Temporal Link Prediction: Techniques and Challenges

Sogol Haghani
Department of Computer Engineering and
Data mining laboratory
Tehran, Iran
e-mail: s.haghani@student.alzahra.ac.ir

Mohammad Reza Keyvanpour
Alzahra University
Tehran, Iran
e-mail: keyvanpour@alzahra.ac.ir

ABSTRACT
Link prediction is a problem of interest in many application domains. Various researches have been done about utilizing information on the networks entities and relations. Temporal Link prediction is a task of predicting unseen links that will be formed in the future based on different snapshots of the network. It has various challenges. In this paper we propose a framework, in which the concerns and challenges of temporal link prediction are discussed, and then all major and recent techniques are classified, and also the presented techniques based on the discussed challenges are analyzed. This framework provides an accurate resolution of different challenges and techniques which is essential to understand various aspects of the problem. We think that this paper can help future research in a more efficient way.

Keywords
Link Prediction, Dynamic Network, Graph, Benefits, Challenges.

1. INTRODUCTION
In the form of relational data, one often wants to be able to predict the behavior or temporal interactions between two entities [5] [7] . Network can be visualized as a graph in which nodes represent entities and edges corresponding to links or interactions [3] [11]. One of the issues in relational data is Link prediction problem [11] [3] [12]. In fact, link prediction is categorized in two tasks: Missing Link prediction and Temporal Link prediction [5]. The goal of missing link prediction is to predict missing link with a given current state of network in order to describe a more complete picture of the overall network [5]. Moreover, in temporal link prediction it is assumed that the network is complete and the aim is to predict link, which would appear in the next state of the network [15] [7]. One of the notable characteristics in the dynamic network in link prediction task is increasing and decreasing links and entities over time [17]. The dynamic of such a network makes the study of these graphs, is a challenging task. In addition, a large dynamic network may be complicated by multi-relational data[4]. Most of the time networks are heterogeneous, and dealing with multiple links and nodes, they may be complicated. Another challenge is about the sparsity of linked data [9]. When a network is sparse, it is sensitive to noise, and in a dynamic and heterogeneous network noise rate can be easily changed with time [16]. Nonlinear transformations over time are commonly seen in dynamic networks with seasonal fluctuations [15] [17]. Catching such a nonlinearity is expensive and time-consuming.

Wide range of techniques are presented to overcome some of the mentioned challenges. However, choosing an appropriate method could be challenged because there are already trouble in dynamic network link prediction. We introduce an analytical framework, which discussed and categorized temporal link prediction challenges and classified link prediction techniques. Moreover, we evaluate each technique based on our discussed challenges. The rest of the paper is organized as follow: in the next section, we review the related works. Section 3, describes dynamic network. In Section 4, the formal definition of the link prediction problem is presented. In Section 5, the first component of analytical framework is presented. Section 6 includes challenges as the second component of analytical framework and in Section 7 we describe the third component of framework. Section 8 includes conclusion.

2. RELATED WORK
There are already several outstanding surveys on link prediction problem [11] [3]. An early study of predicting links is proposed by Liben [11], which provide useful information and insight for link prediction problem. Lu and Zhu [12] summarize popular link prediction algorithms for complex network. However, they emphasize the contributions from physical perspective, instead of computer science perspective. In addition, complex network is a more abstract model than a dynamic real world network. Another line of research, Hasan and Zaki [2], investigate feature-based link prediction, probabilistic models, graphical models and linear algebraic models as the most used approaches in link prediction. However, most of these models are designed for static homogeneous network and they are not about temporal link prediction. Mishara and Dhote [18] present a brief study of various temporal link prediction techniques in online social network. This work covers studies on co-authorship network. In conjunction with all the above studies, we present a novel framework which focus on temporal link prediction. We categorize temporal link prediction challenges and present a comprehensive classification for temporal link prediction techniques. Also, this framework analyzes the presented classification based on the discussed challenges.

3. DYNAMIC NETWORK
Formally, networks include a set of entities. The other elements that make sense to the network are relations and interaction. For instance, in Facebook, each user is viewed as an entity, and the link types are friendship, follower, liking on a post etc [3] [2] [6] . In fact, net-

work structure is a graph, the nodes of which represent entities and edges corresponding to relations and interactions between them [12]. Data structure in graph depends on a number of links between two nodes. It means that, in single-relational data, it has just one type of link between the network and in multi-relational data; the type of links is more than one. Almost in every network, all the entities are not available in and by the time they are added to it. The formal expression of a dynamic network is a network, in which the rates of changes at the same time intervals are not the same and always change [9]. In other words, nodes and edges shrink and grow quickly. Such a dynamic network can typically be modeled as a dynamic graph $G = \langle V, E \rangle$, where V is the set of nodes, and $E : N^+ \rightarrow 2^{V \times V}$ is a dynamic edge function assigned to each round $r \in N^+$ by a set of edges $E(r)$ for that round. A round is occurring between two times; round $r \in N^+$ occurs between the times $r-1$ and r. $G(r) = \langle V, E(r) \rangle$ is instantaneous in round r. In the literature, such dynamic graphs have also been termed as evolving graphs [9] [14]. Figure 1 is a simple description of homogeneous and heterogeneous dynamic networks.

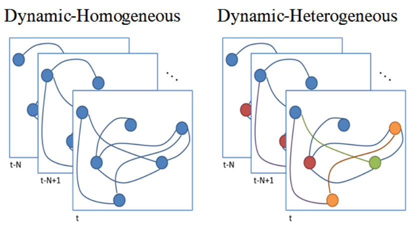

Figure 1. An overview of dynamic network

4. PROBLEM DEFINITION

Link prediction is a task of forecasting relations in a network. Predicting of unknown links falls into two categories in accordance with the linked data: (i) Missing Link Prediction and (ii) Temporal Link prediction [16]. Missing Link prediction is a task of predicting an unseen link with the current state of network, in order to complete the network [5]. In this paper, we focus on the temporal link prediction.

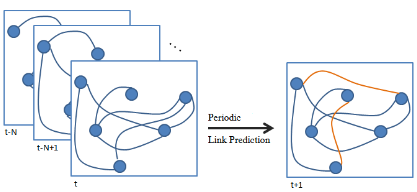

Figure 2. Temporal link prediction

4.1 Temporal Link prediction

Given a series of snapshots $\{G_{t-N}, G_{t-N+1}, \ldots, G_t\}$ of an evolving graph $G_t = \langle V, E_t \rangle$ in which each $e = (u, v) \in E_t$ represents a link between u and v that took place at a particular time t. We seek to predict the most likely link state in the next time step G_{t+1}. The temporal link prediction also called a *periodic link prediction* when $N \gg 0$ [7] [15](Fig. 2). In almost every method that we study in this paper, assume that nodes V remain the same at all time steps but edges E_t changes for each time t. In this paper we focus on this type of link prediction.

5. TEMPORAL LINK PREDICTION TECHNIQUES

Classification of temporal link prediction techniques is the first component of the proposed framework. Review of a wide range of methods offered in this paper represents that temporal link prediction techniques can be classified as shown in Figure 3. Next, we describe each technique in details.

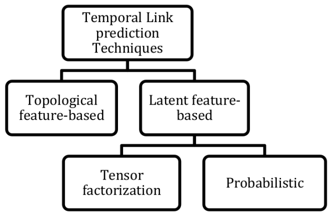

Figure 3. Temporal link prediction techniques

5.1 Topological feature-based

The most widespread techniques are based on topological patterns. Basically, they consist of applying topological similarity metrics to pairs of nodes in the network at a time t in order to predict if a link will occur at a time $t_0 (t_0 > t)$. Such metrics provide scores to each pair of nodes which are then used to perform the prediction task either by an unsupervised or a supervised technique. Table 1 summarizes some of the well-known topological metrics. These topological metrics calculate a score function which demonstrates the similarity of two nodes in different ways [18][4][10]. The pairs of non-connected nodes are ranked by their similarity scores and the top ranked ones are predicted to be connected in an unsupervised manner. In the supervised strategy, the link prediction is treated as a classification problem. In this approach, pairs of nodes are assigned to the positive class if they are connected, or the negative class otherwise. The similarity scores related to the chosen set of topological metrics are adopted in this approach as features which are used by a supervised classifier to perform the link prediction task [5] [2] [11].

5.2 Latent feature-based

An underlying assumption of a latent feature-based model is to build a model, which can discover the latent features from the structure of the graph, especially the evolutionary pattern. Nevertheless, the researchers in this domain believe that the structure of the graph and the combination of these two types of information have a latent characteristic, which cannot be achieved with simple techniques [15]. The goal of the latent feature-based approach is to learn a model from observed dynamic links that can predict the values of unobserved entries. The latent representation of each node corresponds to a point on the surface of a unit hypersphere [5].

5.2.1 Tensor factorization

Significantly, tensor factorization is known as an approach for structured data in different learning contexts. The success of tensor factorization in link prediction problem is due to the fact that it has a high ability to model and analyze the relational data [13]. Tensor-based methods usually consist of in two (Matrix [13]) and three orders. In Temporal link prediction, the third domain is considered as a different time snapshot. Hereupon, this approach has a reasonable ability to detect latent features over time. More formally, given a

Table 1. Popular similarity indices in the link prediction. The set $\Gamma(x)$ consists of the neighbors of x in G_t.

Similarity Metrics	Score Function	Pros & Cons
Common neighbors	$\|\Gamma(x) \cap \Gamma(y)\|$	Reasonable result on most data set Not Normalized Scalable Cold start
Jaccard Coefficeint	$\frac{\|\Gamma(x) \cap \Gamma(y)\|}{\|\Gamma(x) \cup \Gamma(y)\|}$	Normalized Poor Performance
Adamic/Adar	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log \|\Gamma(z)\|}$	Use Minor Information Reasonable performance
Preferential Attachment	$\|\Gamma(x)\|.\|\Gamma(y)\|$	Work well in evolutionary network It has least Computational Complexity
Katz	$\sum_{l=1}^{\infty} \beta^l \cdot \|path_{x,y}^l\| = \beta A + \beta^2 A^2 + \beta^3 A^3 + \cdots$	Unscalable Work well in weighted networks High Computational Complexity
SimRank	$\begin{cases} 1 & \text{if } x{=}y \\ \gamma . \frac{\sum_{a \in \gamma(x)} \sum_{b \in \gamma(y)} simRank(a,b)}{\|\Gamma(x)\|.\|\Gamma(y)\|} & \text{otherwise} \end{cases}$	The similarity score considers both the graph topological and attribute based similarity. Unscalable Computational feasibility
Rooted Page Rank	$(1-\beta)(I-\beta N)^{-1}$	Random walks on a graph Effective for capturing various relations Well defined for any given graph

sequence of words:

$$z(i,j,t) = \begin{cases} 1 & \text{node } i \text{ links to node } j \text{ at time } t. \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

which shows that the link from node i to j was appeared at time t. Factorization techniques like the CP (CanDecomp/Parafact) and Tucker model can be considered as a higher-order generalization of the matrix singular value decomposition (SVD) and principle component analysis (PCA). However, the CP model is more advantageous in terms of interoperability, uniqueness of solutions and parameter determination. A three dimensional tensor $\underline{\mathbf{Z}}$ is defined as $m * n * t$, its k-component CP factorization is defined as:

$$\sum_{k=1}^{k} \lambda_k a_k \circ b_k \circ c_k \quad (2)$$

Symbol $\circ$ stands for an outer product, $\lambda_k \in \mathbb{R}^+, a_k \in \mathbb{R}^m, b_k \in \mathbb{R}^n, c_k \in \mathbb{R}^t$, where $k = 1, \cdots, K$. Each summand $(\lambda_k a_k \circ b_k \circ c_k)$ is called a component, each vector is called a factor [19].

5.2.2 *Probabilistic*

In probabilistic model, it is assumed that the probability of a link between two nodes depends only on their latent positions. In principle, given a dynamic graph $\{G_{t-N}, G_{t-N+1}, \ldots, G_t\}$, it is needed Z_{t+1} to predict the future graph G_{t+1}. Z_{t+1} is the temporal latent space representation for the node V_{t+1} at time t. Z_{t+1} is approximated by Z_{t-N}, Z_{t-N+1} , ... , Z_t. Therefore, given a dynamic graph$G = \{G_{t-N}, G_{t-N+1}, \ldots, G_t\}$, it can be inferred Z_{t-N}, Z_{t-N+1} , ... , Z_t based on $G_{t-N}, G_{t-N+1}, \ldots, G_t$, and use Z_{t-N}, Z_{t-N+1} , ... , Z_t to approximate Z_{t+1} and to predict G_{t+1} [17] [15] [16] . Probabilistic models have a high ability to explore the evolutionary pattern, especially seasonal fluctuations [16].

6. CHALLENGES

There are some challenges in link prediction regarding its requirement. We classify the link prediction challenges in three categories: *dynamic network challenges*, *linked data challenges*, and *link prediction specific challenges*(Fig. 4). In this categorization, not only dynamic network challenges are dependent on link prediction, but their challenges are also primary for link prediction problem.

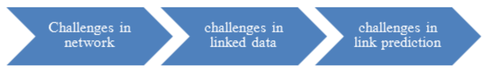

Figure 4. Link prediction Challenges

6.1 Dynamic network challenges

One of the main concern in a dynamic network is dealing with node failures. This phenomenon is a serious challenge in temporal link prediction when in different snapshots of graphs nodes joining and leaving the network. Unfortunately, most of the methods ignore this and assume that nodes V remain the same at all time steps [9] [4]. Unfortunately, if ignore the dynamic nature of the nodes and assume that all nodes in all snapshots are presented, then the information is missed which corresponds to creating a new link due to appearance of one of its nodes, and vice versa [8]. Large dynamic networks may be complicated by the high dimensionality of responses, the large number of observations and the complexity of choices to be made among the explanatory variable [1].

6.2 Linked-data challenges

Mining linked data is about how to effectively and efficiently utilize information from both nodes attribute and link structure. Many existing models seek to represent link structures using the selected statistic on networks, then combine these statistics with node attributes [1]. Copping with noisy data is a main concern in link data. In some networks, relations can be missing or be invalid [10]. On the other hand, sparse dynamic networks are sensitive to noise. Precisely, noise-to-signal ratio can be easily changed on sparse networks [13].

6.3 Link prediction challenges

Seasonal fluctuation and ill-behaved node are two phenomena that skew the prediction proceeding. In seasonal fluctuations nodes add or delete a massive number of connections in a very short period of time, may indicate something significant, such as a shocking event happening in a social network, or a serious detection of cancer [17]. Ill-behaved nodes referring to nodes with random behavior, usually provide less information [17]. They randomly start and end relationship with other nodes, which sabotages the network stableness. The point is, that this circumstance has a nonlinear transformation over time and is commonly considered in dynamic networks, which has an expensive cost for catching this nonlinearity [16].

7. ANALYZING TEMPORAL LINK PREDICTION TECHNIQUES

This section is the third component of our analytical framework. We evaluate temporal link prediction techniques based on the introduced challenges. Our evaluation is summarized in Table 2. As observed in Table 2 the columns heading are challenges and the rows headings are the techniques, which are based on the classification tree (Fig. 2). In table 2 we want to express the strength and weakness of models from different points of view.

Table 2: Analytical framework for evaluation of ink prediction models

	Feature-base Model	Tensor Factorization	Probabilistic Model
Node-failure	Average	Average	Low
Scalability	High	Low	High
Noise-resident	Low	Average	Average
Ill-behaved node	Low	High	Average
Seasonal flections	Low	High	High

8. CONCLUSION

This work was motivated by our requirements to understanding the potential advantages of temporal link prediction techniques compared to each other. In this paper, we have presented a triplex analytical framework for temporal link prediction and to illustrated that there are different challenges and techniques. We classify the link prediction techniques from different points of view. We have tried to collect all the current major works done and then evaluate this classification based on the presented evaluation criteria. This framework could lead to future works.

REFERENCES

[1] Aggarwal, Charu and Subbian, Karthik, "Evolutionary network analysis: A survey", *ACM Computing Surveys (CSUR)* , pp. 10, 2014

[2] Al Hasan, Mohammad and Zaki, Mohammed J, "A survey of link prediction in social networks", *Social network data analytics*, pp. 243-275, 2011.

[3] Al Hasan, Mohammad and Chaoji, Vineet and Salem, Saeed and Zaki, Mohammed, "Link prediction using supervised learning", *SDM06: workshop on link analysis, counter-terrorism and security*, 2006.

[4] Bliss, Catherine A and Frank, Morgan R and Danforth, Christopher M and Dodds, Peter Sheridan, "An evolutionary algorithm approach to link prediction in dynamic social networks", *Journal of Computational Science*, pp. 750-764, 2014.

[5] Dunlavy, Daniel M and Kolda, Tamara G and Acar, Evrim, "Temporal link prediction using matrix and tensor factorizations", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, pp. 10, 2011.

[6] Getoor, Lise and Diehl, Christopher P, "Link mining: a survey", *ACM SIGKDD Explorations Newsletter*, pp. 3-12, 2005.

[7] Heaukulani, Creighton and Ghahramani, Zoubin, "Dynamic Probabilistic Models for Latent Feature Propagation in Social Networks", *ICML (1)*, pp. 275-283, 2013.

[8] Keyvanpour, Mohammad Reza and Azizani, Fereshteh, "Classification and analysis of frequent subgraphs mining algorithms", *Journal of Software*, pp. 220-227, 2012.

[9] Kuhn, Fabian and Oshman, Rote, "Dynamic networks: models and algorithms", *ACM SIGACT News*, pp. 82-96, 2011.

[10] Lee, Conrad and Nick, Bobo and Brandes, Ulrik and Cunningham, Pádraig, "Link prediction with social vector clocks", *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 784-792, 2013.

[11] Liben-Nowell, David and Kleinberg, Jon, "The link-prediction problem for social networks", *Journal of the American society for information science and technology*, pp. 1019-1031, 2006.

[12] Lü, Linyuan and Zhou, Tao, "Link prediction in complex networks: A survey", *Physica A: Statistical Mechanics and its Applications*,pp. 1150-1170, 2010.

[13] Menon, Aditya Krishna and Elkan, Charles, "Link prediction via matrix factorization", *joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 437-452, 2011.

[14] Ngonmang, Blaise and Viennet, Emmanuel and Tchuente, Maurice and Kamga, Vanessa, "Community Analysis and Link Prediction in Dynamic Social Networks", *Computing in Research and Development in Africa*, pp. 83-101, 2015

[15] Sarkar, Purnamrita and Chakrabarti, Deepayan and Jordan, Michael, "Nonparametric link prediction in dynamic networks", *arXiv preprint arXiv:1206.6394*, 2012.

[16] Sarkar, Purnamrita and Chakrabarti, Deepayan and Jordan, Michael and others, "Nonparametric link prediction in large scale dynamic networks", *Electronic Journal of Statistics*, pp. 2022-2065, 2014.

[17] von Winckel, G, "A deep learning approach to link prediction in dynamic networks", *SIAM*, 2014.

[18] Dhote, Yugchhaya and Mishra, Nishchol and Sharma, Sanjeev, "Survey and analysis of temporal link prediction in online social networks", *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pp. 1178-1183, 2013.

[19] Acar, Evrim and Dunlavy, Daniel M and Kolda, Tamara G, "Link prediction on evolving data using matrix and tensor factorizations", *2009 IEEE International Conference on Data Mining Workshops*, pp. 262-269, 2009