ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 12, No. 4, 2017, *pp*.255-263

Temporal link prediction algorithm based on local random walk

Yuanxiao Fan[1], Pei-ai zhang[2]
[1]*School of Information and Science Technology, Jinan University, Guangzhou, 510632, China, E-mail: 1329791691@qq.com.*
[2]*School of Information and Science Technology, Jinan University, Guangzhou, 510632, China, E-mail: tzhangpa@jnu.edu.cn.*
(Received July 15, 2017, accepted September 25, 2017)

Abstract. Link prediction is an important part of complex network research. Traditional static link prediction algorithm ignores that nodes and links in network are added and removed over time. But temporal link prediction can use the information of historical network to make better prediction. Based on local random walk, this paper proposes a time-series random walk algorithm. Given link data for times 1 through T, then we predict the links at time T+1. The algorithm first computes the Markov probability transfer matrix at each time, then combines them into a transformation matrix, and applies the local random walk algorithm to obtain the final prediction result. The experimental results on real networks show that our algorithm demonstrates better than other algorithms.

Keywords: Strong and Weak solutions, temporal link prediction, Markov probability transfer matrix, local random walk.

1. Introduction

Many social, biological, and information systems can be well described by networks, where nodes represent individuals, biological elements, computers, web users, and so on, and links denote the relations or interactions between nodes. The study of complex networks has therefore become a common focus of many branches of science. Link prediction [1] aims at estimating the likelihood of the existence of links between nodes. The prediction of existent yet unknown links is similar to the data mining process, while the future links relates to the network evolution.

Link prediction has huge theoretical and practical value. It has attracted a great deal of interest of many scientists from different fields. In many biological networks, such as food webs, protein–protein interaction networks and metabolic networks, whether a link between two nodes exists must be demonstrated by a large number of laboratorial experiments, which are usually very costly. Instead of blindly checking all possible interactions, to predict based on known interactions and focus on those links most likely to exist can sharply reduce the experimental costs if the predictions are accurate enough [2,3]. In rapid development of social networks and e-commerce platforms, link prediction is widely used in personalized recommendation system. Specifically, in online social network, based on some information of the current network, link prediction can recommend friends to users by predicting their possible relationship [4]. In e-commerce platform, link prediction can recommend goods or information to users. It can not only improve the users' experience, but also solve the problem of data sparseness in commodity recommended system [5]. In addition, in scientific collaboration network, link prediction can be used to identify the possibility of cooperation [6]; In the aeronautical network, link prediction can provide transportation management strategy for airlines [7].

In recent years, the research of link prediction algorithms have attracted much attention. Document [8] is a summary literature of link prediction in complex networks, and link prediction algorithms are classified into three categories: (1) similarity-based algorithm; (2) the maximum likelihood algorithm; (3) probabilistic model for link prediction. This kind of classification reflects different modeling ideas of link prediction problem. The similarity- based algorithm is the most commonly used algorithm, which is characterized by simple algorithm, low computational complexity and widely apply. In this approach, some of the basic properties of nodes can be used to define their similarities, such as common features or topologies between nodes [9].

Most of the existing link prediction algorithms are still based on static network, that is, according to the static network at a certain time, the possibility of the link between nodes at next time is predicted, which is called static link prediction method [10]. In fact, these methods ignore the temporal characteristics of

network evolution, which is unreasonable in many specific application scenarios. Huang et al. [10] introduced temporal link prediction method, taking temporal evolutions of link occurrences into consideration to predict link occurrence probabilities at a particular time. Bliss et al. [11] proposed an evolutionary algorithm to integrate the topological features and node attribute characteristics in the network to improve the link prediction accuracy. Dunlavy et al. [12] considered bipartite graphs that evolved over time and considered to use matrix and tensor-based methods to predict future links. In addition, a weight-based approach is proposed to integrate multiyear data into a single matrix, and future links can be predicted by using a truncated singular value. Behnaz Moradabadi et al. [13] proposed a new time series link prediction based on learning automata, for each link that must be predicted there is one learning automaton and each learning automaton tried to predict the existence or non-existence of the corresponding link. Deng et al. [14] proposed the concept of a temporal link prediction method based on the prediction error correction. Ahmed et al. [15] proposed a method for link prediction in temporal uncertain networks, which formalized the predicting problem by designing a random walk in temporal uncertain networks to obtain more accurate results.

This paper is organized as follows: Section 2 describes the temporal link prediction problem. Section 3 provides some concepts, presents the local random walk method, and then gives a temporal link prediction algorithm. Section 4 presents and analyzes the experimental results for link prediction algorithm on real datasets. Finally, Section 5 offers conclusions and possibilities for future work.

2. Problem description

Given a network $\mathrm{G} = \{G_1, G_2, \cdots, G_t, \cdots\}$ represents an evolution of network (as shown in Figure 1). A temporal network can be described by snapshots $G_t = (V_t, E_t, A_t)$ for $t = t_0, t_0 + 1, \cdots, t_0 + T - 1$, where T is the window size, V_t is the node set of G_t, E_t is the edge set of G_t, and A_t is the adjacency matrix. The temporal link prediction problem is to predict the probability of occurrence of edges between two nodes in the network at time $t_0 + T$. At time t, a link prediction algorithm is given to predict the probability s_{xy} that any nodes pair (v_x, v_y) will generate a new link at the next moment, that is, a representation of a s_{xy} corresponds to one link prediction method, temporal link prediction method (TLPM) can be represented by a mapping:

$$\mathrm{TLPM}: G' \to S$$

Where $G' \subset G$ is historical network topology information that has been observed, and contains multiple continuous network topological graphs; S represents the link probability matrix, which is a guess about network topology in the future. Lv Linyuan and Zhou Tao [8] summed up different link prediction algorithms, and this paper selects common neighbors (*CN*) index, Adamic-Adar (*AA*) index, priority link (*PA*) index, *Katz* index as a reference.

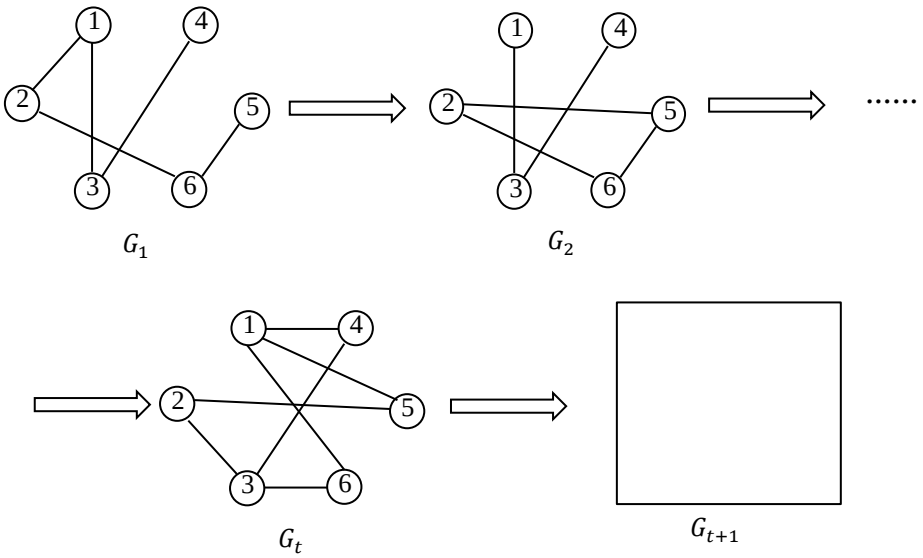

Fig. 1: temporal link prediction diagram of undirected simple network

3. Algorithm

3.1 Basic concepts

Definition 1. Given a probability space $(\Omega, \mathfrak{F}, P)$ and an index set $\mathbb{T}(\subset \mathbf{R}^1)$, a stochastic process is a family of random variables $\{X(t), t \in \mathbb{T}\}$ or $\{X_t\}_{t\in\mathbb{T}}$, and it also can be written as $\{X(w,t), t \in \mathbb{T}\}$ to reflect that it is actually a function of two variables, $t \in \mathbb{T}$ and $w \in \Omega$.

$$A_{\propto} = \{x : \mu_{\tilde{A}}(x) \geq \propto \ \forall x \in X\}.$$

In many problems from the natural sciences a point $t \in \mathbb{T}$ has the meaning of time, so $X(t)$ is random variable representing a value observed at time t. The random variables $X(w,t)$ take values in the space E, which is called the state space of the stochastic process [16].

Definition 2. $\{X_t, t = 0,1,2,\cdots\}$ is a random sequence defined in the probability space $(\Omega, \mathfrak{F}, P)$, if any negative integer t and the state $i_0, \cdots, i_t, i_{t+1} \in E$, have

$$P(X_{t+1} = i_{t+1} | X_t = i_t, X_{t-1} = i_{t-1}, \cdots, X_0 = i_0) = P(X_{t+1} = j | X_t = i_t)$$

then the random sequence $\{X_t, t = 0,1,2,\cdots\}$ is called a Markov chain, recorded as $\{X_t, t \geq 0\}$.

The conditional distribution of any future state i_{t+1} given the past states $i_0, i_1, \cdots, i_{t-1}$ and present state i_t, is independent of the past states and depends on the present state only [17].

The probability $P(X_{t+1} = j | X_t = i) = p_{ij}$ represents the probability that the process will make a transition to state i given that currently the process is in state j. Clearly one has

$$\begin{cases} 0 \leq p_{ij} \leq 1, \ i,j = 0,1,2,\cdots,n \\ \sum_{j=0}^{\infty} p_{ij} = 1, \ \ i,j = 0,1,2,\cdots,n \end{cases}$$

where n is the number of nodes.

Definition 3. The matrix containing p_{ij},the transition probabilities

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

is called the one-step transition probability matrix of the process.

Definition 4. Define the l-step transition probability

$$p_{ij}^{(l)} = P(X_{t+l} = j | X_t = i)$$

to be the probability that a process in state j will be in state i after l additional transitions.

The matrix $P^l = (p_{ij}^{(l)})_{n\times n}$ is the l-step transition probability matrix.

3.2 Local random walk

The concept of random walk first appeared in Karl Pearson's article in Nature, where Pearson described the random walk problem and gave the corresponding solution, then Lord Rayleigh extended the random walk. Since then random walk has been widely used in many research fields. In fact, the random walk

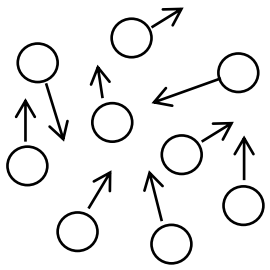

Fig. 2: Brownian movement of molecules in the air

phenomenon is common in life, which has been reflected in many areas, such as the Brownian movement in physics; the simulation of biological individual activities in ecology; Markov chain Monte Carlo algorithm in computer science etc.. Figure 2 shows the Brownian motion of molecules in the air.

Random walk is a Markov chain that describes the random walk particle access node sequence. At any node, it is assumed that the random walk particle randomly selects the edge adjacent to the next vertex with a certain probability at each step, or returns to the initial position with a certain probability. The global random walk algorithm is often computationally complex and therefore difficult to be applied to large-scale networks. While the local random walk has a smaller complexity, it is suitable for more real network with high complexity. Liu Weiping and Lvlin Yuan [18] proposed a local random walk (*LRW*) based on the local random walk of the network, which only considers the random walk process with finite number of steps.

At time t, consider an undirected simple network $G_t = (V_t, E_t, A_t)$, where $|V_t| = n$ is the set of nodes, $|E_t| = m$ is the set of links, and A_t is the adjacency matrix. Multiple links and self-connections are not allowed. Random walk is a Markov chain that describes the sequence of nodes visited by a random walker. This process can be described by the transition probability matrix P, with $P_{xy} = a_{xy}/k_x$ presenting the probability that a random walker staying at node v_x will walk to v_y in the next step, where a_{xy} equals 1 if node v_x and node v_y are connected, 0 otherwise, and k_x denotes the degree of node v_x. Given a random walker starting from node v_x, denoting by $\pi_{xy}(l)$ the probability that this walker locates at node v_y after l steps, we have

$$\vec{\pi}_x(l) = P'\vec{\pi}_x(l-1), l \geq 0$$

Where $\vec{\pi}_x(0)$ is an $N \times 1$ vector with the element equal to 1 and others to 0. The initial resource is usually assigned according to the importance of nodes. Here we simply set the initial resource q_x of node v_x proportional to its degree k_x, that is $q_x = k_x/2m$. Then after normalization the similarity between node v_x and node v_y is

$$s_{xy}^{\mathrm{LRW}}(l) = q_x\pi_{xy}(l) + q_y\pi_{yx}(l)$$

3.3 Time series random walk algorithm

In the actual complex network, the relationship between nodes will keep changing over time, the probability of nodes' connection will also keep changing, if link prediction can detect this dynamic change, then it can improve the prediction accuracy, so it is necessary to consider the effect of time variation on link prediction.

In this work, we exploit temporal and topological information to predict potential links. In our proposed time series random walk method(*TS-RWM)*, we first compute the Markov probability transfer matrix $\overline{P}_{t_0}$, $\overline{P}_{t_0+1}$,…, $\overline{P}_{t_0+T-1}$ for the given temporal networks $\mathrm{G_{t_0}}, \mathrm{G_{t_0+1}}, \cdots, \mathrm{G_{t_0+T-1}}$ with window size T, then combine them into a transformation matrix $\tilde{P}$, and apply the local random walk algorithm to obtain the final prediction result. In the evolution of the temporal network, recent snapshots are more reliable for future link prediction; they should be emphasized to obtain more accurate prediction results. In our algorithm, a damping factor is used to assign greater importance to more recent information. Based on the damping factor $\gamma(0 < \gamma < 1)$, transformation matrix $\tilde{P}$ is defined as

$$\tilde{P} = \sum_{t=t_0}^{t_0+T-1} \gamma^{t_0+T-1-t}\,\overline{P}_t$$

Then it applies the local random walk based on transformation matrix $\tilde{P}$ to obtain the similarity score. The framework of algorithm *TS-RWM* is as follows.

Algorithm *TS-RWM*(link prediction in temporal networks)
Input:
$G_t = (V, E_t, A_t)(\ t = t_0, t_0 + 1, \cdots, t_0 + T - 1)$: sequence of networks;
A_t: the adjacent matrix of G_t;
γ : damping factor $(0 < \gamma < 1)$;
$1 - \lambda$: the probability of random walking of particles back to the initial position;
$steps$: steps of random walk;
Output:
S: time series random walk similarity matrix;
Begin

DOI https://doi.org/2024-JICS-22468 | Generated on 2025-04-20 05:55:20

1. Initial $\tilde{P}$ as a zero $n \times n$ matrix, and initial S as a unit matrix I;
2. For $t = t_0$ to $t_0 + T - 1$:
   Execute the Markov probability transfer matrix $\overline{P}_t$;
   $\tilde{P} = \gamma\tilde{P} + \overline{P}_t$;
   End
3. While $(stepi < steps)$
   Calculate LRW index: $S = (1 - \lambda)I + \lambda\tilde{P}'S$;
   $stepi = stepi + 1$;

End
S = S + S';
End

4. Experiment

4.1. Dataset

In this paper, we use two different data sets to evaluate the proposed algorithm, namely the Enron E-mail dataset and the Hep-th dataset (High-energy particle physics coauthor-ship dataset).

The Enron data set, which originated from Enron's internal mail contact network, is the most widely used public data set in email related research. Its data is collected from mails of 150 Enron senior managers. This data set contains 252,759 e-mails, we select 21,254 e-mails during the period from May 1999 to June 2002 for link prediction analysis. The nodes of the network are the mail address of 150 employees. An edge (i, j) represents that there has been at least one e-mail communication between i and j (either i sending at least one e-mail with recipients including j, or j sending at least one e-mail with recipients including i). We performed the link prediction analysis mainly on the monthly e-mail graphs.

The Hep-th dataset is based on data from the arXiv archive, which is widely used in the study of dynamic structure of complex network.. The complete data set contains 29,555 papers by 9,200 authors with 87,794 coauthor-ship relationships during the period from 1992 to 2003, and we select 1,172 papers for link prediction analysis. The nodes of the network are authors. An edge (i, j) represents that there has been at least one paper coauthored by i and j. We performed the link prediction analysis on the quarterly coauthor-ship graphs.

There are three standard metrics AUC, Precision and Ranking Score to quantify the accuracy of link prediction algorithms. AUC evaluates the overall ranking resulted from the algorithm [19]; Precision focuses on the top-L candidates [20]; Ranking Score is more concerned with the sorting of the predicted edges [21]. This paper uses the AUC index to measure the accuracy of the link prediction algorithm.

AUC can be interpreted as the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link. In the implementation, among n independent comparisons, if there are n' times the missing link having a higher score and n'' times being of the same score, we have

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

If all the scores are generated from an independent and identical distribution, the AUC should be about 0.5. Therefore, the degree to which the AUC exceeds 0.5 indicates how much better the algorithm performs than pure chance.

The training set and the test set are selected by moving window slice algorithm to calculate the AUC values (as shown in Fig. 3). First, we select the network snapshot of the pre-T time as the training set, and the network snapshot is used as the test set at time $T + 1$, then calculate the AUC value, and move the window size back one unit for the next AUC value calculation, finally take the average of all AUC values.

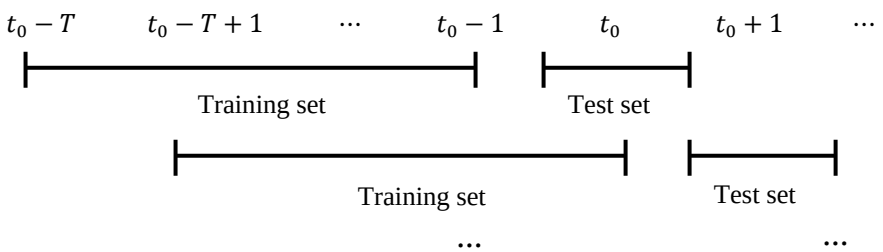

Fig. 3: Selection of training set and test set

The AUC values at each time were calculated by using the common neighborhood (*CN*), Adamic-Adar (*AA*), preferred attachment(*PA*) and *Katz(KZ)* index for each time, and we took the average of all AUC values as the prediction results, and then compared them with the proposed algorithm.

4.2. Experimental results and analysis

Experiment 1 overall performance verification of the algorithm

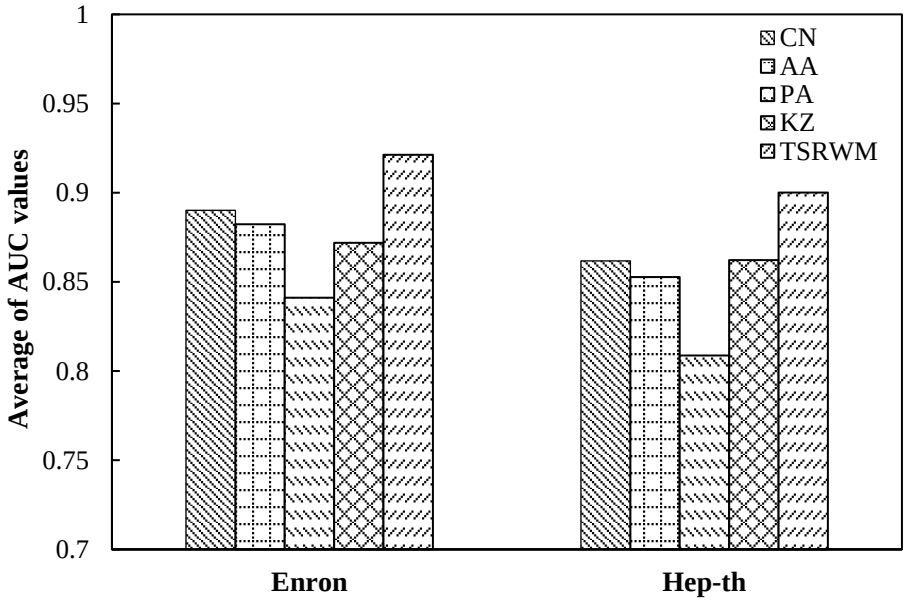

Fig. 4: Performance of TS-RWM and other methods

In this experiments, we tested and compared the performance of proposed time series method *TS-RWM* with that of methods *CN*, *AA*, *PA* and *Katz* on the two datasets, and results were obtained by fixing window size $T = 20$ and damping factor $\gamma = 0.8$. Fig. 4 shows the AUC values of results by *TS-RWM* and the AUC values of the results by other methods. From the figure, we can see clearly that *CN*, *AA* and *Katz* have similar effects and *PA* is the worst, while the results obtained by the *TS-RWM* model are better than those obtained by the previous four methods. In addition, the results of the Enron dataset are better than the results of the Hep-th dataset, which may be due to the fact that the Hep-th dataset is a bit sparse, and more papers will be considered for further research to improve prediction accuracy.

Experiment 2 The effect of window size

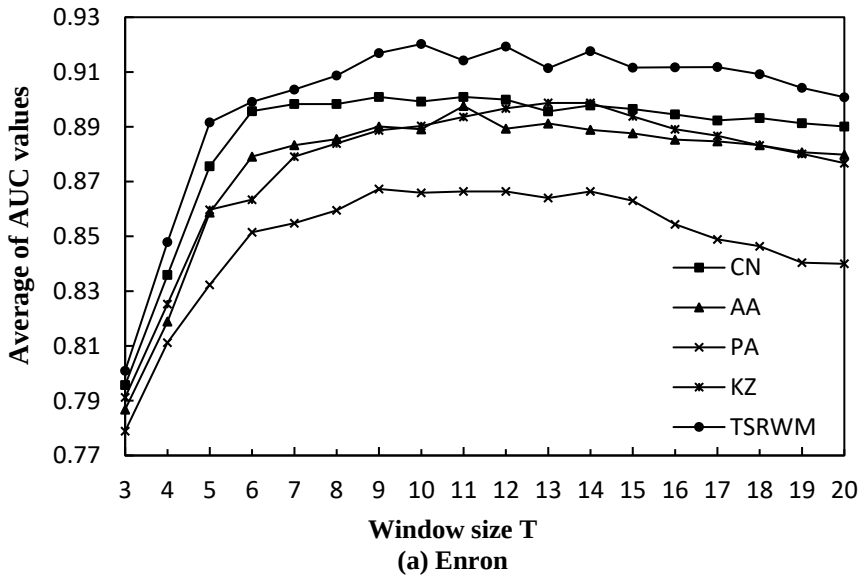

Fig. 5: The effect of window size

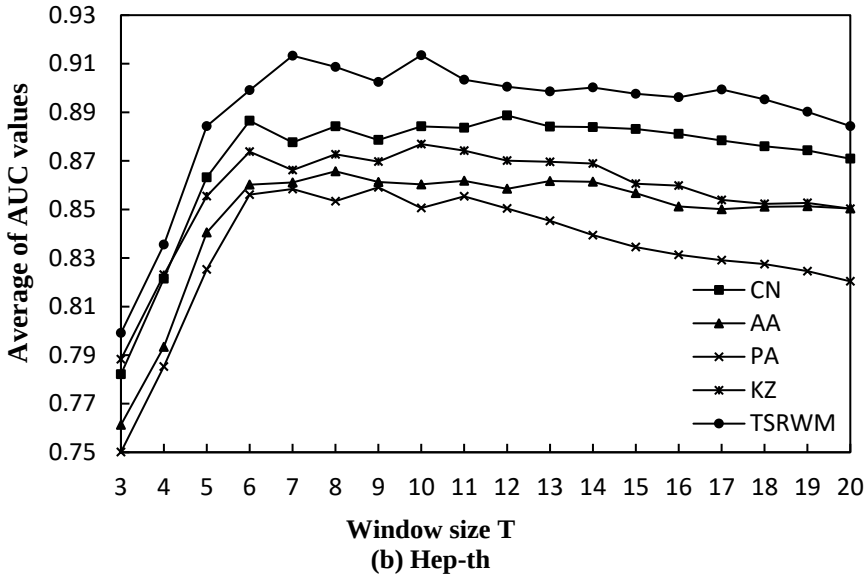

Fig. 6: The effect of window size

Window size T is an important parameter in temporal link prediction. We tested our method by varying the value of window size T and fixing the damping factor $\gamma = 0.8$, the experimental results are shown in Figure 5 and Figure 6. As can be seen from the figures, the performance of this algorithm is better than other algorithms. As the window size increases, the AUC values corresponding to different algorithms will increase first, and then there will be a relatively stable phase. When the window size continues to increase, the AUC value of some algorithms will decrease. This shows that the prediction accuracy of the algorithm is related to the window size. At the beginning, as the window size increases, the algorithm can use more historical information, so the prediction accuracy of the algorithm will gradually increase. However, when the window size increases to a certain extent, the distance between the initial time and the predicted time is far, too much historical information will have a negative impact, the accuracy of the algorithm will decrease. Therefore, it is important to find a suitable window size for the algorithm.

Experiment 3 The effect of damping factor

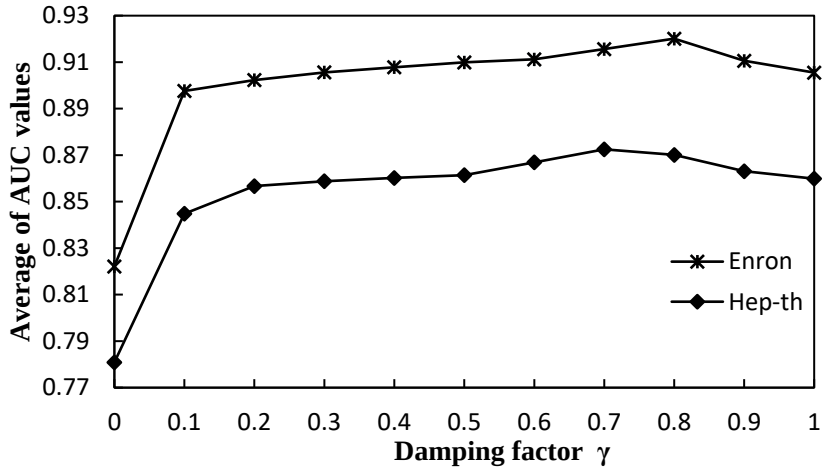

Fig. 7: The effect of damping factor

Damping factor γ is an important parameter in this paper. We tested our method by varying the value of the damping factor γ and fixing window size $T = 7$, the experimental results are shown in Figure 7. From the figure, we see that when $\gamma = 0$, the algorithm achieved the lowest performance on the two datasets. If $\gamma = 0$, only the latest snapshot is used to predict its next graph, which shows the importance of historical information. The AUC values improves when historical information is introduced. After the AUC value has reached the maximum, it deteriorates when the value of γ continues to increase. This indicates that damping factor γ has an impact on the prediction accuracy. Because the window size is 7, which is relatively small, all

inclusive historical information is useful, and the following AUC value declines more slowly.. In further experiments, we can vary the value of the window size T and do more experiments.

5. Conclusion

This paper summarizes some methods of static link prediction and temporal link prediction, which have received widespread attention in recent years. Based on the static link prediction algorithm based on local random walk, a time series random walk algorithm is proposed to solve the problem of time series link prediction in the non-directional network. Based on local random walk, this paper proposed a time-series random walk algorithm, first computed the Markov probability transfer matrix at each time, then combined them into a transformation matrix, and applied the local random walk algorithm to obtain the final prediction result. The experimental results on real datasets-- Enron E-mail dataset and High-energy particle physics coauthor-ship dataset show that the proposed algorithm can effectively improve link prediction accuracy. In addition, the effect of window size and damping factor on experimental results is analyzed.

The next research focuses on two aspects: (1) the applicability of the proposed method to directed, weighted networks; (2) how to predict the relationship between nodes for networks with multiple types of links.

6. References

[1] Wang, P. et al. Link prediction in social networks: the-state-of-the-art. Science China Information Science, 2015, 58: 1-38.

[2] Chengwei Lei, Jianhua Ruan. A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. Bioinformatics original paper, 2013, 29(3):355–364.

[3] Yuriy H, Ryan W. Solava, Tijana M. Revealing Missing Parts of the Interactome via Link Prediction. PLoS ONE, 2014, 9(3): e90073.

[4] Nazpar Yazdanfar, Alex Thomo. LINK RECOMMENDER: Collaborative-Filtering for Recommending URLs to Twitter Users. Procedia Computer Science,2013,19:412-419.

[5] Chiluka N,Andrade N,and Pouwelse J.A link prediction approach to recommendations in large-scale user-generated content systems. Proceedings of the 33rd European conference on Advances in Information Retrieval, Ireland, 2011: 189-200.

[6] Bulent Ozel. Link and Node Analysis of Gender Based Collaborations in Turkish Social Science. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Minings. IEEE Computer Society. 2012:15-19.

[7] Dimitrios T, Serafeim P. Decomposing multilayer transportation networks using complex network analysis: a case study for the Greek aviation network. Journal of Complex Networks, 2015, 3: 642 -670.

[8] Linyuan Lü, Tao Zhou. Link prediction in complex networks: A survey . Physical A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[9] Lin Yao, Luning Wang, Lv Pan, Kai Yao. Link Prediction Based on Common-Neighbors for Dynamic Social Network. Procedia Computer Science, 2016, 83:82-89.

[10] Huang Z，Lin D K J．The time-series link prediction problem with applications in communication surveillance.INFORMS Journal on Computing，2009，21(2): 286-303．

[11] Bliss C A, Frank M R, Danforth C M, Dodds P S. An evolutionary algorithm approach to link prediction in dynamic social networks.Journal of Computational Science, 2014,5(5):750-764.

[12] Dunlavy D M, Kolda T G, and Acar E. Temporal link prediction using matrix and tensor factorizations. ACM Transactions on Knowledge Discovery from Data, 2011, 5(2):1-27.

[13] Behnaz Moradabadi, Mohammad Reza Meybodi. A novel time series link prediction method: Learning automata approach. Physica A, 2017, 482:422-432.

[14] Deng Zhi-hong, Lao Song-yang, Bai Liang. A Temporal link prediction algorithm based on link prediction error correction. Journal of Electronics& Information Technology,2014,36(2):325-331．

[15] Ahmed, N.M., Chen, L.An efficient algorithm for link prediction in temporal uncertain social networks.Information Sciences, 2016,331, pp. 120-136.

[16] Tian Zheng, Qin Chaoying et al. Random Process and Application . Beijing: Science Press, 2007.

[17] W.-K. Ching et al., Markov Chains, International Series in Operations Research&Management Science 189, DOI 10.1007/978-1-4614-6312-2

[18] Liu W, Lu L. Link prediction based on local random walk .Euro physics Letters, 2010, 89(5): 58007-580013.

[19] HANELY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982, 143: 29-36.

[20] HERLOCKER J L, KONSTANN J A, TERVEEN K, et al. Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst, 2004, 22(1): 5-53.

[21] HOU T,REN J,MEDO M, et al. Bipartite network projection and personal recommendation. Phys Rev E, 2007, 76: 046115.