# Proximity Measures For Link Prediction In Dynamic Network

Shubhangi R Urkude

***Abstract---*** *Now a day's growth of social network is gaining more attention from the user in different age group, profession, culture and geographical area in the world. This increasing use of social network attracted industries and academician to study the evolvement of people over time. The social network like Facebook, Twitter, Instagram and Flicker have more complex ties between the people and requires more efficient algorithm to recommend friends to their users in the network. Friend recommendation is one of the applications of link prediction. Link prediction is about forming future links in the social network. In this paper, we discussed various link prediction measures in context to the structural information and other high-level measures.*

***Keywords---*** *Link prediction, Social network, Friend Recommendation.*

## I INTRODUCTION

Social network (public network) is popularly used to represents the interaction between the people in the community. Any network is characterized as a graph consisting of peoples in the different geographic areas. Each person in the graph is represented by node and the association between more than two in represented as edge. Social networks are temporary / virtual in nature. These networks will grow and shrink with respect to time and can be visualize as a graph. Analyzing the social network is a complex problem because of dynamic nature of social network which is based on different parameters.

According to Liben-Nowell and Kleinberg [1], social network is represented as a graph G. This graph is a collection of vertices and connections among them, so $G = (V, E)$ V denotes set of vertices/users, E denotes edges/ connections between two nodes. Where $V = \{v1, v2, \dots \dots, vn\}$ and$E = \{e| < v_i, v_j >, v_i, v_j \varepsilon V$ are collection of vertices and edges respectively. The graph can be directed graph or undirected graph. We consider only undirected social network. The interaction between two neighbors in edge E is represented as $< v_i, v_j > \varepsilon E$ at particular time t [4,6]. We can calculate multiple interactions by considering particular time to form an edge. Let us take a time tand$t'$, where t is less than$t'$. Assume a graph $G[t, t']$ represents a subgraph in time period tand$t'$.Figure 1 shows the social network with some nodes connected to each other via edge. Each edge represents the relationship between the existing nodes. There can be direct relationship/ connection between two vertices or can be indirect relationship.

Link prediction is way of identifying the future connection among the existing objects in social network. In other word it is described as problem of how many new linksare possiblebetween any random nodes? The future links can

*Department of Computer Science and Engineering, Faculty of Science and Technology. The ICFAI Foundation for Higher Education (IFHE), Hyderabad – 501203, Telangana, India. Email: ushubhu@ifheindia.org*

**Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020**

be detected in two ways, (i) identifying unnoticed link in the present state of network (ii) to predict which link will occurs at period t+1 time with respect to given time period t, referred as time series problem [5].

To analyze the social network, we should find the links that may appear in the near future. Link prediction is useful in many domains like, e-commerce, medical, information retrieval, social recommendation, bioinformatics, in security domain, where criminals can be identified based on their interaction and so on.

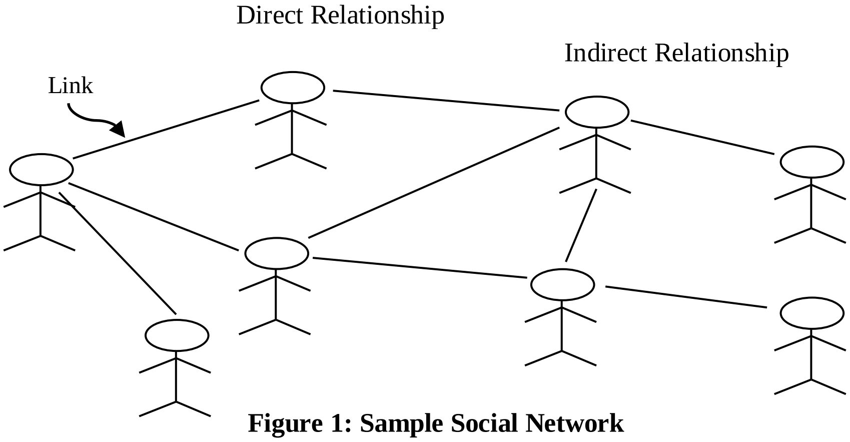

**Figure 1: Sample Social Network**

## II LITERATURE SURVEY

Liyan Dong [10] et al. proposed a new link prediction approach using global information, which in turn will consider topological information about the network. They designed a new algorithm based on the assumption that, there are more paths between two nodes;there is greater chance of forming new links. In the existing Katz algorithm path length is not considered for predicting the new links. Linyuan Lü et. al. [11] proposed, predicting the missing links in the incomplete network using common neighbors and Katz index. In this study path length is playing a major role in reducing CPU time and memory space. Carter Chiu et.al. [20] Proposed, probability-based approach for dynamic link prediction. According to Chiu, the weak estimator is working with more accuracy as compared with the traditional similarity measures. Yao et al. [21] proposed a modified version of common neighbors to find the new links between the nodes separated by two hops. He proposed the idea of giving more weights to the latest subgraph considering the time decomposition factor. Popescul et. at [22,23] proposed future link prediction in the internet domain using the relational data. There exists a complex relationship between the datasets and system is designed to predict the availability of the relationship and category of the new relationship. There are many social networks exists is today’s world such as Facebook, Twitter, Instagram and so on. One of the biggest challenges in this type of network is their dynamic nature, which will change with respect to time and huge number of interactions among the users may be added or removed from the network. These issues cannot be handled by earlier algorithms. Jingwei Wang et. al. [24] used vertex similarity index to improve link prediction accuracy. They proposed new technique of link prediction by combining topological information and community information. According to Jingwei, joining both the features accuracy can be improved irrespective of the type of algorithm used to detect community. Time varying probabilistic model for social network was designed by using Markov prediction model considering time scale, local information and structural information of nodes. Barabasi et. al. proposed [15] a simple model to grab

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

network evolutions with respect to time and identifies role of internal links. They studied the structural properties of the network which is changing time to time, with this effect of link evolution on internal and external links.

## III THE LINK PREDICTION PROBLEM

Forecasting the future interaction among the available vertices in the network is the primary function of link prediction. Many techniques are available to perform link prediction such as using proximity measures, supervised, unsupervised machine learning algorithms and so on. In supervised machine learning techniques network data is divided into training and testing data. Training data is used to train the model and testing data is used to predict the future links with the help of present links in the network.In unsupervised technique different patterns are identified and based on the pattern future links are predicted in the network. The link prediction can also be done periodically and non-periodically [8,9]. Periodic link prediction is predicting the links at specific time intervals and non-periodic link prediction is done at any random time period. Figure 2 shows a spanshot of social network at different time and can be represented as subgraph. Given a series of subgraphs $\{G_1, G_2, \dots\dots. G_t\}$ of a graph $G_t = (V, E_t)$ in that each edge, $e = (v_i, v_j)\ \varepsilon\ E_t$ shows a link among $v_i$ and $v_j$that occur at specific time t [7,12,13].

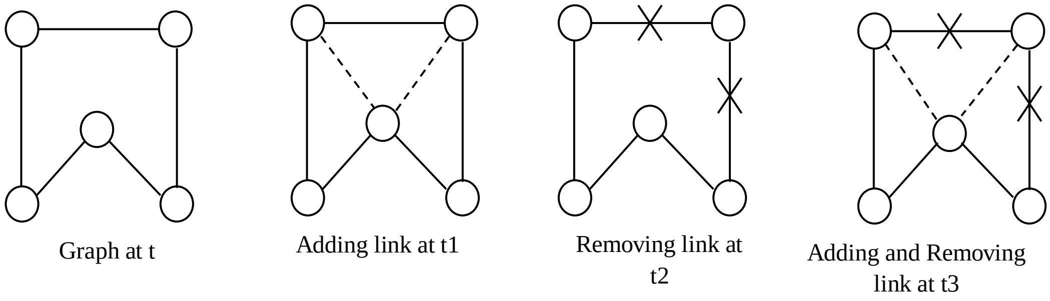

**Figure2: Social network at different time (a) at time t (b) link added at time t1 (c) link removed at time t2 (d) link is added and another link is removed at time t3**

The non-periodic technique is predicting the links at any random time in the existing network [4]. The task of predicting the unknown links are divided into four categories [4,5]: (i) Link existence (ii) Link weight (iii) Link cardinality (iv) Link type. Figure 3 shows different types of link prediction.

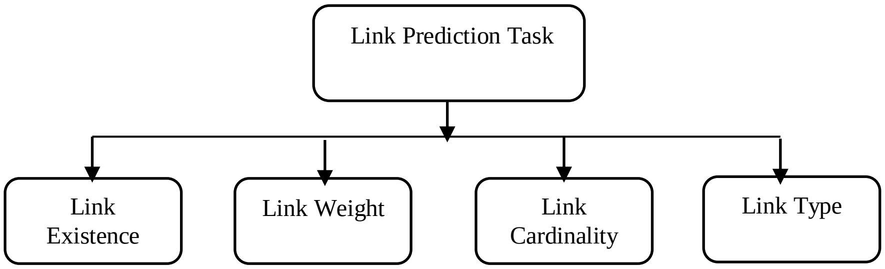

**Figure 3: Link prediction types**

## IV METHODS OF LINK PREDICTION

Many methods of link prediction are discovered by different researchers. This section is going to discuss some of them. Huge amount of information is generated by the people while communicating to each other on social media.

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

This information is used by many researchers and proposed different methods that will enhance the efficiency of new link appear in the graph at some time period. These methods are based on different approaches such as (i) Based on local similarity (ii) based on global similarity [2].Some of these techniques are discussed below:

**Based on local similarity**

In this paper we are reviewing some well-known methods to predict future links using topological information present in the networks. To represent the structural information, the entire section uses following naming conventions: x, y represents the vertices, N gives total number of vertices present in the network, k gives average degree of vertex,$\Gamma(x)$and $\Gamma(x)$represents neighboring nodes of x and y, $k_x$&$k_y$ gives average degree of vertex x and y respectively.

**Common Neighbor (CN):**

Common neighbor is a similarity measure based on node neighborhood. This technique will consider the common neighbors of two vertices and if both are having same neighbor there is greater chance of having relationship in future. Common neighbor for two vertices, x and y is shown in equation (1) and it is defined as, consider vertex x is having interaction with vertex z and vertex y also having interaction with vertex z, then there is a greater probability of establishing an interaction among vertex x & vertex y. In the collaboration network, common neighbor coefficient is used to give association among neighboring vertex x & y at specific time tin future [14]. The complexity of this method is estimated as$O(n^{k^2}$).

$$CommonNeighbor(x,y) = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

**Adamic-Adar coefficient (AA) [3]:**

Adamic Adar proposed this technique as metrics to find closeness among the web pages. This coefficient value is greater, more the similarity between the web pages. It gives likelihood of strongly connected pages.It finds the similarity between two vertices based on common feature between the vertices. For Adamic Adar common features are calculated and similarity is defined as below in equation (2).

$$AdamicAdar(x,y) = \sum_{z\epsilon\Gamma(x)\cap\Gamma(y)} \frac{1}{log|\Gamma(z)|} \tag{2}$$

**Jaccard coefficient:**

Paul Jaccard introduced Jaccard coefficient. Jaccard coefficient is statistical similarity measure. Jaccard coefficient is used to measure the similarity or non-similarity between the binary variable for finite sample set. It gives the probability of common neighbor for a two-vertex x & y selected randomly from the union of neighbor of x and y. It is ratio of intersection of sample set to union of sample set. It always lies between 0 and 1. It is defined as below in equation (3).

$$Jaccardcoefficient\,(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{3}$$

**Based on global similarity:**

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

**Katz Index:**

Leo katz proposed this index [16]. Katz index is similar to shortest path and also performs well for predicting links in future. It is one of node centrality measure. In the graph theory, relative degree of influence of vertex gives the katz index. It is calculated as total number of paths among two vertices. It is used in directed network to compute centrality. Katz index is defined as below in equation (4).

$$Katzindex(x,y) = \sum_{l=1}^{\infty} \beta^{l} * \left|paths_{xy}^{(l)}\right| \quad (4)$$

Where, $paths_{xy}^{(l)}$ is defined as combination of all possible paths having length l and joining two vertex x & y and β is an unrestricted constraint monitoring the path weights.

**Simrank:**

Simrank is a similarity measure for nodes. As per this measure is concern, it says there is similarity between two vertices when they are referenced by the similar vertex. It is widely used in graph analytics. Similarity value is 1 for the same vertex and similarity for two different vertices is stated in equation (5).

$$similarity(x,y) = \gamma * \frac{\sum_{a\epsilon\Gamma(x)} \sum_{b\epsilon\Gamma(y)} similarity(a,b)}{|\Gamma(x)||\Gamma(y)|} \quad (5)$$

Decay factor $\gamma\epsilon[0,1]$ for some networks. For uncertain graphs Simrank is interpreted as random walk [11, 17].

**Hitting Time:**

Hitting time come from random walk on graph. Hitting time is defined as on a random walk, how much time taken by vertex x to travel to vertex y? Shorter distance indicates that, the both vertices are highly similar to each other and probability of becoming friends is more in the coming future. Hitting time has many useful properties [18] such as cover time and expected time. It is easy to compute and can be calculated by some random walk on graph. For example, consider a vertex u placed at more distance from vertex x and vertex y and affecting the hitting time for both vertices. There exists a problem to find the hitting time among vertex x & vertex y, in case vertex u is a steady vertex with higher probability value. This problem can be solved by restarting the journey again and returning back to same vertex x by considering the fixed probability value α at every step. Because of the scale free nature of social network, some of the vertex will have high probability value in steady case on random walk. By multiplying the steady probability value of respective vertices hitting time can be normalized, to overcome this issue. Hitting time is defined as below in equation (6).

$$normalizedhittingtime(x,y) = H_{x,y} \,.\, \pi_{y} + H_{y,x} \,.\, \pi_{y} \quad (6)$$

**Rooted PageRank:**

According to Chung et. al. [19] Pagerank algorithm is used to rank web pages along with hitting time. Pagerank is one of the centrality measures used is link analysis. The Pagerank algorithm gives probability distribution for a user to arrive at particular page after clicking some random pages in web. Pagerank can be applied to collection of documents of any size. It is based on two assumptions: 1) Consider some fixed probability value w, a user in a network jump from one web-page to another web-page with probability w and having 1-w as hyperlink value. 2) Significance of vertex v is defined as sum of significance of total web pages u that link to v, on random walk. For link analysis purpose the random walk assumption of innovative Pagerank algorithm is modified as: proximity value of both vertices x & vertex y can be calculated as steady probability of vertex y restarting from the same vertex with probability 1 – β every time and passing through some arbitrary neighbor having probability value as β. This will give asymmetric matrix for vertex x & vertex y. By performing the reverse operation on role of vertex x & vertex y it can be converted to symmetric matrix value. This is known as rooted Pagerank [19]. The rooted Pagerank is derived as, Let W is diagonal matrix having $W[i,j] = \sum_j A[i,j]$ among the pair of vertices and it is represented as RPR. Let, $\mathrm{N} = \mathrm{W} - 1\mathrm{A}$ is an adjacency matrix having row normalized value as 1. Then, Rooted Pagarank is defined as in equation (7).

$$RPR = (1-\beta)(I-\beta N)^{-1} \tag{7}$$

**Higher-level approaches:**

**Low-rank approximation:**

In the collaboration network$G_{\mathrm{collab}}$ many adjacency matrices M are used to represent the link prediction techniques. It was observed that, the basic neighbor's technique comprises of integrating every vertex x to its row $r(x)$in adjacency matrix M. Then the $score(x,y)$is calculated as row multiplication operation on rows $r(x)$ and $r(y)$.

A very common technique while discussing the enormous structure of matrix $M$ is to take any moderately modest number $k$ and find rank-k matrix of M-K. It is best way of approximating M for any number of standard matrices. This should be possible by effective utilization of core methods in data recovery for semantic investigation. Instinctively, working with $M_K$ rather than $M$ is type of noise-reduction, that gives matrix structures in highly simplified manner. This paper uses low rank estimation is used in ranking (i) Katz score uses$M_k$ instead of M. (ii) common neighbor's use row multiplication operation as $M_k$ and (iii) in defining $score(x,y)$ in the form of matrix $M_k$.

**Unseen bigrams:**

Unseen bigram is evaluating the frequencies in the language modeling. Both the problems like link prediction and unseen bigram are look similar to each other. Unseen bigram is applied to the pair of words in the testing dataset and it is different from training dataset. According to the literature survey, it is suggested that you can enlarge score for vertex x & vertex y using the score among the vertex z & vertex y, where vertex z is a neighboring vertex of x & y. Consider the $score(x,y)$ computed using one of the technique above. Let Sx denote the δ vertex are related to x for the value $score(x,y)$ for $\delta \in Z\ +$. The weighted and unweighted score can be defined as follows.

$$score^*_{unweighted}(x,y) := \left|\left\{z : z\epsilon\Gamma(y) \cap S_x^{\langle\delta\rangle}\right\}\right| \tag{8}$$

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

$$score^{*}_{weighted} := \sum_{z\epsilon\Gamma(y)\cap S_x^{(\delta)}} score(x,z) \qquad (9)$$

### Clustering:

Clustering is a one of the data mining technique and it is also used in the graph theory. It is used to identify the similar nodes in the social network. Clustering is one of the measures to find the likelihood on the node. It will give the association between two nodes. In graph theory clustering coefficient is defined as degree of likelihood of similar nodes in the network. The$score(x,y)$for all pair of $< x, y >$in this subgraph determines node proximities using only edge. Clustering coefficient of vertex v is defined as below in equation (10).

$$clusteringcoef.(v) = \frac{3*\# trianglesadjacenttou}{\# possibletriplesadjacenttou} \qquad (10)$$

## V CONCLUSION AND FUTURE WORK

Link prediction algorithms based on local structure compute the score depending upon the common neighbors that may occur in future interactions among pair of vertices. These techniques will work only for interaction among the two vertices not more than that. Due to that some of the interesting link may be missed and it is time consuming process to compute common neighbors for every pair of nodes, because of huge size of network. To overcome this issue other parallel algorithms are explored.

Global structure-based algorithms are depending upon the global structure of network and work for a greater number of nodes. With the help of this global information some useful links may be found out, but finding the path length for greater number of vertices in large network is very tedious task. Every social network where terabytes of data is generated every fraction and it has to analyze for prediction.

Other high-level approaches discussed in the paper consider the entire network to find the similarity score. But all these methods are not explored further. In future all these methods can be applied on somewell-known datasets and tested for link prediction. This paper assists a starting journey for researchers and understands the basics ofanalyzingthe social network.

## REFERENCES

[1] D. Liben-Nowell and J. Kleinberg.: The link prediction problem for social networks. Proceedings of the 12th ACM International Conference on Information and knowledge Management (CIKM ’03)ACM, New York, NY, USA.pp. 556--559 (2003).

[2] Yingying Liang, Lan Huang, and Zhe Wang.: Link prediction in social network based on local information and attributes of nodes. IOP Conference Series: Journal of Physics. vol. 887(1). (2017).

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

[4] Sogol Haghani and Mohammad Reza Keyvanpour.: A systemic analysis of link prediction in social network. Springer Science+Business Media B.V. Springer Nature (2017).

[5] Mohammad Hasan.: Survey of Link Prediction in Social Networks.Social network data analytics. pp 243--275(2011).

[6] Ajay Kumar Singh Kushwah and Amit Kumar Manjhvar.: A Review on Link Prediction in Social Network. International Journal of Grid and Distributed Computing. vol. 9(2). pp.43--0(2016).

[7] Brandes U and Wagner D.: Analysis and visualization of social networks. In: Jünger M, Mutzel P (eds) Graph drawing software. Mathematics and visualization. Springer Berlin. pp 321—340(2004).

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

[8] W. Cukierski, B. Hamner, and B. Yang.: Graph-based features for supervised link prediction. Proceedings of the International Joint Conference on Neural Network (IJCNN '11). pp. 1237—1244(2011).

[9] M. E. J. Newman.: Clustering and preferential attachment in growing networks. Physical Review Letters E. vol 64(025102). (2001).

[10] Liyan Dong, Yongli Li, Han Yin, Huang Le, and Mao Rui.: The Algorithm of Link Prediction on Social Network. Mathematical Problems in Engineering. vol 1. pp. (2013).

[11] Linyuan Lu, Ci-Hang Jin and Tao Zhou.: Similarity index based on local paths for link prediction of complex networks. PHYSICAL REVIEW E, vol. 80, pp (2009).

[12] Miller K, Jordan MI and Griffiths TL.: Nonparametric latent feature models for link prediction. Advances in Neural Information processing systems. pp 1276--1284(2009).

[13] Tylenda T, Angelova R and Bedathur S.: Towards time-aware link prediction in evolving social networks. Proceedings of the 3rd workshop on social network mining and analysisACM. pp 9(2009).

[14] D. Sharma, U. Sharma, and Sunil Kumar Khatri.: An Experimental Comparison of the Link Prediction Techniques in Social Networks. vol. 4(1). pp (2014).

[15] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek.: Evolution of the social Network of scientific collaboration. Physica A. vol. 311(3-4). pp. 590--614(2002).

[16] L Katz.: A new status index derived from sociometric analysis. Psychmetrika 18. pp. 39--43(1953).

[17] Rong Zhu, Zhaonian Zou and Jianzhong Li.: SimRank Computation on Uncertain Graphs. Arxiv. (2015).

[18] Shravas K Rao.: Finding hitting times in various graphs. Arxiv. pp. 1--9 (2012).

[19] Chung, Fan, and Zhao and Wenbo. PageRank and random walks on graphs. Proceedings of the "Fete of Combinatorics" conference in honor of Lovasz. pp. 43--62(2010).

[20] Carter Chiu and Justin Zhan.: Deep Learning for Link Prediction in Dynamic Networks Using Weak Estimators. IEEE Access.(2018).

[21] L. Yao, L. Wang, L. Pan, and K. Yao.: Link prediction based on common-neighbors for dynamic social network. Procedia Computer Science, the 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops. vol. 83. pp. 82--89(2016).

[22] Popescul, Alexandrin and Ungar, Lyle H.: Statistical Relational Learning for Link Prediction. In Proceedings of Workshop on Learning Statistical Models from Relational Data at IJCAI Conference (2003).

[23] S. Velliangiri, P. Karthikeyan, I. T. Joseph and S. A. P. Kumar, "Investigation of Deep Learning Schemes in Medical Application," *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, United Arab Emirates, 2019, pp. 87-92.

[24] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar Joseph, "A Review of Dimensionality Reduction Techniques for Efficient Computation," *Procedia Comput. Sci.*, vol. 165, pp. 104–111, 2019.

**Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020**