scientific reports

OPEN

Path-based extensions of local link prediction methods for complex networks

Furqan Aziz[1,2,3,4]✉, Haji Gul[5], Irfan Uddin[6] & Georgios V. Gkoutos[1,2,3,4,7,8,9]

Link prediction in a complex network is a problem of fundamental interest in network science and has attracted increasing attention in recent years. It aims to predict missing (or future) links between two entities in a complex system that are not already connected. Among existing methods, local similarity indices are most popular that take into account the information of common neighbours to estimate the likelihood of existence of a connection between two nodes. In this paper, we propose global and quasi-local extensions of some commonly used local similarity indices. We have performed extensive numerical simulations on publicly available datasets from diverse domains demonstrating that the proposed extensions not only give superior performance, when compared to their respective local indices, but also outperform some of the current, state-of-the-art, local and global link-prediction methods.

The study of complex networks is a relatively new, but rapidly growing field of interdisciplinary scientific research that aims at modelling and analysing real-world complex systems[1,2]. The interest in network science has emerged from the empirical study of networks that are obtained as a result of modelling real-world complex systems[3]. Example of such networks include ecological networks[4], social networks[5], transportation networks[6], and biological networks[7]. A complex network provides a convenient way of representing a complex system where nodes of the network represent entities of the complex system and links represent interactions between entities. However, the process of acquiring networks from complex systems may introduce noise which can result in missing links in a network[8]. To tackle this issue, link prediction has attracted the attention of researchers from a diverse scientific disciplines. It aims to estimate the likelihood of the existence of a link between disconnected nodes based on node attributes, neighbour information, and network structures. The problem is of both theoretical interest and has broad applications. Some of its applications include friend recommendation in social networks such as Facebook[9], predicting interactions between proteins[10], product recommendation to users[11], and drug target interaction prediction[12].

Motivated by the practical significance of link prediction, numerous link prediction algorithms have been proposed for unweighted networks. Among the various categories of link prediction algorithms, the most popular ones are the structural based similarity indices that are based solely on the structural properties of the underlying complex network. One of the most commonly used structural based similarity indices is the common neighbour[13], which measures the similarity between two nodes by counting the number of common neighbouring nodes. This method, however, does not take into account the degree information of the two nodes or their common neighbours. To overcome this problem, many variations of common neighbours have been proposed. For example, Adamic-Adar[14] and resource allocation[15], that penalise the high-degree common neighbour and perform better than common neighbour in most practical situations. Other local indices include preferential attachment[16] Jaccard coefficient[17], Sørensen index[18], and Salton index[19]. Cannistraci et al.[20] have combined a local link prediction algorithm with a local community structure to define a new set of link prediction indices, namely the CAR-based indices, demonstrating the application of CAR-based indices in predicting links in brain connectomes. However, these similarity indices are defined for an unweighted and undirected network. In order

[1]Centre for Computational Biology, University of Birmingham, Birmingham B15 2TT, UK. [2]College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT, UK. [3]Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, Birmingham B15 2TT, UK. [4]MRC Health Data Research UK (HDR), Midlands, UK. [5]City University of Science and Technology, Peshawar, Pakistan. [6]Kohat University of Science and Technology, Kohat, Pakistan. [7]NIHR Experimental Cancer Medicine Centre, Birmingham B15 2TT, UK. [8]NIHR Surgical Reconstruction and Microbiology Research Centre, Birmingham B15 2TT, UK. [9]NIHR Biomedical Research Centre, Birmingham B15 2TT, UK. ✉email: f.aziz@bham.ac.uk

to predict links in weighted networks, Zhao et. al.[21] have extended the unweighted similarity indices to weighted ones that can, not only, predict the missing link in a network but also estimate the weight of the missing link. Furthermore, Ghorbanzadeh et al.[22] have defined a measure, based on common neighbour, that can be applied to directed networks.

The structure-based similarity indices discussed so far are also called local similarity indices (or node dependent similarity indices) as they are based on the information of the immediate neighbours of the two query nodes. An alternative approach to link prediction is to consider the overall structure of the network. Such type of similarity indices are called global similarity indices. Global methods are also sometimes called path-dependent similarity indices since they are generally based on path information between nodes. One example is the Katz index[23] which considers the set of all paths between the two query nodes. Recently, Yant et al.[24] and Ahmad et al.[25] have proposed methods that take advantages of both the local and the global properties of a network by combining common neighbour and distance information to estimate the likelihood of the formation of link between two nodes. Other global similarity indices include hitting (or commute) time[26], Matrix-Forest Index[27], Linear Optimization[28], SimRank[29], and similarity-popularity based methods[30].

To provide a tradeoff between accuracy and computational time, quasi-local indices[31,32] are introduced that consider paths with wider horizon. To that end, Lü et al[33] have defined local path (LP) index that considers local paths of shorter lengths between the two query nodes. They have empirically shown that the LP index performs better when compared to common neighbour and gives comparable performance to Katz index[23]. They have also demonstrated that LP has low computational time as compared to Katz index. Just like common neighbour, LP index suffers from the problem that it does not take into account the degree information of the two nodes and the nodes on the local paths. In this paper we propose novel global and quasi-local measures that extend the existing local measures and can be used to predict missing link with higher accuracy. The idea is to use the node information on local paths. We commence by providing vectorised implementation of several local structural based similarity indices. For each of those similarity indices, we propose their global and quasi-local extensions. In the experimental evaluation section, we compare the local indices to their global and quasi-local extensions and empirically demonstrate that the global (and quasi-local) indices usually give better performance (but have higher computational cost) as compared to their corresponding local indices. In particular, we show that the difference in performance is significant when the noise in the data is high. We also demonstrate that some of the global indices introduce in this paper outperform the Katz index.

Overview of link prediction

In this section, we introduce some of the state-of-the-art local link prediction indices. We also present the local path index[33] and the Katz index[23], considered as the corresponding quasi-local and global extensions of the common neighbour index[13]. A *network* $G = (V, E)$ is defined as a set V of *nodes* and a set E of *links*, where $E \subseteq V \times V$. A network is *directed*, if the links it contains are directed, and *undirected* if the links it contains have no direction. A network is termed *weighted*, if the links it contained are assigned different weights. Otherwise, it is termed *unweighted*. A *simple* network is a network where multiple links between nodes as well as links between same nodes (self-loops) are not allowed. In this work, we considered simple, undirected and unweighted networks. The *degree* of a node $v \in V$, represents the number of connections that a node has with other network nodes. We denote the degree of the node v by $|\Gamma(v)|$, where $\Gamma(v)$ represents the set of all neighbours of v. A *walk* w in a network $G = (V, E)$ is defined as a sequence of alternating nodes and edges $v_0, e_1, v_1, e_2, v_2, ..., e_k, v_k$ where $v_i \in V$ and $e_i = (v_{i-1}, v_i)$. This walk has *length* k, where k is the number of links in the walk. An *adjacency matrix* provides a compact way of representing a network $G = (V, E)$. It is a square matrix of size $|V| \times |V|$, whose (u, v) th entry is 1 if u and v are linked and 0 otherwise. The (u, v)th entry of the k^{th} power of the adjacency matrix , $(A^k)_{uv}$, represents the number of walks of length k from u to v.

We now present some of the commonly used local similarity indices. In the next section we present their global and quasi-local extensions.

Common Neighbour (CN)[13]	A common neighbour is a simple but effective measure based on the number of shared neighbours between two nodes. In other words, two nodes are more likely to have a link if they share many common neighbours.
Adamic Adar (AA)[14]	AA is a variant of the CN that assigns more weight to neighbours with lower degrees. It captures the notion that neighbours with fewer links are more influential in facilitating the formation of future interactions.
Resource Allocation (RA)[15]	RA is defined in a similar way to AA. However, compared to AA, it assigns a lower score to the node pairs whose common neighbours have a high node degree. The only difference in the mathematical representation of RA and AA indices is that the later takes the logarithm of the denominator.
Sørensen (SO)[18]	Sørensen Index was proposed to establish equal amplitude groups in plant sociology based on the similarity of species. It is also used to calculate similarities of nodes in complex networks. It is determined by common neighbours of node pairs relative to their sum of individual degrees.
Salton (SA)[19]	This measure, proposed by Salton and McGill, is based on cosine angle between rows of adjacency matrix having query nodes u and v. This index is also called Salton Cosine Index.
Leicht Holme Newman (LHN)[34]	This measure gives higher score for node pairs having more common neighbours in proportion to their expected number of neighbours.

Local index	Matrix form	Global extension	Quasi-local extension
$CN(u,v) = \lvert\Gamma(u) \cap \Gamma(v)\rvert$	A^2	$(I - \beta A)^{-1} - I$	$A^2 + \beta A^3$
$RA(u,v) = \sum_{w \in \{\Gamma(u) \cap \Gamma(v)\}} \frac{1}{\lvert\Gamma(w)\rvert}$	$AD^{-1}A$	$A\left(I - \beta D^{-1}A\right)^{-1} - A$	$AD^{-1}A + \beta AD^{-1}AD^{-1}A$
$AA(u,v) = \sum_{w \in \{\Gamma(u) \cap \Gamma(v)\}} \frac{1}{\log\lvert\Gamma(w)\rvert}$	$A\left(\log D\right)^{-1}A$	$A\left(I - \beta(\log D)^{-1}A\right)^{-1} - A$	$A\left(\log D\right)^{-1}A + \beta A\left(\log D\right)^{-1}A\left(\log D\right)^{-1}A$
$SO(u,v) = \frac{2\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\lvert\Gamma(u)\rvert + \lvert\Gamma(v)\rvert}$	$2\left(D\left(A^2\right)_{ij}^{-1} + \left(A^2\right)_{ij}^{-1}D\right)_{ij}^{-1}$	$2\left(D\left((I - \beta A)^{-1} - I\right)_{ij}^{-1} + \left((I - \beta A)^{-1} - I\right)_{ij}^{-1}D\right)_{ij}^{-1}$	$2\left(D\left(A^2 + \beta A^3\right)_{ij}^{-1} + \left(A^2 + \beta A^3\right)_{ij}^{-1}D\right)_{ij}^{-1}$
$SA(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\sqrt{\lvert\Gamma(u)\rvert \times \lvert\Gamma(v)\rvert}}$	$D^{-\frac{1}{2}}A^2D^{-\frac{1}{2}}$	$D^{-\frac{1}{2}}\left((I - \beta A)^{-1} - I\right)D^{-\frac{1}{2}}$	$D^{-\frac{1}{2}}A^2D^{-\frac{1}{2}} + \beta D^{-\frac{1}{2}}A^3D^{-\frac{1}{2}}$
$LHN(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\lvert\Gamma(u)\rvert \times \lvert\Gamma(v)\rvert}$	$D^{-1}A^2D^{-1}$	$D^{-1}\left((I - \beta A)^{-1} - I\right)D^{-1}$	$D^{-1}A^2D^{-1} + \beta D^{-1}A^3D^{-1}$
$HP(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\min(\lvert\Gamma(u)\rvert, \lvert\Gamma(v)\rvert)}$	$\left(\min\left(D\left(A^2\right)_{ij}^{-1}, \left(A^2\right)_{ij}^{-1}D\right)\right)_{ij}^{-1}$	$\left(\min\left(D\left((I - \beta A)^{-1} - I\right)_{ij}^{-1}, \left((I - \beta A)^{-1} - I\right)_{ij}^{-1}D\right)\right)_{ij}^{-1}$	$\left(\min\left(D\left(A^2 + \beta A^3\right)_{ij}^{-1}, \left(A^2 + \beta A^3\right)_{ij}^{-1}D\right)\right)_{ij}^{-1}$
$HD(u,v) = \frac{\lvert\Gamma(u) \cap \Gamma(v)\rvert}{\max(\lvert\Gamma(u)\rvert, \lvert\Gamma(v)\rvert)}$	$\left(\max\left(D\left(A^2\right)_{ij}^{-1}, \left(A^2\right)_{ij}^{-1}D\right)\right)_{ij}^{-1}$	$\left(\max\left(D\left((I - \beta A)^{-1} - I\right)_{ij}^{-1}, \left((I - \beta A)^{-1} - I\right)_{ij}^{-1}D\right)\right)_{ij}^{-1}$	$\left(\max\left(D\left(A^2 + \beta A^3\right)_{ij}^{-1}, \left(A^2 + \beta A^3\right)_{ij}^{-1}D\right)\right)_{ij}^{-1}$

Table 1. Local, Quasi-local, and Global Similarity indices. Here A represents the adjacency matrix of the network, I is the identity matrix with size equal to the size of the matrix A, and D represents the diagonal degree matrix whose ith diagonal element is the degree of the ith node of the graph. Furthermore, A^{-1} represents the inverse of the matrix A, while A_{ij}^{-1} represents the element-wise inverse operation.

Hub Promoted(HP)[35] This index is proposed for quantifying the topological overlap of pairs of substrates in metabolic networks. Here, node pairs adjacent to hub nodes are assigned higher scores.

Hub Depressed(HD)[15] This measure is similar to HP measure but it is affected by higher degree nodes. Any node which has high degree is penalised by this measure.

Table 1 (first column) gives mathematical formulation for each of these local similarity indices. Although local similarity measures can be computed efficiently and perform relatively well, their accuracy cannot generally reach to methods which are based on global information. One typical example of global similarity index is the Katz index which is defined as follows:

Katz Index ($Katz_\beta$)[23] This index computes the similarity scores, based on the set of paths of different lengths, between two query nodes. The paths are exponentially damped by the length of the path so to assign more weight to shorter paths. Mathematically, this index is defined as follows:

$$Katz_{uv} = \sum_{l=1}^{\infty} \beta^l \cdot \left[\text{path}_{uv}^{\langle l \rangle}\right], \tag{1}$$

where $0 < \beta < 1$ is the parameter that controls the weight of the paths of different lengths. The similarity matrix S, whose (u, v)th entry equals $Katz_{uv}$, can also be computed as $(I - \beta A)^{-1} - I$, where I represents the identity matrix of size $|V|$.

Since this method is based on the topology of the whole network, therefore it generally outperforms local similarity indices such as CN. The difference is significant when the network is sparse, or when the network has many missing links. However, global indices generally have higher computational time when compared to local indices. In order to provide a good trade-off between accuracy and Complexity, Lü et al.[33] have introduced path index, which is defined as follows:

Local Paths (LP)[33] This index considers locals paths of shorter length and is generally computed as $A^2 + \beta A^3$, where A is the adjacency matrix of the network. As with Katz index, $\beta < 1$ is set to a small value so that shorter paths get more weights.

Lü et al.[33] have empirically demonstrated that LP index performs remarkably better than the simple CN index. They have also demonstrated that both Katz and LP indices generally give comparable performances, while the computation time of LP is considerably low than Katz index. Note that CN index, LP index, and Katz index have unified form as all the three indices can be expressed using Eq. (1), where for CN $l = 2$, for LP $l = 2, 3$, and for Katz $l = 1, 2, ..., \infty$. Therefore, both LP and Katz indices can be considered as extensions of Common neighbours to local paths.

Methods

In this section, we define the global and the quasi-local extensions of some of the most widely used local similarity indices. For each local similarity index, we first give its vectorised implementation. Our goal is to define global indices similar to Katz index that reduce to local indices for smaller values. Additionally, we also propose quasi-local measures of these indices. In the experimental evaluation section of this paper, we demonstrate that the global and quasi-local indices of RA and AA indices generally outperform all the other indices on most of the datasets. However, for the sake of completeness and experiments, we also define the global and the quasi-local extensions for the remaining local similarity indices. These similarity indices are summarised in Table 1, where the first column gives the mathematical definition of the local similarity index and the second column provides a matrix representation of the local index. The global and the quasi-local extension of the respective

local similarity index are also given in the third and fourth column of the table respectively. In the remaining of this section, we briefly discuss how the global and quasi-local extensions are obtained.

We commence by defining the global and quasi-local extensions of RA index. This index assigns more weight to the less connected neighbour. It can be shown that the RA index can be expressed in the form of matrix multiplication as $AD^{-1}A$, where D is the diagonal degree matrix whose ith diagonal element is the degree of the ith node. In order to define a global extension of the RA index, we not only consider the local paths between two nodes, but also consider the degrees of the nodes along the local paths. The global RA index is then defined as follows:

$$RA_G(u,v) = \beta AD^{-1}A + \beta^2 AD^{-1}AD^{-1}A + \cdots = A\big(I - \beta D^{-1}A\big)^{-1} - A. \tag{2}$$

The global RA index, defined above, can be interpreted as follows: To predict the existence of a link between two nodes u and v, the global RA index considers all simple paths from u to v. Moreover, all the nodes on the paths contribute to the computation of the similarity index, where less connected nodes are assigned higher scores. As with Katz index, a damping parameter β is used that assigns more weights to shorter paths. We also define a quasi-local extension of RA index that considers only the first two terms of the global RA index. Mathematically, this can be expressed as $RA_{QL}(u,v) = AD^{-1}A + \beta AD^{-1}AD^{-1}A$.

We define the global and quasi-local extensions of the AA in a similar way that we defined for the RA index. Note that, as mentioned earlier, the only difference between the AA and RA indices is that AA produces higher score values than RA for node pairs whose common neighbours have high node degree. This is achieved by taking the log of degrees of the common neighbours. The AA index can be expressed in the matrix form as $A(\log D)^{-1}A$, where $\log D$ is the diagonal degree matrix whose ith diagonal element is the log of the degree of the ith node.

Next, we define the global and quasi-local extensions of the remaining five similarity indices, i.e., SO, SA, LHN, HP, and HD. Note that all these five indices can be considered as modified versions of the CN index, that not only consider the the degrees of the common neighbours of the nodes u and v, but also take into account the degrees of the nodes u and v in one way or the other. Here we discuss the global and quasi-local extensions of SO index. The remaining indices can be extended in a similar way. By applications of simple matrix algebra, it can be shown that SO can be expressed in matrix form as follows:

$$SO(u,v) = \frac{2|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u)| + |\Gamma(v)|} = 2\Big(D\big(A^2\big)_{ij}^{-1} + \big(A^2\big)_{ij}^{-1}D\Big)_{ij}^{-1}, \tag{3}$$

where $(A)_{i,j}^{-1}$ represents the element-wise inverse of the matrix A. Using matrix representation, we provide a straightforward global extension of SO by considering all the local paths between the nodes u and v, instead of only common neighbours. Therefore, we define the global extension of the SO as follows:

$$SO_G(u,v) = 2\Big(D\big((I-\beta A)^{-1} - I\big)_{ij}^{-1} + \big((I-\beta A)^{-1} - I\big)_{ij}^{-1}D\Big)_{ij}^{-1}. \tag{4}$$

For the quasi-local extension of SO, we only consider paths up to length two. This index is defined as follows:

$$SO_{QL}(u,v) = 2\Big(D\big(A^2 + \beta A^3\big)_{ij}^{-1} + \big(A^2 + \beta A^3\big)_{ij}^{-1}D\Big)_{ij}^{-1} \tag{5}$$

. The global and quasi-local indices of the remaining four local similarity indices (i.e., SA, LHN, HP and HD) can be defined in a similar way as we defined for the SO index. This is because the only difference between the SO index and each of these indices is the denominator has a different form. These extensions are reported in Table 1.

We conclude this section by discussing the time complexities of the global and quasi-local extensions of the the similarity indices discussed in this paper. We note that the key operations performed, when computing the global and the quasi-local extensions, are the two matrix operations, namely matrix multiplication and the matrix inversion. Both of these operations require cubic time in N, the number of nodes in the network. Therefore, the running time of both the quasi-local, as well as the global index, is bounded by $O(N^3)$. However, in practice, quasi-local index can be performed much faster as it takes into account only the information about the neighbours and the neighbours of the neighbours. Lü et al[33] have demonstrated that the quasi-local extension provides a comparable performance to Katz index and it also requires less CPU time and memory space than Katz index. Furthermore, the computation of the global index requires a matrix inversion that is computationally very expensive (and can be unstable for large networks). In our experimental evaluation, we demonstrate that the quasi-local extensions of other local indices also result in a competitive performance, when compared to respective global extensions. Therefore, although the global extensions are effective for small and average-size networks, the quasi-local extensions are strong candidates for potential practical applications for large networks.

Results and discussion

In this section we present the experimental evaluation results of the proposed methods on real-world datasets and compare the performance of local similarity indices with their global and quasi-local extensions.

Datasets. To evaluate the performance of proposed and alternate methods, we have used various publicly available datasets from diverse domains, most of which are downloaded from KONECT[36]. A brief introduction of each of these datasets is given below. A summary of their topological properties is also presented in Table 2.

Datasets	$\|V\|$	$\|E\|$	C	$\langle k \rangle$	$\langle d \rangle$	ρ	H
Karate[37]	34	78	0.588	4.588	1.204	0.139	7.769
US Roads[6]	49	107	0.507	4.367	2.082	0.091	4.935
Dolphin[38]	62	159	0.303	5.129	1.678	0.084	6.805
Train Bombing[39]	64	243	0.711	7.594	1.345	0.121	12.597
Neurons[40]	279	2287	0.337	16.394	1.218	0.059	25.916
E. coli[41]	329	456	0.222	2.772	2.421	0.008	12.314
Netscience[42]	379	914	0.798	4.823	3.021	0.013	8.021
Infectious[43]	410	17298	0.467	84.38	1.815	0.206	2.992
Metabolic[44]	453	4596	0.782	20.291	1.332	0.045	17.903
US Air[45]	500	2980	0.726	11.92	1.496	0.024	53.785
Email[46]	1133	5451	0.254	9.622	1.803	0.009	18.688
Yeast[47]	2375	11693	0.388	9.847	2.548	0.004	34.223

Table 2. Topological properties of the networks used in experiments. $|V|$ and $|E|$ are the number of nodes and links respectively. C is the clustering coefficient. $\langle k \rangle$ and $\langle d \rangle$ are average degree and average path length. Finally ρ denotes the density of the network while H is the heterogeneity defined as $H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$.

Karate[37] A dataset (also known as the Zachary karate club) consisted of a karate club members, collected in 1977. The nodes of this network represent club members while the links represent ties between two members.

US Roads[6] This network consists of 49 nodes and 107 links. The nodes represent the 48 contiguous states and the District of Columbia (Washington D.C.) of the USA and the links represent drivable roads between two nodes. This network includes all the states except the states of Alaska and Hawaii, which are not connected by land with the other states.

Dolphins[38] A social network of bottlenose dolphins. The dataset consists of a set of links, each link representing frequent associations between dolphins.

Train bombing[39] A dataset containing a list of 64 suspected terrorists who were believed to be involved in the Madrid train bombing on March 11, 2004. The nodes of the network represent the suspected terrorists, while the links between terrorists are established if the are friends or have co-participated in training camps.

Caenorhabditis elegans (neurons)[40] This dataset consists of 279 neurons and 2990 links, including 1584 unidirected and 1406 bidirected links. In our experiments, the direction was ignored resulting in a total of 2287 links.

E. coli[41] A protein-protein interaction network of Escherichia coli that originally consisted of 424 nodes and 519 connections. We have considered the largest connected component (LCC) of the network having 329 nodes and 456 links.

Network Science[42] A network of 1461 scientists working on network theory. In this network, the nodes represent scientists and a link is established between two scientists, if they are co-authors on the same paper. Similarly to the E.Coli dataset, we have only considered the LCC with 379 nodes and 914 links.

Infectious[43] A network of 410 individuals who have attended exhibition, "infectious: stay away" in 2009 in Dublin. Here node represent individuals and a link represent face-to-face contact that was active for at least 20 seconds.

Caenorhabditis elegans (metabolic)[44] This is the undirected metabolic network of the roundworm Caenorhabditis elegans, where nodes represent metabolites (e.g., proteins), and links represent the physical interactions between them.

US Air[45] A network of direct flights among 500 US airports. The nodes represent airports and two nodes are connected if there is a direct flight between the corresponding airports.

Email[46] An email communication network between individuals at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain. Here the nodes represent individuals and a link is established between two individuals, if one of the two users has sent at least one email to the other user. The direction and the frequency of the emails are ignored.

Yeast[47] A yeast protein-protein interaction network, where each protein is a node and the interaction between them is represented by a link.

Evaluation metric. In order to assess and compare the performances of the local similarity indices and their corresponding global and quasi local extensions, we computed their accuracies using the area under the receiver operating characteristic metric *AUC*[48]. Consider a simple network $G = (V, E)$. Her we refer to the set E as the set of observed links. Let E' represents the set of nonexistent links in the network. In other words, $E' = \{(u, v) : u, v \in V, (u, v) \notin E\}$. We note that, if U represents the set of all possible $\frac{|V|(|V|-1)}{2}$ edges that G can have, then $E' = U \setminus E$. In order to evaluate the prediction algorithm's performance, the set of observed links, E, is randomly divided into two disjoint sets, namely, a training set E^T and a probe set E^P. Since E^T and E^P are disjoint, the two sets form a partition of the set E, i.e., $E = E^T \cup E^P$, and $E^T \cap E^P = \phi$. The information in E^T is used to predict missing links while the information in E^P is used to evaluate the performance of the prediction algorithm. To estimate the accuracy of the prediction algorithm, we compute their AUC values. In our case the metric AUC can be interpreted as the probability that a randomly chosen link in E^P gets higher score than a randomly chosen link in E'. If among n independent comparisons, n' is the number of times a missing link has higher score than a non-existent link, and n'' is the number of times a missing link and a nonexistent link having the same score, then the *AUC* is defined as

$$AUC = \frac{n' + 0.5n''}{n}.$$

We note that the value of AUC should be about 0.5, if all the link scores are randomly generated according to an independent identical distribution. Therefore, a value greater than 0.5 indicates how well the prediction algorithm performs when compared to pure chance.

Experimental results. In order to assess the performance of the global and quasi-local similarity indices and compare it with the local similarity indices' performance, we have randomly divided the set of observed links of the network, E, into two sets, namely, a training set E^T and a probe set E^P. In our first experiment, 90% of the observed links were contained in the training set while the remaining 10% were used for the probe set. The performance of all the similarity indices were evaluated using the same training and probe sets. For the quasi-local and the global indices, the value of parameter β was set to 0.001. The experiment was repeated 100 times, and in each run an independent random sampling of the observed links was performed. The average accuracies (along with standard deviations) of all the 100 runs are reported in Table 3. Figure 1 presents a visual representation of these results.

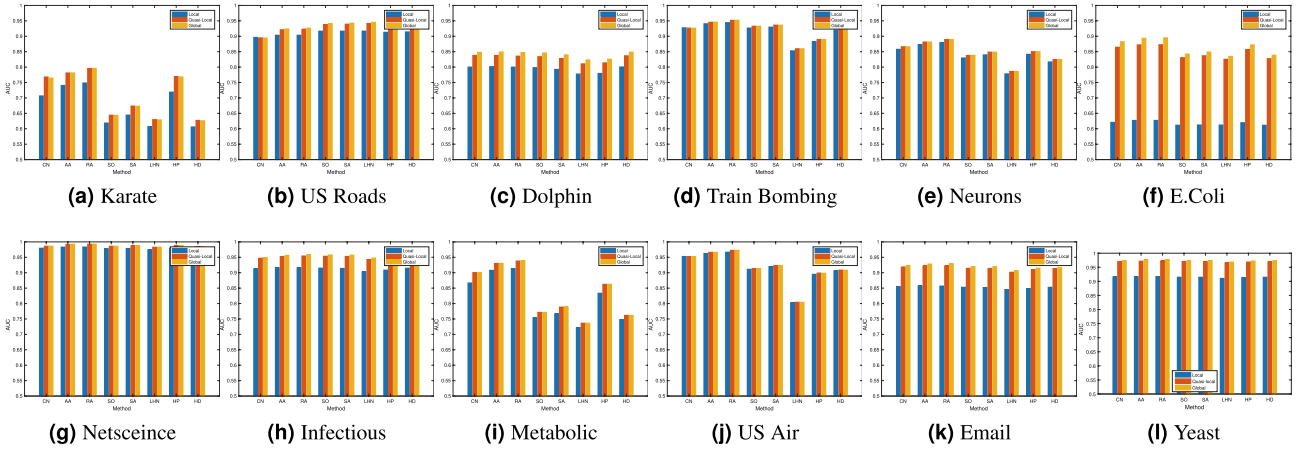
It is evident from the results that both the local and the global extensions can increase the prediction accuracy of the corresponding local indices. These global and quasi-local extensions not only result in high accuracies, but the accuracies' variations are also low when compared to the variations observed in local indices. We note that for some of the networks presented in Table 2, such as E.Coli network, the performance has been significantly increased, when longer paths are considered, whereas for other networks, the difference is not very significant. This improvement in performance may be attributed to the topological properties of the network, in particular, the average clustering coefficient and the density of the network. For a sparse network with a low clustering coefficient, it is unlikely that a similarity index, computed purely based on the degree statistics of immediate neighbour, will predict links with higher accuracy. From the statistics of networks presented in Table 2, one can see that the E.Coli dataset is very sparse and has the lowest clustering coefficient. Train bombing, on the other hand, is denser with a high clustering coefficient. Consequently, the increase in performance for the E.Coli dataset is around 25%, whereas for the Train bombing dataset, the performance has increased by less than 1%. Further information about the difference between the AUC values of global and quasi-local extension from their respective local index is presented in supplementary material (Table S1). In terms of comparison among the global and the quasi-local extensions of different indices, we observed that the path-based extensions of the RA index outperform all the alternative methods (including Katz index) on most of the datasets. Furthermore, the difference between the performances of the global and the quasi-local extensions of the RA index is also not very significant in most cases. Finally, it is also worth noting that the quasi-local extension of both the RA index and AA index always give superior performance when compared to local path index (a quasi-local extension of CN index). These results suggest that by incorporating the degree information of nodes on local paths, the prediction accuracies of local indices can be significantly improved.

To further investigate the performances of the global and quasi-local extensions and compare it to local indices, we evaluate the classification accuracy with different partitioning sizes of training and probe sets. For this purpose, we choose different sizes of the probe sets as 20%, 30%, 40%, 50% respectively. We have chosen eleven datasets in this experiment. For each split, we have computed the accuracies of the local indices, and both their global and quasi-local extensions. To visualise and compare those results, we have plotted the average accuracies of 100 independent runs of each experiment (with independent random splitting of E into E^P and E^T) in Figure 2. For comparison purpose, we have also included the results of the previous experiment, where we have chosen the size of the probe set as 10%, in the plot of Fig. 2. Note that, for large networks, the time required to compute AUC significantly increase with increase in probe size. Therefore, we have excluded the yeast dataset in this experiment.

There are a number of important observations that can be made from the results plotted in Fig. 2. Firstly, the performance of all the methods generally decreases with the increase in training size. This is obvious, as with the decrease in training size, we have less information available to predict links. This reduces the performance of the prediction algorithm. Secondly, in most case, when the structural error is very high, the local similarity indices suffer from low performance, while the global extensions can still give reasonably better performance. This is because of the fact that when we delete more links from the network, the local topology of network is considerably changed. In such cases, the global similarity indices, that take into account the overall topology

Datasets	Method	CN	AA	RA	SO	SA	LHN	HP	HD
Karate	Local	0.7066±0.0698	0.7406±0.0767	0.7485±0.0778	0.6187±0.0614	0.6444±0.0646	0.6071±0.0694	0.7190±0.0862	0.6061±0.0597
	Quasi-local	0.7681±0.0644	0.7811±0.0570	0.7957±0.0581	0.6441±0.0492	0.6737±0.0487	0.6297±0.0552	0.7693±0.0736	0.6272±0.0499
	Global	0.7646±0.0649	0.7813±0.0567	0.7961±0.0578	0.6432±0.0490	0.6732±0.0485	0.6289±0.0549	0.7682±0.0727	0.6260±0.0497
US Roads	Local	0.8963±0.0513	0.9034±0.0512	0.9035±0.0512	0.9167±0.0521	0.9166±0.0517	0.9168±0.0509	0.9131±0.0509	0.9142±0.0525
	Quasi-local	0.8948±0.0355	0.9212±0.0344	0.9231±0.0339	0.9381±0.0334	0.9394±0.0333	0.9416±0.0316	0.9314±0.0328	0.9332±0.0338
	Global	0.8943±0.0338	0.9240±0.0293	0.9263±0.0279	0.9412±0.0275	0.9427±0.0270	0.9454±0.0239	0.9338±0.0270	0.9358±0.0285
Dolphin	Local	0.8002±0.0531	0.8016±0.0544	0.8001±0.0542	0.7983±0.0539	0.7926±0.0530	0.7776±0.0510	0.7795±0.0508	0.8006±0.0545
	Quasi-local	0.8383±0.0478	0.8379±0.0496	0.8357±0.0498	0.8339±0.0483	0.8276±0.0477	0.8106±0.0466	0.8139±0.0461	0.8367±0.0489
	Global	0.8478±0.0405	0.8492±0.0417	0.8474±0.0414	0.8459±0.0404	0.8399±0.0397	0.8233±0.0382	0.8258±0.0377	0.8486±0.0408
Bombing	Local	0.9276±0.0332	0.9407±0.0316	0.9442±0.0312	0.9268±0.0310	0.9298±0.0307	0.8528±0.0267	0.8831±0.0284	0.9205±0.0314
	Quasi-local	0.9266±0.0323	0.9457±0.0279	0.9516±0.0266	0.9326±0.0273	0.9362±0.0267	0.8596±0.0225	0.8897±0.0253	0.9258±0.0278
	Global	0.9263±0.0325	0.9461±0.0274	0.9523±0.0256	0.9327±0.0275	0.9362±0.0268	0.8595±0.0224	0.8895±0.0252	0.9258±0.0280
Neurons	Local	0.8578±0.0110	0.8732±0.0108	0.8796±0.0107	0.8297±0.0110	0.8399±0.0110	0.7782±0.0106	0.8414±0.0109	0.8168±0.0109
	Quasi-local	0.8662±0.0099	0.8814±0.0090	0.8896±0.0086	0.8381±0.0094	0.8487±0.0093	0.7860±0.0093	0.8502±0.0092	0.8246±0.0095
	Global	0.8658±0.0100	0.8814±0.0090	0.8896±0.0086	0.8381±0.0094	0.8487±0.0093	0.7859±0.0093	0.8501±0.0092	0.8245±0.0095
E.coli	Local	0.6208±0.0402	0.6271±0.0421	0.6272±0.0421	0.6116±0.0380	0.6124±0.0383	0.6122±0.0383	0.6198±0.0405	0.6114±0.0379
	Quasi-local	0.8644±0.0335	0.8722±0.0346	0.8727±0.0349	0.8307±0.0313	0.8369±0.0315	0.8255±0.0312	0.8570±0.0332	0.8274±0.0311
	Global	0.8824±0.0281	0.8934±0.0294	0.8951±0.0294	0.8424±0.0296	0.8491±0.0293	0.8349±0.0305	0.8721±0.0292	0.8386±0.0297
Net Science	Local	0.9796±0.0091	0.9831±0.0091	0.9834±0.0091	0.9773±0.0092	0.9788±0.0092	0.9741±0.0091	0.9794±0.0092	0.9760±0.0091
	Quasi-local	0.9860±0.0048	0.9919±0.0047	0.9923±0.0046	0.9859±0.0047	0.9875±0.0046	0.9828±0.0046	0.9883±0.0047	0.9844±0.0047
	Global	0.9862±0.0045	0.9923±0.0043	0.9928±0.0043	0.9861±0.0045	0.9878±0.0044	0.9830±0.0045	0.9884±0.0044	0.9846±0.0045
Infectious	Local	0.9123±0.0062	0.9158±0.0063	0.9156±0.0063	0.9149±0.0062	0.9140±0.0062	0.9031±0.0060	0.9086±0.0062	0.9147±0.0062
	Quasi-local	0.9472±0.0038	0.9526±0.0036	0.9543±0.0036	0.9535±0.0035	0.9529±0.0035	0.9428±0.0033	0.9477±0.0035	0.9527±0.0035
	Global	0.9490±0.0035	0.9563±0.0031	0.9586±0.0029	0.9576±0.0028	0.9571±0.0028	0.9476±0.0025	0.9518±0.0029	0.9567±0.0029
Metabolic	Local	0.8668±0.0110	0.9077±0.0117	0.9134±0.0119	0.7539±0.0091	0.7669±0.0093	0.7224±0.0095	0.8335±0.0107	0.7479±0.0091
	Quasi-local	0.9002±0.0097	0.9301±0.0086	0.9379±0.0086	0.7715±0.0081	0.7888±0.0078	0.7364±0.0085	0.8624±0.0084	0.7617±0.0084
	Global	0.8998±0.0096	0.9304±0.0085	0.9382±0.0085	0.7715±0.0080	0.7890±0.0077	0.7364±0.0084	0.8625±0.0084	0.7616±0.0083
US Air	Local	0.9527±0.0063	0.9625±0.0061	0.9667±0.0061	0.9112±0.0055	0.9198±0.0058	0.8025±0.0054	0.8949±0.0064	0.9065±0.0054
	Quasi-local	0.9526±0.0061	0.9662±0.0050	0.9726±0.0047	0.9136±0.0050	0.9233±0.0052	0.8043±0.0049	0.8990±0.0058	0.9086±0.0048
	Global	0.9519±0.0063	0.9663±0.0049	0.9728±0.0045	0.9134±0.0051	0.9231±0.0053	0.8041±0.0048	0.8984±0.0059	0.9083±0.0049
Email	Local	0.8551±0.0083	0.8571±0.0083	0.8567±0.0083	0.8529±0.0082	0.8521±0.0082	0.8456±0.0080	0.8489±0.0081	0.8528±0.0082
	Quasi-local	0.9187±0.0065	0.9220±0.0065	0.9223±0.0066	0.9142±0.0065	0.9140±0.0064	0.9016±0.0063	0.9092±0.0064	0.9133±0.0065
	Global	0.9236±0.0059	0.9283±0.0057	0.9294±0.0056	0.9188±0.0060	0.9192±0.0059	0.9058±0.0059	0.9148±0.0058	0.9174±0.0061
Yeast	Local	0.9157±0.0056	0.9164±0.0056	0.9167±0.0056	0.9148±0.0056	0.9148±0.0056	0.9106±0.0055	0.9136±0.0056	0.9147±0.0056
	Quasi-local	0.9703±0.0019	0.9718±0.0019	0.9725±0.0019	0.9698±0.0018	0.9699±0.0018	0.9653±0.0018	0.9687±0.0019	0.9696±0.0018
	Global	0.9736±0.0021	0.9763±0.0019	0.9777±0.0018	0.9732±0.0020	0.9733±0.0020	0.9686±0.0020	0.9722±0.0020	0.9730±0.0020

Table 3. The prediction accuracy of the local, quasi-local, and the global indices for each method, measured by AUC. Each experiment was executed 100 times with independent random network division of the training set and the probe set and the average value along with standard deviations of all the 100 runs are reported. The cells highlighted in gray colour present the best performance obtained while the cells highlighted in light-grey colour present the second best performance.



(a) Karate **(b)** US Roads **(c)** Dolphin **(d)** Train Bombing **(e)** Neurons **(f)** E.Coli

(g) Netsceince **(h)** Infectious **(i)** Metabolic **(j)** US Air **(k)** Email **(l)** Yeast

Figure 1. Bar chart comparison of the accuracies, measured by AUC values, resulting from the application of different methods. The difference between the AUC values of the local indices and their respective extensions is significant in most cases.

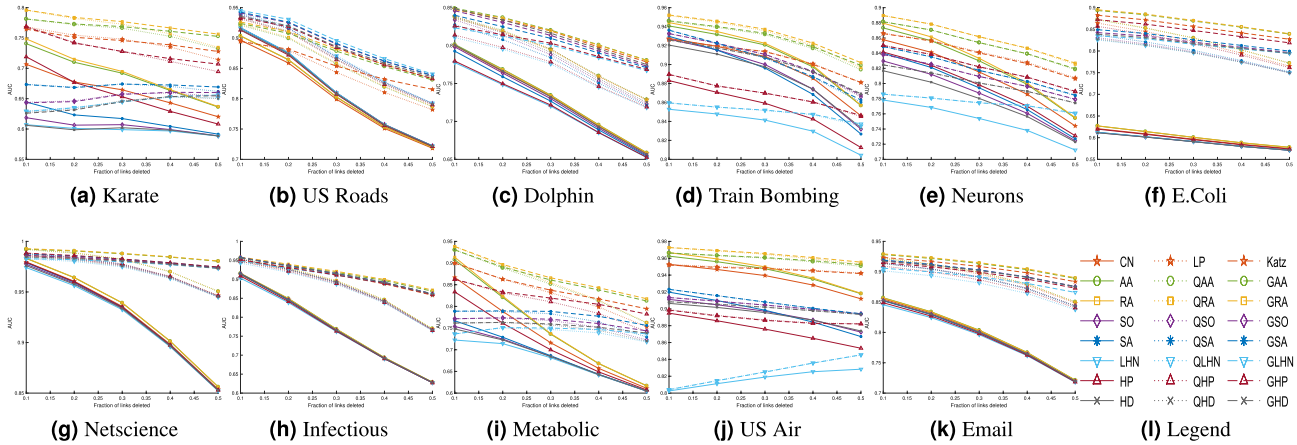


Figure 2. The prediction accuracy of the proposed and the alternative methods, measured by AUC, with different split of training and probe set. As with the previous experiments, each value is obtained by averaging over 100 executions of experiments with independently random divisions of training set and probe set.

Datasets	RA_Q	MFI	LO	CND	CRA	PA
Karate	0.7957±0.0581	0.7456±0.0661	0.6217±0.0897	0.7206±0.0642	0.5907±0.0686	0.7342±0.0815
US Roads	0.9231±0.0339	0.9299±0.0223	0.8963±0.0585	0.9206±0.0231	0.4822±0.0018	0.4264±0.0697
Dolphin	0.8357±0.0498	0.8483±0.0362	0.7202±0.0775	0.8405±0.0423	0.6468±0.0514	0.6717±0.0488
Train Bombing	0.9516±0.0266	0.9239±0.0282	0.9114±0.0413	0.9334±0.0296	0.9112±0.0409	0.7924±0.0413
Neurons	0.8896±0.0086	0.8691±0.0078	0.6852±0.0177	0.8579±0.0140	0.7902±0.0127	0.7339±0.0200
E.Coli	0.8727±0.0349	0.8840±0.0257	0.6822±0.0652	0.8653±0.0242	0.4997±0.0007	0.8543±0.0548
Netscience	0.9923±0.0046	0.9869±0.0070	0.9670±0.0166	0.9882±0.0058	0.8266±0.0263	0.6471±0.0254
Infectious	0.9543±0.0036	.9521±0.0019	0.7063±0.0122	0.9455±0.0033	0.7690±0.0084	0.6978±0.0068
Metabolic	0.9379±0.0086	0.8979±0.0075	0.7367±0.0166	0.8748±0.0098	0.7026±0.0096	0.8460±0.0093
US Air	0.9726±0.0047	0.9393±0.0044	0.7130±0.0182	0.9550±0.0054	0.9133±0.0116	0.9181±0.0091
Email	0.9223±0.0066	0.9217±0.0061	0.7225±0.0093	0.9099±0.0072	0.7039±0.0080	0.8056±0.0075
Yeast	0.9725±0.0019	0.9719±0.0019	0.8126±0.0122	0.9707±0.0023	0.8517±0.0084	0.8643±0.0056

Table 4. The prediction accuracies of our proposed method as well as of the state-of-the-art methods we benchmarked it against, measured by AUC. Each experiment was executed 100 times with independent random network division of the training set and the probe set and the average value of all the 100 runs are reported.

of the network, can outperform the local/quasi-local indices. As expected, when the structural error is high, the performance of a quasi-local index is always higher than the corresponding local index but less than the corresponding global index. Furthermore, the global extension of RA index usually outperforms all the other methods including Katz index for different partition sizes. Finally, for two datasets, namely Karate and US Air, we note that the prediction accuracy of some indices increases with increase in the size of probe test. This may be due to the fact that some link prediction algorithms, such as LHN, SA and SO, depend upon the degrees of query nodes. With more edges deleted from the network, such indices may predict links with higher accuracy for some specific datasets.

In order to assess the performance of the proposed link prediction indices, we compare their accuracies with some state-of-the-art link prediction algorithms. For this purpose, we have applied five alternative methods, namely, the MFI (Matrix-Forest Index)[27], the LO (Linear Optimisation)[28], the CND (Common Neighbour and Distance information)[24], the CAR-based indices proposed by Cannistraci et al.[20], and the PA (Preferential Attachment)[16] across all the twelve datasets we have used for performance assessment. The AUC values, obtained from the application of all these methods, are presented in Table 4. For the CAR-based indices, we have only reported the accuracies of the CAR-based extension of the RA index (CRA), as we observed that it outperforms all the other car-based indices. As discussed earlier, since the quasi-local extension of the RA index can be efficiently computed and gives comparable performance, for comparison purposes, we have also included its accuracies in the table. The results show that the RA_Q gives best or close to best performance when compared to alternate methods for all the datasets that were used for performance assessment. Additionally, it can also be verified from the results that the RA_G outperforms all the alternate methods. To investigate further, we have also computed the precision of prediction accuracies for all the methods. The results are presented in the supplementary material (Fig. S1).

In our final experiment, we investigate the performances of the global and quasi-local indices by varying the values of the parameter β. We have selected five different values of the parameter β, i.e., 0.001, 0.005, 0.01, 0.05, and 0.1. The resulting accuracies for different datasets are plotted in Fig. 3. These results suggest that both the global and quasi-local indices perform well for small values of the parameter β. The prediction accuracy

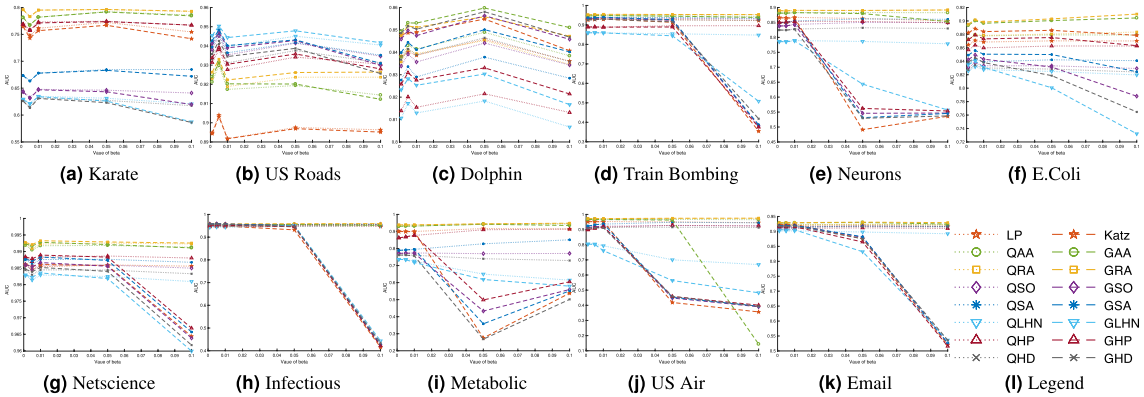


Figure 3. The prediction accuracy of the proposed and the alternative methods, measured by AUC, with different values of beta. As with the previous experiments, each value is obtained by averaging over 100 executions of experiments with independently random divisions of training set and probe set.

generally decreases when the value of the parameter β increases. This is due to the fact that for higher values of β, the longer paths are assigned more weights. This difference is significant for the global indices as it also considers paths with lengths greater than two. Note that a sudden drop in the performances of the global indices is due to the convergence problem of the global indices. In such cases, the performance can be approximated by expanding the series and considering the first few terms.

Conclusion

In this paper, we have proposed quasi-local and global extensions of local similarity indices that are used to predict the likelihood of existence of a link between two nodes in a network. This was achieved by considering local paths of different lengths and the information of the nodes on those local paths. We have also provided vectorised implementation of all the local methods and their proposed extensions. Experimental results on publicly available datasets demonstrate that both the global and the quasi-local extensions can increase the prediction accuracies of local methods. The performance of the proposed similarity indices was also reviewed with respect to different sizes of the probe sets and varying values of the parameter β. In both these cases, our proposed similarity indices achieved higher performance. The proposed method was applied to various domains including chemical networks, biological networks and social networks. In terms of future work, we plan to extend the work presented here to bipartite networks such as drug-target interaction networks. Note that, the experiments performed in this paper were limited to only simple networks, whose edges are unweighted and undirected. However, the proposed similarity indices can be easily extended to more complicated cases such as directed networks or weighted networks.

Received: 16 July 2020; Accepted: 2 November 2020
Published online: 16 November 2020

References

1. Newman, M. E. J. Network structure from rich but noisy data. *Nat. Phys.* **14**, 542–545 (2018).
2. Vallès-Català, T., Guimerà, R. & Sales-Pardo, M. On the consistency between model selection and link prediction in networks. *ArXiv e-prints* (2017).
3. Sales-Pardo, M., Guimerà, R., Moreira, A. A. & Amaral, L. A. N. Extracting the hierarchical organization of complex systems. *Proc. Nat. Acad. Sci.* **104**, 15224–15229 (2007).
4. Gao, J., Barzel, B. & Barabási, A.-L. Universal resilience patterns in complex networks. *Nature* **530**, 307–312 (2016).
5. Dellnitz, A. & Rödder, W. An entropy-based framework to analyze structural power and power alliances in social networks. *Sci. Rep.* **10**, 10697 (2020).
6. Knuth, D. . E. *The Art of Computer Programming* 1st edn, Vol. 4 (Addison-Wesley Professional, Boston, 2008).
7. Sumathipala, M. & Weiss, S. T. Predicting mirna-based disease-disease relationships through network diffusion on multi-omics biological data. *Sci. Rep.* **10**, 8705 (2020).
8. Newman, M. E. Estimating network structure from unreliable measurements. *ArXiv e-prints* (2018).
9. Wang, Z., Liao, J., Cao, Q., Qi, H. & Wang, Z. Friendbook: a semantic-based friend recommendation system for social networks. *IEEE Trans. Mob. Comput.* **14**, 538–551 (2015).
10. Makhatadze, G. I. Linking computation and experiments to study the role of charge–charge interactions in protein folding and stability. *Phys. Biol.* **14**, 013002 (2017).
11. Ai, J., Liu, Y., Su, Z., Zhang, H. & Zhao, F. Link prediction in recommender systems based on multi-factor network modeling and community detection. *EPL (Europhys. Lett.)* **126**, 38003 (2019).
12. Lu, Y., Guo, Y. & Korhonen, A. Link prediction in drug-target interactions network using similarity indices. *BMC Bioinformatics* **18**, 39 (2017).
13. Lorrain, F. & White, H. C. Structural equivalence of individuals in social networks. *J. Math. Sociol.* **1**, 49–80 (1971).
14. Adamic, L. A. & Adar, E. Friends and neighbors on the web. *Soc. Netw.* **25**, 211–230 (2003).
15. Zhou, T., Lü, L. & Zhang, Y.-C. Predicting missing links via local information. *Eur. Phys. J. B* **71**, 623–630 (2009).
16. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

17. Jaccard, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* **37**, 547–579 (1901).
18. Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.* **5**, 1–34 (1948).
19. Salton, G. & McGill, M. J. *Introduction to modern information retrieval* (McGraw-Hill Inc., New York, 1986).
20. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci. Rep.* **3**, 1613 (2013).
21. Zhao, J. *et al.* Prediction of links and weights in networks by reliable routes. *Sci. Rep.* **5**, 12261 (2015).
22. Ghorbanzadeh, H., Sheikhahmadi, A., Jalili, M. & Sulaimany, S. A hybrid method of link prediction in directed graphs. *Expert Syst. Appl.* **165**, 113896 (2021).
23. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).
24. Yang, J. & Zhang, X.-D. Predicting missing links in complex networks based on common neighbors and distance. *Sci. Rep.* **6**, 38208 (2016).
25. Ahmad, I., Akhtar, M. U., Noor, S. & Shahnaz, A. Missing link prediction using common neighbor and centrality based parameterized algorithm. *Sci. Rep.* **10**, 364 (2020).
26. Fouss, F., Pirotte, A., Renders, J. & Saerens, M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans. Knowl. Data Eng.* **19**, 355–369 (2007).
27. Chebotarev, P. & Shamis, E. The Matrix-Forest Theorem and Measuring Relations in Small Social Groups. *Autom. Remote Control.* **58**, 1505–1514 (2006) (**10 p**).
28. Pech, R., Hao, D., Lee, Y.-L., Yuan, Y. & Zhou, T. Link prediction via linear optimization. *Physica A* **528**, 121319 (2019).
29. Jeh, G. & Widom, J. Simrank: a measure of structural-context similarity. *its ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 538–543 (2002).
30. Kerrache, S., Alharbi, R. & Benhidour, H. A scalable similarity-popularity link prediction method. *Sci. Rep.* **10**, 6394 (2020).
31. Bai, M., Hu, K. & Tang, Y. Link prediction based on a semi-local similarity index. *Chin. Phys. B* **20**, 128902 (2011).
32. Liu, S., Ji, X., Liu, C. & Bai, Y. Extended resource allocation index for link prediction of complex network. *Physica A* **479**, 174–183 (2017).
33. Lü, L., Jin, C.-H. & Zhou, T. Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E* **80**, 046122 (2009).
34. Leicht, E. A., Holme, P. & Newman, M. E. J. Vertex similarity in networks. *Phys. Rev. E* **73**, 026120 (2006).
35. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
36. Kunegis, J. Konect: The koblenz network collection. *22Nd International Conference on WWW* 1343–1350 (2013).
37. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
38. Rossi, R. A. & Ahmed, N. K. The network data repository with interactive graph analytics and visualization. *29th Conference on AI* 4292–4293 (2015).
39. Hayes, B. Connecting the dots. can the tools of graph theory and social-network studies unravel the next big plot?. *Am. Sci.* **94**, 400–404 (2006).
40. Jinseop, K. & Marcus, K. From caenorhabditis elegans to the human connectome: a specific modular organization increases metabolic, functional and developmental efficiency. *Phil. Trans. R. Soc. B* **369**, 20130529 (2014).
41. Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of escherichia coli. *Nat. Genet.* **31**, 64–68 (2002).
42. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104 (2006).
43. Isella, L. *et al.* Whats in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.* **271**, 166–180 (2011).
44. Duch, J. & Arenas, A. Community detection in complex networks using extremal optimization. *Phys. Rev. E* **72**, 027104 (2005).
45. Colizza, V., Pastor-Satorras, R. & Vespignani, A. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nat. Phys.* **3**, 027104 (2007).
46. Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, 065103 (2003).
47. von Mering, C. *et al.* Comparative assessment of large- protein-protein interactions. *Nature* **417**, 399–403 (2002).
48. Lü, L. & Zhou, T. Link prediction in complex networks: a survey. *Physica A* **390**, 1150–1170 (2011).

Acknowledgements

GVG and FA acknowledge support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, Nanocommons H2020-EU (731032) and the NIHR Birmingham Biomedical Research Centre and the MRC Heath Data Research UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health.

Author contributions

F.A. and H.G. conceived the main idea of the paper and F.A., H.G. and I.U. have contributed in the development of the main algorithm. F.A., I.U. and G.G. have contributed in writing the manuscript. F.A. and H.G. have performed computational experiments and I.U. and G.G. have contributed in the analysis and interpretation of the results. All authors have reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-76860-2.

Correspondence and requests for materials should be addressed to F.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.