Entropy-based link prediction in weighted networks*

Zhongqi Xu(许忠奇)[1], Cunlai Pu(濮存来)[1,2,†], Rajput Ramiz Sharafat[1], Lunbo Li(李伦波)[1], and Jian Yang(杨健)[1]

[1] *Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing* 210094, *China*
[2] *Department of Industrial and Systems Engineering, University of Florida, Gainesville* 32611, *USA*

(Received 18 July 2016; revised manuscript received 25 October 2016; published online 22 December 2016)

Information entropy has been proved to be an effective tool to quantify the structural importance of complex networks. In a previous work [Xu *et al*. *Physica A*, **456** 294 (2016)], we measure the contribution of a path in link prediction with information entropy. In this paper, we further quantify the contribution of a path with both path entropy and path weight, and propose a weighted prediction index based on the contributions of paths, namely weighted path entropy (WPE), to improve the prediction accuracy in weighted networks. Empirical experiments on six weighted real-world networks show that WPE achieves higher prediction accuracy than three other typical weighted indices.

Keywords: link prediction, weighted networks, information entropy

PACS: 89.75.Hc, 89.75.Fb, 89.20.Hh **DOI:** 10.1088/1674-1056/26/1/018902

1. Introduction

In the field of network science, real-world complex systems are abstracted as complex networks, in which nodes represent individuals and links denote the connections or interactions between individuals.[1–3] Nowadays, although we can obtain abundant data of various complex systems due to advanced technologies, it is demonstrated that large parts of the data of the complex systems are still not available, and there are non-ignorable errors in the data that we obtain.[4,5] Thus, new methods are needed to process, correct, and make predictions from the data. Link prediction methods aim to predict the missing or future links among network data.[6,7] Specifically, they estimate the existence likelihood of links between two nodes based on observed links and nodes' attributes. Link prediction has broad applications,[8] such as detecting potential interactions in protein–protein interaction network,[9] recommending friends and goods in online social networks,[10] exploring potential coauthor relationships in collaboration networks,[11] and so on.

Previous algorithms are basically from the field of machine learning, including supervised learning,[12] Markov chain,[13] and likelihood estimation.[14] These algorithms heavily depend on the attributes of nodes, and do not seriously consider the structural characteristics of networks. Besides, their computation cost is inhibitive for large real-world networks.[15] Recently, the booming network science community has achieved deeper insights into the structure of complex networks,[1] and further stimulates the research of link prediction.[6,16] Lots of prediction algorithms based on structural similarity are proposed, which can be classified into three types: local indices,[17–22] quasi-local indices,[23,24] and global indices.[25,26] For example, common neighbors (CN),[18] preferential attachment (PA),[17] Adamic–Adar (AA),[21] and resource allocation (RA)[22] are local indices, which only use the nearest neighborhood information. Katz,[25] Leicht–Holme–Newman (LHN),[26] and SimRank,[27] which need the knowledge of the whole network topology, are global ones. Quasi-local indices, including local path (LP),[23] local random walk (LRW),[24] and superposed random walk (SRW),[24] need more topological information than local indices, but less topological information than global ones. Generally, local indices have the lowest prediction accuracy, but their computational cost is the smallest among the three types of similarity indices. Global indices are the opposite of local indices, while quasi-local ones are the trade-off between them.

Recently, information entropy has been employed to measure the complexity of the topological structures of complex networks.[28] Results showed that information entropy can better capture the topological difference than the other typical network measurements.[29,30] Moreover, information entropy has a natural connection with link prediction problems, in which the probability of a missing link between two nodes can be transformed into the corresponding information entropy. Thus, researchers began to apply information entropy theory to link prediction problems in complex networks.[31–33] For example, Tan *et al.*[31] reexamined the role of common neighbors in link prediction by using the mutual information, and proposed a mutual information-based similarity index. Xu

*Project supported by the National Natural Science Foundation of China (Grant Nos. 61201173 and 61304154), the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20133219120032), the Postdoctoral Science Foundation of China (Grant No. 2013M541673), and China Postdoctoral Science Special Foundation (Grant No. 2015T80556).

†Corresponding author. E-mail: pucunlai@njust.edu.cn

© 2017 Chinese Physical Society and IOP Publishing Ltd

et al.[32] derived the information entropy of a path, and studied the contributions of paths in link prediction based on path entropy, and finally provided a path entropy based similarity index. Simulation results[31–33] showed that the similarity indices based on information entropy have higher prediction accuracy than the other types of similarity indices.

In real society, the interactions between individuals are of different strengths for many complex networks, which are called weighted networks.[2,34] For instance, in the air traffic network, the weight of a link is measured by the number of passengers in the related flight. In the router-level of the Internet, the weights of links are generally correlated with the bandwidth of the physical connections or the cost for data transmission between routers. In social networks, weights of links are related to the interacting times or frequencies between individuals. Therefore, it is reasonable to consider the link weights when designing link prediction algorithms to further improve the prediction accuracy. So far there have been a few tries in the literature. Murata and Moriyasu[35] improved the CN and AA indices by using the link weights information. Bai *et al.*[36] developed the weighted version of the LP index. Lü *et al.*[37] particularly explored the role of weak ties in link prediction and found that weak ties can improve the prediction accuracy effectively. Simulation results showed that the weighted version of these typical similarity indices indeed have a higher accuracy than the original ones. However, they still have a large space to improve. For example, the weighted and unweighted versions of AA and CN indices only utilize the information of paths with length 2, which can be further improved by considering longer paths, such as paths with length 3. In addition, although the LP index considers paths of length 3, it only cares about the number of paths for unweighted networks or the sum of link weights for weighted networks, while the topological information of those paths can be further explored to improve the prediction accuracy.

In this paper, we apply information entropy to link prediction in weighted networks. Specifically, we reexamine the role of paths in link prediction by considering both the path entropy and the path weight, and propose a weighted similarity index applicable to the weighted networks. Through simulation on real-world weighted networks, we show the prediction accuracy of our index and make comparisons with other typical weighted indices.

2. Path entropy

Quantitative measurement of complex networks is a hot topic in network science,[1] and various measurements are proposed, including degree, betweenness, closeness, k-core number, and so on.[2] However, most of the measurements are for nodes and links, and only a very few are particularly for paths, such as path length[38] and path attack centrality.[39] In our latest work,[32] we applied information theory to measure the importance of paths, and specifically we derived the entropy of a path. Assuming that L^1_{ab} (L^0_{ab}) represents the event that there is (not) a link between nodes a and b in the network without degree correlation. Then, the probability of L^1_{ab} is calculated by

$$\begin{aligned} P(L^1_{ab}) &= 1-P(L^0_{ab}) = 1-\prod_{i=1}^{k_b}\frac{(M-k_a)-i+1}{M-i+1} \\ &= 1-\frac{C^{k_b}_{M-k_a}}{C^{k_b}_{M}}, \end{aligned} \tag{1}$$

where k_a (k_b) is the degree of a (b), and M is the total link number in the network. Based on the definition of entropy and Eq. (1), we obtain the entropy of L^1_{ab} as

$$I(L^1_{ab}) = -\log(P(L^1_{ab})) = -\log\left(1-\frac{C^{k_b}_{M-k_a}}{C^{k_b}_{M}}\right). \tag{2}$$

For a simple path $D = v_0 v_1 \ldots v_\delta$ of length δ, the occurrence possibility can be calculated approximately by

$$P(D) \approx \prod_{i=0}^{\delta-1} P(L^1_{v_i v_{i+1}}), \tag{3}$$

which indicates that the occurrence probability of a path is approximately equal to the product of its links' occurrence probabilities. Thus, the entropy of path D can be calculated as[32]

$$\begin{aligned} I(D) &= -\log(P(D)) \approx \sum_{i=0}^{\delta-1} I(L^1_{v_i v_{i+1}}) \\ &\approx -\log\left(\prod_{i=0}^{\delta-1}\left(1-\frac{C^{k_{v_i}}_{M-k_{v_{i+1}}}}{C^{k_{v_i}}_{M}}\right)\right), \end{aligned} \tag{4}$$

which indicates that the entropy of a path approximates to the sum of its links' entropies. Furthermore, path entropy takes both path length and node degree information into consideration. The longer the path is or the smaller the node degree is, the larger the path entropy is. Through the basic definition of entropy, we know that if a path has a large path entropy, its occurrence probability should be low, in other words, the existence of the path is important to the network. Generally, path entropy is more discriminating than the other measurements. The reason is that when computing the path entropy, we not only consider the path length, but also node degrees, and even the order of the nodes in the path.

3. Prediction index based on path weight and path entropy

Although the weights of links and weighted networks have been well discussed, there is no specific result about the weights of paths. In traffic routing protocols, we usually choose the optimal path, of which the sum of the cost of all

links is the smallest among the candidate paths.[40–42] Enlightened by this, here we define the weight of path D as the sum of its links' weights,

$$W_D = \sum_{t=0}^{\delta-1} W_{v_t v_{t+1}}, \tag{5}$$

where $W_{v_t v_{t+1}}$ is the weight of the link with end nodes v_t and v_{t+1}.

In the framework of information theory, the probability of link existence between two nodes can be expressed with information entropy. Then, the link prediction problem can be defined as the conditional entropy,[31] that is

$$I(L_{ab}^1|G') = I(L_{ab}^1) - I(L_{ab}^1;G'), \tag{6}$$

where G' is the topological structure information we know, based on which we make the prediction of the event L_{ab}^1. $I(L_{ab}^1;G')$ is the joint entropy, which quantitatively measures how much the existence of G' leads to the decrease of uncertainty of the event L_{ab}^1. In this paper, we consider the contributions of simple paths in link prediction. Thus, we have $G' = \bigcup_{i=2}^{l} \{D_{ab}^i\}$, where $\{D_{ab}^i\}$ is the set of all simple paths of length i between a and b, and l is the maximum length of simple paths we consider in the network.

Previous results showed that the longer the path is, the less important the path is in link prediction. Moreover, results demonstrated that the link weights can be used to improve the prediction accuracy. Here we combine the path length, path weight, and path entropy to calculate the contribution of path D with length i in the link prediction:

$$I(L_{ab}^1;D) \approx \frac{I(D)\cdot {W_D}^{\alpha}}{i-1}, \tag{7}$$

where α is a free parameter, which controls the influence of path weights. When $\alpha = 0$, path weights are ignored. When $\alpha > 0$, the path with a large weight is thought to have a large contribution. Otherwise, when $\alpha < 0$, the path with a small weight is thought to have a large contribution. $1/(i-1)$ is the optimal penalty factor that we found for suppressing the contributions of long paths. Then, we give the definition of our prediction index, namely weighted path entropy (WPE), by considering the contributions of all simple paths, which is

$$\begin{aligned} S_{ab}^{\mathrm{WPE}} &= I(L_{ab}^1|\bigcup_{i=2}^{l}\{D_{ab}^i\}) \\ &= I(L_{ab}^1) - I(L_{ab}^1;\bigcup_{i=2}^{l}\{D_{ab}^i\}) \\ &= I(L_{ab}^1) - \sum_{i=2}^{l}\{\frac{1}{i-1}(\sum_{D\in\{D_{ab}^i\}}\{{W_D}^{\alpha}\cdot I(D)\})\}. \end{aligned} \tag{8}$$

Based on Eqs. (2), (4), (5), and (8), we finally obtain the complete expression of WPE index

$$\begin{aligned} S_{ab}^{\mathrm{WPE}} = \log\left(\frac{C_M^{k_a}}{C_M^{k_a}-C_{M-k_b}^{k_a}}\right) - \sum_{i=2}^{l}\left\{\frac{1}{i-1}\sum_{D\in\{D_{ab}^i\}}\right. \\ \left.\times\left\{\left\{\sum_{t=0}^{i-1}W_{v_t v_{t+1}}\right\}^{\alpha}\cdot\sum_{j=0}^{i-1}\log\left(\frac{C_M^{k_{v_j}}}{C_M^{k_{v_j}}-C_{M-k_{v_{j+1}}}^{k_{v_j}}}\right)\right\}\right\}. \end{aligned} \tag{9}$$

4. Problem description and standard metrics

Assuming an undirected network $G(V,E,W)$, where V, E, and W denote the sets of nodes, links, and link weights, respectively. In W, $W_{xy} = W_{yx}$, which means there is no direction for the weight between two nodes. To measure the predicting ability of an index, E is usually randomly divided into two parts: a training set E^{T} (in this paper, 90% of all links) and a probe set E^{P}(10% of all links). Clearly, $E = E^{\mathrm{T}} \cup E^{\mathrm{P}}$ and $E^{\mathrm{T}} \cap E^{\mathrm{P}} = \emptyset$.

Two standard metrics, AUC and precision, are applied to quantify the prediction quality. To calculate AUC, n times of independent score comparisons are made between node pairs in E^{P} and $U-E$. Each time we select a node pair randomly from these two sets respectively and compare their scores. If there are n' times that the score of the link from E^{P} is higher (or smaller in the case of WPE) than the link from $U-E$, and n'' times that they have the same scores, AUC is calculated as

$$\mathrm{AUC} = \frac{n' + 0.5n''}{n}. \tag{10}$$

AUC should be about 0.5 if all the scores are generated from an independent and identical distribution. Precision aims to measure the ability to predict top ranked links. If among the top L links ranked by the scores, there are m links belonging to E^{P}, the precision is calculated as

$$\mathrm{Precision} = \frac{m}{L}. \tag{11}$$

Note that for the WPE index, node pairs with small scores (S^{WPE}) have large probability to be connected by links. Thus, when calculating the precision, we rank the links based on their S^{WPE} from the smallest to the largest, and thus the top ranked links have the smallest scores, which is the opposite to the other prediction indices.

Next, we introduce three indices and their weighted versions that we use for comparison. They are common neighbors (CN),[18] Adamic–Adar index (AA),[21] and local path (LP),[23] which are defined as

$$S_{ab}^{\mathrm{CN}} = |O_{ab}|, \tag{12}$$

$$S_{ab}^{\mathrm{AA}} = \sum_{c\in O_{ab}} \frac{1}{\log(|\Gamma(c)|)}, \tag{13}$$

$$S^{\mathrm{LP}} = \boldsymbol{A}^2 + \varepsilon \boldsymbol{A}^3, \tag{14}$$

where O_{ab} is the set of common neighbors of node a and b, and $\Gamma(c)$ is the set of neighbors of node c. $\boldsymbol{A}$ is the adjacency matrix, and we set $\varepsilon = 0.01$ to obtain a near optimal prediction accuracy. In addition, their parameter-dependent weighted versions, WCN,[35] WAA,[35] and WLP,[36] are respectively defined as

$$S_{ab}^{\mathrm{WCN}} = \sum_{c\in O_{ab}} (W_{ac}{}^{\alpha} + W_{cb}{}^{\alpha}), \tag{15}$$

$$S_{ab}^{\mathrm{WAA}} = \sum_{c\in O_{ab}} \frac{(W_{ac}{}^{\alpha} + W_{cb}{}^{\alpha})}{\log(1+S_c)}, \tag{16}$$

$$S_{ab}^{\mathrm{WLP}} = \sum_{c\in O_{ab}} (W_{ac}{}^{\alpha} + W_{cb}{}^{\alpha}) + \varepsilon \sum_{(i,j)\in l_{a\to b}} (W_{ai}{}^{\alpha} + W_{ij}{}^{\alpha})(W_{ij}{}^{\alpha} + W_{jb}{}^{\alpha}), \tag{17}$$

where $S_c = \sum_{z\in\Gamma(c)} W_{cz}{}^{\alpha}$. $l_{a\to b}$ is the path of length three between nodes a and b. i and j are the intermediate nodes in the path $l_{a\to b}$. CN and AA as well as their weighted versions are all taken as local indices, which are only based on the nearest neighbors. LP and its weighted version are quasi-local indices, since they consider the next-nearest neighbors.

5. Results

Six weighted networks from disparate fields are used to test the accuracy for various prediction indices. The directions of links are ignored. The self-connections and multiple links are deleted from the network data. For the unconnected networks, we select the maximum connected components for experiments. The statistics of these networks are summarized in Table 1. i) Les.[43] This undirected network contains co-appearances of characters in Victor Hugo's novel 'Les Misérables'. A node represents a character and a link between two nodes shows that these two characters appeared in the same chapter of the book. The weight of each link indicates how often such a co-appearance occurred. ii) USAir.[44] This is the network of US air transportation, where nodes represent airports, links represent routes between airports, and the weight of a link is the frequency of flights between two airports. iii) Bomb.[45] This undirected network contains contacts between suspected terrorists involved in the train bombing of Madrid on March 11, 2004 as reconstructed from newspapers. A node represents a terrorist and a link between two terrorists shows that there was contact between the two terrorists. The link weights denote how strong a connection was, by considering the friendship and co-participation in training camps or previous attacks. iv) Bible.[46] This undirected network contains nouns (places and names) of the King James version of the Bible and information about their co-occurrences. A node represents one of the above noun types and a link indicates that two nouns appeared together in the same Bible verse. The link weights show how often two nouns occurred together. v) Florida.[47] This network contains the carbon exchanges in the cypress wetlands of South Florida during the dry season. Nodes represent taxa and a link denotes that a taxon uses another taxon as food with a given trophic factor, which is the basis of link weight. vi) C. elegans.[48] This is the neural network of the nematode worm C. elegans, where a node represents a neuron, a link joins two nodes if the corresponding neurons have synaptic contacts, and the weight represents the number of synapses between two neurons.

Table 1. The topological statistics of six real-world weighted networks. $|v|$ is the number of nodes. $|E|$ represents the number of links. $\langle k\rangle$ is the average degree. H represents the degree heterogeneity, defined as $H = \frac{\langle k^2\rangle}{\langle k\rangle^2}$. C is the clustering coefficient.

Networks	$\|v\|$	$\|E\|$	$\langle k\rangle$	H	C
Les	77	278	7.2338	1.9654	0.4634
USAir	332	2126	12.8072	3.4639	0.7494
Bomb	70	256	7.3429	2.3645	0.7122
Bible	1773	9131	10.3001	4.0115	0.7208
Florida	128	2106	32.9063	1.2307	0.3346
C. elegans	297	2148	14.4646	1.8008	0.3079

To investigate the ability of our prediction index, we perform experiments on the six weighted networks, and make comparisons with three other typical indices. Both the unweighted and weighted versions of these indices are tested. The link weights of the networks are only considered when calculating the weighted versions of these indices. In addition, for the weighted versions of these indices, we adjust the control parameter α to achieve the near optimal performances measured by AUC and precision. For the PE and WPE indices, $l = 2$ means only paths with length of 2 are used in the calculation, while $l = 3$ indicates that paths with lengths of both 2 and 3 are used in the calculation. In theory, we can use the information of all paths in link prediction, but in real experiments, we usually consider only short paths, since short paths are more important than long paths in link prediction. On the other side, the longer the paths are, the larger the computational cost is. Our simulation results are shown in Tables 2 and 3, as well as Figs. 1 and 2. Each value is the average of 100 independent runs. In the tables, the maximum performances for each network are marked in bold font.

Tables 2 and 3 show that it is generally necessary to consider the link weights since the weighted version of these indices achieves higher prediction accuracy than the unweighted ones when the appropriate control parameter is considered (the near optimal parameters α_{opt} are given in the tables). Furthermore, Table 2 indicates that from the perspective of AUC, the prediction accuracy of all the unweighted indices is already high for all the six real-world networks, so that the improvements of the weighted versions are not significant. However, as shown in Table 3, the improvements of the weighted versions are more obvious from the perspective of precision.

Moreover, we can see that for all the unweighted and weighted prediction indices, WPE always has the best performances for all the six real-world networks. Especially, for the networks of Bible, Florida, C. elegans, and USAir, we can see from Table 3 that PE and WPE are much better than the other prediction indices. It is worth mentioning that for PE and WPE, extra consideration of the contributions of longer paths is not always necessary and sometimes adverse for link prediction. For instance, we can see that for some networks $l = 2$ is better than $l = 3$.

Table 2. Prediction accuracy measured by AUC for the unweighted and weighted indices. For the weighted indices, we present the maximum AUC with the corresponding control parameter α_{opt} in the brackets.

Nets\index	WAA	AA	WLP	LP	WCN	CN	WPE($l=2$)	PE($l=2$)	WPE($l=3$)	PE($l=3$)
Les	0.967477(0.6)	0.962911	0.96254(0.3)	0.955112	0.963042(0.3)	0.955129	**0.97055**(0.6)	0.964639	0.956291(0.4)	0.953417
USAir	0.967732(–0.5)	0.965757	0.954597(–0.3)	0.951873	0.963738(–0.7)	0.953757	**0.971917**(–0.2)	0.969813	0.966167(–1)	0.946596
Bomb	0.944603(0.7)	0.944998	0.938329(0.2)	0.937651	0.934543(0.5)	0.934408	**0.957373**(1.2)	0.954435	0.940593(–0.5)	0.939566
Bible	0.985164(0.2)	0.985157	0.978042(–0.8)	0.975619	0.977955(–0.1)	0.976989	**0.988501**(0.2)	0.987872	0.964088(–0.9)	0.95649
Florida	0.611155(0.1)	0.606072	0.783328(–0.3)	0.668619	0.609324(0.1)	0.604758	0.59479(1.4)	0.562404	**0.863607**(0.1)	0.860174
C. elegans	0.868117(0.1)	0.866639	0.868887(0)	0.868344	0.850465(–0.1)	0.850071	0.87321(0.2)	0.871474	**0.885356**(0)	0.885356

Table 3. Prediction accuracy measured by precision (top-100) for the unweighted and weighted indices.

Nets\index	WAA	AA	WLP	LP	WCN	CN	WPE($l=2$)	PE($l=2$)	WPE($l=3$)	PE($l=3$)
Les	0.2212(0.2)	0.2192	0.2077(0.3)	0.2043	0.2162(0.3)	0.2112	**0.2227**(0.5)	0.2205	0.21(0.2)	0.2076
USAir	0.6514(–0.3)	0.6324	0.6207(–0.4)	0.6094	0.633(–0.3)	0.6115	0.5926(0.1)	0.5854	**0.6704**(–0.1)	0.6672
Bomb	0.2026(0)	0.1988	0.1749(–1)	0.1678	0.1886(0.5)	0.188	**0.2157**(1.6)	0.2101	0.1861(–0.9)	0.1814
Bible	0.5853(–1.5)	0.5602	0.3839(–0.4)	0.3798	0.4375(–0.1)	0.4371	**0.8536**(–0.1)	0.8508	0.3859(0.4)	0.3812
Florida	0.144(0.3)	0.0912	0.2797(–0.2)	0.1314	0.1384(0.2)	0.0965	0.0614(0.1)	0.0228	**0.4826**(0)	0.4826
C. elegans	0.1895(1.4)	0.1328	0.1757(0.6)	0.1371	0.1824(1.5)	0.1384	0.1293(0.2)	0.1262	**0.2044**(0.3)	0.1841

Figures 1 and 2 show how the weight affects the prediction accuracy for the weighted indices. For WPE, we focus on path weight defined as the sum of all the link weights in the path, while the other weighted indices only consider link weights. We can see that for WPE, generally both AUC and precision increase first, and then decrease with the control parameter α, and there are optimal parameters α_{opt} (given in Tables 2 and 3) corresponding to the maximum performances. The curves of the other weighted indices have the similar change trends. These results indicate that the prediction accuracy is sensitive to path weights for WPE and link weights for the others, and we should balance the influences of the weak ties and strong ties by appropriately choosing the control parameter to achieve high prediction accuracy. Moreover, we should be slightly biased to the contributions of weak ties, since the simulations results demonstrate that the optimal parameters are less than 1 for all the six networks except Bomb (for WAA and WCN, C. elegans is also an exception).

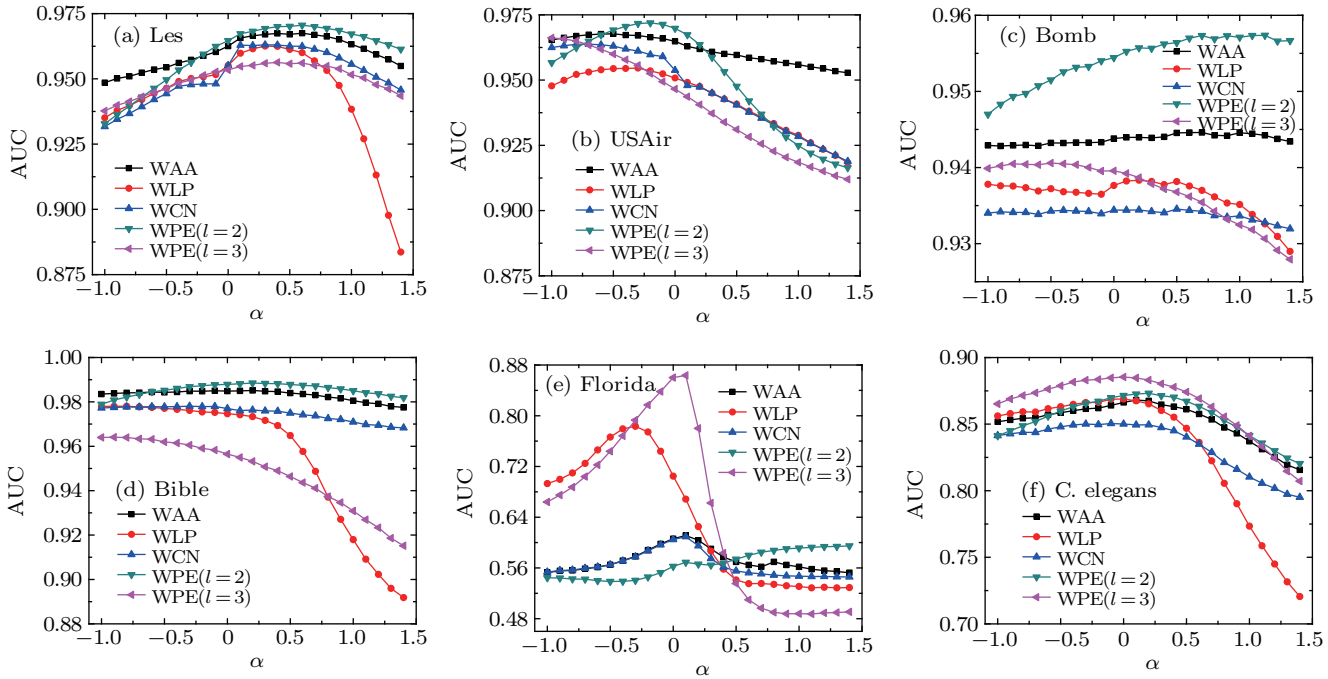


Fig. 1. (color online) AUC versus α for the weighted indices: (a) Les, (b) USAir, (c) Bomb, (d) Bible, (e) Florida, (f) C. elegans.

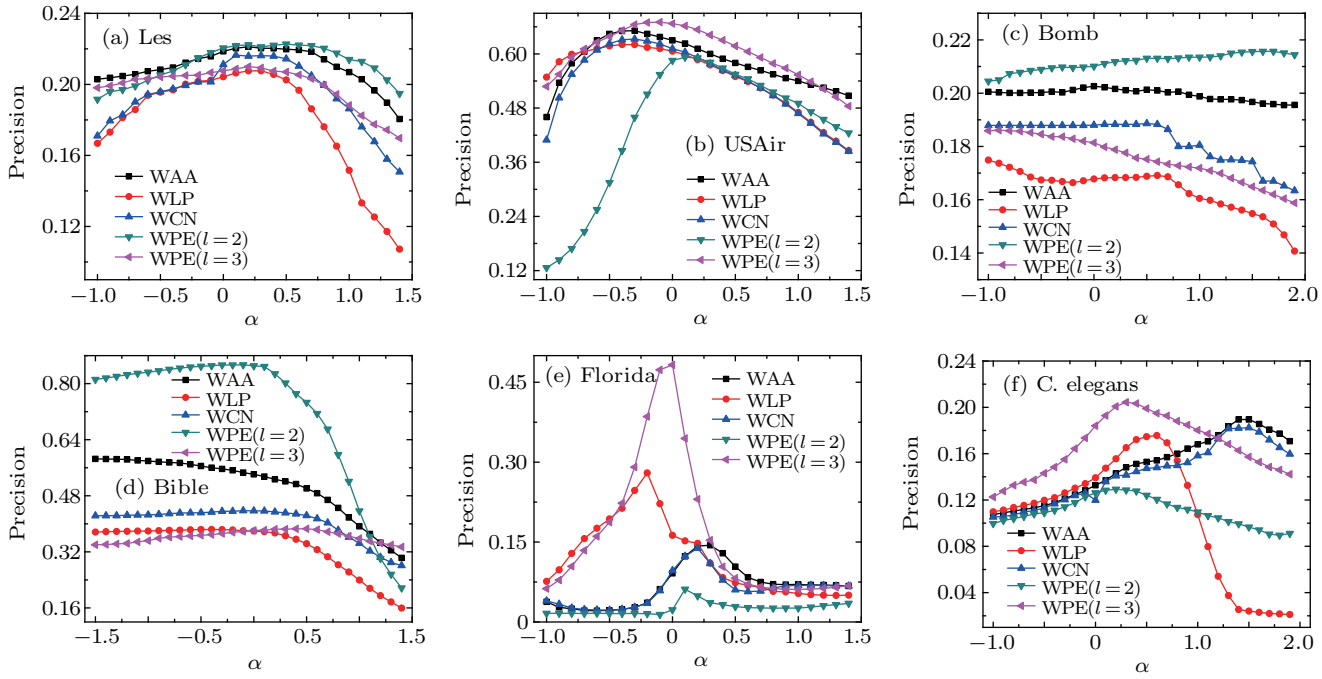


Fig. 2. (color online) Precision (top-100) versus α for the weighted indices: (a) Les, (b) USAir, (c) Bomb, (d) Bible, (e) Florida, (f) C. elegans.

6. Conclusion

In summary, we study the link prediction problem in weighted complex networks from the perspective of information entropy. In fact, the likelihood of a link between two nodes can be converted into entropy, and small entropy corresponds to large probability of link existence. In this paper we consider the contributions of simple paths between node pairs in link prediction. Specifically, we measure the contribution of an existing path by considering its length, entropy, and weight which is further defined as the sum of link weights in the path. Furthermore, we propose a weighted path entropy (WPE) index for link prediction by considering the contributions of all existing simple paths. Through simulation on several real-world weighted networks, we find that WPE has higher prediction ability measured by AUC and precision than the other typical weighted similarity indices. In fact, the PE index proposed in our previous work already has high prediction accuracy. Thus, by appropriately considering the path weight, WPE further improves the prediction accuracy. We also find that a weak tie is more critical than a strong tie in link prediction. Note that in our context, a weak tie refers to a path with small weight, while in the other works a weak tie means a small weight link.

References

[1] Barabási A L 2016 *Network Science* (Cambridge University Press)
[2] Newman M 2010 *Networks: An Introduction* (Oxford University Press)
[3] Chen G R, Wang X F and Li X 2012 *Introduction to Complex Networks: Models, Structures and Dynamics* (Higher Education Press)
[4] Mayer-Schönberger V and Cukier K 2013 *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Houghton Mifflin Harcourt)
[5] Abbasi A, Sarker S and Chiang R H 2016 *J. Assoc. Inf. Syst.* **17** 3
[6] Lü L Y and Zhou T 2011 *Physica A* **390** 1150
[7] Wang P, Xu B W, Wu Y R and Zhou X Y 2015 *Sci. China-Inform. Sci.* **58** 1
[8] Lü L Y, Medo M, Yeung C H, Zhang Y C, Zhang Z K and Zhou T 2012 *Phys. Rep.* **519** 1
[9] Cheng W L and Jian H R 2013 *Bioinformatics* **29** 355
[10] Sherkat E, Rahgozar M and Asadpour M 2015 *Physica A* **419** 80
[11] Newman M E J 2001 *Proc. Natl. Acad. Sci.* **98** 404
[12] Mohammad A H, Vineet C, Saeed S and Mohammad Z 2006 *The Proceedings of the Fourth Workshop on Link Analysis, Counterterrorism and Security*, April 22nd, 2006, Bethesda, USA
[13] Sarukkai R R 2000 *Comput. Netw.* **33** 377
[14] Getoor L and Diehl C P 2005 *ACM SIGKDD Explor. Newslett.* **7** 3
[15] Cui W, Pu C L, Xu Z Q and Yang J 2016 *Physica A* **457** 202
[16] Li Y J, Yin C, Yu H and Liu Z 2016 *Acta Phys. Sin.* **65** 020501 (in Chinese)
[17] Barabási A L and Albert R 1999 *Science* **286** 509
[18] Newman M E J 2001 *Phys. Rev. E* **64** 025102
[19] Kossinets G 2006 *Soc. Netw.* **28** 247
[20] Liben-Nowell D and Kleinberg J 2007 *J. Am. Soc. Inf. Sci. Technol.* **58** 1019
[21] Adamic L A and Adar E 2003 *Soc. Netw.* **25** 211
[22] Zhou T, Lü L Y and Zhang Y C 2009 *Eur. Phys. J. B* **71** 623
[23] Lü L Y, Jin C H and Zhou T 2009 *Phys. Rev. E* **80** 046122
[24] Liu W P and Lü L Y 2010 *Europhys. Lett.* **89** 58007
[25] Katz L 1953 *Psychmetrika* **18** 39
[26] Leicht E A, Holme P and Newman M E J 2006 *Phys. Rev. E* **73** 026120
[27] Jeh G and Widom J 2002 *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, p. 538
[28] Solé R V and Valverde S 2004 *Information Theory of Complex Networks: On Evolution and Architectural Constraints, in Complex Networks* (Springer) pp. 189–207
[29] Halu A, Mukherjee S and Bianconi G 2014 *Phys. Rev. E* **89** 012806
[30] Anand K and Bianconi G 2009 *Phys. Rev. E* **80** 045102
[31] Tan F, Xia Y X and Zhu B Y 2014 *PLoS ONE* **9** e107056
[32] Xu Z Q, Pu C L and Yang J 2016 *Physica A* **456** 294
[33] Zhu B Y and Xia Y X 2016 *PLoS ONE* **11** e0148265
[34] Shen Y 2014 *Physica A* **393** 560

[35] Murata T and Moriyasu S 2007 *IEEE/WIC/ACM International Conference on Web Intelligence* p. 85
[36] Bai M, Hu K and Tang Y 2011 *Chin. Phys. B* **20** 128902
[37] Lü L Y and Zhou T 2010 *Europhys. Lett.* **89** 18001
[38] Pu C L and Cui W 2015 *Physica A* **419** 622
[39] Pu C, Li S, Michaelson A and Yang J 2015 *Phys. Lett.* **379** 1633
[40] Feigenbaum J, Papadimitriou C, Sami R and Shenker S 2005 *Distrib. Comput.* **18** 61
[41] Shen Y, Ren G and Liu Y 2016 *Physica A* **452** 229
[42] Song H Q and Guo J 2015 *Chin. Phys. B* **24** 108901
[43] The data is released by Knuth D E in 1993, available at http://moreno.ss.uci.edu/data.html
[44] The data is released by Batagelj V and Mrvar A in 2006, available at http://vlado.fmf.uni-lj.si/pub/networks/data
[45] Brian H 2006 *AmSci* **94** 400
[46] The Koblenz Network Collection 2015, available at http://konect.uni-koblenz.de/
[47] The data is released by Ulanowicz R E, Bondavalli C and Egnotovich M S in 1998, available at http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/foodweb.htm
[48] Watts D J and Strogatz S H 1998 *Nature* **393** 440