Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 82 – 89

Procedia
Computer Science

The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)

Link Prediction Based on Common-Neighbors for Dynamic Social Network

Lin Yao[a], Luning Wang[a], Lv Pan[a], Kai Yao[b]

[a] *School of Software, Dalian University of Technology. Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China*
[b] *Shenyang University of Technology. Liaoning Province, China*

Abstract

Link prediction is an important issue in social networks. Most of the existing methods aim to predict interactions between individuals for static networks, ignoring the dynamic feature of social networks. This paper proposes a link prediction method which considers the dynamic topology of social networks. Given a snapshot of a social network at time t (or network evolution between $t1$ and $t2$), we seek to accurately predict the edges that will be added during the interval from time t (or $t2$) to a given future time t'. Our approach utilizes three metrics, the time-varied weight, the change degree of common neighbor and the intimacy between common neighbors. Moreover, we redefine the common neighbors by finding them within two hops. Experiments on DBLP show that our method can reach better results.

Keywords: link prediction; Common Neighbors; dynamic network

1. Introduction

A social network is a social structure made up of a set of social actors, whose nodes represent people or other entities embedded in a social context, and whose edges represent interaction, collaboration, or influence between entities. The associations are usually driven by mutual interests that are intrinsic to a group. Because the relationship between people are always changing, new edges and vertices are added to the graph over time and old ones may be deleted. Consequently, social network is generally complex and highly dynamic. As a key issue of social networks, link prediction has attracted more and more attention because link prediction is important for mining and analyzing the evolution of social networks[1].

The existing link prediction approaches can be classified into similarity-based ones and learning-based ones[1]. Similarity-based approaches are to compute the similarities between a pair of nodes by various graph-based similar-

* Corresponding author. Tel.: 008641162274396 ; fax: 008641162274455
E-mail address: yaolin@dlut.edu.cn

doi:10.1016/j.procs.2016.04.102

ity metrics and to use the ranking on the similarity scores to predict the link between two vertices[2,3,4]. Learning-based approaches are to treat the link prediction problem as a binary classification task. Therefore, some typical machine learning models such as classifier and probabilistic model can be used for solving this problem[5,6]. Compared with similarity-based approach, the latter approach usually suffer from model capacity and computational problems[7]. Moreover, most existing methods are designed for static network without considering the dynamics and evolutionary patterns of social networks. They would rather predict links from one static snapshot of the graph. However, graph data sets often show dynamic characteristics because of addition and deletion of nodes and edges in the networks.

In this paper, we propose a link prediction method for dynamic social networks. In a social network, people tend to create new relationships with people that are closer to them. The idea of using the size of Common Neighbors(CN) is just an attestation to the network transitivity property. Therefore, we design our prediction method based on the metric of common neighbors. While, the major problem is that most existing methods based on CN just focus on topological structure alone to predict the links in social networks without considering the dynamic network. In our approach, we consider three metrics, the time-varied weight, the change degree of common neighbor, and the intimacy between common neighbors. The time-varied weight reflects the change of topological structure with time. The closer the time is to us, the bigger the weight is. The change degree of common neighbor reflects the stability of every neighbor in the current period, and the weight of a common neighbor gets more with the smaller degree (i.e. more stable). In the CN algorithm, every common neighbor has the same weight for the final prediction. However, every neighbor may not have the same change degree. By adopting this metric, we can kick off some abnormal nodes. The intimacy between common neighbors is used to judge the similarity of two nodes. If their common neighbors are more closely, the similarity between these two nodes are higher. Furthermore, the common neighbors are defined as some nodes within two hops, while CN considers the common neighbors within only one hop.

The contributions of this work are:

(1) We propose a new link prediction method for dynamic networks. To improve the predictive accuracy, we adopt three metrics, the time-varied weight, the change degree of common neighbor, and the intimacy between common neighbors.

(2) We redefine the common neighbors by considering the nodes within two hops to achieve better performance.

(3) We demonstrate experimental results on the effectiveness of the proposed approach with DBLP database. Results show that our approach is competitive to (and sometimes are better than) those of the exact ones.

The rest of this paper is organized as follows. In Section 2, we discuss the related work. Our proposed method is given in Section 3. Experimental results are presented in Section 4. Finally, we conclude our work in Section 5.

2. Related Work

Most of the existing link prediction approaches can be classified into learning-based ones and similarity-based ones[1]. Learning-based approaches adopt the classifier, such as Markov chains[8,9], SVM[5], etc.. Some methods use probabilistic models such as Markov Random Fields, Bayers model, etc. to predict the link association[10,11]. Compared with similarity-based approaches, learning-based approaches have the difficulties in feature selection and unbalancing output classes and is suffered from computational cost and limitation of capacity, therefore it is not suitable for large-scale and dynamic networks[12]. Similarity-based approaches can adopt nodes' information, network topology, etc. to link prediction. In[13], the keyword distance is adopted to define similarity functions between a pair of users. In[14], users' interests are used to measure the similarity. In[15], authors propose a novel user similarity measure for online social networks, which combines both network and profile similarity. They also propose a method to infer a portion of the missing items from profile of the users contacts. Liben-Nowell and Kleinberg proposed one of the earliest topology-based prediction models that works explicitly on a social network[2]. They tested the predictive power of some proximity metrics, including Common neighbours(CN), Preferential Attachment(PA), Katz measure, etc.. The ranking on the similarity scores is used to predict the link between two vertices. In[16], nine common algorithms of link prediction are compared and the results show that the CN algorithm possesses the best performance.

However, most of the existing link prediction methods are aim to predict links from one static snapshot of the network graph, ignoring the underlying additional temporal information in pace with the evolution of the network. Some researchers have attempted to predict links by changing the dynamic network into several static networks. Then, they try to design algorithms based on static prediction methods. In[17], authors summarized the dynamic graph with a weighted static graph and then incorporated the link weights in a relational Bayes classifier to achieve link prediction. In[18], the time-series link prediction problem is introduced, taking into consideration temporal evolutions of link occurrences to predict link occurrence probabilities at a particular time. Both inter-link structural dependencies and intra-link temporal dependencies are exploited. In[19], the history information available on the interactions is incorporated into the current social network state. Results unequivocally show that timestamps of past interactions significantly improve the prediction accuracy of new and recurrent links. In[20], the time series for each pair of non-connected nodes in the networks are calculated, and then a forecasting model on these time series is deployed.

Though the above papers consider the different topology of different time series, none of these earlier works have taken into account the weight of the links in the past. A hybrid approach utilizing time-varied weight information of links is proposed in[21]. However, the variation degree of every node over a period of time has not been considered. In our method, we adopt three metrics, the time-varied weight, the change degree of common neighbor, and the intimacy between common neighbors. Furthermore, we also redefine the common neighbors.

3. The Proposed Method

3.1. Some Basic Definitions

3.1.1. Graph

Given a social network $G = (V, E)$, where $V = \{v_1, v_2, v_3, ..., v_n\}$ and $E = \{e \mid < v_i, v_j >, v_i, v_j \in V\}$ are sets of nodes and links, respectively. In addition, the graph can be divided into directed graph and undirected graph. In the directed graph, $< v_i, v_j >$ and $< v_j, v_i >$ represent the different edges. In the undirected graph, $< v_i, v_j >$ and $< v_j, v_i >$ represent the same edge. In this paper, we only consider the undirected social networks. Each edge $< v_i, v_j > \in E$ represents an interaction between two neighbors, v_i and v_j. $\Gamma(v_i)$ denotes the neighbor set of node v_i.

3.1.2. Link Prediction

Consider a static network $G = (V, E)$ at a particular time t, the set U is defined as the set of all possible edges and $Z = U - E$ represents the set of edges which do not exist in the network. The link prediction aims to find new links between nodes for a future time $t^{'}(t^{'} > t)$ or missing links or unobserved links, in current network from the set Z[1]. In the dynamic network, the topology data presents a sequence of graph snapshots. Let $G = (G_1, G_2, ..., G_T)$ be the sequence of temporal snapshots of the data at consecutive time steps t. New objects and links may have been added or deleted from G_{t-1} to G_t. The link prediction for dynamic networks is aimed at predicting the occurrence probabilities between edges at time steps $T + 1$ according to G.

3.2. Three Metrics

In this section, we introduce three metrics used to support the prediction process.

3.2.1. Time-Varied Weight

Link prediction for static networks usually only focuses on the presence or absence of links, rather than the link existence time or the frequency of links. However, the dynamic links have a certain degree of influence on the prediction results. For example, the possibility of establishing a link between the two black nodes is relatively low if we only consider the topology at the snapshot t_3 in Figure 1. However, the contact probability between the two black nodes is relatively high if some historical information at t_1 and t_2 is considered. Consequently, we define the time-varied weight $W(t)$ to highlight the importance of past link associations on the link prediction. It is obvious that network topology of the current moment has greater impact on the prediction results. Hence, $w(t)$ is defined as a Time-Decay function:

$$W(t) = e^{-\lambda(T-t)} \tag{1}$$

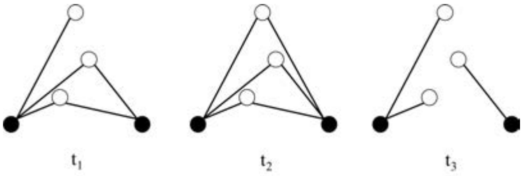

Fig. 1. An example of time-varied weight.

3.2.2. Change Degree of CN

CN is widely used due to its simplicity and good performance. In this paper, we consider the change degree of every common neighbor in dynamic networks. For instance, in a co-authorship network, a researcher usually cooperates with the other researchers whose research field is the same as his. Even if links are changed, the changes will occur only in this research community. However, if one researcher often jumps out of his current research community, namely, often changes his research direction, the change degree of this researcher is big and this researcher can be considered as a outlier.

For a time period $(1, 2, ..., t, ...T)$, the impact of a common neighbor on the link prediction is defined $W_t(v_m)$, where v_m is a common neighbor between v_i and v_j at the snapshot t, T denotes the current time and $\triangle T$ denotes the time interval from 1 to T. $d_{t-1,t}$ denotes the Euclidean distance of v_m between $t-1$ and t.

$$W_t(v_m) = \frac{1}{\sum_{t=2}^{T} d_{t-1,t} \Big/ \triangle T} \tag{2}$$

3.2.3. Intimacy Between Common neighbors

In static networks, the mutual relationship between the common neighbors is not considered. For example, the white nodes in Figure $2(a)$ and $2(b)$ are the common neighbors of the pair of black nodes. After we extract the subgraph containing only common neighbors in Figure $2(c)$ and $2(d)$, it is obvious that the possibility of establishing a link between the two black nodes in Figure $2(b)$ is higher intuitively because the common neighbors have a more intimate or complex relationship. This is also similar to the real communication. If two persons have a lot of mutual friends, and relationships between these friends are also relatively close, then the two persons have a high possibility of meeting. This suggests that the relationship between the common neighbors will affect the probability of the link established.

Based on the above idea, we define the intimacy between common neighbors at the snapshot t, $W_t(v_i, v_j)$, in Equation (3), where $N = \{< v_a, v_b > | < v_a, v_b > \in E, v_a \in \Gamma(v_i) \cap \Gamma(v_j), v_b \in \Gamma(v_i) \cap \Gamma(v_j)\}$, and $|N|$ denotes the number of edges between the common neighbors.

$$W_t(v_i, v_j) = ln(|N|) \tag{3}$$

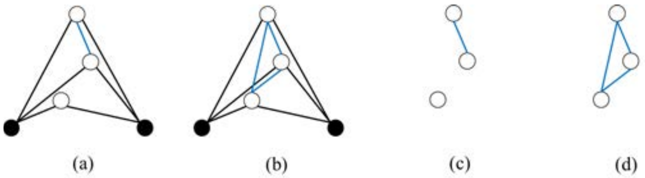

Fig. 2. An example of intimacy.

3.3. Redefined Common neighbors

The core idea of CN is to find the common neighbors between the pair of two nodes. However, CN as well as the most of the improved algorithms based on CN only consider the common neighbors within one hop. For example, there is no common neighbor between node 1 and node 6 in Figure 3 for CN. Therefore, the prediction value between node 1 and node 6 is 0, that is node 1 and node 6 will not establish a link in the future based on CN. But intuitively, the possibility of establishing a link between node 1 and node 6 is high. In order to improve the weakness of CN, our method redefines the common neighbors between node v_i and node v_j according to Definition 1. Our improved method will find the common neighbors within two hops. For example, node 1 and node 6 in Figure 3 will have common neighbors: node 3 and node 4, thus, node 1 and node 6 have the possibility of establishing a link. Therefore, redefining common neighbors will improve the prediction accuracy.

Definition 1. If $F(v_i, v_k) > 0$ and $F(v_j, v_k) > 0$, then node v_k is defined as the common neighbor between v_i and v_j, where $F(v_i, v_j)$ refers to the similarity value between v_i and v_j.

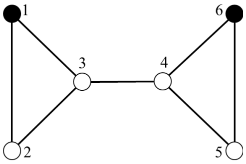

Fig. 3. An example of redefined common neighbors.

3.4. Details

In this section, we will describe the details of our predictive method. The frequently used notations in the paper are listed in Table 1. The algorithm is listed in Algorithm 1.

Table 1. Frequently Used Notations.

Notation	Description
v_i	Node i
$\Gamma(v_i)$	Neighbor set of node v_i
$f_t(v_i, v_j)$	Similarity value between v_i and v_j at t
$F(v_i, v_j)$	Similarity value between v_i and v_j considered the period from 1 to T
$P(v_i, v_j)$	Final prediction value between v_i and v_j
$S_t(v_i, v_j)$	The set of common neighbors between v_i and v_j at t
$W_t(v_i, v_j)$	Intimacy between the common neighbors of v_i and v_j at t
$W_t(v_i)$	Change degree of node v_i from 1 to t
$W(t)$	Time-varied weight at t
$N_T(v_i, v_j)$	The set of v_i, v_j and their redefined common neighbors at T

For every pair of v_i and v_j in G, our method includes the following steps:
Step 1: For every snapshot at t, $S_t(v_i, v_j)$ is found.
Step 2: $W_t(v_i, v_j)$ is calculated among the nodes in the set of $S_t(v_i, v_j)$ in Equation (3).
Step 3: $W_t(v_m)$ of every common neighbor v_m in $S_t(v_i, v_j)$ is calculated based on Equation (2).
Step 4: $f_t(v_i, v_j)$ is calculated in Equation (4).

$$f_t(v_i, v_j) = W_t(v_i, v_j) \cdot \sum_{v_m \in S_t(v_i, v_j)} W_t(v_m) \tag{4}$$

Step 5: Considering the period from 1 to T, $F(v_i, v_j)$ is calculated in Equation (5).

$$F(v_i, v_j) = \sum_{t=1}^{T} W(t) \cdot f_t(v_i, v_j) \tag{5}$$

Step 6: Redefining the common neighbors between v_i and v_j according to F. If $F(v_i, v_k) > 0$ and $F(v_j, v_k) > 0$, node v_k is defined as the common neighbor between v_i and v_j, then $N_T(v_i, v_j)$ is found.

Step 7: $P(v_i, v_j)$ is calculated in Equation (6), where $F(v_x, v_y)$ is considered as the weight of the edges between v_i, v_j and their redefined common neighbors. A bigger $P(v_i, v_j)$ means a higher probability of association between v_i and v_j.

$$P(v_i, v_j) = \sum_{v_x, v_y \in N_T(v_i, v_j)} F(v_x, v_y) \tag{6}$$

Algorithm 1 Link Prediction Based on CN for Dynamic Network

```
Input: $G = (G_1, G_2, ..., G_T)$, $(v_i, v_j)$
Output: $P(v_i, v_j)$
function PREDICTION($v_i, v_j$)
    for $t = 1 \rightarrow T$ do
        $S_t(v_i, v_j) \leftarrow \Gamma(v_i) \cap \Gamma(v_j)$
        $wi \leftarrow W_t(v_i, v_j)$
        $cnd \leftarrow \sum_{v_m \in S_t(v_i, v_j)} W_t(v_m)$
        $f_t(v_i, v_j) \leftarrow wi \cdot cnd$
    end for
    $F(v_i, v_j) \leftarrow \sum_{t=1}^{T} W(t) \cdot f_t(v_i, v_j)$
    $P(v_i, v_j) \leftarrow \sum_{v_x, v_y \in N_T(v_i, v_j)} F(v_x, v_y)$
end function
```

4. Empirical Evaluation

4.1. Data Sets and Features

We evaluate our link prediction method on the bibliographic database DBLP (http://dblp.uni-trier.de/), from which we extract a co-authorship network. In the co-authorship network, a node represents an author and a link indicates that two authors has cooperated with each other at least once. DBLP contains papers spanning more than several decades and the density of the records over time is not consistent - with earlier years sparsely represented. We eliminate these early years, and use the papers from the past 20 years. We further clean the data set by eliminating papers written by only one author since they do not help in link prediction framework based on our method. Table 2 summarizes various features of the data set after data-cleaning.

Table 2. The statistics of the data set.

Time span	Number of authors	Number of papers
1995 – 1999	302612	172121
2000 – 2004	579961	343897
2005 – 2009	1139183	706749
2010 – 2014	1629103	983474

4.2. Experimental Configurations

We compare our method with CN[1], RA[1], PA[1] and tw-CN method[21]. RA is motivated by the physical processes of resource allocation and suppress the contribution of the high-degree common neighbors. PA indicates that new links will be more likely to connect higher-degree nodes than lower ones. In tw-CN, a hybrid approach utilizing time-varied weight information of links is proposed. For every statistics in Table 2, the data of first 4 years is selected

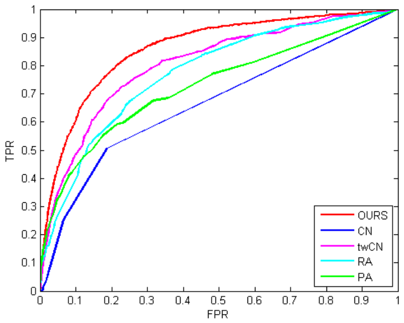

Fig. 4. ROC curves of three methods from 1995 to 1999.

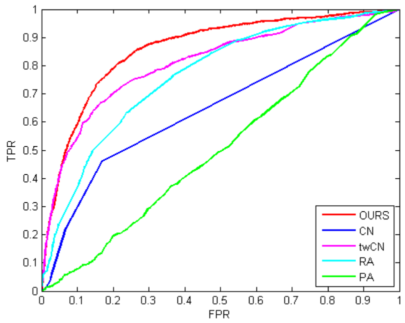

Fig. 5. ROC curves of three methods from 2000 to 2004.

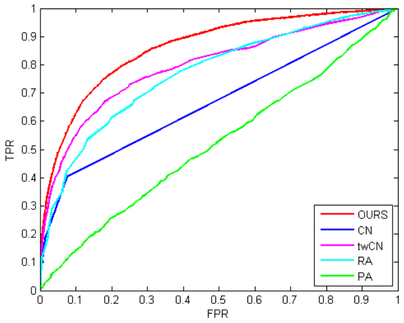

Fig. 6. ROC curves of three methods from 2005 to 2009.

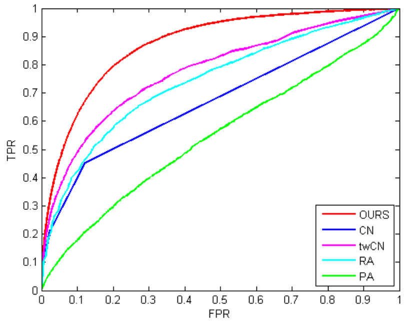

Fig. 7. ROC curves of three methods from 2010 to 2014.

as the training set and the data of the fifth year is used as the testing set. Exceptionally, as CN, RA, PA are the link prediction methods for static network, the data of the fourth year is selected as the training set and the data of the fifth year is used as the testing set. The training set and the testing set are the same in all five methods.

In our experiments, we evaluate our method from two metrics, ROC curve and AUC[21].

• *ROC curve*: A receiver operating characteristic curve is a graphical plot which illustrates the performance of a binary classifier system intuitively.

• *AUC*: The area under the ROC curve is to quantify the accuracy of prediction method.

4.3. Experiment Results and Analysis

Figure 4 to Figure 7 demonstrate the ROC curves of five methods, CN, RA, PA, tw-CN, and ours in each time span. FPR represents false positive rate and TPR represents true positive rate.

Table 3. AUC values of three methods.

Time span	CN	RA	PA	tw-CN	Ours
1995 – 1999	0.6676	0.7740	0.7278	0.8049	**0.8607**
2000 – 2004	0.6386	0.7640	0.5033	0.8100	**0.8586**
2005 – 2009	0.6658	0.7701	0.5202	0.7951	**0.8615**
2010 – 2014	0.6498	0.7413	0.5473	0.7759	**0.8737**

The ROC curves and Table 3 show both tw-CN and our method can achieve better performance than other three methods. These results show that the performance of link prediction has been greatly improved after taking the evolution of network topology into account. Moreover, our performance improves much compared with tw-CN, because only the time-varied weight information of links is considered in tw-CN. Therefore, the performance of tw-CN improves slightly and is very close to RA. While, our method considers three metrics, the time-varied weight, the change degree of common neighbor, the intimacy between common neighbors and considers the common neighbors within two hops, resulting in the improvement of performance.

5. Conclusion

In this work, we propose a link prediction method based on CN for dynamic social network. Different from the static method CN, our method considers the dynamic feature of social networks and present three metrics, the time-varied weight, the change degree of common neighbor, and the intimacy between common neighbors. Furthermore, we redefine the common neighbors within two hops. The experimental results show that our method improves the performance of link prediction. In the future, we would like to test our link prediction method on more data sets, and adapt the three metrics according to different features of different networks to achieve better performance.

Acknowledgements

This research is sponsored in part by the National Natural Science Foundation of China (contract/grant number: No.61173179 and No.61202441), Nuclear high base of major projects(2012ZX01039-004-21) and Program for New Century Excellent Talents in University (NCET-13-0083).

References

1. Wang, P., Xu, B., Wu, Y., Zhou, X.. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences* 2015; **58**(1):1–38.
2. Liben-Nowell, D., Kleinberg, J.. The link-prediction problem for social networks. *Journal of the American society for information science and technology* 2007;**58**(7):1019–1031.
3. Lü, L., Jin, C.H., Zhou, T.. Similarity index based on local paths for link prediction of complex networks. *Physical Review E* 2009; **80**(4):046122.
4. Liu, W., Lü, L.. Link prediction based on local random walk. *EPL (Europhysics Letters)* 2010;**89**(5):58007.
5. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.. Link prediction using supervised learning. In: *SDM06: Workshop on Link Analysis, Counter-terrorism and Security*. 2006, .
6. Benchettara, N., Kanawati, R., Rouveirol, C.. Supervised machine learning applied to link prediction in bipartite social networks. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. IEEE; 2010, p. 326–330.
7. Li, X., Du, N., Li, H., Li, K., Gao, J., Zhang, A.. A deep learning approach to link prediction in dynamic networks. In: *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia*. 2014, p. 289–297.
8. Sarukkai, R.R.. Link prediction and path analysis using markov chains. *Computer Networks* 2000;**33**(1):377–386.
9. Zhu, J., Hong, J., Hughes, J.G.. Using markov chains for link prediction in adaptive web sites. In: *Soft-Ware 2002: Computing in an Imperfect World*. Springer; 2002, p. 60–73.
10. Liu, Z., Zhang, Q.M., Lü, L., Zhou, T.. Link prediction in complex networks: a local naïve bayes model. *EPL (Europhysics Letters)* 2011; **96**(4):48007.
11. Wang, C., Satuluri, V., Parthasarathy, S.. Local probabilistic models for link prediction. In: *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE; 2007, p. 322–331.
12. Wang, T., Liao, G.. A review of link prediction in social networks. In: *Management of e-Commerce and e-Government (ICMeCG), 2014 International Conference on*. IEEE; 2014, p. 147–150.
13. Bhattacharyya, P., Garg, A., Wu, S.F.. Analysis of user keyword similarity in online social networks. *Social network analysis and mining* 2011;**1**(3):143–158.
14. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.. Effects of user similarity in social media. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM; 2012, p. 703–712.
15. Akcora, C.G., Carminati, B., Ferrari, E.. User similarities on social networks. *Social Network Analysis and Mining* 2013;**3**(3):475–495.
16. Zhou, T., Lü, L., Zhang, Y.C.. Predicting missing links via local information. *The European Physical Journal B* 2009;**71**(4):623–630.
17. Sharan, U., Neville, J.. Exploiting time-varying relationships in statistical relational models. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM; 2007, p. 9–15.
18. Huang, Z., Lin, D.K.. The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing* 2009;**21**(2):286–303.
19. Tylenda, T., Angelova, R., Bedathur, S.. Towards time-aware link prediction in evolving social networks. In: *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*. ACM; 2009, p. 9.
20. Soares, P.R.d.S., Prudêncio, R.B.C.. Time series based link prediction. In: *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE; 2012, p. 1–7.
21. Huang, S., Tang, Y., Tang, F., Li, J.. Link prediction based on time-varied weight in co-authorship network. In: *Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 18th International Conference on*. IEEE; 2014, p. 706–709.