SCIENTIFIC REPORTS
nature research

OPEN

Application of hyperbolic geometry in link prediction of multiplex networks

Received: 28 November 2018
Accepted: 19 August 2019
Published online: 30 August 2019

Zeynab Samei[1] & Mahdi Jalili[2]

Recently multilayer networks are introduced to model real systems. In these models the individuals make connection in multiple layers. Transportation networks, biological systems and social networks are some examples of multilayer networks. There are various link prediction algorithms for single-layer networks and some of them have been recently extended to multilayer networks. In this manuscript, we propose a new link prediction algorithm for multiplex networks using two novel similarity metrics based on the hyperbolic distance of node pairs. We use the proposed methods to predict spurious and missing links in multiplex networks. Missing links are those links that may appear in the future evolution of the network, while spurious links are the existing connections that are unlikely to appear if the network is evolving normally. One may interpret spurious links as abnormal links in the network. We apply the proposed algorithm on real-world multiplex networks and the numerical simulations reveal its superiority than the state-of-the-art algorithms.

Many real biological, social and technological systems are modeled as networks in which nodes and links represent entities and different kinds of connections respectively. Network analysis in complex systems such as biology, ecology, computer science and sociology has become very important and applicable[1]. One of the major topics in network science is to predict missing, forthcoming and spurious links[2]. Many different link prediction algorithms have been introduced that use structure information of networks. Most of them are classified in similarity-based prediction methods, which work under the assumption that the probability of existing a link between two nodes is depended to their similarity[3].

There are other types of link prediction methods such as Hierarchical Structure Model and Stochastic Block Model which are based on maximum likelihood analysis[2,4,5]. Recently the study of hyperbolic geometry based on the network structure has become useful in solving the link prediction problem. Considering the hyperbolic geometry of networks, HyperMap method was proposed by Papadopoulos *et al.*[6]. This method first map target networks into hyperbolic space, and then predict the missing links using the hyperbolic coordinates of node pairs[7].

Recent studies[8–10] have shown that many real network systems are modeled better in multiple layers to show different kinds of interactions between individuals[11,12]. Multiplex networks are a special kind of multilayer networks in which the number of nodes in all layers is the same. Some studies have shown that the structural features of different layers in multiplex networks are indeed correlated to each other[1,13]. So it can be supposed that considering the interlayer information can enhance the performance of link prediction in each layer of a multiplex network. In this paper, we investigate the node similarity index based on hyperbolic geometry and the layer relevance of the multiplex networks for predicting the spurious and missing links. In this method, we improve the performance of the link prediction based on hyperbolic distance considering both the popularity and similarity of nodes by combining the similarity indices in multiplex networks.

Using the interlayer information in solving the missing link prediction in multiplex networks has been considered before in a number of works. Pujari *et al.*[14] used a decision tree classifier to predict the interaction of the coauthorship network in a multiplex collaboration network with three layers. Hristova *et al.*[15] used a supervised classifier for link prediction in a two layer network containing Foursquare and Twitter using the interlayer information. In another work, Sharma *et al.*[16] proposed a new method considering weight for each layer of multiplex network and used it to solve the link prediction problem in the target layer. Yao *et al.*[17] proposed a novel method

[1]Department of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran. [2]School of Engineering, RMIT University, Melbourne, Australia. Correspondence and requests for materials should be addressed to Z.S. (email: z.sameie@gmail.com)

based on interlayer and intralayer information to solve the missing link prediction problem in multiplex networks. Hajibagheri *et al.*[18] proposed a holistic method considering the information of all layers simultaneously in link prediction of a target layer in multiplex networks. Also, Guimerà *et al.* used Stochastic Block Models to predict missing and spurious links in noisy networks. Zeng *et al.*[19] studied the impact of spurious link identification methods on distortion of networks' structure and dynamics. In another study, Zhang *et al.*[20] measured the inter-similarity using the local diffusion processes in bipartite networks. Samei *et al.*[21] proposed a method to identify spurious links in multiplex networks. In fact, they proposed a method to employ interlayer information to improve the performance of spurious link prediction in the target layer.

In the context of hyperbolic geometry of network, Krioukov *et al.*[22] introduced the mapping of networks to hyperbolic space. They used the underlying hyperbolic geometry of network to study the functionality and structure of complex networks. They showed that the strong clustering and the heterogeneous degree distribution are natural reflections of the negative curvature and other properties of the hyperbolic geometry of complex networks. Then, Papadopoulos *et al.*[23] studied the impact of popularity and similarity in networks' growth. They developed a framework to suggest that new connections can be made between node pairs with an optimized trade-off between popularity and similarity. In another work, Papadopoulos *et al.*[24] presented the HyperMap method to map a network to its underlying hyperbolic space and used the hyperbolic distance as a similarity measure to solve the link prediction problem. Different from these works, other methods were introduced to infer hidden geometry of complex networks[6,25]. Recently, Muscoloni *et al.*[26,27] introduced a nonuniform popularity-similarity optimization model (*N-PSO*). This model was used to predict the missing links using the community structure of the networks in *N-PSO* that improved the performance of the link predictors significantly. Muscoloni *et al.* also proposed an intelligent machine to infer the network hyperbolic geometry based on an "angular coalescence" phenomenon[28]. A minimum curvilinear automata has been recently proposed to embed hyperbolic geometry of networks and used it for link prediction[29].

In this paper, our proposed similarity indices based on hyperbolic geometry of network benefit both intralayer and interlayer information to solve the spurious and missing link prediction in multiplex networks. Based on that the experimental results on four single layer synthetic networks and six real multiplex networks show that the performance has been improved when the hyperbolic-based methods is used and the node pairs similarity measures are computed considering both interlayer and interlayer information.

Methods

Consider $G = (G^1, G^2, \ldots, G^M)$ as a multiplex network with N nodes in each of M layers, where $G^\alpha = (V^\alpha, E^\alpha)$ represents the network of layer α with V as the set of nodes and E as the set of links[11,12,30]. We can assume $A^{[\alpha]} = \{a_{ij}^{[\alpha]}\}$ as the adjacency matrix of each layer G^α, where for $1 \le \alpha \le M$ and $1 \le i, j \le |V^\alpha|$[31]:

$$a_{ij}^{[\alpha]} = \begin{cases} 1 & if (v_i^{[\alpha]}, v_j^{[\alpha]}) \in E^\alpha \\ 0 & otherwise \end{cases} \quad (1)$$

In the context of unsupervised link prediction, many similarity measures are defined to find the likelihood of link existence between each node pair (i, j). In multiplex networks the similarity score in layer α is shown by s_{ij}^α. After computing the similarity scores for all potential node pairs in each layer, a ranking method can be used to choose the top ranked pairs which have more chance to make a connection. The key issue is how to calculate the similarity scores based on the known topology of networks. Recent studies have shown that the structure of the layers in multiplex networks are mostly dependent[32,33]. Hence, one of the main challenges in solving the link prediction problem in multiplex networks is to find an appropriate similarity measure that can benefit the relevant information of all layers[30]. Based on this, here we use both the information of the target layer (intralayer information) and other layers (interlayer information) and combine them based on layer relevance to improve the performance of link prediction compared with the single-layer based methods.

In the case of missing link prediction, the goal is to estimate the probability of existence of non-observed links based on the current topology of network and available node's features in network G(V, E). Since the missing links are not known, we assume that a fraction of observed links E, is missing and the goal of link prediction is to identify them. In order to do that, in each iteration a fraction of the observed links $_E$, is removed based on k-fold decomposition method and the proposed methods are supposed to predict them. In the case of identifying spurious links, the task is to evaluate whether the observed links are reliable enough based on the current topology of the network. In order to do that, in each iteration, some nonexistent links are randomly added to the link set and the proposed methods are supposed to identify them[34]. Precision is used here to quantify the accuracy of a link prediction method which is defined as:

$$Precision = \frac{|TP|}{|TP| + |FP|} \quad (2)$$

where $|TP|$ is the number of positive predictions that are truly predicted and $|FP|$ is the number of positive predictions that are wrongly predicted.

Node similarity index. Description of the similarity indices is given in the following.

Existing measures.

- **Preferential Attachment (PA):** This index is based on the node degrees and for each node pair i and j is defined as:

$$s_{ij}^{PA} = \|\Gamma_i\| \times \|\Gamma_j\| \tag{3}$$

where $||\Gamma_i||$ indicates the number of neighbors of node i.

- **Common Neighbors (CN)**: For each node pair i and j, this index counts the number of neighbors that are common between them and is defined based on the assumption that node pairs with more common neighbors are more likely to make connection. It is defined as:

$$s_{ij}^{CN} = \left\|\Gamma_i \bigcap \Gamma_j\right\| \tag{4}$$

- **CAR:** This measure considers both common neighbors of each node pair and the number of connections between the common neighbors, and is computed as below:

$$s_{ij}^{CAR} = s_{ij}^{CN} \cdot s_{ij}^{LCL} \tag{5}$$

where s_{ij}^{CN} is the number of common neighbors between (i, j) and s_{ij}^{LCL} is the number of links between nodes in the common neighbors set[35].

- **CJC:** This measure is a modified version of Jacard measure and is defined as below:

$$s_{ij}^{CJC} = \frac{s_{ij}^{CAR}}{\left|\left|\Gamma_i \cup \Gamma_j\right|\right|} \tag{6}$$

where s_{ij}^{CAR} is the similarity measure CAR defined above and $\left\|\Gamma_i \cup \Gamma_j\right\|$ is total number of neighbors of nodes i and j[35].

- **Hyperbolic distance (HP):** This measure computes the hyperbolic distance (Eq. (8)) of each node pair i and j based on *Hypermap* method. *HyperMap* is based on Maximum Likelihood Estimation. It finds the radial and angular coordinates r_i, θ_i for all nodes $i \leq N$, which maximizes the likelihood:

$$L = \prod_{1 \leq i < j \leq N} p(x_{ij})^{\alpha_{ij}} [1 - p(x_{ij})]^{1-\alpha_{ij}} \tag{7}$$

where the product is computed over all node pairs i, j and x_{ij} is defined as the hyperbolic distance between pair i, j:

$$\begin{aligned} x_{ij} &= \operatorname{arccosh}(\cosh r_i \cosh r_j - \sinh r_i \sinh r_j \cos \Delta\theta_{ij}) \\ &\approx r_i + r_j + 2ln \sin(\Delta\theta_{ij}/2) \\ &\approx r_i + r_j + 2ln (\Delta\theta_{ij}/2) \end{aligned} \tag{8}$$

where

$$\Delta\theta_{ij} = \pi - |\pi - |\theta_i - \theta_j|| \tag{9}$$

and $p(x_{ij})$ is the Fermi-Dirac connection probability:

$$p(x_{ij}) = \frac{1}{1 + e^{\frac{1}{2T}(x_{ij} - R)}} \tag{10}$$

where $R \sim \ln N$. The estimated radial coordinate of node i is based on its degree in the network (k_i) via $r_i \sim \ln N - \ln k_i$. Therefore, if node degrees are correlated in different layers so will be the radial coordinates[36].

Proposed measures. Node degree or popularity plays an important role in defining the similarity measures and many of them are based on common neighbors and preferential attachment. The underlying principle behind preferential attachment is that new connections are mainly made to more popular nodes. However, Papadopoulos *et al.*[23] showed that popularity is just one aspect of attractiveness, while similarity could be considered as another aspect. They developed a framework where new connections consider a trade-off between popularity and similarity.

We know that the degree distribution of many real networks follow power-law distribution. However, as it can be seen in As real multilayer networks, we consider six networks (see Table 1). The multilayer networks are converted to multiplex networks by assuming that all layers have the same number of nodes (the maximum number of nodes of all layers). Explanation of these networks is as follow:

Table 1, the degree distribution of the multiplex networks with small size do not follow power-law distribution. The previous experimental results indicated that HP's performance was better in the networks with power-law distribution and less good in those that does not obey power-law degree distribution[7]. The reason for that would be the way the nodes' radial coordinates are calculated. Because one of the parameters which is considered in *HyperMap* method to estimate the radial coordinates is the power-law exponent of the network. Therefore, if the network does not have a power-law degree distribution, the link-prediction accuracy of HP decreases. In order

	i	N	|E|	⟨k⟩	S	H	Γ	T
Vicker	1	29	240	16.5	0.59	1.09	3.5	0.75
	2	29	126	8.69	0.31	1.27	3.5	0.85
	3	29	152	10.48	0.37	1.25	2.64	0.75
Lazega	1	71	717	20.19	0.28	1.16	3.5	0.75
	2	69	399	11.56	0.17	1.31	3.5	0.5
	3	71	726	20.45	0.29	1.16	3.5	0.8
CKM	1	215	480	2.23	0.02	1.61	3.04	0.55
	2	231	565	2.44	0.021	1.44	3.5	0.45
	3	227	504	2.22	0.019	1.33	3.5	0.35
CElegans	1	253	516	4.07	0.016	2.15	3.13	0.65
	2	260	888	6.83	0.026	1.78	3.35	0.85
	3	278	1703	12.25	0.044	1.67	2.79	0.85
Rattus	1	2035	3014	1.48	0.001	5.86	2.62	0.35
	2	1017	1093	1.07	0.002	3.99	2.14	0.25
SacchPomb	1	971	1686	1.73	0.003	2.72	2.92	0.9
	2	347	404	1.16	0.006	1.93	2.85	0.9
	3	2402	7502	3.12	0.002	3.92	2.68	0.25

Table 1. The topological features of six real multiplex networks. In the table, i is the number of layers, N is the number of nodes and E is the number of edges in each layer. $\langle k \rangle$ represents the average degree, S is the density of each layer based on $\left(S = \frac{2E}{N(N-1)}\right)$ and H is the degree heterogeneity obtained as $\left(H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}\right)$, T is the temperature and γ is the power-law coefficient.

to overcome this shortcoming and benefit the advantages of both popularity and similarity features of nodes, we proposed two approaches that are detailed in the following.

- **Weighted Common neighbors (WCN):** We generate a weighted version of CN that computes the weight of common neighbors considering the hyperbolic distance of them with the target node pairs. There are some studies about converting the original similarity measure to the weighted one, however it has been shown that that such conversion may reduce the prediction performance[37]. The pseudo-code of the proposed method is as follows:

  1. Approximate the hyperbolic coordinates of each node.
  2. Compute the matrix H of hyperbolic distance of the existing links in the network.
  3. $h =$ average of H
  4. $\Gamma_{ij} =$ list of common neighbors of node pair (i, j) in the test list of missing or spurious link prediction.
  5. for each $k \in \Gamma_{ij}$
     if $H(i, j) < h$
     node pair (i, k) is a strong tie and has more weight
     $WCN(i, j) = WCN(i, j) + 1 + 1/H(i, k)$;
     else
     node pair (i, k) is a weak tie and takes the weight as CN
     $WCN(i, j) = WCN(i, j) + 1$;
  6. Repeat step 5 for node pair (k, j)
  7. Sort all links in the test list in decreasing (for missing link prediction) or increasing (for spurious link prediction) order

- **Ranking CN and HP (CN-HP)**

This method benefits the advantages of both CN and HP measures. It uses a ranking method to combine the prediction given by both of these measures. In order to do that, one of the well-known classical rank aggregation methods, Borda's method is used[38]. It is based on absolute positioning of the ranked elements rather than their relative rankings. A Borda score for each element is calculated based on the ranking of it in the aggregated list. For a set of full list $L = [L_1, L_2, L_3, \ldots, L_n]$, the Borda's score for element x and list L_k is given by:

$$B_{L_i(x)} = \{count(y) | L_i(y) < L_i(x) \& \, y \in L_i\} \tag{11}$$

and the total Borda's score of element x is:

$$B(x) = \sum_{i=1}^{n} B_{L_i(x)} \tag{12}$$

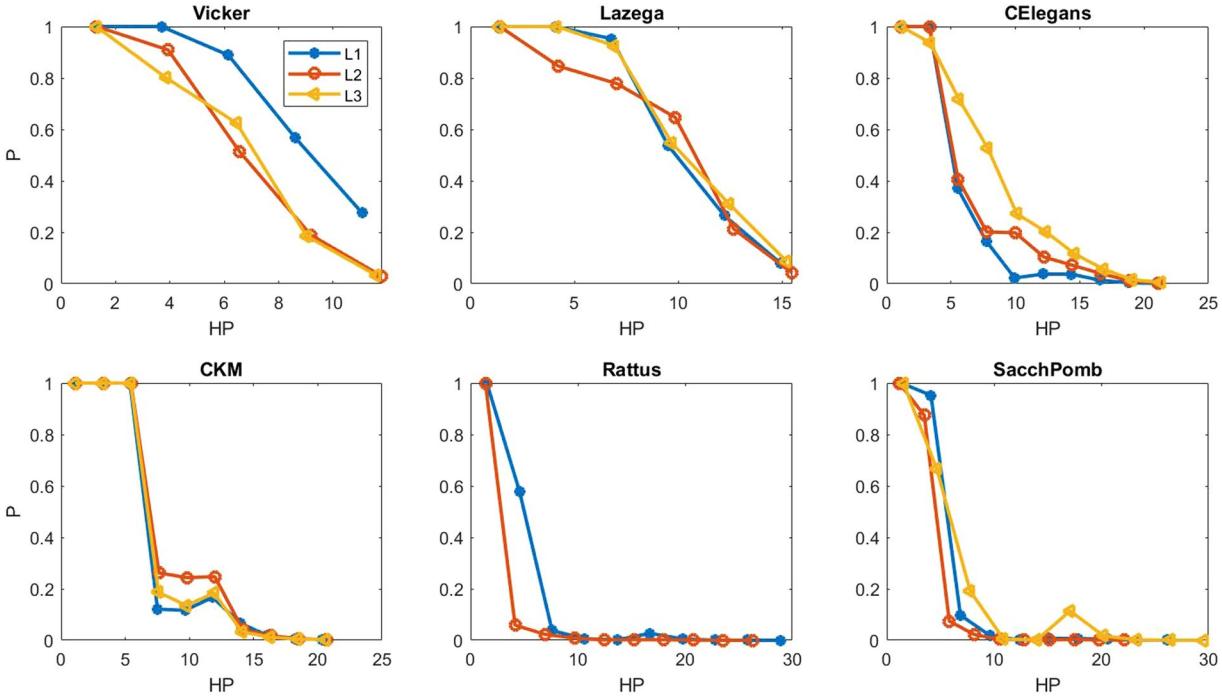

Figure 1. Probability of existing a link between node pairs based on their hyperbolic distance in different layers of multiplex networks denoted by L1, L2 and L3.

The advantage of Borda ranking method is that we can aggregate different kinds of measures with different categories and values and obtain a rank-based score. Also the computational complexity of this method is linear; however it does not satisfy the Condorcet criterion. In the proposed method, two lists of CN and HP scores for any node pair are constructed, and the final score for each node pair is computed by aggregating their ranking score using Borda method, i.e. in the case of missing/spurious link prediction the aggregated scores of CN and HP of all node pairs are computed based on Eq. (12) and are sorted descending/ascending. The top-k elements of the final list are the predicted links (k is the number of expected missing/spurious links).

Impact of layer relevance. In the case of HP as a similarity measure, the procedure of mapping each layer to its hyperbolic space can be done in different directions. One direction would be to jointly embed the different layers of a given multiplex and infer single radial and angular coordinates for each node. A second direction would be to aggregate the different layers using different operations such as those proposed in[39], and then embed the aggregated network to infer single coordinates for nodes. Finally, a third direction would be to infer the node coordinates in each layer independently as considered here.

As it was mentioned above, we map each layer of each real multiplex to its hyperbolic space using the *HyperMap* method[6,24]. The method takes the network adjacency matrix and the network parameters T, γ. It then approximates the angular and radial coordinates of all nodes in the network. Parameter γ is the power law degree distribution exponent which is approximated separately for layers using the method introduced by Clauset *et al.*[40], and T is the temperature. To estimate the values of T, the Nonuniform Popularity × Similarity Optimization *N-PSO* model is used[27]. The *N-PSO* model grows synthetic complex networks and it is equivalent to the hyperbolic H^2 model. The inputs to this model are the final network size N, the average node degree k, power-law coefficient γ and the network parameters T. The *N-PSO* model is used to construct synthetic networks with the same size N and average degree k and power-law exponent γ, using different values for T. The estimated values of T are then the values that best match the degree distribution and average clustering between the layer and the corresponding synthetic network.

In order to test whether this measure can be a good one for the link prediction, we classify the hyperbolic distance of all node pairs, and compute the probability of the existence of a link between the pairs in each bin. To this end, first the hyperbolic distances of all node pairs are sorted in ascending order and divided to k bins. Bin b_i contains the node pairs with the hyperbolic distance in the range of $[d_i, d_{i+1}]$. Then, the probability p_i of having a link between the node pairs of each bin is computed based on the network topology. The results are shown in Fig. 1. As it is shown, the probability of existing a link between each node pairs decreases, while their hyperbolic distance increases. Two nodes have a smaller hyperbolic distance as much as they are popular or similar to each other, in this case the probability of existing a connection between them increases. Thus, this measure can be a candidate for the similarity score for the link prediction problem. It is worth noting that the behavior of different layers are almost the same in all multiplex networks, with being more similar in the bigger networks including Rattus and SacchPomb.

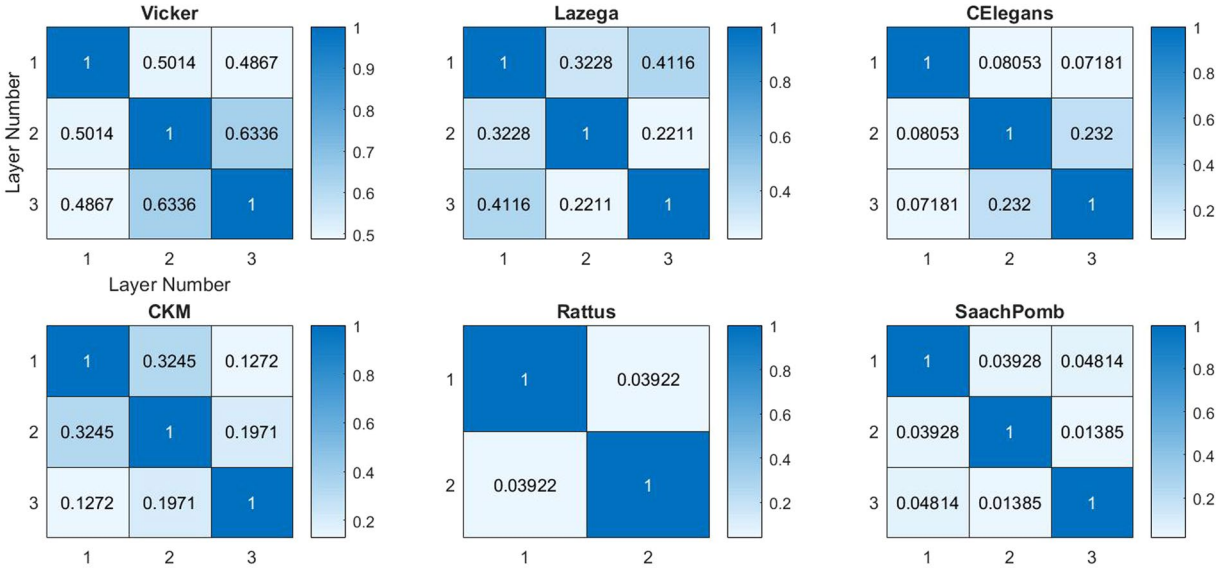

Figure 2. Link Overlap of different layers of six real multiplex networks.

In order to employ the interlayer information, for node pair (i, j) in target layer α, we first calculate its similarity within each layer based on the proposed methods above. This enables us to compare prediction performance of the algorithms. In order to compare the prediction performance of the proposed prediction framework, we exploit different algorithms for quantifying the relevance between layers including link overlap, Pearson correlation, Spearman correlation and hyperbolic angular correlation[12]. The results show that the link overlap has the best effect on the link prediction performance. It is defined in the following.

- **Link Overlap (LO):** This measure identifies the ratio of common links in two layers, i.e. if α and β are two layers in a multiplex network, LO is the fraction of the same node pair that connects in both layers α and β and is defined as:

$$O^{\alpha,\beta} = \frac{2\sum_{i=1}^{N}\sum_{j>i}^{N} A_{ij}^{[\alpha]} \cdot A_{ij}^{[\beta]}}{\sum_{i=1}^{N}\sum_{j>i}^{N} A_{ij}^{[\alpha]} + \sum_{i=1}^{N}\sum_{j>i}^{N} A_{ij}^{[\beta]}} \tag{13}$$

Where $A^{[\alpha]}$ is the adjacency matrix of layer α that takes value of 0 for each disconnected node pair and 1 for each connected node pair, and N is the number of nodes. The value of $O^{\alpha,\beta}$ is in the range of [0, 1], where 0 indicates that the layers are completely irrelevant and 1 indicates that the layers are quite relevant. The similarity measure is defined as:

$$\forall\, i, j \in V\colon S_{ij} = s_{ij}^{\alpha} + \sum_{\beta=1}^{M} \eta\mu^{\alpha\beta} s_{ij}^{\beta} (\alpha \neq \beta) \tag{14}$$

where s_{ij}^{α} is the similarity index of target layer α and s_{ij}^{β} is the similarity index of any other layer β. $\mu^{\alpha\beta}$ represents the correlation between layers α and β (link overlap), which can be explained as the weight of interlayer information involved from any layer β in link prediction in layer α and η is the tunable parameter. The correlations between different layers are shown in the Fig. 2. As it is shown, for all networks the link overlap correlation between different layers is positive. Furthermore, the highest layer relevance belongs to the Vicker network and the lower relevance belongs to larger and sparser networks. Our experiments show that LO is mostly consistent with other correlation metrics, but it has the most positive effect in the extent the interlayer information can improve the link prediction performance.

Results

We perform experiments on four single-layer synthetic networks to evaluate the similarity measures and six real multilayer networks to investigate the impact of interlayer information. The synthetic networks are evolved based on *N-PSO* model described above and their structural features and the precision of spurious and missing link prediction methods are presented in Figs 3 and 4. In the *N-PSO* model, the true node coordinates are generated for the networks. In the case of missing link prediction, we remove a fraction of edges using k-fold decomposition in each iteration and regenerate the node coordinates of the new network using the *Hypermap* method. Similarly, in the case of spurious link prediction, we add a fraction of nonexistent links to the network in each iteration and regenerate the node coordinates of the new network using the *Hypermap* method. There is no restriction in selecting the parameters for *N-PSO* model. It is preferred to generate networks with features that are near to real networks (large and sparse with power-law degree distribution) and temperature is chosen to be 0.3 and 0.6.

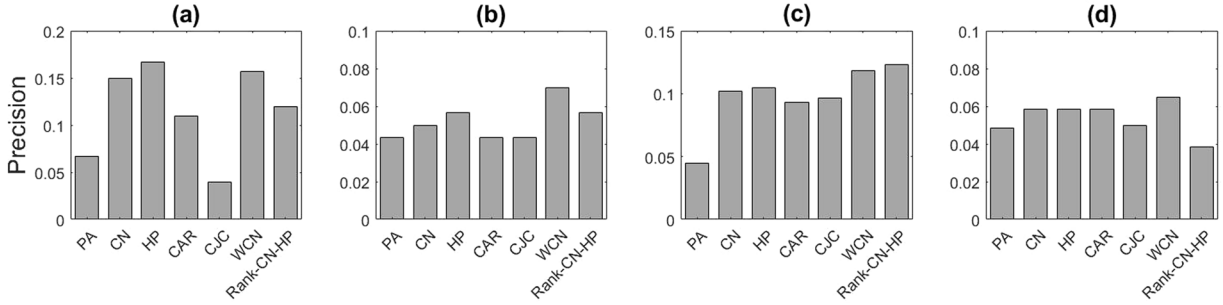

Figure 3. The missing link prediction performance of the synthetic networks based on *N-PSO* model, with (**a**) N = 500, m = 4, λ = 3, T = 0.3, (**b**) N = 500, m = 4, λ = 3, T = 0.6, (**c**) N = 1000, m = 4, λ = 3, T = 0.3, (**d**) N = 1000, m = 4, λ = 3, T = 0.6. Different similarity measures are used, including Preferential Attachment (PA), Common Neighbors (CN), Hyperbolic Distance (HP), CAR, CJC, Weighted Common Neighbors (WCN) and Rank-CN-HP. The results show the mean values over 20 independent experiments.

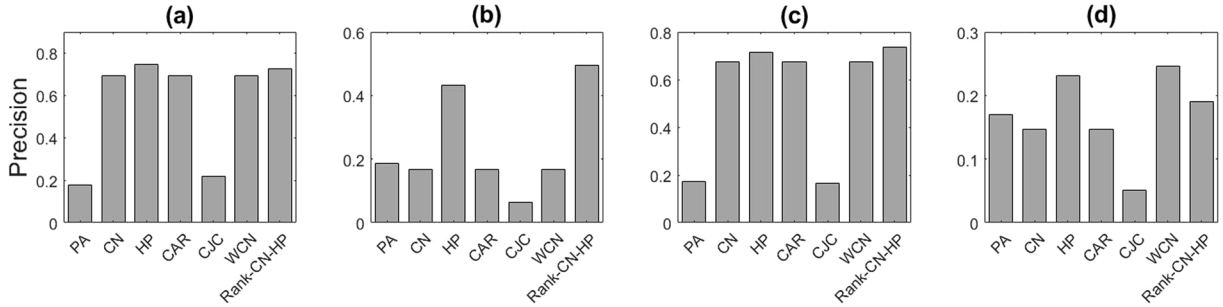

Figure 4. The spurious link prediction performance of the synthetic networks based on *N-PSO* model, with (**a**) N = 500, m = 4, λ = 3, T = 0.3, (**b**) N = 500, m = 4, λ = 3, T = 0.6, (**c**) N = 1000, m = 4, λ = 3, T = 0.3, (**d**) N = 1000, m = 4, λ = 3, T = 0.6. Different similarity measures are used, including Preferential Attachment (PA), Common Neighbors (CN), Hyperbolic Distance (HP), CAR, CJC, Weighted Common Neighbors (WCN) and Rank-CN-HP. The results show the mean values over 20 independent experiments.

Since the two parameters λ and *T* of *Hypermap* are set manually, so the approximation of hyperbolic coordinates is more accurate in synthetic networks.

Based on these reasons as it can be seen, in all cases the performance of hyperbolic distance (HP) is better than the other measures and the precision of the proposed methods (Weighted Common Neighbors (WCN) and Rank-HP-CN) is the highest in most cases. Thus, hyperbolic distance and its derived methods can be good choices as similarity measures for link prediction.

As real multilayer networks, we consider six networks (see Table 1). The multilayer networks are converted to multiplex networks by assuming that all layers have the same number of nodes (the maximum number of nodes of all layers). Explanation of these networks is as follow:

(1) Vicker[41]: It is a 3 layer multiplex network with 29 nodes representing the students of a school in Australia. The layers are defined as the contact relationship, co-working and best friends.
(2) Lazega[42,43]: This multilayer network represents the partnership of corporate law between associates and partners. The layers correspond to co-working, friendship and advice relationship.
(3) CKM[44]: This multilayer network represents the interactions between physicians. It contains 3 layers that correspond to friendship, discussion and asking for advice.
(4) CElegans[45,46]: It is a biological multilayer network in which nodes represent neurons and layers correspond to chemical monadic, chemical polyadic and electric interactions.
(5) Rattus[47,48]: It is a multiplex genetic and protein interactions network of the Rattus Norvegicus. It contains two main layers of physical association and direct interaction.
(6) SacchPomb[47,48]: It is a multiplex genetic and protein interactions network of the Saccharomyces Pombe. It includes three kinds of relationships, including direct interaction, colocalization and physical association.

The experimental results of the proposed link prediction methods on six real networks is presented in this section. For each multiplex network, the layer with the most density is chosen as the target layer. In the case of missing link prediction, 15% of links in the target layer are considered to be hidden and based on k-fold decomposition method, the performance of the similarity measures are examined over 20 independent experiments. For spurious link prediction, random links are added to the network and the performance of the similarity measures

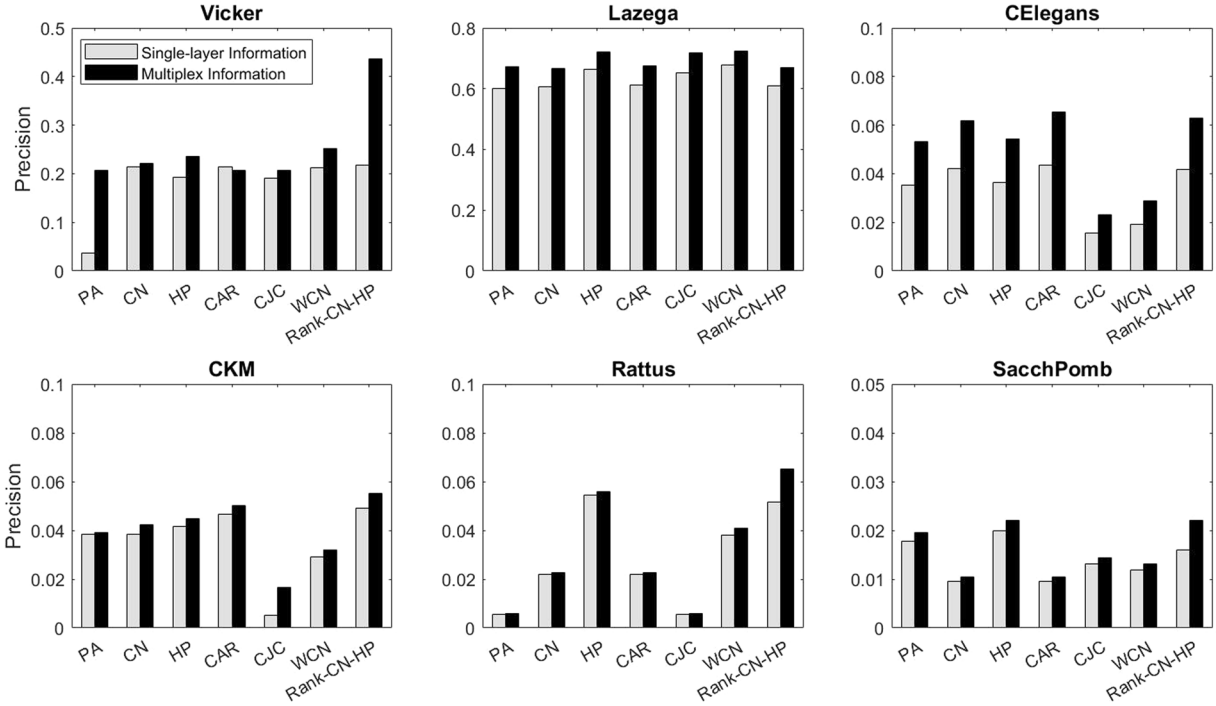

Figure 5. The missing link prediction performance of the multiplex networks based on different similarity measures. The results are based on the mean values over 20 independent experiments. 'Single-Layer Information" corresponds to the case when only intralayer information of the target layer is considered. "Multiplex Information" corresponds to the case when both intralayer information of the target layer and interlayer information of other layers are considered.

are examined over 20 independent experiments. In order to evaluate the impact of employing the layer relevance and the extra information of other layers, we separately study the performance of the algorithms on single-layer (when only information of the target layer is considered) and multiplex (when inter-layer information is also considered) fashions. We employ the layer correlation based on link overlap and compute the similarity measures based on Eq. (14).

Figure 5 shows the precision of missing link prediction of different similarity measures. For each measure there are two bars. The left bar shows the performance of the similarity measure while considering only the intra-layer information of the target layer and the right bar is the performance of the similarity measure while using both intralayer and interlayer information. As it can be seen, in all cases incorporating the interlayer information improves performance of the missing link prediction and this is more pronounced in CElegans. The proposed similarity measures Rank-CN-HP has the best performance in most cases.

Figure 6 shows the performance of the algorithms on spurious link prediction. For this problem, we also consider the cases when only intralayer information of the target layer is considered and when both intralayer and interlayer information are considered. As it can be seen, in most cases, including the interlayer information in the prediction process improves the performance. Furthermore, predictions based on PA similarity measures have the worst performance in most cases. In contrast to the missing link prediction, in this case Rank-CN-HP is not better than CN or HP in some of the networks, but WCN has the best performance in all multiplex networks. Our experiments show that in small networks, in the case of HP, the approximated radial coordinates of nodes in hyperbolic space for both true positive (correctly predicted) links and false negative links are almost in the same range. But the average degree of true positive links is significantly higher than the false negatives. It means that the radial coordinates of nodes which corresponds to their popularity are not precisely approximated, since the degree distribution of the target layers do not obey the power-law. HP mostly represents the similarity of node pairs, and thus it is not expected in most cases to have high performance. Therefore, combining this similarity measure with CN in different ways help to overcome the shortcoming of HP in covering the popularity attribute of each node. On the other hand, in large networks and especially in those with scale-free degree distribution, approximating the underlying hyperbolic geometry is more precise, but these networks are mostly sparse and similarity measures such as CN, CAR and CJC may not be quite successful in link prediction. Thus, in such cases combining the popularity-based measures with HP can improve the link prediction. The Rank-CN-HP and WCN methods both use CN as the popularity factor, and HP as the similarity factor. The difference is that in Rank-CN-HP the proposed similarity measure uses CN and HP independently, i.e. these two measures are first computed independently for each node pair, and then ranked based on Borda rank aggregating algorithm to achieve the final score that considers both CN and HP with the same weight. Whereas in the WCN method, we compute HP-distance for the common neighbors of each node pair and compare them with a threshold. If the HP-distance is less than the threshold, that node pair is assumed to have a strong tie, i.e. they are more similar to

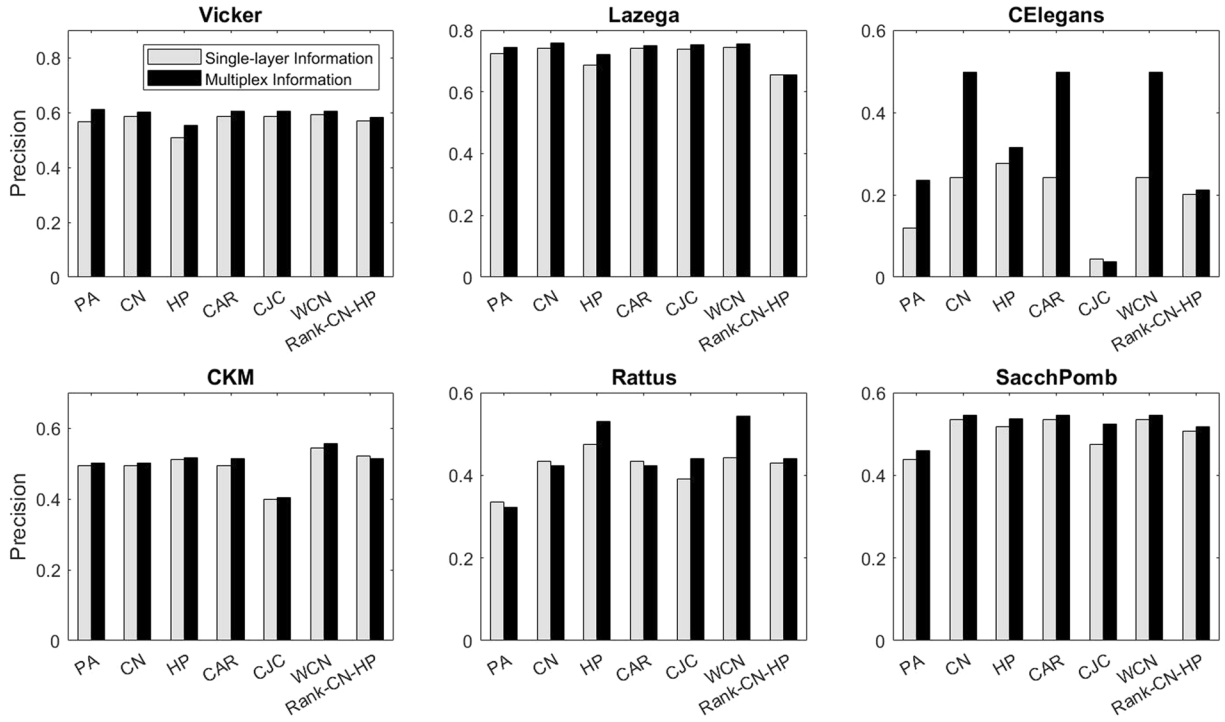

Figure 6. The spurious link prediction performance of the multiplex networks based on different similarity measures. The results are based on the mean values over 20 independent experiments. 'Single-Layer Information" corresponds to the case when only intralayer information of the target layer is considered. "Multiplex Information" corresponds to the case when both intralayer information of the target layer and interlayer information of other layers are considered.

each other, and thus a fraction of HP-distance is added to the weight of that common neighbor; otherwise it is computed as the original CN. Therefore, in the WCN method the two similarity measures are dependent to each other.

Discussion

In this work, two novel methods based on the hyperbolic geometry of the multiplex networks are proposed to discover spurious and missing links in multiplex networks. The hyperbolic underlying of complex networks considers two parameters of popularity and similarity of nodes that both play important role in link prediction problem. Since the common local similarity measures mostly consider only the node degree (popularity), we suggest to enhance their predictability by adding the similarity feature to them. As we can see, in the case of missing link prediction specifically in social networks, each node is more likely to connect to nodes with similar features (his friends) as well as popular nodes (influencers). Another hypothesis is that interlayer relevance can be helpful in link prediction. Based on this hypothesis, recently a new method was proposed that considered the existing similarity measures in both target layer and other layers and combined the similarity measures via a correlation metric (Link Overlap) and obtained a multiplex-based similarity measure for spurious link prediction[21]. Based on this research, new measures are proposed based on the hyperbolic geometry of the network. First, a number of existing similarity measures which are widely used for the link prediction are chosen and then new measures are proposed to solve the spurious and missing link prediction problem. Our experimental results on four synthetic networks and six real-world multiplex networks shows that the new proposed measures outperform in all cases and also incorporating the interlayer information can improve the prediction performance compared with the case that only intralayer information is considered.

References

1. Jalili, M., Orouskhani, Y., Asgari, M., Alipourfard, N. & Perc, M. Link prediction in multiplex online social networks. *Royal Society open science* **4**, 160863 (2017).
2. Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences* **106**, 22073–22078 (2009).
3. Lin, D. An information-theoretic definition of similarity. In *Icml*. 296–304 (1998).
4. Celisse, A., Daudin, J.-J. & Pierre, L. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics* **6**, 1847–1899 (2012).
5. Clauset, A., Moore, C. & Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98 (2008).
6. Papadopoulos, F., Aldecoa, R. & Krioukov, D. Network geometry inference using common neighbors. *Physical Review E* **92**, 022807 (2015).

7. Wang, Z., Wu, Y., Li, Q., Jin, F. & Xiong, W. Link prediction based on hyperbolic mapping with community structure for complex networks. *Physica A: Statistical Mechanics and its Applications* **450**, 609–623 (2016).
8. Cardillo, A. G ómez-Gardenes. *J*., *Zanin*, *M*., *Romance*, *M*., *Papo*, *D*., *del Pozo*, *F*., *& Boccaletti*, *S* (2013).
9. Nicosia, V., Bianconi, G., Latora, V. & Barthelemy, M. Growing multiplex networks. *Physical review letters* **111**, 058701 (2013).
10. Szell, M., Lambiotte, R. & Thurner, S. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences* **107**, 13636–13641 (2010).
11. Kivelä, M. *et al*. Multilayer networks. *Journal of complex networks* **2**, 203–271 (2014).
12. Boccaletti, S. *et al*. The structure and dynamics of multilayer networks. *Physics Reports* **544**, 1–122 (2014).
13. Lee, K.-M., Min, B. & Goh, K.-I. Towards real-world complexity: an introduction to multiplex networks. *The European Physical Journal B* **88**, 48 (2015).
14. Pujari, M. & Kanawati, R. Link prediction in multiplex networks. *NHM* **10**, 17–35 (2015).
15. Hristova, D., Noulas, A., Brown, C., Musolesi, M. & Mascolo, C. A multilayer approach to multiplexity and link prediction in online geo-social networks. *EPJ Data Science* **5**, 24 (2016).
16. Sharma, S. & Singh, A. An efficient method for link prediction in complex multiplex networks. In *11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). IEEE,* 453–459 (2015).
17. Yao, Y. *et al*. Link prediction via layer relevance of multiplex networks. *International Journal of Modern Physics C* **28**, 1750101 (2017).
18. Hajibagheri, A., Sukthankar, G. & Lakkaraju, K. A holistic approach for link prediction in multiplex networks. In *International Conference on Social Informatics. Springer,* 55–70 (2016).
19. Zeng, A. & Cimini, G. Removing spurious interactions in complex networks. *Physical Review E* **85**, 036101 (2012).
20. Zhang, P., Zeng, A. & Fan, Y. Identifying missing and spurious connections via the bi-directional diffusion on bipartite networks. *Physics Letters A* **378**, 2350–2354 (2014).
21. Samei, Z. & Jalili, M. Discovering spurious links in multiplex networks based on interlayer relevance. *Journal of Complex Networks* (2019).
22. Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A. & Boguná, M. Hyperbolic geometry of complex networks. *Physical Review E* **82**, 036106 (2010).
23. Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M. & Krioukov, D. Popularity versus similarity in growing networks. *Nature* **489**, 537 (2012).
24. Papadopoulos, F., Psomas, C. & Krioukov, D. Network mapping by replaying hyperbolic growth. *IEEE/ACM Transactions on Networking (TON)* **23**, 198–211 (2015).
25. Alanis-Lobato, G., Mier, P. & Andrade-Navarro, M. A. Manifold learning and maximum likelihood estimation for hyperbolic network embedding. *Applied Network Science* **1**, 10 (2016).
26. Muscoloni, A. & Cannistraci, C. V. Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction. *New Journal of Physics* (2018).
27. Muscoloni, A. & Cannistraci, C. V. A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *New Journal of Physics* **20**, 052002 (2018).
28. Muscoloni, A., Thomas, J. M., Ciucci, S., Bianconi, G. & Cannistraci, C. V. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nature communications* **8**, 1615 (2017).
29. Muscoloni, A. & Cannistraci, C. V. Minimum curvilinear automata with similarity attachment for network embedding and link prediction in the hyperbolic space. *arXiv preprint arXiv:1802.01183* (2018).
30. Bianconi, G. Statistical mechanics of multiplex networks: Entropy and overlap. *Physical Review E* **87**, 062806 (2013).
31. Battiston, F., Nicosia, V. & Latora, V. Structural measures for multiplex networks. *Physical Review E* **89**, 032804 (2014).
32. Gemmetto, V. & Garlaschelli, D. Multiplexity versus correlation: the role of local constraints in real multiplexes. *Scientific reports* **5**, 9120 (2015).
33. Lee, K.-M., Kim, J. Y., Cho, W.-K., Goh, K.-I. & Kim, I. Correlated multiplexity and connectivity of multiplex random networks. *New Journal of Physics* **14**, 033027 (2012).
34. Pan, L., Zhou, T., Lü, L. & Hu, C.-K. Predicting missing links and identifying spurious links via likelihood analysis. *Scientific reports* **6**, 22955 (2016).
35. Daminelli, S., Thomas, J. M., Durán, C. & Cannistraci, C. V. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. *New Journal of Physics* **17**, 113037 (2015).
36. Kleineberg, K.-K., Boguná, M., Serrano, M. Á. & Papadopoulos, F. Hidden geometric correlations in real multiplex networks. *Nature Physics* **12**, 1076 (2016).
37. Lü, L. & Zhou, T. Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)* **89**, 18001 (2010).
38. de Borda, J. C. Mémoire sur les élections au scrutin (1781).
39. Taylor, D., Shai, S., Stanley, N. & Mucha, P. J. Enhanced detectability of community structure in multilayer networks through layer aggregation. *Physical review letters* **116**, 228301 (2016).
40. Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).
41. Vickers, M. & Chan, S. Representing classroom social structure. *Victoria Institute of Secondary Education*, *Melbourne* (1981).
42. Lazega, E. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. (Oxford University Press on Demand, 2001).
43. Snijders, T. A., Pattison, P. E., Robins, G. L. & Handcock, M. S. New specifications for exponential random graph models. *Sociological methodology* **36**, 99–153 (2006).
44. Coleman, J., Katz, E. & Menzel, H. The diffusion of an innovation among physicians. *Sociometry* **20**, 253–270 (1957).
45. Chen, B. L., Hall, D. H. & Chklovskii, D. B. Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences* **103**, 4723–4728 (2006).
46. De Domenico, M., Porter, M. A. & Arenas, A. MuxViz: a tool for multilayer analysis and visualization of networks. *Journal of Complex Networks* **3**, 159–176 (2015).
47. De Domenico, M., Nicosia, V., Arenas, A. & Latora, V. Structural reducibility of multilayer networks. *Nature communications* **6**, 6864 (2015).
48. Stark, C. *et al*. BioGRID: a general repository for interaction datasets. *Nucleic acids research* **34**, D535–D539 (2006).

Author Contributions

Z.S. conceived the study, performed the experiments, analyzed the data, and wrote the manuscript. M.J. analyzed the results and wrote the paper. Both authors approved the final version of the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

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