Link Prediction in Complex Networks: An Empirical Review

NANDINI, YV, TANGIRALA, Jaya Lakshmi <http://orcid.org/0000-0003-0183-4093> and ENDURI, Murali Krishna

Link Prediction in Complex networks: An Empirical Review

Y.V.Nandini[1], T.Jaya Lakshmi[1], and Murali Krishna Enduri[1]

School of Engineering and Sciences,
SRM University, Andhra Pradesh, India.
{nandini_y, jayalakshmi.t, muralikrishna.e}@srmap.edu.in

Abstract. Any real-world entity with entities and interactions between them can be modelled as a complex network. Complex networks are mathematically modelled as graphs with nodes denoting entities and edges(links) depicting the interaction between entities. Many analytical tasks can be performed on such networks. Link-Prediction (LP) is one of such tasks, that predicts missing/future links in a complex network modelled as graph. Link Prediction has potential applications in the domains of biology, ecology, physics, computer science and many more. Link prediction algorithms can be used to predict future scientific collaborations in a collaborative network, recommend friends/connections in a social network, future interactions in a molecular interaction network. The task of link prediction utilises information pertaining to the graph such as node-neighborhoods, paths etc. The main focus of this work is to empirically evaluate the efficacy of a few neighborhood-based measures for link prediction. Complex networks are very huge in size and sparse in nature. Choosing the candidate node pairs for future link prediction is one of the hardest tasks. Majority of the existing methods consider all node pairs absent of an edge to be candidates; compute prediction score and then the node pairs with the highest prediction scores are output as future links. Due to the massive size and sparse nature of complex networks, examining all node pairs results in a large number of false positives. A few existing works select only a subset of node pairs to be candidates for prediction. In this study, a sample of candidates for LP based are chosen based on the hop-distance between the nodes. Five similarity-based LP measures are chosen for experimentation. The experimentation on six benchmark datasets from four domains shows that a hop distance of maximum three is optimum for the prediction task.

Keywords: Link Prediction · Complex Network · Topological measures

1 Introduction

A network is a representation of any system that has entities and interactions between those entities. Networks, where nodes standing for entities and links for relationships between nodes, can be used to depict a variety of social, technological, ecological, and informational systems [1]. Complex networks are dynamic

due to nodes and links' constant addition and deletion. Link prediction (LP) is the task of predicting future links in a complex network [2]. When an interaction between two nodes does not already exist at this time, the LP problem seeks to estimate the likelihood that it will occur in the future. The problem of link prediction can be relevant in various disciplines.

To predict missing interactions between biological entities, unknown interactions in protein-protein interaction networks, unknown reactions in metabolic networks, expensive laboratory research is required. Link prediction measures help reducing experimental costs significantly in such applications [3]. Link prediction algorithms identify spurious links in computer network. LP methods are used to propose friends on social networks like Facebook and LinkedIn. Users on online websites such as Amazon might receive product recommendations by foreseeing links between users and products in a user-item bipartite graph that indicates purchases. Link prediction in coauthorship networks can suggest future collaborations [4]. Section 2 discusses the existing methods of Link Prediction. Implementation and results are given in section 3. Conclusions of this study are given in section 4.

1.1 Problem statement

Link prediction is defined as follows: Given a network $G(V, E)$ V and E denoting node-set and edge-set, the link prediction task is to generate a list of edges that are not existent at time $G(t_0)$ but are expected to form in the network $G(t_n)$ for $t_0 < t_n$ [2].

Fig.1 depicts the problem. In Fig.1, a new edge has been established between

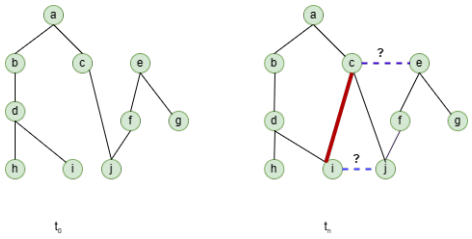


Fig. 1. Link Prediction

nodes c and i at time t_n. The following steps are typically included in link-prediction.

- The network data is divided into train and test sets.
- List out the node pairs without an edge from training set.
- For each pair of such node pairs, assign a prediction-score that determines how likely a link is probable in future.
- After sorting the node pairs in descending order according to the computed scores, the top k nodes will be delivered as the desired list.

– Then evaluate the performance using links in test set.

The following section discusses the existing measures that assign prediction scores for node pairs.

2 Link Prediction Measures

LP measures are mainly categorized into Similarity-based/Neighborhood-based, probabilistic, learning-based [5]. Fig.2 summarises these measures.

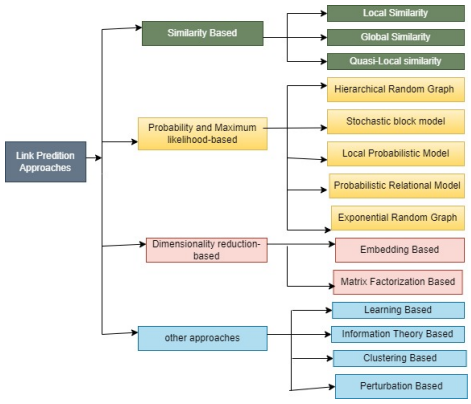


Fig. 2. Classification of LP measures

The following notation is used throughout this study.

– p, q: Two nodes in the network.
– $N(p)$: Neighborhood set of node p.
– n: Number of nodes in network.

Neighbourhood-based methods use simple approach in which the similarity scores are calculated for each pair of nodes p and q. Sort the scores; the pairs with the highest scores may eventually create links in future. These measures are called as local if the computation involves local neighborhood; global if path information is used in computing LP score [6]. Examples of global measures include Katz , Random-Walk-With-Restart [7], Shortest Path etc. Quasi-local measures use a combination of these two. Probabilistic models usually require the information other than graph topology, such as knowledge of the node/edge attributes [8,9]. It is difficult to gather such attribute information due to privacy issues. Dimensionality reduction based measures utilize matrix factorization and embedding techniques.

The main focus of this study is on local neighborhood based LP measures. The following section discusses the neighborhood-based LP measures.

2.1 Neighbourhood based measures

Common Neighbours and node degrees are typically used in the calculation of local indices. Five local neighborhood measures are described below.

1. **Common Neighbours(CN) [10]** : This measure works with the intuition that if two nodes have many neighbors in common, then the probability of link formation increases. $CN(p, q)$ is given in Eq.1.

$$CN(p, q) = |N(p) \cap N(q)| \tag{1}$$

   It is obvious that $CN(p, q) = A^2[p][q]$, where A is Adjacency-Matrix of graph G.
2. **Jaccard Coefficient (JC) [5]**: Jaccard Coefficient is the normalized Common-Neighbour measure. It is described as the fraction of common neighbors among existing neighbors of both the nodes. $JC(p, q)$ is defined as.

$$JC(p, q) = \frac{|N(p) \cap N(q)|}{|N(p) \cup N(q)|} \tag{2}$$

3. **Preferential Attachment (PA) [11]**: The node with the highest degree is expected to connect to other nodes in the future. By multiplying the degrees of nodes p and q, we may calculate the richness of two nodes. $PA(p, q)$ is defined as follows:

$$PA(p, q) = |N(p)| * |N(q)| \tag{3}$$

   Only the degree of the nodes is required for this measurement. As a result, the computational complexity of PA is the lowest.
4. **Adamic/Adar Index (AA) [12]** Adamic and Adar introduced a metric to determine the score of similarity between two web pages based on shared traits. Liben-Nowell et al. [12] modified this metric and used it to predict links between web sites.

$$AA(p, q) = \sum_{r \in N(p) \cap N(q)} \frac{1}{log(N(r))} \tag{4}$$

   where $N(r)$ is degree of a node r. The equation makes it obvious that common neighbours with lower degrees are given more weight. This makes sense in the real world as well; for instance, someone with more friends will spend less time and resources on each friend than someone with fewer friends.
5. **Resource Allocation Index (RA) [13]**: Consider two vertices p and q, which are not neighbouring. Assuming that node p sends some resources to node q through the two nodes' shared nodes, the similarity between the two vertices is evaluated in terms of resources provided from p to q. RA can be mathematically represented as

$$RA(p, q) = \sum_{r \in N(p) \cap N(q)} \frac{1}{N(r)} \tag{5}$$

There are plenty of other measures available in the literature such as Cosine-similarity [14], Sorensen Index [15], CAR-based Common Neighbor Index [16] ,CAR-based Adamic-Adar Index [16] , CAR-based Resource Allocation Index [16] , CAR-based Preferential Attachment Index [16],Hub Promoted Index and Hub Depressed Index , Local Naive Bayes-based Common Neighbors [17], Leicht-Holme-Newman Local Index [10] , Node Clustering Coefficient [18] , Node and Link Clustering coefficient which are variations of the above mentioned measures.

One of the challenges in the problem of link prediction in complex networks is the selection of the candidate node pairs. According to problem description, all those node pairs without an edge between them can be considered as candidate pairs for computing the LP score that can be used for future link formation. But the complex networks being huge in size and very sparse in nature, considering all node pairs induces large number of false positives. To address this problem, many existing works consider only a sample of node pairs to be candidates for prediction [19]. In this work we have experimented the effect of hop distance between candidate node pairs on the prediction accuracy. The following section presents the experimental setting used in this study.

3 Experimentation and Results

3.1 Data set Description

Six network data sets from various disciplines are used in the experimentation. CA-GrQc and ca-netscience [20] are collaborative networks with nodes representing authors and scientific collaborations denoting edges. Web-polblogs [21] is another dataset used, which is related to web graphs in which web pages are nodes and hyperlinks are edges. Bio-celegans [22] is the fourth dataset from the domain of Biology, where nodes denote substrates and metabolic reactions between them are edges. The last dataset used is E-road network [23] which is a road network located mostly in Europe. In E-road network, nodes represent cities and an edge denotes that they are connected by an E-road. And the last dataset we used is powerGrid [23], in this power stations and substations are represented as nodes and the power lines or transformers acts as links between the nodes. A few network statistics of these datasets are given in Table. 1.

Among all the networks, *bio-celegans* is dense network with a negative assortativity coefficient. CA-GrQc is sparse with high assortativity. Clustering coefficient is high for ca-netscience.

3.2 Evaluation metrics

A confusion matrix(Fig.3) can be used to illustrate the evaluation of prediction performance of link prediction measures. [24]

In the confusion-matrix,

Table 1. Dataset statistics

Network	$\|Nodes\|$	$\|Links\|$	**Average clustering-coefficient**	**Assortativity Coefficient**	**Density**
CA-GrQc	5242	14496	0.5296	0.6593	0.0010
ca-netscience	379	914	0.7412	-0.0816	0.0120
web-polblogs	643	2280	0.2320	-0.2178	0.0110
bio-celegans	453	2025	0.6464	-0.2258	0.0197
Euroroad	1174	1417	0.0167	0.1266	0.0020
powerGrid	4941	6594	0.0801	0.0034	0.0005

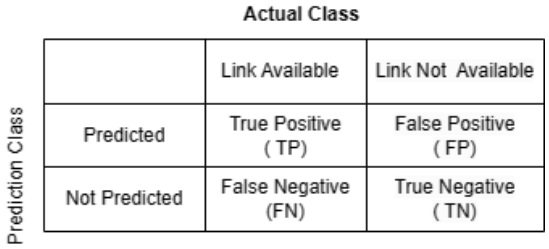


Fig. 3. Confusion Matrix

- True-Positive (TP): Number of node pairs with a predicted a link by the LP measure and the link is also existing in the test set.
- True-Negative (TN): Number of node pairs with a predicted a link by the LP measure and the link is not existing in the test set.
- False-Positive (FP): Number of node pairs between which the link is not predicted by LP measure and the link is actually existing in the test set.
- False-Negative (FN): Number of node pairs between which the link is not predicted by LP measure and the link is not existing in the test set.

The other metrics that are based on confusion matrix are as follows.

$$TPR = \frac{\#TP}{\#TP + \#FN} \quad ; \quad FPR = \frac{\#FP}{\#FP + \#TN} \quad ; \quad Precision = \frac{\#TP}{\#TP + \#FP} \tag{6}$$

There are threshold based metrics to evaluate the performance of LP measures. Area under the receiver operating characteristics curve (AUROC) and Area under the precision–recall curve (AUPR) are two such measures.

- AUROC [25]: The true positive rate (sensitivity) on the Y-axis and the false positive rate (1-specificity) on the X-axis are plotted to form a roc curve. Eqs. (6) and (7) can be used to calculate the true positive rate and false positive rate, respectively. Specificity is the performance of a dataset's entire negative part, and sensitivity is the performance of the entire positive part.

The area under the roc curve is a single-point summary statistic with a range between 0 and 1 [25].
- AUPR [24]: A binary classifier's performance is assessed using AUPR, a single-point summary statistic (predictor). Based on the precision-recall curve, which is a plot between the precision values on the Y-axis and the recall values on the X-axis, this number is calculated. Equations (9) and (6), respectively, can be used to calculate the precision and recall values. The more high the value of aupr, the better the model.

As the task of link prediction is highly imbalanced with huge number of negative (non-existing) links compared to positive (existing) links, AUPR is more appropriate measure [19].

3.3 Results

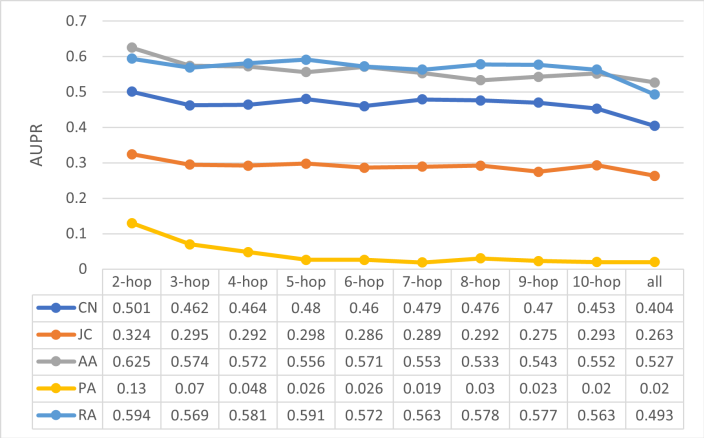
Five similarity-based LP measures discussed in section 2.1 are chosen for experimentation. The challenge of a huge number of false positives due to the selection of the candidate node pairs is specially focused.

1. Step 1: The network is divided into two parts: training-set containing 80% of links and test-set containing remaining 20% of links. Five sets of candidate node pairs from training-set for each network are formed as explained below.
   - *All*: This set contains all node pairs without an edge.
   - 2-*hop*: All node pairs within a distance of 2 -hops.
   - 3-*hop*: All node pairs within a distance of 3 -hops.
   - 4-*hop*: All node pairs within a distance of 4 -hops.
   - 5-*hop*: All node pairs within a distance of 5 -hops.
2. Step 2: Link prediction measures are applied on each of these sets individually.
3. Step3: The performance of LP measures on each network for each candidate node-pair set is evaluated using AUROC and AUPR.

The AUPR results are tabulated in Table 2. It is evident that considering candidate node-pairs within a distance of 2-hop gives better prediction accuracy, claiming the fact that the connection becomes weak as the distance between the nodes increases. To test this claim in depth, we have conducted experiments up to 10-hops. The results of AUPR up to 10-hops for the five LP measures for the network of CA-GrQC is given in Fig.4. The AUPR score is high at 2-hop, decreases slightly with 3-hop and then steady and least when all node-pairs are taken. The node-pairs within 2-hop distance are much less than all node pairs. Therefore, considering node-pairs within 2-hops not only improves prediction performance, but also reduces computation required.

Out of the five LP measures considered, Adamic-Adar predicted future links more efficiently and Preferential Attachment is the least performing for almost all networks. As AUROC being the classical evaluation measure, the best AUROC scores are presented in Table 3.

In terms of AUROC also, Common Neighbors produced accurate predictions and Preferential Attachments is the measure with least prediction performance.



	2-hop	3-hop	4-hop	5-hop	6-hop	7-hop	8-hop	9-hop	10-hop	all
CN	0.501	0.462	0.464	0.48	0.46	0.479	0.476	0.47	0.453	0.404
JC	0.324	0.295	0.292	0.298	0.286	0.289	0.292	0.275	0.293	0.263
AA	0.625	0.574	0.572	0.556	0.571	0.553	0.533	0.543	0.552	0.527
PA	0.13	0.07	0.048	0.026	0.026	0.019	0.03	0.023	0.02	0.02
RA	0.594	0.569	0.581	0.591	0.572	0.563	0.578	0.577	0.563	0.493

Fig. 4. AUPR scores of link prediction measures on CA-GrQC network.

Table 2. AUPR results of LP measures based on hop distance candidate between node pairs

LP measures	Candidate node pairs	CA-GrQc	ca-netscience	web-polblogs	bio-celegans	Euroroad	powerGrid
CN	**all**	0.404	0.451	0.054	0.094	0.001	0.016
	2-hop	0.501	0.459	0.085	0.081	0.064	0.096
	3-hop	0.462	0.386	0.058	0.112	0.003	0.047
	4-hop	0.464	0.442	0.070	0.081	0.022	0.042
	5-hop	0.480	0.407	0.054	0.079	0.004	0.037
JC	**all**	0.263	0.192	0.006	0.031	0.003	0.013
	2-hop	0.324	0.185	0.006	0.021	0.002	0.036
	3-hop	0.295	0.208	0.006	0.035	0.005	0.027
	4-hop	0.292	0.233	0.005	0.014	0.020	0.024
	5-hop	0.298	0.280	0.007	0.029	0.003	0.014
AA	**all**	0.527	0.442	0.046	0.121	0.003	0.011
	2-hop	0.625	0.591	0.074	0.104	0.003	0.029
	3-hop	0.574	0.517	0.058	0.104	0.005	0.032
	4-hop	0.572	0.520	0.055	0.126	0.005	0.026
	5-hop	0.556	0.563	0.066	0.138	0.003	0.028
PA	**all**	0.020	0.005	0.012	0.047	0.000	0.001
	2-hop	0.130	0.079	0.049	0.038	0.009	0.023
	3-hop	0.070	0.246	0.021	0.053	0.005	0.010
	4-hop	0.048	0.021	0.016	0.072	0.003	0.006
	5-hop	0.026	0.011	0.013	0.036	0.003	0.004
RA	**all**	0.493	0.521	0.039	0.177	0.001	0.014
	2-hop	0.544	0.541	0.057	0.187	0.005	0.022
	3-hop	0.581	0.587	0.076	0.143	0.003	0.024
	4-hop	0.569	0.564	0.057	0.136	0.005	0.016
	5-hop	0.591	0.539	0.067	0.130	0.005	0.025

Table 3. AUROC results of LP measures based on hop distance between candidate node pairs.

LP	CA-GrQc	ca-netscience	web-polblogs	bio-celegans	Euroroad	powerGrid
CN	0.978	0.944	0.868	0.903	0.550	0.738
JC	0.964	0.922	0.737	0.782	0.534	0.682
AA	0.971	0.942	0.830	0.919	0.515	0.694
PA	0.650	0.579	0.841	0.791	0.438	0.520
RA	0.966	0.930	0.846	0.943	0.528	0.677

4 Conclusion

Link prediction in complex networks is one of the significant analytical tasks in many domains. In this work five similarity based link prediction measures are evaluated on six networks from various domains. We have taken a sample of node-pairs from the training set as candidate node pairs for which prediction scores are computed. It is observed that node-pairs within 2-hop distance exhibited better prediction accuracy than considering all node-pairs. Limiting the candidate node-pairs based on hop distance not only improves prediction performance, but also significantly reduce the computation required.

References

1. Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.
2. Lin Yao, Luning Wang, Lv Pan, and Kai Yao. Link prediction based on common-neighbors for dynamic social network. *Procedia Computer Science*, 83:82–89, 2016.
3. Michael PH Stumpf, Thomas Thorne, Eric De Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.
4. J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.
5. Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289, 2020.
6. T Jaya Lakshmi and S Durga Bhavani. Link prediction measures in various types of information networks: a review. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1160–1167. IEEE, 2018.
7. Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *Sixth international conference on data mining (ICDM'06)*, pages 613–622. IEEE, 2006.
8. Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. Local probabilistic models for link prediction. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 322–331. IEEE, 2007.
9. Jennifer Neville. *Statistical models and analysis techniques for learning in relational data*. University of Massachusetts Amherst, 2006.

10. Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.
11. Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4):590–614, 2002.
12. L Adamic. & adar, e.(2003). friends and neighbors on the web. *Social Networks*, 25(3).
13. Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, 2009.
14. Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. mcgraw-hill, 1983.
15. Bruce McCune, James B Grace, and Dean L Urban. *Analysis of ecological communities*, volume 28. MjM software design Gleneden Beach, OR, 2002.
16. Carlo Vittorio Cannistraci, Gregorio Alanis-Lobato, and Timothy Ravasi. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific reports*, 3(1):1–14, 2013.
17. Zhen Liu, Qian-Ming Zhang, Linyuan Lü, and Tao Zhou. Link prediction in complex networks: A local naïve bayes model. *EPL (Europhysics Letters)*, 96(4):48007, 2011.
18. AJ Morales, JC Losada, and RM Benito. Users structure and behavior on an online social network during a political protest. *Physica A: Statistical Mechanics and its Applications*, 391(21):5244–5253, 2012.
19. Ryan Lichtnwalter and Nitesh V Chawla. Link prediction: fair and effective evaluation. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 376–383. IEEE, 2012.
20. Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.
21. Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.
22. J. Duch and A. Arenas. Community identification using extremal optimization phys. *Rev. E*, 72:027104, 2005.
23. Ryan Rossi and Nesreen Ahmed. The network data repository with interactive graph analytics and visualization. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.
24. Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 45:751–782, 2015.
25. Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.