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.

© 2014 Elsevier B.V. All rights reserved.

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:

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

Download English Version:

https://daneshyari.com/en/article/7380735

Download Persian Version:

https://daneshyari.com/article/7380735

Daneshyari.com