A Similarity Link Prediction Method in Complex Network Based on Endpoint Clustering

Yang Yang[1], Yuchun Xu[2] and Xin Yang[3]
[1]153 Centoal Hospital, zhengzhou 450002
[2]66389 Troops,wuhan 430071
[3]Air Defense Academy, zhengzhou 450002

***Abstract*—Link prediction aims to predict the probability of the existence of links between two endpoints in complex network. Many methods ignore the clustering of endpoints when calculate the similarity between two endpoints. To distinguish the contribution of endpoints clustering, we propose a similarity link prediction method based on endpoint clustering. In order to improve the link prediction accuracy, the method considers both the common neighbor and endpoint clustering. Empirical study on six real networks has shown that the method we proposed can achieve a good performance, compared with CN, AA, RA, LP and Katz.**

Keywords-complex network; link prediction; agglomeration, similarity

I. INTRODUCTION

In recent years, complex networks as a new discipline - network science, has been widely concerned about the field. Link prediction technology, as an important direction of complex networks, is designed to study the possibility of establishing a connection between any two points in the network.

At the beginning of link prediction research, it was based on machine learning and network information mining, the reality of the network attribute information acquisition is more difficult, so the utility is poor. At present, many scholars are more concerned with the similarity link prediction method based on network topology. Similarity link prediction methods include local similarity and global similarity. (1) is the simplest local similarity index, which mainly describes the similarity between nodes by calculating the number of common neighbors among nodes. The local similarity index (CN) is the simplest local similarity index. How to put forward a relatively low complexity, the effect is better method is the core of the study. As shown in Figure 1, there is a common neighbor for node pairs (x,y), and between (x1,y1). It can be seen that the degree of node (x,y) around the node is obviously higher than that of (x1,y1)., and the similarity between node (x,y) is higher than that of node (x1,y1). The possibility of establishing a connection (x,y) is also higher than that of the node (x1,y1).

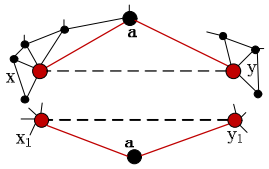

FIGURE I SCHEMATIC DIAGRAM OF SIMILARITY BETWEEN NODES AT DIFFERENT ENDPOINTS

Considering the difference of clustering degree between two endpoints of different node pairs, this paper will re-characterize the similarity between two nodes based on the degree of agglomeration of endpoints and explore the effect of endpoint agglomeration on similarity, and further improve link prediction The prediction accuracy. It is proved that the link prediction method based on endpoint aggregation degree has higher prediction accuracy than CN, AA, RA, LP, Katz and other indicators in a number of actual network data.

II. SIMILARITY INDICATORS

Similarity indicators, including based on local information, global information, now introduce the follow-up article related to the relevant methods are as follows:

- CN [1]: The similarity between nodes is determined by the number of common neighbors between two nodes. The common neighbor indicator is expressed as (the neighbor of the node)

$$s_{xy} = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

- AA [3]: the role of common neighbors to distinguish, the greater the degree of nodes, the smaller the similarity. Specifically defined as:

$$s_{xy} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{\log k_z} \tag{2}$$

- RA [2]: For nodes in the network and according to the process of resource allocation, a unit of resources from the node, through the common neighbors to the amount of resources is similarity quantitative value, the similarity described as :

$$s_{xy} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{k_z} \tag{3}$$

- LP [4] (Local Path): On the basis of CN, the effect of longer path on similarity is considered, and the longer the path is, the smaller the effect is:

$$\boldsymbol{S} = \boldsymbol{A}^2 + \alpha \cdot \boldsymbol{A}^3 \tag{4}$$

Where is the adjacency matrix A,(A3)x y representing the number of paths between nodes x, y and length 3, $0 < \alpha < 1$ as the adjustment parameter.

- Katz: on the basis of the LP index, all the path was calculated, expressed as:

$$s_{xy} = \sum_{l=1}^{\infty} \alpha^l \cdot |path_{xy}^l| = \alpha \boldsymbol{A}_{xy} + \alpha^2 (\boldsymbol{A}^2)_{xy} + \alpha^3 (\boldsymbol{A}^3)_{xy} + \ldots \tag{5}$$

$path_{xy}^l$ is the number of paths between the length of nodes x and y, $0 < \alpha < 1$ is the adjustment parameter.

III. Forecasting Method Based on Endpoint Aggregation

For a pair of nodes, the common neighbors play an important role in their connection, but some of the connection characteristics of the two endpoints also have a significant impact on their similarity. The network empirical and research shows that the higher the agglomeration around the node, the greater the likelihood of the existence of the connection (5,6), and the clustering coefficient is the local topology of the description and reflection of the node. The clustering factor of any node in the network can be expressed as follows:

$$C_i = \frac{\langle E_i \rangle}{k_i (k_i - 1)/2} \tag{6}$$

In the formula, the number of connections between nodes $\langle E_i \rangle$ as nodes for nodes i.

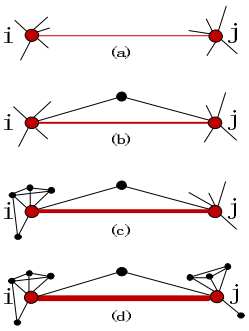

FIGURE II COMPARISON OF DIFFERENT TOPOLOGY CONNECTIONS IN THE NETWORK

Figure 2 shows four typical topological connection cases, which include two kinds of topology comparison: First, the number of common neighbors, that is, (a) and (b); Second, the endpoint clustering coefficient of different contrast, that is (b), (c) and (c), (d). First, for both topologies of (a) and (b), the degrees of the two endpoints are the same, and the only difference is that the number of common neighbors in (b) is greater than (a), so from the similarity (a); Secondly, for both b (b) and (c) topologies, the number of co-neighbors and nodes of the endpoints are the same, but the endpoints of (c) are the same (C) the possibility of establishing a connection between two points is higher than (b); similarly, for (c) the probability of establishing a connection between the two points is higher than that of (b) And (d), from the point of view of similarity (d) the possibility of establishing a connection between two points is higher than (c). It can be seen that for the four typical topologies, the similarity between endpoints is from large to small (d)> (c)> (b)> (a).

Based on the discussion and analysis of the above network structure, for the nodes in the network I and j, the definition of the possibility of connection between them, that is, the similarity between nodes is:

$$s_{xy}^{EC} = \sum_{z \in \Gamma(x) \cap \Gamma(x)} C_x \cdot C_y \tag{7}$$

The clustering coefficients of the endpoints Cx are given in the above formulas, but there is a sparsity in some networks, that is, the clustering coefficient of a small number of nodes is close to or equal to zero. Thus, the similarity of formula (7) is further defined as follows:

$$s_{xy}^{EC} = \sum_{z \in \Gamma(x) \cap \Gamma(x)} (C_x + 1) \cdot (C_y + 1) \tag{8}$$

Through the definition of formula (8), the similarity degree characterization method can describe the number of common neighbors and describe the clustering coefficient. In terms of complexity, the complexity of the clustering coefficient is small and does not affect the overall complexity. The complexity of the method (8) is the same order of magnitude as that of CN- $O(N < k >^2)$

IV. Experimental Results Analysis

A. *Evaluation Indicators*

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

For a given unqualified network $G(V, E)$, which V represents a collection of nodes in the network, E represents a collection of all edges. Given a link prediction algorithm, this algorithm assigns an evaluation score to all unconnected connections in the network, and the higher the fractional value,

the higher the probability of a connection between nodes. In order to evaluate the accuracy of the link prediction algorithm, the connections existing in the network are generally divided into training sets E^T and test sets E^P,and $E^T \bigcap E^P = \varnothing$.

AUC can be simply understood as the probability that a fractional value E^P of one edge is randomly selected in the test set is greater than the fractional value of the unconnected edge [7]. If the test set is greater than the edge of the edge is not connected (n'), then add 1 point, if the two are equal (n''), then add 0.5, expressed as:

$$\text{AUC}=\frac{n'+0.5n''}{n} \tag{9}$$

Among them, the number of times n for comparison. Obviously, if all the results produced by the algorithm are randomly generated, $\text{AUC} \approx 0.5$,the number of AUCs over 0.5 is measured by the accuracy of the current link prediction algorithm compared to the random method.

B. Results Analysis

In order to verify the validity of similarity link prediction method (EC) based on endpoint clustering, this paper compares five typical similarity indexes of CN, AA, RA, LP and Katz respectively.

TABLE I. COMPARISON OF AUC RESULTS SUCH AS LWCN AND CN

Net-work	***PB***	***Router***	***Celegans***	*SciMet*	***Hamster***	*Open Flights*
CN	0.923	0.653	0.851	0.802	0.814	0.970
AA	0.927	0.653	0.868	0.804	0.816	0.972
RA	0.928	0.653	0.872	0.803	0.817	0.973
LP-0.001	0.937	0.945	0.869	0.919	0.934	0.984
LP-0.01	0.938	0.946	0.868	0.920	0.938	0.981
Kate-0.001	0.936	0.973	0.868	0.942	0.934	0.984
Kate-0.01	0.933	0.971	0.869	0.943	0.933	0.983
EC	**0.949**	**0.974**	**0.891**	**0.943**	**0.978**	**0.985**

a. Sample of a Table footnote. (Table footnote)

Table 1 shows the comparison of EC link prediction methods and other methods in different actual networks. It can be seen that the proposed method EC has improved the prediction accuracy of link prediction in six networks. In particular, CN is the simplest link prediction method, and its complexity is the lowest, but it also plays a certain prediction effect. AA and RA are weighted by common neighbors on the basis of CN, and are generally better than CN in the case of lower complexity. Katz considered all the paths between the two nodes, the highest complexity, but the effect is better, many networks are significantly better than other methods such as Router and Hamster network. LP can be said to be a compromise between complexity and precision, which only considers the two-order path, but its prediction accuracy is very close to Katz. On the basis of CN, the method has recalculated the degree of aggregation of two endpoints, and its prediction accuracy has obviously improved greatly, especially in some sparse networks. Network evidence shows that many of the less predictive methods perform well in high agglomeration networks even close to Katz (such as PB and Celegans networks), but perform poorly in sparse networks (such as Router and Hamster networks). The proposed EC method shows a higher prediction accuracy in the sparse network Router and Hamster under the premise that the cost and common neighbor method are the same time complexity. In general, the EC method also considers the common neighbor and node aggregation, and improves the prediction accuracy of link prediction. And its time complexity $O(N<k>^2)$ is the same as that of CN, which can be applied to the link forecasting of large complex networks.

V. CONCLUSION

Link prediction based on similarity between nodes is an important research method in complex network link prediction. In this paper, a similarity method is proposed to eliminate the problem of endpoint clustering when characterizing the similarity. In this paper, a similarity link prediction method based on endpoint aggregation is proposed. Six actual network tests show that compared with CN, AA, RA, LP and Katz, the proposed method has high prediction accuracy. In addition, this method belongs to local similarity method, the complexity is low, and it is better in sparse network, and can be applied to large complex network.

REFERENCES

[1] Liu Y Y, Slotine J J, Barabási A L. Observability of complex systems[J]. Proceedings of the National Academy of Sciences, 2013, 110(7): 2460-2465.

[2] Bianconi G, Barabási A L. Bose-Einstein condensation in complex networks[J]. Physical review letters, 2001, 86(24): 5632.

[3] Albert R, Barabási A L. Topology of evolving networks: local events and universality[J]. Physical review letters, 2000, 85(24): 5234.

[4] Li X, Chen G. A local-world evolving network model[J]. Physica A: Statistical Mechanics and its Applications, 2003, 328(1): 274-286.

[5] Chang H, Su B B, Zhou Y P, et al. Assortativity and act degree distribution of some collaboration networks[J]. Physica A: Statistical Mechanics and its Applications, 2007, 383(2): 687-702.

[6] Noh J D, Rieger H. Random walks on complex networks[J]. Physical review letters, 2004, 92(11): 118701.

[7] Noh J D, Rieger H. Random walks on complex networks[J]. Physical review letters, 2004, 92(11): 118701.