Paper

Similarity Index based Link Prediction Algorithms in Social Networks: A Survey

Pulipati Srilatha and Ramakrishnan Manjula

School of Computer Science & Engineering (SCOPE), VIT University, Vellore, Tamil Nadu, India

Abstract—Social networking sites have gained much popularity in the recent years. With millions of people connected virtually generate loads of data to be analyzed to infer meaningful associations among links. Link prediction algorithm is one such problem, wherein existing nodes, links and their attributes are analyzed to predict the possibility of potential links, which are likely to happen over a period of time. In this survey, the local structure based link prediction algorithms existing in literature with their features and also the possibility of future research directions is reported and discussed. This survey serves as a starting point for beginners interested in understanding link prediction or similarity index algorithms in general and local structure based link prediction algorithms in particular.

Keywords—link prediction, similarity based link prediction, similarity index, social network analysis.

1. Introduction

Social network analysis [1] has gained lot of importance with millions of users being part of one or other online social networking sites. Link prediction problem in social networks has gained considerable interest from researchers of various domains [2]–[6].

Everyone in this world is connected to each other through an average small length was claimed by Stanley Milgram in 1967 itself in his ”small world” experiments [7]. But predicting the link or possibility of potential link among each other based on social networks is also an interesting task as social networks is also a kind of “small world” [8] where in nodes are connected to each other either through a direct or indirect link through intermediate nodes.

Link prediction as a tool helps in understanding potential links, which may happen among nodes in a network over a period of time. Based on individual preference, either through attributes or link structure, recommending friends (or likeminded people), passing information to individuals having similar interests, recommending products [9]–[13] seems to be an interesting and challenging task as the online social network size is growing with time and some of the conventional algorithms find it difficult to analyze such huge data volume.

In general link prediction refers to possible links among a pair of nodes in a social network [14] but it can also be thought as a possible link prediction among a node and commodity, i.e. how likely that a person is interested in having link with the commodity or in simple purchase the commodity [15]. Recommender systems [16], [17] which recommend products or commodities to individuals can also be thought as a special case of link prediction problem. Apart from identifying potential links and recommending commodities, link prediction problem also help to solve similar problems like spurious link detection [18], mitigating e-mail spams in social networks [19] defending against Sybil attacks [20].

In literature many methods have been proposed for link prediction problem. Broadly they can be classified into similarity based methods, maximum likelihood based methods and probabilistic model based methods [21], [22]. Similarity based methods predict links among nodes based on similarity score or similarity index calculated among a pair of vertices or nodes. Pair of nodes having high similarity value will tend to form a link in future. Maximum likelihood methods predict links based on hierarchical structure model or stochastic models [23], [24]. Probabilistic methods of link prediction are defined over Markov chains or Bayesian learning models in turn using principles of artificial intelligence in broad sense [25]–[30].

In this survey, we are concentrating only on the similarity based methods used in link prediction. The methods exploit the nature or structure of graphs how the nodes are connected and attributes associated with the nodes for link prediction. Similarity based methods can be further classified into three types based on the nature of similarity score or index calculation, as follows:

- local structure based,
- global structure based,
- quasi-local structure based.

Investigations have been carried out to understand the link prediction based on the nature of link over a period of time or evolutionary history of links and clustering based link prediction techniques. All the techniques or methods discussed in this survey are proposed over unweighted edges for the results on link prediction over weighted edges [31]–[33].

This paper is organized as follows. Section 2 introduces the preliminaries of graph theory. Section 3 deals with different types of similarity based prediction techniques and how they are defined. Section 4 deals with the discussion on different similarity based prediction techniques and their summary. Section 5 contains conclusions.

2. Preliminaries

2.1. Graph

A graph G is defined as G = (V, E), where V is a set of vertices or nodes and E is a set of edges. The notion of vertices and nodes, links and edges are used interchangeably in the literature.

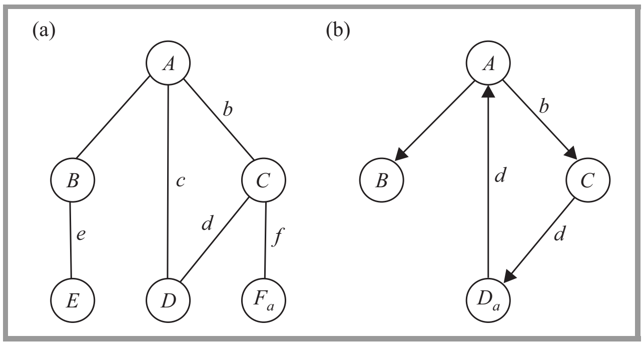

Fig. 1. Graphs: (a) undirected, (b) directed.

Graph shown in Fig. 1a is a simple undirected graph with set {A, B, C, D, E} as vertices and set {a, b, c, d, e, f} as edges. An edge from a node A to node B is called a path and it is of length 1 (or simply path length is 1) where as the path from node A to node D via node C is of length 2. Closed path from node A to node A (via node C and D) is called a circuit or cycle. The girth of a graph is number of edges in shortest cycle. Graph shown in Fig. 1b represents a graph with directed edge, in such graph, traversal is only possible in the direction of edge. Traversal from node A to node D via node C is possible where as traversal is not possible from node B to node C. In an undirected graph, degree d of a node is the number of edges or links connected to that node where as in case of directed graph, outdegree and indegree respectively indicate number of links or edges directed out and directed into a node. In case of Fig. 1a, degree of node A is 3 and of node e is 1. Similarly, for Fig. 1b, indegree for node A is 1 and outdegree is 2.

2.2. Adjacency Matrix

Adjacency matrix is used to understand the existence or non-existence of links or edges between nodes in a graph. Apart from that, other advantages includes that it can be given computed easily, can be given as an input to computer and also to understand the nature of path length by taking powers of adjacency matrix.

Adjacency matrix A_{ij} is defined as:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from node } i \text{ to node } j, \\ 0 & \text{otherwise.} \end{cases}$$

where i and j are nodes in the graph. For the example considered in graph shown in Fig. 1a adjacency matrix is given by

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

where the matrix entry 1 indicates the existence of link or edge between nodes in the graph and a entry 0 indicates no link or edge.

Path length more than 1 can be computed from adjacency matrix by taking its powers. Adjacency matrices A^2, A^3 gives the nodes connected by path length 2 and 3 respectively. But the entries in table will remain either 0 or 1 for any powers of A except the elements of principal diagonal. However, principal diagonal elements have no role to play with regard to link prediction as these elements indicate links or edges to same nodes.

Similarly, matrix A^2 indicates nodes connected with a path length 2. In matrix A^2, nodes A and E, A and F, B and C etc., are connected with path length 2:

$$A^2 = \begin{bmatrix} 3 & 0 & 1 & 1 & 1 & 1 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 1 & 1 & 3 & 1 & 0 & 0 \\ 1 & 1 & 1 & 2 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

2.3. Link Prediction Problem

Link prediction problem is identifying a potential or possible link among vertices or nodes in a network. The link prediction problem can be defined as, given node E, what are the chances that node D and node F would like to connect with node E (Fig. 1). Predicting such links is not an easy task. Many similarity index based link prediction methods have been defined based on nature of computing them, as local structure, global structure and quasi local structure [21], [22].

Path length equal to 1 indicates the existence of direct link between nodes in a network. Local structure based method computes the similarity score using a node and its neighborhood nodes (of path length 2). The aim is to determine the chances of possible links between a pair of vertices where as global structure based method computes similarity score considering over all structure of network (mainly paths of length > 2) by ranking each nodes by similarity score and nodes with highest value of similarity score tends to have more links than the ones having lower value. Quasi-local structure based similarity measure is even though computed similar to the local structure based link prediction method, it uses path of length more than 2. Hence, covering the entire network like global structure based methods.

3. Methods in Similarity Index Based Link Prediction

Link prediction between nodes in a network is given by a score $S(a,b)$ where, a and b are nodes. Higher the value of S between the pair of vertices, higher is the chance of forming a link between them in future.

3.1. Local Structure Based Similarity Index Methods

3.1.1. Common Neighbor

Common Neighbor [34] method is one of the simplest techniques used for link prediction. Two nodes are likely to form a link if they have many common neighbors [35]. This method of link predicition is also called friend-of-a-friend link prediction [36]. Index $S_{CN}(a,b)$ for Common Neighbor method is computed as:

$$S_{CN}(a,b) = |\Gamma(a) \cap \Gamma(b)|, \qquad (1)$$

where a and b are two nodes and $\Gamma(a)$ and $\Gamma(b)$ denote the set of neighbors of nodes a and b respectively. Common Neighbor method is being used extensively in various context to predict future links or collaborations [14], [37], [38].

3.1.2. Salton Index

Salton index [39] also called as Salton Cosine index is used to find the similarity index based on cosine angle between rows of adjacency matrix having nodes a and b [40]. Salton index $S_{Salton}(a,b)$ is computed as follows:

$$S_{Salton}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{\sqrt{k_a \times k_b}}, \qquad (2)$$

where k_a is the degree of node a and k_b is degree of node b.

3.1.3. Jaccard Index

Jaccard [41] in 1901 proposed a statistic to compare similarity and diversity of sample sets. It is the ratio of common neighbors of nodes a and b to the all neighbors nodes of a and b. As a result value of Jaccard index prevents higher degree nodes to have high similarity index with other nodes [42]. Jaccard index $S_{Jaccard}(a,b)$ is computed as:

$$S_{Jaccard}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{|\Gamma(a) \cup \Gamma(b)|}. \qquad (3)$$

3.1.4. Sørenson Index

A measure similar to Jaccard [41] proposed by Sørenson [43] in 1948 to measure similarities among species. Sørenson index is calculated as ratio of twice the common neighbors of nodes a and b to the sum of degrees of nodes a and b or total degree of nodes a and b. Sørenson index $S_{Sørenson}(a,b)$ is computed as:

$$S_{Sørenson}(a,b) = \frac{2|\Gamma(a) \cap \Gamma(b)|}{k_a + k_b}. \qquad (4)$$

3.1.5. Hub Promoted Index

Hub Promoted index [44] is a measure defined as the ratio of common neighbors of nodes a and b to the minimum of degrees of nodes a and b. Hub Promoted index $S_{HPI}(a,b)$ is computed as:

$$S_{HPI}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{\min\{k_a, k_b\}}. \qquad (5)$$

3.1.6. Hub Depressed Index

Hub Depressed index [21], [22] defined as the ratio of common neighbors of nodes a and b to the maximum of degrees of nodes either a or b. Hub Depressed index gives lower score compared to Hub Promoted index as the similarity measured is computed by taking maximum of degrees of nodes a and b [45]. Hub Depression index $S_{HDI}(a,b)$ is computed as:

$$S_{HDI}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{\max\{k_a, k_b\}}. \qquad (6)$$

3.1.7. Leicht Holme Newman Index-1

Leicht *et al.* in [46] proposed a measure to define local structure based similarity measure. It is the ratio of common neighbors of nodes a and b to the product of degrees of nodes a and b. Leicht Holme Newman index-1 is computed as:

$$S_{LNH1}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{k_a \times k_b}. \qquad (7)$$

Equations (7) and (2) differ only in their denominator. As a result, for a same pair of nodes Salton index always assigns a higher score compared to Leicht Holme Newman index-1 [45].

3.1.8. Preferential Attachment Index

Preferential Attachment index [47] is a similarity score calculated independent of the neighborhood of each node. Social networks expand as and when new nodes joins in and the new nodes joins in with the existing nodes having higher degree compared to lower degree nodes [48]. Barbasi *et al.* showed this result for the evolution co-authorship links [49]. Preferential Attachment index $S_{PA}(a,b)$ between nodes a and b is computed as:

$$S_{PA}(a,b) = k(a) \times k(b). \qquad (8)$$

3.1.9. Adamic-Adar Index

Adamic-Adar index [50] proposed by Adamic and Adar is calculated by adding weights to the nodes which are connected to both nodes a and b. Adamic and showed how similar nodes are linked analyzing text, out-links, in-links and mailing lists from homepages of students of Massachusetts

Institute of Technology (MIT) and Stanford University [50]. The index is computed as:

$$S_{AAI}(a,b) = \sum_{z \in \Gamma(a) \cap \Gamma(b)} \frac{1}{\log k_z} \,. \tag{9}$$

where z is a common neighbor to nodes both a and b and k is the degree of node z.

3.1.10. Resource Allocation Index

Resource Allocation index [22] is calculated based on the intermittent nodes connecting node a and b. The similarity index S_{RAI} is defined as the amount of resource node a receives from node b through indirect links and each intermediate link contributes an unit of resource. Also, $S_{RAI}(a,b)$ is symmetric to $S_{RAI}(b,a)$:

$$S_{RAI}(a,b) = \sum_{z \in \Gamma(a) \cap \Gamma(b)} \frac{1}{k_z} \,. \tag{10}$$

Equations (9) and (10) differs only in their denominator as the former takes the logarithm of the denominator. As a result, for a same pair of nodes Adamic-Adar index always assigns a higher score compared to Resource Allocation Index.

3.2. Global Structure Based Similarity Index Methods

3.2.1. Katz Index

Katz index [51] is calculated considering all paths instead of individual paths among node pairs. Like all global structure based similarity indices Katz index considers paths of length ≥ 2.
Katz score between two nodes a and b is calculated as a sum of products of path length (of all length) and β a free parameter used to control path weights. It is computed as:

$$S_{Katz}(a,b) = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{ab}^{<l>}| =$$
$$= \beta A_{ab} + \beta^2 (A^2)_{ab} + \beta^3 (A^3)_{ab} + \ldots \tag{11}$$

where A is the adjacency matrix of the graph, $|paths_{ab}^{<l>}|$ is the set of all paths with length l connecting nodes a and b. When β is very small Katz index will be similar to Common Neighbor method of link prediction defined in Subsection 3.1.1. Scalability of Katz index is discussed in [52] implementing it on Hadoop parallel Map/Reduce tool [53], [54].

3.2.2. Average Commute Time

Average Commute Time (ACT) [14] is defined as the average number of steps required for a random walker to reach node b starting from node a. For computation simplicity pseudo-inverse of Laplace matrix is used for representing the commute time between node a and node b. It is computed as:

$$S_{ACT}(a,b) = \frac{1}{l_{aa}^{+} + l_{bb}^{+} - 2 l_{ab}^{+}} \,, \tag{12}$$

where a and b are nodes, l_{aa}^{+} is corresponding entry in Laplacian Matrix, L^{+}.

3.2.3. Random Walk with Restart

Random Walk with Restart [55] Index is calculated for similarity between nodes. Other variations to this approach include SimRank [56], Neighborhood Formation and anomaly detection [57], PageRank algorithm [58], Local Random Walk [59]. Random Walk with Restart is computed as:

$$S_{RWR}(a,b) = q_{ab} + q_{ba} \,, \tag{13}$$

where a and b are nodes, q_{ab} is the probability with which a random walker starting at node a comes back to node a itself visiting a random neighbor, mathematically stated if a random walker visits random neighbor of node a with probability c and returns to node a with probability $1-c$. q_{ab} and q_{ba} are not symmetric. It is similar to Rooted PageRank index [14].

3.2.4. SimRank

SimRank [56] is calculated for node similarity between nodes, intuition behind SimRank similarity index is two nodes are similar if they are referred by similar nodes [60]. SimRank $S_{SimRank}(a,b)$ is recursively computed as:

$$S_{SimRank}(a,b) = C \frac{\sum_{z \in \Gamma(a)} \sum_{z' \in \Gamma(b)} S_{SimRank}(z,z')}{k_a \cdot k_b} \,, \tag{14}$$

where z is the set of neighbors of node a ($\Gamma(a)$) and z' is the set of neighbors of node b ($\Gamma(b)$). $C \in [0 \ldots 1]$ is the decay factor. $S(a,a) = 1$ is assumed and indicates each node is similar to itself [60].

3.2.5. Escape Probability

Escape Probability [61]–[63] is a global structure based link prediction algorithm computed as:

$$S_{EP}(a,b) = \frac{Q[a,b]}{Q[a,a]Q[b,b] - Q[a,b]Q[b,a]} \,, \tag{15}$$

where a and b are nodes, $Q = \frac{RPR}{1-\beta_{RPR}}$, and RPR is the rooted page rank algorithm, which is similar to Random Walk with Restart similarity measure [63].

3.2.6. Leicht Holme Newman Index-2

Leicht *et al.* [46] proposed a measure to define global structure based similarity measure and it is a path dependent method covering entire network. Leicht Holme Newman index-2 S_{LNH2} is defined as:

$$S_{LNH2}(a,b) = \delta_{ab} + \frac{2M}{k_a k_b} \sum_{l=0}^{\infty} \phi l \lambda^{1-l} (A^l)_{ab} , \qquad (16)$$

where δ_{ab} is Kronecker's function, k_a is the degree of node a, k_b is degree of node b, ϕ is a free parameter, A is an adjacency matrix, M is the total number of edges in the network, l is the path length and λ is an Eigen vector. Two measures are proposed to predict links between nodes in a network, one is local structure based link prediction method (shown in Subsection 3.1.7) and other is global structure based link prediction method (Subsection 3.2.6).

3.3. Quasi-Local Structure Based Similarity Index Methods

3.3.1. Local Path Index

Local Path index [22], [64] is a similarity score calculated between nodes similar to local structure based methods but by considering paths of length > 2. Local Path index S_{LPI} is computed as

$$S_{LPI}(a,b) = A^2 + \varepsilon A^3 + \varepsilon^2 A^4 + \ldots + \varepsilon^{n-2} A^n , \qquad (17)$$

where A is the adjacency matrix, ε is a free parameter and $n > 2$. Equation (17) will be equal to common neighbors similarity index if ε is zero. If $\varepsilon = 0$, Eq. (17) reduces to $S_{LPI}(a,b) = A^2$, which is similar to Common Neighbors similarity based index defined in Subsection 3.1.1. Local Path index is computed for path with length > 2 with undirected edges. Recently Wang *et al.* [65] proposed a variation to local path index method to predict links in a directed network.

3.3.2. Local Random Walk

While computing Local Random Walk index [59] S_{LRW}, random walker is put on node a and initial density is given by $\vec{\pi}_a(0) = \vec{e}_a$, which will evolve as $\vec{\pi}_a(t+1)$ for $t \geq 0$. Local Random Walk index S_{LRW} is computed as:

$$S_{LRW}(a,b)(t) = q_a \pi_{ab}(t) + q_b \pi_{ab}(t) , \qquad (18)$$

where a and b are nodes, q is the initial configuration t represents the initial density at time $t = 0$ and later for $t \geq 0$.

3.3.3. Superposed Random Walk

Superposed Random Walk proposed by Liu and Lü in [59] is the sum of all Local Random Walk in the network. Superposed Random Walk $S_{SRW}(a,b)(t)$ is computed as:

$$S_{SRW}(a,b)(t) = \sum_{\tau=1}^{t} S_{LRW}(a,b)(\tau) = \sum_{\tau=1}^{t} \left[q_a \pi_{ab}(t) + q_b \pi_{ab}(t) \right] , \qquad (19)$$

where a and b are nodes, S_{LRW} is defined in Eq. (18) and t denotes the time steps.

3.3.4. Extended Jaccard Index

The Jaccard index is defined as a similarity index to identify possible link between nodes in common neighborhood. Jaccard index is a local structure based algorithm and considers paths of length 2. With a notion that the nodes tend to connect in path length are more likely to connect in path length ≥ 2, Chartsias [52] proposed extended Jaccard algorithm and implemented on Hadoop parallel Map/Reduce tool for scalability.
Extended Jaccard index is computed as:

$$S_{Jaccard}(a,b) = \frac{|\Gamma_d(a) \cap \Gamma_d(b)|}{|\Gamma_d(a) \cup \Gamma_d(b)|} , \qquad (20)$$

where a and b are two nodes and $\Gamma_d(a)$ and $\Gamma_d(b)$ denote the set of extended neighbors of nodes a and b respectively at hops $1 \ldots d$ for each of the node.

3.3.5. FriendLink

Papadimitriou *et al.* [66] defined FriendLink algorithm for measuring similarity between nodes. FriendLink algorithm is a quasi-local structure based algorithm and uses paths of length ≥ 2. Similarity measure $S_{FriendLink}$ is computed as:

$$S_{FriendLink}(a,b) = \sum_{i=2}^{l} \frac{1}{i-1} \cdot \frac{|paths^i_{a,b}|}{\prod_{j=2}^{i}(n-j)} , \qquad (21)$$

where n is the number of vertices in graph, l is the path length considered $l \geq 2$, $\frac{1}{i-1}$ is the attenuation factor that weights path according to length l. $\prod_{j=2}^{i}(n-j)$ is the number of possible length l-paths from a to b.

4. Summary and Discussion

Local structure based methods of link prediction computes the similarity score based on common neighbors which gives an accurate measure to know link structure arising between nodes. Such measure is computed only between nodes of path length 2 and not beyond that. As a result some interesting and potential links may be missed and also as a matter of fact it will be difficult and time consuming for computing similarity score for all nodes in network. Developing parallel algorithms for such task seems to be interesting but not yet explored.
Global structure based methods of link prediction computes similarity score based on global link structure of graph and computed for nodes having path length > 2. As a result many interesting and potential links can be indentified which are being missed in local structure based link prediction methods. But calculating similarity measure based on global structure is time consuming and difficult in case of large networks such as online social networks where petabytes of data has to be analyzed for predicting links. Developing scalable global structure based algorithms seems to be interesting field but not yet investigated in full pledge except a handful of attempts [52], [66]–[68].

Table 1
Summary of similarity based link prediction algorithms

Link prediction technique	Local	Global	Quasi-local	Scalability
Common Neighbor	✓			Not investigated
Salton index	✓			
Jaccard index	✓			
Sørenson index	✓			
Hub Promoted index	✓			
Hub Depressed index	✓			
Leicht Holme Newman index-1	✓			
Preferential Attachment index	✓			
Adamic-Adar index	✓			
Resource Allocation index	✓			
Katz index		✓		Investigated
Average Commute Time		✓		Not investigated
Random Walk with Restart		✓		
SimRank		✓		
Escape Probability		✓		
Leicht Holme Newman index-2		✓		
Local Path index			✓	
Local Random Walk			✓	
Superposed Random Walk			✓	
Extended Jaccard index			✓	Investigated
FriendLink			✓	

Quasi-local structure based methods of link prediction seems to be more accurate than local link structure based link prediction algorithms. Starting with neighborhood of nodes similar to local structure based methods, quasi-local structure based methods cover entire network by considering all paths between nodes of length more than 2. Some of the methods such as FreeLink, Extended Jaccard are been verified for scalability issues. Further developing algorithms able to compute similarity index for entire network with more accuracy and scalability seems to be interesting problem to work on.

In all the methods discussed in Section 3, information about the local communities are not considered while calculating similarity index. By considering local information about community while calculating similarity index, precision of link prediction increases as shown by Soundarajan and Hopcraft [69] for Common Neighbor and Resource Allocation. But the method proposed by them is not being further explored over other link prediction methods and other bench mark datasets. It will be interesting to investigate similarity based link prediction with local information about community as it can be further extended to Louvain method [70] proposed by Blondel *et al.*

5. Conclusions

In this survey, authors have reported similarity based link prediction algorithms existing in literature and possibility of future research issues. Summary of these methods is given in Table 1. This paper serves as a starting point for researchers and novice readers interested in understanding basics of link prediction in online social networks in general and local structure based link prediction methods in particular.

References

[1] J. Scott, *Social Network Analysis*. Sage, 2012.

[2] Z. Huang and D. K. Lin, "The time-series link prediction problem with applications in communication surveillance", *INFORMS J. Comput.*, vol. 21, no. 2, pp. 286–303, 2009.

[3] K. Jahanbakhsh, V. King, and G. C. Shoja, "Predicting missing contacts in mobile social networks", *Pervas. & Mob. Comput.*, vol. 8, no. 5, pp. 698–716, 2012.

[4] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks", *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106, no. 52, pp. 22073–22078, 2009.

[5] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic block models for relational data with application to protein-protein interactions", in *Proc. Int. Biometrics Society Annual Meeting ENAR 2006*, Tampa, FL, USA, 2006.

[6] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Prediction and validation of gene-disease associations using methods inspired by social network analyses", *J. PLOS One*, vol. 8, no. 9, 2013.

[7] S. Milgram, "The small world problem", *Psychology Today*, vol. 2, no. 1, pp. 60–67, 1967.

[8] S. Goel, R. Muhamad, and D. Watts, "Social search in small-world experiments", in *Proc. 18th Int. Conf. World Wide Web WWW'09*, Madrid, Spain, 2009, pp. 701–710.

[9] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks", in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining ASONAM 2011*, Kaohsiung, Taiwan, 2011, pp. 121–128.

[10] M. E. Newman, "The structure and function of complex networks", *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.

[11] Z. Yin, M. Gupta, T. Weninger, and J. Han, "Linkrec: a unified framework for link recommendation with user attributes and graph structure", in *Proc. 19th Int. Conf. World Wide Web WWW'10*, Raleigh, NC, USA, 2010, pp. 1211–1212.

[12] U. Shardanand and P. Maes, "Social information filtering: algorithms for automating «Word of Mouth»", in *Proc. SIGCHI Conf. Human Fact Comput. Syst. CHI'95*, Denver, CO, USA, 1995, pp. 210–217.

[13] S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: identifying likely adopters via consumer networks", *Statistical Sci.*, vol. 21, no. 2, pp. 256–276, 2006.

[14] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks", *J. American Soc. for Inform. Sci. & Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[15] X. Li and H. Chen, "Recommendation as link prediction in bipartite graphs: a graph kernel-based machine learning approach", *Decision Support Syst.*, vol. 54, no. 2, pp. 880–890, 2013.

[16] F. Ricci, L. Rokach, and B. Shapira, *Introduction to Recommender Systems Handbook*. Springer, 2011.

[17] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender systems", *Phys. Reports*, vol. 519, no. 1, pp. 1–49, 2012.

[18] A. Zeng and G. Cimini, "Removing spurious interactions in complex networks", *Phys. Rev. E*, vol. 85, no. 3, article ID 036101, 2012.

[19] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazières, and H. Yu, "Re: Reliable email", in *Proc. 3rd Symp. Netw. Syst. Design & Implement. NSDI'06*, San Jose, CA, USA, 2006, vol. 6, pp. 22–22.

[20] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: defending against sybil attacks via social networks", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 36, no. 4, pp. 267–278, 2006.

[21] L. Lü and T. Zhou, "Link prediction in complex networks: A survey", *Physica A: Statist. Mechan. & its Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[22] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information", *The Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, 2009.

[23] M. Girvan and M. E. Newman, "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences (PNAS)*, vol. 99, no. 12, pp. 7821–7826, 2002.

[24] R. Guimera, M. Sales-Pardo, and L. A. Amaral, "Classes of complex networks defined by role-to-role connectivity profiles", *Nature Phys.*, vol. 3, no. 1, pp. 63–69, 2007.

[25] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer, "Learning probabilistic relational models", in *Relational Data Mining*, S. Džeroski and N. Lavrač, Eds. Springer, 2001, pp. 307–335.

[26] R. R. Sarukkai, "Link prediction and path analysis using Markov chains", *Computer Networks*, vol. 33, no. 1, pp. 377–386, 2000.

[27] P. Sarkar and A. Moore, "A tractable approach to finding closest truncated-commute-time neighbors in large graphs", in *Proc. 23rd Conf. Uncertainty in Artif. Intellig. UAI 2007*, Vancouver, BC, Canada, 2007 (arXiv preprint arXiv:1206.5259, 2012).

[28] D. Heckerman, C. Meek, and D. Koller, "Probabilistic entity-relationship models, PRMs, and plate models", in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2007, pp. 201–239.

[29] H. Kashima and N. Abe, "A parameterized probabilistic model of network evolution for supervised link prediction", in *Proc. 6th Int. Conf. Data Mining ICDM 2006*, Honk Kong, China, 2006, pp. 340–349.

[30] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction", in *Proc. 7th Int. Conf. Data Mining ICDM 2007*, Omaha, NE, USA, 2007, pp. 322–331.

[31] L. Lü and T. Zhou, "Link prediction in weighted networks: The role of weak ties", *Europhys. Lett. (EPL)*, vol. 89, no. 1, pp. 18001-p1–18001-p6, 2010.

[32] J. Zhao, L. Miao, J. Yang, H. Fang, Q.-M. Zhang, M. Nie, P. Holme, and T. Zhou, "Prediction of links and weights in networks by reliable routes", *Scientific Reports*, vol. 5, article no. 12261, 2015.

[33] N. Sett, S. R. Singh, and S. Nandi, "Influence of edge weight on node proximity based link prediction methods: An empirical analysis", *Neurocomputing*, vol. 172, pp. 71–83, 2016.

[34] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks", *The J. Mathem. Sociol.*, vol. 1, no. 1, pp. 49–80, 1971.

[35] H. Liao, A. Zeng, and Y.-C. Zhang, "Predicting missing links via correlation between nodes", *Physica A: Statist. Mechan. & its Appl.*, vol. 436, pp. 216–223, 2015.

[36] Z. Zhang, Y. Liu, W. Ding, W. W. Huang, Q. Su, and P. Chen, "Proposing a new friend recommendation method, frutai, to enhance social media providers' performance", *Decision Support Syst.*, vol. 79, pp. 46–54, 2015.

[37] M. E. Newman, "Clustering and preferential attachment in growing networks", *Phys. Rev. E*, vol. 64, no. 2, article ID: 025102, 2001.

[38] G. Kossinets, "Effects of missing data in social networks", *Social Networks*, vol. 28, no. 3, pp. 247–268, 2006.

[39] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1986.

[40] A. Rodriguez, B. Kim, M. Turkoz, J.-M. Lee, B.-Y. Coh, and M. K. Jeong, "New multi-stage similarity measure for calculation of pairwise patent similarity in a patent citation network", *Scientometrics*, vol. 103, no. 2, pp. 565–581, 2015.

[41] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901 (in French).

[42] H. Liao and A. Zeng, "Reconstructing propagation networks with temporal similarity", *Scientific Reports*, vol. 5, article no. 11404, 2015.

[43] T. J. Sørensen, *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and its Application to Analyses of the Vegetation on Danish Commons*. Biologiske skrifter, Kongelige Danske videnskabernes selskab. København: I kommission hos E. Munksgaard, 1948.

[44] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks", *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[45] Y.-L. He, J. N. Liu, Y.-X. Hu, and X.-Z. Wang, "OWA operator based link prediction ensemble for social network", *Expert Syst. with Appl.*, vol. 42, no. 1, pp. 21–50, 2015.

[46] E. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks", *Phys. Rev. E*, vol. 73, no. 2, article ID: 026120, 2006.

[47] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering", in *Proc. 5th ACM/IEEE-CS Joint Conf. Digit. Librar.*, Denver, CO, USA, 2005, pp. 141–142.

[48] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks", *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[49] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations", *Physica A: Statist. Mechan. & its Appl.*, vol. 311, no. 3, pp. 590–614, 2002.

[50] L. A. Adamic and E. Adar, "Friends and neighbors on the Web", *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[51] L. Katz, "A new status index derived from sociometric analysis", *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[52] A. Chartsias, "Link prediction in large scale social networks using hadoop", PhD thesis, Technical University of Crete, Greece, 2010.

[53] T. White, *Hadoop: The Definitive Guide*, 3rd ed. O'Reilly Media, 2012.

[54] K.-H. Lee, Y.-J. Lee, H. Choi, Y. D. Chung, and B. Moon, "Parallel data processing with MapReduce: a survey", *ACM SIGMOD Record*, vol. 40, no. 4, pp. 11–20, 2012.

[55] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery", in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. & Data Mining KDD'04*, Seattle, WA, USA, 2004, pp. 653–658.

[56] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity", in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. & Data Mining KDD'02*, Edmonton, AB, Canada, 2002, pp. 538–543.

[57] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, "Neighborhood formation and anomaly detection in bipartite graphs", in *Proc. 5th IEEE Int. Conf. Data Mining ICDM'05*, Houston, TX, USA, 2005, pp. 418–425.

[58] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine", *Computer Networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[59] W. Liu and L. Lü, "Link prediction based on local random walk", *Europhys. Lett. (EPL)*, vol. 89, no. 5, article ID: 58007, 2010.

[60] W. Yu, X. Lin, W. Zhang, and J. McCann, "Fast all-pairs simrank assessment on large graphs and bipartite domains", *IEEE Trans. Knowl. & Data Engin.*, vol. 27, no. 7, pp. 1810–1823, 2015.

[61] P. G. Doyle and J. L. Snell, *Random Walks and Electric Networks*, The Carus Mathematical Monographs no. 22. Mathema Association of America, 1984.

[62] H. Tong, C. Faloutsos, and Y. Koren, "Fast direction-aware proximity for graph mining", in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. & Data Mining KDD'07*, San Jose, CA, USA, 2007, pp. 747–756.

[63] H. H. Song, T. W. Cho, V. Dave, Y. Zhang, and L. Qiu, "Scalable proximity estimation and link prediction in online social networks", in *Proc. 9th ACM SIGCOMM Conf. Internet Measur. Conf. IMC 2009*, Chicago, IL, USA, 2009, pp. 322–335.

[64] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks", *Phys. Rev. E*, vol. 80, no. 4, article ID: 046122, 2009.

[65] X. Wang, X. Zhang, C. Zhao, Z. Xie, S. Zhang, and D. Yi, "Predicting link directions using local directed path", *Physica A: Statist. Mechan. & its Appl.*, vol. 419, pp. 260–267, 2015.

[66] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems", *J. Syst. & Softw.*, vol. 85, no. 9, pp. 2119–2132, 2012.

[67] Y. Dong, C. Robinson, and J. Xu, "Hadoop based link prediction performance analysis".

[68] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features", in *Proc. 3rd IEEE Int. Conf. on Priv., Secur., Risk and Trust and 3rd IEEE Int. Conf. on Social Comput. PASSAT/SocialCom 2011*, Boston, MA, USA, 2011, pp. 73–80.

[69] S. Soundarajan and J. Hopcroft, "Using community information to improve the precision of link prediction methods", in *Proc. 21st Int. Conf. Companion on World Wide Web WWW 2012*, Lyon, France, 2012, pp. 607–608.

[70] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks", *J. Statist. Mechanics: Theory and Experim.*, vol. 2008, no. 10, p. P10008, 2008.

Pulipati Srilatha received her B. Tech. and M. Tech. from Jawaharlal Nehru Technological University, Hdyerabad, India in 2002 and 2006, respectively. Presently she works as Assistant Professor at Anurag Group of Institutions (Formerly CVSR College of Engineering), Hyderabad and pursuing Ph.D. at the School of Computing Science and Engineering, VIT University, Vellore, India. Her interest include data mining, social network analysis and link prediction algorithms.

Email: sreelatha.pulipati@gmail.com
School of Computer Science & Engineering (SCOPE)
VIT University
Vellore, Tamil Nadu, India

Ramakrishnan Manjula received her B.E. in Computer Science & Engineering from University Visvesvaraya College of Engineering (UVCE), Bangalore, Karnataka State, India in 1992 and M.E. in Software Engineering from Anna University, Tamil Nadu, India in 2001 and Ph.D. in Computer Science and Engineering from VIT University, Vellore, India. Presently she is working as Associate Professor at the School of Computing Science and Engineering at VIT University, Vellore, India. Her area of specialization includes software process modeling, software testing and metrics, service oriented architecture, data mining and social network analysis.

Email: rmanjula@vit.ac.in
School of Computer Science & Engineering (SCOPE)
VIT University
Vellore, Tamil Nadu, India