Pre-Print Version

Analysis of link prediction methods based on topological information of the COVID-19 network

Shauban Ali Solangi[1*], Abdul Waheed Mahesar[2], Lachhman Das Dhomeja[3], Khalil-ur-Rehman Khoumbati[4], Bisharat Rasool Memon[5]

Abstract— Link prediction is an important area of research in complex networks, helping to understand the numerous relationships between nodes based on their topological structures. Recently, link prediction has been extensively used to study the spread of COVID-19. Given the diffusion of the virus from one location to another, it is vital to classify af fected locations and predict the links related to the virus's spread. We propose a novel approach for the COVID-19-infected locations based on complex network theory to analyze and explore link prediction. For this, we construct a weighted two-mode network from the COVID-19 dataset, comprising confirmed case records of affected locations during specific weeks. The network comprises nodes representing locations and weeks, these two nodes are linked if a COVID-19 case is confirmed within a location during a particular week. The weight of the links corresponds to the frequency of COVID-19 cases. We apply five local and four global link prediction methods and evaluate their results using the ROC curve. We select the most appropriate link prediction method for the observed network based on our evaluation. The experimental results show that the ACT method has outperformed the other eight local and global similarity methods for the COVID-19 location network. Consequently, our findings reveal better accuracy, signifying the effectiveness and applicability of the ACT link prediction method.

***Keywords*— Complex networks, network modelling, weighted location network, link prediction methods, COVID-19 pandemic.**

INTRODUCTION

Link prediction, one of the main topics in complex networks, strives for missing or unknown links in the network. Usually, it follows the structural features of the under-observed network, searches for nodes and links, and determines where the link is to be established [1]. The complex network theory has been used for devising the network structure, namely nodes and links. The contemporary literature witnesses the applications of link prediction in various complex network areas, such as communication networks, biological networks, dark or terrorist networks, disease networks, citation networks and to name a few [2-4]. Among them, link prediction tactics save time and costs for researchers in biological networks while performing toilsome tasks, for example, exploring the gene interactions and missing links between genes. Similarly, the potential of the link predictions can be observed in exploring the growth of social networks which evolve from time to time [5]. In the context of disease networks, link prediction has been utilized to uncover the hidden relationships among the diverse diseases, and to study the relationships between the disease and medications [6]. Moreover, many other examples are there to support the utilization of link predictions in discovering the drug repositioning, analyzing the connections between the covert or dark criminal groups, exposing the collaborators in the citation networks, and providing robust recommendations for the food and real-time shop-ping strategies best on one's preferences [7–10], [30].

In December 2019, COVID-19 appeared for the first time as a contagious virus with unknown etiological characteristics in Wuhan, province of China [11]. Due to speed and contagiousness and keeping public safety abreast, the World Health Organization (WHO) declared it a pandemic on March 11, 2020. As per WHO reports, the whole world has experienced the wreaking havoc of COVID-19 on almost all continents. Therefore, Pakistan is one of the COVID-19-affected countries, where the first COVID-19 cases appeared in Karachi and Islamabad on February 26, 2020. Since then, COVID-19 cases increased and become a major public health issue in Pakistan [12-14]. In this study, we focus on the two provinces of Pakistan namely Sindh, and Balochistan including the capital territory of Pakistan such as Islamabad. The growth of the cases within these provinces and Islamabad during the years 2020-21 is shown in Fig. 1. Initially, the COVID-19 cases growth steadily increased in March 2020, after this, the cases increased exponentially and more than 6,500 cases were recorded in April 2020. In both months May 2020 and June 2020, the cases jumped to 20,000 and 70,000, which is the peak point of cases record for the year 2020. The confirmed cases decreased in July 2020, and more than 45,000 cases were reported, as shown in Fig. 1. Certainly, these facts and figures point out the serious consequences of COVID-19 in the locations (districts), in which they have been recorded, thus, it has become a serious issue and requires not only to

[1-2 -3-4-5] University of Sindh, Jamshoro
Country: Pakistan,
Email: shauban@scholars.usindh.edu.pk

tackle with the disease within the COVID-19-affected locations but also a comprehensive understanding regarding the cases growth. Therefore, the occurrence of COVID-19 cases within the locations needs to be modelled and analyzed using scientific methods. However, we formalize this problem using a real COVID-19 dataset obtained from [15] and [16] as a weighted two-mode network comprising week-wise occurrences of cases and COVID-19-affected locations information. After this, the weighted two-mode network is projected onto the one-mode network to extract the location-to-location relationships using the standard network projection method. The resulting COVID-19 location network is further analyzed for the link predictions.

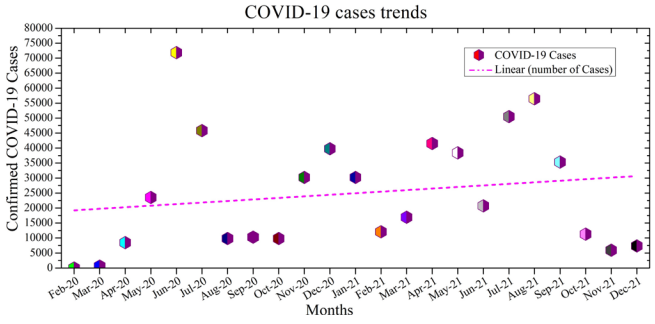

Figure 1: The COVID-19 cases growth from during 2020-21.

Numerous research has suggested a diversity of approaches for link prediction, categorized as local similarity measures and global proximity indexes. Though dissimilar models and approaches for link prediction in COVID-19 networks have been suggested, limited research has focused on the diffusion of the pandemic i.e. COVID-19. It formalizes the COVID-19 diffusion problem as a weighted two-mode network and empirically analyzes a weighted one-mode network using link prediction methods, considering both local and global perspectives of the network. We explore the link predictions of the COVID-19 location network from a local perspective and a global point of view as well. Hence, we apply five-link prediction methods from the local perspective of the network such as Adamic Adar (AA), Common Neighbor (CN), Leicht-Holme-Newman Local (LHN-L), Salton Index or cos index (SI), and Resource Allocation (RA), and four proximity methods from the global structure of the network namely average commute time (ACT) method, matrix forest index (MFI), Katz coefficient, and leicht-holme-newman global (LHN-G) [17-20]. After applying these nine link prediction methods, we compare their results, and based on their performances, the most suitable link prediction method is identified for the COVID-19 diffusion network. Additionally, four global link prediction methods for the COVID-19 diffusion network, such as the average commute time (ACT) method, matrix forest index (MFI), Katz coefficient, and leicht-holme-newman global (LHN-G), are evaluated to analyze the overall structural features of the network. The results obtained by these methods help to discover future links based on the existing topological information of the COVID-19 network. Hence, it provides the insights that a suitable link prediction method can be applied to watch out for future trends and the contributions of the nodes, either from a local or global perspective of the network. In other words, these insights can be viewed as the appropriate application of link prediction methods to find out the particular location nodes that are playing a key role as common neighbour location nodes with which other nodes are connected, resulting in link prediction in the diffusion of the pandemic from one location to another.

This research article is organized into six sections. Section I represents the introduction to the main field of link prediction, and a brief overview of the COVID-19 pandemic cases, especially from the perspective of COVID-19-affected locations in Sindh, Balochistan and Islamabad. Section II explores the related studies and recent approaches to link prediction. Section III comprises a detailed explanation of the modelling of the COVID-19 dataset as a two-mode network and the projection of a weighted location network. Section IV gives an insightful overview of the link prediction methods from the local and global perspective of the weighted location network. The evaluation results are compared and discussed in Section V. Finally, the concluding remarks and future work are provided in Section VI.

LITERATURE REVIEW

Link prediction in complex networks has attracted the attention of many researchers of the day to reveal hidden phenomena based on a particular problem. Mostly, the state-of-the-art link prediction methods comprise the scoring probability regarding the existence of the link between any two nodes [21]. This shows the similarity between the pair of nodes and based on the score i.e. low or high, whether the link exists. The higher the score the higher the similarity between the nodes, indicating higher chances of link presence. So, the similarity can be measured using the link prediction methods for different nodes in the network. Therefore, various research in the literature can be found regarding link prediction utilization for specific uses in observed datasets, based on the fields of interest [22-24], and [31]. In the following studies, the link prediction uses in disease transmission, collaboration networks, transportation networks, social networks, etc. are provided.

The authors in [8] analyzed the COVID-19 network for disease transmission in Korea based on the link prediction methods to predict missing links. They tried to find the missing links in the mesoscopic COVID-19 network for the structural and clustered behaviour of the network. The authors claimed that their results support the clustered network comprising both strong and weak links, also, the network looked tree-shaped when the weak links were excluded. In [25], the link prediction was analyzed on the topological

featured integrated network for the collaboration network. The collaboration between the scholars has been made through forecasting the links based on two factors namely temporality and weight. They claimed that these two factors could be important in the recommendation of the scholars` collaboration. Similarly, the work presented in [26] proposes link predictions for the recommendation of academic collaboration using a two-layer knowledge network. The authors focused on the network topology features and research field of interest. If the research field is somehow similar and the authors are connected topologically, the recommendation for that particular author is high. On the other hand, the link prediction-based study proposed in [27] reveals the missing links in the traffic flow to find out the centrality. The centrality analysis is made for the dynamic large-scale transportation networks to observe the traffic speed on the specific road.

Likewise, the authors in [28] modelled dynamic traffic flow by deploying the RFID (radio-frequency identification). They used RFID to monitor and analyze transportation flow over the roads forming a dynamic traffic flow network. The potential of the link prediction for privacy protection is suggested in [29], by devising a technique called similarity-based target privacy protection. In this method, the privacy concerns are addressed and the successful results. The recommendation engine-based system for finding the expert during the COVID-19 outbreak was proposed by [37] using link prediction. For this, many parameters namely structure of the network, auxiliary attributes, context attributes, co-occurrences, and weights have been considered to recommend the COVID-19 expert. Also, the link prediction-based study using the Twitter dataset is proposed by [38]. The dataset comprises the COVID-19 vaccine-related misinformation shared via tweets between the users. The dataset further analyses the true and false information of the COVID-19 vaccine and compares it with the classification-based techniques. The results claimed that the link prediction strategy bears better results than the classification-based methods. Another study related to the link prediction for COVID-19 transmission is presented in [39], where a link prediction-based biomedical drug discovery technique is utilized by text mining and artificial intelligence (AI) approach. The drugs proposed for COVID-19 have been analyzed and the link prediction is performed between the drugs whether the proposed drug is suitable. Text mining and AI are used to rank between the predicted links. The authors claimed that the link prediction-based results are useful for the re-proposing of drugs for such emergent diseases. In [40], the authors reviewed various link prediction methods proposed by different research scholars, especially for COVID-19 disease transmission. The main focus has been paid to the keywords specifically used for the COVID-19 research for connecting different research concepts. From this brief review, it is clear that link prediction-based research has equal importance from both theoretical and practical perspectives.

The COVID-19 Dataset modelling as a weighted two-mode Network

This section provides a detailed overview of the COVID-19 cases dataset modelling as a weighted two-mode (bipartite) network. The two-mode network comprises two disjoint sets of nodes; such as primary (top) nodes and secondary (bottom) nodes . The nodes in the primary set establish the connection with the nodes of the secondary sets and vice versa. No pair of nodes have a link in the same set of nodes [3]. There are various examples in the real-world network including actors-movie, authors-books, scientific collaboration networks, air transportation, etc. These networks are constructed using the graph theory. In the graph theory, the two-mode network is the triplet graph . Here, is a set of links between top nodes , and a set of bottom nodes [19].

To illustrate an unweighted two-mode network projection, two separate sets of nodes, namely the primary set of nodes 1, 2, 3, 4, 5, and 6, have a connection with nodes of the secondary set namely s, h, u, b, a, and n as shown in Fig. 2 (a). Here, node 1 has links with nodes S, and H, while node 2 has connections with nodes S, B, and N. Similarly, two nodes B and H (secondary set nodes) are linked with node 6 as depicted in Fig. 2 (a). To obtain the links between the nodes either primary nodes or secondary nodes, the two-mode network is projected onto the one-mode network by selecting either set of nodes. An example of the one-mode network projection by selecting a primary set of nodes is shown in Fig. 2 (b). The link is only established between the nodes if they have a common co-occurrence in the other set. Node 1 shares links with nodes 2, 5 and 6, due to co-occurrence with them in the other set. Node 5 is linked with nodes 1 and 4. Node 6 is linked with nodes 1, and 2, as illustrated in Fig. 2 (b). From the perspective of network analysis, it is essential to convert the network into one-mode network by selecting the desired set of nodes.

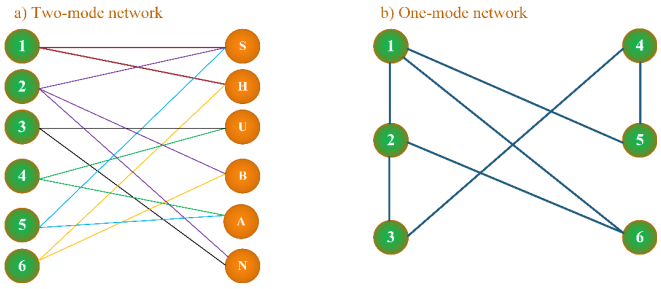

Figure 2. (a) Two-mode network illustration with two disjoint sets of nodes, (b) Projection as the one-mode network.

The complex network approach is a great motivation to model the COVID-19 dataset as the weighted two-mode COVID-19 network by picking useful attributes such as location (district)

and weeks. The location node establishes the link with the week node if the COVID-19 case is recorded in that week. Thus, the frequency of cases formalizes the weighted links between location and week nodes. An example of the weighted two-mode network is illustrated in Fig. 3 (a), while the projection of the weighted one-mode network is shown in Fig. 3 (b). However, we projected the network by selecting the primary set of nodes (locations) to construct a COVID-19 location network, for proper analysis of the case diffusion. It is interesting to note that, the dataset comprises case information from locations in Sindh, Balochistan, and Islamabad from February 2020 to December 2021 [15], [16]. Moreover, the weighted two-mode COVID-19 network has 58 locations, and 96-week nodes and 3105 links. We transform this network onto a weighted one-mode network by transformation using the standard weighted Newman network projection method to gain a COVID-19 location network.

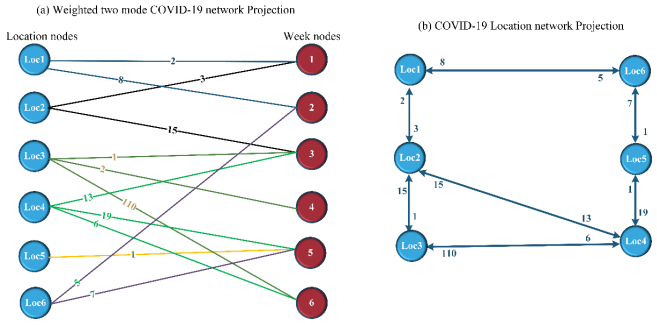

Figure 3. An illustration of a weighted two-mode COVID-19 network (a) Weighted location-weeks projection, (b) Projection as COVID-19 Location Network.

Local and Global topological charateristics based link prediction methods

This section provides a detailed overview of the state-of-the-art link prediction methods especially from both local and global structural information of the COVID-19 location network [2], [19], [20]. First, we analyze five local similarity methods from the local perspective of the network to observe the suitable applicability of these link prediction measures. Second, global link prediction measures are also applied using the topological information COVID-19 location network. In the following sub-sections, a brief definition of the graph theory and link prediction is provided, and then, the description of the five local similarity methods and four global similarity methods is given respectively.

A. Important definition

a) Graph theory

A network, denoted as A=(N,E) is a mathematical representation of nodes and edges. represents the set of nodes, denoted as$N=\{N_1,N_2,N_3,\ldots,N_n\}$, and represents the set of edges, denoted as edges . Networks can be categorized into two types: directed and undirected. In undirected networks, the edges do not have specific directions towards any nodes, whereas in directed networks, the edges have specific directions towards the nodes of interest and may have varying degrees denoted as in-degree and out-degree such as $\leq N_i,N_j\geq$ and $\leq N_j,N_i\geq$ [2-8], [16-20].

Link prediction

In the study of future links, we not only determine the similarity of nodes in the network but also gain insight into the underlying structural patterns of the network. It is significant to note that actual complex networks are not static and do not evolve solely based on the presence or absence of links [2], [24–29].

Let's consider a static network, denoted as $A=(N,E)$, at a specific time T. Here, P represents all the possible edges in the network, and any missing edges are denoted by $Q=E-P$. When we aim to predict missing or unobserved links in the network, the new future links can be represented as $T'(T'>T)$ for the set Q. In the case of an evolving network, the presence or absence of edges can be understood as a sequence of graph iterations over time, $A=(A_{T_1},A_{T_2},A_{T_3},\ldots,A_{T_T})$. The absence of links can be described from A_{T-1} to A_T, and consequently, predicting new edges depends on the probabilities of occurrences between the edges over time, such as A_{T+1}, according to A [9], [19].

A brief overview of the local link prediction methods

Adamic-Adar (AA) method

This link prediction method computes the common neighbours of each node and formalizes the weights of every node as inversely proportional to its degree. The nodes with fewer links are considered as the unique nodes which means that those nodes have trivial acquaintance, so the high weight is attached to them. The AA method was first time used to compute friends and neighbour nodes in the World Wide Web (WWW) network [18], [19]. The equation (1) expresses the functionality of the AA method.

$$S_{x,y}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{log k_z} \quad (1)$$

If z has a degree of 2, then the weight-assigning process in this method is given by the inverse log frequency of the degree.

3.2.2. Common Neighbors (CNs) link similarity method

This method is based on the perception of links between any two nodes which had shared links previously. The main idea was coined by Newman for the scientific collaboration network; in which the scientific collaboration between the authors was explored [32]. This means that any two scientists would like to contribute together if they had collaborated previously. Here the acquaintance regarding scientific collaboration matters for establishing future links between the scientists. Therefore, the CNs method is used to compute relations between the collaborators based on a similar test of literature, interest, or the subject matter. Additionally, this link similarity method quantifies future links between any two scientific collaborators [1], [32]. The CNs method is provided

in the following equation.

$$S_{x,y} = |\Gamma(x) \cap \Gamma(y)| \quad (2)$$

Jaccard Index (Jac)

The Jac method also known as the Jaccard coefficient can be defined as the number of intersections between common neighbours of given nodes divided by the union of the shared neighbors of any two nodes. The sharing of common neighbour nodes poses motivation for this link prediction method. It reaches its maximum, when each of the neighbour nodes is shared and none of the neighbour nodes is left to have any shared link, computed as Γ(x)=Γ(y). Moreover, the link prediction between any pair of nodes solely depends upon a shared quantity of neighbours. The Jac method is expressed in the following equation[3], [20], [32]:

$$S_{x,y} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (3)$$

Leicht-Holme-Newman Local (LHN-L)

The LHN-L method, named after its authors Leicht-Holme-Newman [17], determines the similarity among nodes in a network by considering the structural properties of the network, particularly in the structural corresponding and regular equivalent networks. The basic concept is that the links in the network themselves indicate the similarity between the nodes they connect. This method is represented by an equation that calculates the number of common neighbours of two nodes divided by the product of their degrees [19], [20]:

$$S_{x,y} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x \times k_y} \quad (4)$$

This method evaluates node similarity based purely on the network's structure. It allocates high similarity values to node pairs that shares many common neighbours and also identifies nodes with statistically unlikely connections in their neighborhood. Although the LHN-L technique bears some resemblance to the Salton index method, their computational properties vary.

Salton Index (SI)

In 1983, the SI index was introduced as Salton's formula in the work proposed by Salton. This method was purposefully used for the citation and co-citation networks analysis. In the literature, the SI method is also known as the cosine (cos) index. As far as the functionality of the SI method is concerned, this creates the adjacency matrix of the corresponding nodes in the observed network and then, calculates the cosine of the angle between columns of that adjacency matrix. Care is taken for the varying degrees of the nodes, so all common neighbours of a node are computed in the initial stage, and then, they are divided based on their geometric mean. The resulting score lies within the range of 0 and 1. The 0 score implies that two nodes have no common neighbour, while, the value 1 shows that any two nodes have exactly similar common neighbour nodes [16], [19]. The SI is expressed in the following equation, which can be interpreted as the inner product of the two nodes having common neighbours divided by the product of their lengths.

$$S_{x,y} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}} \quad (5)$$

GLOBAL LINK PREDICTION METHODS OVERVIEW

Average Commute Time (ACT) method

The average commute time (ACT) calculates the overall root length to reach from to . It means that the ACT score is the computation of the average number of roots visited by a random walker, thus, this link prediction method includes the random walk methodology to go from to . By this, it is the expected distance commute by a random walk from node to node and back to node [32].

$$S_{x,y} = \frac{1}{n(x,y)} = \frac{1}{m(x,y) + m(x,y)'} \quad (6)$$

Where shows the number of nodes, and represents the possible links between nodes. The m(x,y) indicates the overall number of links which facilitate a random walker to start from to reach at In ACT, the symmetry is gained by summing up two directional commute times. In other words, any pair of nodes are akin similar to each other if they are close to one another and possess a smaller commute time.

Katz index

In 1953, the Katz index was proposed by Leo Katz which aims to explore link similarity globally [35]. This method calculates all paths between any two nodes namely x and y . After the path computation, the highest weights are assigned to the nodes having optimal or smaller paths between them. The Katz index is given in the following equation [35]:

$$S_{xy} = \sum_{l=1}^{\infty} \beta^l |paths_{xy}^{<l>}| \quad (7)$$

The sum converges when is less than the reciprocal of the leading eigenvalue of the adjacency matrix. The findings are contingent on this condition being met, so the Katz index can be denoted as follows:

$$S = (I - \beta A)^{-1} - I \quad (8)$$

In equation (8), the Katz index has the adjacency matrix and identity matrix , while is the free parameter.

Matrix Forest Index (MFI)

In this link prediction method, two points are essential to be comprehended. (1) the number of spanning roots is evaluated between any two nodes say and , which belong to akin roots

at . All spanning roots considering all rooted forests from a node are given as (2). Therefore, the MFI coefficient is nothing but a ratio between (1) and (2). The study represented in [34] provides the details regarding the formalization of the MFI method. The MFI method is expressed in the equation (9):

$$S = (I - L)^{-1} \quad (9)$$

Leicht-Holme-Newman Global (LHN-G)

The local hitting time network centrality (LHN-G) is a measure of the proximity of a pair of nodes within a network. This means that two nodes will be measured near if their neighboring nodes are themselves near. LHN-G is an extended version of the Katz index. It computes all possible paths between any pair of nodes and assigns a weight based on the expected number of such paths with the same degree distribution in a random graph. The LHN-G matrices can be derived in proportion to equation (10).

Where the main eigenvalue of the adjacency matrix is represented by, and signifies the free parameter.

RESULTS

This section is dedicated to the overview of tools and techniques used for this study, and evaluation of the outcomes of each link prediction method namely five local link similarity methods and four global link prediction methods. However, we used RStudio and the link prediction package for the link prediction analysis, an open-source network analysis tool [36]. Before analysis, the COVID-19 dataset is loaded for the extraction of the case records using the attributes by using the ChainSys Smart Data platform [41]. Furthermore, we used DataZap to extract and transform this dataset and prepare for further network analysis. Moreover, DataZen and DataZense are utilized for the identification and rectification of data quality issues, such as missing values or inconsistencies, which are essential for accurate network modelling and link prediction [41]. We have analyzed the weighted COVID-19 location network comprising a total of 58 location nodes and 3304 links. From this network, we extracted 17% links and the residual 2743 links were used for the akin testing set of the study. Additionally, the evaluation and computation of the empirical results of the corresponding link prediction method are carried out by the area under curve (AUC) and receiver operating characteristics (ROC) precision criteria, which are visualized and analyzed by using visualization features of the dataZense. The precision values of the ROC lie between 0 and 1, the score closer to 1 is considered a robust performance of the applied link prediction method, however, a score below 0.5 or closer to 0 can be interpreted as the equal distribution of classification, which further, implies the weak application the measure.

First, we have applied five local link similarity methods based on the local information of the network while keeping the aforementioned network settings. These methods explore first-order akin proximity for each node to identify latent links between the nodes. The computational results are provided in Table 1 and Fig. 4 (a), where the AA similarity method has outperformed the other link prediction methods namely CN, Jac, LHN-L, and SI, by securing 60.2% accuracy in the COVID-19 location network.

Table 1. Local link prediction methods` score.

Local link similarity method	AUC precision value
Adamic Adar	0.60277
Common Neighbors	0.60095
Jaccard	0.60095
Leicht-Holme-Newman Local	0.59906
Salton Index	0.5991

While both methods CN and Jac have identical outcomes as shown in Fig. 4 (a). Interestingly, no local similarity has produced a link prediction score below 0.5, but based on the accuracy and performances, the AA method tops the result as given in Table 1 and Fig. 4 (a). Additionally, the TPR the true positive rate also denotes the sensitivity in the AUC outcomes which is always plotted at the y-axis, and the false positive rate (FPR) or 1-specificity is usually depicted at the x-axis of the AUC plot as shown in Fig. 4 (a) and (b).

Table 2 includes the performance scores obtained from the four-link prediction methods based on the global structural information of the COVID-19 location network. These four link similarity methods namely ACT, Katz, MFI and LHN-G endeavour for the hidden links of every node based on the whole topology of the observed network. Therefore, the Katz centrality and LHN-G have performed unsatisfactorily whose scores are below the 50% in the AUC as shown in Fig. 4 (b). Hence, the lowest outcomes of the LHN-G suggest that this method is inappropriate for this network because its computational results incorporate the link exploration strategy based on the same degree distribution in the random graphs while the weights are assigned after calculation of all paths between pairs of nodes. As the COVID-19 location does not bear the characteristics of the random graphs, so this LHN-G is suitable for this research.

Table 2. Global link prediction methods` score.

Global link similarity method	AUC precision value
Average Commute Time	0.60728

Katz coefficient	0.49765
Matrix Forest Index	0.5028
Leicht-Holme-Newman Global	0.40087

Also, the Katz centrality calculates possible paths between any given pair of nodes and assigns maximum weights to the node whose paths are shorter. In this network, only a few nodes have shorter paths which is why this index scores 49.8%, which is a clear indication of false positive rate calculation as shown in Fig. 4 (b). The result of the ACT link prediction measure is 60.7% which implies that this method has, comparatively, gained maximum accuracy and outperformed other methods as shown in Fig. 4 (b). Based on the topological analysis of the link prediction results, it is obvious that some methods have produced ameliorating results on the local information of the nodes in this network, but when it comes to the global information of the nodes only the ACT method has much better performance results as shown in Fig. 4 (a) and (b). Based on the empirical results in the context of the COVID-19 location network, we believe that the ACT method is suitable for this study to predict the links between location nodes. Also, it is better to evaluate link prediction based on the global topological information of the COVID-19 location network. In Fig. 5, the outcome of the ACT method is visualized, where the red-coloured links represent the results of the ACT measure between the location nodes.

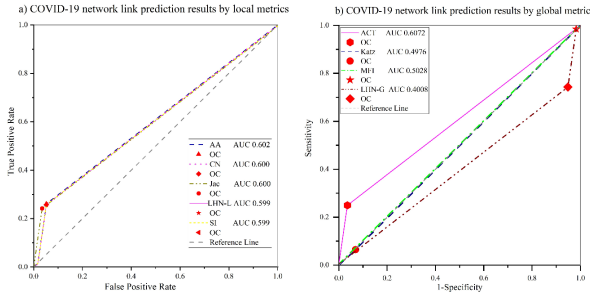

Figure 4. Results of link prediction methods: (a) five local link prediction methods, and (b) Four global link similarity methods.

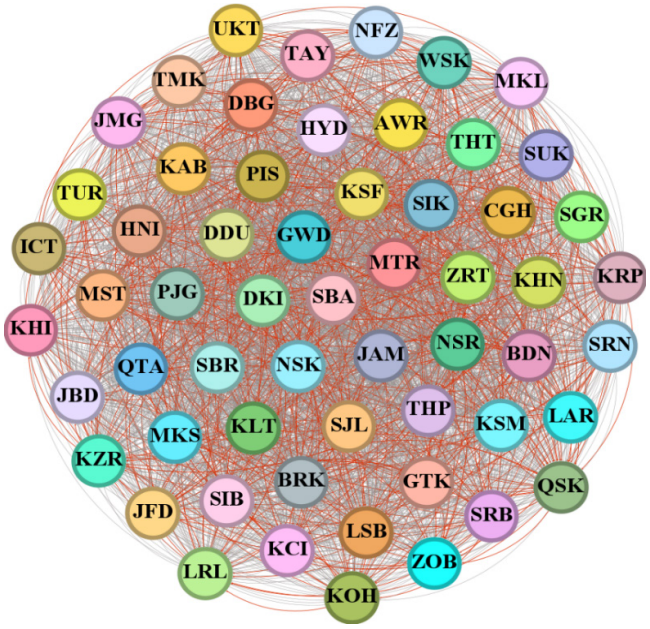

Figure 5. Visualization of predicted links between locations using the ACT method.

CONCLUSION

Link prediction in complex networks is one of the crucial problems, especially when it comes to addressing the COVID-19 disease network. Various endeavours have been made by research scholars to predict the transmission of COVID-19. However, there is a huge research gap which is yet to be filled in the literature from the topological perspective of the COVID-19 network. Therefore, this study explores and compares the link prediction methods based on the topological information of the COVID-19 location network. For this, we used complex network theory to formalize the COVID-19 dataset and modelled the weighted two-mode COVID-19 network comprising locations and weeks, then, this network is projected onto a weighted one-mode location network using the weighted Newman method. Moreover, we applied nine link prediction methods on the obtained one-mode location network to find out the suitable applicability of link prediction methods from local and global perspectives. Among them, five local link prediction methods namely Adamic-Adar index, Common Neighbors method, Jaccard index, Salton Index, and Leicht-Holme-Newman Local are evaluated using the structural information of the network. While, four global similarity methods such as average commute time (ACT), Katz index, matrix forest index, and Leicht-Holme-Newman global, are comparatively evaluated. The empirical results obtained by these nine link prediction methods suggest that the link prediction methods can better be evaluated from the global point of view of the COVID-19 network. However, the ACT similarity method with the highest score such as 60.7% has outperformed the remaining link prediction methods. We will explore link

prediction methods on the tripartite COVID-19 network in the future.

Funding: This research received no external funding.

Acknowledgements: We express our gratitude to the Faculty of Engineering and Technology (FET), University of Sindh, Jamshoro.

Conflicts of Interest: The authors declare no conflicts of interest.

References

L. Yao, L. Wang, L. Pan, and K. Yao, "Link Prediction Based on Common-Neighbors for Dynamic Social Network," in Procedia Computer Science, Elsevier B.V., 2016, pp. 82–89. doi: 10.1016/j.procs.2016.04.102.

H. Wu, C. Song, Y. Ge, and T. Ge, "Link Prediction on Complex Networks: An Experimental Survey," Data Science and Engineering, vol. 7, no. 3. Springer, pp. 253–278, Sep. 01, 2022. DOI: 10.1007/s41019-022-00188-2.

A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," Nature, vol. 453, no. 7191, pp. 98–101, May 2008, doi: 10.1038/nature06830.

D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," Journal of the American Society for Information Science and Technology, vol. 58, no. 7, pp. 1019–1031, May 2007, doi: 10.1002/asi.20591.

M. Newman, "The structure and function of complex networks," SIAM, pp. 167-256, 2003.

J. Menche et al., "Uncovering disease-disease relationships through the incomplete interactome," Science (1979), vol. 347, no. 6224, p. 841, Feb. 2015, doi: 10.1126/science.1257601..

S. Li, X. Song, H. Lu, L. Zeng, M. Shi, and F. Liu, "Friend recommendation for cross marketing in online brand community based on intelligent attention allocation link prediction algorithm," Expert Syst Appl, vol. 139, Jan. 2020, doi: 10.1016/j.eswa.2019.112839.

O. Kwon and H. H. Jo, "Clustering and link prediction for mesoscopic COVID-19 transmission networks in Republic of Korea," Chaos, vol. 33, no. 1, Jan. 2023, doi: 10.1063/5.0130386.

A. W. Mahesar, A. Waqas, N. Mehmood, A. Shah, and M. Ridza Wahiddin, "Analyzing the Weighted Dark Networks using Scale-Free Network Approach.".

S. Li et al., "Link Prediction Based on Heterogeneous Social Intimacy and Its Application in Social Influencer Integrated Marketing," Mathematics, vol. 11, no. 13, Jul. 2023, doi: 10.3390/math11133023.

H. Aghamirza Moghim Aliabadi et al., "COVID-19: A systematic review and update on prevention, diagnosis, and treatment," MedComm (Beijing), vol. 3, no. 1, pp. 1–42, 2022, doi: 10.1002/mco2.115.

O. J. Peter et al., "A New Mathematical Model of COVID-19 Using Real Data from Pakistan," Results Phys, p. 104098, 2021, doi: 10.1016/j.rinp.2021.104098.

I. Ahmad and S. M. Asad, "Predictions of coronavirus COVID-19 distinct cases in Pakistan through an artificial neural network," Epidemiol Infect, vol. 148, no. 4, pp. 1–10, 2020.

"WHO (COVID-19)." [Online]. Available: https://www.who.int/publications/m/item/covid-19-epidemiological-update-2023.

"Government of Sindh (COVID-19)." [Online]. Available: www.sindhhealth.gov.pk.

"Government of Balochistan (COVID-19)." [Online]. Available: www.health.balochistan.gov.pk.

E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," 2005.

L. Hamers et al., "Similarity Measures in Scientometric Research: The Jaccard Index Versus Salton's Cosine Formula," 1989.

M. Newman, Networks, vol. 1. Oxford University Press, 2018. doi: 10.1093/oso/9780198805090.001.0001.

M. Bojanowski and B. Chroł, "Proximity-based Methods for Link Prediction in Graphs with R package 'linkprediction,'" Ask: Research and Methods, vol. 29, no. 1, pp. 5–28, 2020, doi: 10.18061/ask.v29i1.0002.

K. ke Shang, M. Small, and W. sheng Yan, "Link direction for link prediction," Physica A: Statistical Mechanics and its Applications, vol. 469, pp. 767–776, Mar. 2017, doi: 10.1016/j.physa.2016.11.129.

L. Lu and T. Zhou, "Link Prediction in Complex Networks: A Survey," Oct. 2010, doi: 10.1016/j.physa.2010.11.027.

F. Aziz, L. T. Slater, L. Bravo-Merodio, A. Acharjee, and G. V. Gkoutos, "Link prediction in complex network using information flow," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-41476-9.

S. D. Ghiassian, J. Menche, and A. L. Barabási, "A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome," PLoS Comput Biol, vol. 11, no. 4, pp. 1–21, 2021, doi: 10.1371/journal.pcbi.1004120.

E. Butun and M. Kaya, "Predicting Citation Count of Scientists as a Link Prediction Problem," IEEE Trans Cybern, vol. 50, no. 10, pp. 4518–4529, Oct. 2020, doi: 10.1109/TCYB.2019.2900495.

W. Liu, H. Duan, Z. Li, J. Liu, H. Huo, and T. Fang, "Entity Representation Learning with Multimodal Neighbors for Link Prediction in Knowledge Graph," in 2021 7th International Conference on Computer and Communications, ICCC 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1628–1634. doi: 10.1109/ICCC54389.2021.9674496.

C. Lee and Y. Yoon, "Context-Aware Link Embedding with Reachability and Flow Centrality Analysis for Accurate Speed Prediction for Large-Scale Traffic Networks," Electronics (Basel), vol. 9, no. 11, p. 1800, Oct. 2020, doi: 10.3390/electronics9111800.

Y. Liu et al., "Dynamic traffic demand uncertainty prediction using radio-frequency identification data and link volume data," IET Intelligent Transport Systems, vol. 13, no. 8, pp. 1309–1317, Aug. 2019, doi: 10.1049/iet-its.2018.5317.

S. Yu, M. Zhao, C.Fu, et al., Target defense against link-prediction-based attacks via evolutionary perturbations, IEEE Trans. Knowl. Data Eng. 33, 2021, 754–767, https://doi.org/10.1109/TKDE.2019.2933833

F. Guo, W. Zhou, Z. Wang, C. Ju, S. Ji, and Q. Lu, "A link prediction method based on topological nearest-neighbors similarity in directed networks," J Comput Sci, vol. 69, May 2023, doi: 10.1016/j.jocs.2023.102002.

S. Mallek, I. Boukhris, Z. Elouedi, and E. Lefèvre, "Evidential link prediction in social networks based on structural and social information," J Comput Sci, vol. 30, pp.
98–107, Jan. 2019, doi: 10.1016/j.jocs.2018.11.009.

P. Jaccard "The Distribution of the Flora in the Alpine Zone." New Phytologist 11 (2): 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x 1912.

A. K. Chandra, P. Raghavan, W. L. Muzzo~, R. Smolensky, and P. Tiwari, "the electrical resistance of a graph captures its commute and cover times," 1996.

P. Y. Chebotarev and E. V Shamis, "the matrix-forest theorem and measuring relations in small social groups 1," 1997.

L. Katz, "A New Status Index Derived from Sociometric Analysis." Psychometrika 18 (1): 39–43. 1953. https://doi.org/10.1007/BF02289026

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com.

H. Wang and Z. C. Le, "Expert recommendations based on link prediction during the COVID-19 outbreak," Scientometrics, vol. 126, no. 6, pp. 4639–4658, Jun. 2021, doi: 10.1007/s11192-021-03893-3.

M. A. Weinzierl and S. M. Harabagiu, "Automatic detection of COVID-19 vaccine misinformation with graph link prediction," J Biomed Inform, vol. 124, Dec. 2021, doi: 10.1016/j.jbi.2021.103955.

K. McCoy et al., "Biomedical text link prediction for drug discovery: A case study with covid-19," Pharmaceutics, vol. 13, no. 6, Jun. 2021, doi: 10.3390/pharmaceutics13060794.

X. Wang et al., "Dynamic Link Prediction for Discovery of New Impactful COVID-19 Research Approaches.," IEEE J Biomed Health Inform, vol. 26, no. 12, pp. 5883–5894, Dec. 2022, doi: 10.1109/JBHI.2022.3212863.

ChainSys Smart Data Platform: Data Migration by ChainSys Corporation. Available at: http://www.chainsys.com/.