# Hybrid influence based on diversity of degree and H-Index of neighbors

Mingzhe Zhao, Yang Tian, Hui Tian*
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

## ABSTRACT

At present, many scholars have done a great amount of research in link prediction. Researchers have found that the attributes of neighbors cannot be ignored in the study of expressing node similarity. Some scholars synthetically consider the degree and H-index to better express node influence. On this basis, the paper introduces a weight factor $\beta$ to evaluate the role of degree and H-index, then proposes a mixed influence based on diversity of degree and H-index of neighbors model, and carries out experiments on twelve data sets. The experimental results indicate that the link prediction performance can be improved by the weight factor $\beta$ to the hybrid influence of neighbors.

**Keywords:** Complex network, link prediction, node degree, H-index, weight, hybrid influence

## 1. INTRODUCTION

Predicting unobserved connections or future connections based on existing connections in the network topology as a link prediction problem, it is an important topic in physics and computer science, the basic idea is that the existing connection structure of a network can indicate that new connections are more likely to occur in an evolving network, or that unobserved connections are missing in a partially known network[1-4]. Link prediction has many applications. For example, friends recommendation in social network, product recommendation in shopping website[5-7], pre-experimental analysis in protein interaction networks and metabolic networks[8]. In addition, link prediction can also reveal the structural growth or formation mechanism of complex networks[9].

Link prediction has been studied by scientists from different fields because of its wide application value and theoretical significance. Researchers have achieved great success in many similarity algorithms, especially topological similarity algorithms[10]. There are three kinds of topology similarity algorithms based on path length: local similarity algorithm, global similarity algorithm and semi-local similarity algorithm. The local similarity index considers the influence of the common neighbors of two disconnected nodes on the connection. For example, CN Index[11] considers the amount of common neighbors. Adamic-Adar (AA) Index[12] study the contribution of the degree of the common neighbor node to the connection; Resource Allocation Index[13] considers suppressing the common neighbors of large nodes. The global similarity algorithm considers the topology of the whole network, Katz Index[14] computes all paths among two disconnected points under the condition that short paths are given priority. Local Random Walk (LRW) Index[15] limits the random number in the semi-local range; Superposed Random Walk Index[15] superposed LRW contributions to paths of different lengths. Some researchers can reflect the influence of nodes better by using mixed indicators and improve the prediction results when applying them to SRW. For example, Zhu et al.[16] proposed SHI model and HHI model, using node degree and H-index to jointly reflect node influence. In the DCHI model and HCHI model proposed by Tian Yang et al.[17], degrees mixed with coreness and H-index mixed with coreness are respectively used to jointly reflect the influence of node. In the HIN model proposed by Gao et al.[18], the value of degree and H-index is used to reflect node influence.

So far, there is a lot of research on mixed indicators to express node influence, these studies have not fully explored the value of mixed indicators. Through in-depth study, based on the HIN model, we introduce a weight factor $\beta$ to balance the degree and H-index, so as to propose the DHIN model, which can reflect the node influence more accurately.

Figure 1 describes some of the attributes of the node, such as H-index and degree. The degree of node $a$ is 3, the degrees of the three neighboring nodes of node a are 3, 4, and 5 respectively, so the H-index is 3. When only study the degree of endpoints, the influence of endpoint $a$ is 3. When the influence of node a is multiplied by the degree and H-index, the

[*] tianhui@bupt.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250621
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661777

influence of node $a$ is 9. By contrast, the hybrid influence promotes the connection of the two nodes.

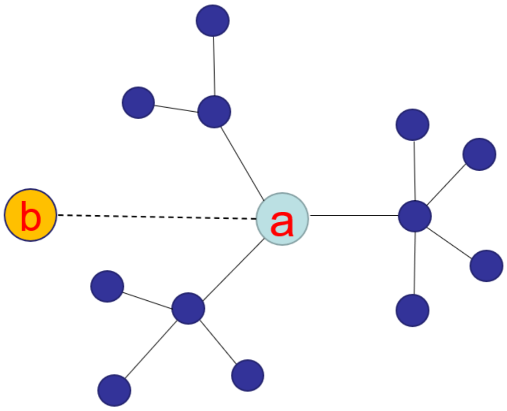

Figure 1. (Color online) Node influence graph based on H-index and degree.

Note: Endpoint $a$ represents a source node with H-index and degree are both 3. Node $b$ represents a destination node. All solid and dotted lines represent existing and potential links, respectively.

In this article, we propose a DHIN model. Through the depth investigation of twelve reference data sets and extensive experiments show that DHIN can improve predictive performance compared to these mainstream standards and HIN.

In Section 2, we establish a DHIN model. Sections 3 introduce twelve benchmark experimental data sets. Sections 4 show the methods. Section 5 shows the discussion and results. Section 6 summarizes this research.

## 2. MODEL

The model in this study is oriented to an undirected network $G(V,E)$, where $V$ and $E$ represents the group of points and connections, respectively. For each pair of nodes $x,y \subset V$, there is a fraction $s_{xy}$ that represents the probability of connection between them. In this paper, the value of similarity is directly used as the score of $s_{xy}$. After the score of all non-existent connection is arranged in descending order, the connection with the highest score is most likely to exist in the future or already exist but not detected.

### 2.1 HIN model

According to Lu et al.[15], Zhu et al.[18] introduced the mixed influence index as a metric into the SRW proposed by Lu et al.[15], then constructing the HIN model. Lu et al.[15] established a similarity index using a random walk. The probability of one-step transition of two endpoints X and Y through a random walk using a Markov chain is $p_{xy} = a_{xy}/k_x$, where $k_x$ represents the degree of $x$, $a_{xy} = 1$ when $x$ is connected to $y$, and $a_{xy} = 0$ if not. When the step size is $t$, the nodes $x$ and $y$ are represented as $\{x = x_0 = y_t, x_1 = y_{t-1}, \dots, x_{t-1} = y_1, x_t = y_0 = y\}$. Therefore, the t-step transition probability from node X to node Y is denoted as $\pi_{xy}(t) = \prod_{i=0}^{t-1} p_{x_i x_{i+1}}$and $\pi_{yx}(t) = \prod_{i=0}^{t-1} p_{y_i y_{i+1}}$. Then, Zhu et al. replaced the degree effect in SRW by combining the H-index and degree, so as to realize the mixed effect HIN of neighbor index.

$$s_{xy}^{HIN}(t) = \sum_{l=2}^{t} \left[ \frac{\sqrt{\overline{k_x} \times \overline{h_x}}}{2|E|} \pi_{xy}(l) + \frac{\sqrt{\overline{k_y} \times \overline{h_y}}}{2|E|} \pi_{yx}(l) \right] \tag{1}$$

Among them, $\overline{k_x} = \frac{1}{|\Gamma(x)|} \sum_{z \in \Gamma(x)} k_z$, $\overline{h_x} = \frac{1}{|\Gamma(x)|} \sum_{z \in \Gamma(x)} h_z$, node $y$ in the same way.

### 2.2 DHIN model

On the basis of HIN, we introduce the idea of weighting, and use the weight factor to balance the H-index and degree, so as to propose the DHIN model.

$$S_{xy}^{DHIN}(t) = \sum_{l=2}^{t} \left[ \frac{\sqrt{\overline{k_x}^{1-\beta} \times \overline{h_x}^{\beta}}}{2|E|} \pi_{xy}(l) + \frac{\sqrt{\overline{k_y}^{1-\beta} \times \overline{h_y}^{\beta}}}{2|E|} \pi_{yx}(l) \right] \tag{2}$$

Among them, $\overline{k_x} = \frac{1}{|\Gamma(x)|} \sum_{z \in \Gamma(x)} k_z$, $\overline{h_x} = \frac{1}{|\Gamma(x)|} \sum_{z \in \Gamma(x)} h_z$, node $y$ in the same way.

## 3. EXPERIMENTAL DATA

Our dataset of 12 real networks in this experiment：(1) US Air97(USAir)[19] uses the data from the United States airline network. (2) Yeast PPI(Yeast)[20] uses data from the yeast networks on protein-protein relationships. (3) Food Web (Food)[21] uses data from carbon exchange relationships. (4) Power Grid (Power)[22] uses data from the Electric transmission Network in the Western US. (5) NetScience (NS)[23] uses data based on the collaboration of scientists by paper in scientific networks. (6) Jazz[24] uses the data from the musicians network. (7) e-mail network (e-mail)[25] uses the data from the e-mail communication network of URV. (8) Slavko[26] uses the data from Slavko Zitnik's friendship network on social network. (9) UC social network (UCsocial)[27] uses the data from an online social network formed by UC Irvine students. (10) Infectious (Infec)[28] uses the data from a network of offline contacts made by visitors through an exhibition called "Infectious: Stay Away". (11) EuroSiS web (EuroSiS)[29] uses data from an interactive network of "social science" participants. (12) C. elegans (CE)[22] uses the data from the C. elegans worm neuron network. Table 1 gives the basic topology characteristics of the twelve networks.

In order to realize the preprocessing, the arc is changed into a directionless link. To ensure that networks have no authority and no direction, we need to eliminate loops and multilaterals. Then, on the premise of ensuring link connectivity, the maximum link is extracted to simplify the network subgraph.

Table 1. Seven fundamental topological characteristics of the 12 benchmark networks.

| **Nets** | $\|V\|$ | $\|E\|$ | $\langle k \rangle$ | $\langle d \rangle$ | $C$ | $r$ | $H$ |
|---|---|---|---|---|---|---|---|
| USAir | 332 | 2128 | 12.81 | 2.74 | 0.749 | -0.208 | 3.36 |
| Yeast | 2370 | 10904 | 9.2 | 5.16 | 0.378 | 0.469 | 3.35 |
| Food | 128 | 2075 | 32.42 | 1.78 | 0.334 | -0.112 | 1.24 |
| Power | 4941 | 6594 | 2.669 | 15.87 | 0.107 | 0.003 | 1.45 |
| NS | 1461 | 2742 | 3.75 | 5.82 | 0.878 | 0.461 | 1.85 |
| Jazz | 198 | 2742 | 27.7 | 2.24 | 0.633 | 0.02 | 1.4 |
| Email | 1133 | 5451 | 9.62 | 3.61 | 0.254 | 0.078 | 1.94 |
| Slavko | 334 | 2218 | 13.28 | 3.05 | 0.488 | 0.247 | 1.62 |
| Ucsocial | 1893 | 13825 | 14.62 | 3.06 | 0.138 | -0.188 | 3.81 |
| Infec | 410 | 2765 | 13.49 | 3.63 | 0.467 | 0.226 | 1.39 |
| EuroSiS | 1272 | 6454 | 10.15 | 3.86 | 0.382 | -0.012 | 2.46 |
| CE | 453 | 2025 | 8.94 | 2.66 | 0.655 | -0.225 | 4.49 |

Notes: $|V|$ describes the number of nodes, $|E|$ represents links, $\langle k \rangle$ represents the average degree, $\langle d \rangle$ denotes the average distance, $C$ denotes the clustering coefficient, $r$ describes the assortativity coefficient, $H$ defined as $H = \langle k^2 \rangle / \langle k \rangle^2$ , denotes the degree heterogeneity.

First, the network connection set is randomly divided into two parts, 90% of the training set $E^T$ and 10% of the test set

$E^P$,while the connectivity of $E^T$ is ensured[1]. In addition, there are 30 identical and independent branches on the network. Next, to achieve average accuracy in a statistical manner, we performed the experimental procedure on 30 separate training and test sets to, and recalled over 30 implementation measures.

## 4. EXPERIMENTAL METHODS

### 4.1 Metric

We use AUC[30] to measure the accuracy of algorithms. It can be interpreted as the probability that a potential link (a link in EP) score higher than a non-existent link (a link in U-E,). In algorithm implementation, we usually calculate a score for each unobserved link. Then, each time we randomly select a missing link and a non-existent link to compare their scores, if in $n$ independent comparisons, there are $n'$ missing links with higher scores and $n''$ missing links with the same scores, then the AUC value is

$$AUC = \frac{n' + 0.5n''}{n} \tag{3}$$

For purely probabilistic algorithms, the AUC tends to 0.5 when each result is independent. Therefore, as long as the accuracy of the algorithm is greater than 0.5, the performance of the algorithm is better than the pure probability algorithm.

### 4.2 Baselines

We describe the following six basic models:

CN Index[11] focusing the number of common neighbors, which is defined as

$$S_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)| \tag{4}$$

where $\Gamma(X), X \in \{x, y\}$, describes the set of neighbors of $X$ and $|\Gamma(x) \cap \Gamma(y)|$ represent the common neighbors between nodes $x$ and $y$.

AA Index[12], which is improved from CN, using the inverse logarithm to suppress the contribution of a large degree of common neighbors, it defined as

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(k_z)} \tag{5}$$

where $k_z$ describe the endpoint degree of $z$.

Resource-Allocation (RA)[13], originated from AA. The main idea is that if the degree of neighbor is larger, the possibility of nodes being connected is smaller, it defined as

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{6}$$

Local Path (LP) Index[15] takes into account the similarity between specific step-size paths(two and three step), which is defined as

$$S^{LP} = A^2 + \varepsilon A^3 \tag{7}$$

where A describes the adjacency matrix, $\varepsilon$ represents a penalty parameter.

SRW Index[15] limits the random number within the quasi-local range then superposed contributions to paths of different lengths, which is defined as

$$s_{xy}^{SRW}(t) = \sum_{l=2}^{t} \left[\frac{k_x}{2|E|}\pi_{xy}(l) + \frac{k_y}{2|E|}\pi_{yx}(l)\right] \tag{8}$$

The probability of one-step transition of two endpoints *X* and *Y* through a random walk using a Markov chain is $p_{xy} =$

$a_{xy}/k_x$, where $k_x$ describe the degree of node *X*, $a_{xy} = 1$ when *x* is connected to *Y*, and $a_{xy} = 0$ if not. When the step size is *t*, the nodes $x$ and $y$ are represented as $\{x = x_0 = y_t, x_1 = y_{t-1}, \dots, x_{t-1} = y_1, x_t = y_0 = y\}$. Therefore, the *t*-step transition probability from node X to node Y is denoted as $\pi_{xy}(t) = \prod_{i=0}^{t-1} p_{x_i x_{i+1}}$and $\pi_{yx}(t) = \prod_{i=0}^{t-1} p_{y_i y_{i+1}}$.

HIN[18] is showed in Section 2.

## 5. RESULTS

To examine the performance of our model, we ran simulations on twelve data sets and compared the resulting data with the primary baseline, and the results are discussed below.

According to the description in Sections 1 and 2, we believe that integrating the endpoints H-index and degree can better describe node influence and improve link prediction performance. To verify our thoughts, we propose a new model DHIN. First, we obtain the value of random walk $t$ corresponding to the maximum of the predicted results when there is no weight (excluding the influence of inhibiting factor $\beta$). The corresponding values of t are shown in parentheses after DHIN in Figure 2; DHIN gets the optimal AUC value with the least steps $t$, which is 3 in Figure 2a USAir, 3 in Figure 2b Yeast, 3 in Figure 2c Food, 14 in Figure 2d Power, 8 in Figure 2e NS, 2 in Figure 2f Jazz, 7 in Figure 2g Email, 4 in Figure 2h Slavko, 9 in Figure 2i UCsocial, 4 in Figure 2j Infec, 5 in Figure 2k Eurosis and 4 in Figure 2l CE. Then we set the weighting factor $\beta$ from 0.1 to 0.9 at 0.1 intervals.

In Figure 2, DHIN shows its optimal values of AUC at certain inhibitory factor of $\beta \in [0,1)$ on random walk steps $t$ in the different datasets, i.e., $\beta = 0.3$ in Figure 2a USAir, $\beta = 0.7$ in Figure 2b Yeast, $\beta = 0.9$ in Figure 2c Food, $\beta = 0.8$ in Figure 2d Power, $\beta = 0.4$ in Figure 2e NS, $\beta = 0.6$ in Figure 2f Jazz, $\beta = 0.6$ in Figure 2g Email , $\beta = 0.8$ in Figure 2h Slavko, $\beta = 0.1$ in figure 2i UCsocial, $\beta = 0.4$ in figure 2j Infec , $\beta = 0.7$ in Figure 2k Eurosis and $\beta = 0.1$ in Figure 2l CE.

DHIN introduced the idea of weight into hybrid influence index based on HIN. We study the relationship between weights $\beta$ and AUC of two endpoints with typical predictive length *L*=100, under the condition that the maximum random walk step $t$ ($t$ is from 1 to 15) of each data set obtained by HIN remains unchanged. The optimal values are calculated by the parameter $\beta$ from 0.1 to 0.9.

The detailed data in Figure 2 are shown in Table 2 and compares the AUC on WHIN with six models. DHIN obtained the optimal value on seven data sets, four for SRW (Power, Slavko, Ucsocial, EuroSiS) and one for LP (Food). By comparison, DHIN improved predictive performance.

Table 2. AUC of seven models on twelve benchmark network under the condition of $L = 100$.

| AUC | CN | AA | RA | LP | SRW | HIN | DHIN |
|---|---|---|---|---|---|---|---|
| USAir | 0.977781 | 0.984263 | 0.986586 | 0.973134 | 0.989705 (3) | 0.989851 (3) | 0.99002 (3, 0.3) |
| Yeast | 0.736897 | 0.737023 | 0.737053 | 0.743085 | 0.744265 (3) | 0.744233 (3) | 0.744292 (3, 0.7) |
| Food | 0.616496 | 0.617357 | 0.619666 | 0.827907 | 0.77069 (3) | 0.769437 (3) | 0.769316 (3, 0.9) |
| Power | 0.679672 | 0.679644 | 0.679649 | 0.764137 | 0.949493 (14) | 0.949299 (14) | 0.94927 (14, 0.8) |
| NS | 0.990236 | 0.990331 | 0.990353 | 0.994177 | 0.995597 (8) | 0.995593 (8) | 0.9956 (8, 0.4) |
| Jazz | 0.972277 | 0.976377 | 0.981308 | 0.947589 | 0.981307 (2) | 0.982827 (2) | 0.983208 (2, 0.6) |
| Email | 0.881974 | 0.883095 | 0.8824 | 0.945157 | 0.956093 (6) | 0.956229 (7) | 0.95646 (7, 0.6) |
| Slavko | 0.964003 | 0.965942 | 0.965657 | 0.965124 | 0.971646 (4) | 0.971492 (4) | 0.971568 (4, 0.8) |
| Ucsocial | 0.813189 | 0.817415 | 0.817553 | 0.948598 | 0.950135 (7) | 0.948185 (9) | 0.949157 (9, 0.1) |
| Infec | 0.962356 | 0.964272 | 0.964238 | 0.975218 | 0.980498 (4) | 0.980572 (4) | 0.980811 (4, 0.4) |
| EuroSiS | 0.955322 | 0.956548 | 0.956051 | 0.980574 | 0.985396(5) | 0.985197 (5) | 0.985201 (5, 0.7) |
| CE | 0.951563 | 0.977061 | 0.979053 | 0.9558 | 0.985181 (3) | 0.985212 (4) | 0.985287 (4, 0.1) |

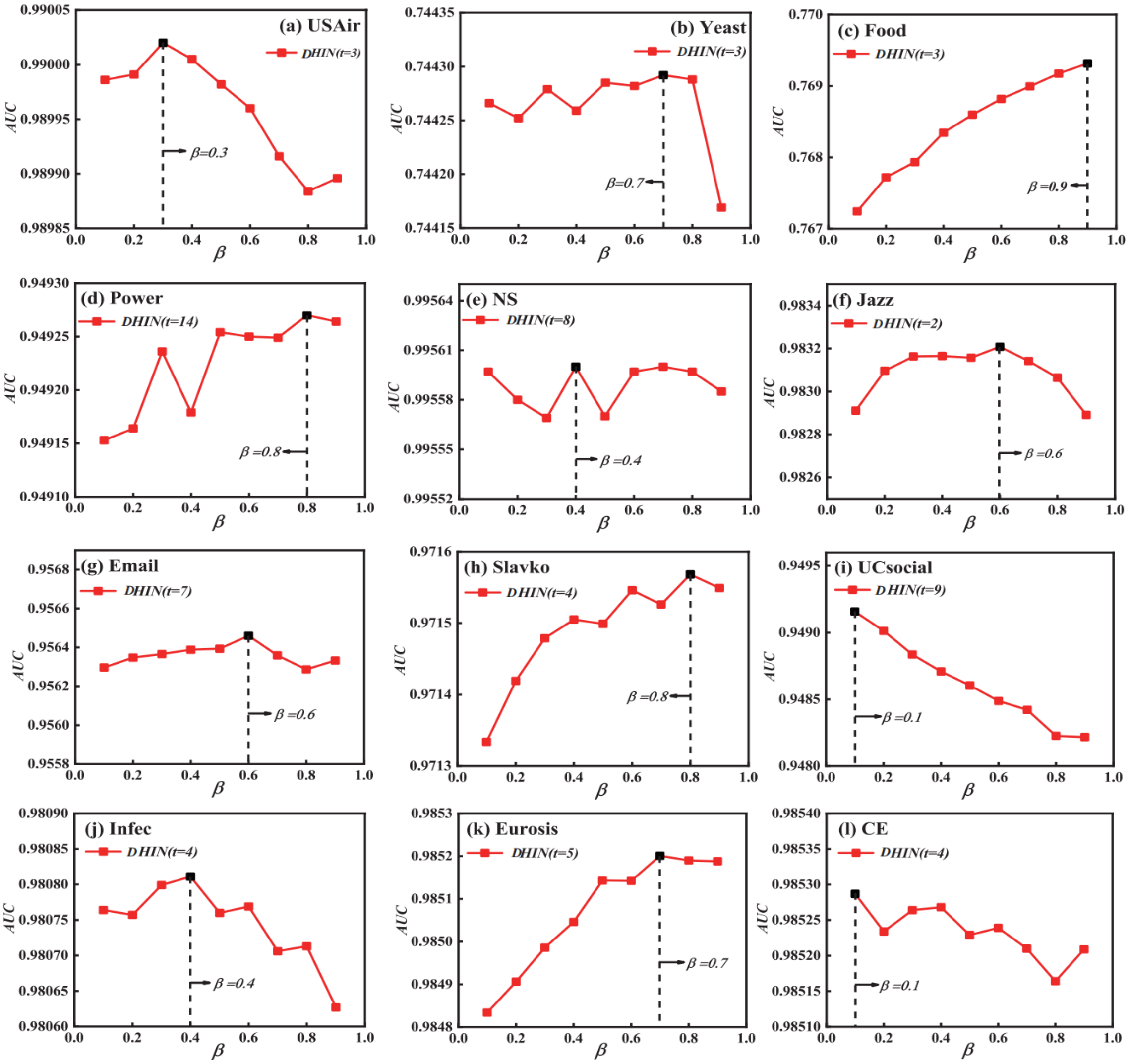

Figure 2. (Color line) AUC of the DHIN (red square) and the random walk step $t$ (black square).

In Table 2, the value of $L$ represents the number of candidate links. Each data point is an average of over 30 separate implementation processes, and each point represents a random partition of 90-10% of the training and test sets. The maximum values are bold. On SRW, HIN and DHIN the first value in the parentheses represent the corresponding optimal random walk step $t$. On DHIN the second value is optimal weight $\beta$. It can be concluded from Table 2 that DHIN's prediction results are better than other models in more than half of the data sets. All results represent the optimal situation by adjusting the coefficient.

In addition, computational complexity needs to be considered on describe link prediction performance. The time complexity of CN, AA, RA have $O(N3)$ and LP, SRW, HIN, DHIN have $W \times O(N3)$ with coefficient $W$. To sum up, DHIN showed a significant improvement with remain the time complexity.

## 6. CONCLUSIONS

The value of the mixing coefficient has not been fully explored in the existing link prediction studies using node hybrid influence. Through analysis, we propose a new DHIN index focusing on DHIN. By comparing the test results on twelve datasets with six indices, we investigate the utility of the mixed effects of tenure weight factors in link prediction. Therefore, the accuracy of DHIN proposed in this paper is obviously superior to other indexes.

# REFERENCES

[1] Lü, L. and Zhou, T., Physica A: Statistical Mechanics and Its Applications, 390(6), 1150-70(2011).

[2] Lü, L., Medo, M., Yeung, C. H., Zhang, Y. C., Zhang, Z. K. and Zhou, T., Physics Reports, 519(1), 1-49(2012).

[3] Chi, K., Yin, G., Dong, Y. and Dong, H., Knowledge-Based Systems, 181, 104792(2019).

[4] Wang, Y., Wang, Y., Lin, X. and Wang, W., Discrete Dynamics in Nature and Society, 6148273(2020).

[5] Wang, W., Liu, Q. H., Liang, J., Hu, Y. and Zhou, T., Physics Reports, 820, 1-51(2019).

[6] Gurini, D. F., Gasparetti, F., Micarelli, A. and Sansonetti, G., Future Generation Computer Systems, 78, 430-39(2018).

[7] Xiong, F., Wang X, Pan S, Yang H, Wang, H. and Zhang, C., IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50(10), 3804-16(2018).

[8] Battiston, F., et al., Physics Reports, 874, 1-92(2020).

[9] Zhu, B. and Xia, Y., PloS one, 11(2), e0148265(2016).

[10] Lü, L., Pan, L., Zhou, T., Zhang, Y. C. and Stanley, H. E., Proceedings of the National Academy of Sciences, 112(8), 2325-30(2015).

[11] Newman, M. E., Physical Review E, 64(2), 025102(2001).

[12] Adamic, L. A. and Adar, E., Social Networks, 25(3), 211-30(2003).

[13] Zhou, T., Lü, L. and Zhang, Y. C., The European Physical Journal B, 71(4), 623-30(2009).

[14] Katz, L., Psychometrika, 18(1), 39-43(1953).

[15] Liu, W. and Lü, L., EPL (Europhysics Letters), 89(5), 58007(2010).

[16] Zhu, X., Li, W., Tian, H. and Cai, S., EPL (Europhysics Letters), 122(6), 68003(2018).

[17] Tian, Y., Wang, Y., Tian, H. and Cui, Q., Complexity, 1544912(2021).

[18] Gao, T. and Zhu. X., International Journal of Modern Physics B, 34(05), 2050018(2020).

[19] Batagelj, V. and Mrvar, A., Connections, 21(2), 47-57(1998).

[20] Bu D, et al., Nucleic Acids Research, 31(9), 2443-50(2003).

[21] Melián, C. J. and Bascompte, J., Ecology, 85(2), 352-358(2004).

[22] Yan, G., Zhou, T., Hu, B., Fu, Z. Q. and Wang, B. H. Physical Review E, 73(4), 046108(2006).

[23] Holme, P. and Newman, M. E., Physical Review E, 74(5), 056108(2006).

[24] Gleiser, P. M. and Danon, L., Advances in Complex Systems, 6(04), 565-73(2003).

[25] Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. and Arenas, A., Physical Review E, 68(6), 065103(2003).

[26] Opsahl, T. and Panzarasa, P., Social Networks, 31(2), 155-63(2009).

[27] Blagus, N., Šubelj, L. and Bajec, M., Physica A: Statistical Mechanics and Its Applications, 391(8), 2794-802(2012).

[28] Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J. F. and Van den Broeck, W., Journal of Theoretical Biology, 271(1), 166-80( 2011).

[29] Ermiş, B., Acar, E. and Cemgil, A. T., Data Mining and Knowledge Discovery, 29(1), 203-36(2015).

[30] Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E. and Makse, H. A., Nature Physics, 6(11), 888-93(2010).