Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **2154:2167***

A Comprehensive Review of Similarity Based Link Prediction Methods for Complex Networks including Computational Biology

Nirmaljit Singh[1], Dr. Harmeet Singh[2]

[1]Research Scholar, CSA, [2]Asst. Professor, CSE

[1,2]Sant Baba Bhag Singh University, Jalandhar

**Corresponding author's E-mail: nirmaljit_singh@rediffmail.com*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 02 Dec 2023

Abstract

Information retrieval is one of the most challenging tasks for the mankind and to retrieve information interaction is required, which ultimately leads to the formation of networks. Universe is packed with different type of networks. Networks with complex topological properties are called complex network. Such types of networks are major tools for learning the connection between the organizations and finding the purpose of complex systems. The link prediction problems in complex networks facilitate predictions about the future organization of the network. Network is represented as a graph. The data in the network is signified by nodes, and the relations are represented by links. The future of non-connected links amid node pairs is predicted. This paper reviews the methods used to predict links for complex networks using similarity-based heuristics. Previous reviews, despite having a clear outline of the link prediction study, only described the prediction approaches. Research gaps between the similarity-based link prediction techniques, however, were not explicitly stated. With the help of chronological findings and a research gaps approach, this review seeks to give a continuing review and introduce the link prediction.

Keywords: *Complex Networks, Link Prediction, Prediction Approaches, Prediction Features, Similarity Based Link Prediction*

1. Introduction

Complex Networks and their need for realistic networks are increasing continually. Such types of networks contain enormous amount of nodes and very similar attributes. These networks use complex communication techniques to interact with one another. Such types of networks are major tools for learning the connection between the organizations and finding the purpose of complex systems. The complexity present in the system is directly proportional to two factors; the interactions between the parts of the system and the amount of information details required for defining the system [1]. Complex networks are all over the place inside our universe. They are a critical part of our living system. It is a forthcoming multidisciplinary territory of research that is scattered to many disciplines such as engineering, biology, sociology, physics, social interacting species, along with chemical systems, neural networks, and the Internet. Complex networks are characterized by vastly heterogeneous allotment of links, often spread through the presence of key properties such as bulkiness. From the research evidence in recent years, it has become obvious that complex networks are contender for becoming the new approach towards the research in the study of methods of communication between nodes. The network representation of these systems are labeled as complex networks as there are properties that appear as a result of the global topological structure of the organization that cannot be described with the help of arbitrary or standard graphs. Basically complex networks are described by quad properties; First, Complex Networks are disorganized i.e. spread everywhere, In other words, the number of their edges is relative to the number of nodes, Second, Nodes are present in clusters i.e. if two nodes have a common neighbor, there is a great chance of connection between them, Third, small world, implies clusters have significantly shorter path between the nodes, i.e. Most of the nodes in a component are at a very short distance from one another; Fourth, they are scale free networks that follow the power law [1]. Power law implies that a relative change in one attribute or object gives rise to a proportional relative change in another attribute or object. In complex networks there are a lot of characteristics that emerge as a consequence of the global organizational structure of the network. Predictions are needed almost in every walk ofexistence. Link prediction means predicting the future links between the nodes of a complex network. Facebook's "People You May Know"(PYMK), Linkedln's "People You May Want to Hire" and Google+'s "You May Know" are most popular examples of link prediction in complex social networks [2]. The universe elements are generally very difficult to be completely understood; therefore it's a better option to think of the universe as a network, by specifying the components as the nodes and the relationships as the links [3].

Link Prediction

Complex networks are all over the place or in case we model real-world situations in conditions of networks, we repeatedly find out novel things. The Internet is an illustration of a complex network, which can be defined as huge group of interconnected nodes. With the help of Internet, The world emerges to be becoming smaller, and people are fitting ever more linked. The use of high end telecommunication services has implied that people are not more strongly connected than ever before. The world is very rightly called a global village now due to these local and global connections. Being connected has very big effect on the broadcasting of information. While the first property appears in randomly generated networks, the second "emerges" as a consequence of a characteristic feature of many complex systems in which relations display a high level of transitivity [4]. The link prediction problems in complex networks facilitate predictions about the future organization of the network. Network is represented as a graph. The data in the network is signified by nodes, and the relations are represented by links. The future of non-connected links amid node pairs is predicted. Link prediction methods are planned and implemented by using relationships between nodes based upon their attributes. A node can be whatever thing: a human being, an institute, a computer, a biological cell, etc., Interconnected implies that two nodes may be linked, for example, two people recognize each other, two organizations trade goods or two computers contain a wire connecting the two of them.

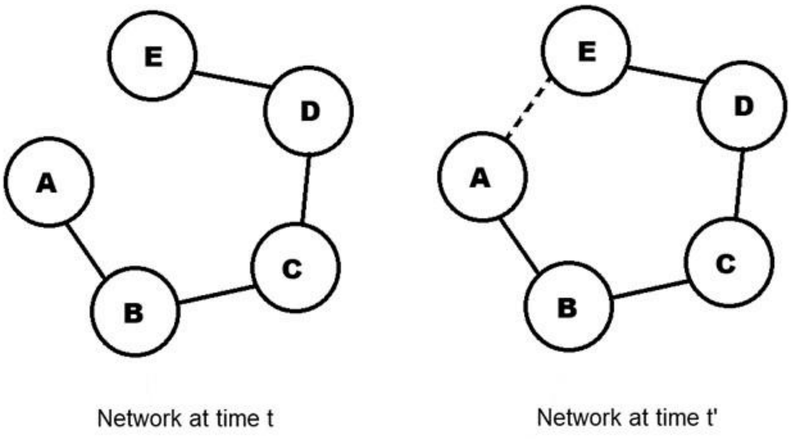

Fig 1: Link Prediction Methods predict the possibility of link between nodes A and E in future i.e. at time t'(t<t')

Calculations such as likelihood of link formation in the future can be done according to the common neighbor or shortest path or link state between the two nodes etc. The topologies of networks are extensively applied to learn the link-prediction problem in recent times. Link prediction is a phenomenon that predicts the future behaviors in complex network for instance to predict future collaborators of researchers in a co-authorship network, some other applications include recommender system, community detection etc [5]. The cost of performing experiments for developing new methods for link prediction is very high as reveling the hidden interactions among the nodes can be very expensive, therefore cost a a very big deciding factor while innovating new link prediction techniques. There are my special usages of link prediction techniques. For example, in natural networks, such as protein-protein communication, link accessible between the nodes indicate they have an association.

Link Prediction in Computational Biology

Link prediction is a computational method used to identify missing or potential links between entities in a complex network. In the context of computational biology, link prediction is applied to biological networks, which represent interactions between various entities such as genes, proteins, diseases, and drugs. By identifying potential links, link prediction can provide valuable insights into the underlying mechanisms of biological systems and aid in the development of new therapeutic strategies. Applications of Link Prediction in Computational Biology:

- Gene-Disease Association Prediction: Identifying potential associations between genes and diseases can help prioritize genes for further investigation and facilitate the development of new diagnostic and therapeutic approaches.
- Protein-Protein Interaction Prediction: Predicting interactions between proteins can aid in understanding cellular signaling pathways and identifying potential drug targets.
- Drug-Target Interaction Prediction: Predicting interactions between drugs and their targets can accelerate the drug discovery process and improve drug efficacy.
- Disease-Drug Association Prediction: Predicting associations between diseases and drugs can help in repurposing existing drugs for new indications and identifying potential side effects.

Further advancements in data collection, network modeling, and machine learning techniques are essential to overcome these challenges and fully harness the power of link prediction in advancing our understanding of biology and other complex networks.

Similarity based Link Prediction

This approach involves computing the similarity between pairs of nodes based on their structural properties, such as their degree centrality, clustering coefficient, or shortest-path distance. Nodes that are structurally similar are more likely to be connected. With many uses in areas like social network analysis, recommendation systems, and biological networks, similarity-based link prediction is a potent instrument for revealing hidden relationships in complex networks. Some Examples of similarity based link prediction methods are Common neighbors [15] refers to the set of nodes that two nodes have in common. Two nodes are more likely to be connected if they have numerous common neighbours, Jaccard coefficient [18][7] measures the ratio of the number of common neighbors to the total number of neighbors of two nodes. The likelihood of the two nodes being connected increases if this ratio is large, Adamic-Adar Index[19] measures the inverse logarithms of the degrees of the common neighbours of two nodes are added to determine the value of the Adamic-Adar index. Due to their greater predictive value for links, this measure gives more weight to common neighbours with fewer degrees and Preferential attachment [15] advocates that nodes with higher degrees are more likely to be connected in future, preferential attachment measures the degree distribution of the network.

Review Of Literature

Euler's [6] ‘Bridges of Konigsberg’ is generally regarded as the origin of graph theory and complex networks, in which the problem is to cross the seven bridges of the city only once. Jaccard [7] put forward a set of values to compare resemblance and variety of sample sets called Jaccard Coefficient, which was later used to predict links as proposed in Proceedings of The Twelfth International Conference on Information and Knowledge Management, 2003, using the formula of the ratio of common neighbors of the nodes a and b to total number of neighbors nodes of a and b. Rényi [8] revolutionized the graph theory by providing the branching process theory, in which random graphs are generated by starting at every node and adding adjoining neighbors. Strogatz [9] ‘Milgram’s Small World’, [10] [11] is one of the major property of complex networks which implies that majority of nodes have very short distance between them. Faloutsos et al. [12] proposed ‘Power Law-Internet’ that describes in brief the skewed distributions of graph attributes such as the node out degree and approximate neighborhood size. Brin & Page [13] proposed a method RPR (Rooted Page Rank) that ranks the WebPages depending upon quality specially used by Google to perform search queries upon them. Albert [14] proposed ‘Scale-free’ property that advocates that networks get bigger constantly and new nodes are attached substantially to sites that are well connected(Preferential Attachment). At this point of time, it was a generally accepted phenomenon about complex networks that they vigorously interact with each other. In physics elements are represented by sites, in computer science as nodes and vertex in graph theory. Link prediction is all about finding the similarity between two nodes. Computing the similarity score between two disconnected nodes is the most important phase or task in link prediction in complex networks. That score can determine whether there will be a link between those nodes in future or not.

Newman [15] proved that nodes having common neighbor are very likely to make a link between them. For example that probability of correlation between scientists in collaborations system is generally linear which strongly supported the previously established theories of complex networks regarding clustering and preferential attachment (Barabasi & Albert) [14] in growing networks. There are number of methods, metrics and approaches that are used for the problem of predicting future links between nodes, common neighbor index is most popular. This algorithm's fundamental principle is that two nodes are more likely to be linked if they have a large number of neighbours in common defined as follows:

$$J(i,j) = |N(i) \cap N(j)|$$

Where N(i) and N(j) are the sets of neighbors of nodes i and j, respectively, and |.| denotes the size of a set. As published in same Physical Review E., The preferential attachment index (PA) measures the likelihood of a new link forming between nodes i and j based on the degree of these nodes. The formula for the PA index between nodes i and j is: PA (i,j) = ki * kj / M, where ki and kj are the degrees of nodes i and j, respectively, and M is the total number of edges in the network. The PA index assumes that new links are more likely to form between nodes with high degrees, reflecting the idea that nodes that already have many connections are more attractive to new connections. Widom [16] proposed SimRank which is a recursion-based algorithm which determines whether a couple of nodes will have connections in future, depends on neighbours i.e. if two nodes are connected to alike nodes then this node couple are similar i.e. they have same type of characteristics. The SimRank score between two nodes i and j is defined as follows:

$$SimRank(i,j) = C / (|N(i)| * |N(j)|) * sum(SimRank(x,y))$$

where C is a damping factor that controls the rate of convergence of the algorithm, N(i) and N(j) are the sets of neighbors of nodes i and j, respectively, and x and y are nodes that are neighbors of both i and j. The SimRank score ranges from 0 to 1, with higher values indicating a higher degree of similarity between the two nodes. The algorithm converges when the SimRank scores between all pairs of nodes reach a steady state.

Ravasz [17] Proposed hub promoted index in which the nodes close to the hub have higher chance of connect as a link. The Hierarchical organization of modularity in metabolic networks was greatly by this effort of as it was very logically to think, pre traffic among nodes close to hub. The HPI score between two nodes i and j is defined as follows:

$$HPI(i,j) = sum(h\_k) / sqrt(k\_i * k\_j)$$

where h_k is the hub score of node k, k_i and k_j are the degrees of nodes i and j, respectively, and the sum is over all nodes k that are neighbors of both i and j.

The HPI score ranges from 0 to 1, with higher values indicating a higher likelihood that nodes i and j will be connected in the future. The algorithm converges when the scores of all pairs of nodes reach a steady state. Nowell and Kleinberg [18] in early 21st century performed very important work of comparing different link prediction methods, when they used Jaccard Coefficient as similarity Index. The objective of a link prediction algorithm is to guess most accurately the chance of the presence of missing links based on the currently existing links in a network. They concluded that a new link can be formed between two nodes by using straightforward topological features in complex network.The Jaccard coefficient can be used to gauge how similar the sets of neighbours of two nodes are in the setting of link prediction.The Jaccard correlation between nodes I and j has the following fo rmula:

$$J(i,j) = |N(i) \cap N(j)| / |N(i) \cup N(j)|$$

where N(i) and N(j) are the sets of neighbors of nodes i and j, respectively, and |.| denotes the size of a set Higher values of the Jaccard coefficient signify greater similarity between the sets of neighbours of the two nodes, with a number ranging from 0 to 1. Higher values in the context of link prediction imply that it is more probable that nodes I and j will link in the future.

Adamic-Adar [19] is closely related to Jaccard's coefficient measure; it filters the plain counting of common features by weighting rarer features more greatly. It is one of the most popular and efficient methods of link prediction. The formula for Adamic-Adar is as follows:

$$AA(i,j) = \sum (k \in N(i) \cap N(j)) 1/log(deg(k))$$

Where AA(i,j) is the Adamic-Adar similarity score between nodes i and j.

N(i) is the set of neighbors of node i.

N(j) is the set of neighbors of node j.

deg(k) is the degree of node k, i.e., the number of edges incident to node k.

This formula computes the similarity between nodes i and j by summing up the inverse logarithm of the degrees of their common neighbors. The intuition behind this formula is that nodes with more common neighbors that have lower degrees are more likely to be connected to each other, hence the higher similarity score. Taskar et al. [20] successfully forecasted and categorized a complete set of links in a link node graph. The Relational Markov networks work well with collective approach of using more than one link prediction methods simultaneously. They tried to learn patterns of cliques in web pages with relatively good result. This method is more domains oriented and powerful as it tried to use actual webpage attribute like text. Taskar's Relational Markov Networks (RMN) link prediction between nodes i and j:

P(y_ij=1|G,X,θ) = 1 / (1 + exp(-z))

where z is a linear combination of features that capture the similarities and differences between nodes i and j, as well as the patterns of connectivity in the graph.

Popescul & Ungar [21] developed a system that used structural logistic regression using statistical learning that extended inductive logic programming and formed citation prediction systems. Such type of system helped propositional learners to handle relational representation using some exceptional features in link prediction and bibliometrics. The link prediction using such technique is not easy to implement but it’s somehow more innovative than some of the other contemporary methods. Zhou et al. [22] come up with old and popular graph problems with different kind of approach in classifications of nodes in link prediction according to their rank. This concept is more of theoretical interest rather than providing empirical results. Popescul et al. [23] used clustering to make efficient link prediction of documents of authors in a network. The relationships and characteristics of different documents are represented by author name and subject topics. Such type of clustering provided great accuracy to prediction using relational models. Getoor et al. [24] advocated that Link Mining is a great approach to find new links. Almost all of today’s networks of interest can be described by a collection of linked objects. Data mining algorithms can be very successfully used to find the missing links by modeling similar type of patterns in database. Classification, ranking, group detection etc can be easily and efficiently performed using link mining techniques described in their paper. Leicht et al. [25] compared the H-Index with Sorenson Index [26]. The Sorenson Index can be further improved in its application to citation networks.

Page proposed Page Rank Citation Ranking [27] that was innovated by Liben-Nowell – Kleinberg [4] for link prediction problem. The proposed that the most popular algorithm used by most search engines i.e. PageRank Algorithm can be modified for link prediction and converted into rooted page rank. Page Rank Citation Algorithm [28] calculates the probability that node can be touched by a random walk in the graph is directly proportional to the specific rank given to the node. Ou et al. [29] proposed Resource Allocation Method RA Index for link prediction. It was inspired by the task of resource allocation of physical devices. It only used the common neighbors to compute the resource of target nodes obtained from source node. RA is better for high degree common neighbors than Adamic Adar Method but both works just as same for small degree neighbor networks. The Resource Allocation Method (RAM) Index formula developed by Ou et al. is given by:

RA Index = ∑ (Ci / Di)

Where Ci is the number of connections established by node i with other nodes in the network, and Di is the degree of node i, which is the sum of the number of connections established by node i with other nodes and the number of connections established by its neighboring nodes. Fouss et al. [30] proposed Hitting Time extracted from graph theory, defined as the expected number of steps required from a random walk from x to y will describe the Hitting time which is better if shorter.

Lu et al. [31] defined local path metric which is a very efficient metric that uses local path of length 2 and 3. The Local Path (LPC) metric formula developed by Lu is used to measure the similarity between two nodes in a network based on the number of shared common neighbors. It is given by:

LPC(i, j) = ∑ k∈N(i)∩N(j) [1/log(d_k+1)]

Where N(i) is the set of neighbors of node i in the network,

N(j) is the set of neighbors of node j in the network, and

d_k is the degree of node k, which is the number of connections established by node k with other nodes in the network.The LPC formula calculates the sum of the inverse logarithm of the degree of each common neighbor between nodes i and j. A higher value of LPC(i, j) indicates a higher similarity between nodes i and j, while a lower value indicates the opposite. Zhou et al. [32] proposed a Hub Depressed (HD) metric that uses higher degrees of nodes therefore its better and more efficient than Hitting Time Algorithm. Kat [33] proposed Katz Index which is also a very good global similarity Index that gives better results than most algorithms; it proves that more paths lead to more chances of forming a link. The Katz Index is a centrality measure used to identify the most important nodes in a network based on their proximity to other nodes. It is calculated using the following formula:

Katz(i) = $\alpha \sum$(j=1 to n) Aij*Katz(j) + β

Where Aij is the element in the adjacency matrix corresponding to the connection between nodes i and j,n is the total number of nodes in the network, α and β are constants that control the weight of the paths of different lengths. Lichtenwalter et al. [34] proposed PropFlow metric that is more localized than Rooted Page Rank that helps in extracting scores that is able enough to provide as an assessment of the probability of new links. Sarkar et al. [35] advocated that even if common neighbor is a great method of finding a link between two nodes but the weighted count of common neighbors is better method for accurate link prediction than other complex methods. They empirically proved that common neighbor bounds the alikeness of node pair; they further proved that weighted count outperforms unweighted count of common neighbors with justification. Lü- Zhou [36] proposed Salton Index (Cosine Similarity Index) that used cosine based prediction to compare documents in text mining using dot product of two vectors.

Lichtenwalter et al. [37] proposed Vertex Collocation Profile (VCP) which is basically described as a vector that defines common member graphs embedding two nodes. Although VCP does not represents similarity between nodes but it can be classified as path based metric because of its supervised learning based classification approach. Papadimitriou et al. [38] proposed Friend Link uses a method of link prediction that uses paths of greater length to analysis the nodes neighbors that have unique pathways. The unique pathways provide a great chance of getting associated by a link in future by traversing every path of a bounded distance. Chen et al. [39] proposed Relation Strength Similarity (RSS) that describe the comparative degree of resemblance between neighboring nodes. It is an asymmetric metric that is generally used on a weighted network. They can be used by assigning same weights to edges; if the importance of weights is not accurately known in advance. The primary job of link prediction algorithm is to predict potential links amongst entities in the graph, with reference to its present state.

Zhu et al. [40] proposed Parameter-Dependent (PD) that is a generic type of metric containing free parameter, When its value is zero it is converted in common neighbor when its 50% it is converted to Salton metric. Chen et al. [41] proposed a metric based upon the cosine similarity time metric for calculating sameness of two vectors. Martínez et al. [42] proposed commute time that basically finds and counts the expected number of hops from node 1 to 2 and 2 to 1. Srilatha et al. [43] proposed Similarity Index based Link Prediction Algorithms in Social Networks in which KA is used which counts all paths between two nodes. Key here is bigger path lengths have less weight while calculating the final similarities. Zeng [44] stated preferential attachment method implies that if a node is connected to many nodes then there is a very big chance for that node to develop new links with other node, it enhanced the concept that was proposed by Barabási–Albert [14]. The paper proposed that common neighbors plus preferential attachment index is a superior to approximation to the probability of the being of a link between two nodes on the basis of local data of the closest neighbors.

Jia –Qu [45] advocated that the metric proposed by Salton and McGill that can be used as a cosine metric that measures the similarity between two nodes x and y, the bigger value of cosine, the higher chance is there of link between nodes. Mohan et al. [46] proposed a Parallel Similarity Measure formula used for link prediction in large-scale networks. It is based on the Jaccard Similarity Coefficient and is given by:

PS(i, j) = $(|N(i) \cap N(j)| + \lambda) / (|N(i) \cup N(j)| + \lambda)$

Where N(i) is the set of neighbors of node i in the network,

N(j) is the set of neighbors of node j in the network,

$|N(i) \cap N(j)|$ is the number of common neighbors between nodes i and j, and

$|N(i) \cup N(j)|$ is the total number of neighbors of nodes i and j.

The λ is a smoothing factor to handle cases where $|N(i) \cup N(j)| = 0$. The Parallel Similarity Measure formula is designed for parallel processing on distributed computing environments and can handle large-scale networks efficiently. A higher value of PS(i, j) indicates a higher similarity between nodes i and j, and thus a higher likelihood of a connection forming between them. Nandi et al. [47] proposed a new link prediction method that incorporates the advantages of two methods, firstly preferential attachment is implemented and then Adamic-Adar approach is used to rank the top similarity scores. The Link Prediction technique proposed by Nandi et al. is given by:

$LP(i, j) = \sum k \in (N(i) \cap N(j)) (w(i, k) * w(j, k))$

where N(i) is the set of neighbors of node i in the network, N(j) is the set of neighbors of node j in the network, and k is a common neighbor of nodes i and j. The weights w(i, k) and w(j, k) are calculated using a formula based on the degree of node k. The Link Prediction formula calculates the sum of the product of the weights assigned to the common neighbors of nodes i and j. A higher value of LP(i, j) indicates a higher likelihood of a connection forming between nodes i and j. Zhou et al. [48] proposed a technique to enhance the strength of similarity-based link prediction by providing a constrained group reliable queries which precisely measure the presence and credibility of queried links. Such type of method works great to find links that are hidden for malicious purpose by advocating the fact that deleting links purposely will change the state of the effected nodes.

Iftikhar et al. [49] proposed CCPA - Common Neighbor and Centrality based Parameterized Algorithm that uses Common neighbor and closeness centrality based prediction for finding missing edges (links). The algorithm uses three parameters, α, β, and γ, to weight the contribution of different features in the prediction. The formula for CCPA is given as:

$CCPA(i,j) = \alpha * CN(i,j) + \beta * SC(i,j) + \gamma * CC(i,j)$

where CN(i,j) is the common neighbor score between nodes i and j, SC(i,j) is the similarity measure based on the shortest path between i and j, and CC(i,j) is the centrality score of nodes i and j. The parameters α, β, and γ are used to adjust the relative weight of each feature in the prediction.

Wang et al. [50] developed an improved spatial graph convolution network (SGCN) for link prediction in heterogeneous information networks (HINs) using local community discovery and an optimizable kernel layer to measure pairwise vertex embeddings. The method was tested on four real-world HINs and was found to be effective in handling the complexity of link predictions in these networks. Longjie et al. [51] proposed a hybrid similarity model that combines Grey Relation Analysis and State of Art Similarity based LP Models. The proposed model was tested on three real-world networks, and the results showed that the hybrid model outperformed other traditional link prediction models, such as common neighbors, resource allocation, and Jaccard coefficient, in terms of both accuracy and efficiency. Szyman et al. [52] hypothesized that similar vertices belonging to the same community would be more likely to be connected than those that were not similar. Comprehensive R Archive Network (CRAN) guide is used for link prediction including techniques like similarity-based methods, graph-based methods, and matrix factorization methods. The research presented a study on using link prediction as a tool for community detection in business social networks based on employee email exchanges. The primary objective of the computational experiment was to predict potential new connections between network vertices by assigning similarity scores to each pair of unconnected vertices and selecting those with the highest scores. The secondary goal was to determine if resolving the link prediction problem could aid in the social network's community detection process. The study hypothesized that similar vertices belonging to the same community would be more likely to be connected than those that were not similar.

Zhao et al. [53] proposed model, called HGE, uses a hypergraph neural network to learn the embedding representation of hyperedges and nodes. The model also employs a joint loss function that considers both hyperedge and node-level proximity. The experiments conducted on several real-world datasets demonstrate that HGE outperforms state-of-the-art methods for link prediction in hypergraphs

Blocker et al [54] proposed a novel method for link prediction called MapSim that introduces similarity-based link prediction from modular compression of network flows. MapSim utilizes a hierarchical modular structure to capture node similarities, outperforming traditional embedding-based methods in various network types. The algorithm employs modular compression to represent a network's hierarchical structure. This involves compressing the flow of information between nodes by identifying redundant paths and assigning them to modules. Unlike embedding-based methods that project nodes into a metric space, MapSim positions nodes in a non-metric latent space based on their positions within the coding tree. This allows for asymmetric similarities, capturing directed interactions. Node similarities are calculated based on the compressibility of transitions between nodes. Nodes with highly compressible transitions are considered more similar, indicating stronger connections. The Formulas used in Map sim are as follows:

Compressibility Formula: $C_{uv} = \log \left(\frac{L_u + L_v}{L_{uv}} \right)$

This formula calculates the compressibility of a transition between two nodes u and v. Compressibility measures the redundancy in the flow of information between nodes. A higher compressibility indicates a stronger connection between the nodes.

Similarity Formula: $S_{uv} = \frac{1}{K} \sum_{i=1}^{K} C_{u_iv_i}$

This formula calculates the similarity between two nodes u and v. Similarity measures the degree of connection between nodes. Higher similarity indicates a more likely connection between the nodes. The compressibility formula measures the redundancy of information flow between nodes, while the similarity formula averages compressibility across multiple paths to capture the overall strength of connection. These formulas form the basis of MapSim's link prediction capabilities. MapSim consistently outperforms traditional embedding-based methods in link prediction tasks across various network types. It effectively handles directed networks, capturing asymmetric relationships between nodes. MapSim's computational complexity is lower than embedding-based methods, making it more scalable for large networks.

2. Results and Discussion

Modern research emphasize on numerous other higher level Methods for link prediction such as clustering and low-rank approximation, which can be joined with the above mentioned link prediction techniques to shape up a hybrid approach. Major methods using similarity score heuristics are summarized in tabular form with key characteristics as follows:

Table 2: Similarity based Link Prediction Methods

Year	Link Prediction Method	Proposed By	Publisher	Findings	Research Gap
2001	Preferential Attachment Index[15]	Newman	Physical Review E	Better for quantifying the functional significance of nodes on the bases of network dynamics. Nodes with high degree are more likely to get new links	Only degree of the nodes is used to rank link between the nodes, need to explore new methods that use different attributes for link prediction.
2001	Common Neighbor Index[15]	Newman	Physical Review E	CN consumes very little time and does well as compared to other local indices.	Basic method, need to explore new methods for Similarity or proximity measure.
2002	Hub Promoted Index[17]	Ravasz et al.	Science	Links adjacent to hubs are likely to obtain a higher similarity score. Used for analyzing metabolic	Links away from the nodes are discarded. Need to explore more efficient methods

				networks	
2003	Adamic Adar Index[19]	Adamic A &Adar E	Networks	Nodes with rarer attributes are more likely to get new links. Yields very good results as compared to some other popular methods	Research gap for finding better method using degree of nodes as non-prominent attribute links are not always influential in facilitating the formation of future interactions. Need to explore relative mean between nodes based similarity measure.
2003	Katz Index [18][33]	Liben-Nowell and Kleinberg	Proceedings of The Twelfth International Conference on Information and Knowledge Management	Good performance path-based link prediction method. It proposed that greater the number of paths between two nodes, greater the chances of forming a link.	Too complex as information about all paths is required and algorithm did not distinguish the contribution of the paths with the same path length.
2003	Jaccard Index[18][7]	Liben-Nowell and Kleinberg	Proceedings of The Twelfth International Conference on Information and Knowledge Management	JC is used to measure similarity over the multiplicity. Solves the problem of ranking if there are multiple common neighbors.	Need to explore and understand better measures of “proximity” that lead to accurate predictions, as It treat all products’ common neighbours equally, no matter what the products’ position are in product category hierarchy.
2006	Leicht-Holme-Newman Index[25]	Leichtet al.	Physical Review E	Nodes are similar if their immediate neighbors in the network are themselves similar. It uses common neighbor approach relative to their geometric mean.	Need to explore to generalize the methods using path based metrics.
2006	Sorenson Index[26][25]	Sørensen T	Biologiske Skrifter	Sample Similarity based prediction used for analysis of plant sociology. It uses common neighbor	Need to explore other mean based approaches

				approach relative to their arithmetic mean.	
2009	Hub Depressed Index[32]	Zhou et al.	The European Physical Journal B	Links adjacent to hub given a lower score. Proposed to measure and index nodes with the opposite effect of hubs	Links adjacent to hub given sometimes have more chances of getting connected by a link.
2009	Local path Index [31]	Lu et al.	Physical Review E	Shortest Path based prediction. Uses the path of length 2 or3.	Need to explore methods that include factor information on paths between a node and its neighbors.
2010	PropFlow Index[34]	Lichtenwalter et al.	Proceedings Of The 16th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining	Random walk-based prediction. Extract the connectivity properties of pairs of nodes in social networks using random walk	Need to explore more efficient global method for link prediction.
2011	Salton Index(Cosine Similarity Index)[36]	Lü & Zhou	Physica A: Statistical Mechanics and Its Applications	Cosine based prediction, Compares documents in text mining using dot product of two vectors	Works best with text mining base prediction only
2012	Friend Link Index [38]	Papadimitriou et.al.	Journal of Systems and Software	Uses unique pathways that provides greater chance of getting associated by a link in future	Unique pathway doesn't always provide best prediction.
2016	Commute time similarity Index[42]	Martínez et al.	ACM Computing Surveys (CSUR)	Similarity time metric based prediction that finds and counts the expected number of hops from node 1 to 2 and 2 to 1	Need to explore hybrid similarity indices for better efficiency

2017	Parallel Similarity Measure Index[46]	Mohan et al.	Journal of Parallel and Distributed Computing	Hybrid common neighbor based Prediction that uses Adamic-Adar metric, community information and degree centrality of common neighbors.	Involves lot of message generation and transfer which can affect the system performance.
2018	LinkGyp Index [47]	Nandi - Das	International Journal of Advanced Computer Science and Applications(IJ ACSA)	Hybrid common neighbor based Prediction in which Adamic-Adar metric and preferential attachments are used to rank similarity scores.	Need to explore new methods for links that are hidden for malicious purpose.
2019	Strong Stackelberg Equilibrium[48]	Zhou, K., Michalak, T. et al.	IEEE Xplore	Enhances the strength of similarity-based link prediction by providing a constrained group reliable queries which precisely measure the presence and credibility of queried links.	Need to explore link prediction that uses other features and topological structure including signed networks.
2020	CCPA - Common Neighbor and Centrality based Parameterized Algorithm Index[49]	Iftikhar et al.	Nature Research, Scientific Reports	Hybrid of common neighbor and closeness centrality based prediction	A thorough analysis is needed to explore better methods for multi - dimensional data sets to obtain a general inference based algorithm.
2021	GRA: Grey Relation Analysis and Tradition LP Methods	Longjie et al	Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology	Hybrid similarity model that combines GRA, node similarity and link similarity measures to capture both local and global structural information.	Lack of explanation of how the proposed hybrid model is better than the state-of-the-art methods
2022	Comprehensive R Archive Network (CRAN) based Index[52]	Szyman et al.	Proceedings of the 26th International Conference on Knowledge-	Hybrid of similarity-based methods, graph-based methods, and matrix	Lack of investigation especially dedicated to assessing the

			Based and Intelligent Information & Engineering Systems (KES 2022)	factorization methods implemented in R	effectiveness of link prediction algorithms in the setting of dynamic social networks.
2023	HGE: Hyper graph Neural Network [53]	Zhao et al.	*Appl. Sci.* 2023	Use a similarity measure based on the cosine similarity of the node and hyper edge embedding	Only evaluated their approach on two real-world datasets, which may not be representative of all hyper graph structures. Lack of comparison to state-of-the-art hypergraph-based link prediction methods.
2023	MapSim	Blöcker .et al,	Proceedings of the Machine Learning Research Press	A hierarchical modular structure to capture node similarities, employs modular compression to represent a network's hierarchical structure. This involves compressing the flow of information between nodes by identifying redundant paths and assigning them to modules.	Lack of investigation especially dedicated to assessing the effectiveness of link prediction algorithm by using a hybrid approach i.e. combining Mapsim with other major similarity based algorithms and study results.

4. Conclusion

The link prediction problems in complex networks facilitate predictions about the future organization of the network. The future of non-connected links amid node pairs is predicted. Link prediction methods are planned and implemented by using relationships between nodes based upon their attributes. In conclusion, by offering comparisons based on prediction approaches and prediction features, this review's primary goal is to give a continued link prediction review and close any gaps left by earlier literature review studies. The tabularization of link prediction provided in this paper can help researchers come up with fresh concepts and developments that advance link prediction research. Link prediction holds immense promise in the field of computational biology, offering the ability to unveil hidden connections between biological entities within complex networks. Similarity-based link prediction methods are versatile tools that can be applied to a wide range of fields to identify potential connections between entities in complex networks including social media friend recommendations, community detection, influence analysis and other recommendation systems.

References:

1. Estrada, E., 2014, 'Introduction To Complex Networks: Structure And Dynamics', Book Chapter,www.Estradalab.Org › Uploads › 2015/10 › Bookchapter_11
2. Gupta, A., 'Analysis and Improvement of Link Prediction Techniques in Online Social Networks', Doctoral Dissertation, Jaypee Institute Of Information Technology (Declared Deemed To Be University), Noida, India, July 2020, https://shodhganga.inflibnet.ac.in /handle/10603/295019

3. Bellotti, R., 2018,‘Complex network-based quantitative methods applied to the study of neurodegenerative disease’, UniversitàDegliStudi Di Bari, http://phdphysics.cloud.ba.infn.it/wp-content/uploads/2018/03/la_rocca_tesi-compressed.pdf
4. Liben-Nowell, D., Kleinberg, J., ‘The Link Prediction Problem for Social Networks’, Journal of The American Society for Information Science and Technology, 58(7):1019–1031, (May 2007).
5. Daud. N., Hamid,S.et al., 2020, ‘Applications of link prediction in social network: A review’, Journal of Network and Computer Applications (IF 5.570) Pub Date : 2020-05-21 , DOI: 10.1016/j.jnca.2020.102716
6. Euler, L., 1741, ‘Solutioproblematisadgeometriamsituspertinentis’, Commentarii academiaescientiarum Petropolitanae, Volume 8, pp. 128-140.
7. Jaccard, P., 1901. “Etude Comparative De La Distribution FloraleDansUne Portion Des Alpes Et. Des Jura”, Bulletin Del La Soci´Et.´EVaudoise Des Sciences Naturelles, Vol. 37, Pp. 547–579, 1901 (In French)
8. Erdős, P. Rényi, A., 1960,‘On The Evolution Of Random Graphs’, Magyar TudományosAkadémiaMatematikaiKutatóIntézeténekKőzleményei [Publications Of The Mathematical Institute Of The Hungarian Academy Of Sciences]. 5: 17–61
9. Watts, D, Strogatz, S., 1998, Small World, Nature, 393:440-442
10. Milgram S, 1967, ‘The Small World Problem’, Psychology Today, Vol. 2, Pp. 60–67, 1967.
11. Travers J, Milgram S, 1969, ‘An Experimental Study of the Small World Problem’, Sociometry, Vol. 32, No. 4, Pp. 425–443, 1969.
12. Faloutsos, M, Faloutsos C. et al., 1999, ‘On Power-Law Relationships Of The Internet Topology’, ACM SIGCOMM Computer Communication Review, August 1999, Https://Doi.Org/10.1145/316194.316229.
13. Brin, S.& Page,L., 1998, ‘The Anatomy of a Large-Scale Hypertextual Web Search Engine’ Computer Networks, vol. 30, pp. 107-117
14. Barabási A, Albert A,1999, ‘Emergence of scaling in random networks’, AAAS-American Association For The Advancement Of Science, Science , Oct 1999: Vol. 286, Issue 5439, Pp. 509-512, DOI: 10.1126/Science.286.5439.509
15. Newman, M., 2001, ‘Clustering And Preferential Attachment In Growing Networks’ Physical Review E, 64.
16. Jeh, G. & Widom, J., 2002, ‘Simrank’,Proceedings of The Eighth ACM SIGKDD International Conference on Knowledge Discovery And Data Mining - KDD '02,ACM Press.
17. Ravasz, E., Somera, L. et al., 2002, ‘Hierarchical organization of modularity in metabolic networks,’ Science, vol. 297, no. 5586, pp. 1551–1555, 2002.
18. Liben-Nowell, Kleinberg, 2003, 'The Link Prediction Problem For Social Networks', Proceedings Of The Twelfth International Conference On Information And Knowledge Management, 3-8 November 2003, Pp. 556-559
19. Adamic & Adar , 2003, “Friend and Neighbors On The Web” Social Networks, Networks, 2003, 25: 211{230
20. Taskar, B., Wonng M. et al. 2003, 'Label And Link Prediction In Relational Data', Proceedings Of The IJCAI Workshop On Learning Statistical Models From Relational Data, 11 August 2003.
21. Popescul, A, Ungar, L, 2003, 'Structural Logistic Regression For Link Analysis', Proceedings Of KDD Workshop On Multi-Relational Data Mining, 2003, http://www.cis.upenn.edu/~popescul/publications/popescul03mrdm .pdf.
22. Zhou, D., Scholkopf, B, 2004, 'A Regularization Framework for Learning from Graph Data', Proceedings of Workshop on Statistical Relational Learning’, International Conference on Machine Learning, Banff, 12 June 2006, http://www.cs.umd.edu/projects/srl2004/papers/zhou.pdf.
23. Popescul, A., Ungar, L., 2004, 'Cluster-Based Concept Invention For Statistical Relational Learning', Proceedings Of Conference Knowledge Discovery And Data Mining (KDD-2004), 22-25 August 2004, Viewed 12 June 2006
24. Getoor, L.& Diehl, C., 2005, ‘Link Mining: A Survey’, ACM SIGKDD Explorations Newsletter, Vol. 7, Issue 2, Pp 3-12.
25. Leicht, E., Holme, P. et al., 2006, ‘Vertex Similarity In Networks’, Physical Review E, 2006, 73: 026120
26. Sørensen T., 1948, ‘A Method of Establishing Groups of Equal Amplitude in Plants Ociology Based on Similarity of Species and Its Application to Analyses of The Vegetation on Danish Commons’, vol. 5 of BiologiskeSkrifter, E. Munksgaard, 1948.
27. Berkhout J, 2016, “Google's Pagerank Algorithm For Ranking Nodes In General Networks”, IEEE, 2016 13th International Workshop On Discrete Event Systems (WODES)
28. Page, L., 1999, ‘The PageRank Citation Ranking: Bringing Order to the Web’, WWW 1999
29. Ou, Q., Jin, Y. et al.,2007, ‘Power-Law Strength-Degree Correlation From Resource-Allocation Dynamics On Weighted Networks’ Phys. Rev. E, 75:021102, 2007. Doi: 10.1103/Physreve.75.021102. http://Dx.Doi.Org/10.1103/Physreve.75.021102.
30. Fouss, F., Pirotte A. et al., 2007, ‘Random-Walk Computation of Similarities between Nodes of A Graph with Application to Collaborative Recommendation’, IEEE Transactions on Knowledge and Data Engineering, 19, 355-369.

31. Lu, L., Jin, C. et al., 2009 ‘Similarity Index Based On Local Paths for Link Prediction of Complex Networks’, Physical Review E, 80: 046122
32. Zhou, T., Lu, L.et al., 2009, ‘Predicting Missing Links via Local Information’, The European Physical Journal B, 2009, 71: 623{630
33. Katz L, 1953, ‘New Status Index Derived From Sociometric Analysis’, Psychometrika, 1953, 18: 39{43138{143
34. Lichtenwalter, R., Lussier, J. et al., ‘New Perspectives and Methods in Link Prediction’, Proceedings ofthe 16th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, 2010. 243{252
35. Sarkar, P. Chakrabarti,D. et al, 2011, ‘Theoretical Justification of Popular Link Prediction Heuristics’, Proceedings of the 22nd International Joint Conference on ArtiCial Intelligence, Barcelona, Spain, 2011. 2722-27276
36. Lü, L. and Zhou T, 2011, ‘Link prediction in complex networks: a survey,’ Physica A: Statistical Mechanics and Its Applications, vol. 390, no. 6, pp. 1150–1170, 2011.
37. Lichtenwalter R, Chawla N, 2012, ‘Vertex Collocation Pro_Les: Subgraph Counting For Link Analysis And Prediction’, Proceedings Of The 21st World Wide Web Conference (WWW'12), Lyon, France, 2012. 1019{1028
38. Papadimitriou, A.,Symeonidis, P. et al., 2012, ‘Fast And Accurate Link Prediction In Social Networking Systems’, Journal of Systems and Software, 2012, 85: 2119{2132
39. Chen, H., Liang, G.et al.,2012”, ‘Discovering Missing Links In Networks Using Vertex Similarity Measures’, Proceedings Of The Twenty-Seventh Annual ACM Symposium On Applied Computing (SAC'12), Trento, Italy, 2012.
40. Zhu, Y., Lu, L.et al.,2012, ‘Uncovering Missing Links With Cold Ends’, Physica A, 2012, 391: 5769{5778.
41. Chen, X., Xia, M. et al., 2016 ‘Trend Prediction of Internet Public Opinion Based on Collaborative Filtering’, Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 12th International Conference On, 2016. IEEE, 583-588
42. Martínez, V., Berzal, F. et al., 2016, ‘A Survey of Link Prediction In Complex Networks’, ACM Computing Surveys (CSUR), 49, 69.
43. Srilatha, P. & Manjula, R., 2016, ‘Similarity Index Based Link Prediction Algorithms In Social Networks: A Survey’, Journal of Telecommunications and Information Technology, 2016 | nr 2 | 87--94
44. Zeng, S., 2016, ‘Link Prediction Based On Local Information Considering Preferential Attachment’, Physica A: Statistical Mechanics And Its Applications, 443, 537-542.
45. Jia, Y. & Qu, L., 2016, ‘Improve The Performance Of Link Prediction Methods In Citation Network By Using H-Index’, Cyber-Enabled Distributed Computing And Knowledge Discovery (Cyberc), 2016 International Conference On, 2016. IEEE, 220-223. (ANT 2016), Published By Elsevier B
46. Mohan, A.,Venkatesan, R. et al., 2017, ‘A scalable method for link prediction in large real world networks’, Journal of Parallel and Distributed Computing, http://dx.doi.org/10.1016/j.jpdc.2017.05.009
47. Nandi ,G., & Das, A., 2018, ‘An Efficient Link Prediction Technique in Social Networks based on Node Neighborhoods’, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018
48. Zhou, K., Michalak,T. et al., 2019, ‘Adversarial Robustness of Similarity-Based Link Prediction’, IEEE International Conference on Data Mining (ICDM), Nov,2019,added to IEEE Xplore: 30 January 2020.
49. Iftikhar, A., Akhtar, A. et al., 2020, ‘Missing Link Prediction using Common Neighbor and Centrality based Parameterized Algorithm’, Nature Research, Scientific Reports, Sci Rep 10, 364 (2020). https://doi.org/10.1038/s41598-019-57304-y
50. Wang, X., Chai, Y. ,et al., 2021, ‘Link prediction in heterogeneous information networks: An improved deep graph convolution approach’, Elsevier, Decision Support Systems, Volume 141, February 2021, 113448, https://doi.org/10.1016/j.dss.2020.113448
51. Longjie et al. ‘Towards Effective Link Prediction: A Hybrid Similarity Model’, Journal of Intelligent & Fuzzy Systems, vol. 40, no. 3, pp. 4013-4026, 2021
52. Szyman, P. & Barbucha, D., 2022, ‘Link prediction in organizational social network based on e-mail communication’, Proceedings of the 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)
53. Zhao, Z et al, Link Prediction with Hypergraphs via Network Embedding. Appl. Sci. 2023, 13, 523. https://doi.org/10.3390/app13010523
54. Blöcker, C. et al, Similarity-based link prediction from modular compression of network flows, in the Proceedings of the Machine Learning Research Press (Vol. 198, pp. 1-18), 2023. https://proceedings.mlr.press/v198/blocker22a/blocker22a.pdf