# A Novel Strategy for Link Prediction in Social Networks

Naveen Gupta
Computer Science & Engineering Discipline,
PDPM Indian Institute of Information Technology,
Design and Manufacturing Jabalpur, India
naveen.gupta@iiitdmj.ac.in

Anurag Singh
Computer Science & Engineering Discipline,
PDPM Indian Institute of Information Technology,
Design and Manufacturing Jabalpur, India
anuragsg@iiitdmj.ac.in

## ABSTRACT

The problem of link prediction has gained a lot of attention recently from the research community. It can be formalized as, given a snapshot of a social network at time $t$, can it be predicted which new connections among its members are likely to occur in the future at time $t'$. Apart from analysing social networks, it has also found application in other domains e.g., information retrieval, bio-informatics and e-commerce. Topological information of the network, i.e. the information about the present nodes and links, can be used to predict the future links in the network. As an example, "Common neighbors" method is a trivial but efficient strategy for predicting the possibility of a link between a pair of nodes. Many variants of the common neighbors method have been proposed to address this problem. In this paper, we propose a novel strategy for predicting the missing links which also takes into account the number of links between two sets of uncommon neighbors of given nodes, in addition to their common neighbors.

## Categories and Subject Descriptors

H.2 [**Database Management**]: Database Applications—*Data mining*; G.2 [**Discrete Mathematics**]: Graph Theory—*Network problems*; J.4 [**Social and Behavioral Sciences**]: Sociology

## Keywords

Data mining; Link prediction; Complex networks; Information retrieval

## 1. INTRODUCTION

Social networks are a popular way to model the interactions or associations among the people in the real world. People can be represented as the nodes in the network and the links between these nodes represent the associations between the people. However, social networks are dynamic in nature. They keep evolving over time i.e. new associations keep developing between people over time. Great efforts have been made to understand the evolution of social networks [3]. To understand the evolution of the whole network, it is necessary to analyze the association between each pair of nodes present in the network. In this context, several questions arise: How does the network evolve over time? In predicting a future association between two nodes, what is the role of other nodes and links already present in the network? This article tries to address the problem of predicting the likelihood of a future association between two nodes, using the local topological information of the network. This problem is referred to as the link prediction problem in social networks [2, 8].

More formally, the link prediction method can be formulated as follows: Given a social network, $G(V, E)$, where $V$ is the set of nodes and $E$ is the set of the links. Let $U$ denote the universal set containing all the $\frac{|V|(|V|-1)}{2}$ possible links, where $|V|$ denotes the number of elements in set $V$. Then, the set of non-existent links is $U - E$. We assume that there are some missing links (or the links that will appear in the future) in the set $U - E$, and the task of link prediction is to find out these links [17]. Hence, the task is to predict the probability of a link between two nodes $x$ and $y$, $\forall x, y \in V$, where $x$ and $y$ are not connected at present.

Link prediction is done by utilizing the information about the nodes and the existing links in the network. It is used to study the evolution of social networks, e.g., in a network of individuals, it can be used to predict who might be friends with whom in the future. Apart from analyzing social networks, it can also be applied in other domains. As in bioinformatics, link prediction can be used to discover interactions among proteins, which otherwise needs costly laboratorial experiments. In the field of electronic commerce, it can be used to create the recommender systems; and in the security field, it can help to find the hidden terrorist and criminal gangs. Therefore, in recent years a lot of algorithms have been proposed to solve the problem of link prediction [1, 4, 5, 6, 7, 12, 13, 14, 15, 16, 19, 20, 22].

## 2. PROPOSED METHOD

Given a snapshot of a social network at time $t$, we seek to accurately predict the edges that will be added to the network during the interval $(t, t')$. We propose a method for predicting the likelihood of a link between two nodes, based on their local topological information i.e. the information about their neighbors and connections between them. The score for a link to be there in the future between two nodes $x$ and $y$ is calculated in two parts. The first part, simply counts the number of common neighbors of both the nodes. In the second part, the connections between the sets of uncommon neighbors of both the nodes are counted and


*CoNEXT Student Workshop'14,* December 2, 2014, Sydney, Australia.

http://dx.doi.org/10.1145/2680821.2680839.

divided by the total possible connections. This method is based on the assumption that more the number of common friends of two persons predict a strong possibility of them being friends as well. But on the other hand, the uncommon friends of the two persons also play a crucial role here. The more the connections or acquaintances between these uncommon friends, more is the probability that there will be a connection between $x$ and $y$ in the future. Hence the connections between the uncommon neighbors have also been considered in this work, in addition with the common neighbors of the nodes. It can be formulated mathematically as follows:

Let $N(x, y)$ be the set of total neighbors of the nodes $x$ and $y$, $C(x, y)$ the set of common neighbors and $UC(xy, x)$, $UC(xy, y)$ the set of uncommon neighbors of the nodes $x$ and $y$, respectively. For a node $x$, $\Gamma(x)$ represents the set of neighbors of $x$. $degree(x)$ is the size of the $\Gamma(x)$.

$$N(x, y) = (\Gamma(x) \cup \Gamma(y)) \quad (1)$$

$$C(x, y) = (\Gamma(x) \cap \Gamma(y)) \quad (2)$$

$$UC(xy, x) = (\Gamma(x)) - (\Gamma(x) \cap \Gamma(y)) \quad (3)$$

$$UC(xy, y) = (\Gamma(y)) - (\Gamma(x) \cap \Gamma(y)) \quad (4)$$

$$Score(x, y) = \frac{|C(x, y)|}{|N(x, y)|} + \frac{|e_{jk} : v_j \in UC(xy, x), v_k \in UC(xy, y), e_{jk} \in E|}{|UC(xy, x)| \; X \; |UC(xy, y)|}$$

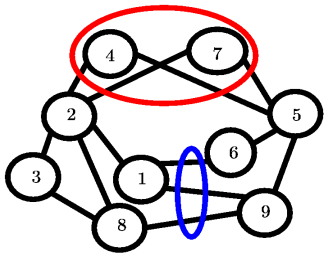


**Figure 1: Calculation of score for a link between a pair of nodes using the proposed method**

Fig. 1 shows an example of how to calculate the score for a link between a pair of nodes using the proposed method. There are 9 nodes in the given network. The procedure to calculate the score for happening a future link between nodes 2 and 5. Nodes 2 and 5 has got five and four neighbors respectively amongst which 2 neighbors are common to both the nodes. So, $\Gamma(2) = (1, 3, 4, 7, 8)$ and $\Gamma(5) = (4, 6, 7, 9)$.
$N(2, 5) = (1, 3, 4, 6, 7, 8, 9)$
$C(2, 5) = (4, 7)$
$UC(25, 2) = (1, 3, 8)$
$UC(25, 5) = (6, 9)$

Given nodes 2 and 5 have two neighbors in common, as depicted by the red circle. Amongst the set of uncommon neighbors of the nodes, there are 3 edges present, where there are total 6 edges possible. These 3 edges are encircled by the blue circle. Hence, the score for a link to be there between nodes 2 and 5 will be

$$Score(2, 5) = \frac{2}{7} + \frac{3}{3 \; X \; 2} = 0.285 + 0.5 = 0.785 \quad (5)$$

The scores for the links would range between 0 and 2, as each part in the above equation can have a minimum value of 0 and maximum of 1. Usually, we do not know that which links are missing or future links, otherwise we do not need to do prediction. Therefore, to test the accuracy of algorithm, the set of observed links $E$, is randomly divided into two parts: the training set, $E^T$, is treated as known edges, while the probe set, $E^P$, is used for testing and no information in this set is allowed to be used for prediction. Clearly, $E^T \cup E^P = E$ and $E^T \cap E^P = \phi$.

To quantify the accuracy of the prediction algorithm, the metric of *area under the receiver operating characteristic curve* (AUC) is used [11]. A detailed introduction of the metric is as follows.

**AUC:** Provided the score for all non-observed links, the AUC value can be interpreted as the probability that a randomly chosen missing link (i.e., a link in $E^P$) is given a higher score than a randomly chosen non-existent link (i.e., a link in $U - E$). We randomly pick a missing link and a non-existent link to compare their scores. If among $n$ independent comparisons, there are $n'$ times the missing link having a higher score and $n''$ times they have the same score, the AUC value is

$$AUC = \frac{n' + 0.5n''}{n} \quad (6)$$

If all the scores are generated from an independent and identical distribution, the AUC value should be about 0.5. Therefore, the degree to which the value exceeds 0.5 indicates how better the algorithm performs than pure chance.

## 3. RESULTS

The proposed algorithm has been run on four real world networks and the AUC values are calculated (Table 1). The simulation has been run for 10 times with independent random partitions for testing set (90%) and probe set (10%). Minimum, average and maximum values of the results are produced in the Table 1.

As it can be observed from the table 1, AUC values are

**Table 1: Results on real world networks**

| Dataset | AUC Values | | |
|---|---|---|---|
| | Min. | **Avg.** | Max. |
| Zachary karate club [21] | 0.603 | **0.675** | 0.711 |
| Dolphin social network [18] | 0.630 | **0.760** | 0.828 |
| American college football [9] | 0.837 | **0.909** | 0.940 |
| Jazz network [10] | 0.841 | **0.882** | 0.891 |

substantially higher than 0.5 for last three networks, which indicates that the algorithm has got a good accuracy in predicting the missing links using the local topological information.

## 4. CONCLUSION AND FUTURE WORK

The proposed algorithm uses only the local topological information and is giving a good accuracy as AUC values obtained are substantially higher than 0.5. For future work, we would apply the proposed algorithm for community detection in social networks. The score would be calculated for all the existing links in the network. The links with the least score would be considered as inter-community links whereas the nodes with large score for link between them would be considered as part of the same community.

# 5. REFERENCES


[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] M. Al Hasan and M. J. Zaki. A survey of link prediction in social networks. In *Social network data analytics*, pages 243–275. Springer, 2011.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[4] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds. An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science*, 2014.

[5] P. Chebotarev and E. Shamis. The matrix-forest theorem and measuring relations in small social groups. *arXiv preprint math/0602070*, 2006.

[6] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80. IEEE, 2011.

[7] M. Fire, L. Tenenboim-Chekina, R. Puzis, O. Lesser, L. Rokach, and Y. Elovici. Computationally efficient link prediction in a variety of social networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):10, 2013.

[8] L. Getoor and C. P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12, 2005.

[9] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[10] P. M. Gleiser and L. Danon. Community structure in jazz. *Advances in complex systems*, 6(04):565–573, 2003.

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

[12] P. Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[13] G. Jeh and J. 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.

[14] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[15] E. Leicht, P. Holme, and M. E. Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.

[16] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.

[17] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[18] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[19] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[20] G. Salton and M. J. McGill. Introduction to modern information retrieval. 1983.

[21] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.

[22] T. Zhou, L. Lü, and Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4):623–630, 2009.