International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 9, No.2 (Mar-2020)
http://dx.doi.org/10.12785/ijcds/090209

A Review of Similarity Measures and Link Prediction Models in Social Networks

Hemkiran S[1]and Sudha Sadasivam G[2]

[1]*Departmentof Computer Science and Engineering, PSG Institute of Technology and Applied Research, Coimbatore, India*
[2]*Departmentof Computer Science and Engineering, PSG College of Technology, Coimbatore, India*

Received 16 Feb. 2019, Revised 31 Jan. 2020, Accepted 23 Feb. 2020, Published 01 Mar. 2020

Abstract: Social network is a web-based platform which enables people to share information, make new connections and explore various events that occur in society. In social networks, link prediction techniques are widely used to discover new indirect relationships that may occur in the future. These techniques are also utilized to effectively detect missing links in any monitored network. This study presents a concise review of the similarity measures, techniques employed in predicting future links and application of link prediction with emphasis on dynamic networks. An analysis of available models for link prediction and their suitability for heterogeneous, large, static or dynamic networks is also presented.

Keywords: Link Prediction, Similarity Measures, Social Networks, Static Networks, Dynamic Networks

1. INTRODUCTION

Social networking sites such as Facebook, LinkedIn, hi5 and Myspace are utilized by several users to post their views on a multitude of topics. Recently, social networks have attracted the attention of researchers aiming to study, analyze and model the interaction among people of a specific group or community [1]. As social networks continuously evolve on a day-to-day basis, novel, time efficient and accurate approaches are required to analyze these networks. Social networks can be visualized in the form of a graph, G = (V, E) where V is the set of vertices (also referred to as nodes) and E is the set of edges (also referred to as links). In social networks, vertices represent people whereas, edges represent the relationships between people [2], [3]. Fig. 1 depicts a sample graph representing the persons a, b, c, d, e in the form of nodes and their relationships. The edges between the nodes indicate that there is relationship between the corresponding nodes. As time progresses, new nodes and edges append to the graph resulting in a progressive increase in the size of a social network. This necessitates a comprehensive study of the entire network structure and various features associated with it. Modelling of social networks is typically performed using graph models [4] in order to enhance ease of visualization and facilitate interpretation of relationships between nodes.

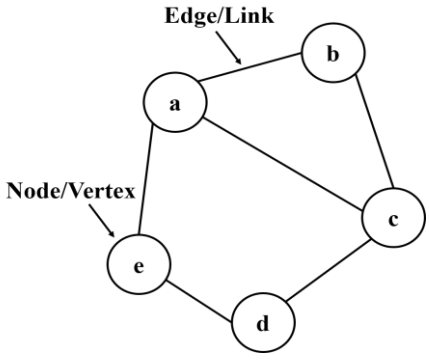

Figure 1. Social network represented as a graph

One crucial area of research in social networks is Link Prediction (LP) [5]. LP can be envisaged as the probability of a new link being added to the network structure at a future time [1], [6]–[8]. By identifying the nodes that will be connected to the structure in the near future, it is possible to determine and predict relationships which may occur in future [9] as shown in Fig. 2.

E-mail: hemkiran@psgitech.ac.in,gss.cse@psgtech.ac.in

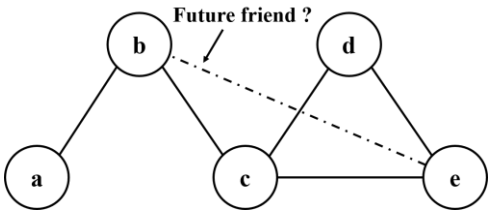

Figure 2. Prediction of future links between nodes

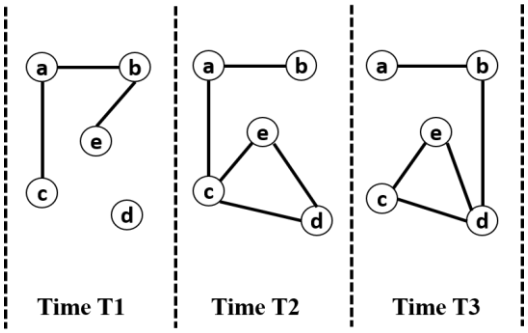

Figure 3. Representation of Temporal Social Network with nodes a, b, c, d and e connected distinctly at various time intervals

The results of link prediction can be utilized for friend recommendation systems like Friendbook, in business for understanding the buying behavior of customers and in military sector.LP acts as a tool to analyze relations in social networks and thereby forecast hidden links which were hitherto non-existent [10]. Most of the existing literatures have focused on LP in static social networks. However, a comprehensive study on social networks requires an investigation into link prediction as applied to dynamically changing nodes [11]. Here, Section 2 presents the existing work on link prediction, Section 3 discusses the utility of link prediction and Section 4 considers the link prediction techniques along with performance metrics. Section 5 presents a discussion on static and dynamic networks, followed by a review of link prediction models in Section 6. Few challenges in the domain of link prediction are stated in Section 7. Section 8 presents the suggested future work and concluding remarks are shown in Section 9.

Generally, a social network is defined as a social structure consisting of many different network nodes where each node stands for a particular individual or organization. A temporal social network can be defined as a sequence of social networks created at different time intervals, 'S' [12]. It is a time varying network whose links are active only at certain point of time as depicted in Fig. 3. The figure indicates that, nodes a, bare connected to c and e respectively at time T1. There exists no connections between nodes c, d and e. However, at time T2,the link between b and e is lost and new connections are established between nodes c, d and e. When we observe later at time T3, the connection between nodes a and c is lost and new relationships occur between node b with d. Thus, as shown in Fig. 3, new relationships appear and disappear between nodes at various time intervals. Since the relationships established between the nodes are transient in nature and exhibit a high degree of impermanence, Fig. 3 constitutes a temporal social network.

2. EXISTING WORK ON LINK PREDICTION

In social networks, many nodes are added or removed from the structure during the time interval't'. Therefore, social networks are dynamic in nature. In order to evaluate the performance of social networks, metrics such as Area Under the Curve (AUCs), precision, accuracy and recall are utilized [13], [14]. Utilizing the aforementioned metrics, the performance of different link prediction algorithms and methods such as Non-negative Matrix Factorization method (NMF), Common Neighbors (CN), Resource Allocation (RA), Preferential Attachment (PA) and local path [15] were compared when implemented on different dynamic datasets such as Irvine message, Enron email, NewsWords, Nodobo and Infectious socio patterns. It was found that NMF was the best among LP algorithms considered for analysis [15].

Prediction of future links between two nodes can also be based on sign. Signed prediction is based on methods such as structural balance theory and status theory. The structural balance theory considers the triads in which three individuals are signed based on the relation between them, as shown in Fig. 4. The status theory suggests the idea of positive directed link which denotes that, the creator of the link views the recipient as having higher status and a negative directed link indicates that the recipient is viewed as having a lower status. A positive sign denotes friendship or trust and a negative sign denotes the unfriendliness or distrust between the nodes considered. Jaccard coefficient is the popular metric utilized to describe the node similarity of signed social networks [16]. The interactions among the social entities extend beyond normal connections. For displaying the social entities in terms of friend or foe, trust or distrust relationships, it is better to represent the social system as a signed network. As reported by Estrada [17], any signed social network containing enormous number of directed nodes were found to be poorly balanced.

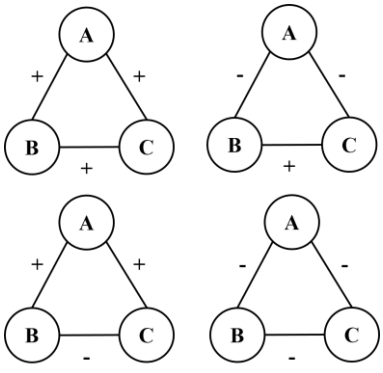

Figure 4. Undirected signed triads

3. UTILITY OF LINK PREDICTION

Dong et al. [10] proposed that LP can be used for friend recommendations. In the domain of Web science and Internet, it can be used for web hyperlink creation, information retrieval and predicting website hyperlinks. Wang et al. [18] reported that LP is used for building recommender systems in e-commerce. In recommendation systems, LP can be used to predict links between users and items referring to purchases or preferences. In citation networks, scholars can find the core papers that will be potentially useful in future [19]. In security related applications, it is used to identify criminals or terrorists [20]. It can be applied to potential collaborators for providing related items in online shopping, recommending patent partners, predicting the mobile contacts in large scale communication network, predicting hyperlinks in heterogeneous social network, understanding the evolution of networks better, in health care and gene expression networks, viral marketing, social mobilization etc [18], [21]. The heterogeneous social network comprises of multiple entities along with complicated interactions among them. For instance, in Digital Bibliography and Library Project (DBLP) network, the authors, papers and conferences act as entities (nodes) whereas the co-author, paper-published-in-conference etc are categorized under interactions (links/edges) [22]. In health care, LP can be used to detect the hidden links between disease and gene to find out the root cause of the disease and recommend the proper treatment [15]. Further, prediction of negative links is performed in the domain of health care medical referral [4] in which selected nodes are removed from the network using LP. In stock market field, it can be used to perform comparative stock performance analysis and effective visualization of correlation patterns [4].

4. TECHNIQUES OF LINK PREDICTION

LP problems can be solved using various techniques such as node based metrics, topology based metrics and social theory [21]. Among these, topology based metrics [21] are widely used to decipher LP problems due to their simplicity of use and applicability to simpler networks without many node and edge attributes. The similarity based methods can be further classified as node based (local features) and path based (global features) methods as shown in Fig. 5. Both these methods are used to find the similarity between two nodes in the network structure by estimating the proximity between two nodes [20], [22], [23]. The similarity or connection between the nodes can be established by studying their common features [24].

A. Topological based metrics

Topological metrics are utilized to determine the similarity between the nodes [25]. The more similar the pair is, the more likelihood the relation between them. Node based metrics can be used for measuring the similarity between the nodes using attributes such as profile in online social network, mail name in email networks, and publication record in academic social network. Path based metrics utilize the whole network topological information in order to measure the path distance between the nodes. A compilation of node and path based metrics discussed in literature is presented in the forthcoming sections.

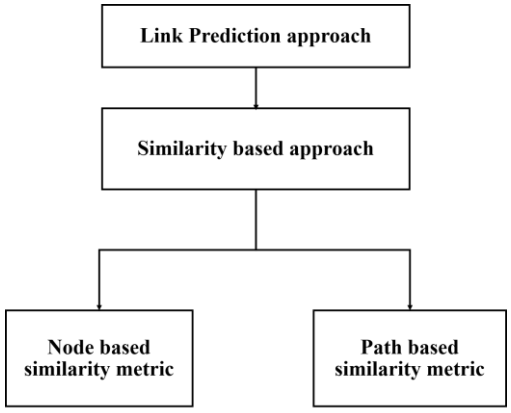

Figure 5. Classification of Link Prediction approaches

B. Node based topological similarity

Node based or local feature based link prediction can be performed using schemes like, Common Neighbors, Jaccard Coefficient, Adamic-Adar, Resource Allocation Index, Preferential Attachment, Friend of a Friend, Locally Adaptive, Sorensen Index, Salton Cosine Similarity, Hub Promoted, Hub Depressed and Leicht-Holme-Newman Parameter Dependent [2], [12]. In these

schemes, all the neighbors are considered as a set and prediction can be made in computing and ranking the similarity of the neighbor set. A concise overview of the node based similarity metrics is presented in the section below.

1) Common Neighbors (CN): In a social network, many people are linked to each other. However, it is not necessary that all people are connected to everyone. If two nodes consist of many mutual friends, they have a high probability of becoming friends in the future. The similarity score based on common neighbors is an accurate measure to understand the link structure between nodes. This measure is computed only between nodes of path length two. Link prediction based on common neighbors can be applied to dynamic networks such as co-authorship network, where nodes represent authors and relationship represents the joint work of authors in atleast one paper. The CN along with metrics such as time varied weight, their changed degree and intimacy between common neighbors were discussed by Yao et al. [26]. CN is commonly preferred for its simplicity and is utilized in Facebook social network.

2) Jaccard Coefficient (JC): JC is a statistic measure used for computing the similarity and diversity of sample sets. It is utilized to find the similarity of two nodes in the graph and is widely used in information retrieval. The Jaccard measure is applied to find the similarity among chemical structures [27]. In co-authorship networks, the keyword match count is measured by using JC.

3) Adamic/Adar (AA): This measure weighs the rarer features more heavily thereby refining the enumeration of common features. For evaluating Adamic Adar index of two people, we find the inverse log of their common neighbor and count its frequency.

4) Preferential Attachment (PA): Users with many friends tend to create more connections in future. It follows the phenomenon of "rich get richer". It is used to find the growth or density of the network. It is calculated by multiplying the number of neighbors of two nodes.

5) Resource Allocation Index (RA): In social network sites such as Facebook, we might find an option of friend suggestion like a page or comment, join a group etc. Thus, a chance of forming link exists between them. For this prediction, we use the resource allocation index.

6) Sorensen Index (SI): It is used for predicting links for different communities based on common neighbor scheme. SI is calculated as the ratio of twice the common neighbors of nodes x and y to the sum of degrees of nodes x and y. This index is used in ecological community data.

7) Salton Cosine Similarity (SC): SC is the statistic measure to compare the similarity and diversity of sample sets.

8) Hub Promoted (HP): HP metric defines the topological overlap and is measured by the lower degree of nodes. It states that edges adjacent to hubs are more likely to get higher similarity score. It is used to analyze metabolic networks.

9) Hub Depressed (HD): HD metric provides the link adjacent to hub with a lower score. It is the ratio of common neighbors of nodes x and y to the maximum of degrees of nodes x and y.

10) Leicht-Holme-Newman (LHN): This metric is used to increase the similarity of nodes in the network. It states that two nodes are similar if their immediate neighbors in the network are similar.

One of the drawbacks of node based metrics is that, potential links may be missed as it cannot compute the similarity between nodes of path length greater than two. It is difficult and time consuming for measuring the similarity scores for all nodes in the network. Parallel algorithms are yet to be developed for performing this.The process of link prediction is a promising framework in thetask. As only local features are considered, they are less efficient.

TABLE I. NODE BASED SIMILARITY METRICS

Name	Score (x,y)	Application of Metrics
Common neighbors [2], [16], [22], [21], [30], [31], [32]	$\lvert \Gamma(x) \cap \Gamma(y) \rvert$	Used in Facebook social network, Predicting potential drug-drug interaction
Jaccard Coefficient [2], [15], [19], [22], [28], [30]	$\frac{\Gamma(x) \cap \Gamma(y)}{\Gamma(x) \cup \Gamma(y)}$	Used in co-authorship network, You-Tube to measure interest similarity, in Epinions and Flixter for user product prediction
Adamic/Adar [12], [2], [21], [22], [30], [32]	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(\Gamma(z))}$	Wikipedia Collaboration graph, Predicting potential drug-drug interaction
Preferential Attachment [12], [2], [21], [22], [24], [30], [32]	$\lvert \Gamma(x) \rvert * \lvert \Gamma(y) \rvert$	Predicting potential drug-drug interaction
Resource Allocation Index [2], [21], [31]	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\Gamma(z)}$	Predicting potential drug-drug interaction
Sorensen Index [24], [33]	$\frac{2 * \lvert \Gamma(x) \cap \Gamma(y) \rvert}{\lvert \Gamma(x) \rvert + \lvert \Gamma(y) \rvert}$	Image Segmentation

Name	Score (x,y)	Application of Metrics
Salton Cosine Similarity [24], [31]	$\frac{\lvert \Gamma(x) \cap \Gamma(y) \rvert}{\sqrt{\lvert \Gamma(x) . \Gamma(y) \rvert}}$	Citation analysis
Hub Promoted [24], [34]	$\frac{\lvert \Gamma(x) \cap \Gamma(y) \rvert}{\min(\lvert \Gamma(x) \rvert, \Gamma(y) \rvert)}$	Metabolic networks
Hub Depressed [24], [31]	$\frac{\lvert \Gamma(x) \cap \Gamma(y) \rvert}{\max(\lvert \Gamma(x) \rvert, \Gamma(y) \rvert)}$	Metabolic networks
Leicht-Holme-Newman [24], [31]	$\frac{\lvert \Gamma(x) \cap \Gamma(y) \rvert}{\lvert \Gamma(x) \rvert . \lvert \Gamma(y) \rvert}$	Measuring vertex similarity

The process of link prediction is a promising framework in the study of complex networks involving the health care sector. The topological information is paramount in predicting novel Drug-Drug Interaction (DDI). In order to predict potential DDI, node based metrics such as CN, JC, AA, PA and RA are used [14].

For predicting the similarity of drug-drug interaction on biological data, chemical data and phenotypic data, metrics like JC, CN, AA, and RA are utilized [29]. In online social network such as YouTube, the Jaccard measure is widely used for measuring the interest similarity [16]. In user-user feature dataset, JC is utilized to pre-process the data for user product prediction in social networks such as Epinions and Flixter [17]. The similarity score of the node based metrics can be summarized as shown in Table 1, where x and y are the nodes that are involved in the social networking structure.

C. Path Based Topological Similarity

Katz [2], [35] proposed a global feature based algorithm to measure the centrality of the given node which is available in the network. This algorithm is used to measure the relative degree of influence of the node in the given network. It is based on the number of walks that occur between the pair of nodes. Other methods such as Random Walk with Restart (RWR) [2], SimRank, and Shortest Path algorithm are also discussed in literature. These algorithms are used for calculating the similarity score between the pair of nodes present in the graph [7], [9]. A concise overview of the path based similarity metrics is presented in the section below.

1) Katz: This metric is the variant of the shortest path measure. The more oaths there are between two vertices and shorter the paths are, the stronger is the connection. It can be used for predicting the similarity of drug-drug interaction on biological data, chemical data and phenotypic data.

*2) Random Walk with Restart (RWR):*This metric is based on the probability that a node will visit its neighborand is calculated for similarity between nodes. It considers the random walker which starts from a node and selects randomly the available edges every time with a probability and returns back to the same node. It can be used for predicting the similarity of drug-drug interaction on medical data.

3) Hitting time: A random walk starts at a node x and recursively progresses to a neighbor of x designated as node y chosen uniformly at random.

4) SimRank: A link analysis algorithm that works on a graph 'G' to measure the similarity between two vertices 'u' and 'v' in the graph.

5) Local Path: This metric uses the information of local path of length 2 and length 3. It includes the information of nodes which is of length 3 distances from the current node. There is an adjustment factor α considered here and A is the adjacency matrix as shown in Table 2.

As these metrics compute similarity scores based on global link structure of graph having path length greater than two, it is possible to trace interesting and potential links. For large networks, it is time consuming and difficult to analyze the data and traverse all paths of network to predict the links. Scalable global structure based algorithms are yet to be developed for handling the critical situations.

Katz measure is applicable in directed networks and World Wide Web. In order to obtain the data from Twitter, the dynamic version of Katz centrality can be used whereas the similarity measure RWR can be used in keyword search in database graphs and for detecting spam. The hitting time metric can be effectively used in query suggestion, recommender networks and for image segmentation problems. The path based metrics can be summarized as shown in Table 2.

TABLE II. PATH BASED SIMILARITY METRICS

Name	Score (x,y)	Application of Metrics
Katz [2],[21], [30],[32], [36]	$\sum^{l} \lvert path_{x,y}^{l} \rvert$ where x, y are the vertices, l is the length and β is the weight factor, $\lvert path_{x,y}^{l} \rvert$ is the set of all paths from x to y with length l, β >0	Directed networks, World Wide Web, Twitter
Random Walk with Restart [21], [23], [34], [37]	$q_{ab} + q_{ba}$ where q_{ab} is the probability of moving random from node a to b.	Database graphs
Hitting time[13], [30], [36]	$- H_{x,y}$	Recommender networks, image segmentation problems

Name	Score (x,y)	Application of Metrics
SimRank [21], [30], [34], [36]	1 if x=y ((∑ a€Γ(x)∑b€Γ(y) score(a,b)) / (|Γ(x)||Γ(y)|))*γ Otherwise	In co authorship networks, world wide web
Local Path [13], [14], [21]	LP = A2 + α A3 where A2 and A3 are the adjacency matrices about the nodes having length 2 and 3, α is the adjustment factor	Directed networks

Among the path-based metrics, Katz measure shows better performance for smaller networks whereas for large networks, random walk is considered better. Certain measures perform better in static rather than dynamic networks. In contrast, hybrid metrics adapt to both local and global features of the network.

When representing the social networks based on health care it is obvious to consider the disease and genes in dynamic manner as the health issues are constantly changing and thus the prediction of new diseases can be made. To predict the gene-disease associations, the inductive matrix completion method is used [38]. In order to predict the gene-disease associations, methods such as Katz measure and Combining dATaAcross species using Positive-Unlabeled Learning Techniques (CATAPULT) can be applied. Katz measure is better in identifying associations between traits and poorly studied genes whereas CATAPULT is better in identifying overall gene-trait association [39]. This method has the advantage of applying to diseases which have not appeared in the training time.

Presently, there is no successful application of the matrix completion approach to recover the gene-disease association matrix. By using inductive method, it is possible to predict the potential genes for a given disease of interest. In gene-disease association matrix, each row corresponds to a gene and each column corresponds to a disease [38]–[43]. When considering the social network of physicians associated with the medical referral process, link prediction models are employed to predict and identify the specialists who are more likely to get future referrals when compared with general practitioners in the same network. In order to predict gene-disease associations, we can apply matrix completion method. The dynamic link prediction method called DyLip was used for generating a series of network evolutionary graphs [11]. Along with the idea of ensemble learning, the dynamic variation of the structural changes in the network was studied and was found to be performing better than static and other state-of-art dynamic methods like moving average, linear regression and simple exponential smoothing [65].

5. Link Prediction in Static and Dynamic Social Networks

A review of literature suggests that a majority of research on LP in social networks has focused on static networks. However, in practice, social networks are dynamic in nature. The ensuing sections summarize link prediction in static and dynamic networks.

A. Link Prediction on static networks

A variety of approaches have been proposed for predicting the links in static networks. Wang et al. [44] suggested the usage of novel probabilistic graphical model which can be scaled substantially to measure the joint co-occurrence probability of two nodes. Benchettara et al. [45] proposed a topological supervised machine learning approach where the link prediction in static networks is obtained by the projection of a bipartite graph. Wang et al. [46] explained the issue of cold-start link prediction by means of latent-feature representation model for the existing users. Incremental learning on latent-feature representation model to adapt the model to a dynamic circumstance can be used. Dunlavy et al. [47] discussed weighted methods for collapsing the temporal data into a matrix and tensor-based approach such as CANDECOMP / PARAFAC (CP) tensor decomposition where temporal patterns are yet to be applied. A combination of the local structural similarity information along with Within and Inter Cluster (WIC) measure was considered by Valverde et al. [48]. This measure is efficient on directed and asymmetric large-scale networks. In [49], Chenbo Fu et al. outlined three supervised learning methods to realize the link weight prediction for single or multi layered network by studying the behaviour of users in offline through their online interaction. The supervised learning methods along with the appropriate network properties results in the success of link weight prediction.

B. Link Prediction on dynamic networks

Normally, the prediction of links occurring in the future were made in static snapshots of networks. But as social networks are evolving widely nowadays, it is necessary to consider the highly dynamic nature of the sequence of snapshots. There is definitely a chance of addition and deletion of nodes and edges during the process of network evolution. In [19], the link prediction in social networks on dynamic networks was considered. The dynamic social network is a series of network snapshots within a time interval which change over time. In social networks while predicting the future links, static snapshots were previously considered. But in contemporary research, predicting links over time depends on the sequence of previous graph snapshots. The naive approach to adapt the static latent space modelling for dynamic networks is to model each node as a single latent representation and update its position whenever the network evolves. Modelling dynamic networks necessitates the consideration of the temporal latent

positions of all nodes involved [50]. In [51], the links among nodes in different time frames are deduced. A few studies have reported link prediction on dynamic networks. In [50], Zhu et al. discussed the local and incremental BCGD (Block Coordinate Gradient Descent) algorithm which models the temporal link probability of node pairs for predicting future links and allowing the model to larger networks without affecting the prediction accuracy. In [51], Oyama et al. concluded that the dimension reduction approach encompassing the time-dependent feature projection Cross Temporal- Locality Preserving Projection (CT-LPP) method gives better accuracy. Fu et al. [49] suggested the supervised learning framework which can effectively study the dynamics of social networks. It yields the higher prediction accuracy than unsupervised and single source supervised models. In [52], Huang et al. proposed the time series models called Autoregressive Integrated Moving Average (ARIMA). Here, the study reveal the combination of static graph link prediction algorithms embedded with time series model considerably improved the prediction accuracy. In [53], Pujari et al. studied the supervised rank aggregation model for predicting the future links in dynamic complex networks which comprises of temporal sequence of graphs. The results stated that this model performed better than the traditional machine learning algorithms for predicting the links. Ahmed et al. [13] suggested non-negative matrix factorization based method to solve the LP problem in transitory networks. This method learns the latent features from the transitory nature of a dynamic network. The incorporation of time and structural information enables the method to achieve better results. Li et al. [54] proposed a novel framework which incorporates the deep learning method called the temporal restricted Boltzmann machine along with the machine learning approach called gradient boosting decision tree. It reduces the computational complexity and permits the algorithm to scale for large networks. This model effectively handles the macroscopic transition and microscopic topology evolution of dynamic networks.

TABLE III. LINK PREDICTION MODELS

S.No	Model Name used for LP	Type of network used	Suitable for Static/ Dynamic Network	Applied Datasets
1	Latent feature model [55]	Large network	Static network	US Patent dataset
2	Dynamic Social Network in latent space model [56]	Large network	Dynamic Network	NIPS Co-publication dataset
3	Information Theoretic model for link prediction [57]	Complex network	Static network	Applied to 12 datasets
4	Ranking factor graph model [22]	Heterogeneous networks	Static network	Epinions, Slashdot, Wikivote, Twitter datasets
5	Temporal Latent Space Model [50]	Large networks	Dynamic Network	Facebook, YouTube, DBLP
6	Three -Level Hidden Bayesian Link Prediction Model [18]	Large networks	Static network	Twitter dataset
7	Markov Link Prediction model [19]	Large networks	Dynamic Network	Twitter, Facebook
8	Balanced modularity maximization link prediction model [20]	Large networks	Static network	Douban, Livemocha
9	Multilevel learning based model [21]	Large networks	Dynamic Network	Facebook, Amazon, Google+

6. LINK PREDICTION MODELS

Many models have been used for predicting the links in social networks. Table 3 categorizes the prevalent models on the basis of their suitability for static or dynamic networks. Zhu et al. [55] discuss the working of improved latent feature model for the task of predicting the links in large scale static networks. Sarkar et al. [56] explore the latent space model which involves the relationship that change over time and results in the linear scaling of computation time and improved performance. In [57], Zhu et al. proposed the information theoretic model for link prediction in multiple structural features. Here, the role of network topology is considered for predicting the missing links. In [22], Dong et al. discusses about predicting the links using random factor graph model on 12 different datasets. It was based on the notion that people on different networks make friends based on their similarity in principles. Zhu et al. proposed the temporal latent space model in dynamic networks which outperforms the existing approaches in terms of scalability and predictive power. Here the temporal latent space representation of nodes is made by Block-Coordinate Gradient Descent (BCGD) algorithm [50]. In [58], the discussion about internal and external factors affecting the formation of links is investigated. A three-level hidden Bayesian Link Prediction model was proposed by integrating the user behavior and relationships. The

experimental results indicated that the model, in addition to mining user latent interest distribution, also effectively improved the performance of link prediction. In [59], Das et al. proposed the novel Markov prediction model over time-varying graph of social network.

This model considers the temporal analysis for predicting the links and its effectiveness lies in the integration of multiple timescales with local and semi-global correlated structural evolution. The balanced Modularity Maximization Link Prediction (MMLP) model which is a partitioned network generative model addresses the limitation of community based link prediction methods as depicted in [60].

7. CHALLENGES IN LINK PREDICTION TECHNIQUES

In order to deal with the highly dynamic nature of social networks, it is essential to have highly time efficient and accurate approaches. The massive size of social networks requires a thorough investigation into scalability aspects. In social network such as twitter, many users are inactive for a longer period of time. Further, link prediction methods have to also consider the negative impact created by many fake users. Fake accounts typically establish random links to other users in the network with malicious intentions such as defaming reputations of celebrities and influencing voting results. Numerous studies have explored machine learning algorithms, graph analysis and classification algorithms for identifying the fake users. Recently, a double layered meta-classifier employing topology based features was proposed by Kagan et al. [61] to detect fake profiles. Similar to link formation, investigating the mechanism of termination of links referred to as link dissolution is paramount. Since the number of links in a social network is essentially large and the possibility of fake profiles existing is high, constructing a representative training set is a challenging task [62]. This is because most link prediction algorithms evaluate the link propensities only over a subset of possibilities. They do not evaluate the entire range. Therefore, the problem of ranking the entire structure in very large networks remains largely unsolved [63]. In security applications, missing an actual link poses severe threat compared to predicting false links. Mostly the data for experimentation comes from the commercial websites and therefore, the quality of the data is unpredictable[64]. Moreover, predicting the links in heterogeneous network is a non-trivial task and to detect the malicious users who develop programs to emulate the real users need to be detected. Most of the existing work considers the indirect connections which add noise and computational complexity to the link prediction problem [37].

8. FUTURE ENHANCEMENT

Very few studies focus on links that may disappear in future. Novel techniques have to be explored to solve the problem of link dissolution. Most of the existing link prediction methods are focused towards static networks. Potential methods have to be developed to handle the dynamic behavior of links. Present-day techniques face a challenge in creating the statistical model and evaluating the same for predicting links as, the preexisting probability of a link is relatively small. The prevalent models for heterogeneous social networks experience high time complexities and are problem-specific. Generic models are essential to solve the aforementioned drawbacks. Further, it is necessary to devise benchmark datasets to compare such novel models.

9. CONCLUSION

This paper presented an overview of link prediction schemes based on node neighborhood and path based techniques. The study revealed that, similarity based approaches were well suited to identify links and it has been depicted here. Numerous studies have addressed static and homogeneous networks whereas very few have investigated dynamic and heterogeneous networks. It is well established that in contemporary social networks, the relationship between nodes is predominantly dynamic in nature. The importance of time dependent dynamic nature over static social network in terms of increasing accuracy and efficiency has been analyzed. Additionally, applications of LP techniques in multifarious domains were reviewed. The main challenge in dealing with complex dynamic networks is their size. This limits the techniques being applied. Techniques to improve dynamic link prediction for large social networks can be proposed.

REFERENCES

[1] N. Gupta and A. Singh, “A novel strategy for link prediction in social networks,” 2014.

[2] S. Gupta, S. Pandey, and K. K. Shukla, “Comparison analysis of link prediction algorithms in social network,” Int. J. Comput. Appl., vol. 111, no. 16, pp. 27–29, 2015.

[3] S. Mishra, “Modeling of social network using Graph Theoretical approach,” pp. 34–37, 2014.

[4] W. Almansoori, S. Gao, and J. Rokne, “Link prediction and classification in social networks and its application in healthcare,” IEEE IRI.

[5] F. Gao, K. Musial, C. Cooper, and S. Tsoka, “Link prediction methods and their accuracy for different social networks and network metrics,” vol. 2015, no. i, 2015.

[6] N. Arora, “An analytical study on link prediction in social networks,” vol. 5, no. 5, pp. 141–143, 2016.

[7] A. Clauset, C. Moore, and M. E. J. Newman, “Hierarchical structure and the prediction of missing links in networks,” vol. 101, pp. 98–101, 2008.

[8] M. Srinivas, Virinchi, Pabitra, “Link prediction in social networks role of power law distribution”, 2016.

[9] J. Yang, L. Yang, and P. Zhang, “A new link prediction algorithm based on local links,” no. 2013, pp. 16–28, 2015.

[10] L. Dong, Y. Li, H. Yin, H. Le, and M. Rui, “The algorithm of link prediction on social network,” vol. 2013, 2013.

[11] Y. Chen and K. Chen, "A link prediction method that can learn from network dynamics," IEEE, pp. 549–553, 2014.

[12] R. Michalski, P. Kazienko, and D. Krol, "Predicting social network measures using machine learning approach," 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., vol. 1056, no. 1, pp. 1056–1059, 2012.

[13] D. Liben-nowell, "The link prediction problem for social networks," no. November 2003, pp. 556–559, 2004.

[14] R. B. C. P. Hially Rodrigues de sa, "Supervised link prediction in weighted networks," Int. Jt. Conf. Neural Networks, pp. 2281–2288, 2011.

[15] N. M. Ahmed, L. Chen, Y. Wang, B. Li, Y. Li, and W. Liu, "DeepEye: Link prediction in dynamic networks based on non-negative matrix factorization," Big Data Min. Anal., vol. 1, no. 1, pp. 19–33, 2018.

[16] W. Yuan, K. He, D. Guan, and G. Han, "Edge-dual graph preserving sign prediction for signed social networks," IEEE Access, vol. 5, pp. 19383–19392, 2017.

[17] E. Estrada and M. Benzi, "Are social networks really balanced?," arXiv:1406.2132v1 [physics.soc-ph], 2014.

[18] H. Wang, W. Hu, and Z. Qiu, "Nodes' evolution diversity and link prediction in social networks," IEEE Trans. Knowl. Data Eng., vol. 29, no. 10, pp. 1–1, 2017.

[19] K.-J. Chen, Y. Chen, Y. Li, and J. Han, "A supervised link prediction method for dynamic networks," J. Intell. Fuzzy Syst., vol. 31, no. 1, pp. 291–299, 2016.

[20] A. K. Rai, R. K. Yadav, S. P. Tripathi, and R. R. Tewari, "A survey on link prediction problem in social networks," vol. 5, no. Ix, pp. 1875–1883, 2017.

[21] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," Sci. China Inf. Sci., vol. 58, no. 1, pp. 1–38, 2015.

[22] S. Nehi and S. Chaudhury, "Link prediction in heterogeneous networks", ACM, 2016.

[23] Y. Dong et al., "Link prediction and recommendation across heterogeneous social networks," 2012.

[24] S. Virinchi and P. Mitra, "Similarity measures for link prediction using power law degree distribution,"In:ICONIP 2013. Lecture Notes in Computer Science, vol 8227, M. Lee, A. Hirose, Z.G. Hou, R.M. Kil Eds. Springer, Berlin, Heidelberg,2013, pp. 257–264.

[25] N. S. Anand Kumar Gupta, "Impact of Topological Properties over Link Prediction Based on Node Neighborhood : A Study," IEEE, no. 2, 2014.

[26] L. Yao, L. Wang, L. Pan, and K. Yao, "Link Prediction Based on Common-Neighbors for Dynamic Social Network," Procedia - Procedia Comput. Sci., vol. 83, pp. 82–89, 2016.

[27] G. Máté, A. Hofmann, N. Wenzel, and D. W. Heermann, "A topological similarity measure for proteins," Biochim. Biophys. Acta - Biomembr., vol. 1838, no. 4, pp. 1180–1190, 2014.

[28] A. Kastrin, P. Ferk, and B. Leskošek, "Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning," PLoS One, vol. 13, no. 5, pp. 1–23, 2018.

[29] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data," BMC Bioinformatics, vol. 18, no. 1, pp. 1–12, 2017.

[30] A. A. Rad, K. E. Ave, and O. O. N. Kn, "Similarity and ties in social networks : a study of the Youtube social network," pp. 1–11, 2013 [Proceedings of the Conference for Information Systems Applied Research].

[31] D. Sharma, U. Sharma, and S. Kumar Khatr, "An experimental comparison of the link prediction techniques in social networks," Int. J. Model. Optim., vol. 4, no. 1, pp. 21–24, 2014.

[32] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," 2011 IEEE Third Int'l Conf. Privacy, Secur. Risk Trust and 2011 IEEE Third Int'l Conf. Soc. Comput., pp. 73–80, 2011.

[33] R. N. Lichtenwalter, N. Dame, J. T. Lussier, N. Dame, and N. V Chawla, "New perspectives and methods in link prediction," ACM, pp. 243–252, 2010.

[34] Z. Wu and Y. Li, "Link prediction based on multi-steps resource allocation," 2014 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol., no. 11172209, pp. 355–360, 2014.

[35] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," IEEE Int. Conf. Soc. Comput., 2011.

[36] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds, "An evolutionary algorithm approach to link prediction in dynamic social networks," J. Comput. Sci., vol. 5, no. 5, pp. 750–764, 2014.

[37] A. K. Gupta and N. Sardana, "Impact of topological properties over link prediction based on node neighborhood: A study," 2014 7th Int. Conf. Contemp. Comput. IC3 2014, no. 2, pp. 194–198, 2014.

[38] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," ACM Comput. Surv., vol. 49, no. 4, pp. 1–33, 2016.

[39] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene-disease associations," Bioinformatics, vol. 30, no. 12, pp. 60–68, 2014.

[40] 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," PLoS One, vol. 8, no. 5, 2013.

[41] J. Fisher and M. Clayton, "Who gives a tweet: assessing patients' interest in the use of social media for health care," Worldviews Evidence-Based Nurs., vol. 9, no. 2, pp. 100–108, 2012.

[42] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, and C. DeLisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," Genome Biol., vol. 10, no. 9, 2009.

[43] W. Almansoori et al., "Link prediction and classification in social networks and its application in healthcare and systems biology," Netw. Model. Anal. Heal. Informatics Bioinforma., vol. 1, no. 1–2, pp. 27–36, 2012.

[44] Z. Dhouioui, H. Tlich, R. Toujeni, and J. Akaichi, "A fuzzy model for friendship prediction in healthcare social networks," Proc. 2016 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2016, pp. 1050–1054, 2016.

[45] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 322–331, 2007.

[46] N. Benchettara, "A supervised machine learning link prediction approach for academic collaboration recommendation," pp. 253–256, 2010.

[47] Z. Wang, J. Liang, R. Li, and Y. Qian, "An approach to cold-start link prediction: establishing connections between non-topological and topological information," IEEE Trans. Knowl. Data Eng., vol. 28, no. 11, pp. 2857–2870, 2016.

[48] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Temporal link prediction using matrix and tensor factorizations," vol. 5, no. 2, pp. 1–27, 2010.

[49] J. Valverde-rebaza, "Structural link prediction using community information on twitter," pp. 132–137, 2012.

[50] C. Fu et al., "Link weight prediction using supervised learning methods and its application to yelp layered network," IEEE Trans. Knowl. Data Eng., vol. 30, no. 8, pp. 1507–1518, 2018.

[51] L. Zhu, D. Guo, J. Yin, G. Ver Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks," IEEE Trans. Knowl. Data Eng., vol. 28, no. 10, pp. 2765–2777, 2016.

[52] S. Oyama, K. Hayashi, and H. Kashima, "Cross-temporal link prediction," Proc. - IEEE Int. Conf. Data Mining, ICDM, no. May, pp. 1188–1193, 2011.

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

[54] R. K. Manisha Pujari, "Supervised rank aggregation approach for link prediction in complex networks," pp. 1189–1196, 2012.

[55] T. Li, B. Wang, Y. Jiang, Y. Zhang, and Y. Yan, "Restricted Boltzmann Machine-based approaches for link prediction in dynamic networks," IEEE Access, vol. 6, pp. 29940–29951, 2018.

[56] J. Zhu and B. Chen, "Latent feature models for large-scale link prediction," Big Data Anal., vol. 2, no. 1, p. 3, 2017.

[57] P. Sarkar and A. W. Moore, "Dynamic social network analysis using latent space models," ACM SIGKDD Explor. Newsl., vol. 7, no. 2, pp. 31–40, 2005.

[58] B. Zhu and Y. Xia, "An information-theoretic model for link prediction in complex networks," Sci. Rep., vol. 5, pp. 1–11, 2015.

[59] Y. Xiao, X. Li, H. Wang, M. Xu, and Y. Liu, "3-HBP: A three-level Hidden Bayesian link prediction model in social networks," IEEE Trans. Comput. Soc. Syst., vol. 5, no. 2, pp. 430–443, 2018.

[60] S. Das and S. K. Das, "A probabilistic link prediction model in time-varying social networks," IEEE Int. Conf. Commun., 2017.

[61] D. Kagan, Y. Elovichi and M.Fire, "Generic anomalous vertices detection utilizing a link prediction algorithm", arXiv:1610.07525v4 [physics.soc-ph], 2017.

[62] J. Wu, G. Zhang, and Y. Ren, "A balanced modularity maximization link prediction model in social networks," Inf. Process. Manag., vol. 53, no. 1, pp. 295–307, 2017.

[63] P. K. Sharma, S. Rathore, and J. H. Park, "Multilevel learning based modeling for link prediction and users' consumption preference in Online Social Networks," Futur. Gener. Comput. Syst., 2017.

[64] E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter, "Using friendship ties and family circles for link prediction," In: Advances in Social Network Mining and Analysis, vol. 5498, L. Giles, M. Smith, J. Yen, H. Zhang (Eds) Springer, Berlin, Heidelberg, pp. 97–113, 2010.

[65] Y. Zhuang, X. Wang, H. Zhang, J. Xie, and X. Zhu, "An ensemble approach to link prediction," IEEE Transactions On Knowledge And Data Engineering, vol. 20, no. 10, pp. 1–14, 2017.

S.Hemkiran is working as Assistant Professor (Sr.Gr) in Department of Computer Science and Engineering at PSG Institute of Technology and Applied Research, India. Her areas of interest include data mining and social network analysis. She completed her B.Tech (IT) and M.E. (CSE) from Anna University.

Dr. G. Sudha Sadasivam is working as Professor and Head in Department of Computer Science and Engineering at PSG College of Technology, India. Her areas of interest include Distributed Systems, Compiler Design, Software Engineering, Theory of Computation and Data Analytics. She has published more than 70 papers in refereed international and national journals, and at conferences. She has published five books in her areas of interest. She has coordinated two AICTE RPS projects in distributed and grid computing arena. She is also the coordinator for PSG-Yahoo RESEARCH on grid and cloud computing.