The Intelligent System Review

Volume 1, Issue 1, 2024

OPTIMAL LINK PREDICTION IN COMPLEX HUMAN NETWORKS VIA STRUCTURAL PATTERN ANALYSIS

Zulfiqar Ali

City University of Science and Information Technology, Peshawar, Pakistan

Keywords
Link Prediction, Social Networking, Complex Network

Article History
Received: 29 October, 2024
Accepted: 12 December, 2024
Published: 31 December, 2024

Corresponding Author: * **Zulfiqar Ali**

Abstract
This study investigates optimal link prediction methods in complex human networks by applying eight similarity-based algorithms across five real-world human interaction datasets. Networks were converted into adjacency matrices, with links removed for testing. Each method was evaluated using AUC scores to assess predictive performance. Results showed that the Resource Allocation Index (RAI) consistently outperformed other techniques, particularly on large and complex networks, highlighting its effectiveness for predicting missing or future connections in human-centric systems.

INTRODUCTION

Complex networks provide a robust framework for describing real-world networks [1]. These networks can be accurately represented using graphical models [2], where nodes correspond to social entities or individuals, and links signify interactions or relationships between these nodes [3]. Complex networks involve nodes representing entities (biological components, individuals, etc.) and links indicating their interactions [4]. The applications of complex networks are widespread, particularly in applied sciences. They have played a crucial role in shaping legislative efforts focused on citizen engagement and optimizing road networks for efficient routes [5]. Because of over time, new edges and vertices are constantly added due to their inherent adaptability. [6]. Research into these networks has gained significant attention, leading to practical applications that address various network-related inquiries [7]. Because complex networks are dynamic, the problem of missing link prediction (LP) has grown more difficult. [8].

The core concept of the LP problem involves predicting potential links between nodes within a given complex network [9]. In essence, LP predicts linkages that do not yet exist between nodes by estimating the likelihood of links between pairs of nodes that are not yet linked. [10]. Solving the problem of predicting link existence holds promise not only for filling in missing data gaps within complex networks, such as predicting protein-protein interactions in biological networks, but also for forecasting network evolution (predicting future link existence) [11], [12]. The underlying premise of LP is that if two nodes are similar, they probably have a link. Calculating node similarity is a critical aspect, and the original method used to address this problem was the Common Neighbors (CN) technique [13]. But there were drawbacks to this approach, especially its dependence on nodes with higher degrees. Several CN variations, such as the Jaccard Index [14], have been used to minimize this bias to get around this restriction. Other techniques for calculating node

similarity, such as the Katz Index, have also shown promise [15].

In the realm of social networks, LP has exhibited progress, effectively revealing hidden connections that contribute to a deeper understanding of social dynamics [16]. Similarly, LP's application extends to transportation networks, shedding light on potential links that impact route optimization and overall network efficiency [17]. Furthermore, in the 2 realms of biological networks, LP has been harnessed to uncover latent relationships, enhancing our comprehension of complex molecular interactions [18]. Beyond these domains, LP techniques have proven beneficial for path analysis within the World Wide Web (WWW), enhancing link navigation and information retrieval [19]. This method has also found application in hyperlink creation and prediction, contributing to the dynamic evolution of web content and connectivity.

[20] [21]. Additionally, LP has been harnessed to delve into intricate protein-protein interactions, offering a more nuanced perspective on biological processes [22].

The rest of the paper is organized in this manner. Section 2 discusses the literature review. Section 3 discusses methodology, Section 4 discusses results and analysis, and Section 5 discusses the conclusion and Future Work.

1. LITERATURE REVIEW

LP is an emerging research problem, and work has been done in this area. In 2011, a survey was performed by Tao Zhou and Linyuan Lu¨ in which the LP techniques were implemented on different datasets. Local, global, and quasi-local similarity indices were the similarity indices that were used. Also, the Maximum likelihood methods were implemented, and probabilistic models were applied. As an evaluation metric, they employ accuracy and precision. The evaluation of the research led to the conclusion that the similarity indices performed effectively in the provided LP situation. [4]. In 2014, LP techniques were implemented using the information theory perspective for the role of network topology. Fei Tan, Boyao Zhu, and Yongxiang Xia use data mining techniques. The measurement metric for evaluation was accuracy. The results concluded that a mutual information approach was introduced [30].

The use of the Pearson correlation coefficient approach on high-order pathways was discovered to be successful in 2015. The data mining techniques were implemented on the 9 empirical networks [31]. In 2016, another survey was performed, whose main focus was the computational complexity analysis of other techniques. The methods used were those of data mining, and these were implemented on 7 of the empirical networks with different perspectives and backgrounds. Accuracy and Precision were used as evaluation metrics. The comparison was made with the old techniques [5]. In 2017, in the area of graph theory, a quick way to resolve the LP problem was developed. The problem covered was that LP can do more than just determine which edges will appear in the network. The techniques used were those of data mining. The evaluation metrics used were Accuracy and Precision [32].

In 2018, a network's nine missing links were located using a multiple attribute decision-making process. In this study, a novel approach was developed and applied to 10 real-world networks. A metric for evaluation was accuracy. The MADM method was effective in solving the problem [33]. A survey of LP methods, applications, and performance was carried out in 2019. They were implemented in the graph theory domain, and the techniques used were those of data mining. 8 different datasets were used, and the accuracy was the evaluation metric [34]. In 2019, common neighbor degree penalization was performed on the real world as well as the synthetic networks. Accuracy and precision were employed as evaluation metrics. The research covered two main challenges: one was that of large data, and the second was of low computational complexity [35]. The LP on local path research was carried out in the field of graph theory in 2020. Twelve separate datasets were subjected to data mining techniques. The evaluation metric employed was accuracy and precision. When compared to current approaches, the results produced offer greater accuracy [10]. The parameterized approach based on the centrality and common neighbors was created in 2020 and was intended to be used on eight networks to address the missing LP issue. The evaluation metric employed was accuracy and precision. It demonstrates that accuracy outperforms precision using the employed approach [36].

2. THE PROBLEM OF LINK PREDICTION

One of the main issues with LP is that the networks are dynamic and incomplete, which makes it possible for nodes to form and disappear at any time in the future. In terms of the static networks, the nodes are static, but in the case of real-world networks, the situation is opposite, and the links can be changed dynamically with time. For example, we have the graph G (V, E) as shown in Figure 1 [41]. It is essential to note that although networks in certain contexts might be considered static, the same does not hold for real-world networks. This characteristic poses a significant challenge for LP [23].

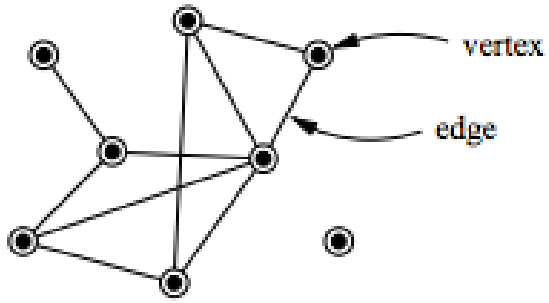

Figure 1: Graphical Representation [41]

2.1 Graph Types

The variety of network structures and ideas, with the following graph categories, which are shown in Figure 2.

An Undirected Graph is a mathematical structure consisting of nodes or vertices connected by edges, where the edges lack directionality, representing mutual relationships or connections between the nodes.

- **Directed Graph** is a directed graph is a mathematical representation in which nodes or vertices are connected by directed edges, signifying one-way relationships between the nodes, indicating a clear direction from one node to another. It shows the direction of edges.
- **A Weighted Graph** is a graph in which each edge has a real-numbered value associated with it
- **A Connected Graph** is a graph in which all the vertices are connected to all the other vertices.

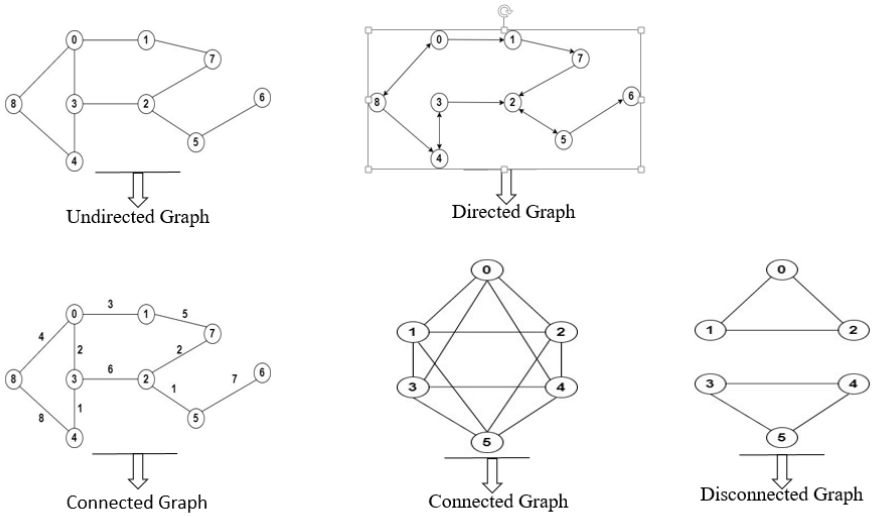

Figure 2: Types of Graphs

3. APPLICATION OF LINK PREDICTION

LP approaches have been used extensively in a variety of fields. Any area where entities interact with one another can gain from LP. The following are some of the common uses for LP.

- **Identification of Anomalous Email**

The feature of detecting anomalous email is that communication acts occur in a network-based structure. One-to-one or one-to-many communication is possible. One-to-many approach detection and anomaly detection are performed using the LP technique. A strategy put out by Huang and Zeng views email communication as a network with nodes and uses LP techniques to forecast future conversations [24].

- **Co-participation in Event Prediction**

In social network analysis, persistent relationships and discrete events are included in the data. In these social networks, the data is changing over time. Such linkages and events can be used to forecast the co-participation in an event. How likely is it that X will call Y next week, for instance? [25].

- **Items recommended to users.**

A recommendation system that suggests items to the users from a Bipartite Network. Bipartite networks consist of user-tags, item-tags, and some other networks. With these, the LP technique is applied, and the recommendation system works in that manner [26].

- **Yeast proteome**

A protein function that takes advantage of LP methods to improve protein interaction. After testing, the role-similarity measure is applied to the yeast proteome [27].

4. METHODOLOGY

The detailed research methodology begins with the collection of datasets, which are in numeric form. The data sets are in the raw form and needed to be presented in an understandable form, which was done by the tool (MATLAB). Using Matlab, after being stored in an array, the data sets were individually given a graphic representation, so that the datasets could be understood properly. The adjacency matrix is obtained from the graphical representation, which represents the links between the nodes. There are two instances. The phrase "Adjacency Matrix" refers to a matrix where either a link exists, signified by a 1, or a link does not exist, denoted by a 0. The Adjacency matrix is then divided into Training and Probe sets. Different division classes, including 90-10, 80-20, 70-30, 60-40, and 50-50, were used to divide the training and probe sets. To determine the best division value, the implementation performed on each of the division classes will produce particular results that will be compared. The AUC value that is closest to "1" will be chosen for further processing. Each of the chosen approaches then runs the prediction phenomenon after choosing the train-probe split percentage. A post-prediction analysis is carried out to determine the effectiveness of each technique on each set of data. In the end, the results are compared, and the best predicting technique is identified. The overall methodology of the research can be shown in Figure 3.

To evaluate the performance of the selected algorithms in our analysis, a comprehensive set of experiments was conducted on four distinct datasets for each algorithm. The underlying process is systematically outlined below and can also be visualized in Figure 3 for a clearer understanding. Once the datasets are refined, the subsequent step involves the creation of an adjacency matrix. This matrix serves as a fundamental representation of the network, capturing the relationships between nodes. To construct this matrix, the process begins by identifying the maximum number of nodes present in the dataset. This value, denoted as max(max(Dataset())), essentially indicates the total count of nodes within the dataset. Subsequently, a square matrix of dimensions N x N is generated, where N corresponds to the previously determined maximum node count. Initially, this matrix is populated with zeros, signifying the absence of any connections between nodes.

The heart of the process lies in populating this matrix with the actual interactions present in the dataset. The values within the matrix are updated based on the

existence of connections between nodes, thereby effectively reflecting the network's underlying structure. This transformed matrix serves as the final adjacency matrix, providing a comprehensive overview of the relationships among entities within the network. The datasets are randomly split into a training set and a testing set. For a single experiment, there would be no accurate results due to the bias in data selection. To overcome this effect for each algorithm on the respective data set, we perform 100 experiments and then calculate the average AUC. Through this procedure, the total number of experiments is 4800. After 4800 successful experiments, the results would be more accurate. And the error rate would become negligible.

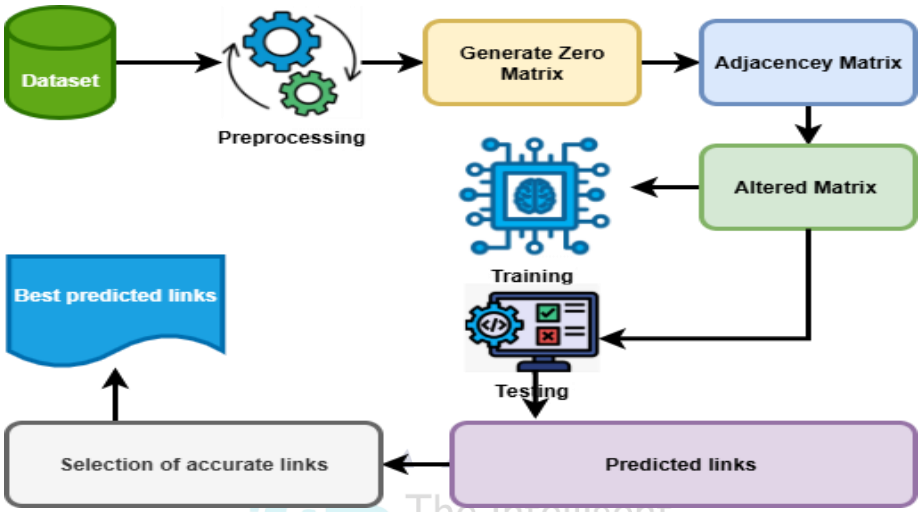

Figure 3. Methodology

5.1. Terminology and Notations

The size of the network may alternatively be expressed in terms of a link between the i and j nodes, denoted by the pair $e\_i.j.$ $|V|$, which stands for "number of nodes." The number of linkages is represented by the letter E. LP uses undirected graphs; therefore, instead of arcs, they contain edges. In essence, the elements of the vertex set that are represented as $e = (vi,vj) \in E$, where $(vi,vj) \in V$, are called edge sets. When the edge E is split into two sections, ET and EP, which stand for Training Set and Probe Set, respectively, and $E^T \cup E^P = E$, the standard LP may be expressed as follows. The set of nodes that are connected by the edge $x \in V$ is known as x's neighbors and is symbolized by the symbol Γx. The degree of node x in the undirected graph is denoted by $|\Gamma x|$.

5. TOOLS AND TECHNIQUES

The following tools and techniques are employed for this work.

5.1 MATLAB

MATLAB is a tool used for numeric and programming platforms developed by MathWorks. It allows matrix manipulation, implementation of data and algorithms, and plotting of data. It will help to visualize the datasets and convert them into an adjacency matrix for further processing of the results. Using MATLAB, after being stored in an array, the data sets were individually given a graphic representation, so that these datasets could be understood properly. The adjacency matrix is obtained from the graphical representation, which represents the links between the nodes. There are two instances. The phrase "Adjacency Matrix" refers to a matrix where either a link exists, signified by a 1, or a link does not exist, denoted by a 0. The Adjacency matrix is then divided into Training and Probe sets.

5.2 Jaccard Index

A local similarity index, the Jaccard Index, is comparable to the CN but more effective because it normalizes the score. The probability of choosing pair-

wise vertices from a node's neighbors is what is known as, and it is represented as

$$S_{ab}^{Jaccard} = \frac{|\Gamma(a) \cup \Gamma(b)|}{|(\Gamma(a) \cap \Gamma(b)|} \quad (1)$$

Where Γ(*a*) is the collection of a node's neighbors [37].

5.3 Preferential Attachment Index

PA is a local similarity index and depends on the degree of nodes a and b. It is defined as the probability that the nodes a and b, where the link exists, are proportional to k_a and k_b [38] and is represented as

$$SabPA = ka \times kb \quad (2)$$

5.4 Resource Allocation Index

RA is a local similarity index, and it works based on the intermittent nodes connecting node x and node y. It can be defined as the amount of resources occupied by node x via an indirect relation from node y.

$$S_{ab}^{RA} = \sum_{z \in \Gamma(i,j)} \frac{1}{k(z)} \quad (3)$$

Where *k*(*z*) is the degree of node z [39].

5.5 Common Neighbors

The concept that two strangers who have a friend are more likely to be introduced than those who don't is encapsulated by the phrase "common neighbors." [16].

$$cn(i,j) = |\Gamma(i) \cap \Gamma(j)| \quad (4)$$

(4)

5.6 Adamic Adar

Based on their shared neighbors, Adamic Adar is a measure used to determine how close two nodes are to one another. [16].

$$aa(i,j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(k)|} \quad (5)$$

5.7 Hub Promoted Index

Due to the denominator's dependence on the minimum degree of the vertices of interest, this measure gives linkages near hubs (high-degree vertices) better scores. [33].

$$hpi(i,j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\sqrt{|\Gamma(k)|}} \quad (6)$$

5.8 Salton Index

When provided with vertices, it calculates the cosine of the angle between the columns of the adjacency matrix. Information retrieval frequently employs this measure. [36].

$$salton(i,j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{\sqrt{|\Gamma(i)||\Gamma(j)|}} \quad (7)$$

5.9 Hub Depressed Index

This measure gives connections that are close to hubs lower ratings than the Hub Promoted Index does. The large neighborhoods are penalized. [34].

$$hdi(i,j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \sqrt{|\Gamma(k)|} \quad (8)$$

6. DATASETS

The following are the various datasets used in this work.

6.1 Human Contact Network

People or humans are represented as nodes or vertices in a network model called the Human Contact Network, and the contacts they have—such as phone calls, texts, or interactions—are represented as links or edges between the nodes. Each node symbolizes a different person, and the edges connecting the nodes indicate that there has been interaction or communication between the various people. This data set provides important details regarding interpersonal relationships, communication preferences, and social network organization. Researchers can look at how information moves across this network dynamically, identify key nodes or central figures, and understand how disease or information spreads through social interactions [7].

6.2 Kindergarten Network

A structured collection of information and

observations that captures the communication, cooperation, and behavioral exchanges among young children within a kindergarten setting. This dataset includes records of how children engage with each other, their interactions during group activities, and their responses to various social situations. It provides insights into the development of social skills, cooperation, conflict resolution, and the formation of social relationships among kindergarten-aged individuals [22].

6.3 Human Wireless Contact Network

In a network model called the Human Wireless Contact Network, people or humans are shown as nodes or vertices, while wireless contacts or connections between them are shown as links or edges. Each node in this dataset represents a distinct individual, and the edges between nodes show that the corresponding individuals have interacted or communicated wirelessly [22].

6.4 Women's Social Event Interaction Network

A network representation that was inspired by a particular women's event. The women participants in this network are shown as nodes or vertices, and their interactions with one another throughout the event—such as speaking or meeting—are shown as linkages or edges between the nodes. This data set offers insightful information about the social dynamics and communication styles of women at the event. By examining this network, researchers can learn how social interactions are structured, spot important players or pivotal figures in the group, and investigate how information or social influence moves among the participants [32].

6.5 Karate Network

A karate club's members are represented as nodes or vertices in the network representation known as the" Karate Network," while their interactions or relationships with one another are shown as links or edges connecting the nodes. Due to a study conducted in the 1970s by social psychologist Wayne W. Zachary, this dataset rose to fame. The network captures the ties between members, such as friendships, social contacts, or training alliances. The interactions of the karate club were studied over time. Researchers have learned a lot about social networks and community systems by examining the Karate Network. The network is a well-known example of a community that was finally divided into two groups after being widely examined in the field of social network analysis [7]

The average density, average degree of nodes, vertices, and edges of the aforementioned datasets are given in Table 1.

The table presents information about four different datasets, each representing a social network with nodes (V) and edges (E). Additionally, it provides two average measures, namely the average degree (¡k¿) and the average shortest path length (¡d¿).

Table 1: Datasets Details

Datasets	|V |	|E|	< k >	< d >
Human Contact Network	43	336	15.628	0.80
Kindergarten Network	111	225	16.87	0.311
Human Wireless Contact Network	274	2124	15.5	0.41
Karate Network	34	78	4.588	0.137
Woman Social Event Network	18	63	7	0.592

7. EVALUATION METRICS

The AUC has been widely used for prediction accuracy [40]. The author uses AUC as an LP accuracy measure. The information from E^T is allowed to record the performance score $S_{x.y}$. The prediction score will be based on which n pairs of nodes from E^P and E^- are randomly selected.

If the score measured from EP is bigger than E^-, then n′, and if EP is equal to E^-, then n″, AUC can be calculated by Equation (9).

$$AUC = \frac{(n' + 0.5n'')}{n} \quad (9)$$

7.1 Working of AUC

We use the LP method, or PA Index, as an example of the AUC for easier understanding. The real

network is represented by set (a) in Figure 4, whereas the training and probing sets are represented by sets (b) and (c), respectively, and the undetected links are included in set (d). Next, a pair of nodes from unobserved linkages and BD from the probe set were selected. The only set that can be utilized to compute node degrees is the training set. For instance, we select the AB pair from links that are not visible. Applying the PA Index now, the similarity between nodes BD and CD is $ScoreAB = kA \times kB = 0$ and $ScoreBD = ScoreAB$; therefore, $n' = 0$ and $n'' = 1$, and the AUC will be equal to, in Equation (10).

$$AUC = \frac{(0 + 0.5 \times 1)}{1} = 0.5 \quad (10)$$

As the AUC value is closer to 1, it means that the LP technique is efficient.

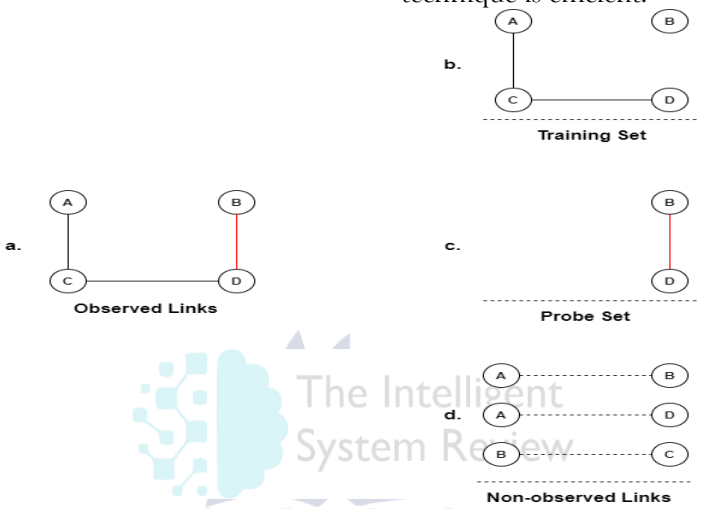

Figure 4: AUC Calculation

7.2 Why do we use AUC

The flexibility of the Area under the ROC Curve (AUC-ROC) to assess the performance of binary classifiers across a range of classification thresholds makes it a popular statistic in a variety of machine learning applications, including link prediction. Link prediction algorithms assign probability scores to potential links, estimating their likelihood of existence. AUC-ROC (Area under the Receiver Operating Characteristic curve) is a metric that gauges classifier performance without requiring a predetermined probability threshold. This is important as the best threshold can differ based on the application or data. AUC-ROC provides a comprehensive assessment of the classifier's ability to distinguish between positive and negative instances across all potential thresholds. In scenarios with imbalanced datasets, where non-links outnumber actual links, AUC-ROC takes into account both true positive rate (sensitivity) and false positive rate, offering a balanced performance assessment not skewed by class distribution.

In link prediction, probabilities rank links. A higher AUC-ROC signifies better differentiation between real links and non-links, crucial for skewed data and recommendations. AUC-ROC allows algorithm comparison and handles varied score calibrations, supporting reliable performance assessment. By measuring a model's ability to prioritize actual links without setting a threshold, it ensures a comprehensive evaluation of link prediction effectiveness in diverse contexts.

However, it's important to note that while AUC-ROC is a valuable metric, it might not capture the full picture in cases where the specific classification threshold is critical or when the class distribution is extremely imbalanced. In such cases, other metrics like Precision-Recall curves and F1-score might provide more nuanced insights. Therefore, the choice of evaluation metric should be aligned with the

specific goals and characteristics of the link prediction problem at hand.

8. RESULTS AND DISCUSSIONS

The primary objective of this study is to investigate the effectiveness of eight Link Prediction (LP) techniques within the realm of human complexity networks. The focal point is to discern the LP approach that exhibits superior performance when applied to real-world networks. Through a series of well-designed experiments, the study sheds light on the capacity of LP techniques to identify missing or forthcoming connections, particularly evident in the context of Human Contact Networks (HCN). The experimentation phase involves a deliberate variation of Probe Set (EP) percentages to enhance the precision of Area Under the Curve (AUC) measurements. The careful manipulation of EP percentages allows for a more accurate assessment of the LP techniques' predictive capabilities. In pursuit of this objective, multiple experiments were meticulously carried out with varying Training and Probe sets. These sets encompassed proportions of 10%, 20%, 30%, 40%, and 50%, each contributing to a comprehensive evaluation of the LP techniques' efficacy. The comparative analysis of the results obtained from these experiments enabled the selection of the most promising LP technique based on its AUC performance. These results are aptly summarized in Tables 2, 3, 4, 5, and 4.5, providing a clear overview of the AUC values obtained for different EP percentages. Furthermore, for a more intuitive understanding, the study includes graphical representations of each dataset. These visualizations, presented in Figures 5, 6, 7, and 8, offer an insightful depiction of the LP techniques' performance across different scenarios, highlighting their proficiency in capturing network dynamics and predicting linkages. In conclusion, this study ventures into the realm of human complexity networks to unravel the potential of LP techniques in predicting links.

Table 2: Average AUC Results of 90%:10%

Techniques	Human Contact Network	Human Wireless Contact Network	Karate Network	Kinder garden	Women's Social Event Interaction Network
AA	0.8581	0.9345	0.7074	0.7858	0.8346
CN	0.846	0.9333	0.6845	0.7742	0.8154
PA	0.7095	0.9362	0.7123	0.5531	0.6909
HDI	0.8193	0.8859	0.5997	0.7884	0.801
HPI	0.8313	0.7005	0.6971	0.7326	0.734
JC	0.847	0.934	0.6693	0.8265	0.7708
RA	0.8588	0.9342	0.7224	0.8006	0.8356
SALTON	0.8553	0.8905	0.6194	0.7885	0.8036

Divide the datasets into 90%,10% We give 90% of the nodes to training and 10% to test. From the average AUC, we conclude that the Resource Allocation Index (RAI) works best among all other algorithms we use.

Table 3: Average AUC Results of 80%:20%

Techniques	Human Contact Network	Human Wireless Contact Network	Karate Network	Kinder garden	Women's Social Event Interaction Network
AA	0.834	0.9307	0.6869	0.7529	0.818
CN	0.8283	0.9308	0.6593	0.7319	0.7878
PA	0.7069	0.933	0.6798	0.5511	0.6884
HDI	0.8004	0.8854	0.583	0.7607	0.7547
HPI	0.8114	0.7284	0.6584	0.7176	0.7039
JC	0.8289	0.9271	0.6575	0.7505	0.788
RA	0.8451	0.9307	0.6984	0.7431	0.8139

SALTON	0.8272	0.8846	0.6154	0.7402	0.7634

Divide the datasets into 80%,20% We give 80% of the nodes to training and 20% to test. From the average AUC, we conclude that the Resource Allocation Index (RAI) works best among all other algorithms we use.

Table 4: Average AUC Results of 70%:30%

Techniques	Human Contact Network	Human Wireless Contact Network	Karate Network	Kinder garden	Women's Social Event Interaction Network
AA	0.8187	0.9263	0.6553	0.7115	0.7907
CN	0.8053	0.9247	0.6406	0.6986	0.7483
PA	0.7049	0.9284	0.6641	0.5284	0.6769
HDI	0.774 8	0.8851	0.5849	0.7122	0.7287
HPI	0.7799	0.7537	0.6207	0.6821	0.6936
JC	0.8056	0.9249	0.6286	0.6952	0.7543
RA	0.8195	0.926	0.6512	0.7053	0.7989
SALTON	0.7942	0.8817	0.6021	0.7012	0.724

Divide the datasets into 70%. 30% we give 70% of the node to training and 30% to test. From the average AUC, we conclude that the Resource Allocation Index (RAI) works best among all other algorithms we use.

Table 5: Average AUC Results of 60%:40%

Techniques	Human Contact Network	Human Wireless Contact Network	Karate Network	Kinder garden	Women's Social Event Interaction Network
AA	0.7873	0.9189	0.6149	0.6639	0.7477
CN	0.7777	0.9182	0.6081	0.6501	0.7202
PA	0.6917	0.9267	0.6611	0.5259	0.6752
HDI	0.7492	0.8836	0.5834	0.6621	0.6884
HPI	0.7476	0.7723	0.5974	0.6591	0.6604
JC	0.7752	0.9197	0.6121	0.6551	0.7174
RA	0.7933	0.9202	0.6195	0.6678	0.7377
SALTON	0.7584	0.8756	0.5875	0.6684	0.6884

Divide the datasets into 60%. 40% we give 60% of the node to training and 40% to testing. From the average AUC, we conclude that the Resource Allocation Index (RAI) works best among all other algorithms we use.

Table 6: Average AUC Results of 50%:50%

Techniques	Human Contact Network	Human Wireless Contact Network	Karate Network	Kinder garden	Women's Social Event Interaction Network
AA	0.7569	0.9116	0.5983	0.6222	0.6948
CN	0.7441	0.9104	0.5836	0.6184	0.6641
PA	0.6886	0.9234	0.6357	0.5154	0.6597
HDI	0.7167	0.8809	0.5754	0.6268	0.6583
HPI	0.708	0.7912	0.5843	0.6167	0.6539
JC	0.7411	0.9112	0.5804	0.6201	0.6843
RA	0.7584	0.9124	0.593	0.6191	0.6938
SALTON	0.7208	0.8679	0.5739	0.6148	0.6478

Divide the datasets into 50,50 %. We give 50% of the nodes to training and 50% to testing. From the average AUC, we conclude that the Resource Allocation Index (RAI) works best among all other algorithms we use.

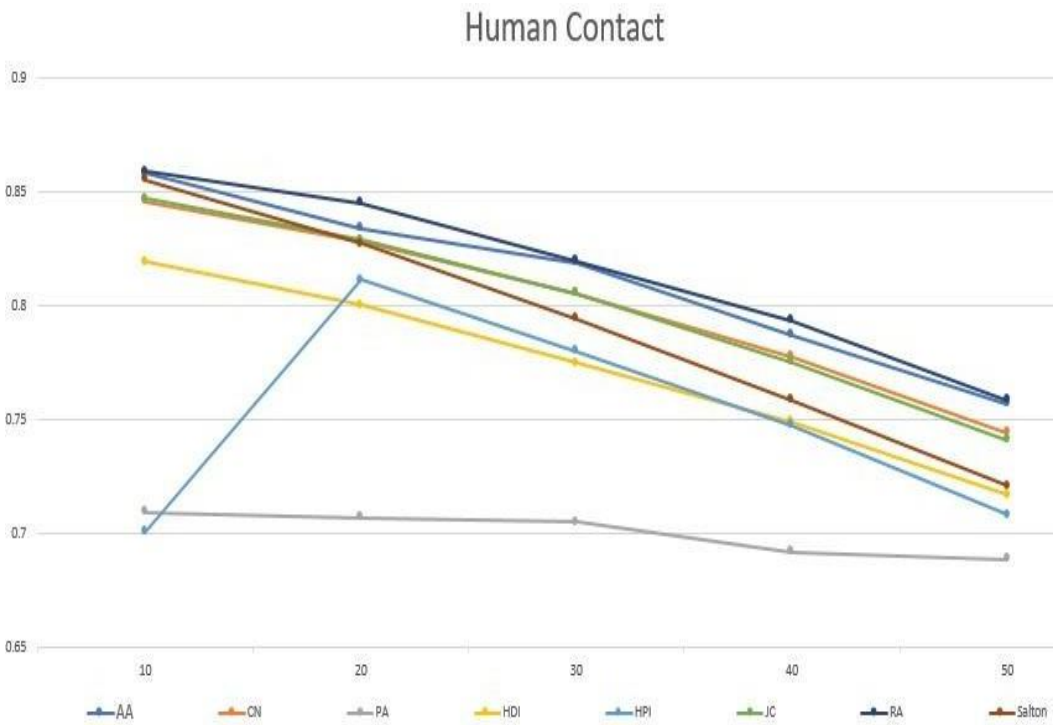

Figure 5: Human Contact Network Percentage Results

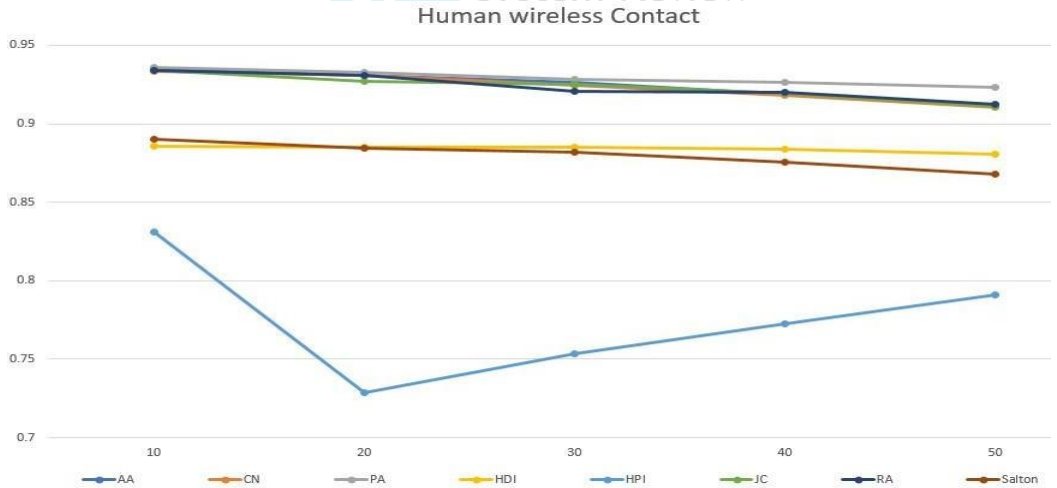

Figure 6: Human Wireless Contact Network Percentage Results

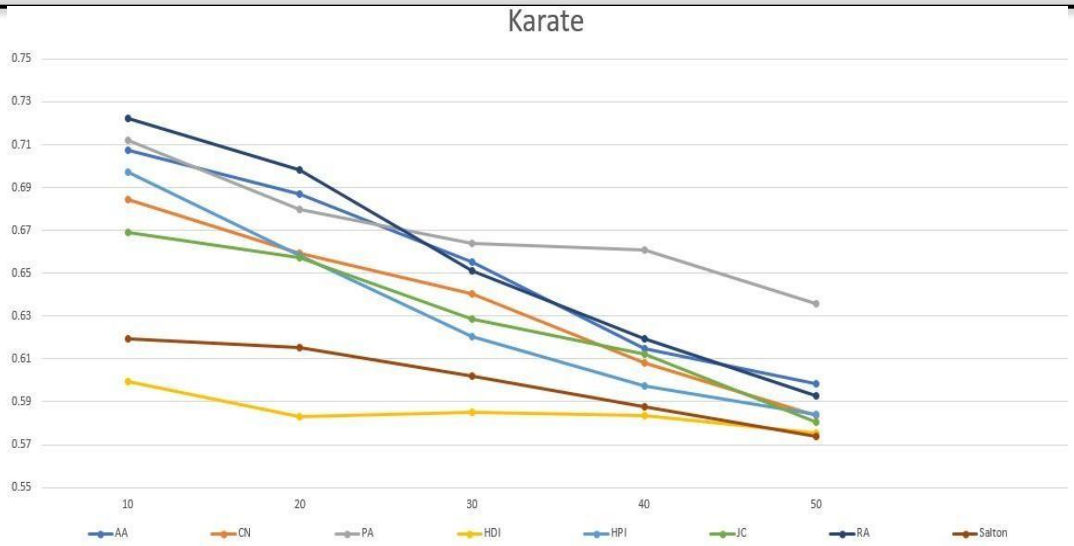

Figure 7: Karate Network Percentage Results

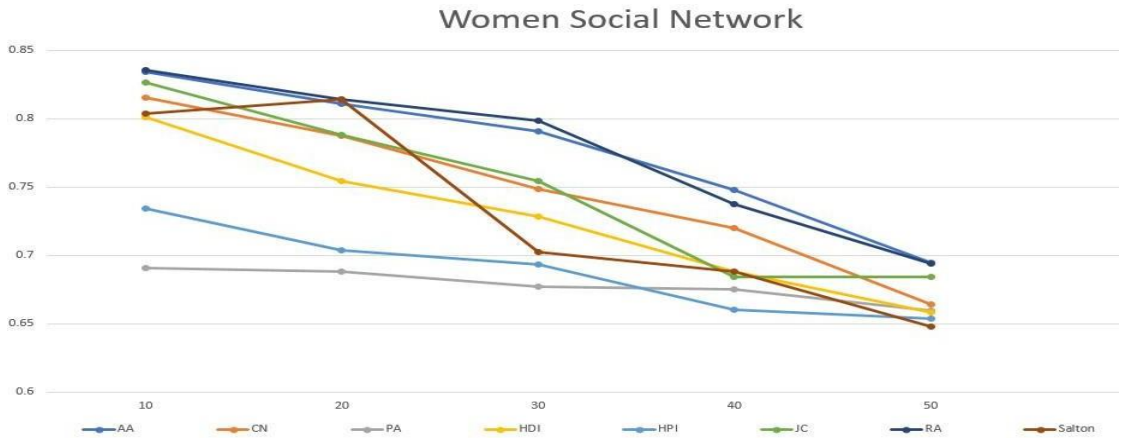

Figure 8: Women's Social Event Interaction Network Percentage Results

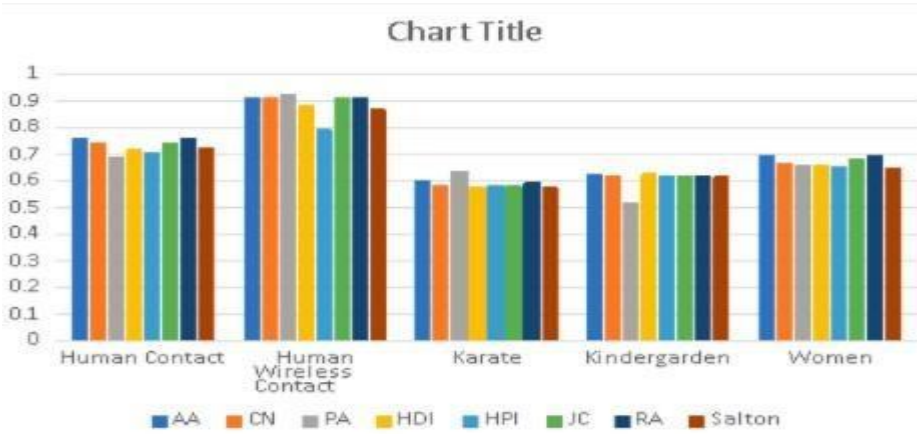

Figure 9: Average AUC

8.1 Critical Discussion

According to the Figures and Tables, the 10% Probe sets' AUC values were quite close to 1. Therefore, the Probe set taken into account was 10%. Finding the best HCN prediction method with greater accuracy was the major goal. In terms of the outcomes, the 10% Probe set Table 4.1 may demonstrate that the RA Index brings each dataset's outcomes closer to "1." As a result, we can conclude that in the domain of human complexity, the RA Index performs better than other LP approaches.

The central focus of this work revolves around the exploration and identification of the most suitable method for actual networks in the intricate field of human interactions. By conducting comprehensive experiments, the study aimed to assess the LP techniques' capabilities in discovering missing or potential links within the HCN. To ensure the accuracy and reliability of the evaluation process, the experimentation incorporated varying percentages of Probe Set (EP), which allowed for a more precise measurement of the AUC. This approach was implemented to gain deeper insights into the performance of each LP technique across different proportions of training and probe data. The creation of multiple sets of Training and Probe sets, ranging from 10% to 50%, allowed for a systematic comparison of the AUC values achieved.

The findings of the experiments underscored the significance of LP techniques in the realm of Human Complex Networks. These techniques demonstrated their potential in predicting and uncovering connections between nodes, thereby revealing latent relationships within the network structure. Moreover, the varying proportions of the Probe Set enabled a meticulous assessment of the LP techniques' robustness and adaptability to different data scenarios. As a result of the meticulous analysis, the LP technique that exhibited the highest AUC value was identified as the most promising approach for further investigations. This technique's superior performance could have substantial implications for enhancing our comprehension of complex human interactions and network dynamics. Overall, the research undertaken in this study contributes significantly to the field of link prediction in real-world networks, particularly in the Human Complex domain. By shedding light on the strengths and weaknesses of different LP techniques and their sensitivity to data proportions, the study paves the way for future advancements in understanding and analyzing intricate human networks and social systems. These insights hold the potential to drive innovations in various domains, including social sciences, data mining, and network analysis, opening new avenues for research and practical applications.

The methodology for the entire research is provided in this paper; however, there are several limitations that are not addressed, such as the network size, which can make predictions take longer if the network is large, and thus requires more computer power.

9. CONCLUSION

The LP is an emerging research topic, and it has received much attention in the last two decades. Many different disciplines have benefited from this research area. In addition to helping with the analysis of missing links in biological elements like PPI, Yeast Proteome, Gene Natures, and many more, the application of LP to real-world complex networks has also advanced computer science by predicting website navigation, hyperlink navigation, social elements like Facebook and Twitter, and many other areas. Additionally, it has advanced sports by providing some future predictions, such as score predictions in cricket. The real-world network in the field of human complex networks was the focus of this study. Much work has been performed for predicting the links that are either missing or deleted, or for future prediction of the links, but some of the human complex networks are left untouched, which can help to provide opportunities in LP in the domain. This study offers the whole methodology of the entire study of "Link Prediction in Human Complex Networks," which is essentially a comparative examination of the selected LP approaches. A literature review was included in the paper to better comprehend the research subject." Furthermore, LP and complicated networks were described. Furthermore, the author assessed the performance of eight different algorithms using five distinct datasets. The evaluation metric used was the AUC. The results indicate that among all the algorithms, the Resource Allocation Index (RAI) demonstrated the best performance on large and complex datasets.

In recent years, LP has garnered attention across fields like physics, biology, and computer science. Research outcomes vary based on field-specific characteristics, advancing LP's significance and problem-solving potential. Despite progress, the challenge remains to reconcile static LP techniques with the dynamic nature of real-world complex networks.

REFERENCES

I. Ahmad, M. U. Akhtar, S. Noor, and A. Shahnaz, "Missing link prediction using common neighbor and centrality based parameterized algorithm," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

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

A. Kumar, S. S. Singh, K. Singh, and B. Biswas, "Link prediction techniques, applications, and performance: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 553, p. 124289, 2020.

T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

V. Martınez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM computing surveys (CSUR)*, vol. 49, no. 4, pp. 1–33, 2016.

N. Z. Gong, A. Talwalkar, L. Mackey, *et al.*, "Joint link prediction and attribute inference using a social- attribute network," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 2, pp. 1–20, 2014.

A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, 2010.

Z. Liu, Q.-M. Zhang, L. Lu¨, and T. Zhou, "Link prediction in complex networks: A local naıve bayes model," *EPL (Europhysics Letters)*, vol. 96, no. 4, p. 48007, 2011.

P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, "Wtf: The who to follow service at twitter," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 505–514.

F. Aziz, H. Gul, I. Muhammad, and I. Uddin, "Link prediction using node information on local paths," *Physica A: Statistical Mechanics and its Applications*, vol. 557, p. 124980, 2020.

A. Brasoveanu, M. Moodie, and R. Agrawal, "Textual evidence for the perfunctoriness of independent medical reviews,"

L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011,pp. 635–644.

L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1100–1108.

F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *The Journal of mathematical sociology*, vol. 1, no. 1, pp. 49–80, 1971.

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

S. Zhou and R. J. Mondrag´on, "Accurately modeling the internet topology," *Physical Review E*, vol. 70, no. 6, p. 066108, 2004.

Y. Hadas, G. Gnecco, and M. Sanguineti, "An approach to transportation network analysis via transferable utility games," *Transportation Research Part B: Methodological*, vol. 105, pp. 120–143, 2017.

J. S. Kim and M. Kaiser, "From caenorhabditis elegans to the human connectome: A specific modular organization increases metabolic, functional and developmental efficiency," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1653, p. 20130529, 2014.

R. R. Sarukkai, "Link prediction and path analysis using markov chains," *Computer Networks*, vol. 33, no. 1- 6, pp. 377–386, 2000.

S. F. Adafre and M. de Rijke, "Discovering missing links in wikipedia," in *Proceedings of the 3rd international workshop on Link discovery*, 2005, pp. 90– 97.

J. Zhu, J. Hong, and J. G. Hughes, "Using markov models for web site link prediction," in *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*, 2002, pp. 169–170.

P. Jaccard, "Etude comparative de la distribution florale dans une portion ´ des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

Z. Huang and D. D. Zeng, "A link prediction approach to anomalous email detection," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, vol. 2, 2006, pp. 1131–1136.

J. O'Madadhain, J. Hutchins, and P. Smyth, "Prediction and ranking algorithms for event-based network data," *ACM SIGKDD explorations newsletter*, vol. 7, no. 2, pp. 23–30, 2005.

J. Kunegis, E. W. De Luca, and S. Albayrak, "The link prediction problem in bipartite networks," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, Springer, 2010, pp. 380–389.

P. Holme and M. Huss, "Role-similarity based functional prediction in networked systems: Application to the
yeast proteome," *Journal of the Royal Society Interface*, vol. 2, no. 4, pp. 327–333, 2005.

B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook,"
Proceedings of the 2nd ACM workshop on Online social networks, 2009, pp. 37–42.

C. Lin, Y.-r. Cho, W.-C. Hwang, P. Pei, and A. Zhang, "Clustering methods in protein-protein interaction
network," *Knowledge Discovery in Bioinformatics: techniques, methods and application*, pp. 1–35, 2007.

F. Tan, Y. Xia, and B. Zhu, "Link prediction in complex networks: A mutual information perspective," *PloS one*, vol. 9, no. 9, e107056, 2014.

H. Liao, A. Zeng, and Y.-C. Zhang, "Predicting missing links via correlation between nodes," *Physica A: Statistical Mechanics and its Applications*, vol. 436, pp. 216–223, 2015.

B. Pachev and B. Webb, "Fast link prediction for large networks using spectral embedding," *Journal of Complex Networks*, vol. 6, no. 1, pp. 79–94, 2018.

L. Li, S. Bai, M. Leng, L. Wang, and X. Chen, "Finding missing links in complex networks: A multiple- attribute decision-making method," *Complexity*, vol. 2018, 2018.

A. Kumar, S. S. Singh, K. Singh, and B. Biswas, "Link prediction techniques, applications, and performance: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 553, p. 124289, 2020.

S. Rafiee, C. Salavati, and A. Abdollahpouri, "Cndp: Link prediction based on common neighbors degree penalization," *Physica A: Statistical Mechanics and its Applications*, vol. 539, p. 122950, 2020.

I. Ahmad, M. U. Akhtar, S. Noor, and A. Shahnaz, "Missing link prediction using common neighbor and centrality based parameterized algorithm," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

L. Lu¨, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, p. 046122, 2009.

L. Lu¨ and T. Zhou, "Link prediction in weighted networks: The role of weak ties," *EPL (Europhysics Letters)*, vol. 89, no. 1, p. 18001, 2010.

[40] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[41] https://en.wikipedia.org/wiki/ Vertex_%28 graph_theory%29#, Access date: 05/08/2025