International Journal of Engineering & Technology, 7 (4.10) (2018) 274-277

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Weighted path based Link Prediction in Social Networks using Bounded Length of Separation between Nodes

Srilatha P[1]* and Manjula R[2]

[1]School of Computer Science and Engineering, Vellore Institute of Technology,
Vellore, Tamil Nadu - 632014
**Corresponding author E-mail: sreelatha.pulipati@gmail.com*

Abstract

The problem of link prediction in online social networks like facebook, myspace, Hi5 and in other domains like biological network of molecules, gene network to model disease have became very popular because of the structural connections and relationships among the entities. The classical methods of link prediction based on the topological structure of the graph exploit all different paths of the network which are being computationally expensive for large size of networks. In this paper, incorporating the small world phenomenon, the proposed algorithm traverses all the paths of bounded length by considering clustering information and the connection pattern of the edges as weights on the edges in the graph. As a result, the proposed algorithm will be able to predict accurately than the existing link prediction algorithms. Our analysis and experiment on real world networks shows that our algorithm outperforms other approaches in terms of time complexity and the prediction accuracy.

Keywords: *Social Networks; Link Prediction; Bounded Length.*

1. Introduction

The study of complex networks has become very popular in many branches of science. One such type of networks is the social networks and analysis of social networks has become an interesting and challenging issue that has recently attracted attention from researchers in various domains. Possible link prediction of new friendship (can be thought of as new connection that was not before) between two individuals in the social networks had been studied widely as it applications in many fields apart from only in online social networks. Using the notion of graph theory, the prediction of missing links between two nodes based on the topology of the graph and the node attribute information of existing nodes and the edges in the graph is commonly known as the link prediction problem [1]. In other words, given a snapshot of a graph in a time t, the goal of link prediction is to predict non existing edges in the graph at time t+1.

The link prediction problem has various applications which includes friend recommendations [2,3], protein- protein interaction in biological networks, finding co-author or experts [4], recommender systems like e-commerce websites in which a prediction is made on customer preferences in purchasing the items and also in other domains like metabolic networks, disease-gene network and biological networks. Several link prediction methods defined, considers topological information and the node attribute information. However the clustering information i.e., the node belonging to the same cluster tends to be more similar and so plays an important role in prediction of links. In [5,6] authors showed that the accuracy of similarity based link prediction methods are greatly improved by the inclusion of the clustering information. However the complexity of the similarity based metrics increases with the increase in the size of the network. To address the complexity of the algorithm, along with the cluster information and the connection pattern of the graph, we considered the bounded path length l to calculate the likelihood score of all the non observed links.

The rest of the paper is organized as follows. Section 2 formulates the problem definition. Section 3 reviews the relevant methods in the area of link prediction in social networks. Section 4 proposes the new algorithm for finding the likelihood scores of all the non existing edges. Section 5 discusses the experimental setup and evaluation metrics. Also, Section 5 gives the experimental results obtained by proposed algorithm and compares their performance with other similar methods. Section 6 presents the conclusion and the future work.

2. Problem Formulation

Generally social networks are considered as a graph. An undirected graph G = (V,E) is considered where V is the number of the nodes and E is the number of the edges. Let n = |V| and m = |E| then n(n - 1)/2 be the number of possible edges by removing self loops and multiple links. We denote the set of all possible edges by U. Then U - E gives set of all possible non observed or non existing edges in the graph. Thus the link prediction problem is formulated as finding the likelihood scores of all the non observed links. The high likelihood score between the two nodes represents the more likeliness that they are likely to connect by a link in near future. The likelihood scores of all the non existing links are calculated by incorporating clustering information in addition to the importance of the edge. For a given graph G, the clustering information C, and the given bounded path length l the new link prediction problem is to find the likelihood score $I(e_{ij}|G,C,l)$ for all $e_{ij} \in U - E$.

3. Related Work

The existing methods of link prediction are classified into 3 categories similarity based methods, probability based maximum likelihood algorithms and machine learning based algorithms [7]. Similarity based methods are popularly used methods for the link prediction. In the similarity based methods, a similarity score (also called as structural similarity score) is computed for all non existing methods. Based on the topological structural information of the network, the structural similarity indices can be classified as local indices, global indices and quasi local indices. The local indices are calculated by using local graph topological structure (node neighbourhood) [8,9,10]. Typical local indices include Common Neighbours, Jaccard Index, Adamic-Adar Index, Sorensen Index, Salton Index. Global indies are calculated by considering the global graph [11,12]. Typical global includes Katz index, Sim rank score, random walk with restart and average commute time. Quasi local indices are calculated by limiting the global information [12,13,14,15] and also considering the local information. Quasi local indices include local path index, local random walk, super imposed random walk and friend link score. However the accuracy of the link prediction methods based on the local information may be less effective because of insufficient information. In the second method[16], maximum likelihood estimation, the likelihood of any non existing links is calculated by using the probabilistic rules and the defined parameters. The probabilistic models estimate the likelihood by using the conditional probability. However the computational complexity of local information methods is lesser than the global information methods. The above traditional methods have the following disadvantages. First, they won't consider the cluster information as the nodes within the same cluster tends to be more similar than the node belonging to the different cluster. Second, they have higher complexity so they are not suitable for large scale networks. Considering the features of the real world networks, and the clustering information we propose a weighted path based link prediction algorithm that reduces the computational cost and can achieve higher prediction accuracy.

4. Weighted Path based Link Prediction (WLP)

In the proposed algorithm we incorporate the cluster information and the importance of the edge. The popular methods used to know the importance of the edge includes edge betweenness centrality values developed by Grevan and Newman [17]. It considers all the shortest paths between nodes in the graph going through the edge. K path edge betweenness centrality [18] is the another measure where the edge between values are calculated only for the specified length paths. Formal notion of the edge centrality measure are defined below:

Definition 1: Edge betweenness Centrality

For a given Graph G(V,E) edge betweenness centrality is given as follows:

$$B(e_{ij}) = \sum_{s \neq t \in V} \frac{\sigma_{st}(e_{ij})}{\sigma_{st}}$$

where, σ_{st} is the number of shortest paths from node s to node t and $\sigma_{st}(e_{ij})$ is the number of shortest paths from node s to node t passing through the edge e_{ij}.

Definition 2:

K – path edge centrality

For a given graph G(V,E), K-path edge centrality is given as follows

$$B_k(e_{ij}) = \sum_{s \in V} \frac{\sigma_s{}^k(e_{ij})}{\sigma_s{}^k}$$

where S is the start node having at least K-path length. $\sigma_s{}^k(e_{ij})$ is the number of all shortest paths going through the edge e_{ij} from the start node S and extending to the path length K.

Definition 3:

Cluster Information

Given a graph G (V,E) and a set of K community labels, weights are assigned to all the edges E based on the locality of the node in the community structure. All the nodes belonging to the same community are given the same labels. That is, if C is cluster then C_i is the community label given to all the nodes in the cluster C such that that i ∈ K. We assign a positive weight if the edge belongs to the same cluster otherwise negative weight is assigned. The weights assigned to the existing edges have positive and negative impact on the calculation of likelihood scores of the non existing edges.

Therefore cluster information (CI) on the edge is given as follows:

$$CI(e_{ij}) = \begin{cases} +\frac{SC_i}{n} & if\ C_i = C_j \\ -\frac{SC_i}{n} & Otherwise \end{cases}$$

where SC_i represents the size of the cluster that contains the node i. n represents the number of nodes in the Graph G. C_i represents the community label for node i. Cj represents the community label for node j.

By considering the cluster information and the edge betweenness centrality value the total importance of the edge eij from the node i to node j is computed as

$$I(e_{ij}) = CI(e_{ij}) \times B(e_{ij})$$

To find the total likelihood score of the non existing links i.e., e_{ij} ∈ U - E between the nodes i and node j , we considered the importance of all the paths of bounded length l and is defined as follows:

$$TotallikelihoodScore(e_{ij}) = \sum_{k=2}^{l} paths^k{}_{(ij)} \sum_{x \in paths^k{}_{(ij)}} I(x)$$

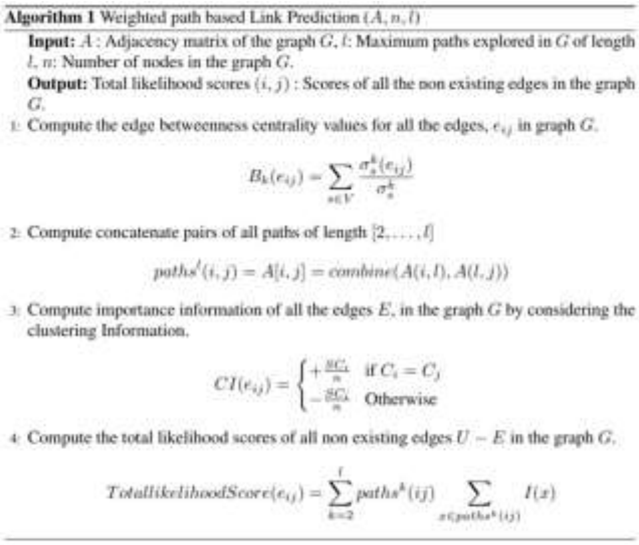
$paths^k$ is defined as the list of all k-length paths from the node i to node j.

Incorporating the above definitions we explain the weighted path based link prediction algorithm. The pseudo code of proposed method is given in algorithm 1. The proposed algorithm has five steps.

1. For the a graph G(V,E) compute the edge between centrality values for all the existing edges i.e., e_{ij} ∈ E.
2. Compute the adjacency matrix where each entry A[i,j] is a list of all paths of bounded length 2 to l from the node i to node j.
3. For the graph G(V,E) find the communities and assign the labels to all the nodes based on the belongingness property.
4. Assign the weight to all the edges in such a way that edges gets positive value if the node i and node j belongs to the same community and gets negative value if the node i and node j belongs to the different community.
5. For the graph G(V,E) compute the total likelihood score of all the non existing edges i.e., e_{ij} ∈ U – E where U is the total possible edges and E represents the existing edges in the graph G.

4.1 Complexity of Proposed WLP algorithm

The time complexity of the proposed algorithm is dependent on the computation cost of centrality values, cluster information and the likelihood scores. The computation cost of edge betweenness centrality values is O(n^2) [18] and the \$k\$-path edge centralities values is O(mn) [19]. We have considered the Fast Newman algorithm for detecting the communities. The Newman algorithm costs O(m^2n) [18]. To compute the likelihood scores we have explored all the paths from 2 to l and concatenated all the paths. Each entry of the adjacency matrix A[i,j] contains concatenated paths from node i to node j. The computation cost thus reduces to O(nh)[14]. h is the average degree in a network. Thus the total complexity of the proposed algorithm is O(n^2).



```
Algorithm 1 Weighted path based Link Prediction $(A, n, l)$
    Input: $A$ : Adjacency matrix of the graph $G$, $l$: Maximum paths explored in $G$ of length $l$, $n$: Number of nodes in the graph $G$.
    Output: Total likelihood scores $(i, j)$ : Scores of all the non existing edges in the graph $G$.
1: Compute the edge betweenness centrality values for all the edges, $e_{ij}$ in graph $G$.
        $B_k(e_{ij}) = \sum_{s \in V} \frac{\sigma_s^k(e_{ij})}{\sigma_s^k}$
2: Compute concatenate pairs of all paths of length $[2, \ldots, l]$
        $paths^l(i,j) = A[i,j] = combine(A(i,l), A(l,j))$
3: Compute importance information of all the edges $E$, in the graph $G$ by considering the clustering Information.
        $CI(e_{ij}) = \begin{cases} +\frac{\delta C_i}{n} & \text{if } C_i = C_j \\ -\frac{\delta C_i}{n} & \text{Otherwise} \end{cases}$
4: Compute the total likelihood scores of all non existing edges $U - E$ in the graph $G$.
        $TotallikelihoodScore(e_{ij}) = \sum_{k=2}^{l} paths^k(ij) \sum_{x \in paths^k(ij)} I(x)$
```

5. Experimental Setup and Evaluation Metrics

The algorithm is implemented using the R tool by making use of igraph package. We used standard precision and AUC metrics as accuracy measure for link prediction. To test the accuracy of the proposed algorithm, the existing edges and the observed edges in the graph are divided into 2 sets. The training set ET and the test set EP where ET U EP = E and ET Ո EP = Φ.

AUC: AUC scores[20] are interpreted as probability that a randomly selected nonexistent link U-E is less than a randomly selected link in the test set EP. At each time we randomly choose a non existing link and a missing link to compare their scores. We perform n independent comparisons and if there are n' times missing edges having higher score and n" times both have the same scores then AUC score is calculated as

$$AUC = \frac{n' + 0.5 \times n''}{n}$$

Precision: Given the ranking of all the non existing edges by the algorithm, the precision is defined as the ratio of m right links taken from the top L predicted links precision[21] is calculated as

$$precision = \frac{m}{L}$$

Data Sets: In this paper we have considered real world datasets[22] like Network of US political Blogs (PB), US airport network (USAir), electrical power grid of the western US (power grid).

The following table summarizes basic topological features of the networks.

Table 1: Topological features of giant components

Network	N	E	Nc	e	C	a
USAir	232	1365	232/1	0.44	0.749	-0.028
PB	1224	19090	1222/2	0.397	0.346	-0.221
Grid	4941	6594	4941/1	0.056	0.107	0.003

Table 1 summarizes the topological features of the large networks. In the table, N is the total number of nodes, M is the total number of edges, Nc is the number of components connected together and the size of largest one, e is the efficiency of the network, C is the clustering coefficient and a is the assortative coefficient.

Table 2 and 3 gives the AUC and precision values of the proposed method -- WLP and comparison with other methods defined in the literature. From the experimental values we can infer that the proposed method provides better output.

Table 2: AUC values by proposed WLP algorithm along with other similarity based methods

	USAir	PB	Grid
CN [23]	0.939	0.926	0.638
Salton [24]	0.926	0.878	0.612
Jaccard [25]	0.899	0.865	0.622
Sorenson [26]	0.917	0.885	0.633
WLP	0.919	0.9	0.65

Table 3: Precision values by proposed WLP algorithm along with other similarity based methods

	USAir	PB	Grid
CN [23]	0.6585	0.2356	0.0364
Salton [24]	0.0976	0.0012	0.0121
Jaccard [25]	0.1037	0.0407	0
Sorenson [26]	0.0976	0.0024	0.0121
WLP	0.75	0.34	0.4

6. Conclusions and Future Scope

In this paper, weighed path based link prediction algorithm is proposed by incorporating the edge betweenness values , clustering information and the connected information of the bounded paths of length [2,…,l] for finding the likelihood scores of all the non observed links. The comparative analysis is performed on various real world datasets. AUC and precision are the evaluation methods used in the paper to evaluate the accuracy of the algorithm. The experiment results on real world data sets tabulated in Table 2 and 3, shows that clustering information, and the connected information can improve the accuracy of the link prediction over other unsupervised link prediction metrics. In the future work we will consider the other network that follows power law distribution and also we will include the node attributes information to achieve higher accuracy in link prediction.

References

[1] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla, “New perspectives and methods in link prediction”. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.

[2] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Chaosheng Fan, and Xueqi Cheng. “Informational friend recommendation in social media”. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1045–1048. ACM, 2013.

[3] Jacob W Bartel and Prasun Dewan. “Evolving friend lists in social networks”. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 435–438. ACM, 2013.

[4] Maryam Fazel-Zarandi, Hugh J Devlin, Yun Huang, and Noshir Contractor. “Expert recommendation based on social drivers, social network analysis, and semantic data representation”. In *Proceedings of the 2nd international workshop on information heterogeneity and fusion in recommender systems*, pages 41–48. ACM, 2011.

[5] Mark EJ Newman. “Clustering and preferential attachment in growing networks”. *Physical review E*, 64(2):025102, 2001.

[6] Yangyang Liu, Chengli Zhao, Xiaojie Wang, Qiangjuan Huang, Xue Zhang, and Dongyun Yi. “The degree-related clustering coefficient and its application to link prediction”. *Physica A: Statistical Mechanics and its Applications*, 454:24–33, 2016.

[7] Tsuyoshi Murata and Sakiko Moriyasu. “Link prediction based on structural properties of online social networks”. *New Generation Computing*, 26(3):245–257, 2008.

[8] Abir De, Niloy Ganguly, and Soumen Chakrabarti. “Discriminative link prediction using local links, node features and community structure”. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1009–1018. IEEE, 2013.

[9] Jing Wang and Lili Rong. “Similarity index based on the information of neighbor nodes for link prediction of complex network”. *Modern Physics Letters B*, 27(06):1350039, 2013.

[10] Ismail Güne¸s, ¸Sule Gündüz-Ö˘güdücü, and Zehra Çataltepe. “Link prediction using time series of neighborhood-based node similarity scores”. *Data Mining and Knowledge Discovery*, 30(1):147–180, 2016.

[11] Glen Jeh and Jennifer Widom. “Simrank: a measure of structural-context similarity”. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[12] Bai Meng, Hu Ke, and Tang Yi. “Link prediction based on a semi-local similarity index.” *Chinese Physics B*, 20(12):128902, 2011.

[13] Weiping Liu and Linyuan Lü. “Link prediction based on local random walk”. *EPL (Europhysics Letters)*, 89(5):58007, 2010.

[14] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. “Fast and accurate link prediction in social networking systems.” *Journal of Systems and Software*, 85(9):2119–2132, 2012.

[15] Pulipati Srilatha and Ramakrishnan Manjula. “Similarity index based link prediction algorithms in social networks: A survey.” *Journal of Telecommunications and Information Technology*, (2):87, 2016.

[16] Linyuan Lü and Tao Zhou. “Link prediction in complex networks: A survey.” *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[17] Michelle Girvan and Mark EJ Newman. “Community structure in social and biological networks.” *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[18] Pasquale DeMeo, Emilio Ferrara, Giacomo Fiumara, and Angela Ricciardello. “A novel measure of edge centrality in social networks”. *Knowledge-based systems*, 30:136–150, 2012.

[19] Ulrik Brandes. “A faster algorithm for betweenness centrality”. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[20] James A Hanley and Barbara J McNeil. “The meaning and use of the area under a receiver operating characteristic (roc) curve”. *Radiology*, 143(1):29–36, 1982.

[21] Jesse Davis and Mark Goadrich. “The relationship between precision-recall and roc curves”. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[22] Bolun Chen, Ling Chen, and Bin Li. “A fast algorithm for predicting links to nodes of interest”. *Information Sciences*, 329:552–567, 2016.

[23] Francois Lorrain and Harrison C White. “Structural equivalence of individuals in social networks”. *The Journal of mathematical sociology*, 1(1):49–80, 1971.

[24] Gerard Salton and Michael J McGill. “Introduction to modern information retrieval”. 1986.

[25] Paul Jaccard. “Étude comparative de la distribution florale dans une portion des alpes et des jura”. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

[26] Thorvald Sørensen. “A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons”. *Biol. Skr*., 5:1–34, 1948