# Community-Based Link Prediction in Social Networks

Rong Kuang(✉), Qun Liu, and Hong Yu

Chongqing Key Laboratory of Computational Intelligence,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, People's Republic of China
kuangrongcom@163.com

**Abstract.** Link prediction has attracted wide attention in the related fields of social networks which has been widely used in many domains, such as, identifying spurious interactions, extracting missing information, evaluating evolving mechanism of complex networks. But all of the previous works do not considering the influence of the neighbors and just applying in small networks. In this paper, a new similarity algorithm is proposed, which is motivated by the herd phenomenon taking place on network. Moreover, it is found that many links are assigned low scores while it has a longer path. Therefore, if such links the longer path has not been taken into account, which can improve the efficiency of time further, especially in large-scale networks. Extensive experiments were conducted on five real-world social networks, compared with the representative node similarity-based methods, our proposed model can provide more accurate predictions.

**Keywords:** Link prediction · Community structure · Common neighbor · Herd phenomenon

## 1 Introduction

The graphs are always used to describe the complex networks in which nodes stand for individuals or organizations and edges represent the interaction among them. The purpose of link prediction is to estimate the possibility of unknown links according to the known links [1]. And the link prediction has been applied in many relation networks such as underground relationships between terrorists [2], prediction of being actor [3], recommendation of friends for new members [4], and so on.

Finding the similar nodes based on similarity calculation are basic methods in link prediction which have low space complexity and low time consumption. In these methods, for each pair of nodes, *x*, *y*, the *s(x, y)* is used to define the similarity between *x* and *y* [5]. The core of these methods is to find a good criterion of similarity which plays an effective role on providing appropriate result in link prediction.

This paper is organized as follows. A short description of studies that had been done is presented in Sect. 2. In Sect. 3, the problem of link prediction and typical evaluation methods are described. A novel method is put forward in Sect. 4. Section 5 contains the results analysis of experiments by comparing our method with other

Y. Tan et al. (Eds.): ICSI 2016, Part II, LNCS 9713, pp. 341–348, 2016.
DOI: 10.1007/978-3-319-41009-8_37

previous works. Finally in Sect. 6, we summarize the features of the proposed model and the prospect for the future.

## 2 Related Work

The mainly methods used in link prediction can be classified into two categories: topology-based methods and learning-based methods. Topology-based methods only use the local node features which have low space complexity and low time consumption. Zhang and Wu [6] put forward that 3 hops common neighbor give valuable contributions to the connection likelihood. Yin et al. [7] proposed the node link strength algorithm (*SA*), which considered both common neighbors and link strength between each common neighbor node. The accurate community structure [8] can also improve the accuracy of prediction. Valverde-Rebaza and Lopes [9] combined topology with community information by considering users' interests and behaviors. The learning-based methods often have better performance than topology-based algorithms, many studies [10–12] show that using attributes of nodes and links (such as users' ages, interests, characteristics and friends) can significantly improve the link prediction performance. Menon and Elkan [13] treated link prediction as matrix completion problem and extend matrix factorization method to solve the link prediction problem. Ozcan and Sule [14] utilizes the network temporal information along with modeling the combination of topological metrics and link occurrences information, this method can predict new link information and repeat occurrences of existing links.

Our approach is based on similarity by considering the information of communities. Meantime we introduce a new method which based on herd phenomenon to show the influence of neighbors. The result shows that our model has good performance.

## 3 Problem Description and Evaluation Method

### 3.1 Problem Description

Given an undirected simple network $G\ (V,\ E)$, where $V$ represents the set of vertices, $E$ represents the set of edges. For each pair of nodes, $x, y \in V$, every algorithm referred to in this paper assigns a score $s_{xy}$. This score can be viewed as a measure of similarity between nodes $x$ and $y$, higher score implies higher likelihood that two nodes are connected, and vice versa.

Empirically, the known link set E would normally be randomly divided into a training set of $E_p$ and a test set of $E_t$ before the experiment, which the training set contains 90 % links and the remaining 10 % of links are constituted in the test set.

### 3.2 The Algorithm Based on Node Similarity

We will apply four representative traditional algorithms which based on node similarity to five data sets, and compared the accuracy with our method later.

Common Neighbors (*CN*) [15]: it means the possibility of a link between two nodes is equal to the number of their common neighbors. Let $\Gamma(x)$ stand for the neighbor of *x*, then the score of *CN* can be calculated as follows:

$$S_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|. \tag{1}$$

Adamic-Adar Index (*AA*) [16]: at first, it is used to calculate the similarity of two pages. Both common neighbor nodes and the degrees of common neighbor nodes are taken into account.

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log\Gamma(z)}. \tag{2}$$

Resource Allocation Index (*RA*) [17]: namely, the problem whether the node *x* and node *y* generate a link in the network is converted to analyze the process of *x* transferring resources to *y*, and the calculation of the similarity of *x* and *y* is converted to calculate the amount of resources that *y* has received from *x*.

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\Gamma(z)}. \tag{3}$$

Cohesive Common Neighbors (*CCN*) [6]: the three hops common neighbor can also give valuable contributions to the connection likelihood.

$$S_{xy}^{CCN} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\Gamma(z)} + \sum_{\substack{m \in \Gamma(x), n \in \Gamma(y) \\ m \in \Gamma(n)}} \frac{1}{\Gamma(m) * \Gamma(n)}. \tag{4}$$

### 3.3 Evaluation Metrics

The area under the receiver operating characteristic curve (*AUC*) [18] and precision [19] are widely used to quantify the accuracy of link prediction. Providing the rank of all non-observed links, the *AUC* value can stand for the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link.

$$AUC = \frac{n' + 0.5n''}{n}. \tag{5}$$

Different from *AUC*, *Precision* can be defined as the ratio of the number of relevant links to selected links. In our experiments, all the missing and nonexistent links are ranked in decreasing order firstly. The formula of *Precision* is as following:

$$Precision = m/L. \tag{6}$$

In which, we concentrate on the *top-L* (here $L = 100$) links in recovery, and $m$ represents the numbers of actual links in the testing set $E_p$.

# 4 Community-Based Evolution Algorithm

## 4.1 Problem Presentation and Evolution Model

Social networks are highly structured, and now the scale of the network is very huge. Compared with traditional algorithm model, it is difficult to apply to large datasets. Inspired by birds of a feather flock together, and Abir et al. [8] proved that constructing communities can improve the accuracy of prediction, and it can also greatly advance time efficiency, that is the reason we introduce community.

We assume that the scale of a community will not change along with the evolution. The evolution model of the communities is defined as follows:

**Community Detection.** We introduce density-based link clustering algorithm [20] to divide the network into communities. For convenience, we denote the community set as $C$ in the following text.

**Evolution.**

(a) Starting from a node p in community $C_i$ (one of community in $C$).
(b) Selecting nodes around p which $L_{pq} < AVG_L$ ($L_{pq}$ represents the distance between node $p$ and node $q$, $AVG_L$ represents the average distance of the network.) and $D_q \geq D_p$ ($D_x$ represents the degree of node x). $Q$ denotes the set of the selected node.
(c) Calculating the similarity between $p$ and $Q_i$ by Eq. (7).
(d) Inserting the link to network whose score are relatively high.
(e) Repeat steps (a) to (d) until every node in $C_i$ is considered, as well as every community in $C$.

## 4.2 *SN* (Similar to the Neighbors) Algorithm

In this paper, the difference with previous algorithms is that we start from a single node $p$, and then we just find possible related nodes around $p$, rather than compare to all other nodes. In addition, the relationship of the node *p's* neighbor with node $q$ can also make important influence on the result when we analyze whether node $p$ has a connection with node $q$. Therefore we need consider not only the relationship of $p$ and $q$, but also the potential influence between neighbors of $p$ and $q$. Finally, we proposed a method based on the similarity with neighbors which we call it Similar to the Neighbors (*SN*). The definition of the influence of neighbors can be expressed as follow:

$$S_{xy}^{SN} = (1-\delta) \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\Gamma(z)} + \delta \max_{z \in \Gamma(x)} \sum_{t \epsilon \Gamma(z) \cap \Gamma(y)} \frac{1}{\Gamma(t)}. \tag{7}$$

# 5 Experiments and Results Analysis

## 5.1 Data Analysis

We collect five real-world social network datasets, which are different from previous. All these networks contain large amounts of data. (1) *Facebook* [21]: Facebook is a social networking service website. The data set contains 4039 users and 88234 friendship links. (2) *Twitter* [21]: Twitter is an American social networking and micro-blog services, and it is one of the ten most visited Internet sites over the world. And it includes 81304 users and 1768149 connections. (3) *YouTube* [22]: YouTube is the world's largest video site, which contains 1134890 nodes and 2987624 edges. (4) *Sina web*: Sina web provides micro-blog serve, users can use one sentence to express what he saw, heard and thought or sent a picture to share with friends. The data set we obtained contains 60955 nodes and 311056 edges. (5) *DBLP* [22]: DBLP integrates English literature in computer field. It contains 317080 authors and 1049866 connected. All figures are typical of social networks and their topological features are shown in Table 1.

**Table 1.** The basic topological features of five example networks.

| *Networks* | $\|V\|$ | $\|E\|$ | <*k*> | <*d*> | *C* |
|---|---|---|---|---|---|
| *Facebook* | 4039 | 88234 | 43.691 | 3.693 | 0.606 |
| *DBLP* | 317080 | 1049866 | 11.256 | 4.416 | 0.632 |
| *YouTube* | 1134890 | 2987624 | 5.265 | 4.069 | 0.081 |
| *Sina web* | 60955 | 311305 | 3.802 | 4.251 | 0.083 |
| *Twitter* | 81306 | 1768149 | 43.494 | 3.583 | 0.565 |

$|E|$ and $|V|$ are the total numbers of links and nodes respectively. <*k*> is the shortest average degree of the example networks. <*d*> is the average shortest distance between node pairs. *C* stands for the clustering coefficient of every network.

## 5.2 Results

In order to get the proportion of parameters in Eq. (7). Figure 1 shows the performance of SN measured by AUC as a function of $\delta$. According to Fig. 1, AUC changes with $\delta$. We choose $\delta$ when AUC obtain the highest, and then compare with other methods later.

From the Tables 2 and 3, the simulation results of our algorithm and other previous algorithms are shown, where the black and bold items represent the highest accuracies. We simplify the computational complexity by introducing community information.

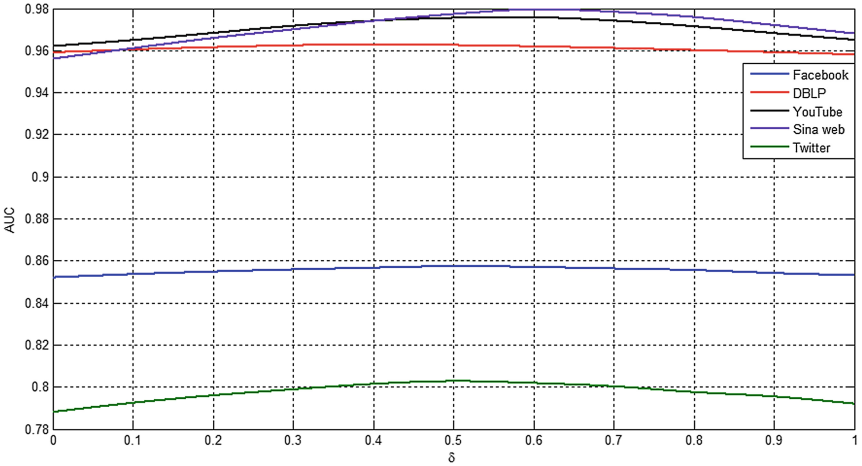

**Fig. 1.** The performance of SN measured by AUC as a function of $\delta$ in Eq. (7) (Colour figure online)

**Table 2.** The prediction accuracy measured by *AUC* on five real-world networks. Each number is obtained by averaging over 10 implementations with independently random partitions of testing set and training set. The abbreviations, *AA*, *CN*, *RA* and *CCN*, stand for Adamic-Adar Index, Common neighbors, Resource Allocation Index and Cohesive Common Neighbors, respectively. The parameter for *SN*, ε and δ are divergence within different data sets, the parameter values get from Fig. 1.

| *Measures* | *CN* | *AA* | *RA* | *CCN* | *SN* |
|---|---|---|---|---|---|
| *Facebook* | **0.861** | 0.849 | 0.852 | 0.856 | 0.858 |
| *DBLP* | 0.950 | **0.963** | 0.959 | 0.961 | **0.963** |
| *YouTube* | 0.956 | 0.974 | 0.962 | 0.975 | **0.976** |
| *Sina web* | 0.953 | 0.965 | 0.956 | 0.978 | **0.980** |
| *Twitter* | 0.729 | 0.748 | 0.788 | **0.790** | 0.789 |

**Table 3.** The prediction accuracy measured by the precision metric on five real-world networks.

| *Measures* | *CN* | *AA* | *RA* | *CCN* | *SN* |
|---|---|---|---|---|---|
| *Facebook* | 0.149 | **0.367** | 0.326 | 0.315 | 0.356 |
| *DBLP* | 0.132 | 0.526 | 0.229 | **0.531** | 0.416 |
| *YouTube* | 0.207 | 0.349 | 0.326 | 0.478 | **0.504** |
| *Sina web* | 0.413 | 0.469 | 0.545 | 0.538 | **0.547** |
| *Twitter* | 0.139 | **0.212** | 0.181 | 0.203 | 0.199 |

As can be seen form Table 2, our algorithm outperforms other algorithms in *YouTube* and *Sina web*, and the accuracies of *CCN* are close to our method in these two networks, our model also performs well in *DBLP*. As can be seen from Table 2, all method performance are poor in *Facebook* and *Twitter*, just because the method based on common neighbors will lost the meaning of original resource allocation in rich-club [23] networks.

As the result Table 3 we can see, the algorithm we purposed, can give the better forecast results than other algorithms in the first type of network. From the above discussion, we know that fewer common neighbors of two nodes will cause about lower scores of these both nodes, though they have connection among them. Therefore, the accuracy of the forecasts will cut down when the pair of nodes, which get lower score, are ignored. In our model, motivated by the herd phenomenon which means any node likes to follow the leader neighbors as in the reality, we consider the influence of the neighbors when analyzing whether there is a relationship between two nodes. Thus, it can improve the score among many pair of nodes. We can also see from Table 3, *AA* get better score than other methods in the second type of network.

## 6 Conclusion

In this paper, we propose a new algorithm to predict the missing links in the network, and also compare it with some other typical link prediction algorithms based on nodes similarities. The difference between our approach and other previous work is that we introduced the community message which motivated by birds of a feather flock together phenomenon, and it can greatly improve time efficiency, and ensure the accuracy.

Numerical results on the five real-world data sets of social network indicate that: (1) There is no such algorithm which can be applied to all kinds of networks due to the different characteristic of networks. (2) Many social networks can be roughly divided into three categories, we called Hub network (clustering coefficient is very small and have low average degree), Uniform network (high clustering coefficient and large degree) and Trend center network (contains small average degree of node and have high clustering coefficient). The algorithms based on common neighbors have better performance in Hub network, but all existing algorithms get poor scores in Uniform network called rich-club network.

In this paper, we found that the social network can be divided into three types roughly by its average degree and clustering coefficient. In the future, we will committed to improve the prediction accuracy of each type of network through using different methods respectively.

**Acknowledgments.** This work is partly funded by the National Nature Science Foundation of China (61075019) and the Natural Science Foundation of Chongqing (CSTC2014jcyjA40047) and the Municipal Education Commission research project of Chongqing (KJ1400403).

## References

1. Lü, L.Y., Zhou, T.: Link prediction in complex networks: a survey. Phys. A: Stat. Mech. Appl. **390**, 1150–1170 (2011)
2. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature **453**, 98–101 (2008)

3. O'Madadhain, J., Hutchins, J., Smyth, P.: Prediction and ranking algorithms for event-based network data. ACM SIGKDD **7**(2), 23–30 (2005)
4. Han, X., Wang, L.Y., Chen C., Farahbakhsh, R.: Link prediction for new users in social networks. In: ICC, pp. 1250–1255 (2015)
5. Lü, L.Y., Pan, L.M., Zhou, T., Zhang, Y.C., Stanley, H.E.: Toward link predictability of complex networks. PANS **112**, 2325–2330 (2015)
6. Zhang, W.Y., Wu, B.: Accurate and fast link prediction in complex networks. In: Natural Computation (ICNC), pp. 653–657 (2014)
7. Yin, G., Yin W.S., Dong, Y.X.: A new link prediction algorithm: node link strength algorithm. In: SCAC, pp. 5–9 (2014)
8. Abir, D., Nilloy, G., Soumen, C.: Discriminative link prediction using local links, node features and community structure. In: Data Mining (ICDM), pp. 1009–1018 (2013)
9. Valverde-Rebaza, J., Lopes, A.: Exploiting behaviors of communities of twitter users for link prediction. Soc. Netw. Anal. Min. **3**, 1063–1074 (2013)
10. Li, X., Chen, H.: Recommendation as link prediction in bipartite graphs: a graph kernel-based machine learning approach. Decis. Support Syst. **54**, 880–890 (2013)
11. Scellato, S., Noulas, A., Mascolo, C.: Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, pp. 1046–1054 (2011)
12. Chen, Z., Zhang, W.: A marginalized denoising method for link prediction in relational data. In: Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, pp. 298–306 (2014)
13. Menon, K., Elkan, C.: Link prediction via matrix factorization. In: Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, pp. 437–452 (2011)
14. Ozcan, A., Sule, O.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)
15. Zhu, M., Zhou, Y.: Density-based link clustering algorithm for overlapping community detection. J. Comput. Res. Dev. 2520–2530 (2013)
16. Newman, M.E.J.: Clustering and preferential attachment in growing networks. Phys. Rev. E **64**, 025102 (2001)
17. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**, 211–230 (2003)
18. Zhou, T., Lü, L., Zhang, Y.C.: Predicting missing links via local information. Eur. Phys. J. B **71**, 623–630 (2009)
19. Hanley, J.A., McNeil, B.J.: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology **148**, 839–843 (1983)
20. Zhu, M., Meng, F.R., Zhou, Y.: Density-based link clustering algorithm for overlapping community detection. 2520–2530 (2013)
21. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM TOIS **22**, 5–53 (2004)
22. McAuley, J., Leskovec, J.: Learning to discover social circles in ego networks. Adv. Neural Inf. Process. Syst. **25**, 548–556 (2012)
23. Zhou, S., Mondragon, R.J.: The rich-club phenomenon in the internet topology. IEEE Common. Lett. **8**, 180–182 (2004)