A Supervised Learning Approach to Link Prediction in Dynamic Networks

Shuai Xu, Kai Han(✉), and Naiting Xu

School of Computer Science and Technology/Suzhou Institute for Advanced Study,
University of Science and Technology of China, Hefei, People's Republic of China
{sa615527,sa615237}@mail.ustc.edu.cn, hankai@ustc.edu.cn

Abstract. Link prediction, as one of fundamental problems in social network, has aroused the vast majority of research on it. However, most of existing methods have focused on the static networks, although there exist some machine learning methods for the dynamic networks, they regard either link structures or node attributes captured from a single snapshot of the network as the features, thus cannot achieve high accuracy. In this paper, following the supervised learning framework, we innovatively propose a new approach to this problem in dynamic networks. In particular, our features are captured from the variation of the structural properties and a lot of important metrics considering the long-term graph evolution of network, instead of a single snapshot. For each feature, we use an optimization algorithm to calculate the corresponding weight of each classifier, and then can determine whether there is a connection between a pair of nodes. In addition, we execute our method on two real-world dynamic networks, which indicate that our method works well and significantly outperforms the prior methods.

Keywords: Dynamic network · Link prediction
Supervised learning · Social network analysis

1 Introduction

With the rapid development of the social network, many relevant problems have raised along with it, *link prediction* is a typical one among them. In particular, the topology of network is always changing, new nodes and edges are added, meanwhile, old nodes may be deleted. Due to this highly dynamic nature, link prediction on the online social network becomes such an extremely challenging and urgent research issue.

In this field, many research works have concentrated on proposing novel and efficient methods for link prediction. Until now, most of the existing methods no matter based on the unsupervised learning or the supervised learning only focus on the static networks. In other words, they barely take account of the long-term graph evolution of the networks, which cannot achieve good performance when applied in the real dynamic networks. In view of this situation, we

S. Chellappan et al. (Eds.): WASA 2018, LNCS 10874, pp. 799–805, 2018.
https://doi.org/10.1007/978-3-319-94268-1_70

creatively put forward some new techniques to construct an approach to the link prediction problem in dynamic networks. The main contributions of this work are summarized as below:

- We innovatively propose an efficient approach with high accuracy to solve link prediction problem for online dynamic social network.
- To improve the prediction accuracy, we propose a different feature selection which fully takes account of the long-term graph evolution of social networks.
- We execute our method on two real-world dynamic social networks and the related experimental evaluations indicate that our method works surprisingly well and significantly outperforms the prior methods.

2 Related Work and Problem Definition

2.1 Related Work

Motivated by the similarity-based methods, Liben-Nowell and Kleinberg [7] originally proposed a link prediction method based on dynamic topology of networks. Later, several approaches to link prediction have been subsequently proposed either regarding the keyword distance as the property between a pair of nodes [4], or based on the user interests [3], or combining both profile similarity and social network topology [2]. In particular, in the work of Zhou et al. [12], experimental evaluations show that the link prediction algorithm based common neighbors achieves the best performance among others.

Based on the idea of Al Hasan et al. [5] that treating the link prediction problem as a binary classification problem, Da Silva Soare and Prudncio [10] proposed a link prediction algorithm which is confirmed to achieve the better performance than the traditional similar methods. Subsequently, Richard et al. [9] proposed a different method by evaluating the evolution process of dynamic network, which shows the better performance when compared with the traditional heuristics approaches. Besides, the approaches of [11] resolve the link prediction problem in dynamic/static networks, whereas they predict the future relationship by only using the link structure information but ignoring the node attributes.

2.2 Problem Definition

Given an evolving social network $G^t = (\boldsymbol{V}, E^t, P^t)$ at time t, then a sequence of graph snapshots $\{G^t\}_{t=1,\ldots,T} = \{G^1, G^2, \ldots, G^T\}$ can be produced at following time stamps. The link prediction aims to predict the most likely link state for a future time $t'(t' > T)$, where $\boldsymbol{V}$ remain the same across all time steps but E^t and P^t change for each time. For each pair of nodes (v, w), we compute their probability $P^{t+1}(v, w)$ of establishing link and predict their link state $L^{t+1}(v, w)$, where $P^{t+1}(v, w)$ is a element of a $n \times n$ matrix P^{t+1}, and $L^{t+1}(v, w)$ denotes the label of establishing link between v and w at time $t + 1$. In detail, if $e(v, w) \in E^{t+1}$, then $L^{t+1}(v, w) = 1$, otherwise $L^{t+1}(v, w) = 0$. Finally the network snapshot G^{t+1} can be obtained in the next time step $t = t + 1$.

3 Method Formulation

3.1 Feature Selection

In this paper, in order to improve the prediction accuracy, we propose a different feature selection which fully takes account of the long-term graph evolution of social networks. Moreover, these features are captured from the variation of the structural properties and a lot of important metrics including the change degree of common neighbors, the time-varied weight and the intimacy between common neighbors, instead of a single snapshot. In detail, these structure properties used are: Preferential Attachment (PA) [8], Common Neighbor (CN) [8], Adamic Adar (AA) [1] and Jaccards Coefficient (JC) [10], where $PA(v,w) = |N(v)| \times |N(w)|$, $CN(v,w) = |N(v) \cap N(w)|$, $AA(v,w) = \sum_{u \in N(v) \cap N(w)} \frac{1}{\log(|N(u)|)}$ and $JC(v,w) = |\frac{N(v) \cap N(w)}{N(v) \cup N(w)}|$. In addition, three important metrics used are: Time Weight (TW), Change Degree of CN (CD) and Closeness Between CNs (CB), where $TW(t) = \exp(-\psi(T-t))$, $CD_t(cn) = \frac{t}{\sum_{i=2}^{t} ed_{i-1,i}}$ and $CB_t(v,w) = \ln(|ACN_t|)$.

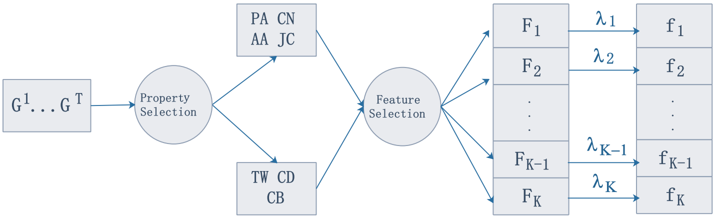


Fig. 1. The model of *SLM-Vp*.

3.2 Model of *SLM-Vp*

To address this problem, we innovatively propose a new approach *SLM-Vp*, which is a supervised learning method based on flat SVM classification methods, and these features are captured both from the variation of the structural properties and a lot of important metrics considering the long-term graph evolution of network, instead of a single snapshot. In detail, the model of *SLM-Vp* is shown in Fig. 1, where $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ is a weight vector of classifiers $(f_1, f_2, \ldots, f_k)$.

Firstly, for each $G^t \in \{G^t\}_{t=1,\ldots,T}$, we can represent G^t as a $n \times n$ adjacency matrix M^t, where M^t is symmetric and each element $M^t(v,w)$ equals $L^t(v,w)$ or $P^t(v,w)$. Moreover, for an arbitrary pair of node v and w, a $K \times T$ matrix $M^t(v,w)$ can be calculated, where each element equals $F_k^t(v,w)$. Then we can calculate the change values of all its features, thus a $K \times (T-1)$ matrix $VM^{T-1}(v,w)$ can be calculated, where each element equals $\triangle F_k^t(v,w)$. Secondly, for each structure feature F_k, we can train a classifier by using a set of data composed of instances annotated with structural features and the label

$L(v, w)$ of establishing link between v and w. We denote this set of instances as $VM_k^{T-1}(v, w)$ and select both the negative and positive training instances by using a random selection algorithm. Finally, we employ an optimization algorithm to calculate the weight of each classifier. After the above three steps, we can eventually obtained a probability of establishing a link between node v and w. In this paper, we use seven classifiers for seven features respectively, which are captured from the variation of four structural properties and three important metrics.

3.3 Implementation of *SLM-Vp*

The pseudo code of *SLM-Vp* algorithm is described as in Algorithm 1 below, where TS_k^T is a training set which is generated for each F_k, and TS_k^T can be calculated in Eq. (1). Then we train a data instance TS_k^T to learn a classifier f_k^T. Note $\overrightarrow{\lambda}$ is a vector of weights and can be represented as $\overrightarrow{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$, where λ_k is the weight of classifier f_k^T. We can give an *example*[1] to illustrate.

$$TS_k^T = \{(\{\triangle F_k(v, w)\}_1^T, L_T(v, w))\}. \tag{1}$$

Algorithm 1. SLM-Vp algorithm

```
Input: ${G^t}_{t=1,...,T}, P^T, \{F_1, F_2, \ldots, F_K\}, (v, w)$;
Output: $P^{T+1}(v, w)$;
for each $F_k$ in $\{F_1, F_2, \ldots, F_K\}$ do
    for each $(v, w)$ in the train set do
        Calculate $F_k^1(v, w), F_k^2(v, w), \ldots, F_k^T(v, w)$;
        Get $\{\triangle F_k^t(v, w)\}_1^{T-1}$
    end for
    Train a classifier $f_k^T$ by using a new training set $TS_k^T$;
end for
Optimize $\min \sum(\sum \lambda_k \cdot f_k^T(v, w) - L_{T+1}(v, w))^2$ and get weight $\overrightarrow{\lambda}$;
return $P^{T+1}(v, w) = \Sigma\lambda_k \cdot f_k^{T+1}(v, w)$;
```

4 Experiment and Analysis

In this section, we conduct extensive experiments to validate our proposed methods on two specific co-authorship networks, i.e., *hep-th*, *hep-lat* from two sections of *Arxiv* (www.arxiv.org) which are also used by lots of prior works (e.g. [7,11]). The *hep-th*[2] network is formed by authors in theoretical high energy physics area, and the *hep-lat*[3] is composed by authors in high energy physics-lattice area.

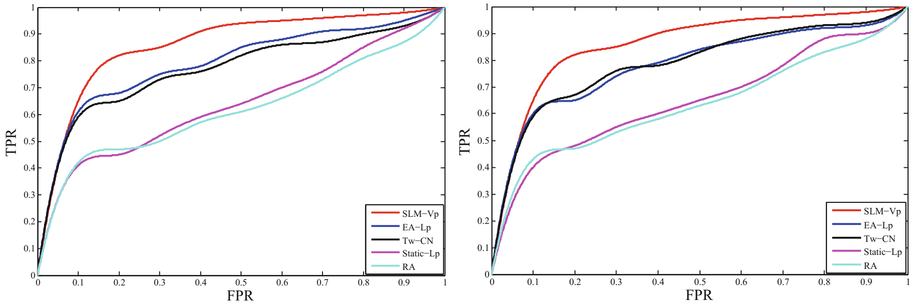
[1] https://github.com/ustcxs/wasa/blob/master/SLM-Vp.pdf.

[2] http://arxiv.org/archive/hep-th.

[3] http://arxiv.org/archive/hep-lat.

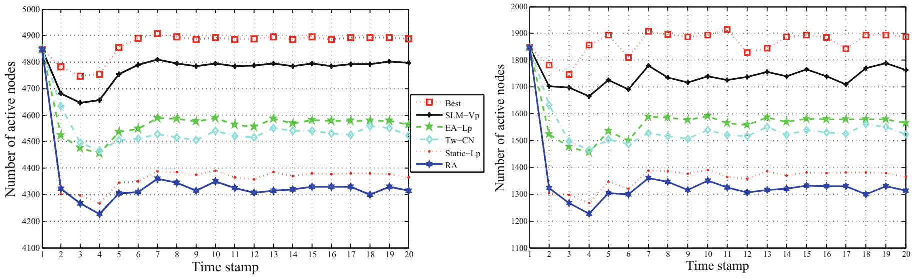
For convenience of experiment setup, we extracted data sets from year 1995 to 2015 for *hep-th* and *hep-lat*. In detail, *hep-th* has 19258 authors and 132568 collaborations, *hep-lat* has 5136 authors and 72354 collaborations. We select training instances and testing instances by running random algorithm with about 1:1 ratio. In detail, *hep-th* has 8236 training instances and 7998 testing instances, *hep-lat* has 2125 training instances and 2204 testing instances.

For verifying our proposed method *SLM-Vp*, we make a comparison among four baseline algorithms: *RA* [12], *Tw-CN* [6], *Static-Lp* [5], *EA-Lp* [9].



(a) *hep-th* (b) *hep-lat*

Fig. 2. The ROC curves of all the methods in two networks



(a) *hep-th* (b) *hep-lat*

Fig. 3. The maximization influence of all methods in two networks

We test all the methods in two real-life social networks, and all the settings are decided by experimentation. As shown in Fig. 2, we demonstrate the *ROC* curves of all the methods in two networks. Table 1 lists the *AUC* values of all the methods in two networks. In Fig. 3, we predict the link state of all nodes in graph by using above five methods, then we obtain five predicted graph snapshots and select top-k most influential node sets where $k = 50$. Finally, we calculate the

Table 1. The AUC values of all the methods in two networks

Network	RA	Tw-CN	Static-Lp	EA-Lp	SLM-Vp
hep-th	0.6523	0.7885	0.6768	0.7968	**0.8585**
hep-lat	0.6456	0.7778	0.6623	0.7883	**0.8503**

influence diffusion of these node sets in a real graph snapshot. Note that the Best method shown in Fig. 3 selects top-k most influential node sets and calculates the influence diffusion in a real graph snapshot without the prediction.

The experimental results show that our *SLM-Vp* algorithms can achieve better performance than other four methods. Moreover, the *ROC* curves and the maximization influence indicate that *SLM-Vp* can achieve better performance than other existing methods in the following graph snapshots. These results show that both supervised learning and unsupervised learning can be employed to link prediction problem. However, the link prediction methods under supervised learning usually achieve better performance. Thus supervised machine leaning is a key research direction for link prediction problem in dynamic networks.

5 Conclusion

In this paper, we propose a supervised learning method called *SLM-Vp*, which is based on flat SVM classification methods to deal with link prediction problem in dynamic networks. In particular, we produce a distinctive feature selection way which fully considers the long-term graph evolution of network and achieves high accuracy. Moreover, experimental evaluations on two social networks show that our proposed method result in better performance compared with other four methods. We believe that our results are meaningful contribute to the research.

Acknowledgment. This work is partially supported by National Natural Science Foundation of China (NSFC) under Grant No. 61472460, No. 61772491, Natural Science Foundation of Jiangsu Province under Grant No. BK20161256. Kai Han is the corresponding author.

References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**(3), 211–230 (2003)
2. Akcora, C.G., Carminati, B., Ferrari, E.: User similarities on social networks. Soc. Netw. Anal. Min. **3**(3), 475–495 (2013)
3. Anderson, A., Huttenlocher, D.P., Kleinberg, J.M., Leskovec, J.: Effects of user similarity in social media. In: Proceedings of the Fifth International Conference on Web Search and Web Data Mining, pp. 703–712 (2012)
4. Bhattacharyya, P., Garg, A., Wu, S.F.: Analysis of user keyword similarity in online social networks. Soc. Netw. Anal. Min. **1**(3), 143–158 (2011)

5. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: Proceedings of SDM 06 Workshop on Link Analysis, Counterterrorism and Security (2006)
6. Huang, S., Tang, Y., Tang, F., Li, J.: Link prediction based on time-varied weight in co-authorship network. In: Proceedings of the 18th International Conference on Computer Supported Cooperative Work in Design, pp. 706–709 (2014)
7. Liben-Nowell, D., Kleinberg, J.M.: The link-prediction problem for social networks. JASIST **58**(7), 1019–1031 (2007)
8. Newman, M.: Clustering and preferential attachment in growing networks. Phys. Rev. E - Stat. Nonlinear Soft Matter Phys. **64**(2), 251021–251024 (2001)
9. Richard, E., Gaïffas, S., Vayatis, N.: Link prediction in graphs with autoregressive features. J. Mach. Learn. Res. **15**(1), 565–593 (2014)
10. Da Silva Soares, P.R., Prudncio, R.B.C.: Time series based link prediction. In: The 2012 International Joint Conference on Neural Networks, pp. 1–7 (2012)
11. 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, San Francisco, CA, USA, 13–17 August 2016, pp. 1225–1234 (2016)
12. Zhou, T., Lü, L., Zhang, Y.C.: Predicting missing links via local information. Eur. Phys. J. B **71**(4), 623–630 (2009)