Exploiting Node Similarity Based on Graphical Markov Models for Link Prediction

Jing Sun$^1$ and Zhijie Lin$^2$

1 Big Data Platform, East China University of Political Science and Law, Room 513, Building Mingshi, Road Longyuan, District Songjiang, Shanghai, China
2 School of Electric Information Engineering, Shanghai Dianji University, Shanghai, China
Email: jingsuncs@126.com

Abstract. Link prediction has been attached more attention in recent years. In this paper, we develop a link prediction method which has a unique perspective on using network structures. The key idea is to exploit the relationship information based on Graphical Markov models (GMM) for designing similarity indices. Specifically, networks with GMM were modeled to capture the relational influence of nodes by taking multi-hop neighbors into consideration. Then, link ties are measured for supporting relationship prediction based on the theory of weak and strong ties. Moreover, the proposed method can be used to predict the emergence of future relationships between the nodes. Finally, empirical studies on real-world dataset demonstrate that the benchmark of the proposed method improves with a significant margin compared with other methods.

Keywords. Link prediction; graphical Markov models; similarity indices; weak ties.

1. Introduction

In this paper, the problem of link prediction from a network structure perspective is focused. The present goal is to design efficient similarity measures for predictions. One previous study found a correlation between the number of common neighbours of $v_i$ and $v_j$ at time $t$ and the probability that they would collaborate in the future [1]. Nodes with more common neighbours have been found to be more similar to each other [2]. However, most existing methods [3-7] compute the similarity of objects directly from their local nearest neighbours, which limits the scope of extended predictions. For example, the complex characteristics of the traffic networks leads to the nodes be affected by higher hop neighbours. There is a need to mine the comprehensible relational influence taking the influence of global into account rather than only that of the local neighbourhood. Graphical Markov models (GMM) are usually constructed using the global structure of the graph, thereby producing efficient modeling, inferences, and probabilistic calculations [8]. In this way, all the link predictions shown in this paper are based on GMM. Additionally, the weak ties theory demonstrates that people usually get useful information and learn of opportunities through their acquaintances (weak ties) instated of through close friends (strong ties) [9]. Motivated by this theory, we defined the role of link ties in link prediction, which is used to regulate the relative contributions of link ties to the predictors. After that, the proposed predictors were used for link prediction. All non-existent links are sorted in a descending order according to their predicted scores, and the links with the highest scores are determined to be the existing links.
2. Related Work
In this section, we review some research work related to the link prediction. In order to improve effectiveness of product recommendation, Ref. [10] used the rating information to adjust linkage-weight between nodes. Ref. [11] used a community-structure and the attribute information of the items for predicting probable links, proposed a recommendation algorithm. Ref. [12] proposed a new recommendation algorithm which incorporate domain knowledge with topological property. Ref. [13] took the product domain knowledge and time context knowledge into consideration in recommendation and proposed a domain knowledge-based link prediction algorithm in customer-product bipartite network to improve effectiveness of product recommendation in retail.

However, most existing metrics ignore the non-fixation path of weighted networks and the fact that the information about multi-hop neighbours is also beneficial for link prediction. Intuitively, if one can flexibly collect more information of the neighbours with multiple hops in a network, then it seems that the better performance would be achieved.

3. Preliminaries
In this section, we introduce several related notions and definitions which are necessary for presenting our approach.

3.1. Graphical Markov Models
Graphical Markov models (GMM) are multivariate statistical models which generate processes in single and joint response variables and conditional independences captured by graphs [12].

Problem definition: In this paper, the network is represented by an undirected graph \( G = (V, E, X) \). \( V \) is the set of nodes, \( E \subseteq V \times V \) is the set of edges among nodes. \( X \) is an attribute matrix \( (|V|^2 \times d) \) with each row \( x_i \) corresponding to the link between a pair of nodes (e.g. node \( v_i \) and \( v_j \)). The attributes captured by \( X \) include the degrees of \( v_i \) and \( v_j \), the number of common neighbors, the weight between \( v_i \) and \( v_j \) etc. \( W \) is the set describing the weights of edges in \( E \). Each edge \( e = (i, j) \in E \) represents an interaction between \( v_i \) and \( v_j \). Multiple links and self-connections are not allowed in \( E \). In this paper, according to the given similarity indices, a score \( f_{ij} \) was obtained. This score indicates the similarity between \( v_i \) and \( v_j \). For a particular node vs, the task of prediction is to find a predictive function such that the nodes can be linked to vs:

\[
f: (V, E, X, v_s) \rightarrow R
\]

Here, \( R \) is a set of inferred results which suggest that whether \( v_s \) will create links with the nodes in \( V \). The predictive function can output the probability for the existence of links between user \( v_s \) and \( v_i \). Any potential links that will be estimated are sorted in a descending order by such score, and the link with the highest score is considered the best prediction results.

3.2. Link Ties
In this paper, we introduce a parameter \( \delta \) to depict the role of link ties in relationship prediction. The value \( \delta \) can be changed from positive to negative. The formally definition is described as follows.

Definition: (Link Ties) Link ties mainly maintain the connectivity in communication or relationship networks. Here, parameter \( \delta \) is used to control the relative contributions of link ties and measure the similarity.

\[
\begin{align*}
\delta > 0; & \text{Strong ties} \\
\delta = 0; & \text{Do not consider the strength of link ties} \\
\delta < 0; & \text{Weak ties}
\end{align*}
\]

When \( \delta = 0 \), the present methods degenerate to obtain the topology information for each node.
4. Exploiting Node Similarity Based on GMM for Link Prediction

In this section, we present the proposed link prediction method in details. First, the method on measuring the social influence of nodes (RecScore) is given, and then the likelihood of link existence between \( i \) and \( j \) is measured based on the RecScore value. Finally, all potential links were sorted by score in a descending order, and the links with the highest scores were considered most likely to exist.

4.1. The RecScore

Here, we define RecScore to represent the influence of node in networks.

Definition: (RecScore) The RecScore of node \( v_i \), denoted as RecScore\(_i\), is recursively defined as follows:

\[
\text{RecScore} = \sum_{v_j \in \text{neighbours}(i)} \text{RecScore}_j \cdot p_{ji}
\]

(1)

The parameter to indicate the strength of link ties plays an important role for measuring the influence of nodes. For the measurement of the influence of each node, the must be present during the process to determine RecScore for each node in each network, which is resulted from that different networks may have different types of link ties. These ties must be regulated during the actual operation. Equation is here rewritten as follows:

\[
\text{RecScore} = \sum_{v_j \in \text{neighbours}(i)} \text{RecScore}_j \cdot \left( \frac{w_{ji} + 1}{\sum_{v_k \in \text{neighbours}(j)} (w_{kj} + 1)} \right)
\]

(2)

Calculation process. The calculation of RecScore can be regarded as a random walk procedure over the weighted networks. Before achieved to a steady-state, several iterations that perform random walk is required. RecScore\(_{n}^{(0)}\) is the \( n \)-th iteration of RecScore. First, assign the initial value of RecScore\(_{n}^{(0)}\), to 1 which can be changed for any real number based on the theory of weak and strong ties. Each node updates its RecScore\(_{n}^{(1)}\), using initial RecScore of its neighbors in the first round. After the \( n \)-th round, node \( v_i \) has searched the \( n \)-th hop neighbors and updated RecScore with RecScore\(_{n}^{(n)}\).

4.2. Numbering Measuring Similarity Based on the RecScore

After calculating the influence of each node (i.e., RecScore), the similarity of each pair of nodes can be calculated based on the RecScore. The present predictor indices are designed based on some existing similarity indices. Hence, those pre-existed indices are first introduced, which are described as follows (note that, \( f(i, j) \) denotes the similarity of each pair of nodes \( i \) and \( j \)).

As mentioned above, the present predictors are designed according to the RecScore value, and the present methods are also from the perspective of common neighbours. The main difference we also considered is that the information received from those common neighbours may constitute the information from nearest neighbours, second-hop neighbours, and the neighbours further away. Here, the strength of link ties in prediction is also taken into account (the parameter).

RecCN predictor. Having a large number of common neighbours, the two nodes are likely from a relationship in the future, the extended similarity index RecCN can be calculated as follows:

\[
f_{\text{RecCN}}(i, j) = \sum_{z \in \text{neighbours}(i) \cap \text{neighbours}(j)} \text{RecScore}(z)
\]

(3)

Here, \( (i) \) and \( (j) \) indicate the sets of nearest neighbours of node \( i \) and node \( j \) respectively. RecScore\((z)\) denotes the RecScore value of the common neighbour \( z \).

RecAA predictor. The weighted version of RA and RecAA is designed with the vector considering the weight of the edge, the RecScore of the node. RecAA is determined as follows:

\[
f_{\text{RecAA}}(i, j) = \sum_{z \in \text{neighbours}(i) \cap \text{neighbours}(j)} \frac{(w_{iz} + 1)^{\delta} + (w_{zj} + 1)^{\delta}}{\log(1 + \text{RecScore}(z))}
\]

(4)
RecRA predictor. Similarity measurement RecRA is defined by extending the RA metric. Three factors were considered, the weight of edges, the influence of nodes (RecScore), and the link ties coefficient. The RecRA is defined as follows:

\[
f_{\text{RecRA}}(i, j) = \frac{\sum_{z \in \text{nei}(j)} (w_{i,z}+1)^\delta + (w_{z,j}+1)^\delta}{\text{RecScore}(z)}
\]

When \( \delta = 0 \), the weight values were not considered. In this way, the RecRA can be used in unweighted networks. Only degree information of different hops’ neighbours need to be obtained for each node. Then links prediction can be made.

In summary, the link prediction process includes three steps. The procedure of the proposed method is shown in algorithm 1.

Algorithm 1: Exploiting node similarity for link prediction

1: Require: \( G; V, n_0 \) node set; \( E \) edge set; \( W \) weight set; node \( v_i \); \( \epsilon \): iterations threshold; \( \delta \): value of link ties.
2: Ensure: the prediction result set \( R \)
3: Initialize \( \text{RecScore}(0)_{v_i} = 1 \)
4: for \( n = 1 \) to \( \epsilon \) do
5: 
6: \[ p_{r_{i,j}} = \frac{\sum_{z \in \text{nei}(j)} (w_{i,z}+1)^\delta}{\sum_{z \in \text{nei}(j)} (w_{i,z}+1)^\delta} \]
7: send \( \text{RecScore}^{(n-1)}_{v_i} \), \( p_{v_i} \) to \( v_j \)
8: end for
9: receive all \( \text{RecScore}^{(n-1)}_{v_i} \), \( p_{v_i} \) from every \( v_i \rightarrow v_j \) \( (v_i \in \text{nei}(v_j)) \)
10: \( \text{RecScore}^{(n)}_{v_i} = \sum_{v_i \in \text{nei}(v)} \text{RecScore}^{(n-1)}_{v_j} \)
11: end for
12: for each \( v_i \in V \) do
13: for each \( v_j \in \text{nei}(v_i) \) do
14: calculate the \( f(i, j) \) value
15: \( R = R \cup v_j \)
16: end for
17: end for
18: Sort \( R \) by \( f(i, j) \) score, here \( f(i, j) = f_{\text{RecCN}}(i, j) \) or \( f(i, j) = f_{\text{RecAA}}(i, j) \) or \( f(i, j) = f_{\text{RecRA}}(i, j) \)
19: return \( R \)

5. Experimental Results

In this section, to verify the effectiveness of the proposed methods, we compare the performances of our methods against some others. The comparison algorithms are list as follows: Common Neighbours (CN) [3], Adamic-Adar (AA) [5], Resource Allocation (RA) [14], Weighted Resource Allocation (WRA), Weighted Common Neighbours (WCN), Weighted Adamic-Adar (WAA) [14]. We provide an empirical evaluation of the performances for these methods on real-world network data (e.g., Interactome). The evaluation metric contains NDCG, AUC, Precision, Recall and F-Score.

The codes are implemented in JAVA (algorithms) and Matlab (visualization). The experiments environment is x64 machine, Intel i5 2.26GHz CPU, 4GB RAM, Windows 7 operation system.

Figure 1 presents the results of overall performance of the proposed methods and six baseline algorithms in terms of five metrics. As can be seen, RecAA and RecRA consistently achieve highest prediction accuracy in terms of NDCG, AUC, Precision, Recall and F-Score. However, we also found some surprising cases in the reported results on Interactome dataset. For example, one of our proposed
methods, RecCN, shows high Also, WCN surprisingly obtain the highest recall accuracy and outperform other methods with a significant margin on this dataset. On the other hand, this also shows that our methods are able to achieve a balance between prediction accuracy and volatility.

![Figure 1. The overall performances on the interactome dataset.](image)

6. Conclusion
This paper focuses on the study of exploiting node similarity for link prediction. Specifically, the real networks were first modeled as Graphical Markov models, and then the relationship influence (RecScore) of each node was mined by considering the multi-hop neighbors based on GMM. The proposed methods also incorporate link ties into the calculation of RecScore. The final similarity measurements, which are based on RecScore, render these link prediction results more accurate. Experiments on real dataset showed how all the information used can improve the prediction, and the present methods were found to outperform several baseline methods.

Acknowledgments
This research was partly supported by 2019 Shanghai University Young Teacher Training Funding Project. Also, it was supported by the scientific research project of East China University of Political Science and Law.

References
[1] Newman M E J 2003 The structure and function of complex networks SIAM Review 45 (2) 167-256.
[2] Liben-Nowell D and Kleinberg J 2007 The link-prediction problem for social networks Journal of the American Society for Information Science and Technology 58 (7) 1019-1031.
[3] Lorrain F and White H C 1977 Structural equivalence of individuals in social networks Social Networks 1 (1) 67-98.
[4] Jaccard P 1901 Etude comparative de la distribution florale dans une portion des alpes et des jura Bulletin del la societe vaudoise des sciences naturelles 37 (142) 547-579.
[5] Adamic L A and Adar E 2003 Friends and neighbors on the web Social Networks 25 (3) 211-230.
[6] Barabási A L and Albert R 1999 Emergence of scaling in random networks Science 286 509-512
[7] Ravasz E, Somera A L and Mongru D A et al. 2002 Hierarchical organization of modularity in metabolic networks Science 297.
[8] Levitz M and Madigan P D 2003 Correction: Separation and completeness properties for amp chain graph Markov models The Annals of Statistics 31 (1) 348.
[9] Granovetter M S 1973 The strength of weak ties American Journal of Sociology 78 (6) 1360-1380.
[10] Cui Y, Zhang L, Wang Q, Chen P and Xie C 2016 Heterogeneous network linkage-weight based link prediction in bipartite graph for personalized recommendation Procedia Computer Science 91 953-958.
[11] Talasu N, Jonnalagadda A, Pillai S S A and Rahul J 2017 A link prediction based approach for recommendation systems 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) pp 2059-2062.
[12] Li J, Zhang L L, Meng F and Li F 2014 Recommendation algorithm based on link prediction and domain knowledge in retail transactions International Conference on Information Technology and Quantitative Management (ITQM) pp 875-881.
[13] Zhang L, Li J, Meng F, Zhang Q and Teng W 2019 Domain knowledge-based link prediction in customer-product bipartite graph for product recommendation International Journal of Information Technology and Decision Making 18 (01) 311-338.
[14] Zhou T, Lü L and Zhang Y C 2009 Predicting missing links via local information European Physical Journal B 71 (4) 623-630.