Physica A 410 (2014) 439–449

Contents lists available at ScienceDirect

Physica A

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

Information loss method to measure node similarity in networks

Yongli Li [a,b], Peng Luo [a,*], Chong Wu [a]

[a] School of Management, Harbin Institute of Technology, Harbin 150001, PR China
[b] Dipartimento Di Economia Politica E Statistica, Università Di Siena, Siena 53100, Italy

HIGHLIGHTS

- Our method defines the entropy-based information loss to measure node similarity.
- Two nodes are more similar if less is the information loss of seeing them as the same.
- The new method has the algorithm complexity $O(n^2)$.
- The method performs well based on artificial examples and synthetic networks.
- The method can be applied to predict network's evolution and nodes' attributions.

ARTICLE INFO

Article history:
Received 14 January 2014
Received in revised form 6 April 2014
Available online 23 May 2014

Keywords:
Node similarity
Information theory
Information loss
Complex network
Prediction
Statistical physics

ABSTRACT

Similarity measurement for the network node has been paid increasing attention in the field of statistical physics. In this paper, we propose an entropy-based information loss method to measure the node similarity. The whole model is established based on this idea that less information loss is caused by seeing two more similar nodes as the same. The proposed new method has relatively low algorithm complexity, making it less time-consuming and more efficient to deal with the large scale real-world network. In order to clarify its availability and accuracy, this new approach was compared with some other selected approaches on two artificial examples and synthetic networks. Furthermore, the proposed method is also successfully applied to predict the network evolution and predict the unknown nodes' attributions in the two application examples.

1. Introduction

The network model is a powerful tool to analyze the relationships. Some nodes in networks usually appear to be similar, which leads to this paper's research topic. For example, the nodes can represent people and the links can represent friendship between individuals in the context of social network [1–3]; if two people have similar interests, backgrounds, or friends, they may be similar to each other. Accordingly, mining the similar nodes can be economical and applicable to make friend recommendation [4–6], link prediction [7,8], peer-effect analysis [9,10], and so forth. Therefore, precise assessment of the similarity of nodes in networks seems to be very necessary and useful in the modern world, and measurement of node similarity is also one of the key issues in the research field of networks. From the perspective of application, an applicable method for measurement of network node similarity could capture some important aspects of various applications, such as

* Corresponding author. Tel.: +86 15804519346.
E-mail addresses: yongli.0440004@gmail.com (Y. Li), luopeng_hit@126.com (P. Luo).

http://dx.doi.org/10.1016/j.physa.2014.05.056
0378-4371/© 2014 Elsevier B.V. All rights reserved.

in the field of physics [11,12], transport [13], biology [14], bibliometrics [15]. Thus, a good and stylized method for network node similarity measurement is needed not only in the field of statistical physics but also in large scopes of other applications.

The proposed method has two distinctive features compared to the other existing methods. First, our complementary approach is based on the information theory [16,17], which gives an information measure for a network and focuses on the information loss when two nodes are seen as the same. Second, the method proposed in this study has the algorithm time complexity $O(n^2)$ (n is the node number of a given network) which makes it possible to be applied in the large scale networks which are often faced in the real-world problem. To sum up, the new method has the advantages of higher reasonability and lower algorithm complexity.

To demonstrate the method's reasonability and present its calculation process clearly, we organize this paper as follows: in the following Section 2, the related work is reviewed in brief; in Section 3, the model and its algorithm is explained and showed in detail with mathematical proof and algorithm analysis; in Section 4, an artificial example and two simulation-based tests are given to uncover our method's properties and compare it with the selected existing methods; in Section 5, two applications are given to illustrate the method's applicability, and the last Section 6 concludes.

2. Related work

The most common approach adopted in the previous work is to count how many neighbors two nodes have in common. It is in the sense that nodes are similar to the extent that their neighborhoods overlap. Let N_i be the neighborhood number of vertex i in a network, i.e., the set of nodes that are directly connected to i via an edge. Then the similarity measure for two nodes i and j is

$$\sigma_1(i,j) = \left|N_i \cap N_j\right|. \tag{1}$$

However, there is a drawback of this measure that the nodes with high degree are favored to be more similar than the low-degree nodes, because the high-degree vertices would have a large value even if only a small fraction of their neighbors are shared. Therefore, this method is not entirely satisfactory.

There are many ways proposed to overcome such a problem. One of these is to normalize the number of shared nodes based on the size of its two neighborhoods' unions:

$$\sigma_2(i,j) = \frac{\left|N_i \cap N_j\right|}{\left|N_i \cup N_j\right|}. \tag{2}$$

This is commonly called the *Jaccard index* which was proposed in Ref. [18]. Then, the *cosine similarity* was proposed by Salton and was widely used in the literature on citation networks [19]. It is the cosine of the angle between the characteristic vectors of the two neighborhoods. We left out the vertices i and j when counting the size of their neighborhoods because of a better measure on loop-less graphs.

$$\sigma_3(i,j) = \frac{\left|N_i \cap N_j\right|}{\sqrt{|N_i|\left|N_j\right|}}. \tag{3}$$

Besides, there are many other ways of improving the common similarity measure (1), like Ravasz et al. [20], Burt [21], and Goldberg and Roth [22] as following:

$$\sigma_4(i,j) = \frac{\left|N_i \cap N_j\right|}{\min\left(|N_i|, \left|N_j\right|\right)}, \tag{4}$$

$$\sigma_5(i,j) = \sqrt{\left|N_i \cap N_j\right|}, \tag{5}$$

$$\sigma_6(i,j) = \left(\left|N_i \cap N_j\right|\right)^2. \tag{6}$$

On the other hand, many researchers are working to propose new node similarity measure methods in other ways. Symeonidis et al. [23] defined a new way to calculate the similarity between vertices on the basis of the Tanomoto coefficient [24]. First, they define the similarity measure as follows:

$$\sigma_7(i,j) = \frac{\left|N_i \cap N_j\right|}{|N_i| + \left|N_j\right| - \left|N_i \cap N_j\right|}. \tag{7}$$

However, it is not reasonable enough since the similarity values between all non-neighbor nodes are zero based on formula (7). Then, they define a transitive node similarity which is calculated by the product of basic similarity between the nodes appearing in the shortest path. As a result, they get the following method:

$$\sigma_8(i,j) = \begin{cases} 0, \quad \text{if there is no path between the two nodes;} \\ \sigma_7(i,j), \quad \text{if } i,\ j \text{ are neighbors;} \\ \prod_{k=1}^{t} \sigma_7(v_k, v_{k+1}), \quad \text{otherwise} \end{cases} \tag{8}$$

where $v_1 = i,\ \ v_{t+1} = j$ and the nodes $v_k\ (k = 2, \ldots, t)$ are all the intermediate nodes that the shortest path from i to j passes through. They also use the transitive node similarity to predict the link in social networks.

Recently, Chen et al. [25] introduced another new vertex similarity measure called *relation strength similarity* (RSS) which could better capture the potential relationships of a real-world network structure. First, they define the *relation strength*, a normalized edge weight score reflecting the relative degree of similarity between neighbor vertices. The relation strength between two nodes i and j can be calculated as follows:

$$\sigma_9(i,j) = \begin{cases} \dfrac{\alpha_{ij}}{\sum\limits_{x\in n_i} \alpha_{ix}}, & \text{if } i \text{ and } j \text{ are adjacent,} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where α_{ij} is the weight of the edge between nodes i and j, and n_i is the set of neighbor vertices of node i. Second, if nodes i and j are not adjacent and they are linked by a path, the *generalized relation strength* is defined as

$$\sigma_{10}(i,j) = \begin{cases} \prod\limits_{k=1}^{t} \sigma_9(v_k, v_{k+1}), & \text{if } t < r, \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $v_1 = i,\ \ v_{t+1} = j$ and the nodes $v_k\ (k = 2, \ldots, t)$ are all the intermediate nodes that form the path. The r is a discovery range parameter to control the maximum degree of separation for a generalized relation strength calculation. For example, if assumed that there are M simple paths from i to j with path length shorter than r, then the relation strength similarity between i and j is calculated by summing all the generalized relation strengths:

$$\sigma_{11}(i,j) = \sum_{m=1}^{M} \sigma_{10}(i,j). \tag{11}$$

Moreover, there are also many other researchers who make contributions to the node similarity. A stochastic approach for determining the similarity of vertices was presented by Nowicki and Snijders [26]. Also, Leicht et al. [2] considered that two nodes are similar if their neighbors in the network are similar. Accordingly, they constructed a method for quantifying the similarity of nodes in networks based on this idea. Furthermore, Penner et al. [14] proposed a biologically motivated quantity, twinness, to evaluate the local similarity of network nodes. Besides, Thiel and Berthold [27] took advantage of the spreading activation and gave two different kinds of node similarity measures.

3. Model and algorithm

The proposed model is established based on the idea that "If two nodes are more similar than the others, then the information loss of seeing them as the same is less than that of the others". In this section, we first present how to use a probability function to describe an undirected network, and then give the definition of "information loss" to measure the loss of seeing two different nodes as the same from the viewpoint of information theory.

Considering an undirected network $\mathbf{G}$ with n nodes, its links can be described by the symmetric matrix $\mathbf{A} = [a_{ij}]_{n\times n}$, where

$$a_{ij} = \begin{cases} \text{weight between nodes } i \text{ and } j, & \text{when } i \neq j \\ \sum\limits_{i=1}^{j-1} a_{ij} + \sum\limits_{i=j+1}^{n} a_{ij}, & \text{when } i = j. \end{cases}$$

Especially, as for an unweighted network, when $i \neq j$,

$$a_{ij} = \begin{cases} 1, & \text{if there exists a link between nodes } i \text{ and } j \\ 0, & \text{otherwise;} \end{cases}$$

when $i = j$,

$$a_{ij} = \sum_{i=1}^{j-1} a_{ij} + \sum_{i=j+1}^{n} a_{ij};$$

here, a_{ij} is the degree of the node i.

Then, the joint probability density function of nodes i and j can be defined as

$$p(i,j) = \frac{a_{ij}}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} a_{ij}}. \tag{12}$$

The above given $p(i,j)$ is the description of an undirected network from the angle of probability function. Also, the $p(i,j)$ can be checked to satisfy the definition of a probability function. Besides, we have $p(i) = \sum_{j=1}^{n} p(i,j)$, and so is the $p(j)$.

The definition of "Mutual information" in the field of information theory can be applied to measure how much information is contained in a complex network. The information of a given network with n nodes denoted by $I(\mathbf{N};\mathbf{N})$ is defined as

$$I(\mathbf{N};\mathbf{N}) = \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}} p(i,j)\cdot\log\frac{p(i,j)}{p(i)\cdot p(j)}, \tag{13}$$

where $\mathbf{N}$ is the set of nodes. When two nodes (for example n_1 and n_2) are seen as the same, the new node set denoted by $\mathbf{N}^*$ is

$$\mathbf{N}^* = \mathbf{N} - \{n_1\} - \{n_2\} + \{\langle n_1, n_2\rangle\}, \tag{14}$$

which has $n-1$ nodes. Then, the mutual information $I(\mathbf{N};\mathbf{N}^*)$ between the original node set ($\mathbf{N}$) and the new node set ($\mathbf{N}^*$) is

$$\begin{aligned} I(\mathbf{N};\mathbf{N}^*) &= \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}^*} p(i,j)\cdot\log\frac{p(i,j)}{p(i)\cdot p(j)} \\ &= I(\mathbf{N};\mathbf{N}) - \sum_{i\in\mathbf{N}}\sum_{k=1}^{2} p(i,n_k)\cdot\log\frac{p(i,n_k)}{p(i)\cdot p(n_k)} + \sum_{i\in\mathbf{N}} p(i,\langle n_1,n_2\rangle)\cdot\log\frac{p(i,\langle n_1,n_2\rangle)}{p(i)\cdot p(\langle n_1,n_2\rangle)}. \end{aligned} \tag{15}$$

The difference between $I(\mathbf{N};\mathbf{N}^*)$ and $I(\mathbf{N};\mathbf{N})$ uncovers the information loss when the two nodes n_1 and n_2 are seen as the same. That is to say, the information loss is the cost of taking the two nodes undifferentiated, which means the less the information loss is, the closer or the more similar the two nodes are. In detail, the *information loss* denoted by $\Delta I(n_1,n_2)$ is

$$\begin{aligned} \Delta I(n_1,n_2) &= I(\mathbf{N};\mathbf{N}) - I(\mathbf{N};\mathbf{N}^*) \\ &= \sum_{i\in\mathbf{N}}\sum_{k=1}^{2} p(i,n_k)\cdot\log\frac{p(i,n_k)}{p(i)\cdot p(n_k)} - \sum_{i\in\mathbf{N}} p(i,\langle n_1,n_2\rangle)\cdot\log\frac{p(i,\langle n_1,n_2\rangle)}{p(i)\cdot p(\langle n_1,n_2\rangle)}. \end{aligned} \tag{16}$$

Because the two nodes n_1 and n_2 are seen as the same, namely they are merged into one node, it holds that

$$p(i,\langle n_1,n_2\rangle) = p(i,n_1) + p(i,n_2), \tag{17}$$

$$p(\langle n_1,n_2\rangle) = p(n_1) + p(n_2), \tag{18}$$

$$p(i|\langle n_1,n_2\rangle) = \frac{p(i,n_1)+p(i,n_2)}{p(n_1)+p(n_2)}. \tag{19}$$

Accordingly, $\Delta I(n_1,n_2)$ can be calculated to be of the following form, where it is noted that when $p(k|n_k) = 0$, we let $p(k|n_k)\cdot\log p(k|n_k) = 0$ $(k = 1, 2)$:

$$\Delta I(n_1,n_2) = p(n_1)\cdot\sum_{k\in\mathbf{N}} p(k|n_1)\cdot\log\frac{p(k|n_1)}{p(k|\langle n_1,n_2\rangle)} + p(n_2)\cdot\sum_{k\in\mathbf{N}} p(k|n_2)\cdot\log\frac{p(k|n_2)}{p(k|\langle n_1,n_2\rangle)}. \tag{20}$$

Besides, $\Delta I(n_1,n_2)$ must be no less than zero which is given and proved in Property 1. The property is consistent with the nature rule that the information loss of regarding two different nodes as the same should be above or at least equal to 0.

Property 1. $\Delta I(n_1,n_2) \geq 0$.

Proof. Based on that $x\log x$ is the convex function, when $\beta_1+\beta_2 = 1$, it holds that

$$\beta_1\cdot a\log a + \beta_2\cdot b\log b \geq (\beta_1 a + \beta_2 b)\cdot\log(\beta_1 a + \beta_2 b).$$

When $p(n_1)+p(n_2) \neq 0$, let $\beta_1 = p(n_1)/(p(n_1)+p(n_2))$, $\beta_2 = p(n_2)/(p(n_1)+p(n_2))$, $a = p(k|n_1)$ and $b = p(k|n_2)$. As a result, for any $k\in\mathbf{N}$, the above inequality is

$$\frac{p(n_1)}{p(n_1)+p(n_2)}\cdot p(k|n_1)\cdot\log\frac{p(k|n_1)}{p(k|\langle n_1,n_2\rangle)}\cdot + \frac{p(n_2)}{p(n_1)+p(n_2)}\cdot p(k|n_2)\cdot\log\frac{p(k|n_2)}{p(k|\langle n_1,n_2\rangle)} \geq 0,$$

which means $\Delta I(n_1,n_2) \geq 0$ when summing the above formula based on all the $k\in\mathbf{N}$.

When $p(n_1)+p(n_2) = 0$, we have $p(n_1) = 0$ and $p(n_2) = 0$. Then, the two nodes n_1 and n_2 are isolated nodes which are not necessary to use the above model to analyze. Even so, in this case, it holds that $\Delta I(n_1,n_2) = 0$. □

In order to present the whole calculation process clearly, we summarize the corresponding algorithm in the following Table 1. The total time complexity can be estimated as $O(n^2)$ from the following algorithm, where n is the number of nodes in the given network $\mathbf{G}$. While, the time complexity of Symeonidis' method [23] and Chen's method [25] are both $O(n^3)$ which is higher than our method. In this viewpoint, our method is more suitable for a large network than the other two methods.

To make the expression clear, we further define $1/\Delta I(i,j)$ as the node similarity measure between two nodes i and j based on the above *information loss* proposed in this paper. The definition is consistent with the common definition of network node similarity measure, for which the larger value means higher similarity between the corresponding two nodes.

Table 1
Algorithm.

```
Input: the given undirected network G (n nodes and its relationship matrix A);
Output: the information loss of any two nodes in the given network G;
Initialization:
        (1) delete the isolated nodes;
        (2) calculate $p(i, j)$ of the given network G;
        (3) calculate $p(i)$ and $p(j)$;
        (4) $a \leftarrow 0$ and $b \leftarrow 0$.
Loop:
        for $i = 1$ to $n$
          for $j = i + 1$ to $n$
            for $k = 1$ to $n$
              $p(k|i) = p(k, i)/p(i)$;
              $p(k|j) = p(k, j)/p(j)$;
              $p(k|\langle i, j\rangle) = (p(k, i) + p(k, j))/(p(i) + p(j))$;
               if $p(k|i) \neq 0$
                 then $a \leftarrow a + p(k|i) \cdot (\log p(k|i) - \log p(k|\langle i, j\rangle))$;
               else
                 CONTINUE;
               end
               if $p(k|n_2) \neq 0$
                 then $b \leftarrow b + p(k|j) \cdot (\log p(k|j) - \log p(k|\langle i, j\rangle))$;
               else
                 CONTINUE;
               end
            end
            $\Delta I(i, j) = p(i) \cdot a + p(j) \cdot b$;
          end
        end
print: the information loss $\Delta I(i, j)$ of any two nodes $(i, j)$.
```

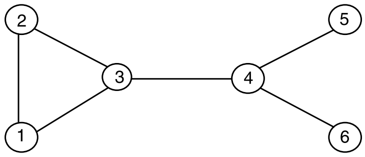

Fig. 1. The artificial network without weight.

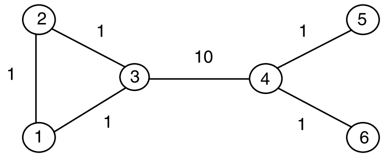

Fig. 2. The artificial network with weights.

4. Examples

4.1. Simple artificial examples

Our method can be used for both networks with or without weights. However, except Chen's method which can be used for both cases, the other methods introduced in the part of related work can only be applicable in the case of networks without weight. Here, we design two artificial networks and compare our method (ILM, for short) with the cosine similarity method (CSM) shown by formula (3), Symeonidis's method (SYM) shown by formulas (7)–(8) and Chen's method (CHM) showed by formulas (9)–(11).

The two simple artificial networks are shown in Figs. 1 and 2 respectively, where the weights in Fig. 2 reflect relation forces between nodes. The so-called relation force has many manifestations in real life, for example, the number of common hobbies between friends, the times of collaborations between authors, the degree of finance flows between companies, and so forth.

Intuitively, the node similarity measure is not only dependent on the nodes' positions but also the relation forces between these nodes. First, we give three definitions satisfying some intuitive properties that a similarity measure should have. The three definitions are the position's effect, the relationship force's direct effect and the relationship force's indirect effect. Also, their rationality can be easily understood based on our common judgment.

Table 2
The artificial example's results.

Figure	Fig. 1				Fig. 2	
Node pair	ILM	CSM	SYM	CHM	ILM	CHM
(**1**, **2**)	**70.6315**	**1.000**	**0.3333**	0.6667	**176.5848**	0.5417
(1, 3)	20.2544	0.7071	0.2500	**0.7500**	12.8702	0.7500
(1, 4)	4.4912	0.4082	0.0000	0.2500	7.3780	0.6250
(**1**, **5**)	6.2842	0.0000	0.0000	0.0833	15.7107	0.0521
(**1**, **6**)	6.2842	0.0000	0.0000	0.0833	15.7107	0.0521
(2, 3)	20.2544	0.7071	0.2500	**0.7500**	12.8702	0.7500
(2, 4)	4.4912	0.4082	0.0000	0.2500	7.3780	0.6250
(**2**, **5**)	6.2842	0.0000	0.0000	0.0833	15.7107	0.0521
(**2**, **6**)	6.2842	0.0000	0.0000	0.0833	15.7107	0.0521
(**3**, **4**)	6.2842	0.0000	0.0000	0.3333	20.3070	**0.8333**
(3, 5)	7.7111	0.0000	0.0000	0.1111	16.2167	0.0694
(3, 6)	7.7111	0.0000	0.0000	0.1111	16.2167	0.0694
(4, 5)	27.8087	0.0000	0.0000	0.3333	28.0481	0.0833
(4, 6)	27.8087	0.0000	0.0000	0.3333	28.0481	0.0833
(5, 6)	17.3124	0.0000	0.0000	0.3333	43.2807	0.0833

Definition 1 (*The Position's Effect*)**.** If two nodes have more shared neighbor nodes and less different neighbor nodes, the two nodes should be more similar.

Definition 2 (*The Relationship Force's Direct Effect*)**.** If the relationship force between two nodes increases, the two nodes become more similar.

Definition 3 (*The Relationship Force's Indirect Effect*)**.** If the relationship force between two nodes increases, the similarity between the two nodes' neighbor nodes should also increase.

Based on Definitions 1 and 2, node 1 and node 2 should be the most similar node pair in Fig. 1, because the two nodes have the most shared neighbor nodes and the least different neighbor nodes, and they also have larger relation force than that of node pair (5, 6) which also satisfies Definition 1. Besides, according to Definition 2, the similarity measure value between node 3 and node 4 in Fig. 2 should be higher than that in Fig. 1. Furthermore, deduced from Definition 3, similarity measure values of nodes pairs (1, 5), (1, 6), (2, 5) and (2, 6) in Fig. 2 should be larger than their counterparts in Fig. 1. If a similarity measure method is a good one, it should be consistent with the above intuitive judgment. Let us compare the similarity measure values resulting from the four listed methods (see Table 2).

The results demonstrate that only our method's result completely agrees with the three definitions. First, CSM and SYM can only deal with the unweighted case, and their results are not elaborate, or in other words, undistinguished, since many node pairs' similarity measure values are zero. Second, the result of CHM violated the first and the third definition, which means that its result is not consistent with the intuitive judgment. Thus, in this artificial example, our method is prior to the other three to some extent.

4.2. Comparisons based on simulation analysis

We continue to test our method on a series of networks generated by computers. Here, two kinds of simulation analyses were designed. The first one considers the community structure as the benchmark for deciding nodes' similarity; namely, the node pair's linkage is generated with a higher probability within the same community than between different communities. The second one adopts a reverse engineering approach, in which nodes are created in a multi-dimensional environment of characteristics and then are linked with some rule that assigns more weight values for these nodes that are closer in the characteristics space. The first test is used for unweighted networks and the second is for weighted networks. We will test our method's accuracy in the two kinds of simulation tests with the aim to achieve comprehensive results.

4.2.1. Simulation analysis based on a given community structure

As for generating the computer-based networks, we considered the *Community Structure* to be the benchmark for deciding which nodes are similar and which ones are not. That is to say, if two nodes were in one community, we regarded them to be similar nodes, and vice versa. Based on this idea, we followed Girvan and Newman's way [28] to produce randomly artificial networks that have the known community structure, which has been a famous and widely used simulation method for testing [29]. Each generated graph is constructed with 128 nodes divided into 4 communities of 32 nodes each. Edges are placed between node pairs independently at random, with the probability P_{in} for nodes belonging to the same community and P_{out} for nodes in different communities. When the P_{in} is given, the P_{out} can be calculated to keep the average degree of nodes as 28, since keeping the same average degree of nodes can make different networks comparable. Then, we denote the computer-generated graphs as $RN(4, 32, 28, P_{\text{in}})$ where P_{in} varies from 0.60 to 0.95 by 0.05 in each step, and we can get 100 artificial networks for each P_{in} and average the results to make the comparison with SYM and CHM mentioned above.

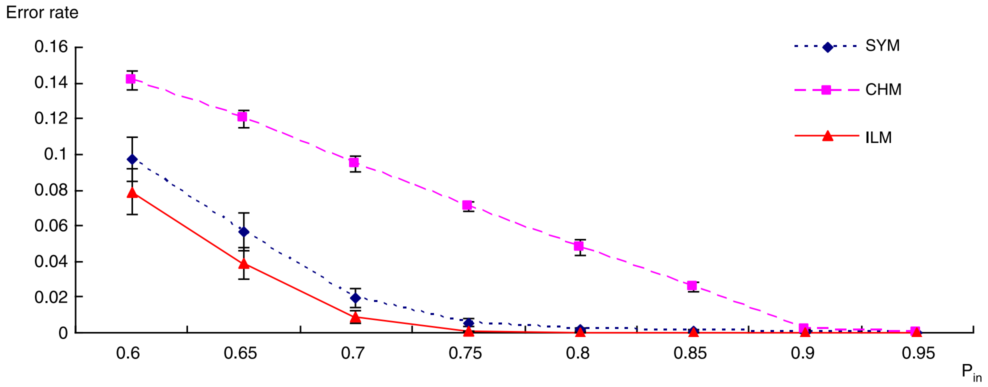

Fig. 3. The accuracy and its standard deviation of three compared methods.

As for the index, we have known nodes within one community similar to each other; thus each node has 31 similar nodes in the network. We can rank all these other nodes based on their similarity measures with the given one and compare the first 31 nodes with our known similar nodes of the given one. Let us denote the number of wrong similar nodes to node i as N_i, then we can get the index N by summing all these N_i according to i. Of course, the smaller the N is, the better the corresponding result is. Accordingly, the *error rate* can be defined as $N/(128 \times 31)$. Based on these 100 generated artificial networks for each P_{in}, we can get a comprehensive index value by averaging and its corresponding standard deviation. The whole result is shown in the following Fig. 3.

Fig. 3 illustrates that when $P_{\text{in}} \geq 0.75$, our method performs nearly perfectly with classifying nearly 100% nodes correctly. With P_{in} varying from 0.6 to 0.75, the error rate is below 0.08 which is much less than those of CHM and SYM. The result indicates that our approach is much more efficient in terms of finding the similarity nodes. Based on the above tests and comparisons, we can conclude that our approach is efficient in searching for similar nodes, and in these experiments, it performs better than the listed approaches in terms of accuracy. Although these artificial networks in the above experiments cannot represent all the real-world networks, they can prove the availability and superiority of our approach to some extent.

4.2.2. Simulation analysis based on given node characteristics

Here, we design a test of this paper's method in the context of weighted network. Note that in a real situation, almost every node has interactions with all others in a group or called a network, and the difference lies in the strength of the interactions between these nodes. Namely, some node pairs have strong relationships so as to have a high value of interaction between them, and some have very weak relationship so that it can be ignored. Based on this common phenomenon, we adopt a reverse engineering approach to design the simulation analysis as follows. A full connected and weighted network with 100 nodes is constructed and every node in the network owns three independent characteristics whose values meet the uniform distribution between 0 and 1. We next assign the weight values to these links by considering the premise that if two nodes are closer in the characteristics space, their link's weight is higher. Accordingly, the weight value could be given by the following formula:

$$w_{ij} = 1 - \frac{1}{3}\sum_{k=1}^{3}\left|x_i^k - x_j^k\right|, \tag{21}$$

where w_{ij} denotes the weight between nodes i and j, and x_i^k and x_j^k denote the k-th characteristics value of nodes i and j, respectively. In the above design, we have known these nodes' similarity measure which can be expressed by their closeness in the characteristics space or the defined w_{ij} in formula (21).

As for the test, we compare the nodes' similarity rankings obtained by the tested method with the known nodes' similarity rankings implied in their closeness of characteristics. If the two rankings obtained by one method are more consistent, the method will be much better. To compare two series of values' rankings, the *Kendall Tau rank-correlation coefficient* (τ, for short) is a widely adopted index, whose definition can be found in https://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient, and also the Appendix for convenience.

The results shown in Fig. 4 are from the 50 generated artificial networks. Recall that only CHM and ILM can be used for the weighted networks; thus we compare the two methods in this simulation analysis. The left part of Fig. 4 shows the mean and quartile of τ values obtained from the two methods, which demonstrates that ILM is much better than CHM. Furthermore, we make the paired-sample t-test for the two methods' τ values shown in the right part of Fig. 4, and the t-statistics value is 26.3730 (with p-value as 2.66×10^{-64}) which is significant at the 0.05 confidence coefficient. Thus, the ILM is better than CHM in the statistical sense in our designed simulation test.

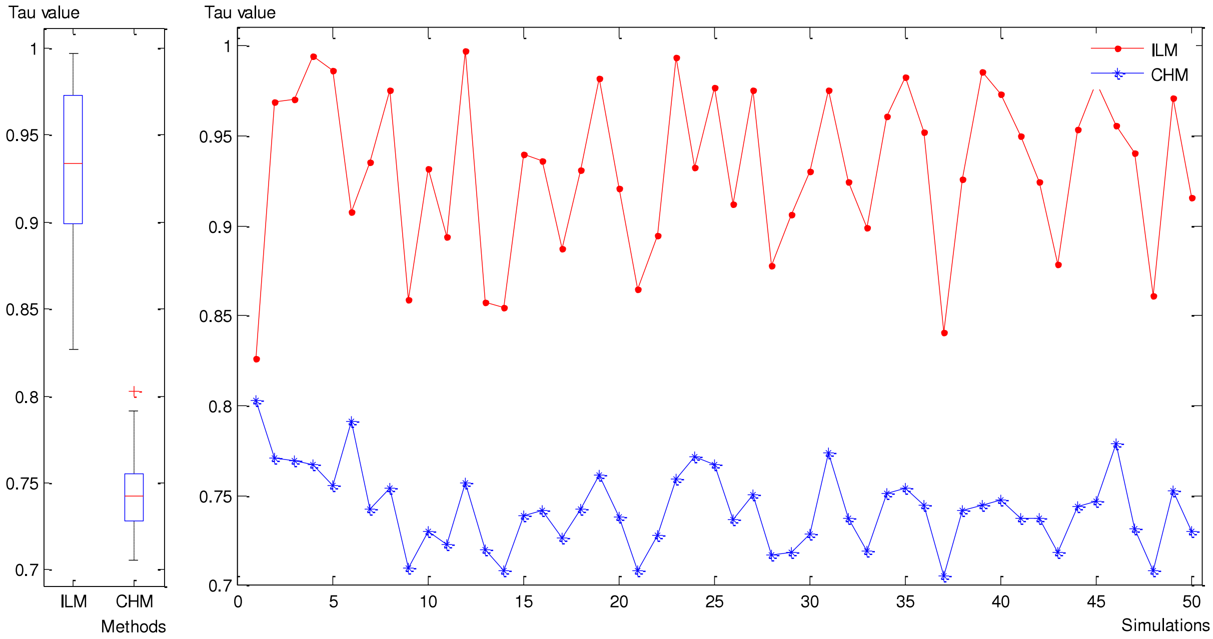

Fig. 4. The boxplot and lineplot of τ values obtained from CHM and ILM.

5. Applications

The proposed method was applied to two real networks with the given node characteristics but without the node similarity values. In these cases, based on the given characteristics, we can infer which nodes are similar, then check the results of our method with our inference. The application can help us to further test our method and to understand the reasons and origins of node similarity in the complex, real-world and tangled datasets. Our first example is a friendship network from Zachary's Karate club; our second is a collaboration network of scientists.

5.1. Friendship network

The first application here is drawn from the well known Zachary's Karate club [29]. The 34 members in this club form a friendship network which is a weighted network. In Zachary's study, he observed the 34 members over two years and found a disagreement between the administrator of the club and the club's instructor. As a result, the instructor left and started a new club, taking some of the club's members away. One interesting thing is whether the friendship network takes effect on forming the new club. In this paper, we will study this real-world problem from the viewpoint of node similarity. First of all, we draw Fig. 5 to show the friendship network and the split two clubs.

Then, our method is applied to measure the similarity between any two members, namely any two nodes in the network. We infer that the two nodes with large similarity measure are much likely to get together when the two clubs form. Thus, similar to Section 4.2.1, we can test whether the first four similar nodes of each node belong to the same club with the tested node. The whole results are shown in Table 3, where N_i means the wrong number of the obtained node i's similar nodes. From this table, we can find that most of the nodes and their first four similar nodes come from the same community. The total percentage of the first four similar nodes that do not belong to the same club with the tested node is below 10%. We can use our method to predict which nodes get together into the same club when the two clubs form. Thus, in this application, our new node similarity measure performs well in predicting the network evolution.

5.2. Collaboration network

The second application is based on a collaboration network of scientists [30]. There are 1589 nodes in this network which represent researchers from a broad variety of fields. An edge formed between a pair of researchers if they coauthored at least one paper. From the statistical analysis, every scientist has approximately four authors on average. Besides, the main research subject of every researcher is also included in this dataset. Here, we want to use our method to predict the research subject of one researcher; namely, given one researcher's research subject, we use our method and the collaboration network to find his or her similar node in order to predict the similar node's research subject based on the given information. To test our method, we can check the predicted result with the fact since everyone's research subject is given in this network. If this method goes well in this application, we can focus on studying only several nodes in one network and then infer that

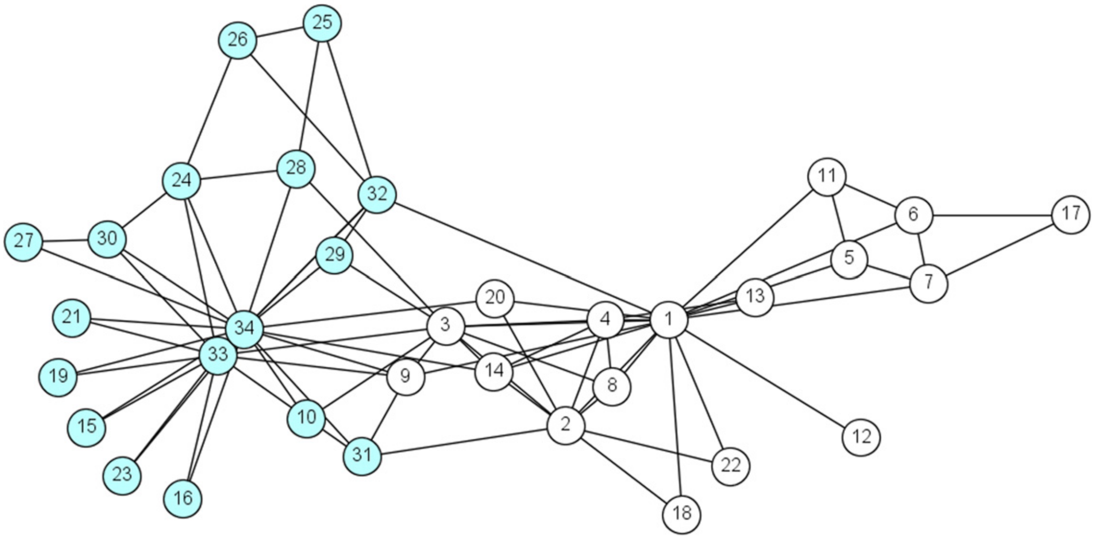

Fig. 5. The friendship network from Zachary's Karate club.

Table 3
Results of the first application.

Node	The first four similar nodes	N_i	Node	The first four similar nodes	N_i
01	12, 18, 22, 20	0	18	22, 12, 20, 13	0
02	22, 18, 20, 12	0	19	23, 21, 10, 15	0
03	18, 10, 12, 22	1	20	18, 22, 12, 10	1
04	13, 18, 08, 12	0	21	19, 15, 23, 10	0
05	11, 12, 18, 22	0	22	18, 12, 20, 13	0
06	17, 07, 11, 12	0	23	19, 21, 15, 10	0
07	17, 06, 12, 18	0	24	19, 21, 10, 15	0
08	18, 22, 12, 13	0	25	26, 28, 10, 19	0
09	10, 19, 31, 12	3	26	25, 32, 10, 19	0
10	19, 29, 23, 21	0	27	30, 10, 19, 12	1
11	05, 12, 18, 06	0	28	10, 19, 25, 12	1
12	18, 22, 01, 20	0	29	10, 19, 12, 18	2
13	04, 12, 18, 22	0	30	27, 19, 21, 10	0
14	18, 10, 12, 22	1	31	19, 21, 10, 15	0
15	19, 21, 23, 10	0	32	19, 29, 10, 21	0
16	19, 21, 23, 15	0	33	21, 19, 15, 23	0
17	07, 06, 12, 18	0	34	19, 10, 23, 15	0

its similar nodes would have the same attributions with our known one based on the responding formed network and our new method. First of all, the collaboration network is shown in Fig. 6 with distinguishing the scientists by different colours according to their research subjects.

Next, we randomly select one node and apply the proposed method to find its two most similar nodes. We inferred the similar nodes' research subject using the given node's information and compare it with the fact one. After repeating it 500 times, the total accuracy is approximately 83%. From this result, our method thus can infer the attributions of the similar nodes precisely based on several nodes that we have studied. Accordingly, we only need to research a few nodes when facing a new network; then, the other nodes' attributes can be inferred by their similar nodes. It can be a time-saving and labor-saving way to study a new network, especially a large-scale network.

6. Conclusions and future work

This study proposes a new method to measure network node similarity. In contrast to the existing methods, the new method has the following features: (1) the new method utilizes the concept of information and information loss to measure the node similarity. Based on this concept, we point out that if two nodes are more similar than the others, then the information loss of seeing them as the same is also less. Accordingly, several mathematical formulas are deduced based on information theory; (2) the new method has the algorithm complexity $O(n^2)$, for which some large real-world networks can be solved with a low time cost; (3) illustrated from the artificial examples and applications, the new method can give more reasonable results which are more consistent with our common judgment.

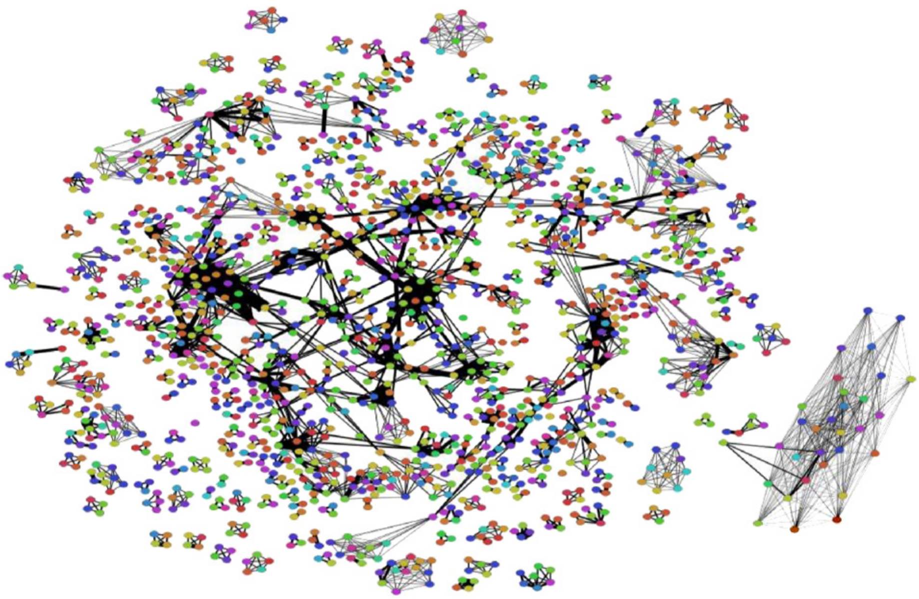

Fig. 6. Structure graph of the collaboration network.

The new method can be further extended in theory and in practice, which could be the future work. In theory, the method can be the first step to detect the community structure in one network. Since the concept of information loss can be used to find the similar nodes, the concept can also be used to measure the quantitative loss of getting nodes together into one community. Following this idea, one new method for exploring the communities of networks can be created. Besides, the input of our model is only the network in itself without any other characteristic of the nodes, but in many real-world cases, such nodes' characteristic information is often available; thus it may be a good direction for introducing such information into our established model, which would likely improve the method's accuracy and applicability. Although we have partly done something in our application, it is not enough and this work needs to be further elaborated on in the near future. Furthermore, the result's normalization, for the node's similarity comparisons between networks rather than within a single network, can also be an interesting topic, for which Vitanyi et al. (2009) [31] could be a useful reference. In practice, we have given some new ideas for using the new method in the application part of this paper. Especially, finding the similar nodes can predict the network evolution as the first application shows and also can predict the other unknown nodes' attributions as the second application shows. There is also a large space left for using the new method in many fields such as goods recommendation, link prediction, epidemic control and so forth. Last but not the least, we deeply hope the idea and the method presented here will prove applicable and useful in the analysis of many other networks.

Acknowledgments

We greatly thank the two anonymous reviewers for their helpful comments and constructive suggestions. We also appreciate Mr. Wen Zhao for his help in polishing the language. This study was partly funded by National Natural Science Foundation of China (No. 71271070) and China Scholarship Council (No. 201306120159).

Appendix

Kendall Tau rank-correlation coefficient, τ for short. Given two vectors $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$, for example $\{a_i\}_{i=1}^n$ express the every two nodes' similarity calculated by different similarity measures and $\{b_i\}_{i=1}^n$ express the link's weight between the corresponding nodes. Then we have

$$C_{ij} = \begin{cases} 1 & \text{If } (a_i < a_j \text{ and } b_i < b_j) \text{ or } (a_i > a_j \text{ and } b_i > b_j) \text{ or } (a_i = a_j \text{ and } b_i = b_j), \\ 0.5 & \text{If } (a_i = a_j \text{ and } b_i \neq b_j) \text{ or } (a_i \neq a_j \text{ and } b_i = b_j), \\ 0 & \text{Other cases} \end{cases}$$

and the number of concordant pairs is $C = \sum_{i=1}^n \sum_{j=i+1}^n C_{ij}$. Accordingly, Kendall's Tau rank-correlation coefficient is $\tau = (4C/(n(n-1))) - 1$.

The above defined τ has the following properties: (1) if the two vectors imply the same ranking, then $\tau = 1$. Otherwise, if the ranking implied by one vector is the reverse of the other, then $\tau = -1$. (2) For all the cases, τ lies between -1 and 1, and the larger the value is, the more the agreement between the rankings implied by the two vectors is.

References

[1] X. Ye, R. Dan, Proceedings of the 5th Ph.D. Workshop on Information and Knowledge 25, 2012.
[2] E.A. Leicht, P. Holme, M.J. Newman, Phys. Rev. E 73 (2006) 026120.
[3] G. Bianconi, P. Pin, M. Marsili, Proc. Natl. Acad. Sci. 106 (2009) 11433.
[4] J. Kwon, S. Kim, Int. J. Comput. Sci. Netw. Secur. 10 (2010) 116.
[5] S. Currarini, M.O. Jackson, P. Pin, Proc. Natl. Acad. Sci. 107 (2010) 4857.
[6] S. Currarini, M.O. Jackson, P. Pin, Econometrica 77 (2009) 1003.
[7] L. Lü, T. Zhou, Physica A 390 (2011) 1150.
[8] O. Allali, C. Magnien, M. Latapy, Intell. Data Anal. 17 (2013) 5–25.
[9] Y.T. Bian, L. Xu, J.S. Li, J.M. He, Y.M. Zhuang, Discrete Dyn. Nat. Soc. (2012).
[10] E. Gould, E. Winter, Rev. Econ. Stat. 9 (2009) 188.
[11] B.B. Wang, L. Gao, Y. Gao, J. Stat. Mech. Theory Exp. (2012) P04011.
[12] Á. Corral, J. Stat. Mech. Theory Exp. (2009) P01022.
[13] R.K. Ahuja, C. Liebchen, Networks 57 (2011) 1.
[14] O. Penner, V. Sood, G. Musso, K. Baskervile, P. Grassberger, M. Paczuski, Physica A 387 (2008) 3801.
[15] D. Roessner, A.L. Porter, N.J. Nersessian, S. Carley, Scientometrics 94 (2013) 439.
[16] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons, 2012.
[17] J. Rissanen, Automatica 14 (1978) 465.
[18] P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et du Jura, Impr. Corbaz, 1901.
[19] G. Salton, Science 168 (1970) 335.
[20] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabasi, Science 297 (2002) 1551.
[21] R.S. Burt, Soc. Forces 55 (1976) 93.
[22] D.S. Goldberg, F.P. Roth, Proc. Natl. Acad. Sci. 100 (2003) 4372.
[23] P. Symeonidis, E. Tiakas, Y. Manolopoulos, Proceedings of the Fourth ACM Conference on Recommender Systems, 2010, pp. 183–190.
[24] T. Tanimoto, Internal Report: IBM Technical Report Series, NY: IBM, 1957.
[25] H.H. Chen, L. Gou, X.L. Zhang, C.L. Giles, Proceedings of the 27th Annual ACM Symposium on Applied Computing, 2012, pp. 138–143.
[26] K. Nowicki, T.B. Snijders, J. Amer. Statist. Assoc. 96 (2001) 1077.
[27] K. Thiel, M.R. Berthold, IEEE 10th International Conference on Data Mining, 2010, pp. 1085–1090.
[28] M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. 99 (2002) 7821–7826.
[29] W.W. Zachary, J. Anthropol. Res. (1977) 452–473.
[30] M.E.J. Newman, Phys. Rev. E 74 (2006) 036104.
[31] P.M.B. Vitanyi, F.J. Balbach, R.L. Cilibrasi, M. Li, Information Theory and Statistical Learning, Springer, 2009.