information, BMC Bioinformatics 15 (1) (2014) 1–13.
[10] B. Schwikowski, P. Uetz, S. Fields, A network of protein–protein interactions in yeast, Nature Biotechnol. 18 (12) (2000) 1257–1261.
[11] P. Wang, B. Xu, Y. Wu, X. Zhou, Link prediction in social networks: The state-of-the-art, Sci. China Inf. Sci. 58 (1) (2015) 1–38.
[12] M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich, A review of relational machine learning for knowledge graphs, Proc. IEEE 104 (1) (2015) 11–33.
[13] Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative filtering, in: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 2005, pp. 141–142.
[14] L.A. Adamic, E. Adar, Friends and neighbors on the web, Social Networks 25 (3) (2003) 211–230.
[15] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (4) (2009) 623–630.
[16] Z. Wu, Y. Lin, J. Wang, S. Gregory, Link prediction with node clustering coefficient, Physica A 452 (2016) 1–8.
[17] Z. Wu, Y. Lin, Y. Zhao, H. Yan, Improving local clustering based top-L link prediction methods via asymmetric link clustering information, Physica A 492 (2018) 1859–1874.
[18] H. Yin, A.R. Benson, J. Leskovec, Higher-order clustering in networks, Phys. Rev. E 97 (5) (2018) 052306.
[19] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.
[20] A. Popescul, L.H. Ungar, Statistical relational learning for link prediction, in: IJCAI Workshop on Learning Statistical Models from Relational Data, Vol. 2003, Citeseer, 2003.
[21] M.E. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2) (2001) 025102.
[22] F. Tan, Y. Xia, B. Zhu, Link prediction in complex networks: A mutual information perspective, PLoS One 9 (9) (2014) e107056.
[23] A. Kumar, S.S. Singh, K. Singh, B. Biswas, Level-2 node clustering coefficient-based link prediction, Appl. Intell. 49 (7) (2019) 2762–2779.
[24] Z. Huang, Link prediction based on graph topology: The predictive value of generalized clustering coefficient, 2010, Available at SSRN 1634014.
[25] Z. Liu, Q.-M. Zhang, L. Lü, T. Zhou, Link prediction in complex networks: A local naïve Bayes model, Europhys. Lett. 96 (4) (2011) 48007.
[26] Y. Liu, C. Zhao, X. Wang, Q. Huang, X. Zhang, D. Yi, The degree-related clustering coefficient and its application to link prediction, Physica A 454 (2016) 24–33.
[27] Z. Wu, Y. Lin, H. Wan, W. Jamil, Predicting top-L missing links with node and link clustering information in large-scale networks, J. Stat. Mech. Theory Exp. 2016 (8) (2016) 083202.
[28] T. Zhou, Y.-L. Lee, G. Wang, Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms, Physica A 564 (2021) 125532.
[29] T. Zhou, Prodoigresses and challenges in link prediction, iScience 24 (11) (2021) 103217.
[30] C.V. Cannistraci, G. Alanis-Lobato, T. Ravasi, From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks, Sci. Rep. 3 (1) (2013) 1613.
[31] M. Wang, X. Lou, B. Cui, A degree-related and link clustering coefficient approach for link prediction in complex networks, Eur. Phys. J. B 94 (1) (2021) 33.
[32] S. Rafiee, C. Salavati, A. Abdollahpouri, CNDP: Link prediction based on common neighbors degree penalization, Physica A 539 (2020) 122950.
[33] B. Zhu, Y. Xia, An information-theoretic model for link prediction in complex networks, Sci. Rep. 5 (1) (2015) 13707.
[34] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[35] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632.
[36] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 107–117.
[37] D.J. Klein, M. Randić, Resistance distance, J. Math. Chem. 12 (1) (1993) 81–95.
[38] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, IEEE Trans. Knowl. Data Eng. 19 (3) (2007) 355–369.
[39] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (4) (2009) 046122.
[40] W. Liu, L. Lü, Link prediction based on local random walk, Europhys. Lett. 89 (5) (2010) 58007.
[41] Z. Xu, C. Pu, J. Yang, Link prediction based on path entropy, Physica A 456 (2016) 294–301.
[42] A. Clauset, C. Moore, M.E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98–101.
[43] R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. 106 (52) (2009) 22073–22078.
[44] L. Pan, T. Zhou, L. Lü, C.-K. Hu, Predicting missing links and identifying spurious links via likelihood analysis, Sci. Rep. 6 (1) (2016) 22955.
[45] H.C. White, S.A. Boorman, R.L. Breiger, Social structure from multiple networks. I. Blockmodels of roles and positions, Am. J. Sociol. 81 (4) (1976) 730–780.
[46] E.M. Airoldi, D. Blei, S. Fienberg, E. Xing, Mixed membership stochastic blockmodels, Adv. Neural Inf. Process. Syst. 21 (2008).
[47] L. Getoor, N. Friedman, D. Koller, A. Pfeffer, Learning probabilistic relational models, in: Relational Data Mining, Springer, 2001, pp. 307–335.
[48] D. Heckerman, C. Meek, D. Koller, Probabilistic entity-relationship models, PRMs, and plate models, in: Introduction to Statistical Relational Learning, Vol. 2007, MIT Press Cambridge, MA, USA, 2007, pp. 201–238.
[49] K. Yu, W. Chu, S. Yu, V. Tresp, Z. Xu, Stochastic relational models for discriminative link prediction, Adv. Neural Inf. Process. Syst. 19 (2006).
[50] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 701–710.
[51] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 855–864.
[52] M. Radmanesh, A.A. Rezaei, N. Al Khafaf, M. Jalili, Topological deep network embedding, in: 2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC, IEEE, 2020, pp. 476–481.
[53] R.M. Gray, Entropy and Information Theory, Springer Science & Business Media, 2011.
[54] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inform. Theory 37 (1) (1991) 145–151.

[55] A. Lesne, Shannon entropy: A rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics, Math. Struct. Comput. Sci. 24 (3) (2014).
[56] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[57] T.M. Cover, Elements of Information Theory, John Wiley & Sons, 1999.
[58] M.Á. Serrano, M. Boguna, Clustering in complex networks. I. General formalism, Phys. Rev. E 74 (5) (2006) 056114.
[59] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, et al., Topological structure analysis of the protein–protein interaction network in budding yeast, Nucleic Acids Res. 31 (9) (2003) 2443–2450.
[60] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.
[61] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 233–240.
[62] M.E. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (20) (2002) 208701.
[63] A. Muscoloni, C.V. Cannistraci, Early retrieval problem and link prediction evaluation via the area under the magnified ROC, 2022, Preprints, http://dx.doi.org/10.20944/preprints202209.0277.v1.
[64] T. Zhou, Discriminating abilities of threshold-free evaluation metrics in link prediction, SSRN Electr. J. (2022).
[65] M. Al Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning, in: SDM06: Workshop on Link Analysis, Counter-Terrorism and Security, Vol. 30, 2006, pp. 798–805.
[66] H. Schütze, C.D. Manning, P. Raghavan, Introduction to Information Retrieval, Vol. 39, Cambridge University Press Cambridge, 2008.
[67] R.A. Rossi, N.K. Ahmed, The network data repository with interactive graph analytics and visualization, in: AAAI, 2015, URL: https://networkrepository.com.
[68] L.A. Adamic, N. Glance, The political blogosphere and the 2004 US election: Divided they blog, in: Proceedings of the 3rd International Workshop on Link Discovery, 2005, pp. 36–43.
[69] B. Rozemberczki, R. Sarkar, Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models, in: Proceedings of the 29th ACM International on Conference on Information and Knowledge Management, CIKM '20, ACM, 2020.