EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems
Research Article EAI.EU

A Comprehensive Survey of Link Prediction Techniques for Social Network

Abdul Samad [1,*], Mamoona Qadir[2], Ishrat Nawaz[3], Muhammad Arshad Islam[4], Muhammad Aleem[4]

[1]Capital University of Science and Technology, Islamabad Pakistan
[2]Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan Pakistan
[3]The Islamia University of Bahawalpur, Bahawalpur Pakistan
[4]FAST-National University of Computer and Emerging Sciences, Islamabad Pakistan

Abstract

A growing trend of using social networking sites is attracting researchers to study and analyze different aspects of social network. Besides many problems, link prediction is a fascinating problem in the field of social network analysis (SNA). Link prediction, in social network analysis, is a task of identifying the missing links and predicting the new links. Several researchers have proposed solutions for the link prediction problem during the past two decades. However, there is a need to provide comprehensive overview of the significant contributions for a thorough analysis. The objective of this review is to summaries and discuss the existing link prediction algorithms in a common context for an unbiased analysis. The extensive review is presented by constructing the systematical category for proposed algorithms, selected problems, evaluation measures along with selected network datasets. Finally, applications of link prediction are discussed.

Received on 28 January 2020; accepted on 16 April 2020; published on 17 April 2020

Keywords: Link Prediction, Social Network, Survey

doi:10.4108/eai.13-7-2018.163988

1. Introduction

Social network (SN) platforms enable social actors to perform various activities, i.e., information sharing, exchanging views, and learn from other social actors by following them [68]. Interaction of people on public places for e.g., students in universities, customers in restaurants and cafe's can be considered as examples of offline SN. Likewise, SN can be online that are supported by social networking websites (i.e., Twitter [69], Facebook [49] etc.). In graph theory, SN is represented as a social graph, where people are the nodes and their relationships/ interactions are edges (i.e., ties or links).

With the unprecedented growth of the WWW, the tendency of humans to interact, communicate and form relationships with each other has grown manifold. In just a few years, online SNs have become an essential part of our lives and provide us with opportunities to stay connected. This focus towards SNs, have created a lot of opportunities for researchers from different disciplines to study and analyze the various aspects of human behaviors as well as characteristics of the SN. Nevertheless, Analysis of SN is an exceptional task that is facing many difficulties. A lot of problems correspond to SN analysis are being studied, including community detection [28], structural analysis of SN [17], network visualization [71], and finding influential users [105]. Besides, link prediction in SN is one of the most interesting problems, i.e., to predict the formation of a new or unknown link during a given time interval textitT-T+1 using the concept of network mining. Consider a SN in Figure 1, at time T, there are three persons with two edges. The solid link between these persons shows that "Ana" is a common friend of both "James" and "Jack". While, "James" and "Jack" are not connected as there is no link. On the other hand, it becomes interesting to consider at time *T+1* that a new link will appear between "James" and "Jack" or not. The task of predicting friendship between "James" and "Jack" is called link prediction. Additionally, link prediction has variety of applications, i.e., friend recommendation [96], citation recommendation [83], identification of collaborators in co-authorship network

*Corresponding author. Email: writetosamadalvi@gmail.com

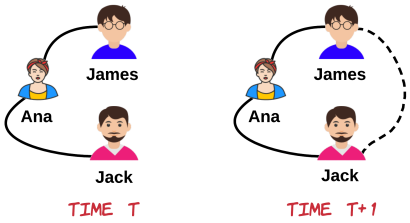

Figure 1. Example of Link Prediction in Social Network

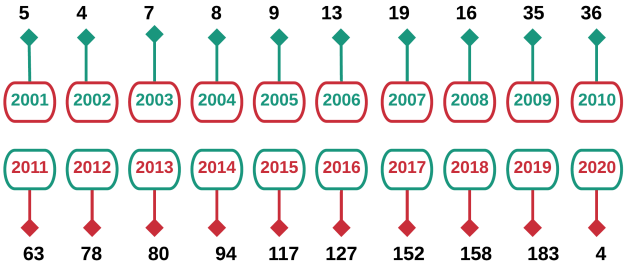

Figure 2. Yearly Publications on the topic of Link Prediction

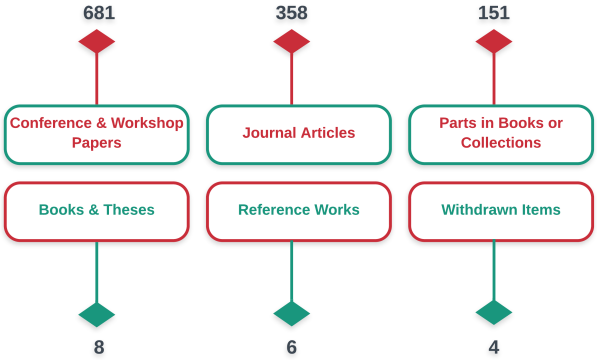

Figure 3. Types of research items on the topic of Link Prediction

[43], identification of criminals in criminal network [59], item recommendation [99] etc.

Figure 2 represents the number of published research papers with search keyword "Link Prediction" on DBLP (Computer Science Bibliography). The significance of link prediction in various domains can be seen during last two decades as shown in Figures 2 and 3. Surprisingly, researchers from different disciplines have done a tremendous job in the last ten years by publishing hundreds of papers on the topic of link prediction. Even, Most of the publications happened in the last year 2019, where 183 research paper are published. This growing number of publications shows that link prediction is a challenging and interesting task for researchers. Besides, 681 out of 1216 research papers are published in conferences and workshops, which means new as well as expert researchers are conducting their research in this area as shown in Figure 3.

In the past, several useful survey have been conducted on link prediction in social network [55] [5] [64]. Liben and Kleinberg [55] give a useful insight and information for link prediction by using classical measures of prediction and topological features of network and can be considered among the pioneer significant work on the topic of link prediction. Lu et al., extended the survey by considering popular algorithms of link prediction for complex networks [5]. However, they have considered their contributions from social sciences, physicial point of view. Although, collection of algorithms considered by Lu et al., is valuable, it still requires deep analysis for the assessment of link prediction techniques. Hassan et al., categorized link prediction methods [64] by considering three types of models: probabilistic, binary classification and linear algebraic. Although, it is best for experts, but not suitable for new researchers who want to learn about link prediction problem.

In oder to include recently proposed link prediction techniques and fulfil the above mentioned shortcomings of the previous surveys, this paper provides a systematic and comprehensive survey on the topic of link prediction in SN. The systematic means, our focuss will be on that type of studies which used various link prediction methods to conduct critical research studies. In addition, we have proposed a taxonomy to categorize the link prediction methods. To the best of our knowledge, this is the first in the last 5 years study that provides a complete picture of the topic of link prediction.

The organization of this article is as follows: In section 2, definition of link prediction problem is explained in detail. Different types of networks that are used for link prediction are explained in section 3. In section 7, link prediction applications are presented. State-of-the-art link prediction methods are discussed in section 4. Detailed discussion on evaluation measures is presented in section 5. In section 6, different kinds of features are explained. Finally, the conclusion is presented in section 8.

2. Problem Definition

- **Definition (Homogenous Network)**: For a given network $G = (V, E)$ where E represents the set of identical links between nodes and V is the set of same type of nodes, then G is called a *Homogeneous Network*.

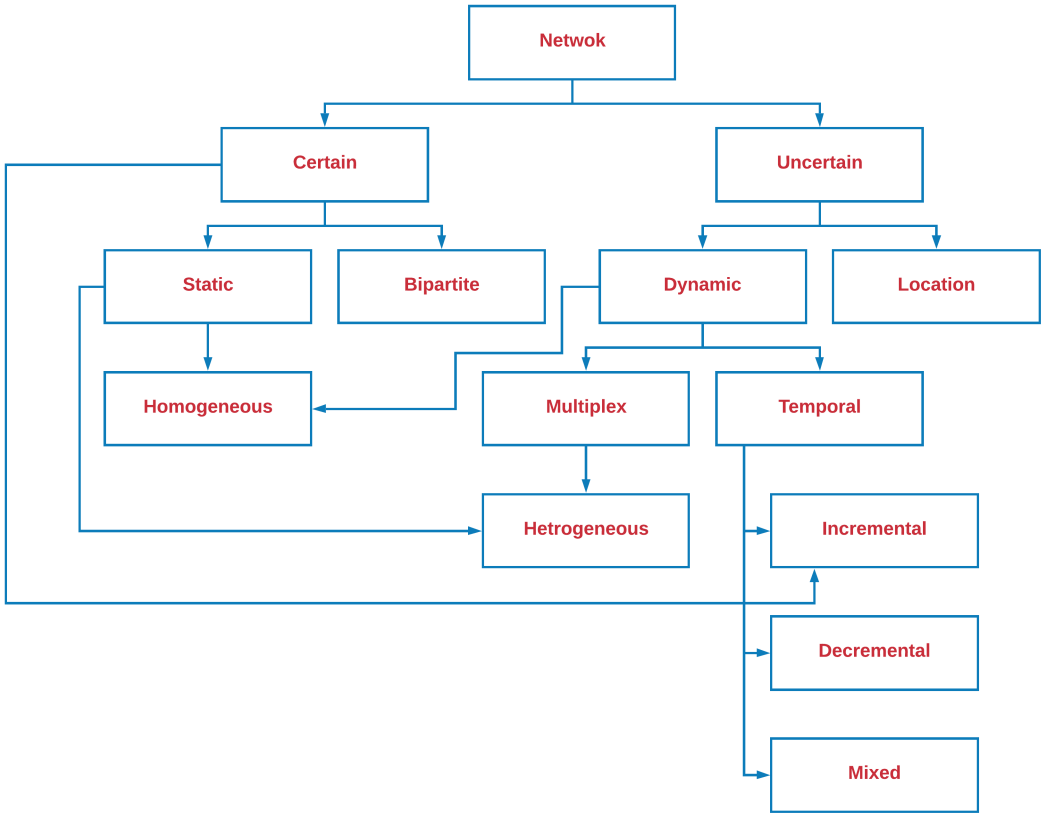

Figure 4. Taxonomy of Networks Used in Link Prediction

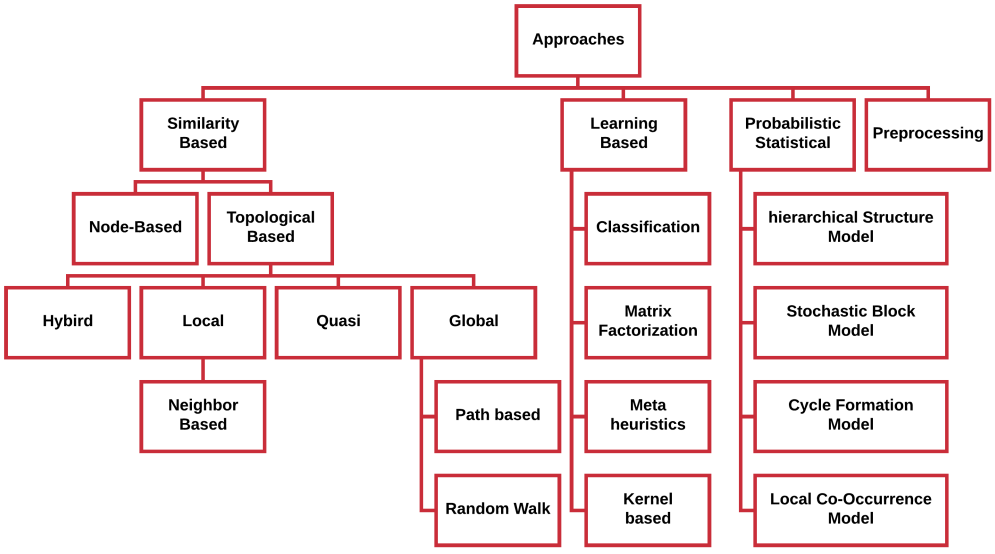

Figure 5. Taxonomy of Link Prediction Approaches

- **Definition (Heterogeneous Network)**: For a given network $G = V(V, E)$ where E represents the set of distinguish links and V is the set of different kinds of nodes, then G is called a *Heterogenous Network*.
- **Definition (Increment Network)**: For a given network $G = (V, E)$ at time t, if at time t+1,

new nodes and edges appeared in G, then new *increment network* at time t+1 is $G_t = (V_t, E_t)$ where $E_t = E_t \bigcup E_{t+1}$ and $V_t = V_t \bigcup V_{t+1}$.

- **Definition (Decrement Network):** For a given network $G = (V, E)$ at time t, if at time t+1, new nodes and edges appeared in G, then new *decrement network* at time t+1 is $G_t = (V_t, E_t)$ where $E_t = E_t \bigcap E_{t+1}$ and $V_t = V_t \bigcap V_{t+1}$.
- **Definition (Mixed Network)**: For a given network $G = (V, E)$ at time t, if at time t+1, some nodes and edges appeared (V_a, E_a) and disappeared (V_d, E_d) in G, then new incremental network at time t+1 is $G_t = (V_t, E_t)$ where $E_t = (E \bigcup E_a) \bigcup (E \bigcap E_d)$ and $V_t = (V \bigcup V_a) \bigcup (V \bigcap V_d)$.

Rely on the different kinds of link prediction approaches and networks, we can formulate the link prediction problem in various ways. Link prediction problem falls into two categories, missing and future links prediction. The formal definition of missing link prediction problem is defined as: consider undirected network graph $G(V, E)$ where E is a set of ties/links and V is a set of nodes/vertices. Moreover, U denotes the set of all possible ties/links $\frac{|V|\times(|V|-1)}{2}$, where $|V|$ is number of nodes in V. Then, $U_n = (U - E)$ is the set of those ties/links, which are not exists. In other words, there are some missing ties/links in U_n. In this case, task of link prediction is to find out those links.

Furthermore, future link prediction problem can be classified into two categorize: periodic and non-periodic link prediction. Periodic link prediction emphasizes on dynamic networks, on the other hand, non-periodic consider the current state of network for prediction.

- **Periodic**: given a graph $G_t = (V, E_t)$ with different snapshots $G_1, G_2, G_3 ... G_t$, where each $e = (u, v) \in E_t$ link took place at time t (as shown in Figure 6). Here, the goal of link prediction is to predict the link state at time step G_{t+1}. In the other words, the objective is to prediction next snapshot of graph.
- **Non-Periodic**: In this case, we have current state of the graph G with only one snapshot G_t instead of series of snapshots (as shown in Figure 7). Consider a graph $G = (V, E_t)$, where E_t is the set of links $E \subseteq (V \times V)$ and V denotes its nodes. Consider subgraphs of G, future G_{t+1} and G_t that $E_t \cap E_{t+1} = E$, $E_t \cup E_{t+1} = \Theta$. Here, objective link prediction is to predict next state of graph i.e., G_{t+1}.

3. Types of Networks

Two types of networks (as shown in Figure 4) are considered for link prediction in the literature: (1) certain and (2) uncertain. The property of certain network is that there is not concept of deletion of nodes and edges. Once, node or edge is added, it will remain there forever and will not be deleted. Co-authorship network is an example of certain network, where authors are represented by nodes and edges between them represents the collaboration of authors. Besides, weight associated with edges represents the number of collaborations among authors. On the other hand, in uncertain networks, probability is attached with each link that a link exists for specified time slot. Further categorize of certain and uncertain networks are as follows.

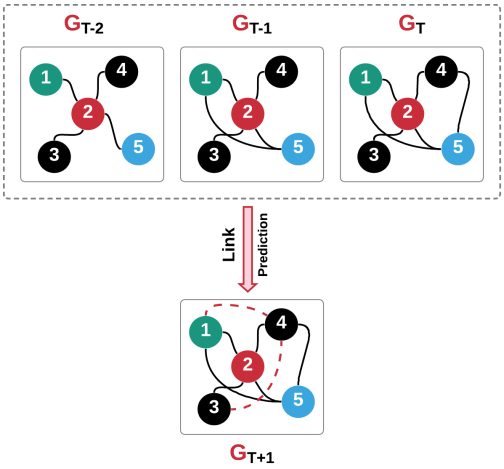

Figure 6. Priodic link prediction: The inputs are graph snapshots in different time intervals

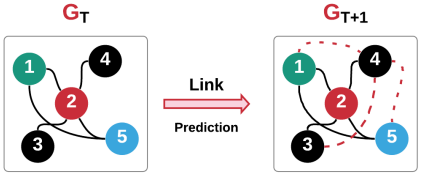

Figure 7. Non-Priodic link prediction: The input is just a snapshot of current time

Static Network:. Static networks (as shown in Figure 9) are those type of networks in which node does not coins its position nor crashes. The whole structure of network along with nodes and edges will remain same. A single snapshot of SN on specific time slot is an example of static network. Static network further classified into Homogeneous and Heterogeneous networks.

Dynamic Network:. Dynamic network (as shown in Figure 8) reshapes its whole structure with the passage of time. Facebook is an example of dynamic network. Due to the dynamics property, network can be change as follows: (1) Edges appear and disappear, but nodes fixed; (2) Both nodes and edges appear

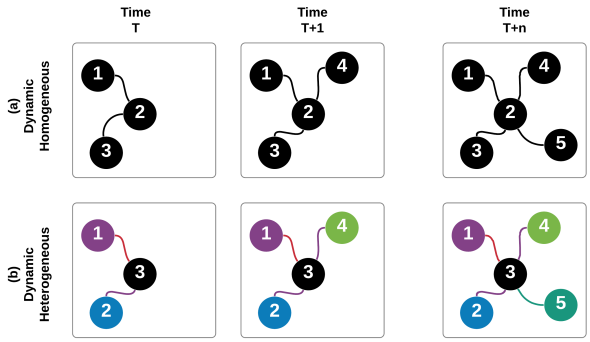

Figure 8. Example of Dynamic Networks

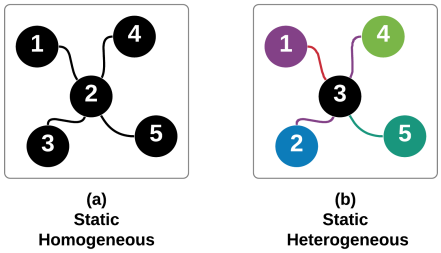

Figure 9. Example of Static Networks

and disappear; (3) Positions of nodes changes, but number of nodes remain same. Dynamic network can be classified further into Homogeneous, Temporal and Heterogenous via Multiplex network.

Homogeneous Network:. Homogeneous networks are those types of networks where nodes and edges are of same type. Co-authorship network is an example of Homogeneous networks, where nodes represents the authors and edges are their collaborations.

Heterogeneous Network:. Heterogeneous networks can be defined as same as Homogeneous networks, but there are different types of nodes and edges exists. Facebook is an example of Heterogenous network, where nodes are connected through different relations i.e., Family, Friend, close fiend etc.

Temporal Network:. Temporal networks are those type of dynamic networks in which nodes and their connections appear and disappear with the passage of time. Besides, information is associated with each link during activation time. Three common depictions of temporal networks are as follows: (1) series of snapshots of network; (2) contact sequence or time of interaction; (3) interval graphs. Temporal networks are further classified into three categorize: Increment, Decrement and Mixed.

Increment Network:. In increment network, number of edges increased due to the appearance of new nodes.

Decrement Network:. In decrement network, number of edges decreased due to the disappearance of existing nodes.

Mixed Network:. In mixed network, number of edges increased and decreased as number of nodes appeared and disappeared.

4. Link Prediction Approaches

4.1. Similarity-Based

Similarity-based approaches believe that nodes try to make links with other similar nodes . These approaches works on the hypothesis that nodes are similar if they have a common connected node or they have a shortest distance in the network. A similarity function $S(u, v)$ is used by these approaches which allocates similarity score to each non-connected pair of nodes u and v. Finally, pair of nodes sorted in descending order according similarity score. A high score represents high probability that nodes will be linked near in a future, while low score shows that nodes will not be linked.

Node-Based. Similarity computation between pair of node is an interesting solution in for the task of link prediction. It builds on the simple idea: as much as the pair is similar, the more chances a link between them. This reflects the fact that people try to make relationship with those people who are similar in religions, language, educations, locations and interest. This relationship can be measure by computing similarity, where score (known as similarity between u and v) is assigned to each pair of nodes (u,v). A high similarity score means u and v will b linked, while low similarity score means u and v will not be linked.

In a practical SN, a node (i.e., people) has profile in online SN containing set of attributes such as gender, age, location, language, interest, bio, country and city. These attributes values can be used to compute similarity between pair of nodes. Most of time, these attributes are in textual form, where textual-based similarities [83] are used. Discussing similarity based approaches is against the purpose of this study, reader can read some comprehensive survey [32].

Samad et al. , in the area of citation network, evaluated both textual and topological similarity measures in order to predict the link between research papers [83]. Where, they have used profiles of research papers containing textual attributes including title and abstract. Their observation is that predicting link between node through topological similarity is better than textual similarity. They also observe that increasing text in attributes lowers the similarity between nodes. Bhattacharyya et al. define tree model with multiple categorize to study and analyze the keywords of profiles, then compute distance of keywords to find similarity between users [11]. They

have found that similarity between users are almost equal except for direct friends. In addition, as much as keywords and friends increased, similarity between users decreased. Akcora et al. observe that most of the user profiles are not publicly available in current social networks or missing. Keeping this limitation in mind, they invent a method that before computing users similarity, estimate the portion of missing values of profile [2]. Anderson et al. use the commonality of user's interest to measure similarity [6]. User's interests are actions that user takes, such as asking question, editing article, reading blog and bookmark items. All these actions take by user can be represented as vector, and user's similarity is the cosine between action vectors. Samad et al. , in the context of face-to-face contact networks, evaluate six different social attributes in order to predict the link [84]. They have found that, language and country are such attributes that plays an important role in contact prediction. They have observe that people tend to contact those people who are similar in language and country.

In conclusion, actions and attributes are mostly used in node-based similarity approaches. These actions and attributes reflect the personal behaviors and interests. In case of having social attributes and behaviors, node-based approaches are useful.

Topological-Based. There are a lots of metrics are exist to compute similarity between two nodes even without node or edge attributes. These metrics used topological information and known as topological-based measures. These metrics are further categorized into local and global metrics.

Local. In a SN, to estimate the similarity of each node with other nodes, local similarity-based methods relies on structural information like neighborhood. These methods are faster, effective and highly parallelizable as compare to nonlocal methods. Moreover, local methods enable us to adequately deal with link prediction issue in changing and dynamic networks like online SN. The primary defect of these methods is that local information (such as neighborhood) restricts nodes to find contacts within neighbors of neighbors. In real-world networks, it is shown that many connections between nodes are formed at greater distance (i.e., more than two) [55]. Nevertheless, local methods have shown competitive prediction results as compare to complex methods. In addition, it is noticeable that, although these approaches are restricted to two-hop, their time complexity is $O(xk^2 f(m))$, where $O(xk^2)$ is spatial complexity and $f(m)$ is similarity computing.

1. *Common Neighbors:* This method is widely used in link prediction. It works same as its name, the more common neighbors, the more chances to linked in future [72]. It bases on the hypothesis that, if two nodes share maximum common neighbors, it increases the chances that their will be link between them than nodes without common neighbors. Most of the researchers agreed on this hypothesis [55]. Similarity cam be computed as follows in Equation 1:

$$CN(u,v) = |\Gamma u \bigcap \Gamma v| \tag{1}$$

Where, Γu and Γv represents neighbor nodes of u and v.

2. *Jaccard Coefficient:* Jaccard Coefficient is known as Jaccard Index, and is basically the normalizes the similarity score of common neighbors by considering intersection over union [36]. For similarity of two nodes u and v, it take in account the common neighbors and total neighbors of both nodes. Besides, Liben et al. showed that Jaccard produced worst results as compare to common neighbors. It can be defined as in Equation 2:

$$JC(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u) \bigcup \Gamma(v)|} \tag{2}$$

3. *SAM:* This method is recently published by Samad et al. [85]. This works on the simple idea that both nodes have their own similarity, i.e., it is possible that one node is 100% similar to another node, but at the same time other node is not similar as first node. SAM similarity can be defined as Equation 3

$$SAM(u,v) = \frac{\frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u)|} + \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(v)|}}{2} \tag{3}$$

4. *Adamic Adar:* Initially, this method was proposed to find similarity among two pages [1]. Later, Liben et al. [55] used the customize version for link prediction problem as shown in Equation 4. In fact, this measure torchere the common neighbors along with hight degree. It can be defined as in Equation 4.

$$AA(u,v) = \sum_{z \in \Gamma(u) \bigcap \Gamma(v)} \frac{1}{log|\Gamma(z)|} \tag{4}$$

5. *Resource Allocation:* This measure is inspired by the process of resource allocation in operating systems. Resource allocation is same as adami adar, but it gives more punishment to common neighbors along with high degree [106]. This is why, both resource allocation and adamic adar have close results. Its foremost feature is that it consider neighbors of neighbors along with direct neighbors. It is defined as in Equation 5

$$RA(u,v) = \sum_{z \in \Gamma(u) \bigcap \Gamma(v)} \frac{1}{|\Gamma(z)|} \tag{5}$$

6. *Preferential Attachment:* This method is proposed by Barbasi et al. [8]. Its main feature is new node will be connected with node having high degree instead of node with low degree. Method can be defined as in Equation.

$$PA(u,v) = |\Gamma(u)| \cdot |\Gamma(v)| \tag{6}$$

7. *Sørensen Index:* This method was proposed by Thorvald Sørensen to find similarity between data samples of ecological community [90]. The foremost objective of this method is to motivate the lower degree nodes in order to find their links. Similarity can be computed as in Equation 7.

$$SI(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u) + \Gamma(v)|} \tag{7}$$

8. *Salton Cosine:* This method is also known as cosine similarity [82]. This method is similar as Sørensen Index and Jaccard Index. Through some studies, it is found that value produces by Salton Cosine is twice the Jaccard Index [34]. Value can be computed as in Equation 8

$$SC(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{\sqrt{|\Gamma(u) \cdot \Gamma(v)|}} \tag{8}$$

9. *Hub Promoted:* Hub Promoted measure proposed by Ravasz et al. during the study of metabolic network [79]. It defines overlap between nodes u and v on the base of topology. Similarity computation defined as in Equation 9.

$$HP(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{Min(|\Gamma(u)|, |\Gamma(v)|)} \tag{9}$$

10. *Hub Depressed:* This measure is same as Hub Promoted, but the similarity value can be computed by nodes with higher degree [107]. Similarity can be defined as in Equation 10.

$$HD(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{Max(|\Gamma(u)|, |\Gamma(v)|)} \tag{10}$$

11. *Leicht-Holme-Nerman:* This measure assigns high similarity score to pair of nodes with more common neighbors [50]. This method take in account the number of actual paths and number of expected paths of length two between two nodes. The authors claimed that it is more sensitive than others in terms of structural equivalence. Similarity can be computed as in Equation 11.

$$LHN(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u)| \cdot |\Gamma(v)|} \tag{11}$$

12. *Parameter-Dependent:* This measure improves the accuracy of link prediction for both unpopular and popular [108]. Here, λ have many goodness that, in case $\lambda = 0$, this measure debased to Common Neighbors. Besides, if $\lambda = 1$ and $\lambda = 0.5$, it debased to Salton Cosine and Leicht-Holme-Nerman, respectively. Formula is shown in Equation 12.

$$PD(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{(|\Gamma(u)| \cdot |\Gamma(v)|)^{\lambda}} \tag{12}$$

13. *Individual Attraction:* This method is same as resource allocation, but it take in account the connections of shared neighbors [27]. It works on the hypothesis that pair of nodes are likely to be connected if they have highly connected neighbors. Similarity can be estimated as in Equation 13

$$IA(u,v) = \sum_{z \in \Gamma(u) \bigcap \Gamma(v)} \frac{|\ell_{z,\Gamma(u) \bigcap \Gamma(v)}| + 2}{|\Gamma(z)|} \tag{13}$$

Where, $\ell_{z,\Gamma(u) \bigcap \Gamma(v)}$ represents the links between nodes of set $\Gamma(u) \bigcap \Gamma(v)$ and node z.

14. *Local Naive Bayes:* This measure works on the hypothesis that every shared neighbor has unique role or influence [52]. This influence or role of node can be computed using statistical theory. Similarity can be estimated as in Equation 14

$$LNB(u,v) = \sum_{z \in \Gamma(u) \bigcap \Gamma(v)} f(z) Log(oR_z) \tag{14}$$

Where, o, R_z and $f(z)$ representing constant for network, role of node and influence measuring function, respectively.

15. *CAR-Based:* This measure is build on the assumption that, there are more chances that two nodes will be linked, if their neighbors are strongly connected in local community [15]. This CAR-Based method is estimated as in Equation 15.

$$CAR(u,v) = \sum_{z \in \Gamma(u) \bigcap \Gamma(v)} 1 + \frac{|\Gamma(u) \bigcap \Gamma(v) \bigcap \Gamma(z)|}{2} \tag{15}$$

16. *Functional Similarity Weight:* This method is a variant of Sørension index. It considers the probability of interaction of both nodes u and v independently in directed network [21]. Nevertheless, this probability score can also be used in undirected networks. This method can be estimated as in Equation 16.

$$FSW(u,v) = \left(\frac{2|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u) - \Gamma(v)| + 2|\Gamma(u) \bigcap \Gamma(v)| + \lambda}\right)^2 \tag{16}$$

Table 1. Comparisons of Local Similarity Measures

Reference	Measure	Time Complexity	Normalization	Remarks
[72]	CN	$O(V^2)$	No	Simple and intuitive. Common neighbors are necessary to predict links.
[36]	JC	$O(2V^2)$	Yes	Intersection over union. Normalizes common neighbor similarity with respect to total neighbors.
[85]	SAM	$O(V^2)$	Yes	Connect lower degree nodes to higher degree nodes. Both nodes have own similarity.
[1]	AA	$O(2V^2)$	No	Give high weight to shared neighbors having few neighbors. Poor results in dense communities
[106]	RA	$O(2V^2)$	No	New nodes likely to be connected with nodes having higher degree. It also gives poor results in dense area.
[8]	PA	$O(2V^2)$	No	It prefers to connect high degree nodes. Not suitable to find links between lower degree nodes.
[90]	SI	$O(V^2)$	Yes	More links would be predict between lower degree nodes. Links between higher degree nodes would get poor results
[82]	SC	$O(V^2)$	Yes	Simple cosine metric. Value produces by Salton Cosine is twice the Jaccard Index.
[79]	HP	$O(V^2)$	Yes	Similarity is computed by node having lower degree
[107]	HD	$O(V^2)$	Yes	Similarity is computed by node having higher degree
[50]	LHN	$O(V^2)$	Yes	High similarity score to pair of nodes having more common neighbors. Not suitable for lower degree nodes or new nodes.
[108]	PD	$O(V^2)$	Yes	Provide better results for predicting popular and unpopular links.
[27]	IA	$O(2V^2)$	No	It works better if shared neighbors are strongly connected. if clustering coefficient is low it will give poor results.
[52]	LNB	$O(V^3)$	No	Every shared neighbors have unique role and influence.
[15]	CAR	$O(V^3)$	No	Depend on the shared neighbors degree. If clustering coefficient is high, nodes will be connected.
[21]	FSW	$O(V^2)$	No	Link likelihood is estimated by the interaction of both nodes.

Global. In order to estimate the similarity between pair of nodes, global similarity-based methods relies on whole structure of network. These methods are not restricted to two node distance as local methods, however, their complexity make them impractical for large networks. In addition, their parallelization is more complex as whole topology of network may not be known by computational agent. Regardless, they shows very diverse time complexities, $O(k2)$ is their spatial complexity as they store similarity score of each pair. Global similarity-based methods are further categorized into path-based and random walk.

Path-Based. Besides neighbor's and nodes information, path is another feature that can be used to estimate similarity between nodes, and this feature is used in path-based methods.

1. *Local Path:* Local path [61] measures uses information about paths with length 2 an 3. Unlike local measures that relies on the nearest neighbors, it takes into account additional information of neighbors of length 2 and 3. Since, neighbors at length 2 are more important than at length 3, so α is used as adjustment factor in measure. This measure can be defined as in Equation 17.

$$LP = A2 + \alpha A3 \tag{17}$$

Where, A2 represents adjacency matrix of nodes with length 2 and A3 denote Adjacency matrix with length 3. Therefore, LP is the adjacency matrix of nodes with length 2 and 3.

2. *Katz:* Katz method [42] is based on ensemble of all paths between two nodes. The paths are damped exponentially by length that can give more importance to shorter paths. The expression is defined as in Equation 18.

$$Katz(u,v) = \sum_{k=1}^{\infty} \beta^k \cdot |path_{u,v}^k| = \beta A + \beta^2 A^2 + \beta^3 A^3 + ... \tag{18}$$

Where, $path_{u,v}^k$ represents all paths of length k that are connecting u and v, and β is the damping factor that is controlling the weights of all paths. In case of very small β, Katz method behaves like Common Neighbors, since short paths perform extra ordinary in final similarity.

3. *Relation Strength:* Relation strength similarity [18] is a kind of asymmetric measure that is suitable for weighted social networks. It takes

into account the strength of relation R(u,v), a normalized weighting score. Considering L paths $(pth_1, pth_2, pth_3, ..., pth_l)$ shorter than r from node u to v, Where, K $(k_1, k_2, k_3, ..., k_z)$ nodes are occurring on pth_l. Then Relation Strength from u to v defined as in Equation 19.

$$RS(u,v) = \sum_{l=1}^{L} R^*_{pl}(u,v) \tag{19}$$

$$R^*_{pl}(u,v) = \begin{cases} \prod_{k=1}^{K} R(z_k, z_{k+1}), & \text{if } K \leq r \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

4. *Shortest Path:* This is Simplest and easy to use global measure. It determined the similarity of u and v by takes into account the shortest path between u and v [3]. The expression looks like as in Equation 21.

$$SP(u,v) = min(|Pu \rightarrow v|) \tag{21}$$

Where, $Pu \rightarrow v$ is a path between u and v, Whereas, $|P|$ denotes the length of path p.

5. *FriendLink:* This method finds similarity by traversing all the paths [73]. It works on the hypothesis that social network users can use all the paths between them. Therefore, similarity between pair of nodes u and v can be estimated as in Equation 22.

$$FL(u,v) = \sum_{i=1}^{l} \frac{1}{i-1} \cdot \frac{|path^i_{u,v}|}{\prod_{j=2}^{i}(n-j)} \tag{22}$$

Where, n is the size of network, l is the path length between u and v, $path^i_{u,v}$ denotes all paths between u and $v$with i length. In addition, higher l will cause for the poor performance.

Compared with neighbor-based methods and node-based methods, which are restricted to local community information, path-based methods takes into account additional topological information. Where, they consider not only local, but also global information such as paths between pair of nodes. However, path-based methods are more expensive than local methods in terms of time complexity.In addition, longer paths are rarely used and not more useful. Sarkar et a. [86], in their study, shows that if shorter paths are not enough, longer paths will be useful. Therefore, path-based methods can produce better results in case of removing too longer paths.

Random Walk. Similarity between nodes in SN can also be calculated by random walk. Random walk takes into account the amount of transition from current node to its neighbors. There are few similarity measures that uses random walk to find similarity between nodes.

1. *Random Walk:* In 1905, Random walk were coined by Karl Pearson [75] and have been adopted by many researchers from various disciplines such as physics, economics and biology. Consider a graph along with starting node, suppose we randomly pick one of its neighbor and proceed to it, then repeat step for every reached node. This series of randomly picked nodes is called random walk [60]. Let p^u is probability vector of starting node u to reaching any anode in the network, thus, the probability of starting node to reaching any node is iteratively estimated as in Equation 23.

$$\vec{p^u}(t) = Mat^T \vec{p^u}(t-1) \tag{23}$$

Where, Mat is the matrix of transition probability computed by adjacency matrix Am, with $Mat_{i,j} = Am_{i,j} / \sum_k Am_{i,k}$. In addition, $p^u(0)$ assigned 0 to all its elements, except $p^u_u(0)$, where value is 1.

2. *Randome Walk with Restart:* Following the definition of random walk, if the walker come to the point with probability $(1-\alpha)$ where he started, this is known as Random Walk with Restart [94]. The updated Equation is 24.

$$\vec{p^u}(t) = \alpha Mat^T \vec{p^u}(t-1) + (1-\alpha)s^u \tag{24}$$

Where, $p^u(0)$ assigned 0 to all its elements.

3. *Hittime Time:* Hittime time [29], takes into account the average steps required to reach at node v from node u. Usually, it is asymmetric measure which means $HT(u,v) \neq HT(v,u)$. Let $Mat = D_A^{-1} Am$, Where D_A is estimated as $(D_A)_{i,j} = \sum_j Am_{i,j}$, . Based on probability matrix Mat, Hitting Time can be defined as in Equation 25.

$$HT(u,v) = 1 + \sum_{k \in \Gamma(u)} Mat_{u,k} HT(k,v) \tag{25}$$

4. *Commute Time:* Commute Time [60] consider the Hitting Time value of both nodes u and v such as, $HT(u,v)$ and $HT(v,u)$. It takes into account the expected steps of walk from u to v and v to u. Commute Time is defined as in Equation 26.

$$CT(u,v) = HT(u,v) + HT(v,u) \tag{26}$$

5. *Cosine Similarity Time:* Cosine similarity time method is used to find the similarity of two vectors [29]. It is based on $Q^\dagger$, where $Q^\dagger$ is pseudo-inverse of $Mat = D_A - Am$. It is estimated as follows in Equation 27.

$$CST(u,v) = \frac{Q^\dagger_{u,v}}{\sqrt{Q^\dagger_{u,u} Q^\dagger_{v,v}}} \tag{27}$$

6. *SimRank:* SimRank is a unique method that takes into account the point where two random walkers meet [38]. It works on the hypothesis that if two random walkers are meet at node, then they are similar to each other. It is estimated as follows in Equation 28.

$$SR(u,v) = \begin{cases} 1, & \text{if } u = v \\ \gamma \cdot \frac{\sum_{a\in\Gamma(u)}\sum_{b\in\Gamma(v)} SR(a,b)}{\Gamma(u)\cdot\Gamma(v)}, & \text{otherwise.} \end{cases} \tag{28}$$

Where, parameter γ is used to control the weight of connected random walkers.

7. *Rooted PageRank:* Rooted PageRank [55] is another variant of PageRank centrality, which is used to rank the search results.The rank is decided on the random walk of node in the graph. Moreover, factor γ represents the visit of starting node to its neighbors. Let D consist of diagonal values of adjacency matrix $Am, D_{i,i} = \sum_j Am_{i,j}$. Then, Rooted PageRank is estimated as follows in Equation 29.

$$RPR = (1-\gamma)(I - \gamma D^{-1} Am^{-1}) \tag{29}$$

8. *PropFlow:* This measure is same as Rooted PageRank, however, it is localized more than that [57]. It restricts the steps of random walker to l steps. In other words, If random walker starts its walk from u and going to v, then it takes no more than l steps. It pick links on the base of weight. It can be defined as follows in Equation 30.

$$PF(u,v) = PF(x,u)\frac{w_{u,v}}{\sum_{k\in\Gamma(u)} w_{u,k}} \tag{30}$$

9. *SepctralLink:*SpectraLink method is proposed by Symeonidis et al. [92], which is used to capture the proximity of node by enhancing the method of spectral clustering. It takes into account the Laplacian matrix, and produced noise free matrix, which is more compact and smaller. Therefore, it predicts more accurate links. They also extend there work to predict negative and positive links in social networks [93].

Quasi. Quasi approaches have recently appear to force the balance between global and local similarity methods. Quasi methods are almost as effective to compute similarity as local methods, besides, also consider additional structural information, as global methods consider. Some Quasi methods consider the whole structural information, but their time complexity is still below than global methods. spatial complexity of quasi methods is $O(uk^{2+s})$, where s relies on the parameters that set the length of path or number of iterations.

1. *Local Random Walk:* Local random walk [60] measure uses the random walk from source to destination, but restrict the iterations to a small number k. Similarity is estimated as follows in Equation 31.

$$SRW(u,v) = \frac{|\Gamma(u)|}{2|E|}\vec{p_v^u}(t) + \frac{|\Gamma(v)|}{2|E|}\vec{p_u^v}(t) \tag{31}$$

Where, $\vec{p_v^u}(t)$ represents the probability vector, estimated on iteration t.

2. *Supervised Random Walk:* Supervised random walk uses the topological information of node and link features [60]. The foremost objective of this method is to releasing the random walker continuously at the starting node. It can be defined as follows in Equation 32.

$$SRW(u,v) = \sum_{i=1}^{t} \frac{|\Gamma(u)|}{2|E|}\vec{p_v^u}(i) + \frac{|\Gamma(v)|}{2|E|}\vec{p_u^v}(i) \tag{32}$$

Hybrid.

1. *Evidential Measurement:* Yin et al. proposed evidential measurement method [102]. This is an hybrid technique which requires both node and local similarity. It is estimated as follows in Equation 33.

$$EM_{i,j} = \sum_{z\in\Gamma(u)\bigcap\Gamma(v)} \frac{\varphi_{i,j}}{\phi z} \tag{33}$$

2. *Methods in Weighted Networks:* Link prediction also has been used in weighted networks. Some of measures are: weighted adamic/adar, weighted common neighbors, weighted resource allocation [63]. Measures are given in Equations 34, 35 and 36. weighted version of adamic/adar.

$$WAA(u,v) = \sum_{x\in\Gamma(u)\bigcap\Gamma(v)} \frac{(w(u,v)^\alpha + w(x,v)^\alpha)}{\log(1+s(x))} \tag{34}$$

weighted version of common neighbors.

$$WCN(u,v) = \sum_{x\in\Gamma(u)\bigcap\Gamma(v)} w(u,v)^\alpha + w(x,v)^\alpha \tag{35}$$

weighted version of resource allocation.

$$WRA(u,v) = \sum_{x\in\Gamma(u)\bigcap\Gamma(v)} \frac{(w(u,v)^\alpha + w(x,v)^\alpha)}{s(k)} \tag{36}$$

Where, $\Gamma(u)\bigcap\Gamma(v)$ represents the common neighbors of nodes u and v, $w(x,v)$ represents the weight of link between x and v i.e., $S(u) = \sum_{x\in\Gamma(u)} w(u,v)^\alpha$.

4.2. Learning-Based

Classification. Let $u, v \in V$ are nodes from the graph $G(E, V)$ and $l^{(u,v)}$ is the label of pair of nodes (u, v). In the link prediction problem, using classification, we denote every pair of node (non-connected) as instance with class label. If the nodes are connected, label says it positive, otherwise says it negative. In general, label of pair (u,v) defined as in Equation 37

$$l^{(u,v)} = \begin{cases} 1, & \text{if(u,v)} \in E \\ 0, & \text{if(u,v)} \notin E. \end{cases} \tag{37}$$

Classification is powerful concept to deal with link prediction problem, even it can use any kind of similarity measures as features as shown in Figure 10. However, this kind of approach have to deal with serious problem which is known as class imbalance [46]. A lot of classifier-based methods have been developed, and any kind of classifier can be a part of such approaches. Few of the researchers have compared many classifiers for the link the link prediction i.e., support vector machine, decision trees, multilayered perception, k-nearest neighbors, naive Bayes and many ensembles of these classifiers [4]. While, other researchers have found random forest as a good one [23]. For building an effective and efficient classifier, it is crucial to extract and define desirable feature set from SN. From the past studies, topological-based and node based features are proved as important for classification models i.e., VCP measures is a distinctive feature which represents topological information [56]. Li et al. proposed graph-based learning model using profile features (i.e., book title, education, age, introduction and keywords etc.) to predict a link between user and item in bipartite network [53]. Likewise, Scellato et al. [87] developed a classification model based on place features, social features and global features. In addition, supervised learning-based framework is used for link prediction in location-based network. Similarly, Ichise et al. have considered co-authorship network for link prediction and proposed semantic-based approach which uses title, abstract and information of event to predict links between authors [74] [97]. For the link prediction, Scripps et al. developed discriminative classification model based on matrix alignment [89]. The foremost objective of this model is to determine the efficient and predictive attributes and features. It computes the weighted similarity measures using node and topological features for the alignment of adjacency matrix. People usually perceive that weight as feature plays a important role in link prediction. However, the previously statement is not verified yet, even in few studies, performance is damaged. Few of the research works claimed that weights would be helpful for the improvements of prediction results in supervised link prediction [25]. On the other hand, few studies show that weights are futile for unsupervised link prediction [62]. So, it still needs to explore the datasets to find the importance of weights. Kunegis et al. developed a framework which learns the functions of edge weight and link prediction [48]. This framework efficiently estimates the parameters and generalizes both dimensionality reduction and graph kernel methods. First of all, it obtain alternatives that apply to weighted, undirected, unweighted, bipartite and unipartite graphs, then combine the link prediction functions. Pujari et al. introduced dyadic link prediction method on the base of social choice supervised algorithm [76]. In signed social network, social links represents the social behaviour of users to each other i.e., trust, friendship or hostile etc. Wang et al. observed that a link prediction method can be derived through social imbalance in SN [20]. They proposed a link prediction method based on supervised learning and uses the features obtained from the larger cycles in SN. Leskovec et al. investigated the signed social network, where links can be either negative(i.e., enemies) or positive(i.e., friends) [51]. During the investigation, they have found that we can achieve the high prediction accuracy using classifications models that takes into account the basic rules of signed links formation. Cao et al. talked about the data sparsity in link prediction by contemplating prediction of multiple links from heterogeneous domains i.e., link prediction between users and various items namely a problem of collective link prediction [16]. Lue et al. have developed a supervised learning based classification framework [65]. The foremost object of this framework is to effectively learn the dynamics of SN and construct different kinds of path-based features for link prediction. Wu et al. have proposed classification framework of interactive learning [98], which divides in three steps: (1) Applying similarity methods on different features (i.e., homophily) to find candidate nodes. (2) Rank the nodes according to RankFG model. (3) Allow users to send feedback.

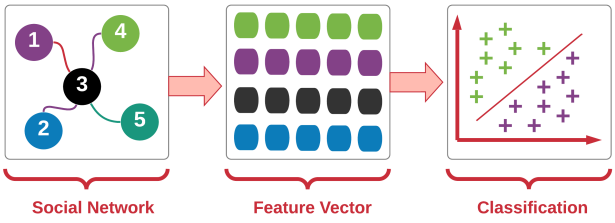

Figure 10. Flowchart of Classification Model

All of the above were supervised learning based frameworks for link prediction. Besides, few of the research works have shown that semi-supervised

learning frameworks can also be a part of link prediction. Kashima et al. used semi-supervised based learning methods to developed link propagation classification framework for link prediction [41]. Where, the main objective was to predict the unrevealed chunks of the network using similarity of nodes. In addition, since it can fill the missing chunks of the network, it enable us to predict different kinds of links simultaneously. As a variant, another fast algorithms of link propagation is proposed to answer the linear equations in the method [80]. Brouard et al. [14] tried to predict the links through kernel regression, a semi-supervised learning approach.

Matrix Factorization. Matrix factorization are such kind of approaches that extract and uses additional or latent features for link prediction and have been used by various recommender systems [45]. Menon et a. have proposed a learning method to learn latent features for link prediction [70]. This learning method considers a vector $\vec{l_i}$of latent features for every node in the network, scaling factor SFu, v for ever pair of node, node feature's weights W_n and link feature's weights $\vec{w_l}$. Moreover, feature's vector $b_{u,v}$ corresponds to link and $\vec{a_i}$ corresponds to node. This model computes prediction score for nodes u and v as follows in Equation 38.

$$MF(u,v) = \frac{1}{1 + exp(-\vec{l_u}^T F \vec{l_v} - \vec{a_u}^T W_n \vec{a_v} - \vec{w_l}^T \vec{b_{u,v}})} \tag{38}$$

Meta-Heuristics. Involvement of plenty of factors makes links formation a complicated process. Most of the link formation methods are heuristics as they try to give high accuracy than other predictors by making hypothesis in the network. Bliss et al. have proposed a method for link prediction on the hypothesis that, different heuristics of link formation can cooperate and coexist [12]. It optimize various link predictors (i.e., global and local similarity measures) by adopting evolution policy. Ever solution x is represented by vector $w^{(x)}$ of real numbers as heuristics number. A similarity function for each candidate predictor is as follows in Equation **??**.

$$S(u,v) = \sum_{i=1}^{|w^{(x)}|} w_i^{(x)} s_i(u,v) \tag{39}$$

Kernel-Based. Kunegis et al. developed a kernel-based method for link prediction that integrates different graph kernels as well as methods of dimensionality reduction [48]. The learn ability of this method makes it unique, since it is capable to learn F function which exerts adjacency or Laplacian matrix. Let there are training and testing sets of X and Y adjacency matrices for link prediction. Now, consider a function F(spectral transformation) which maps adjacency matrix X to adjacency matrix Y with least error using optimization problem as follows in Equation 40

$$Min_F \|F(X) - Y\|_F \tag{40}$$
$$s.t. F \in S$$

Where, $\|F(X) - Y\|F$ corresponds to *Frobenius norm*. The constrain in this norm ensure that spectral transformation function F is the property of another function S(known as function of spectral transformation). Consider a symmetric matrix $X = M\Lambda M^T$ for function F, then we have $F(X) = MF(\Lambda)M^T$, where function $F(X)$ applies on every eigenvalue. Moreover, optimization problem, as shown in Equation 40, can be resolved by calculating the eigenvalue $X = M\Lambda M^T$ as shown in Equation 41.

$$\begin{aligned} &\|F(X) - Y\|F \\ &= \|MF(\Lambda)M^T - Y\|F \\ &= \|F(\Lambda) - M^T Y M\|F \end{aligned} \tag{41}$$

Since, the entries other than diagonal are independent from spectral function F, therefore, optimization function can be converted from matrix to real numbers as follows in Equation 42.

$$Min_f = \sum_i (f(\Lambda_{ii}) - M_i^T Y M_i)^2 \tag{42}$$

Spectral function F can used many kernels as described below.

1. *Exponential Kernel:* Consider a unweighed graph G Along with adjacency matrix Am. Now, Am^n corresponds to count of paths with length n. On the base of hypothesis that nodes that are connecting through more paths are more similar than nodes that are connecting through few paths. So, function F can be estimated as follows in Equation 43.

$$F_{EK}(Am) = \sum_{i=0}^{e} \beta_i Am^i \tag{43}$$

   Thus, Exponential kernel defined as below in Equation 44

$$EK(\beta Am) = \sum_{i=0}^{\infty} \frac{\beta^i}{i!} Am^i \tag{44}$$

2. *Von-Neuman Kernel:* Von-Neuman is same as exponential kernel as it also count the number of paths. it can be expressed as follows in Equation 45

$$(I - \beta Am)^{-1} = \sum^{\infty} \beta_i Am^i \tag{45}$$

3. *Laplacia Kernel:* The basic idea behind this method is use the functions on Laplacian matrix Lm instead of adjacency Am. The Laplacia matrix can be expressed as $Lm = D - Am$, here, D is corresponds to matrix of diagonal degree. In addition, another normalized version of Laplacian matrix is computed as $Ł = I - D^{-1/2}LmD^{-1/2}$. Most of the graph kernels have been defined on Lm i.e., by taking pseudo-inverse of Lm commute time kernel can be defined as:

$$Fc(Lm) = Lm^{+}$$

$$Fc(Ł) = Ł^{+}$$

Similarly, regularized version of commute time kernel can be defined as follows.

$$Fcr(Lm) = (I + \beta Lm)^{-1}$$

$$Fcr(Ł) = (I + \beta Ł)^{-1}$$

Moreover, diffusion kernel can also be estimated as follows.

$$Fd(Lm) = \exp(-\beta Lm)$$

$$Fd(Ł) = \exp(-\beta Ł)$$

4.3. Probabilistic Model

In the literature, a number of network formation models have been studied and discussed in terms of probabilistic and statistical approaches [31]. These approaches stepped into the problem of link prediction on the base of probability and statistical analysis. These probabilistic methods usually suppose that the network which is going to be studied has a known structure. In addition, set of model parameters are estimated in order to build a model. Furthermore, for each missing link, formation probability is computed on the base of these parameters. Finally, formation probability values are sorted the important links as we did in similarity based approaches.

Hierarchical Structure Model. According to the literature, most of the real networks are organized as hierarchically, such as protein-protein interaction network, metabolic networks, social networks like actor network and internet domains [78]. Where lower degree nodes are expected to have higher clustering coefficient than higher degree nodes. In 2008, Clauset et al. have proposed a method that delineates hierarchical network by a dendrogram with [v]-1 internal nodes and [v] leaves [22] as shown in Figure 11. Where, each leaf corresponds to network node and each internal node corresponds to relationship between its descendant nodes. Moreover, probability Pro_n is attached with every internal node, which is uniform to the probability of link

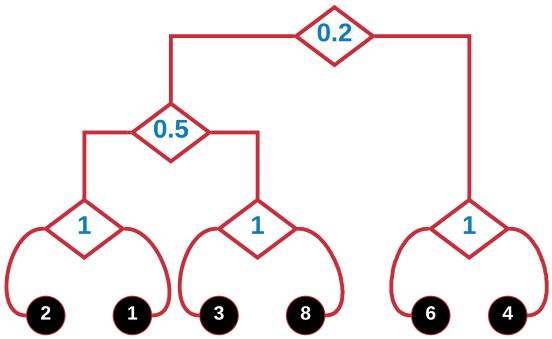

Figure 11. Example of Network's Dendrogram

among those nodes which are descending from it. Consider a dendrogram in Equation 1, where Den is the representation of network and e_k represents the number of those links which are connecting internal nodes k in Den. The likelihood of network can be estimated as in Equation 46.

$$L(D_{en}, P_k) = \prod_{k \in D_{en}} P_k^{e_k}(1 - P_k)^{l_k r_k - e_k} \tag{46}$$

Where, r_k and l_k representing the number of leaves from right and left subtrees along with root k. Consider Figure 10, where a dendrogram of hierarchical network is shown. As per to dendrogram, there is 0.5 connecting probability of nodes 2 and 3. On the other hand, nodes 1 and 6 have 0.2 connecting probability.

A Markov Chain Monte Carlo [30] approaches is employed to sample a set of dendrograms with a probability corresponding to their likelihood. The goal is to regroup subtrees of current dendrogram in another order.

Stochastic Block Model. In reality, not all networks meet the requirements of hierarchical schema. A common approach is to assume that nodes in the network are distributed in blocks and communities, where nodes belongs the same group or community have same status [33]. The chances of link formation between two nodes depends on the community or block they belongs. A stochastic model usually consist of two parts P and $P_{ro}M$, such as $M_{od} = (P, P_{ro}M)$. Where, P is the partition method, and $P_{ro}M$ is the probability matrix of two nodes belongs to two different communities. Let $P_{ro}M_{\alpha\beta}$ be the probability among two blocks α and β and G is the network. Likelihood can be computed as follows in Equation 47.

$$M_{od}(G|P, P_{ro}M) = \prod_{\alpha,\beta \in P_{ro}M} P_{ro}M_{\alpha,\beta}^{\ell_{\alpha,\beta}}(1 - P_{ro}M)^{\gamma_{\alpha,\beta} - \ell_{\alpha,\beta}} \tag{47}$$

Where, $\ell_{\alpha,\beta}$ represents links between nodes in block α and β, while $\gamma_{\alpha,\beta}$ represents those links that exists

between both blocks. The foremost feature of this model is that it allow us to identify spurious as well as missing links from noisy data in the network. In addition, provide better prediction results than hierarchical structure model. However, its computation complexity is high and have not much ability to present possible overlapping. To overcome shortcomings of previous model, Chen and Zhand proposed marginalized deonising model [19]. Its features are to consider problem of link prediction as matrix denoising and learning mapping function. This mapping function is able to convert the matrix of observe links to unobserve links.

Cycle Formation Model. Huang et al. proposed a model on the based on the hypothesis that networks have the inclination towards close cycles in their link formation process [35]. This hypothesis is same as other methods like common neighbors, which take in account the number of cycles that would be shaped if the link existed. Moreover, this approach make an effort to detain longer cycles by increasing clustering coefficient to make it generalized. The generalized clustering coefficient can be computed as in Equation 48

$$C(k) = \frac{Noof CyclesLengthk}{Noof Pathslengthk} \tag{48}$$

Where, k representing the length cycles being analyzed. Furthermore, cycle formation model can be defined as $CF(k)$ with $k > 0$, distinguishes every formation mechanism (i.e., $g(1), g(2), ..., g(k)$) by single coefficient (i.e., $c_1, c_2, ..., c_k$). The anticipated clustering coefficient cam computed as in Equation 49

$$\ell(c_1, c_2, c_3, ..., c_k) = \sum_j |G_j| P_r(G_j) P_r(e_{1,k+1} \in E|G_j) \tag{49}$$

Where, Gj representing possible connected graphs along with i nodes. On the base of this given coefficient probability for link existence is estimated as in Equation 50

$$P_{u,v}(c_1, c_2, c_3, ..., c_k) = \left\{ \frac{c1 \prod_{i=2}^{k} c_i^{|paths^i_{u,v}|}}{c1 \prod_{i=2}^{k} c_i^{|paths^i_{u,v}|} + (1-c1) \prod_{i=2}^{k} (1-c_i)^{|paths^i_{u,v}|}} \right. \tag{50}$$

Local Co-Occurrence Model. Probabilistic models, discussed earlier, are restrictive to large networks, since their computational complexity is very high. Wang et al. proposed probabilistic model based on three types of local topological features: topology features, co-occurrence probability features and semantic features [95]. To obtain link probability among two nodes, tend to create local probabilistic model using MRF (Markov Random Field). For the prediction of link, three steps are performed: (1) A central neighborhood set $S_{x,y}$ is identified, (2) then t nodes that lie on the frequent path

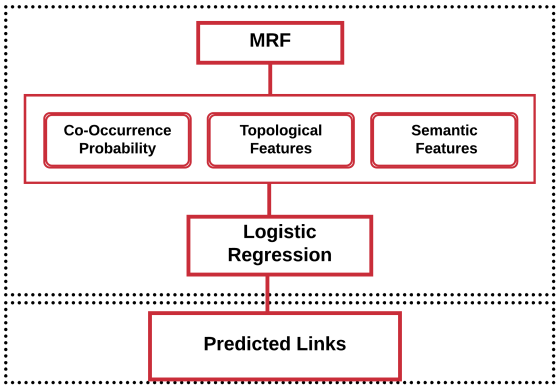

Figure 12. Local Co-Occurrence Model

are selected for training data to train the model, (3) find out the co-occurrence probability features. For the classification, logistic regression isused over three types of features discussed above. This local co-occurrence method is described in Figure 11.

4.4. Preprocessing

Preprocessing approaches are also called meta-approaches or high-level approaches, since they tend to work by combining with other methods. The foremost objective of these approaches is minimize the noise that exists in the networks as "false" or "weak" links. In addition, enhance the performance of approaches described earlier.

Low Rank Approximation. This method works on the network structure to simplify it by solving a well know problem namely low rank approximation. It uses adjacency matrix Am to make the network noiseless [48]. The optimization problem tends to reduce the cost function that estimates the fit among original and approximation matrix of minimized rank. This can be solved efficiently through SVD of original matrix as follows in Equation 51.

$$Am = A \sum B^T \tag{51}$$

Where, M^T and M denotes unitary matrices, while $\sum$ represents the diagonal matrix with positive elements. Most of the methods to estimate SVD are available. Most widely used approach focusses on the fact that eigenvalues of $(AmAm^T)$ as a square roots represents the singular values. In fact, considering decomposition expression, It ca be defined as follows in Equation 52.

$$AmAm^T = (A \sum B^T)(B \sum A^T) = A \sum^2 A^T \tag{52}$$

Here, columns of A corresponds to eigenvectors of Am, can be estimated via computing eigenvectors of

matrix. Demmel et al. proposed most widely used SVD algorithm, that provide high accuracy [26]. Consider SVD of Am, $\vec{Am}$ (low rank matrix) can be computed as follows in Equation 53.

$$\vec{Am} = A_{1:|B|,1:k} \sum_{1:k,1:k} B^{T}_{1:k,1:|B|} \tag{53}$$

Unseen Bigrams. Consecutive or adjacent two elements from the string are called bigram (also known as digram). The concept of bigram have been taken in various applications i.e., speech recognition, linguistics, or cryptography. Unseen bigrams are such kind of bigrams which are valid and not observed in a string collection. Let "a flower", "a room", "the flower", "the room" and "a bike" are observed bigrams, then we can noticed that "the bike" is a kind of unseen bigram. The strategy given by bigrams can be adjusted for link prediction to minimize the noise by replacing similar nodes [54]. In this way, similarity can be expressed as follows in Equation 54

$$UB(u,v) = |X^{\infty}_{u} \bigcap \Gamma(v)| \tag{54}$$

Where, X^{∞}_{u} is corresponds to the set of ∞ nodes similar to u.

Filtering. Also known as clustering [54], to avoid ambiguity, we called it as filtering. This is another kind of noise reduction method, which removes the weakest ties between nodes in order to improve the link prediction results. weakest ties are those kind of observed links which have no shared neighbors or small number of neighbors. It has another feature that it can also used for observed links to find their worth or strength. Therefore, filtering approach is used to assign a similarity score to every connected pair in order to remove γ weakest links and clean the network.

5. Performance Evaluation Measures

Evaluation measures used in the area of link prediction are embraced from other research areas i.e., classification, information retrieval [39]. These evaluation measures can be classified into two categorize: (1) threshold curves and (2) fixed threshold [58][101][24]. Fixed threshold measures have some imperfections that few of the estimates of sensible threshold accessible in score space. To overcome these flaws, threshold curve measures are an alternative.

5.1. Threshold Curves

1. *ROC:* ROC is abbreviation of receive operation characteristics. It narrates fragments of false positive rate versus true positive rate on different thresholds. Where, true positive rate is

$$TPR = \frac{TruePositive}{TruePositive + FalseNegative}$$

and false positive rate is as follows.

$$FPR = \frac{FalsePositive}{TrueNegative + FalsePositive}$$

Where, TPR estimates the portion of correctly predicted positive links. While, FPR estimates the misinterpreted negative links. Although these measure have made a big contribution to link prediction, in spite of this some researchers proves that both AUC and ROC can be illusive [101]. Furthermore, they have stated that, for the reason of acute class imbalance, PR curves and $PRAUC$ are better than ROC and AUC for the performance evaluation.

2. *PR:* PR is abbreviation of precision-recall curve. It represents precision along with recall at different thresholds [24]. It only considers the positive links for instead of negative links. Since, in periodic link prediction, it is required to predict removed links for that PR curve is not suitable [39].

3. *AUC:* AUC is abbreviation of area under the ROC. Here, high AUC represents the superior results of classification, while, low AUC corresponds to poor results.

5.2. Fixed Threshold

1. *Accuracy (Classification):* Accuracy, which is pure from classification, is the most widely and commonly used measure. It is estimated as follows in Equation 55.

$$Accuracy_{c} = \frac{TruePositive + TrueNegative}{P + N} \tag{55}$$

Where, N and P are the overall negative and positive links. Imbalanced data can makes accuracy deceptive. Usually, SNs are too large and existing link just add up to < 10 percent of all possible links, which means it is not meaningful measure [56].

2. *Accuracy (Graph):* Graph accuracy [84] is same as classification accuracy. However, graph accuracy takes into account the original graph and predicted graph. It is estimated as follows in Equation 56.

$$Accuracy_{g} = 1 - \frac{E(G_{\rho}) + E(G_{o}) - 2E(G_{\rho} \bigcap G_{o})}{Max(E(G_{\rho}), E(G_{o}))} \tag{56}$$

Where, E corresponds to the edges of graph, G_ρ corresponds to the predicted graph and G_o denotes the original graph.

3. *Recall:* In the context of link prediction, recall takes into account the number of positive predicted links and total positive links. It can expressed as follows in Equation 57

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (57)$$

4. *Precision:* Precision takes into account the correctly predicted links and total predicted links. It can be estimated as follows in Equation 58

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (58)$$

5. *F1-Measure:* It is also known as harmonic mean of recall and precision. It is defined as follows in Equation 59.

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \quad (59)$$

Where, R represents the recall and P denotes precision.

6. *NDCG:* It is another type of evaluation measure, which measures the accuracy. It uses the top k prediction scores. it is computed as follows in Equation 60.

$$NDCG_k = \frac{DCG_k}{IDCG} \quad (60)$$

Where, DCG_k is follows.

$$DCG_k = \sum_{i=1}^{k} \frac{2^{r_i} - 1}{log_2(i+1)}$$

And I DCG is as follows.

$$DCG_k = \sum_{i=1}^{|r|} \frac{2^{r_i} - 1}{log_2(i+1)}$$

7. *MR(Mean-Rank):* This measures is specifically used for missing link prediction. In order to evaluate following steps should be perform.

   - First, the dataset should separated in two set (i.e., training and testing) without any negative link,
   - For every test link, removed the node and replaced it by another node,
   - Computes the dissimilarity values of corrupted links,
   - After that sort the nodes in descending order,
   - In this way, the correct nodes are sorted according to their rank,
   - Finally, the mean is estimated for predicted ranks.

8. *Hit@n:* This measure is same as *mean rank*. The difference is, it works on the top n nodes. From the many studies, it is shown that researchers uses Hit@10 [13].

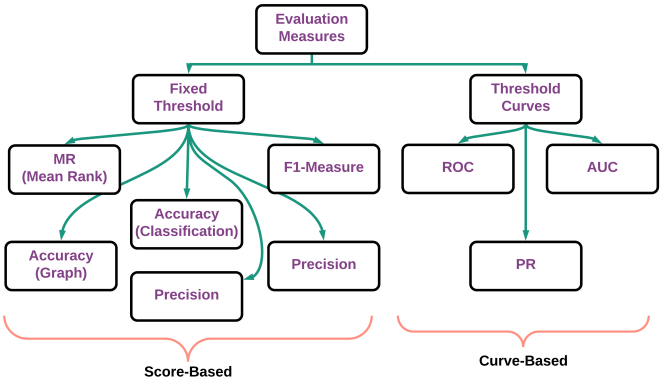

Figure 13. Taxonomy of Performance Evaluation Measures

6. Link Prediction Features

In this study, we have categorized the features used for link prediction as shown in Figure 14. Usually, there are two type of features are used by majority of link prediction approaches: (1) node-based features and (2) link-based features. Node-based features includes node's in-degree, out degree, level and distance. While, link-based features includes Link's level, type, label, weight and path.

Studying the link prediction features, it is experienced that major part of the research is done utilizing node-based features [104][66][103][44][9]. This research claims that features related with nodes plays significant role in link prediction, since individuals in SN have their own attributes (i.e., sex, age, gender, city, country and language) which are helpful in further relationship with new individuals. The previous statement is further justified by Samad et al. [84] by evaluating node features in order to predict the link between nodes. They have observed that language and country are key features that plays an important role in association between nodes in SN. Node-based features are further classified into two categorize: (1) Attributes (i.e., age, gender, school, interest, or location etc.), (2) topological (i.e., degree, neighbors, level, or distance

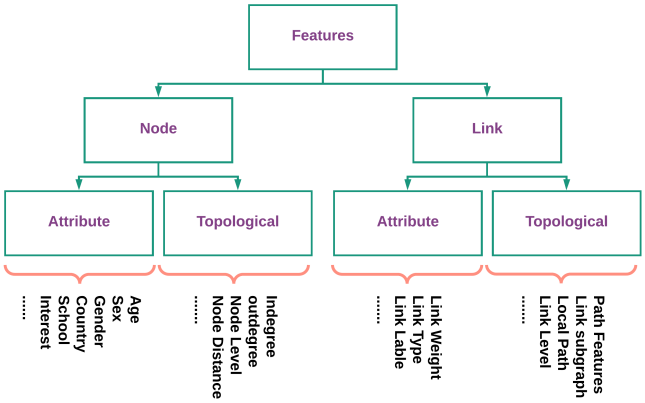

Figure 14. Taxonomy of Features used for link prediction

etc.). Node attributes are used in the situation, when we have a graph without edges. On the other hand, topological features are used, when we have a partial graph to infer the links. For example, Jahanbakhs et al [37] divided their work of link prediction into two parts: (1) In first phase, they predicted the links using graph without eges, (2) in second phase, they have considered a partial graph for link prediction.

Moreover, link-based features are also used by most of the researchers for link prediction [40][91][7][67]. Link weight considers as foremost features for link prediction in weighted networks. In addition, nodes distance is another important feature and estimated through the random walk and shortest paths methods. Few of the researchers have combined both node and link-based features in order to predict the future links [37]. Furthermore, link-based features are classified into two types: (1) attributes(i.e., weight, type, or label etc), (2) topological (i.e., level, path, or subgraph etc.).

7. Applications

On a large scale in different areas number of applications of link prediction technique had been found. Interaction of entities in a structured way from any sort of domain is capable to gain assistance from link prediction. Few compulsive or commonly used applications of link predictions are described shortly.

Link prediction methods facilitates in refinement for selection between similar users from a system using a collective approach, preceding an effective recommendation outcomes [88]. Users of such systems anticipate to have an effective and easy way to find user they are familiar to, as there are huge amount of users registered. To gain high degree of accuracy majority of social networks implements link prediction techniques which instinctively recommends similar users.

In the domain of biology, using protein-protein interaction network, link prediction methods are being used to detect potential interactions among the proteins particles [77]. To examine the interactions of protein by test-tube experiment is costly in terms of time and money, so with the help of results from initial experiments, target could be set computationally.

From collaboration prediction another use of link prediction is found in scientific c o-authorship networks. It is easy to access collaboration data, since collections from journal indexing sites are publicly available. For better knowledge to understand that how some research areas make progress, link prediction methods act as tool by predicting which authors or groups could associate in the upcoming time [74].

Record linkage (namely Entity Resolution) consists of searing identical records or references in a dataset. Traditionally, record linkage, focussed only on similarity of attributes among entries. Recently, few authors have considered structured information to improve record linkage by using link prediction methods along with similarity between the references [10].

Another widely used application of link prediction in social network is to explore structure of terrorist network (namely criminal network) in order to find out the way to fight a gainst t he c rimes [47]. F or instance, authors in [100], claimed that if we reinsert the small portion of links using link prediction methods, structure of few terrorist networks does not change. These outcomes support that the link prediction techniques can reveal important links in criminal networks, creating a way to investigate definite terrorist actions.

Ultimately, network can be useful to predict the likelihood of expansion across society. Marketing studies can also be improve with the help of network analysis. Some authors also reveal that in order to gain high marketing plans, link prediction can be used for vigorous marketing [81].

8. Conclusion

In fact, link prediction have gain more attention in recent decade as new algorithms and applications are emerging rapidly. This article presented comprehensive review on link prediction and expressed that various challenges and techniques are exist. In the context of link prediction, categories of techniques, problems, networks, evaluation measures and features are proposed. Different techniques of link prediction are explained i.e., similarity-based, learning-based, probabilistic models and preprocessing. Node-based and link-based features are also illustrated. Finally, link prediction applications are also discussed.

References

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

[2] Akcora, C.G., Carminati, B. and Ferrari, E. (2013) User similarities on social networks. *Social Network Analysis and Mining* **3**(3): 475–495.

[3] Al Hasan, M., Chaoji, V., Salem, S. and Zaki, M. (2006) Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.

[4] Al Hasan, M., Chaoji, V., Salem, S. and Zaki, M. (2006) Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.

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

[6] Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec, J. (2012) Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining* (ACM): 703–712.

[7] Ayoub, J., Lotfi, D., El Marraki, M. and Hammouch, A. (2020) Accurate link prediction method based on path length between a pair of unlinked nodes and their degree. *Social Network Analysis and Mining* **10**(1): 9.

[8] Barabâsi, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. and Vicsek, T. (2002) Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications* **311**(3-4): 590–614.

[9] Bayrak, A.E. and Polat, F. (2018) Mining individual features to enhance link prediction efficiency in location based social networks. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE): 920–925.

[10] Bhattacharya, I. and Getoor, L. (2007) Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **1**(1): 5.

[11] Bhattacharyya, P., Garg, A. and Wu, S.F. (2011) Analysis of user keyword similarity in online social networks. *Social network analysis and mining* **1**(3): 143–158.

[12] Bliss, C.A., Frank, M.R., Danforth, C.M. and Dodds, P.S. (2014) An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science* **5**(5): 750–764.

[13] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. and Yakhnenko, O. (2013) Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*: 2787–2795.

[14] Brouard, C., d'Alché Buc, F. and Szafranski, M. (2011) Semi-supervised penalized output kernel regression for link prediction.

[15] Cannistraci, C.V., Alanis-Lobato, G. and Ravasi, T. (2013) From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific reports* **3**: 1613.

[16] Cao, B., Liu, N.N. and Yang, Q. (2010) Transfer learning for collective link prediction in multiple heterogenous domains. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (Citeseer): 159–166.

[17] Chen, G. and Kotz, D. (2005) Structural analysis of social networks with wireless users. *Dartmouth College Computer Science Technical Report TR2005-549* .

[18] Chen, H.H., Gou, L., Zhang, X.L. and Giles, C.L. (2012) Discovering missing links in networks using vertex similarity measures. In *Proceedings of the 27th annual ACM symposium on applied computing* (ACM): 138–143.

[19] Chen, Z. and Zhang, W. (2014) A marginalized denoising method for link prediction in relational data. In *Proceedings of the 2014 SIAM International Conference on Data Mining* (SIAM): 298–306.

[20] Chiang, K.Y., Natarajan, N., Tewari, A. and Dhillon, I.S. (2011) Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (ACM): 1157–1162.

[21] Chua, H.N., Sung, W.K. and Wong, L. (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics* **22**(13): 1623–1630.

[22] Clauset, A., Moore, C. and Newman, M.E. (2008) Hierarchical structure and the prediction of missing links in networks. *Nature* **453**(7191): 98.

[23] Cukierski, W., Hamner, B. and Yang, B. (2011) Graph-based features for supervised link prediction. In *The 2011 International Joint Conference on Neural Networks* (IEEE): 1237–1244.

[24] Davis, J. and Goadrich, M. (2006) The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*: 233–240.

[25] De Sá, H.R. and Prudêncio, R.B. (2011) Supervised link prediction in weighted networks. In *The 2011 international joint conference on neural networks* (IEEE): 2281–2288.

[26] Demmel, J. and Kahan, W. (1990) Accurate singular values of bidiagonal matrices. *SIAM Journal on Scientific and Statistical Computing* **11**(5): 873–912.

[27] Dong, Y., Ke, Q., Wang, B. and Wu, B. (2011) Link prediction based on local information. In *2011 International Conference on Advances in Social Networks Analysis and Mining* (IEEE): 382–386.

[28] Fortunato, S. and Barthelemy, M. (2007) Resolution limit in community detection. *Proceedings of the national academy of sciences* **104**(1): 36–41.

[29] Fouss, F., Pirotte, A., Renders, J.M. and Saerens, M. (2007) Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on knowledge and data engineering* **19**(3): 355–369.

[30] Geyer, C.J. (1992) Practical markov chain monte carlo. *Statistical science* : 473–483.

[31] Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M. *et al.* (2010) A survey of statistical network models. *Foundations and Trends® in Machine Learning* **2**(2): 129–233.

[32] Gomaa, W.H. and Fahmy, A.A. (2013) A survey of text similarity approaches. *International Journal of Computer Applications* **68**(13): 13–18.

[33] Guimerà, R. and Sales-Pardo, M. (2009) Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences* **106**(52): 22073–22078.

[34] Hamers, L. *et al.* (1989) Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing and Management* **25**(3): 315–18.

[35] Huang, Z. (2010) Link prediction based on graph topology: The predictive value of generalized clustering coefficient. *Available at SSRN 1634014* .

[36] Jaccard, P. (1901) Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat* **37**: 547–579.

[37] Jahanbakhsh, K., Shoja, G.C. and King, V. (2010) Human contact prediction using contact graph inference. In *2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing* (IEEE): 813–818.

[38] Jeh, G. and Widom, J. (2002) Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM): 538–543.

[39] Junuthula, R.R., Xu, K.S. and Devabhaktuni, V.K. (2016) Evaluating link prediction accuracy in dynamic networks with added and removed edges. In *2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom)(BDCloud-SocialCom-SustainCom)* (IEEE): 377–384.

[40] Ka-Wei Lee, R. and Lim, E.P. (2017) Friendship maintenance and prediction in multiple social networks. *arXiv preprint arXiv:1703.00857* .

[41] Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M. and Tsuda, K. (2009) Link propagation: A fast semi-supervised learning algorithm for link prediction. In *Proceedings of the 2009 SIAM international conference on data mining* (SIAM): 1100–1111.

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

[43] Kiliç, M., Uyar, A. and Koseoglu, M.A. (2019) Co-authorship network analysis in the accounting discipline. *Australian Accounting Review* **29**(1): 235–251.

[44] Kim, D.W., Kwon, H., Lee, S.K., Jeong, W. and Yang, S.i. (2018) Social link prediction and feature analysis in mobile game. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)* (IEEE): 906–909.

[45] Koren, Y., Bell, R. and Volinsky, C. (2009) Matrix factorization techniques for recommender systems. *Computer* (8): 30–37.

[46] Kotsiantis, S., Kanellopoulos, D., Pintelas, P. *et al.* (2006) Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* **30**(1): 25–36.

[47] Krebs, V.E. (2002) Mapping networks of terrorist cells. *Connections* **24**(3): 43–52.

[48] Kunegis, J. and Lommatzsch, A. (2009) Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning* (ACM): 561–568.

[49] Lai, Y.Y., Neville, J. and Goldwasser, D. (2019) Transconv: Relationship embedding in social networks .

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

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

[52] Li, R.H., Yu, J.X. and Liu, J. (2011) Link prediction: the power of maximal entropy random walk. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (ACM): 1147–1156.

[53] Li, X. and Chen, H. (2013) Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. *Decision Support Systems* **54**(2): 880–890.

[54] Liben-Nowell, D. (2005) *An algorithmic approach to social networks*. Ph.D. thesis, Massachusetts Institute of Technology.

[55] Liben-Nowell, D. and Kleinberg, J. (2007) The link-prediction problem for social networks. *Journal of the American society for information science and technology* **58**(7): 1019–1031.

[56] Lichtenwalter, R.N. and Chawla, N.V. (2012) Vertex collocation profiles: subgraph counting for link analysis and prediction. In *Proceedings of the 21st international conference on World Wide Web* (ACM): 1019–1028.

[57] Lichtenwalter, R.N., Lussier, J.T. and Chawla, N.V. (2010) New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM): 243–252.

[58] Lichtnwalter, R. and Chawla, N.V. (2012) Link prediction: fair and effective evaluation. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (IEEE): 376–383.

[59] Lim, M., Abdullah, A., Jhanjhi, N. and Supramaniam, M. (2019) Hidden link prediction in criminal networks using the deep reinforcement learning technique. *Computers* **8**(1): 8.

[60] Liu, W. and Lü, L. (2010) Link prediction based on local random walk. *EPL (Europhysics Letters)* **89**(5): 58007.

[61] Lü, L., Jin, C.H. and Zhou, T. (2009) Similarity index based on local paths for link prediction of complex networks. *Physical Review E* **80**(4): 046122.

[62] Lü, L. and Zhou, T. (2009) Role of weak ties in link prediction of complex networks. In *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management* (ACM): 55–58.

[63] Lü, L. and Zhou, T. (2010) Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)* **89**(1): 18001.

[64] Lü, L. and Zhou, T. (2011) Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* **390**(6): 1150–1170.

[65] Lu, Z., Savas, B., Tang, W. and Dhillon, I.S. (2010) Supervised link prediction using multiple sources. In *2010 IEEE international conference on data mining* (IEEE): 923–928.

[66] Madahali, L., Najjar, L. and Hall, M. (2019) Exploratory factor analysis of graphical features for link prediction in social networks. In *International Workshop on Complex Networks* (Springer): 17–31.

[67] Malla, R. and Bhavani, S.D. (2019) Link weight prediction for directed wsn using features from network and its dual. In *International Conference on Pattern Recognition and Machine Intelligence* (Springer): 56–64.

[68] Mallek, S., Boukhris, I., Elouedi, Z. and Lefèvre, E. (2019) Evidential link prediction in social networks based on structural and social information. *Journal of computational science* **30**: 98–107.

[69] Mateen, M., Iqbal, M.A., Aleem, M. and Islam, M.A. (2017) A hybrid approach for spam detection for twitter. In *2017 14th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (IEEE): 466–471.

[70] Menon, A.K. and Elkan, C. (2011) Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases* (Springer): 437–452.

[71] Moody, J., McFarland, D. and Bender-deMoll, S. (2005) Dynamic network visualization. *American journal of sociology* **110**(4): 1206–1241.

[72] Newman, M.E. (2001) Clustering and preferential attachment in growing networks. *Physical review E* **64**(2): 025102.

[73] Papadimitriou, A., Symeonidis, P. and Manolopoulos, Y. (2012) Fast and accurate link prediction in social networking systems. *Journal of Systems and Software* **85**(9): 2119–2132.

[74] Pavlov, M. and Ichise, R. (2007) Finding experts by link prediction in co-authorship networks. *FEWS* **290**: 42–55.

[75] Pearson, K. (1905) The problem of the random walk. *Nature* **72**(1867): 342.

[76] Pujari, M. and Kanawati, R. (2012) Link prediction in complex networks by supervised rank aggregation. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence* (IEEE), **1**: 782–789.

[77] Qi, Y., Bar-Joseph, Z. and Klein-Seetharaman, J. (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics* **63**(3): 490–500.

[78] Ravasz, E. and Barabási, A.L. (2003) Hierarchical organization in complex networks. *Physical review E* **67**(2): 026112.

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

[80] Raymond, R. and Kashima, H. (2010) Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs. In *Joint european conference on machine learning and knowledge discovery in databases* (Springer): 131–147.

[81] Richardson, M. and Domingos, P. (2002) Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM): 61–70.

[82] Salton, G. and McGill, M.J. (1983) *Introduction to modern information retrieval* (mcgraw-hill).

[83] Samad, A., Islam, M.A., Iqbal, M.A. and Aleem, M. (2019) Centrality-based paper citation recommender system. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems* **6**(19).

[84] Samad, A., Islam, M.A., Iqbal, M.A., Aleem, M. and Arshed, J.U. (2017) Evaluation of features for social contact prediction. In *2017 13th International Conference on Emerging Technologies (ICET)* (IEEE): 1–6.

[85] Samad, A., Qadir, M. and Nawaz, I. (2019) Sam: a similarity measure for link prediction in social network. In *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)* (IEEE): 1–9.

[86] Sarkar, P., Chakrabarti, D. and Moore, A.W. (2011) Theoretical justification of popular link prediction heuristics. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[87] Scellato, S., Noulas, A. and Mascolo, C. (2011) Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM): 1046–1054.

[88] Schafer, J.B., Frankowski, D., Herlocker, J. and Sen, S. (2007) Collaborative filtering recommender systems. In *The adaptive web* (Springer), 291–324.

[89] Scripps, J., Tan, P.N., Chen, F. and Esfahanian, A.H. (2008) A matrix alignment approach for link prediction. In *2008 19th International Conference on Pattern Recognition* (IEEE): 1–4.

[90] Sørensen, T., Sørensen, T., Sørensen, T., SORENSEN, T., Sorensen, T., Sorensen, T. and Biering-Sørensen, T. (1948) 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 .

[91] Srinivas, V. and Mitra, P. (2016) *Link prediction in social networks: role of power law distribution* (Springer).

[92] Symeonidis, P., Iakovidou, N., Mantas, N. and Manolopoulos, Y. (2013) From biological to social networks: Link prediction based on multi-way spectral clustering. *Data & Knowledge Engineering* **87**: 226–242.

[93] Symeonidis, P. and Mantas, N. (2013) Spectral clustering for link prediction in social networks with positive and negative links. *Social Network Analysis and Mining* **3**(4): 1433–1447.

[94] Tong, H., Faloutsos, C. and Pan, J.Y. (2006) Fast random walk with restart and its applications. In *Sixth International Conference on Data Mining (ICDM'06)* (IEEE): 613–622.

[95] Wang, C., Satuluri, V. and Parthasarathy, S. (2007) Local probabilistic models for link prediction. In *Seventh IEEE international conference on data mining (ICDM 2007)* (IEEE): 322–331.

[96] Wang, Z., Liao, J., Cao, Q., Qi, H. and Wang, Z. (2014) Friendbook: a semantic-based friend recommendation system for social networks. *IEEE transactions on mobile computing* **14**(3): 538–551.

[97] Wohlfarth, T. and Ichise, R. (2008) Semantic and event-based approach for link prediction. In *International Conference on Practical Aspects of Knowledge Management* (Springer): 50–61.

[98] Wu, S., Sun, J. and Tang, J. (2013) Patent partner recommendation in enterprise social networks. In *Proceedings of the sixth ACM international conference on Web search and data mining* (ACM): 43–52.

[99] Xie, F., Chen, Z., Shang, J., Feng, X. and Li, J. (2015) A link prediction approach for item recommendation with complex number. *Knowledge-Based Systems* **81**: 148–158.

[100] Xu, J. and Chen, H. (2008) The topology of dark networks. *Communications of the ACM* **51**(10): 58–65.

[101] Yang, Y., Lichtenwalter, R.N. and Chawla, N.V. (2015) Evaluating link prediction methods. *Knowledge and Information Systems* **45**(3): 751–782.

[102] Yin, L., Zheng, H., Bian, T. and Deng, Y. (2017) An evidential link prediction method and link predictability based on shannon entropy. *Physica A: Statistical Mechanics and its Applications* **482**: 699–712.

[103] Yu, C., Zhao, X., An, L. and Lin, X. (2017) Similarity-based link prediction in social networks: A path and node combined approach. *Journal of Information Science* **43**(5): 683–695.

[104] Zhang, Y., Shen, S. and Wu, Z. (2018) Improve link prediction accuracy with node attribute similarities. In *International Conference on Computer Engineering and Networks* (Springer): 376–384.

[105] Zhao, K., Yen, J., Greer, G., Qiu, B., Mitra, P. and Portier, K. (2014) Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association* **21**(e2): e212–e218.

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

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

[108] Zhu, Y.X., Lü, L., Zhang, Q.M. and Zhou, T. (2012) Uncovering missing links with cold ends. *Physica A: Statistical Mechanics and its Applications* **391**(22): 5769–5778.