# *Common neighbour structure and similarity intensity in complex networks*

# Common neighbour structure and similarity intensity in complex networks

Lei Hou[a], Kecheng Liu[a,b]

[a]*Informatics Research Centre, Henley Business School, University of Reading, Reading, Berkshire RG6 6UD, United Kingdom.*
[b]*Data Science and Cloud Service Research Centre, Shanghai University of Finance and Economics, Shanghai 200433, China.*

**Abstract**

Complex systems as networks always exhibit strong regularities, implying underlying mechanisms governing their evolution. In addition to the degree preference, the similarity has been argued to be another driver for networks. Assuming a network is randomly organised without similarity preference, the present paper studies the expected number of common neighbours between vertices. A symmetrical similarity index is accordingly developed by removing such expected number from the observed common neighbours. The developed index can not only describe the similarities between vertices, but also the dissimilarities. We further apply the proposed index to measure of the influence of similarity on the wring patterns of networks. Fifteen empirical networks as well as artificial networks are examined in terms of similarity intensity and degree heterogeneity. Results on real networks indicate that, social networks are strongly governed by the similarity as well as the degree preference, while the biological networks and infrastructure networks show no apparent similarity governance. Particularly, classical network models, such as the Barabási-Albert model, the Erdös-Rényi model and the Ring Lattice, cannot well describe the social networks in terms of the degree heterogeneity and similarity intensity. The findings may shed some light on the modelling and link prediction of different classes of networks.

*Keywords:* Complex networks; vertex similarity; common neighbours

## 1. Introduction

Networks can efficiently describe a wide range of complex systems, such as social systems, biological systems and infrastructure systems [1, 2, 3, 4]. Since most real-world networks are either incomplete or evolving, to understand the dynamics and growing patterns of networks has attracted increasing attentions [5, 6, 7, 8, 9, 10, 11].

The degree preference has been considered as the key attractiveness driving the evolution of networks [12, 13, 14, 15] since the finding of scaling phenomena [16]. However, real networks are also found to be highly clustered [17] and with dense community structure [18, 19] which cannot be explained by the preferential attachment mechanism alone. Accordingly, vertex similarity is also argued to be a driver for networks [20] and has been applied to study the formation and evolution of different networks [21, 22, 23, 24]. While the ground-truth similarities among vertices are mostly unknown, a number of similarity indices have been developed by evaluating either the adjacency matrix or the common neighbour structure of the network [25, 26, 27, 28, 29]. Normally, the vertices that share the same neighbours (adjacent vertices) are considered to be similar to each other. However, the similarities quantified by these indices mostly have systematic bias regarding the vertex degree [6, 26, 30] that hub vertices tend to have more common neighbours with others due to their rich connectivities. As a consequence, it is difficult to determine whether the common neighbours are due to the similarity between vertices or just random mechanism. Additionally, most indices give only positive values without an indication of neutral similarity. Even with a same similarity value, the meaning would be different in different scenarios such as the degrees of the measured vertices and the degree distribution of the given network. For example, two vertices $\alpha$ and $\beta$ having 5 common neighbours could indicate that they are extremely similar to each other if their degrees are $k_\alpha = k_\beta = 5$, but could also be interpreted as extremely dissimilar if their degrees $k_\alpha \approx N, k_\beta \approx N$ where $N$ is the network size, because they are expected to have a lot more common neighbours. Therefore, the key question needs to be answered is that how many common neighbours two particular vertices are expected to share in a given network. Finally, to what extent does the similarity shape the structure and evolution of a given network is still an open question due to the lack of an unbiased and symmetrical similarity index.

In this paper we study the expected number of common neighbours between two vertices which is shown to be determined by the degree heterogeneity of the network. A vertex similarity index is thereby proposed by comparing the number of common neighbours with the expected number so that the random factors are removed. We further define the similarity intensity to quantify the governance of similarity in complex networks as the average similarity over all the connected vertex pairs in the network. The similarity intensities and degree heterogeneities of fifteen real networks are investigated and the social networks are found to be a special class which has both high degree heterogeneity and similarity intensity.

*Email addresses:* l.hou@pgr.reading.ac.uk (Lei Hou), k.liu@henley.ac.uk (Kecheng Liu)

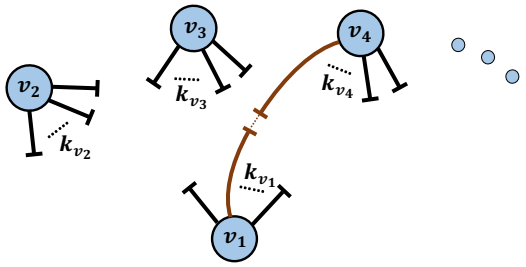


Fig. 1: (Color online.) Illustration of the random rewiring. Each vertex $v$ in the network has $k_v$ half-edges to be paired with others' and each pair of half-edges has equal chance to be connected. Obviously, vertices with more half-edges are more likely to be connected to each other.

## 2. A Balanced vertex similarity index

The vertices that share common neighbours are usually considered to be similar to each other. However, two vertices $x$ and $y$ that are not similar to each other at all, especially these with large degrees, could still have common neighbours by chance. For example, in a network of 10 vertices, $x$ and $y$ with degrees $k_x = k_y = 6$ should have at least 3 common neighbours, but having 3 common neighbours does not mean that they are similar. In other words, every pair of vertices $x$ and $y$ with no similarity are expected to have a certain amount of common neighbours $n_{xy}^{exp}$ due to pure random mechanism. In a given network, if the observed number of common neighbours $n_{xy} = n_{xy}^{exp}$, we can consider these two vertices $x$ and $y$ to be neutral to each other. Accordingly, the difference between the observed and expected number of common neighbours $n_{xy} - n_{xy}^{exp}$ can be used to describe the tendency of $x$ and $y$ to connect the same vertices, which we argue is a more meaningful way to represent their similarity. Therefore, we calculate the expected number of common neighbours between two vertices with given degrees in a given network, so that we can remove the random-caused common neighbours from the observed number to estimate their similarity.

Consider a network consisting of a set of $N$ vertices $V = \{v_1, v_2, \cdots, v_N\}$, and a set of $M$ edges $E = \{e_1, e_2, \cdots, e_M\}$. The expected number of common neighbours between two vertices can be calculated by considering a random rewiring process of a network. Assume all the edges are broken into two half-edges (stubs) and thus each vertex $v$ has $k_v$ half-edges to be paired again with others as shown in Fig. 1. This process is normally referred as the configuration model [2, 31] which generates random networks with a given degree sequence. In the rewiring process of the present paper, for each of $v$'s half-edges, the paired half-edge is chosen randomly but from another vertex that has not been connected by $v$ to avoid multi edges or self-loops. Therefore, the probability of the paired half-edge coming form vertex $j$ is $k_j / \sum_v k_v$. Considering all the $k_i$ edges that vertex $i$ possessing, we have the probability of two random vertices $i$ and $j$ connecting with each other [32, 33],

$$p(i \leftrightarrow j) = \frac{k_i k_j}{\sum_v k_v}. \tag{1}$$

Accordingly, the probability of a vertex $i$ being a common neighbour for vertices $x$ and $y$, i.e. connecting to both $x$ and $y$, can be written as,

$$p(i \leftrightarrow x, y) = \frac{k_i(k_i - 1)}{(\sum_v k_v)^2} \cdot k_x k_y. \tag{2}$$

Considering all the possible common neighbours, we then have the expected number of common neighbours for $x$ and $y$ which reads,

$$n_{xy}^{exp} = \sum_i p(i \leftrightarrow x, y) = \frac{\sum_v k_v(k_v - 1)}{(\sum_v k_v)^2} \cdot k_x k_y. \tag{3}$$

Therefore, as suggested by Eq. (3), the neighbourhood size for two vertices $x$ and $y$ is expected to have a linear relation with the product of their degrees, i.e. $n_{xy}^{exp} \propto k_x k_y$. We test such relation using the Barabási-Albert (BA) network model [16]. The BA model is a random network model in which the edges are attached randomly according to the degree preference without predefined similarity. Accordingly, the vertices in a BA network are expected to be with no similarity and thus we should have $n_{xy}^{exp} = n_{xy}$. As shown in Fig. 2, the averaged number of common neighbours for two vertices $x$ and $y$ has the linear correlation with the product $k_x k_y$ as predicted by the Eq. (3).

Actually, one can find that, in Eq. (3), $\sum_v k_v$ can be given by the product of the network size and the average degree, $N\langle k\rangle$. Accordingly, we have also $\sum_v k_v(k_v - 1) = N\left(\langle k^2\rangle - \langle k\rangle\right)$. Therefore, we can rewrite the expression for the expected number of common neighbours as

$$n_{xy}^{exp} = \frac{\langle k^2\rangle - \langle k\rangle}{N\langle k\rangle^2} \cdot k_x k_y. \tag{4}$$

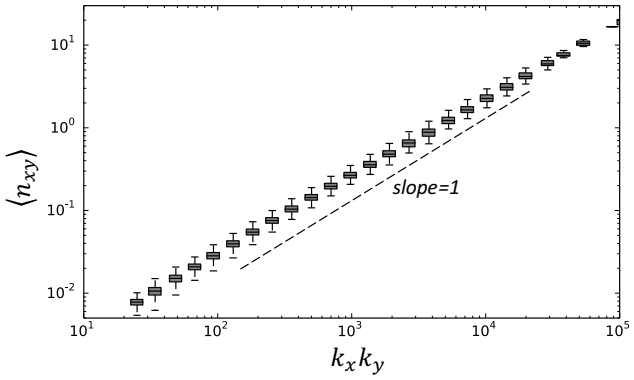


Fig. 2: Number of common neighbours between two vertices $x$ and $y$, $n_{xy}$ versus the product of the corresponding vertices' degrees $k_x k_y$ in BA networks. The dashed straight line has a slope of 1 in the log-log plot. The simulated network starts from a complete network of $m_0 = 6$ vertices. At each of the following step, one vertex is added to the network to connect to $m = 5$ existing vertices. The probability of each vertex being connected is proportional to its current degree, i.e. $p(v) \propto k_v$. Vertices are added continuously until the network size reach $N = 10^4$. Considering most vertex pairs would have no common neighbour at all in a single realisation of network, we average $n_{xy}$ over $10^4$ realisations of the generated BA network. We rewire the generated BA network as follows: a) select two from $N\langle k\rangle/2$ edges uniformly at random; b) chose one vertex from each edge and switch if this will not result in multi edges or self-loops; c) repeat a) and b) for $N\langle k\rangle$ times. In such way, the degree of each vertex will not be changed and we can average the number of common neighbours between two specific vertices accordingly.

The parameter for the product of the degrees basically describes the degree distribution feature of the network. The component $\langle k^2\rangle/\langle k\rangle^2$ is usually used to described a network's degree heterogeneity $H$ [29, 34]. With a unified degree for each vertex, a network has $\langle k^2\rangle = \langle k\rangle^2$ and thus heterogeneity $H = 1$. The more heterogeneous the network's degree distribution is, the higher the value $H$ would generally be. The BA network with the applied settings in this paper has a degree heterogeneity $H = 2.79 \pm 0.08$. The parameter here is thus a function of the degree heterogeneity. Here we define it as a heterogeneity parameter denoting with $\mathcal{H}$, which consequently reads,

$$\mathcal{H} = \frac{1}{N}\cdot\left(\frac{\langle k^2\rangle}{\langle k\rangle^2} - \frac{1}{\langle k\rangle}\right) = \frac{1}{N}\cdot\left(H - \frac{1}{\langle k\rangle}\right). \tag{5}$$

Introducing Eq. (5) to Eq. (4) gives us the final expression for the expected number of common neighbours for two randomly given vertices $x$ and $y$,

$$n_{xy}^{exp} = \mathcal{H}\cdot k_x k_y. \tag{6}$$

Basically, the more heterogeneous the degree is, the more common neighbours two vertices with given degrees would share, and on the other hand, vertices with higher degrees are likely to have more common neighbours with others. With the expected number of common neighbours $n_{xy}^{exp}$ as the estimation for the random component $n_{xy}^{rand}$, we can then define the similarity between vertices $x$ and $y$ as

$$s_{xy} = n_{xy} - n_{xy}^{exp} = |\Gamma_x \cap \Gamma_y| - \mathcal{H}\cdot k_x k_y, \tag{7}$$

where $\Gamma_v$ is the set of vertices that are connecting to vertex $v$ and $|\Gamma|$ gives the number of vertices in the set. Thus, the defined similarity $s_{xy}$ indicates how many more (or less) common neighbours are vertices $x$ and $y$ sharing than expected. If the number of common neighbours is the same to expected, i.e. $s_{xy} = 0$, one can then consider $x$ and $y$ to be neutral to each other. On the other hand, if the vertices $x$ and $y$ share more (less) neighbours, i.e. $s_{xy} > 0$ ($s_{xy} < 0$), they are suggested to be similar (dissimilar) to each other.

Actually in *ref.* [26], a similarity index normally referred as the LHN index was proposed with considerations similar to that in this paper. They derived the expected number of paths between two vertices with length of two, which is in another word the number of common neighbours. Although the same expression for the expected number of common neighbours was derived, they defined the vertex similarity $s_{xy}^{LHN}$ by dividing the real number by the expected number, i.e. [26],

$$s_{xy}^{LHN} = \frac{|\Gamma_x \cap \Gamma_y|}{k_x k_y}. \tag{8}$$

While such definition has shown accuracy in estimating the similarities in many networks, we believe our definition shown in Eq. (7) has advantages in following aspects. The real networks are usually extremely sparse, and thus a significant amount of vertex pairs will share no common neighbours at all. For such vertices, the LHN index considers the similarities uniformly to be zero. However, two hub vertices having no common neighbours has apparently different meaning from two low-degree vertices sharing no neighbours. The Eq. (7) is able to estimate the similarity for vertex pairs sharing even no neighbours. Additionally, the index defined in this paper may yield negative values when the number of common neighbours are less than expected (random case), which can be regarded as the dissimilarity between the measured vertices. Especially, with the random case as the baseline, we can apply the defined similarity index to explore that whether, or to what extent, is the similarity governing the complex networks (to be discussed in the following section).

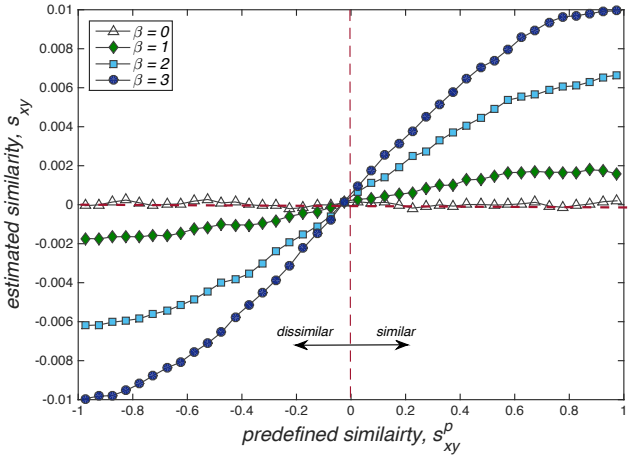


Fig. 3: (Color online.) Estimated similarity using the proposed index in the networks where the similarities are predefined. The vertical and horizontal dashed lines shows the neutral case for the predefined and estimated similarity respectively. The results are achieved with a network size of $N = 10^4$. With the same angular position for each vertex (thus same predefined similarity between each vertex pair), we generated $10^3$ networks. Accordingly, the estimated similarities are averaged over all vertex pairs with the same predefined similarity level.

To test the accuracy of the proposed similarity index, we introduce the influence of similarities into the BA network model. We randomly assign an angular position $\theta$ to each vertex. Vertices near to each other (with small angular distance), are considered to be similar to each other. Therefore, the predefined similarity between two vertices $x$ and $y$ can be written as $s_{xy}^p = 1 - 2\Delta\theta_{xy}/\pi$, where $\Delta\theta_{xy}$ is the angular distance between the two vertices, i.e. $\Delta\theta_{xy} = \pi - |\pi - |\theta_x - \theta_y||$. Thus the larger the similarity $s_{xy}^p$ is, the more similar the vertices are considered to be. Instead of letting new vertex attach each of its $m$ edges to an existing vertex $i$ with probability proportional to only the degree, i.e. $\Pi \propto k_i$, we define the probability of connecting $i$ as $\Pi \propto k_i/(1 + e^{-\beta s_{xy}^p})$, where $\beta$ is a parameter controlling the influence of similarity. The case $\beta = 0$ gives the standard BA model where the edges are attached according to only the degree with no enhancement from the similarity. For any positive $\beta$, the similar vertices are more likely to connect with each other and the larger the parameter $\beta$ is, the stronger the influence of similarity would be governing the network evolution. Additionally, with such mechanism, positive predefined similarities $s_{xy}^p > 0$ will enhance the probability of attachment while negative values reduce such probability. As shown in Fig. 3, the proposed

similarity index can recover the predefined similarity by examining the network wiring patterns. The larger the parameter $\beta$ is, the more distinguishable the estimated similarity would be. Most importantly, the proposed index is accurate in detecting whether the vertices are similar, dissimilar or neutral to each other. On the other hand, most other similarity indices, though can tell which vertex pair is more similar in comparison to other vertex pairs, cannot show it is similar or dissimilar for a specific pair of vertices.

## 3. Similarity intensity of networks

Since we have defined a symmetrical similarity index which can be used to detect whether two vertices are similar or dissimilar to each other in comparison to the random case, in this section we examine the connected vertices in a given network to explore whether and to what extent the edges are established according to the similarity. For each edge $e$, we examine the similarity between the two vertices $e_x$ and $e_y$ on the ends, $s_{e_xe_y}$. Note that, as $e_x$ and $e_y$ have already connected to each other, in the calculation of similarity, we exclude this edge from the vertex degrees, leading the similarity to $s_{e_xe_y} = |\Gamma_{e_x} \cap \Gamma_{e_y}| - \mathcal{H}(k_{e_x} - 1)(k_{e_y} - 1)$. Accordingly, we define the similarity intensity of the network $\mathcal{S}$ as the average similarity of every pair of connected vertices, which reads,

$$\mathcal{S} = \frac{1}{|E|} \sum_{e \in E} s_{e_xe_y}. \tag{9}$$

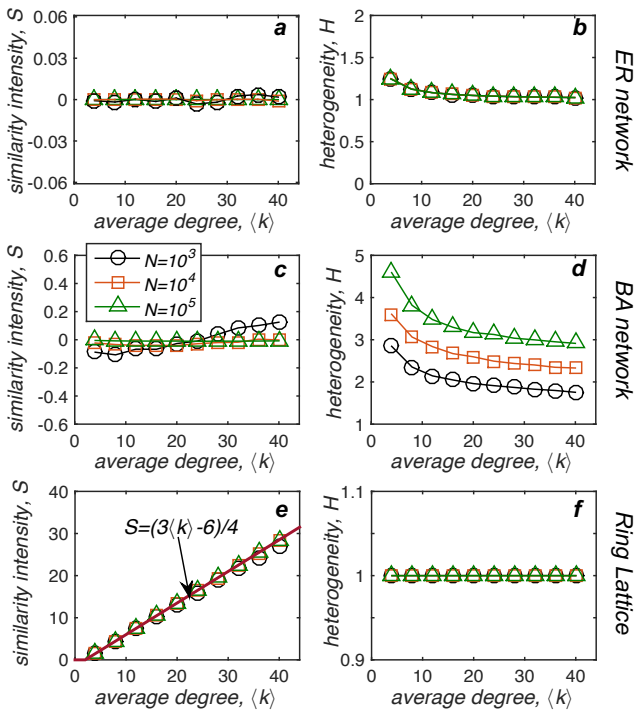


Fig. 4: (Color online.) Similarity intensity $\mathcal{S}$ and degree heterogeneity $H$ in ER networks, BA networks and Ring Lattices. For the ER and BA networks, the results for each size and average degree are averaged over 50 independent realisations.

Therefore, a positive value of $\mathcal{S}$ suggests that the connected vertices share more common neighbours than expected which implies that the evaluated network is shaped by the similarity. On the other hand, a neutral value $\mathcal{S} = 0$ indicates that the formation of the network is irrelevant to the similarity. Additionally, larger values suggest strong governance of similarity in the network evolution.

We firstly analyse the similarity intensity $\mathcal{S}$ and degree heterogeneity $H$ of artificial networks, including the ER network [36], BA network [16], and Ring Lattice. In particular, we study the influence of edge densities [35] on the two features.

The ER random network takes a fixed probability for each vertex pair to establish an edge. Since the edges are established randomly, the ER networks have no similarity preference and thus one should expect a neutral similarity intensity $\mathcal{S} = 0$. As expected, the similarity intensity of ER networks is shown by Fig. 4(a) to be neutral regardless of the network size and average degree. Additionally, following a Poisson degree distribution, the ER random networks' degree heterogeneity $H$ is very close to the lower-limit 1 as shown in Fig. 4(b).

The BA network which introduces the degree preference to model the power-law degree distribution observed in real networks, has been considered as a standard heterogeneous network and thus has a relatively high degree heterogeneity. As shown in Fig. 4(d), the degree heterogeneity of BA networks is correlated with the network size and the average degree, but always takes value that is significantly larger than 1. On the other hand, the edges in BA networks are attached purely according to the degree preference, the similarity intensity $S$ takes a neutral value (close to 0) similar to the ER random network as shown in Fig. 4(c).

Different from the ER and BA networks, the ring lattice is a regular network which places $N$ vertices evenly on a circle and lets each vertex connect to its $\langle k \rangle$ nearest neighbours. Therefore, every vertex has exactly the same degrees and thus the degree heterogeneity is $H = 1$ for ring lattice regardless of the size and average degree (Fig. 4(f)). On the other hand, since the edges are established according to the positions, the vertices near (similar) to each other will have many common neighbours leading to a high degree heterogeneity as shown in Fig. 4(e). Assuming the vertices are numbered according to their positions, the neighbours of an arbitrary vertex $i$ will be $V_i = \{i - \langle k \rangle/2, \cdots, i-1, i+1, \cdots, i + \langle k \rangle/2\}$. The number of common neighbours between $i$ and $j$ ($j \in V_i$) can be given by $n_{ij} = \langle k \rangle - |j - i| - 1$. Accordingly, the average number of common neighbours for vertex pairs involving $i$ is

$$\begin{aligned} \langle n_i \rangle &= \frac{\sum_{j \in V_i} (\langle k \rangle - |j - i| - 1)}{\langle k \rangle} \\ &= \frac{\langle k \rangle^2 - \langle k \rangle - 2\sum_{m=1}^{\langle k \rangle/2} m}{\langle k \rangle} = \frac{3\langle k \rangle - 6}{4}. \end{aligned} \tag{10}$$

Accordingly we can theoretically have the similarity intensity of an ring lattice to be

$$\mathcal{S}^{ring} = \frac{3\langle k \rangle - 6}{4} - \mathcal{H} \cdot \langle k_x k_y \rangle = \frac{3\langle k \rangle - 6}{4} - \frac{\langle k \rangle^2}{N}\left(1 - \frac{1}{\langle k \rangle}\right). \tag{11}$$

Table 1: Statistics of networks applied in this paper. In the table, $N$ and $M$ represent the number of vertices and edges respectively; $C$ is the clustering coefficient [17]; $r$ is the degree assortativity coefficient [37]; $H$ represent the degree heterogeneity, i.e. $H = \langle k^2\rangle/\langle k\rangle^2$; and the $S$ is the defined similarity intensity. In the Coauthorship network, vertices are authors and an edge represents at least one common publication between two authors. The Facebook, Yelp, Gowalla and Flixster are social networking websites where users (vertices) can establish online friendships (edges) with others. The Trust network is based on an encryption program, entitled Pretty-Good-Privacy (PGP) where vertices are certificates and an edge represents authorisation from the owner of a certificate to that of another. The Email network describes the email exchanges (edges) between employees (vertices) of the company Enron. The Yeast and PDZBase networks are the metabolic interactions (edges) between proteins (vertices). For the Road networks of Pennsylvania (PA.) and California (CA.), a road is an edge connecting intersections as vertices. For the power grid, either a generator, a transformer or a substation is regarded as a vertex while the supply lines are regarded as edges. The animal networks regards animals, i.e. dolphins, zebras and kangaroos respectively, as vertices and there will be an edge connecting two individuals if they have at least one interaction during observation. All the empirical networks are considered as simple graphs, i.e. unweighted, undirected.

| Network Type | Network | $N$ | $M$ | $\langle k\rangle$ | $C$ | $r$ | $H$ | $\mathcal{S}$ |
|---|---|---|---|---|---|---|---|---|
| Social | Coauthorship [38] | 18771 | 198050 | 21.1 | 0.63 | 0.45 | 3.09 | 19.65 |
| | Facebook [39] | 63731 | 817035 | 25.64 | 0.22 | 0.42 | 3.43 | 12.36 |
| | Trust (PGP) [20] | 10680 | 24316 | 4.55 | 0.26 | 0.42 | 4.14 | 6.58 |
| | Email [40] | 36692 | 183831 | 10.02 | 0.49 | 0.13 | 13.97 | 7.1 |
| | Yelp [1] | 174097 | 1288077 | 14.79 | 0.11 | 0.18 | 15.79 | 9.03 |
| | Gowalla [41] | 196591 | 950327 | 9.66 | 0.23 | 0 | 31.71 | 3.41 |
| | Flixster [42] | 2523386 | 7918801 | 6.27 | 0.08 | 0.11 | 35.07 | 2.73 |
| Biological | Yeast [43] | 1846 | 2203 | 2.38 | 0.06 | 0.04 | 2.72 | 0.28 |
| | PDZBase [44] | 212 | 242 | 2.28 | 0 | 0 | 2.33 | -0.08 |
| Infrastructure | Rooad (PA.) [40] | 1088092 | 1541898 | 2.83 | 0.04 | 0.26 | 1.12 | 0.13 |
| | Power grid [17] | 4941 | 6594 | 2.66 | 0.08 | 0.18 | 1.45 | 0.29 |
| | Road (CA.) [40] | 1965206 | 2776607 | 2.82 | 0.04 | 0.99 | 13.86 | -2.35 |
| Animal | Dolphin [45] | 61 | 159 | 5.16 | 0.26 | 0.24 | 1.32 | 1.17 |
| | Zebra [46] | 27 | 111 | 8.22 | 0.87 | 0.81 | 1.33 | 3.51 |
| | Kangaroo [47] | 17 | 91 | 10.7 | 0.82 | 0.11 | 1.13 | 1.88 |

For any ring lattice in which $N \gg \langle k\rangle^2$, we can approximately have $\mathcal{S} = (3\langle k\rangle - 6)/4$. Therefore, as shown by Fig. 4(e), the similarity intensity of ring lattice is closely correlated with the average degree, but generally irrelevant to the network size.

We further examine the similarity intensity and degree heterogeneity of empirical networks. Table 1 shows the statistics of fifteen studied real networks including social networks, biological networks, infrastructure networks, animal networks. For each of the empirical networks, we calculate its similarity intensity $\mathcal{S}$ and degree heterogeneity $H$ and the results are shown in Fig. 5. Additionally, we also compare the empirical networks with artificial networks with controlled network size and total links as shown in Table 2.

The biological networks are shown to have high degree heterogeneity and neutral similarity intensity which is very similar to the BA network. The infrastructure networks show different results on the degree heterogeneity that, while the road network in Pennsylvania as well as the power grid are less heterogeneous, the road network in California has extremely heterogeneous degrees. On the other hand, the similarity of infrastructure networks are neutrally or even negatively shaping the structure. Actually, the wiring patterns of such networks are constructed according to the geographical locations of the vertices (intersections in road network, generators/transformers/substations in power grid) which can be regard as location similarity. However, to achieve high efficiency, vertices in infrastructure networks, even geographically near to each other, would not share many common neighbours. Especially in road networks, the vertices (intersections) are mostly organised in squares resulting in second-order common neighbours rather than in triangles resulting in direct common neighbours. For the animal networks, the features are opposite to the biological networks and BA networks in terms of the degree

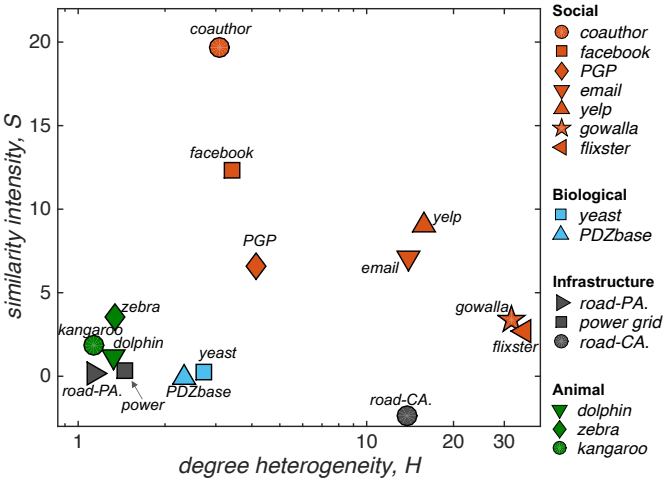


Fig. 5: (Color online.) The similarity intensity $\mathcal{S}$ versus the degree heterogeneity $H$ of networks. While large $H$ means the vertex degrees are very different (heterogeneous) from each other, the lower-limit $H = 1$ represents the case where each vertex has the same degree $k_v = \langle k\rangle, \forall v$. Large (positive) $\mathcal{S}$ indicates the edges tend to establish between similar vertices while small (negative) values suggest the edges tend to connect dissimilar vertices.

Table 2: Comparison of degree heterogeneity $H$ and similarity intensity S between empirical networks and artificial networks including ER, BA and Ring Lattice. For each empirical network, the artificial networks are generated according to its network size $N$. For ER networks, the probability of each vertex pair connecting each other is set to be $p = 2M(N-1)/N$. For BA networks, we set $m_0 = 2M/N + 1$ and $m = M/N$. As to the ring lattices, we let each vertex to connect 2M/N nearest neighbours.

| | | Degree Heterogeneity $H$ | | | | Similarity Intensity $\mathcal{S}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Empirical | ER | BA | Ring | Empirical | ER | BA | Ring |
| Social | Coauthorship | 3.09 | 1.02 | 2.46 | 1 | 19.65 | 0 | 0 | 29.91 |
| | Facebook | 3.43 | 1.02 | 2.71 | 1 | 12.36 | 0 | 0 | 35.96 |
| | Trust (PGP) | 4.14 | 1.12 | 2.96 | 1 | 6.58 | 0 | -0.03 | 4.49 |
| | Email | 13.97 | 1.04 | 2.87 | 1 | 7.1 | -0.01 | -0.02 | 13.49 |
| | Yelp | 15.79 | 1.03 | 3.17 | 1 | 9.03 | 0 | -0.01 | 19.49 |
| | Gowalla | 31.71 | 1.05 | 3.44 | 1 | 3.41 | 0 | -0.01 | 11.99 |
| | Flixster | 35.07 | 1.11 | 4.93 | 1 | 2.73 | 0 | -0.01 | 2.99 |
| Biological | Yeast | 2.72 | 1.08 | 3.03 | 1 | 0.28 | 0 | -0.03 | -0.01 |
| | PDZBase | 2.33 | 1.14 | 2.33 | 1 | -0.08 | 0 | -0.07 | -0.01 |
| Infrastructure | Rooad (PA.) | 1.12 | 0.99 | 6.62 | 1 | 0.13 | -0.01 | -0.01 | 0 |
| | Power grid | 1.45 | 1.13 | 4.41 | 1 | 0.29 | 0 | -0.03 | -0.01 |
| | Road (CA.) | 13.86 | 1.13 | 8.19 | 1 | -2.35 | 0 | -0.01 | 0 |
| Animal | Dolphin | 1.32 | 1.05 | 2.24 | 1 | 1.17 | 0 | -0.45 | 1.38 |
| | Zebra | 1.33 | 1.03 | 1.33 | 1 | 3.51 | 0 | 0.55 | 2.91 |
| | Kangaroo | 1.13 | 1.19 | 1.09 | 1 | 1.88 | 0 | 1.66 | 1.71 |

heterogeneity and similarity intensity. Though with low degree heterogeneity, the similarity is shown to be playing a part in the interactions among animals. However, the sizes of the studied animal networks are quite small which may cause influences on their similarity intensities and degree heterogeneities.

Particularly, we address the social networks which are shown to be a special class of networks in terms of the degree heterogeneity and similarity intensity. The social networks have very heterogeneous degree distributions, normally more heterogeneous than the BA networks. While BA model can generate a power-law degree distribution with slope of 3, social networks in many cases may have much smaller slopes for the degree distribution leading to higher heterogeneities. A more interesting feature of social networks is the high similarity intensity. One can find that, the similarity intensities of social networks are sometimes similar or even higher (Trust network) than the ring lattice. In other words, for social networks, each connected vertex pair shares much more common neighbours than the random case on average. Such result suggests the extreme strong governance of similarity in human interactions. Social networks with both high degree heterogeneity and high similarity intensity, stand alone as a special class of networks in comparison to others. More importantly, one may find from Table 2 that the ER, BA and ring lattice cannot well describe the social networks in terms of the degree heterogeneity and similarity intensity simultaneously.

## 4. Conclusions and Discussion

Networked systems, though with high complexity, always show regularities. Mining the vertex similarities from the wiring pattens is of significance for understanding the structure and evolution of networks. In this paper, we theoretically studied the expected number of common neighbours between two vertices in random networks. Thereby, a new symmetrical similarity index was proposed by comparing the observed number of common neighbours with the expected number. We further defined the similarity intensity of networks and studied fifteen empirical networks as well as artificial network models. It was shown that, the social networks in general have high degree heterogeneities and are largely governed by similarity.

One of the major applications of similarity index is the link prediction. By evaluating the similarity between each unconnected vertex pair in a network, the most similar vertex pairs are considered to be the predictions of emerging edges. However, the hubs naturally yield more new edges than those poorly connected vertices. As a consequence, the similarity indices with emphasis on the high-degree vertices are found to be generally accurate for link predictions. Even the product of the vertex degrees $k_x k_y$ can be an accurate predictor of new links in some networks [29, 48]. Given the fact that the evolution of most real networks are normally driven by multiple mechanisms including preferential attachment and similarity-based attachment [20, 49, 50], to distinguish and separately predict links that resulted from different mechanisms is a better way to further improve the prediction accuracy and understand the evolution pattern. The proposed similarity index distinguishing the similarity-based common neighbours and the expected common neighbours, may provide new perspectives for the link prediction of networks with multiple mechanisms. The present paper defined the similarity intensity $\mathcal{S}$ to explore the governance of similarity on the structure of complex networks. This may be an indicator for the selection of link prediction method that whether to use a similarity-based method or a popularity-based

one. While similarity measures are found to have different performances when predicting links in different networks, the similarity intensity and the degree heterogeneity of the given network may be able to explain the different behaviours.

A number of artificial network models have been developed over the years, and many studies have been carried out based on these models to try to make implications for the understanding and control of the dynamics in real networks. But only if the network models can reveal the structural features of real networks, these theoretical studies could contribute to the knowledge of real-world systems. The examination of similarity intensity in this study provides a method to match the real networks with network models so that we can pick up the appropriate model according to the match to study with to make contributions to the target networks.

## Acknowledgments

This work is partially supported by a Key Project of National Natural Science Foundation of China (NSFC) with grant number 71532002.

[1] Albert R. and Barabási A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74** (2002) 1.

[2] Newman M. E. The structure and function of complex networks. *Soc. Ind. Appl. Math.*, **45** (2003) 167.

[3] Gao, J., Zhou, T. and Hu, Y. Bootstrap percolation on spatial networks. *Sci. Rep.*, **5** (2015) 14662.

[4] Holme, P. Core-periphery organization of complex networks. *Phys. Rev. E*, **72(4)** (2005) 046111.

[5] Dorogovtsev S. N., Mendes J. F. F., and Samukhin A. N. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, **85** (2000) 4633.

[6] Clauset A., Moore C., and Newman M. E. Hierarchical structure and the prediction of missing links in networks. *Nature*, **453** (2008) 98.

[7] Holme P. and Saramäki J. Temporal networks. *Phys. Rep.*, **519** (2012) 97.

[8] Lü L., Pan L., Zhou T., Zhang Y.-C., and Stanley H. E. Toward link predictability of complex networks. *Proc. Natl. Acad. Sci. USA*, **112** (2015) 201424644.

[9] Ren Z.-M., Zeng A., Chen D.-B., Liao H., and Liu J.-G. Iterative resource allocation for ranking spreaders in complex networks. *Europhys. Lett.*, **106** (2014) 48005.

[10] Liu J.-G., Hou L., Pan X., Guo Q., and Zhou T. Stability of similarity measurements for bipartite networks. *Sci. Rep.*, **6** (2016) 18653.

[11] Liu J.-G., Hu Z.-L., and Guo Q. Effect of the social influence on topological properties of user-object bipartite networks. *Eur. Phys. J. B* **86** (2013) 478.

[12] Fortunato S., Flammini A., and Menczer F. Scale-free network growth by ranking. *Phys. Rev. Lett.*, **96** (2006) 1.

[13] Ratkiewicz J., Fortunato S., Flammini A., Menczer F., and Vespignani A. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.*, **105** (2010) 8.

[14] Bagrow J. P. and Brockmann D. Natural emergence of clusters and bursts in network evolution. *Phys. Rev. X*, **3** (2013) 1.

[15] Pan, X., Hou, L., and Liu, K. Social influence on selection behaviour: Distinguishing local- and global-driven preferential attachment. *PLoS One*, **12(4)** (2017) e0175761.

[16] Barabási A.-L. and Albert R. Emergence of scaling in random networks. *Science*, **286** (1999) 509.

[17] Watts D. J. and Strogatz S. H. Collective dynamics of 'small-world' networks. *Nature*, **393** (1998) 440.

[18] Girvan M. and Newman M. E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, **99** (2002) 7821.

[19] Cui X.-M., Kim W. S., Hwang D.-U., and Kee S. Estimation of intermodular connectivity from the local field potentials in a hierarchical modular network. *Europhys. Lett.*, **110** (2015) 38001.

[20] Papadopoulos F., Kitsak M., Serrano M. A., Boguna M., and Krioukov D. Popularity versus similarity in growing networks. *Nature*, **489** (2012) 537.

[21] Crandall D. J., Backstrom L., Cosley D., Suri S., Huttenlocher D., and Kleinberg J. Inferring social ties from geographic coincidences. *Proc. Natl. Acad. Sci. USA*, **107** (2010) 22436.

[22] Gibson S. M., Ficklin S. P., Isaacson S., Luo F., Feltus F. A., and Smith M. C. Massive-scale gene co-expression network construction and robustness testing using random matrix theory. *PLoS One*, **8** (2013) e55871.

[23] Chen L.-J., Zhang Z.-K., Liu J.-H., Gao J., and Zhou T. A vertex similarity index for better personalized recommendation. *Physica A*, **466** (2017) 607.

[24] Zeng A., Vidmer A., Medo M., and Zhang Y.-C. Information filtering by similarity-preferential diffusion processes. *Europhys. Lett.*, **105** (2014) 58002.

[25] Tsourakakis, C. E. Toward quantifying vertex similarity in networks. *Internet Mathematics*, **10(3-4)** (2014) 263-286.

[26] Leicht E. A., Holme P., and Newman M. E. J. Vertex similarity in networks. *Phys. Rev. E*, **73** (2006) 026120.

[27] Shang, M. S., Zhang, Z. K., Zhou, T., and Zhang, Y. C. Collaborative filtering with diffusion-based similarity on tripartite graphs. *Physica A*, **389(6)** (2010) 1259-1264.

[28] Adamic L. A. and Adar E. Friends and neighbors on the web. *Soc. Networks*, **25** (2003) 211.

[29] Zhou T., Lü L., and Zhang Y. C. Predicting missing links via local information. *Eur. Phys. J. B*, **71** (2009) 623.

[30] Hou L., Liu K., Liu J.-G., and Zhang R. Solving the stability-accuracy-diversity dilemma of recommender systems. *Physica A*, **468** (2017) 415.

[31] Kang M. and Seierstad T. G. Phase transition of the minimum degree random multigraph process. *Random Struct. Algorithms*, **17** (2007) 1.

[32] Chung F. and Lu L. Connected components in random graphs with given expected degree sequences. *Ann. Comb.*, **6** (2002) 125.

[33] Chung F. and Lu L. The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA*, **99** (2002) 15879.

[34] Vespignani A. Modelling dynamical processes in complex socio-technical systems. *Nat. Phys.*, **8** (2012) 32.

[35] Vogt, M., Stumpfe, D., Maggiora, G. M., and Bajorath, J Lessons learned from the design of chemical space networks and opportunities for new applications. *J. Comput. Aided Mol. Des.*, **30(3)** (2016) 191-208.

[36] Erdös P. and Rényi A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, **5** (1960) 17.

[37] Newman M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.*, **89** (2002), 208701.

[38] Leskovec J., Kleinberg J., and Faloutsos C. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **1** (2007) 2.

[39] Viswanath B., Mislove A., Cha M., and Gummadi K. P. On the evolution of user interaction in facebook. *in WOSN'09 (ACM, New York, NY, USA, 2009)*, pp. 37-42.

[40] Leskovec J., Lang K. J., Dasgupta A., and Mahoney M. W. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathmatics*, **6** (2009) 29.

[41] Cho, E., Myers, S. A., and Leskovec, J. Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. (ACM, San Diego, California, USA, 2011) pp. 1082-1090.

[42] Zafarani, R., and Liu, H. *Social computing data repository at ASU, in School of Computing, Informatics and Decision Systems Engineering* (Arizona State University, 2009), http://socialcomputing.asu.edu.

[43] Stumpf M. P. H., Wiuf C., and May R. M. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci. USA*, **102** (2005) 4221.

[44] Beuming T., Skrabanek L., Niv M. Y., Mukherjee P., and Weinstein H. PDZBase: A protein-protein interaction database for PDZ-domains. *Bioinformatics*, **21(6)** (2005) 827.

[45] Lusseau D., Schneider K., Boisseau O. J., Haase, P. Slooten E., and Dawson S. M., *Behav. Ecol. Sociobiol.*, **54** (2003) 396.

[46] Sundaresan S. R., Fischhoff I. R., Dushoff J., and Rubenstein D. I. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Oecologia*, **151** (2007) 140.

[47] Grant T. R. Dominance and association among members of a captive and

a free-ranging group of grey kangaroos (Macropus giganteus). *Anim. Behav.*, **21** (1973) 449.

[48] Cannistraci C. V., Alanis-Lobato G., and Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci. Rep.*, **3** (2013) 1613.

[49] Zhang Q.-M., Xu X.-K., Zhu Y.-X., and Zhou T. Measuring multiple evolution mechanisms of complex networks. *Sci. Rep.*, **5** (2015) 10350.

[50] Zhang J. Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction. *Inform. Process. Manag.*, **53** (2017) 42.