Application of Link Prediction in Temporal Networks

Haihang Xu, Lijun Zhang
Department of Computer Science,
Beijing University of Aeronautics and Astronautics,
Beijing, 100191, China
e-mail: xuhaihang@gmail.com, ljzhang@buaa.edu.cn

***Abstract*—Link prediction is an important research hotspot in complex networks. Correlational studies merely use static topology for prediction, without considering the influence of network dynamic evolutionary process on link prediction. We believe that the links are derived from the evolutionary process of network, and dynamic network topology will contain more information, Moreover, many networks have time attribute naturally, which is apt to combine the similarity of time and structure for link prediction. The paper proposes the concept of active factor using time attribute, to extend the similarity based link prediction framework. Then model and analysis the data of citation network and cooperation network with temporal networks. Design the active factors for both network sand verify the performance of these new indexes. The results shows that the indexes with active factor perform better than structure similarity based indexes.**

Keywords: active factor; complex networks; link prediction; temperal networks;

I. INTRODUCTION

Watts and Strogtz raised small world network model [1] in 1998, then Barabási and Albert put forward scale-free network model in 1999, which explained the mechanism of production of power-law distribution [2]. Both of them have established the theoretical foundation of complex networks which is an emerging research field. Many practical networks, afterwards, were proved to be of characters of complex networks, such as neural network [1], Internet routing network [3], WWW (World Wide Web) [4] and social relationship network [5], etc. In recent years, the research hotspots of complex networks have already been transferred to aspects of structure mining and mechanism of transmission from network modeling, where link prediction is an important hot research point. Link prediction, is to evaluate the possibility of generating new edge between unconnected nodes under circumstances of known network structure and some other information. On the one hand, the existing yet unfound edges can be discovered; on the other hand, new edges possibly generated in the future may be predicted [6]. Link prediction is of important research value in both practical application and theoretical research.

The existing link prediction algorithms are mainly divided into three categories: link prediction algorithms based on similarity [7], link prediction algorithms based on probability statistics [8] and link prediction algorithms based on machine learning. The first kind of algorithms is to define a similarity index, evaluate a value for every pair of nodes, the bigger value the index is, the more probably two nodes connect each other. Liben-Nowell and Kleinberg [7] have concluded a number of similarity-based indexes, and divided them into two categories: node-based and path-based. The performances of different indexes were analyzed in empirical studies on large-scale cooperation network. Zhoutao, et al [9] have compared 9 local information based indexes in 6 practical networks. The link prediction algorithm based on similarity is of simple computation, and gets good performance, which is one of the most used algorithms at present, while the latter two algorithms fail to be applied in large networks owing to complex computation.

Pre-existing researches are focusing on static network structure, without considering the influence of dynamic evolutionary process on link prediction. We believe that the links are derived from the evolutionary process of network, dynamic network topology will contain more information, and many networks have time attribute naturally. In [10] Holme has summarized temporal networks: a kind of dynamic network structure, in which each edge has a time attribute, marking its generating time respectively, the edge with multiple time attributes indicates that this edge appeared repeatedly. Temporal networks can be used to predict not only the missing edges in the evolutionary process, but also the possibly generated edges in the near future.

Section.2 reviews the definition of link prediction problem, evaluation method, some existing prediction indexes and temporal networks. Section.3 adopts temporal networks to model the data sets, introduce active factor as per design, test and verify new performance of temporal and structure similarity based index.

II. RELATED WORK

Current correlational studies describe network structure by adopting static topology, which discard the information concerning dynamic evolution of network completely. The link prediction problem is defined in [11] in the way: given an undirected network *G (V,E)*, where *V* represents the node set, *E* for edge set, *N* for total number of nodes, and *M* for number of edges; this network has *N(N-1)/2* node pairs, constituting universal set *U*. Given a link prediction method, assign value p for every pair of nodes, which indicates the possibility of the pair nodes may be connected.

AUC method (area under the receiver operating characteristic curve) is adopted to measure the accuracy of link prediction algorithm. Divide the edge set into two parts: the training set E^T, is treated as known information, while the testing set E^P, is for testing. Only the information in

Published by Atlantis Press, Paris, France.

training set can be used in calculation. $U - E^T$ is known as unknown edge set. AUC can be interpreted as that, to what extent, the prediction probability of edge in E^P is higher than that of edges randomly selected from $U - E^T$.

Randomly select one edge from E^P and $U - E^T$ for comparison every time. Do this for *n* times independently. If there are n_1 times that the edges from E^P have higher score and n_2 times that the edges from E^P and $U - E^T$ have the same score, then the AUC value is:

$$\text{AUC} = (n_1 + 0.5n_2)/n. \quad (1)$$

Obviously, if the probability distribution of connecting edges is random, then AUC=0.5. Therefore, the degree of AUC $>$ 0.5 measures how much the prediction algorithm performs better than random selection.

Our work selects four similarity indexes to extend: Preferential Attachment (PA), Common Neighbors (CN), Jaccard's Coefficient (JC) and Adamic-Adar (AA).

Preferential Attachment: PA doesn't consider the common neighbors, directly uses nodes' degree for prediction [2]. s_{xy} represents the score of node *x* and *y* in this prediction index and $k(x)$ denotes the degree of node *x*.

$$s_{xy} = k(x) * k(y). \quad (2)$$

Common Neighbors: if two nodes share many neighbors, then they are inclined to be connected. $\Gamma(x)$ indicates the set of neighbors of node x.

$$s_{xy} = |\Gamma(x) \cap \Gamma(y)|. \quad (3)$$

Jaccard's Coefficient: intersection of neighbors set of two nodes divides their union set. Here use the percentage of common neighbors for prediction [12]:

$$s_{xy} = |\Gamma(x) \cap \Gamma(y)|/|\Gamma(x) \cup \Gamma(y)|. \quad (4)$$

Adamic-Adar: AA no longer simply calculates the amount of common neighbors, but assigns higher weight on nodes with lower degrees in common neighbors [13].

$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} 1/log k(z). \quad (5)$$

III. DATA AND EXPERIMENTS

A. Date Sets

The data of citation network derives from [14], [15], including the citation data in high energy physical phenomenon between 1992 and 2001 on Arxiv.org site, while the data of cooperation network comes from thesis information concerning high energy physical theory part between 1991 and 2012, each paper only selects the first two authors as one cooperative relationship. Both networks are meeting the scale-free characteristic of complex network.

Table.1 describes the specific scales of the two networks. Citation network contains 28892 papers and 322200 citation relationships, while cooperation network has 25818 scientists and 57958 cooperative relationships, 32848 of which are non-repeated cooperative relationships.

TABLE I. SCALE OF THE TWO NETWORKS

Networks	Num. of Nodes	Num. of Edges
Citation Network	28892	322200
Cooperation Network	25818	57958 (32848)

Fig.1 describes the evolution of citation network (a), (b), (c) and cooperation network (d), (e), (f), including nodes, edges and clustering coefficient of maximum connected component (MCC). The numbers of nodes in both networks increase rapidly, and clustering coefficient of MCC in citation network increases rapidly and maintains at a higher level owing to the faster increasing speed of edges, while the edge number of cooperation network increases slower, and its MCC's clustering coefficient keeps at a very low level.

B. Experiments on Citation Network

Citation network has such features: each directed edge and node has only one time attribute; the edges with the same initial node have the same time attribute; and the time attributes of these edges and this initial node are identical.

Define citation network at time *T* as:

$$G_T(V_T, E_T) = \cup\{G_t(V_t, E_t) | t \le T \text{ and } t \in Z\}, \quad (6)$$

Then there are the following properties:

$$V_{t0} \subseteq V_{t1} \text{ and } E_{t0} \subseteq E_{t1}, \quad iff: t0 \le t1,$$
$$\forall u \forall v((u, v \in V_T \text{ and } \langle u, v\rangle \in E_T) \rightarrow (t(u) \ge t(v) \text{ and } t(u) = t(\langle u, v\rangle))).$$

Where $G_T(V_T, E_T)$ denotes the temporal networks at time *T*, V_T represents the node set and E_T represents the edge set respectively. *t*(*u*) and *t*(⟨*u*,*v*⟩) get the time attributes of node *u* and edge ⟨*u*,*v*⟩.

We have analyzed the distribution of citations with the related papers' time attributes in citation network. Fig.2a describes the percentages of cited number of papers published in 1992 accounts for the newly-increased citation relationship in the next 9 years, and this figure shows power-low diminishing characteristic. Furthermore, classify the yearly-increased cited number within 10 years according to citation time. The value of position(*x*,*y*) in Fig.2b indicates the number of citations that papers published in the *x* year cited papers in the *y* year, which further verifies that the number of citations and nodes' time difference have positive correlation.

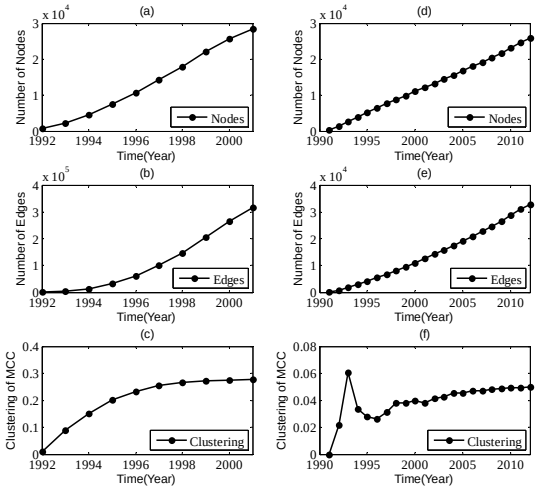

Figure 1. Statistics in the evolution of the networks.

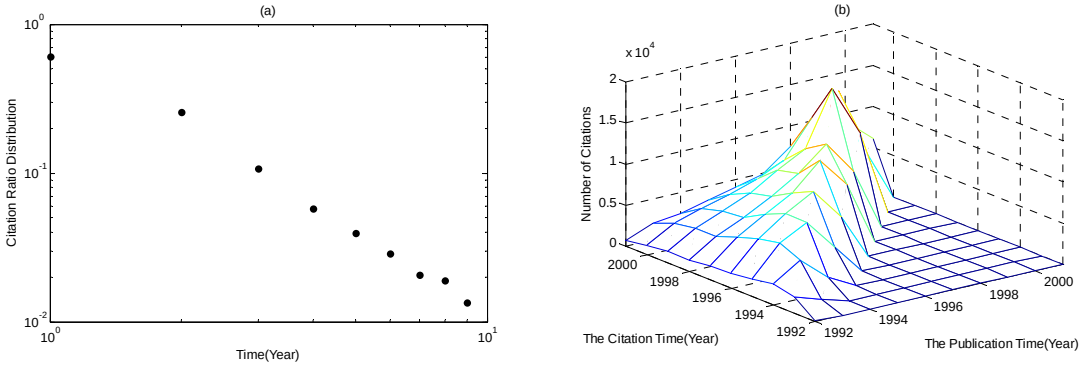

Figure 2. Relationship between citations and the related papers' time.

An active factor γ is defined as following:

$$\gamma = \begin{cases} 0, & t < 0 \\ \alpha, & t = 0 \\ t^{\beta}, & t > 0 \end{cases}, \quad (7)$$

$$t = t(x) - t(y), 0 < \alpha < 1, \beta < 0.$$

$$S_{xy}^{new} = \gamma * S_{xy}^{old}. \quad (8)$$

(7),(8) are used to extend the exsiting indexes PA, CN, JC, AA to 4 new indexes which are called A-PA, A-CN, A-JC, A-AA.

PA, CN, JC and AA were supposed to be applied in undirected networks. In order to make them adapting directed networks, we assume that degree includes out-degree and in-degree, while neighbors includes predecessors and successors.

With regard to citation network, take samples randomly in different rates as training set, apply old and new indexes for comparison respectively, with $\alpha = 0.5$, $\beta = -1.7$ in experiment. Fig.3 indicates that PA has a great promotion after introducing the active factor, PA and A-PA perform better than the rest of indexes when sampling rate is lower, but the performance of other 6 indexes is promoted rapidly as the rise of sampling rate.

C. Experiments on Cooperation Network

Cooperation network is a network with scientists as nodes and cooperative relationships as edges. Each edge has a sorted time attribute list, more than one time attribute shows the existence of multiply cooperative relationships between two scientists. T_{Create} is the time when the node appeared firstly in the network, while T_{Last} is the time when the last cooperation happened.

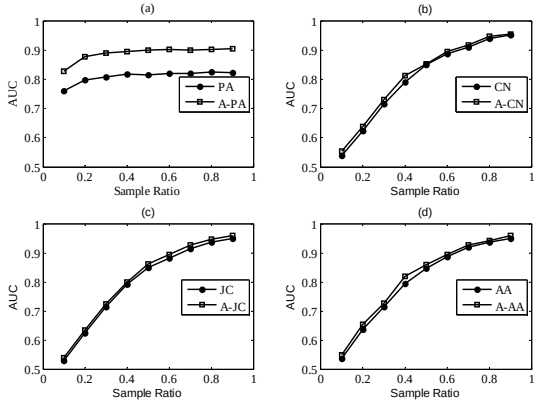

Figure 3. AUC of the eight indexes in the citation network.

The time distribution of new edges is illustrated in Fig.4. It describes the relationship between cooperation happened in 2012 and T_{Last} of cooperators. People with higher T_{Last} are more inclined to have new cooperative relationship, which presents exponential relationship.

Define an active factor γ as following:

$$\gamma = e^{\alpha(x+y)}, \alpha > 0. \quad (9)$$

Where x and y denote the two nodes' T_{Last}.

(8), (9) are used to construct the new indexes based on PA, CN, JC, AA. The new ones are called CA-PA, CA-CN, CA-JC, CA-AA.

Prediction problem on cooperation networks usually means that given the network at certain time, predicting the probability of new edge appearing in near future. Network is keeping growing, which would increase many nodes and edges, we obviously get little information about new nodes, thus is unable to predict the edge having new node(s) effectively [7]. Though two cooperators probably cooperate again, this is not our goal. Therefore, the prediction objective focus on new edges (non-repeated) generated in future period between nodes existed in original network.

Given a temporal network $G_T(V_T, E_T)$ and divide it into two parts. The network $G_t(V_t, E_t)$, at time t (t<T), is treated as the training set, and the new edges in (10) with time attribute larger than t constitute the testing set.

$$\{\langle u, v\rangle | \langle u, v\rangle \in (E_T - E_t) \; and \; u, v \in V_t\}. \quad (10)$$

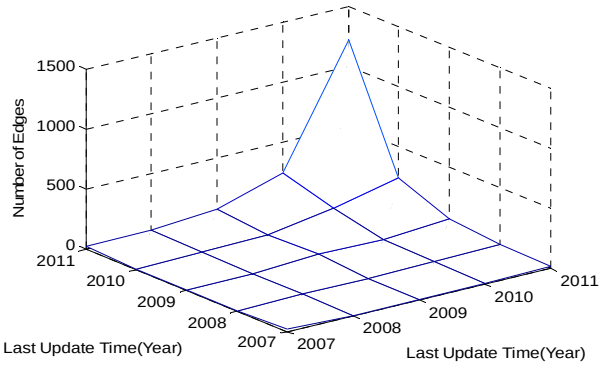

Figure 4. Time distribution of cooperative relationships.

In our experiments, use the data from 1991 to 2000 as the training set and the new edges in the next three years as the testing set. Once complete the prediction task, and then increase training set year by year. Fig.4 shows the AUC values of these eight indexes. The eight indexes stay immune to the scale of the training set, PA is increased significantly after introducing active factor, while CA-CN, CA-JC and CA-AA are promoted slightly comparing with original indexes.

D. Result and Explaination

The results of our experiments show that the active factor has remarkable effect on PA, an index based on node degree, meanwhile three indexes as CN, JC and AA are promoted very limited but consistent after introducing the active factor.

PA and its derivations have balance performances which are not sensitive to the scale of the training set in both networks. In the citation network, the common neighbors based indexes are very sensitive to the sampling rate, while they all perform not very well in the cooperation network.

Xu Feng et al. have studied the influence of clustering coefficient on various prediction indexes in [16], and finding out that clustering coefficient has restricted the performance of prediction indexes which are based on local information. Low clustering coefficient means the correlation between nodes is not strong enough, so local information may plays an insignificant role in prediction.

In our work, the lower sampling rate, the sparser of the training set network, lead to lower clustering coefficient. Moreover, Fig.1(f) indicates the clustering coefficient of the cooperation network basically keeps below 0.05, far below the stable value of clustering coefficient in the citation network, so the prediction results would not be promoted significantly even if keep increasing the training set.

IV. CONCLUSION

Link prediction is a current open research hotspot and many scholars have introduced kinds of methods and thoughts to solve this problem. Similarity based link prediction algorithm is an algorithm framework on which new link prediction algorithm can be constructed by projecting different similarity indexes. PA uses node degree as measurement, CN, JC and AA are indexes based on common neighbors, moreover, all of these are belonging to the category of network structure.

The paper has proposed a method to model the data with temporal networks, and the concept of the active factor which is used to extend existing structure similarity indexes into temporal structure similarity indexes.

The empirical studies on the citation network and the cooperation network show that the introduction of the active factor has greatly promoted the performance of PA index, but CN, JC and AA indexes only got minor lifting due to the restriction by network clustering. On the whole, active factor contains positive information, which is useful to improve link prediction.

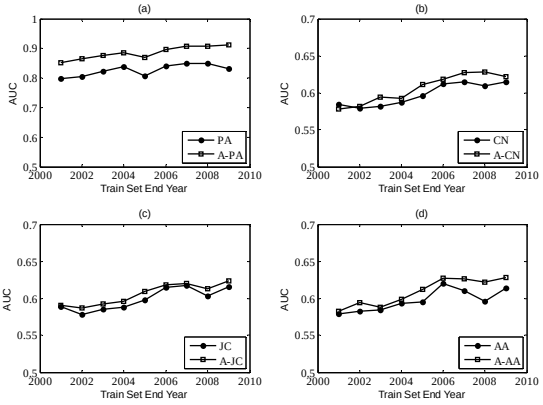

Figure 5. AUC of the eight indexes in the cooperation network.

ACKNOWLEDGEMENT

This work was supported in part by Graduate Innovative Practice Fund of Beijing University of Aeronautics and Astronautics.

REFERENCES

[1] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature. 1998, 393: 440~442.

[2] Barabási A L, Albert R. Emergence of scaling in random networks. Science. 1999, 286: 509-512.

[3] Faloutsos M,Faloutsos P,Faloutsos C. On power-law relationships of the Internet topology. ACM S1GCOMM Computer Communication Review, 1999, 29(4): 251.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University. 1998.

[5] L.A. Adamic, E. Adar. R E. Friends and neighbors on the web Social Networks. 2003, 25, 211-230.

[6] Guimera R, Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks. Proc Natl Sci Acad US. 2009. l06(52) : 22073-22078.

[7] Liben-Nowell, D,Kleinberg J. The link prediction problem for social networks. J Am Soc Inform Sci Technol. 2007, 58(7):1019–1031.

[8] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. Nature. 453:98–101, 2008.

[9] Zhou T, Lü L Y, Zhang Y C. Predicting missing links via local information. Eur Phys J B. 2009, 71(4):623~63.

[10] Petter Holme, Jari Saramäki , Temporal Networks, Phys. Rep. 519, 97-125 (2012).

[11] Lu L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks. Phys. Rev. E. 2009, 80: 046122.

[12] P. Jaccard. Bulletin de la Societe Vaudoise des Sciences Naturelles. 37, 547 (1901).

[13] L.A. Adamic, E. Adar. R E. Friends and neighbors on the web Social Networks. 2003, 25, 211~230.

[14] J. Leskovec, J. Kleinberg and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2005.

[15] J. Gehrke, P. Ginsparg, J. M. Kleinberg. Overview of the 2003 KDD Cup. SIGKDD Explorations 5(2): 149-151, 2003.

[16] Xu Feng, Jichang Zhao, Ke Xu, Link prediction in complex networks: A clustering perspective, European Physica B, VOL 85, Jan 2012.