



# A Network Embedding Approach for Link Prediction in Dynamic Networks

Aswathy Divakaran<sup>(✉)</sup> and Anuraj Mohan

Department of Computer Science and Engineering, NSS College of Engineering, Palakkad, India  
aswathydiv36@gmail.com, anurajmohan@gmail.com

**Abstract.** Dynamic networks and their evolving nature have gained the attention of researchers with its ubiquitous applications in a variety of real-world scenarios. Learning the evolutionary behavior of such networks is directly related to link prediction problem as the addition and removal of links or edges over time leads to the network evolution. With the rise of large-scale dynamic networks like social networks, link prediction in such networks or otherwise temporal link prediction has become an interesting field of study. Existing techniques for enhancing the performance of temporal link prediction leverages the notion of matrix factorization, likelihood estimation, deep learning and time series based techniques. However, building a framework for temporal link prediction that preserves the non-linear varying temporal properties of dynamic networks is still an open challenge. Here, we propose a unified framework that incorporates Network Representation Learning (NRL) and time series analysis for temporal link prediction. Our experimental results on various real-world datasets show that the proposed framework outperforms the state-of-the-art works.

**Keywords:** Dynamic networks · Temporal networks · Link prediction · Network Representation Learning (NRL) · Time series

## 1 Introduction

In the past few years, there have been intensive researches dealing with the study of highly dynamic networks or temporal networks [1] whose topologies or characteristics change as a function of time. Almost all the real-world complex phenomena can be modeled as dynamic networks since they can model the evolving nature quite efficiently. For instance, social networks, communication networks, biological networks etc. have an underlying structure of dynamic networks where entities and relationships are relatively short and instantaneous. Recently, the evolutionary behavior of such networks gained the attention of researchers with its ubiquitous applications in a variety of real-world scenarios. Moreover, learning the evolutionary behavior is directly related to the link prediction problem [5] as the addition and removal of edges or links over time leads to the network evolution. With the rise of large-scale dynamic networks, link prediction in such networks also known as temporal link prediction has become an interesting field of study. The goal of this task is to predict the links in the network that would appear in its

future state of time under the assumption that the network is complete. Unlike missing link prediction in static networks, temporal link prediction is a challenging task driven by its ubiquitous applications in a variety of scenarios. Recommending new products in e-bay or amazon, friend suggestions in online social networks are some of the obvious examples. In biological networks, predicting the interactions between molecules at a specific time stamp can help us better understand the temporal interaction between them. This can provide useful temporal information that indicate the stage of a specific disease such as cancer. Therefore temporal link prediction plays an important role in disease prediction task. In addition, this task can be used to predict the academic collaborations in co-authorship and citation networks. Furthermore, temporal link prediction in terrorist communication networks help us to predict and capture the most important information related to the issue of national security.

The advancements in deep learning has shown its outstanding performance in various fields like financial services, health care, etc. to find better and faster decisions in today's data-driven world. The rapid growth of deep learning techniques extended its utility towards the area of social network analysis. Using the deep layers of non-linear transformation, deep learning integrated this field to better capture the non-linear varying temporal patterns in dynamic networks. Recent trends in exploring such patterns leverages the notion of Network Representation Learning (NRL) techniques [2–4] that embeds nodes in the network into a low-dimensional vector space by preserving structural proximities of the network. The key idea behind this technique is to generate continuous vector space representations for nodes in the network in such a way that the structural proximity is preserved. Such representations of real-world networks encode social relations in a continuous vector space and enables the original network to be exploited easily for further analysis. This lead to the emergence of various network embedding approaches for temporal link prediction rather than the computationally intensive Matrix Factorization (MF), Maximum Likelihood (ML) approaches. Furthermore, time series analysis is a well-studied area that aims at revealing significant statistics and characteristics of data. The key idea is to extract the underlying structure of the observed data. Time series can best capture the change over time under the assumption that past events are good starting points for the prediction of future. Time series forecasting aims at predicting the future scores based on the previously observed time series scores. Moreover, the frequently evolving nature of dynamic networks makes time series a promising option for temporal link prediction. Several works deployed time series forecasting for temporal link prediction [6–8].

In general, the movements to enhance the performance of temporal link prediction depends on the effectiveness in capturing the evolving nature of dynamic networks and extracting the non-linear varying temporal patterns. However, building a unified model that preserves all the complex non-linear varying patterns in dynamic networks is an open challenge. To address this challenge, we propose a unified framework that incorporates NRL techniques and time series analysis for link prediction in dynamic networks. Initially, we take snapshots of the network at regular intervals of time to capture the evolving nature. Inspired by NRL techniques, we extract the complex features in dynamic networks by preserving the network properties. This information is incorporated

into time series analysis where the time series for each node pair is constructed and future scores are predicted. Link prediction task is performed based on the predicted scores.

## 2 Problem Definition

This section provides a formal definition for temporal link prediction. “Let  $G = (V, E)$  be a dynamic network, where  $V$  is the set of vertices and each edge  $(u, v) \in E$  represents a link between  $u$  and  $v$ . Given the snapshots of  $G$  represented as  $G_1, G_2, \dots, G_t$  from time step 1 to  $t$ , how can we predict the network for a next time step  $G_{t+1}$ ?” Fig. 1 depicts an overview of temporal link prediction.

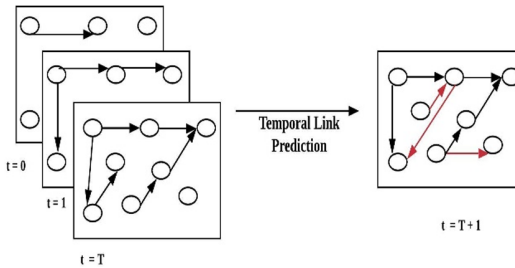


Fig. 1. Overview of temporal link prediction

## 3 Related Works

The literature in the field of temporal link prediction can be broadly classified into six based on the techniques used: Matrix Factorization (MF) models, probabilistic models, Deep Learning (DL) models, time series based models and others. MF or otherwise called matrix factorization techniques aims at decomposing a matrix into its factors and thereby makes complex operations easier. Majority of the works on matrix factorization based temporal link prediction deploy Non-negative Matrix Factorization (NMF) technique [13–16]. Probabilistic models deploy maximum likelihood approaches or probability distributions instead of fixed values. There exists several probabilistic models for temporal link prediction [17, 18]. A few works on temporal link prediction rests on spectral graph theory, which is the study of properties of a graph in relationship to the eigenvalues and eigenvectors associated with the graph [9, 10].

Time series based temporal link prediction deploys various time series forecasting models for predicting links in the network for a future time period. Time series score is constructed by computing various similarity measures between each node pairs in the network. Time series forecasting aims at predicting the future scores based on the previously observed time series scores. Time series based temporal link prediction frameworks take the adjacency and occurrence matrices corresponding to each snapshot network as input and performs temporal link prediction in three steps: node similarity score computation, node similarity score prediction and link prediction. Univariate time

series based temporal link prediction [6] takes into account node’s local neighborhood based similarity measures. Unlike univariate time series models, multivariate time series link prediction models [7, 8] integrate temporal evolution of the network, node similarities and node connectivity information. Deep learning (DL) also called deep structured learning has shown its outstanding performance in various real-world scenarios. Using the deep layers of non-linear transformation, deep learning integrated this field to better capture the non-linear varying temporal patterns in dynamic networks. Recent advancements in DL leverages the notion of NRL for temporal link prediction. NRL or otherwise graph embedding techniques eliminated the need for painstaking feature engineering. The goal of this approach is to represent a graph in a low-dimensional vector space by preserving all the network properties. Different algorithms for graph embedding differs in the way they preserve all the network properties. A very few works in temporal link prediction concentrated on modelling an RBM [11].

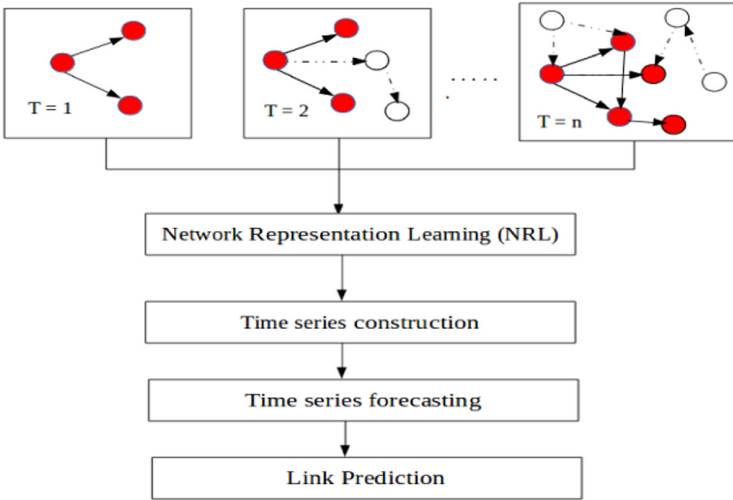
This study revealed that there exists several NRL techniques which gives the latent representations for nodes in the network by preserving the local and global properties. In addition, the frequently changing nature of dynamic networks make time series a promising option for temporal link prediction. There exists several techniques based on time series analysis for temporal link prediction. However, all of them deploy neighborhood based similarity measures and thereby ignores the global properties of the network. Here, we propose a unified framework that incorporates NRL techniques and time series analysis for temporal link prediction.

## 4 Proposed Method

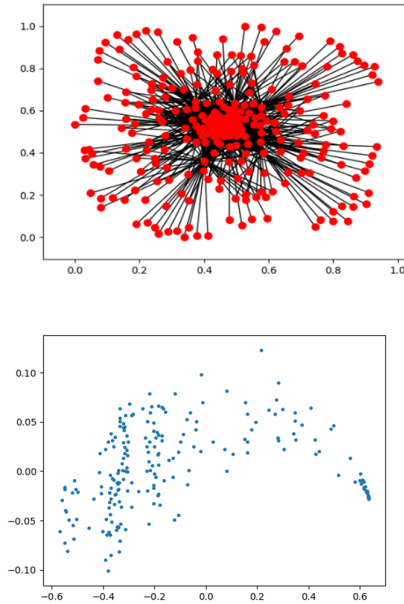
In this section, we introduce the proposed network embedding approach for time series based temporal link prediction. Our framework incorporates NRL based techniques and time series for temporal link prediction. The general architecture of proposed framework given in Fig. 2 is composed of four major phases: Network Representation Learning, Time Series Construction, Time Series Forecasting, Link Prediction. Initially, snapshots of the evolving network is taken at regular intervals of time. This enables to analyze the network structure for consecutive time periods.

### 4.1 Network Representation Learning (NRL)

NRL has been inspired from the language modeling techniques where words are replaced by nodes in the network. This methodology maps network vertices into a low-dimensional vector space, where all the network properties are preserved. Given a network  $G = (V, E)$ , NRL finds a mapping function  $\Phi: v \in V \rightarrow \mathbb{R}^{|V| \times D}$ , where  $D \ll |V|$ , such that every node  $v \in V$  is mapped into a D-dimensional vector space by preserving the structural proximity among nodes. Such latent representations of real-world networks encode social relations in a continuous vector space. This facilitates the original network to be easily deployed for further analysis. In the proposed framework, we deploy the most recent NRL techniques such as Node2Vec [3], SDNE [2] and DNGR [4]. Figure 3 depicts the latent representation of a network obtained using SDNE method.



**Fig. 2.** Architecture diagram



**Fig. 3.** A network and its latent representation

Node2Vec is an algorithmic framework that leverages the notion of random walks that preserves the network neighborhood of nodes to learn continuous feature representations for nodes in the network. The feature learning framework is introduced by extending the SkipGram architecture which optimizes the objective function given by

Eq. 1, where  $N_S(u)$  is the neighborhood of node  $u$  and  $f$  is the feature representation of the corresponding node.

$$\max_f \sum_{u \in V} \log \Pr(N_S(u) | f(u)) \quad (1)$$

SDNE is a semi-supervised framework that captures the highly non-linear structure of the networks. Inspired from the recent advancements in DL, this framework utilized deep autoencoders for learning latent representation of the network. Autoencoders have a deep neural network architecture and is composed of two parts: encoder and decoder. The encoder module is composed of multiple non-linear functions that maps the input data into its corresponding representation space. Decoder also consists of multiple non-linear functions that map the representations into a reconstruction space. SDNE exploits the first and second order proximities of the network to distinguish between the global and local network structure. This enables to learn the latent representations by preserving the structural proximities of the network.

DNGR is also an autoencoder based NRL framework. The model consists of a random surfing and context weighting module that generates the probabilistic distribution of the co-occurrence matrix and Stacked Denoising Autoencoder (SDAE) for dimensionality reduction. Given a network, DNGR performs a random surfing process (similar to PageRank) to generate a weighted co-occurrence matrix followed by the construction of Positive Pointwise Mutual Information (PPMI). This matrix contains the structural information of the network and it is given to SDAE to generate the latent representation for the network by optimizing the following objective function (see Eq. 2), where  $x_i$  is the input data and  $h(y_i; \theta)$  is the reconstructed data.

$$\operatorname{argmin}_{\theta} \sum_{i=1}^n \|x_i - h(y_i; \theta)\| \quad (2)$$

## 4.2 Time Series Construction and Forecasting

Time series construction phase takes as input the node embeddings obtained in the previous phase. For each pair of nodes, a similarity score is computed based on their low-dimensional node vectors. Let  $\Phi_t(u)$  and  $\Phi_t(v)$  be the embeddings of two nodes  $u$  and  $v$  respectively at time  $t$ , cosine similarity is defined as:

$$\operatorname{Cos}_{\text{sim}} = \frac{\Phi_t(u) \cdot \Phi_t(v)}{|\Phi_t(u)| |\Phi_t(v)|} \quad (3)$$

In addition to the similarity score computation, we analyze the change over time by modeling a time series for each pair of nodes. The cosine similarity scores of node pairs over time represented as time series enables to characterize the change in position of nodes in the embedding space.

The time series thus constructed is taken as input for time series forecasting phase. In the proposed system, we deploy ARIMA model [12] which maximizes the likelihood function. Once the time series is constructed, the future score values are predicted using

ARIMA (p, d, q) model. For a pair of nodes (u, v), the model which is applied to predict the score for time t by considering p autoregressive terms and q moving average terms is given by Eq. 4, where  $\Phi_i$  and  $\theta_j$  represents constant terms and  $\epsilon_t$  is the white noise. ARIMA model is applied with different p, d, q values. The parameter values giving minimum Akaike Information Criteria (AIC) value are utilized for predicting the future score values for each node pair.

$$\text{Score}(u, v, t) = \sum_{i=1}^p \Phi_i \text{Score}(u, v, t - i) + \sum_{j=1}^q \theta_j \epsilon_{t-j} + \epsilon_t \quad (4)$$

### 4.3 Link Prediction

In this phase, the future time series scores estimated in the previous phase are used to predict how likely two given nodes are to connect in future. First, each node pair are sorted based on the predicted similarity score. The sorted list is compared with actual links in the network for a future time.

## 5 Experiments

In this section we conduct experiments on several real-world datasets to evaluate the performance of the proposed temporal link prediction framework. Here, we utilize suitable evaluation measures to compare the accuracy of the method with the baseline methods under different scenarios. All the experiments were conducted on a machine with 15.6 GiB RAM and hexa-core processor with 3.2 GHz speed.

### 5.1 Datasets Used

Various standard real-world datasets are available to evaluate the performance of temporal link prediction. The following datasets were used in our experiments.

1. **Enron:** This dataset consists of emails between the employees in Enron Inc. from January 1999 to July 2002. Each node in the network represents a user and a link represents email communication between them.
2. **Haggle:** This network describes human contact information where contacts between people are measured by some wireless devices. Nodes represents users and links between them indicates a contact.
3. **Hep-ph:** This a collaboration graph of authors of scientific papers from Hep-Ph section of arXiv archive. The data covers papers in the period from January 1993 to April 2003.
4. **Radoslaw:** This network represents the email communication between employees in a mid-sized manufacturing company. Nodes in the network represents employees and edges between them are individual emails.

Table 1 shows the statistics of the datasets used. For Hep-ph dataset, we consider only the most popular nodes and it consists of 265 nodes and 19,736 edges. All the other datasets are used as it is.

**Table 1.** Statistics of the datasets used

Dataset	#Nodes	#Edges	#Timestamps
Enron	150	150	27
Haggle	274	274	6
Hep-ph	28,093	28,093	9
Radoslaw	167	167	10

## 5.2 Results and Analysis

The proposed framework is compared with some of the state-of-the-art works to evaluate the performance. The evaluation metrics used are Area Under the Curve (AUC) [16, 19] and Mean Average Precision (MAP) [2]. First, the system is compared with static link prediction techniques. Second, the evaluation of the proposed framework with state-of-the-art time series based temporal link prediction techniques is performed. Moreover, the effect of various network embedding techniques on the proposed framework is also observed. In this paper, static techniques are denoted as st-cn, st-jc, st-aa and the proposed time series based framework is denoted as ts-node2vec, ts-sdne and ts-dngr.

### Comparison with Static Link Prediction Techniques

On comparing the time series based framework which deploy local similarity indices and proposed framework on static link prediction techniques, it was found that the time series based approaches gives a better prediction results. Figure 4(a) shows that time series based local similarity metrics (ts-aa) for temporal link prediction improves the AUC scores for static link prediction using local similarity metrics (st-aa) by 14.75%, 29.09%, 18.3% and 32.7% for Enron, Haggle, Hep-ph and Radoslaw datasets respectively. In addition, the proposed framework (ts-node2vec, ts-sdne, ts-dngr) gives better AUC scores than that for static network embedding techniques (st-node2vec, st-sdne, st-dngr). The result shows that the time series based temporal link prediction techniques performs better than static link prediction techniques which depends solely on static network at a particular time period.

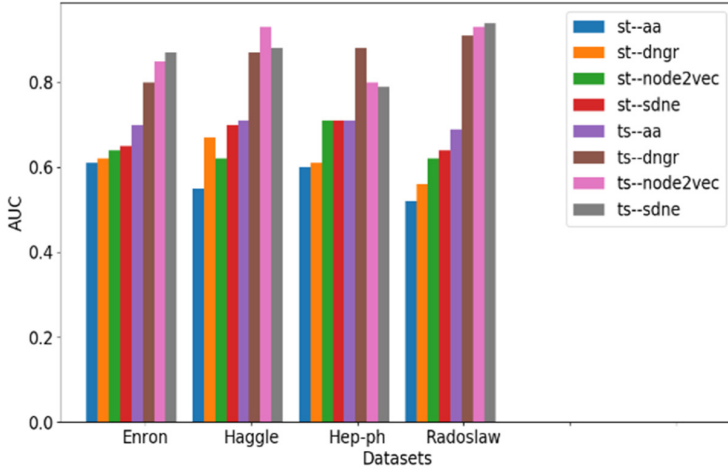
### Comparison with Time Series of Neighborhood Based Similarity Metrics

The MAP scores obtained on comparing the proposed framework with state-of-the-art time series based techniques is shown in Table 2. Better prediction results are obtained by taking top 20% links as connected and the rest as disconnected links. The observed results on evaluating the performance of proposed framework in terms of the AUC value computed is depicted in Fig. 4(b). The proposed system shows better results than time series based method using neighborhood based similarity measures for all the four real-world datasets. This confirms that the ability of NRL techniques to generate deep and latent representations of the network improves the prediction results.

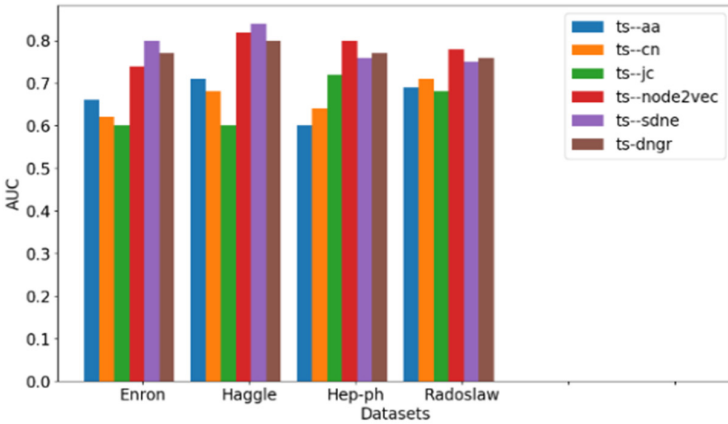
### Effect of Various Network Embedding Approaches

The performance of the system on three recent network embedding techniques are compared here. The observation of the prediction results on various embedding techniques





(a)



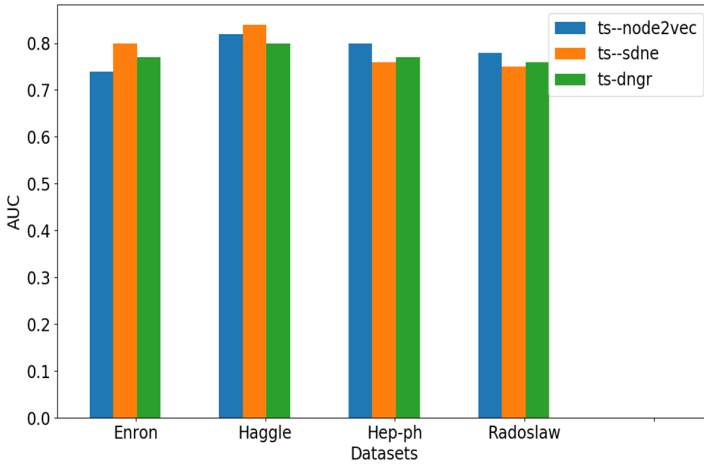
(b)

**Fig. 4.** Comparison of the proposed system with (a) static link prediction techniques (b) time series based link prediction techniques

is shown in Fig. 5. Among the three network embedding techniques, SDNE gives better prediction results for Enron and Hagggle datasets. The feature dimension for SDNE is set as  $d = 16$  for both the datasets. Since SDNE is found to be suitable for capturing non-linear patterns, it confirms that joint objective function of autoencoder designed for SDNE better captures the local and global structures in Enron and Hagggle networks quite efficiently. Moreover, Node2Vec framework gives a better prediction result for Hep-ph and Radoslaw datasets. For this experiment, the feature dimension for Node2Vec is set as  $d = 128$  for both the datasets. It confirms that the random walk based approach in Node2Vec better captures the community structure in these networks effectively and hence gives a better prediction result.

**Table 2.** Comparison of MAP scores of proposed system with baseline methods

Method	Enron	Haggle	Hep-ph	Radoslaw
Ts-cn	0.04	0.07	0.03	0.04
Ts-jc	0.06	0.10	0.04	0.03
Ts-aa	0.04	0.06	0.02	0.04
Ts-node2vec	<b>0.09</b>	<b>0.11</b>	<b>0.07</b>	<b>0.09</b>
Ts-sdne	<b>0.08</b>	<b>0.30</b>	<b>0.05</b>	<b>0.08</b>
Ts-dngr	<b>0.12</b>	<b>0.19</b>	<b>0.06</b>	<b>0.17</b>

**Fig. 5.** Effect of various network embedding approaches

## 6 Conclusion

In this paper, we proposed a unified framework for temporal link prediction which incorporated NRL based techniques and time series analysis. One of the key idea of our framework is to capture the non-linear temporal patterns in dynamic networks using network embedding techniques. Moreover, the framework is extended to incorporate time series forecasting models for prediction, since time series best captures the change over time. Experiments conducted on four real-world datasets show that the proposed system outperforms the state-of-the-art works. In future, the static network embedding techniques can be extended to incorporate dynamic behavior of networks. Dynamic network embeddings techniques can be deployed to perform the temporal link prediction task. The strength of dynamic network embedding techniques can be incorporated for time series construction to yield better prediction results. Moreover, leveraging different neural network models like LSTM for time series forecasting is also an interesting direction towards enhancing the performance of time series based temporal link prediction.

## References

1. Holme, P., Saramäki, J.: Temporal networks. *Phys. Rep.* **519**(3), 97–125 (2012)
2. Wang, D., Cui, P., Zhu, W.: Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1225–1234 (2016)
3. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)
4. Cao, S., Lu, W., Xu, Q.: Deep neural networks for learning graph representations. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
5. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **58**(7), 1019–1031 (2007)
6. Güneş, İ., Gündüz-Öğüdücü, Ş., Çataltepe, Z.: Link prediction using time series of neighborhood-based node similarity scores. *Data Min. Knowl. Discov.* **30**(1), 147–180 (2015). <https://doi.org/10.1007/s10618-015-0407-0>
7. Özcan, A., Öğüdücü, Ş.G.: Temporal link prediction using time series of quasi-local node similarity measures. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 381–386 (2016)
8. Özcan, A., Öğüdücü, Ş.G.: Multivariate temporal link prediction in evolving social networks. In: 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), pp. 185–190 (2015)
9. Wu, T., Chang, C.S., Liao, W.: Tracking network evolution and their applications in structural network analysis. *IEEE Trans. Netw. Sci. Eng.* (2018)
10. Ralescu, A., Kohram, M.: Spectral regression with low-rank approximation for dynamic graph link prediction. *IEEE Intell. Syst.* **26**(4), 48–53 (2011)
11. Li, T., Wang, B., Jiang, Y., Zhang, Y., Yan, Y.: Restricted Boltzmann machine-based approaches for link prediction in dynamic networks. *IEEE Access* **6**, 29940–29951 (2018)
12. Brockwell, P.J., Davis, R.A., Calder, M.V.: *Introduction to Time Series and Forecasting*, vol. 2. Springer, New York (2002)
13. Lei, K., Qin, M., Bai, B., Zhang, G.: Adaptive multiple non-negative matrix factorization for temporal link prediction in dynamic networks. In: Proceedings of the 2018 Workshop on Network Meets AI & ML, pp. 28–34 (2018)
14. Ma, X., Sun, P., Qin, G.: Nonnegative matrix factorization algorithms for link prediction temporal networks using graph communicability. *Pattern Recognit.* **71**, 361–374 (2017)
15. Ma, X., Sun, P., Wang, Y.: Graph regularized nonnegative matrix factorization for temporal link prediction in dynamic networks. *Phys. Stat. Mech. Appl.* **496**, 121–136 (2018)
16. Dunlavy, D.M., Kolda, T.G., Acar, E.: Temporal link prediction using matrix and tensor factorizations. *ACM Trans. Knowl. Discov. Data (TKDD)* **5**(2), 10 (2011)
17. Das, S., Das, S.K.: A probabilistic link prediction model in time-varying social networks. In: 2017 IEEE International Conference on Communications (ICC), pp. 1–6 (2017)
18. Ahamed, N.M., Chen, L.: An efficient algorithm for link prediction in temporal uncertain social networks. *Inf. Sci.* **331**, 120–136 (2016)
19. Lü, L., Jin, C.H., Zhou, T.: Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E* **80**(4), 046122 (2009)