Measuring Quadrangle Formation in Complex Networks

Mingshan Jia, Bogdan Gabrys, *Senior Member, IEEE,* and Katarzyna Musial

Abstract—The classic clustering coefficient and the lately proposed closure coefficient quantify the formation of triangles from two different perspectives, with the focal node at the centre or at the end in an open triad respectively. As many networks are naturally rich in triangles, they become standard metrics to describe and analyse networks. However, the advantages of applying them can be limited in networks, where there are relatively few triangles but which are rich in quadrangles, such as the protein-protein interaction networks, the neural networks and the food webs. This yields for other approaches that would leverage quadrangles in our journey to better understand local structures and their meaning in different types of networks. Here we propose two quadrangle coefficients, i.e., the i-quad coefficient and the o-quad coefficient, to quantify quadrangle formation in networks, and we further extend them to weighted networks. Through experiments on 16 networks from six different domains, we first reveal the density distribution of the two quadrangle coefficients, and then analyse their correlations with node degree. Finally, we demonstrate that at network-level, adding the average i-quad coefficient and the average o-quad coefficient leads to significant improvement in network classification, while at node-level, the i-quad and o-quad coefficients are useful features to improve link prediction.

Index Terms—clustering coefficient, closure coefficient, quadrangle coefficient, network classification, link prediction.

1 Introduction

Complex systems across various domains, such as biology, ecology, physics and social science, can be modelled as networks that abstract the interactions between system's components [1], [2], [3]. Different from a simple grid graph or a line graph for image or text modelling respectively, the complexity of networks comes from their intricate topological structures. Therefore, the study of network structure, especially local structure, underlies a number of representative and analytical applications such as representation learning of graphs [4], [5], node-type classification [6], [7], link prediction [8], [9] and anomaly detection [10], [11].

One fundamental and classic statistical metric to assess the local structure of complex networks is the *local clustering coefficient* [12], [13]. It is defined as the percentage of the number of triangles formed with a focal node to the number of triangles that the focal node could form with all its neighbours. Note that the focal node here serves as the centre node in an open triad (the middle of a length-2 path). Since many of the real-world networks are triangle-rich, the clustering coefficient — a measure of triangle formation — has become a standard metric to describe networks. It has also been used in numerous applications such as malware detection [14], language learning [15] and structural role discovery [16].

A recent study has proposed another interesting measure of triangle formation, i.e., the *local closure coefficient* [17]. With the focal node as the end node of an open triad (the head of a length-2 path), it is quantified as the percentage of twice the number of triangles containing the focal node to the number of all length-2 paths starting from the focal node. Specifically, the classic local clustering coefficient measures the extent to which the 1-hop neighbours of a given node connect to each other, while the local closure coefficient measures the extent to which the 2-hop neighbours of a given node connect to the given node itself. This new metric has been proven to be a useful feature in network analysis tasks such as community detection and link prediction [17].

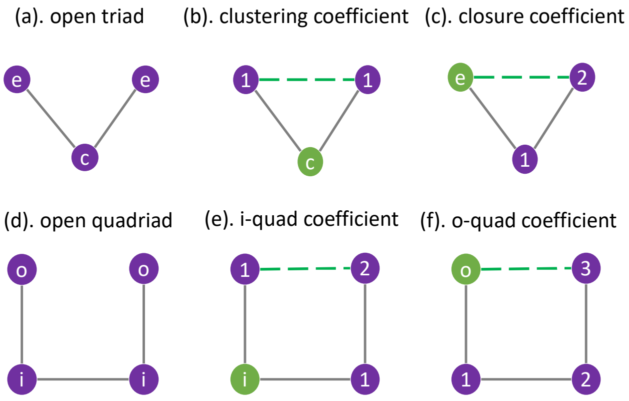


Fig. 1: The i-quad coefficient and the o-quad coefficient in comparison with the clustering coefficient and the closure coefficient. Letters c, e, i and o denote centre node, end node, inner node and outer node respectively. Node in green colour is the focal node in each subfigure. Number on node indicates the node's distance from the focal node in the open triad or the open quadriad, which might be closed by an edge in dotted green line style.

In many types of networks, however, quadrangles appear at a much higher frequency than triangles, and thus become the most dominant motifs [18]. For instance, in gene

- *M. Jia, B. Gabrys and K. Musial are with the School of Computer Science, University of Technology Sydney, Ultimo NSW 2007, Australia. E-mail: mingshan.jia@student.uts.edu.au, {bogdan.gabrys, katarzyna.musial-gabrys}@uts.edu.au*

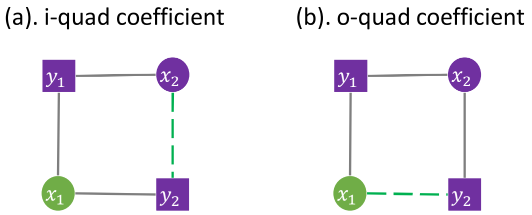


Fig. 2: An example of the i-quad coefficient and the o-quad coefficient in a movie recommender network. Circle nodes represent users, and square nodes represent movies. Node x_1, marked in green, is the focal node. Four nodes and three solid links form an open quadriad, which if is closed by a dotted link will form a quadrangle.

regulatory networks, logical circuits networks and neuron networks, the over-represented ”bi-fan” structure (a specific directed quadrangle) serves to carry information or signals from previous units to following ones; while in food webs, the highly recurring ”bi-parallel” structure (another type of directed quadrangle) describes how energy flows in an ecosystem.

In order to better describe and analyse the local structure of networks, we propose two metrics quantifying the formation of quadrangles, i.e., the *i-quad coefficient* and the *o-quad coefficient*. There are two definitions in that two categories of nodes — the inner node or the outer node — can be distinguished from the node's position in an open quadriad (also called intransitive quadriad in some works [19]). The i-quad coefficient, with the focal node functioning as the inner node of an open quadriad, measures the extent to which the focal node's 2-hop neighbours connect to its 1-hop neighbours. The o-quad coefficient, having the focal node as the outer node of an open quadriad, measures the extent to which the focal node's 3-hop neighbours connect to itself (Figure 1).

Although the focus in this paper lies on the general unipartite networks, the proposed i-quad and o-quad coefficients provide interesting insights into bipartite networks as well. Suppose that in a recommender network where node type x denotes users and node type y denotes movies, an edge between x_i and y_i represents user x_i likes movie y_i. Take the i-quad coefficient for instance (Figure 2a), given x_1, the focal user, likes movies y_1 and y_2, while x_2 likes y_1, it measures whether x_2 likes y_2. In other words, the i-quad coefficient gives the extent to which other users have a similar preference as the focal user. Likewise, for the o-quad coefficient, given x_2 likes y_1 and y_2, while x_1, the focal node, likes y_1, it measures whether x_1 likes y_2 (Figure 2b). That is to say, the o-quad coefficient gives the extent to which the focal user shares a similar opinion with other users. Interestingly, this explanation coincides with the idea of collaborative filtering [20], [21].

In addition to the basic network structure, a deeper understanding of complex systems sometimes requires taking into account the intensity or the strength of interactions between components. This is achieved by assigning weights to links. For instance, in unipartite networks, weighted links are used to represent the frequency of contact in a communication network, or the intensity of the traffic flow in a transportation network; in bipartite networks, especially recommender networks, weights are added to indicate how much a person likes a product or how often he or she purchases it. Accordingly, we introduce the *weighted i-quad coefficient* and the *weighted o-quad coefficient* in order to unveil the quadrangle formation in real weighted networks.

Our empirical study on 16 real-world networks from six domains has revealed several basic and interesting properties of the two proposed coefficients. First, we find that in most types of networks, the average o-quad coefficient is smaller than the average i-quad coefficient, which is also demonstrated through their cumulative density distributions. Secondly, we show that the o-quad coefficient has a strong positive correlation with node degree, whereas the correlation between the i-quad coefficient and node degree is very weak. We then provide a theoretical justification of this phenomenon under the configuration model.

Last but not least, we illustrate how the proposed quadrangle coefficients can be powerful features for network analysis and inference tasks. In a network classification task, we show that different types of real-world networks are significantly better clustered by adding the two quadrangle coefficients. Furthermore, in a link prediction task, we also show that the i-quad and o-quad coefficients can be used as effective predictors to improve the performance, especially in food webs, protein-protein interaction networks and infrastructure networks.

To sum up, in order to measure the formation of quadrangles in networks, we propose the i-quad coefficient and the o-quad coefficient, based on the inner node and the outer node of an open quadriad respectively. We further extend them to weighted networks. Through extensive experiments on real-world networks, we show not only the intrinsic properties of the two coefficients, but also investigate how they can be utilised in common network analysis task and machine learning tasks. The remainder of this paper is organised as follows. Section 2 introduces notations and background knowledge of clustering coefficient and closure coefficient. Section 3 presents and exemplifies the proposed quadrangle coefficients, whereas Section 4 provides details of the evaluation, including the datasets, experiment setups, performance measures, experiment results and our findings. Section 5 briefly contemplates the related works, and finally we conclude this paper in Section 6.

2 Background and Motivating Example

This section first introduces the basic concepts such as the classic clustering coefficient and the recently proposed closure coefficient. We then illustrate how these coefficients are calculated in the case of a small-scale network that serves as an example.

2.1 Clustering Coefficient

The clustering coefficient, or more specifically the local clustering coefficient, was originally proposed in order to measure the cliquishness of a neighbourhood in networks [12]. It has since become one of the most commonly used metrics for network structure, together with such measures as degree distribution, path length, connected components, etc. Let $G = (V, E)$ be an undirected graph on a node set V (the number of nodes is $|V|$) and an edge set E (the number of edges is m), without self-loops and multiple edges. We denote the set of neighbours of node i as $N(i)$, and thus the degree of node i, denoted as d_i, equals to $|N(i)|$. An open triad is a directionless length-2 path. For example, in an open triad ijk, where an edge connects node i and j, and another edge connects node j and k, we do not distinguish between path $i \rightarrow j \rightarrow k$ and path $k \rightarrow j \rightarrow i$.

For any node $i \in V$, its *local clustering coefficient*, denoted $C(i)$, is defined as the number of triangles containing node i (denoted $T(i)$), divided by the number of open triads with i as the centre node (denoted $OTC(i)$):

$$C(i) = \frac{T(i)}{OTC(i)} = \frac{\frac{1}{2}\sum_{j\in N(i)} |N(i) \cap N(j)|}{\frac{1}{2}d_i\,(d_i - 1)}. \tag{1}$$

In other words, it is the fraction of open triads, where the focal node serves as the centre node, that actually form triangles. By definition, $C(i) \in [0, 1]$.

In order to get a network-level measurement, the *average clustering coefficient* is introduced by averaging the local clustering coefficient over all nodes (an undefined local clustering coefficient is treated as zero):

$$\overline{C} = \frac{1}{|V|}\sum_{i\in V} C(i). \tag{2}$$

An alternative way to measure clustering at the network-level is the *global clustering coefficient* [22], which is defined as the fraction of open triads that form triangles in the entire network:

$$C = \frac{\sum_{i\in V}\sum_{j\in N(i)} |N(i) \cap N(j)|}{\sum_{i\in V} d_i\,(d_i - 1)}. \tag{3}$$

Note that the global clustering coefficient is not equivalent to the average clustering coefficient. In Equation 3, we calculate the number of triangles in the entire network, then divided by the number of open triads across the network. Since a node with high degree forms more open triads and also tends to form more triangles, the global clustering coefficient thus puts more weight on hub nodes. On the contrary, in Equation 2, we first calculate the sum of local clustering coefficient of each node, then average over the number of nodes, which gives equal weight on each node.

2.2 Closure Coefficient

Different from the ordinary centre node based perspective in the clustering coefficient, another interesting measure of triangle formation, i.e., the local closure coefficient, has recently been proposed [17]. The focal node in the closure coefficient serves as the end node of an open triad. As Yin et al. [17] has revealed, this subtle difference in measurement leads to very different properties from those of the clustering coefficient.

Adopting the notations of Section 2.1, the local closure coefficient of node i, denoted $E(i)$, is defined as twice the number of triangles formed with i, divided by the number of open triads with i as the end node. (denoted $OTE(i)$):

$$E(i) = \frac{2T(i)}{OTE(i)} = \frac{\sum_{j\in N(i)} |N(i) \cap N(j)|}{\sum_{j\in N(i)}(d_j - 1)}. \tag{4}$$

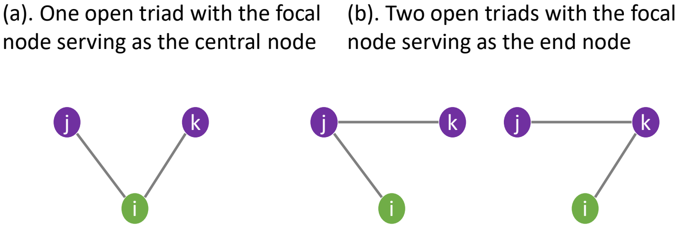

(a). One open triad with the focal node serving as the central node
(b). Two open triads with the focal node serving as the end node
j
k
i
j
k
i
j
k
i

Fig. 3: Two types of open triads in triangle formation. Among three nodes i, j and k, node i, painted in green, is the focal node.

In other words, it is the fraction of open triads, where the focal node serves as the end node, that actually form triangles. $T(i)$ is multiplied by two for the reason that each triangle contains two open triads with i as the end node. When a triangle is actually formed, the focal node can be viewed as the centre node in one open triad or as the end node in two open triads (Figure 3). Obviously, $E(i) \in [0, 1]$.

At the network-level, the *average closure coefficient* is then defined as the mean of the local closure coefficient over all nodes (an undefined local closure coefficient is treated as zero):

$$\overline{E} = \frac{1}{|V|}\sum_{i\in V} E(i). \tag{5}$$

Analogous to the global clustering coefficient (Equation 3), the *global closure coefficient*, denoted E, is defined as:

$$E = \frac{\sum_{i\in V}\sum_{j\in N(i)} |N(i) \cap N(j)|}{\sum_{i\in V}\sum_{j\in N(i)}(d_j - 1)}. \tag{6}$$

The global closure coefficient (Equation 6) is actually equivalent to the global clustering coefficient (Equation 3), as globally the difference of the position of the focal node will not surface.

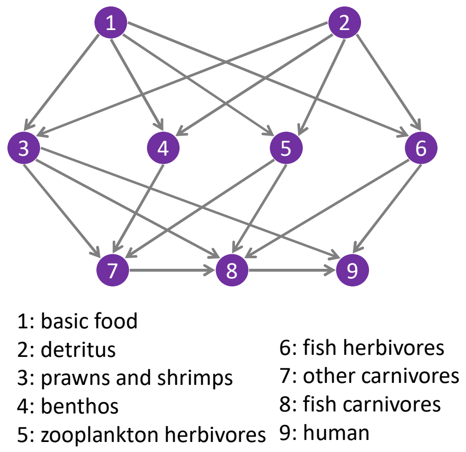
2.3 A motivating example

We illustrate how the two coefficients of triangle formation are calculated via a small yet real network. Figure 4a shows a simplified food web of the backwaters of Kerala, India [23]. It is composed of 9 nodes and 18 edges. Each node represents a species and each edge represents the flow of food energy from one species to another.

Figure 4b gives a detailed table of the number of triangles $T(i)$, the number of centre-node-based open triads $OTC(i)$, the number of end-node-based open triads $OTE(i)$, the local clustering coefficient $C(i)$ and the local closure coefficient $E(i)$ for each node. Also, the last row gives the average clustering coefficient, the average closure coefficient and the global clustering/closure coefficient, all of which are around 0.20.

Different from some triangle-rich networks, we find many more quadrangles than triangles in this food web (23 versus 4), which motivates us to propose measuring quadrangle formation instead. In the next section, new measures to quantify information about quadrangles in complex networks are proposed, and we show how we can leverage the fact that some networks are quadrangle and not triangle rich.

(a). A small food web



(b). Calculation of clustering/closure coef.

	T(i)	OTC(i)	OTE(i)	C(i)	E(i)
①	0	6	12	0	0
②	0	6	12	0	0
③	2	10	15	0.2	0.27
④	0	3	9	0	0
⑤	1	6	13	0.17	0.15
⑥	1	6	12	0.17	0.17
⑦	2	6	13	0.33	0.31
⑧	4	10	15	0.40	0.53
⑨	2	3	11	0.67	0.36
$\bar{C}$ = 0.21		$\bar{E}$ = 0.20		C = E = 0.21	

(c). Calculation of i-quad/o-quad coef.

	Q(i)	OQI(i)	OQO(i)	I(i)	O(i)
①	13	36	37	0.72	0.70
②	13	36	37	0.72	0.70
③	16	56	39	0.57	0.82
④	7	18	31	0.78	0.45
⑤	12	37	38	0.65	0.63
⑥	10	34	36	0.59	0.56
⑦	8	35	36	0.46	0.44
⑧	9	52	36	0.35	0.50
⑨	4	18	32	0.44	0.25
$\bar{I}$ = 0.59		$\bar{O}$ = 0.56		I = O = 0.57	

Fig. 4: A motivating example.

3 Two Quadrangle Coefficients

The clustering coefficient and the closure coefficient provide us two ways of measuring triangle formation. In some networks however, we care more about the formation of quadrangles. Also, triangles do not exist in bipartite networks and the most basic enclosed structure in this representation of networks is quadrangle. In this section, we first propose two coefficients measuring quadrangle formation, based on two different positions of the focal node in an open quadriad. Then, we further extend them to weighted networks.

3.1 I-quad coefficient

Recall that an open quadriad is a directionless length-3 path (Figure 1d). In an open quadriad $ijkl$, for instance, where three edges exist between node pairs (i, j), (j, k) and (k, l), we name nodes j and k as inner nodes. In contrast, nodes i and l are outer nodes. Obviously, an inner node has a degree of two, and an outer node has a degree of one. Further, an open quadriad with the focal node acting as the inner node is called inner-node-based open quadriad of that node; an open quadriad with the focal node acting as the outer node is named outer-node-based open quadriad of that node.

Conforming with the definition of the classic clustering coefficient which measures whether the two endpoints of a centre-node-based open triad are connected by a closing edge, we propose the i-quad coefficient that measures whether the two endpoints of an inner-node-based open quadriad are connected by a closing edge. It is quantified as the fraction of inner-node-based open quadriads that actually form quadrangles. Concretely, the ***i-quad coefficient*** of node i, denoted $I(i)$, is defined as twice the number of quadrangles formed with i (denoted as $Q(i)$), divided by the number of open quadriads with i as the inner node (denoted as $OQI(i)$):

$$\begin{aligned} I(i) &= \frac{2Q(i)}{OQI(i)} \\ &= \frac{\sum_{j\in N(i)} \sum_{k\in (N(j)-i)} |N(k)\cap N(i) - j|}{\sum_{j\in N(i)} \sum_{k\in (N(j)-i)} |N(i) - j - k|}. \end{aligned} \tag{7}$$

In the above equation, j is in i's neighbour set, and k is in j's neighbour set excluding i. $Q(i)$ is multiplied by two because each quadrangle can be viewed as constructed from two open quadriads with i as the inner node. By definition, it is obvious that $I(i) \in [0, 1]$.

Then, we define the ***average i-quad coefficient*** at the network-level, as the mean of the i-quad coefficient over all nodes (undefined ones are treated as zeros):

$$\bar{I} = \frac{1}{|V|} \sum_{i\in V} I(i). \tag{8}$$

In the case of a random network where each pair of nodes is connected with a probability p, the expected value of the average i-quad coefficient is also p, i.e., $\mathbb{E}[\bar{I}] = p$.

An alternative way of measuring quadrangle formation at the network-level is the ***global i-quad coefficient***, which is defined as the fraction of inner-node-based open quadriads that form quadrangles in the entire network:

$$I = \frac{\sum_{i\in V} \sum_{j\in N(i)} \sum_{k\in (N(j)-i)} |N(k)\cap N(i) - j|}{\sum_{i\in V} \sum_{j\in N(i)} \sum_{k\in (N(j)-i)} |N(i) - j - k|}. \tag{9}$$

The numerator of the above equation can be viewed as eight times the number of quadrangles in the entire network (each node of a quadrangle contributes two counts), then divided by twice the number of open quadriads with each node acting as the inner node.

Although both the average i-quad coefficient and the global i-quad coefficient can be used as metrics to describe quadrangle formation in the entire network, they are calculated differently. The average i-quad coefficient adds up the i-quad coefficient of every node then divides it by the number of nodes, giving each node equal weight. In contrast, the global i-quad coefficient gives nodes that form numerous quadrangles more weight, by first totalling the numerator of the i-quad coefficient then dividing it by the sum of the denominator of the i-quad coefficient.

3.2 O-quad coefficient

Inspired by the closure coefficient in measuring triangle formation, we move the focal node from the inner node

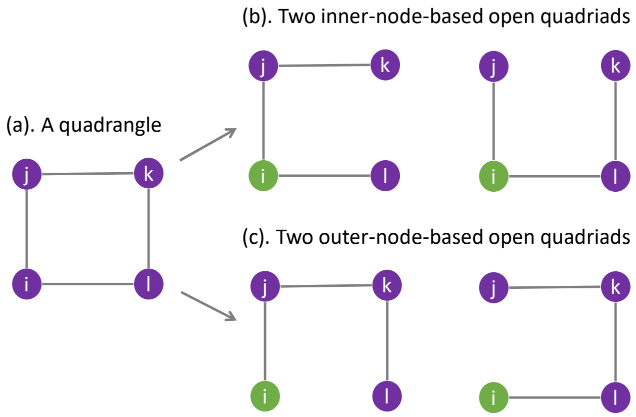

(a). A quadrangle
(b). Two inner-node-based open quadriads
(c). Two outer-node-based open quadriads
j
k
i
l

Fig. 5: Two types of open quadriads in a quadrangle. Node i, depicted in green, is the focal node, among four nodes i, j, k and l.

to the outer node of an open quadriad, thus proposing the o-quad coefficient in order to measure the formation of quadrangle from a different perspective.

The significance of introducing the o-quad coefficient is twofold. First, the o-quad coefficient takes into account length-3 paths emanating from the focal node, and therefore has a larger scope of the network structure. Second, when a quadrangle is formed, the closing edge (the edge that closes the outer-node-based open quadriad) is incident to the focal node. This leads to some special properties, comparing to the i-quad coefficient where the closing edge is not incident to the focal node. We show in Section 4 that the cumulative distribution curve of the o-quad coefficient is above that of the i-quad coefficient, and that the o-quad coefficient tends to increase with node degree.

In a similar way, the ***o-quad coefficient*** of node i, denoted as $O(i)$, is defined as the fraction of open quadriads with i as the outer node that are closed:

$$\begin{aligned} O(i) &= \frac{2Q(i)}{OQO(i)} \\ &= \frac{\sum_{j\in N(i)}\sum_{k\in(N(j)-i)}|N(k)\cap N(i)-j|}{\sum_{j\in N(i)}\sum_{k\in(N(j)-i)}|N(k)-j-i|}, \end{aligned} \tag{10}$$

where $OQO(i)$ is the number of outer-node-based open quadriads of node i, and $Q(i)$ is the number of quadrangles containing i. $Q(i)$ is multiplied by two because each quadrangle contains two open quadriads with i as the outer node. In a quadrangle, the focal node can serve as the inner node in two open quadriads or as the outer node in another two open quadriads (Figure 5). Obviously, $O(i)\in[0,1]$.

In order to measure at the network level, the ***average o-quad coefficient*** is defined by averaging the o-quad coefficient over all nodes (an undefined o-quad coefficient is treated as zero):

$$\overline{O}=\frac{1}{|V|}\sum_{i\in V}O(i). \tag{11}$$

Analogous to the global i-quad coefficient, the ***global o-quad coefficient*** can be defined as the fraction of outer-node-based open quadriads that form quadrangles in the entire network:

$$O=\frac{\sum_{i\in V}\sum_{j\in N(i)}\sum_{k\in(N(j)-i)}|N(k)\cap N(i)-j|}{\sum_{i\in V}\sum_{j\in N(i)}\sum_{k\in(N(j)-i)}|N(k)-j-i|}. \tag{12}$$

As the equivalence between the global clustering coefficient and the global closure coefficient, this definition of global o-quad coefficient is actually not different from the global i-quad coefficient (Equation 9) since globally the difference of the position of the focal node will not arise.

Revisiting the motivating example, Figure 4c gives a detailed table of the number of quadrangles $Q(i)$, the number of inner-node-based open quadriads $OQI(i)$ and the number of outer-node-based open quadriads $OQO(i)$ of each node, based on which the i-quad coefficient $I(i)$ and the o-quad coefficient $O(i)$ are calculated. Also, the last row of this table gives the three network-level measures, i.e., the average i-quad coefficient, the average o-quad coefficient and the global i-quad/o-quad coefficient, which are more than 2.5 times larger than those metrics measuring triangles formation.

3.3 Quadrangle coefficients in weighted networks

Until now, the discussion has been focused on binary networks, where the value of each link is either one or zero. In many networks, however, we need a more accurate representation of the relationships between nodes, such as the frequency of contact in a communication network, or the rating of a product given by a consumer in a recommender network, etc. This kind of information is usually expressed as a strength of the relationship and we use weighted networks to represent it. Therefore, we are interested in extending the two quadrangle coefficients to networks that allow for weights of the relationships.

Several versions of weighted clustering coefficient have been proposed in order to measure triangle formation in weighted networks [24], [25], [26], [27]. For example, Onnela et al. [25] proposed to sum over the geometric averages of the three weights in formed triangles, divided by the number of potential triangles. Alternatively, Zhang and Horvath. [26] chose to sum simply over the products of the three weights in formed triangles, divided by the total of products of the two weights of all open triads, implying the triadic closing edges taking the maximum weight.

Adopting a strategy similar to the one proposed by Zhang and Horvath [26], we introduce the weighted i-quad coefficient and the weighted o-quad coefficient to measure quadrangles formation in weighted networks. Let $G^{\mathcal{W}}=(V,E)$ be a weighted graph without self-loops and multiple edges. The weight of a link between any node i and j is denoted w_{ij} ($w_{ij}\in[0,1]$ after normalisation by the maximum weight). For any node $i\in V$, the ***weighted i-quad coefficient***, denoted as $I^{\mathcal{W}}(i)$, and the ***weighted o-quad coefficient***, denoted as $O^{\mathcal{W}}(i)$, are defined as:

$$I^{\mathcal{W}}(i)=\frac{\sum\limits_{j\in N(i)}\sum\limits_{k\in(N(j)-i)}\sum\limits_{l\in(N(i)\cap N(k)-j)}w_{ij}w_{jk}w_{il}w_{lk}}{\sum\limits_{j\in N(i)}\sum\limits_{k\in(N(j)-i)}\sum\limits_{l\in(N(i)-j-k)}w_{ij}w_{jk}w_{il}}, \tag{13}$$

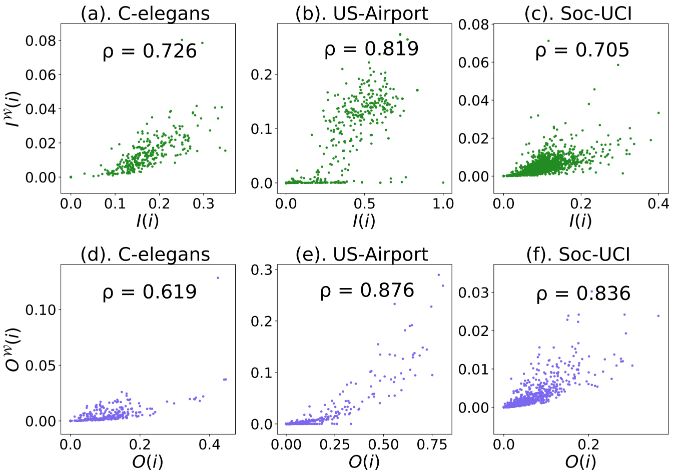


Fig. 6: Correlation of quadrangle coefficients and weighted quadrangle coefficients in three different networks. First row is the correlation of i-quad coefficient $I(i)$ and weighted i-quad coefficient $I^{\mathcal{W}}(i)$, second row is the correlation of o-quad coefficient $O(i)$ and weighted o-quad coefficient $O^{\mathcal{W}}(i)$. The weighted networks are: (1) C-elegans, the neural network of the Caenorhabditis elegans worm [12]; (2) US-Airport, the network of the 500 busiest commercial airports in the United States [28]; (3) Soc-UCI, the social network of online community for students at University of California, Irvine [29].

$$O^{\mathcal{W}}(i) = \frac{\sum\limits_{j \in N(i)} \sum\limits_{k \in (N(j)-i)} \sum\limits_{l \in (N(i) \cap N(k)-j)} w_{ij} w_{jk} w_{il} w_{lk}}{\sum\limits_{j \in N(i)} \sum\limits_{k \in (N(j)-i)} \sum\limits_{l \in (N(k)-j-i)} w_{ij} w_{jk} w_{kl}}. \tag{14}$$

When the graph becomes binary (unweighted), i.e., $w_{ij} = 1$, the above two weighted quadrangle coefficients degrade to their unweighted versions (Equation 7 and Equation 10). The average weighted i-quad coefficient and the average weighted o-quad coefficient are then defined respectively as: $\overline{I^{\mathcal{W}}} = \frac{1}{|V|} \sum_{i \in V} I^{\mathcal{W}}(i)$, $\overline{O^{\mathcal{W}}} = \frac{1}{|V|} \sum_{i \in V} O^{\mathcal{W}}(i)$.

We can see from Figure 6 that in different weighted networks, the correlation of i-quad coefficient and weighted i-quad coefficient (and the correlation of o-quad coefficient and weighted o-quad coefficient) is also different. In other words, when weights are considered in calculating quadrangle coefficients, the weighted i-quad coefficient and the weighted o-quad coefficient capture different information compared to their unweighted counterparts.

3.4 Computational cost

At the end of this section, we give a brief discussion about the computational efficiency of the above mentioned metrics. From Equation 7 and Equation 10, we can see that to compute the i-quad coefficient or the o-quad coefficient for a single node, the worst-case cost is $O((k_{max})^3)$, where k_{max} is the maximum degree of the network. Therefore, the worst-case cost for computing the two coefficients for every node in a network is $O(|V| \cdot (k_{max})^3)$, which is not cheap. Fortunately, however, since most real-world networks are scale-free and exhibit heavy-tailed degree distribution, the actual cost is far less expensive than this. For example, it takes about 22.5 seconds to compute the average i-quad coefficient on the CORA citation network which contains 23,166 nodes and 89,157 edges (test on Intel Xeon Gold 6238R @ 2.2GHz with 180GB of RAM).

4 Experiments and Analysis

In this section, we analyse the proposed quadrangle coefficients on different types of real-world networks and demonstrate their usage in some common applications[1].

4.1 Quadrangle coefficients in real-world networks

Datasets. We run experiments on 16 networks of six categories (collected from Konect [30] and Snap [31]):

1) Food webs. FW-FloridaDry [32] and FW-LittleRock [33]: energy transfer relationships collected from the cypress wetlands of South Florida and the Little Rock Lake of Wisconsin. Nodes represent species and an edge denotes that one species feeds on another (edge direction and weight are ignored).
2) Social networks. EmailEu [34]: a temporal email network from a European research institution (a temporal edge denotes that an email is exchanged between two persons at a certain time); ClgMsg [35]: temporal online message interactions between UCIrvine college students (a temporal edge means that a message is exchanged between two students at a certain time); BTCAlpha [36]: a temporal who-trusts-whom network of users on a Bitcoin trading platform Bitcoin Alpha (edge direction and weight are ignored); TwitchFr [37]: a network of gamers who stream in French, where nodes are the users and edges are mutual friendships between them.
3) Protein-protein interaction (PPI) networks. Stelzl [38], Figeys [39], Vidal [40] and IntAct [41]: four networks of interactions between proteins in Homo sapiens. Nodes represent proteins and an edge denotes the physical contact between two proteins in the cell.
4) Citation networks. DBLP [42] and Cora [43]: two academic publication citation networks. DBLP contains temporal information on edges. Nodes represent papers, and an edge means that one paper cites another paper (direction is ignored).
5) Infrastructure networks. Rd-NewYork and Rd-BayArea [30]: two road networks for New York City and San Francisco Bay Area. Nodes represent intersections and endpoints, and the roads connecting them are represented by edges.
6) Q&A networks. MathOvfl. and AskUbuntu [34]: two temporal Q&A networks derived from Stack Exchange. Nodes represent users, and a temporal edge means that one user answers another user's question at a certain time (edge direction is ignored).

Observations. Table 1 lists some key statistics including the proposed coefficients of these networks. We observe that in

[1] Our code is available at https://github.com/MingshanJia/explore-local-structure.

TABLE 1: Statistics of datasets, showing the number of nodes ($|V|$), the number of edges ($|E|$), the average degree ($\langle k\rangle$), the average clustering coefficient ($\overline{C}$), the average closure coefficient ($\overline{E}$), the average i-quad coefficient ($\overline{I}$) and the average o-quad coefficient ($\overline{O}$). In order to facilitate comparison, the last four columns give the quotient of $\overline{C}$ and $\overline{E}$, the quotient of $\overline{I}$ and $\overline{O}$, the quotient of $\overline{I}$ and $\overline{C}$, and the quotient of $\overline{O}$ and $\overline{E}$ respectively. Datasets having timestamps on edge creation are superscripted by (τ).

Network	$\|V\|$	$\|E\|$	$\langle k\rangle$	$\overline{C}$	$\overline{E}$	$\overline{I}$	$\overline{O}$	$\overline{C}/\overline{E}$	$\overline{I}/\overline{O}$	$\overline{I}/\overline{C}$	$\overline{O}/\overline{E}$
FW-FloridaDry	128	2,106	32.91	0.335	0.261	0.428	0.353	1.280	1.213	1.280	1.351
FW-LittleRock	183	2,452	26.80	0.323	0.208	0.550	0.339	1.553	1.622	1.704	1.631
Soc-EmailEu$^\tau$	986	16,064	32.58	0.407	0.153	0.231	0.102	2.659	2.267	0.568	0.667
Soc-ClgMsg$^\tau$	1,899	13,838	14.57	0.109	0.022	0.081	0.029	5.082	2.806	0.744	1.347
Soc-BTCAlpha$^\tau$	3,783	14,124	7.47	0.177	0.020	0.058	0.013	8.937	4.448	0.326	0.655
Soc-TwitchFr	6,549	113K	34.41	0.222	0.029	0.109	0.034	7.557	3.202	0.493	1.163
PPI-Stelzl	1,706	3,191	3.74	0.006	0.002	0.038	0.021	3.827	1.806	6.332	13.416
PPI-Figeys	2,239	6,432	5.75	0.040	0.005	0.082	0.043	7.321	1.908	2.064	7.918
PPI-Vidal	3,133	6,726	4.29	0.064	0.025	0.040	0.018	2.531	2.291	0.632	0.698
PPI-IntAct	8,077	26,085	6.46	0.083	0.016	0.063	0.021	5.101	2.993	0.750	1.278
Cit-DBLP$^\tau$	12,590	49,651	7.89	0.117	0.026	0.060	0.014	4.529	4.175	0.510	0.553
Cit-Cora	23,166	89,157	7.70	0.266	0.100	0.107	0.047	2.667	2.285	0.402	0.469
Rd-NewYork	264K	365K	2.76	0.021	0.021	0.068	0.069	1.012	0.990	3.291	3.365
Rd-BayArea	321K	397K	2.47	0.017	0.016	0.038	0.038	1.020	0.992	2.284	2.350
QA-MathOvfl.$^\tau$	21,688	88,956	8.20	0.094	0.005	0.031	0.004	17.956	7.305	0.333	0.817
QA-AskUbuntu$^\tau$	138K	262K	3.81	0.015	5e-4	0.004	5e-4	31.708	7.867	0.243	0.981

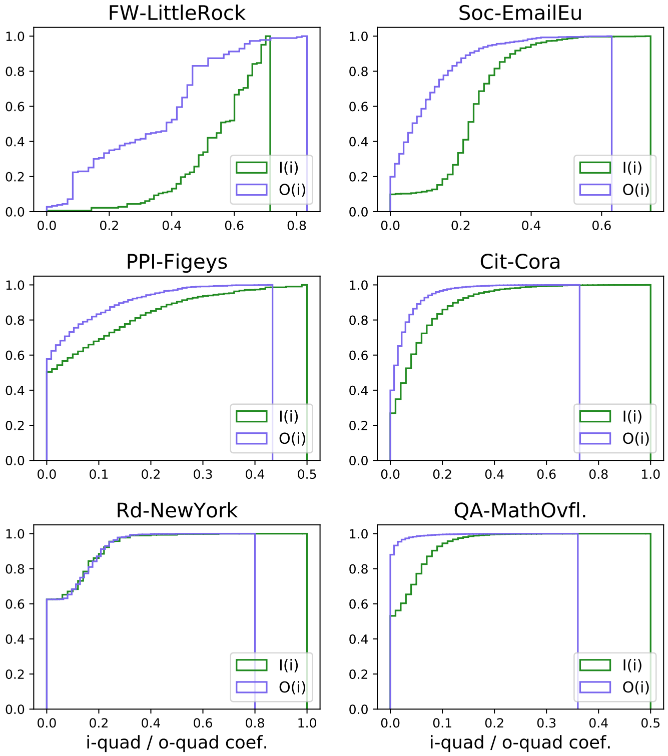


Fig. 7: Cumulative distribution curve of the i-quad coefficient $I(i)$ (in green colour) and the o-quad coefficient $O(i)$ (in purple colour) in six real-world networks of different types.

most types of networks (except road networks), the average o-quad coefficient is smaller than the average i-quad coefficient. That is to say, for the majority of nodes in these types of networks, fewer quadrangles are built from the outer-node-based open quadriads, compared to the number of quadrangles constructed from the inner-node-based open quadriads. This phenomenon is better revealed through the cumulative distribution function (CDF) in Figure 7: the CDF curve of the o-quad coefficient is above that of the i-quad coefficient when the coefficient value is small (except in Rd-NewYork).

We can also observe that in all food webs, two PPI networks (PPI-Stelzl and PPI-Figeys) and all road networks, the average i-quad coefficient is larger than the average clustering coefficient ($\overline{I} > \overline{C}$); and the average o-quad coefficient is larger than the average closure coefficient ($\overline{O} > \overline{E}$). In other words, these networks are more inclined to form quadrangles than to form triangles, which leads us to the following experiments.

4.2 Correlation with node degree

Since node degree is one of the most important and widely used concepts in network science, we study how the two quadrangle coefficients vary with it. We start by conducting an empirical analysis in real networks, followed by a theoretical justification under the degree-preserving random graph model.

We choose one network in each category and plot the correlation of quadrangle coefficients and degree (Figure 8). We observe a strong positive correlation between the o-quad coefficient and the node degree: the average o-quad coefficient is small among nodes with small degree and becomes larger as the average node degree increases. In contrast, the correlation between the i-quad coefficient and the degree is weak: the average i-quad coefficient is large (compared to the average o-quad coefficient) when the average node degree is small and does not change too much as the average degree increases. Since most real-world networks are scale-free and exhibit heavy-tailed degree distribution, it also

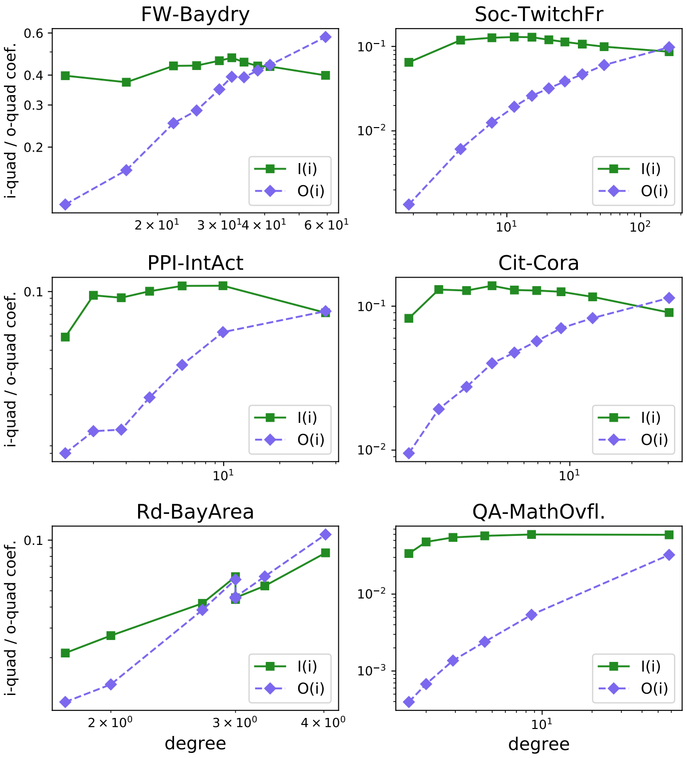


Fig. 8: Correlation of two quadrangle coefficients with node degree in six real-world networks. Nodes are grouped into logarithmic bins in ascending order by degree, then average i-quad and o-quad coefficients are calculated in each bin.

explains why the average i-quad coefficient is bigger than the average o-quad coefficient in most networks studied in our work (Table 1).

To better understand the correlation between the quadrangle coefficients and the node degree, we give a theoretical explanation under the configuration model [44]. Constrained by a given degree sequence, the configuration model generates a network by placing edges between nodes uniformly at random. This can be achieved through a stub-matching process, in which the probability of forming an edge between node i and node j equals $d_i \cdot d_j / 2m$ (assuming $d_i^2 \leqslant 2m$ for all i). Now we give the following proposition.

Proposition 1. *Let V be a set of n nodes with specific degrees $d_1, d_2, ..., d_n$, on which graph G is generated from the configuration model. Let $m = \frac{1}{2}\sum_{i=1}^{n} d_i$ denote the number of edges and $\bar{k} = (\sum_i d_i^2)/(\sum_i d_i)$ be the expected degree when a node is chosen with probability proportional to its degree. As $n \to \infty$, for any node $i \in V$, its local i-quad coefficient satisfies:*

$$\mathbb{E}[I(i)] = \frac{(\bar{k}-1)^2}{2m},$$

and its local o-quad coefficient satisfies:

$$\mathbb{E}[O(i)] = \frac{(d_i - 1) \cdot (\bar{k}-1)}{2m}.$$

Proof. For any open quadriad with node i as an inner node, we denote one outer node by k and another outer node by l (Figure 9a). The probability that this open quadriad is closed equals the probability of having an edge between node k and l, which is $(d_k - 1)(d_l - 1)/2m$ in the configuration mode. The reason of subtracting 1 from d_k and d_l is that one stub of node k (and node l) has already been used in forming the open quadriad.

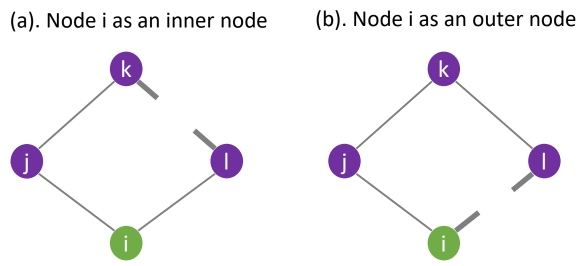


Fig. 9: Two types of quadrangle formation via stub matching. (a) Quadrangle is potentially formed with the focal node i acting as the inner node. The closing edge is between node k and l. (b) Quadrangle is potentially formed with the focal node i acting as the outer node. The closing edge is between node i and l.

Now, we show that as $n \to \infty$, $\mathbb{E}[d_k] = \mathbb{E}[d_l] = \bar{k}$. Via stub matching, any node, other than node i and j, can form an edge with node j and thus become one outer node of the open quadriad. The probability of node k being this node is proportional to its degree, which is $\frac{d_k}{\sum_{k \in V, k \neq i,j} d_k}$. Therefore, we have $\mathbb{E}[d_k] = \sum_{k \in V, k \neq i,j} d_k \cdot \frac{d_k}{\sum_{k \in V, k \neq i,j} d_k}$. When $n \to \infty$, $\mathbb{E}[d_k] = \sum_{k \in V} d_k \cdot \frac{d_k}{\sum_{k \in V} d_k} = \bar{k}$. Similarly, we have $\mathbb{E}[d_l] = \bar{k}$.

In short, we have:

$$\begin{aligned}\mathbb{E}[I(i)] &= \mathbb{E}\left[(d_k - 1)(d_l - 1)/(2m)\right] \\ &= \frac{(\mathbb{E}[d_k] - 1) \cdot (\mathbb{E}[d_l] - 1)}{2m} = \frac{(\bar{k}-1)^2}{2m}.\end{aligned}$$

Likewise, for any open quadriad with node i as an outer node, we denote the other outer node by l (Figure 9b). And we have:

$$\begin{aligned}\mathbb{E}[O(i)] &= \mathbb{E}\left[(d_i - 1)(d_l - 1)/(2m)\right] \\ &= \frac{(d_i - 1) \cdot (\mathbb{E}[d_l] - 1)}{2m} = \frac{(d_i - 1) \cdot (\bar{k}-1)}{2m}.\end{aligned}$$

□

Although Proposition 1 is given under the configuration model, we see from Figure 8 that this property is well preserved in most real-world networks. Only that in road networks, i.e., RD-NEWYORK and RD-BAYAREA, the average i-quad coefficient and the average o-quad coefficient are very similar (Table 1), and they exhibit similar correlations with node degree. This is because the variance of node degree is extremely small (less than one) in this type of network, resulting in d_i close to $\bar{k}$, and thus $\mathbb{E}[O(i)]$ close to $\mathbb{E}[I(i)]$.

4.3 Network classification

In this section, we exhibit how useful the proposed quadrangle coefficients are in classifying different types of networks. Previous works have shown that normalized number of triads and triangles (triad significance profile [45] and clustering signatures [46]) are effective attributes in a network classification task. It motivated us to use the two quadrangle

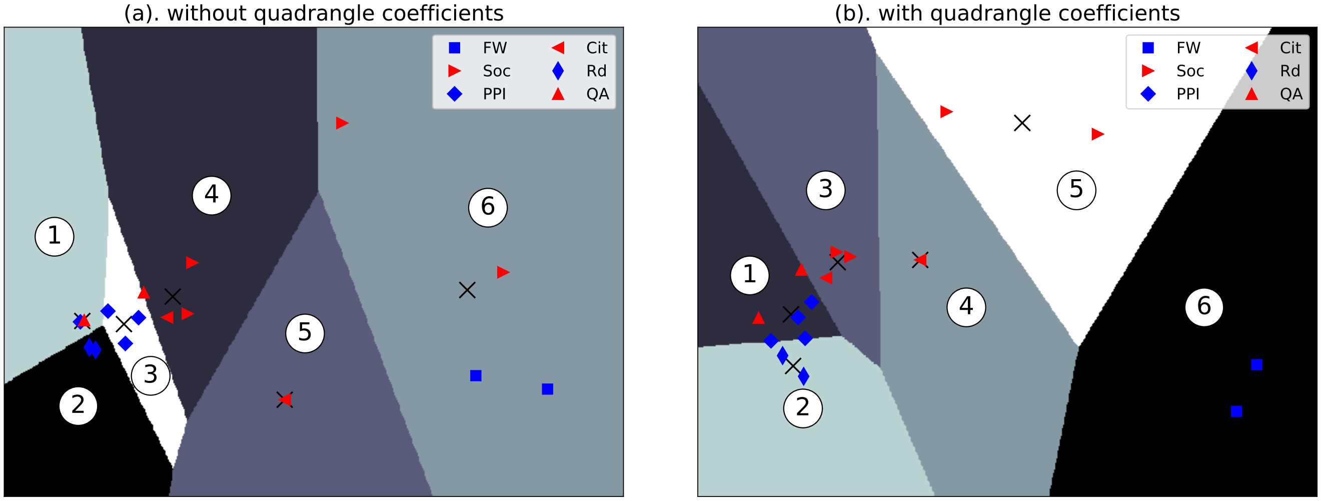


Fig. 10: Two-dimensional visualisation of K-means clustering on PCA-reduced data, without and with quadrangle coefficients (left figure and right figure respectively). Six clusters are labelled from 1 to 6, and painted in different colours. Centroids of clusters are marked as black crosses. Data points are plotted in different shapes and colours representing their ground truth categories, as shown in the legend.

coefficients in the network classification, as they represent a normalized number of quadrangles.

We can see in Table 1 that the quotient of the average i-quad coefficient and the average clustering coefficient ($\overline{I}/\overline{C}$), and the quotient of the average o-quad coefficient and the average closure coefficient ($\overline{O}/\overline{E}$) are contrasting in different types of networks. It is intuitive to expect the two quadrangle coefficients will be able to add useful discriminative information to a set of features, in addition to the average clustering coefficient and the average closure coefficient, for improving of the network classification accuracy.

Setup. We first prepare the data by using the three classic topological features of undirected networks, i.e., the average node degree $\langle k \rangle$, the average clustering coefficient $\overline{C}$ and the average closure coefficient $\overline{E}$. We then employ a K-means clustering algorithm to partition the 16 networks into 6 clusters. The initial centroids are chosen randomly, and we repeat the algorithm with different sets of initial centroids for 1000 times, returning the best results in terms of homogeneity, completeness and V-measure score [47]. The maximum number of iterations for a single run is set to 300. To compare, we keep the experiment setting unchanged, but add the proposed quadrangle coefficients (i.e., the average i-quad coefficient $\overline{I}$ and the average o-quad coefficient $\overline{O}$) to the baseline features.

Results and discussion. The classification results are given in Table 2. Homogeneity measures whether the samples from a single class belong to a single cluster; completeness measures whether all members of a class are assigned to the same cluster; V-measure score is the harmonic mean between homogeneity and completeness. After adding the two quadrangle coefficients, we observe significant improvement in all three measures (13% increase in homogeneity, 10% increase in completeness and 15% increase in V-measure score). It indicates that the information contained in the quadrangle coefficients are complementary to the information contained in the clustering and closure coefficients, making them discriminative features in classifying networks.

TABLE 2: Homogeneity (Homo.), completeness (Compl.) and V-measure score of the K-means clustering on 16 real-world networks, without and with the quadrangle coefficients (first row and second row respectively).

Features	Homo.	Compl.	V-measure
without quadrangle coefs.	0.700	0.764	0.707
with quadrangle coefs.	0.793	0.841	0.816

In order to further analyse the results, we adopt the Principal Component Analysis (PCA) algorithm to compress the data to a two-dimensional space, and thus visualise the classification results (Figure 10). We can see from Figure 10(a) that the networks are poorly classified by just using three classic topological features (without the two quadrangle coefficients). Only two road networks are correctly allocated to cluster 2. Four PPI networks are separated into two clusters, resulting in a low completeness score; and two food webs are grouped together with two social networks, leading to a low homogeneity score. In contrast, when the quadrangle coefficients are included in the feature set, these networks are better clustered, especially the types of networks that are relatively rich in quadrangles (Figure 10(b)). Two food webs and two road networks are perfectly allocated to cluster 6 and cluster 2, respectively. In addition to that, four PPI networks are kept together within the same cluster, increasing, therefore, the completeness score. We observe, however, no obvious improvement in clustering social networks, citation networks and Q&A networks. This is because quadrangles are relatively underrepresented in these types of networks (for example, their average i-quad coefficients are less than their average clustering coefficients).

Since two more dimensions are added in the comparison, is the result statistically significant, i.e., would any added features lead to the same level of improvement? To answer this question, we conduct a significance test on V-measure score. First, we state the null hypothesis: adding two random features to the baseline feature set will achieve at least the performance of adding two quadrangle coefficients. Then we generate two random features from a uniform distribution over 0 to 1, and append them to the baseline feature set. As previously, we employ the same algorithm and the same setup to group these networks and report the best V-measure score.

To get the distribution, we repeat the experiment $1,000$ times with $1,000$ different sets of randomly generated features. There are only 26 out of $1,000$ sets that achieve a score higher than 0.816. Thus, we have the p-value of the null hypothesis equal to 0.026, meaning the probability of achieving such a result with random features is 0.026. As this p-value is lower than the default threshold of 0.05, the null hypothesis is confidently rejected and the statistical significance of the improvement brought by adding quadrangle coefficients is proved.

4.4 Link prediction

As two new metrics measuring quadrangle formation, the i-quad coefficient and the o-quad coefficient provide additional topological features for a node-level network analysis and inference. As an example, we show their utilities in missing link prediction, where significant improvement is brought by adding them.

Many studies have shown that common neighbours index and its variations such as Adamic-Adar index and resource allocation index perform well in the link prediction problem [48], [49], [50]. Besides, the clustering coefficient and the closure coefficient are proven to be useful features to improve the performance [17], [51]. Therefore, we use these five features as the baseline features in our prediction model, and then test the performance by adding the proposed i-quad and o-quad coefficients. XGBoost, the gradient boosted trees, is used as the prediction model due to its speed and performance.

Setup. We model a network as a graph $G = (V, E)$. For networks having timestamps on edges, we order the edges according to their appearing times and select the first 70% edges and related nodes to form an “old graph”, denoted $G_{old} = (V^*, E_{old})$. The remaining 30% edges filtered by node set V^* will form a “new graph”, denoted $G_{new} = (V^*, E_{new})$. For networks not having timestamps, we randomly shuffle the edges then perform the partition, and we repeat 100 times in order to assess variance and reduce the impact of a single partition on the possible conclusions. The test set is built by node pairs, that appear in the old graph, but do not form a link. Each such pair of nodes indicates a positive or a negative example depending on whether a link between them appears in the new graph.

The training set is built on the old graph, on which we fit four XGBoost models with four sets of features: 1) baseline feature set which includes common neighbours, Adamic-Adar, resource allocation, clustering coefficient and closure coefficient; 2) baseline features plus i-quad coefficient; 3) baseline features plus o-quad coefficient; 4) baseline features plus both i-quad coefficient and o-quad coefficients. Then we evaluate their prediction performances on the test set. For large networks ($|V| > 10K$), we perform a randomised breadth first search sampling [52] of $3K$ nodes on the original graph and repeat 10 times.

TABLE 3: Test set performance comparison measured in ROC-AUC score of four XGBoost classifiers with different features. Second column lists the scores with baseline features (BL) , third column adds i-quad coefficient to baseline features, fourth column adds o-quad coefficient to baseline features, and fifth column adds both i-quad and o-quad coefficients to baseline features. An improvement of more than 2% is put in bold type, and an improvement of more than 5% is indicated by dagger. Last row gives the average (over the datasets) ranking of the four classifiers for comparison, where smaller is better. A classifer receives rank 1 if it has the highest ROC-AUC score, rank 2 if it has the second highest, and so on. If two classifiers share the best score, they both get rank 1.5, and so on. The best ranking is put in bold italic.

Network	w/ baseline features (BL)	add I(i) to BL	add O(i) to BL	add I(i) & O(i) to BL
FW-FloridaDry	0.6703	0.6779	0.6834	**0.6886**
FW-LittleRock	0.8077	**0.8357**	**0.8421**	**0.8521**†
Soc-EmailEu$^{\mathcal{T}}$	0.9076	0.9070	0.9090	0.9084
Soc-ClgMsg$^{\mathcal{T}}$	0.7831	0.7873	0.7879	0.7920
Soc-BTCAlpha$^{\mathcal{T}}$	0.8588	0.8601	0.8679	0.8697
Soc-TwitchFr	0.9160	0.9176	0.9192	0.9202
PPI-Stelzl	0.6565	**0.7778**†	**0.7809**†	**0.7764**†
PPI-Figeys	0.8171	**0.8644**†	**0.8668**†	**0.8650**†
PPI-Vidal	0.7566	**0.7973**†	**0.8009**†	**0.7992**†
PPI-IntAct	0.8524	**0.8808**	**0.8839**	**0.8842**
Cit-DBLP$^{\mathcal{T}}$	0.7294	0.7261	0.7336	0.7310
Cit-Cora	0.8700	0.8705	0.8726	0.8734
Rd-NewYork	0.5268	**0.5529**	**0.5538**†	**0.5538**†
Rd-BayAera	0.5218	**0.5353**	**0.5353**	**0.5356**
QA-MathOvfl.$^{\mathcal{T}}$	0.8546	0.8554	0.8541	0.8551
QA-AskUbuntu$^{\mathcal{T}}$	0.8746	0.8791	0.8765	0.8777
Avg. ranking	3.8	2.8	1.9	***1.5***

Results and discussion. Since network link prediction is a highly unbalanced task, we choose the Area Under the ROC Curve (ROC-AUC) as the metric and report the prediction result on the test set, as shown in Table 3. First, we discover that adding the i-quad (3rd column) or the o-quad coefficient (4th column) leads to improvement in most networks. Furthermore, we find that adding the o-quad coefficient outperforms adding the i-quad coefficient in 14 out of 16 networks. One possible explanation of this phenomenon is that the o-quad coefficient looks 3-hop away from the focal node, which is in line with the recent discovery that 3-hop paths are more powerful predictors in link prediction [9], [53]. When both quadrangle coefficients are added to the baseline features (5th column), the performance is improved in all networks. The average ranking (last row) also shows that adding both i-quad and o-quad coefficients at the same time leads to the best overall performance, closely followed by just adding the o-quad coefficient.

Second, we find that the improvement is particularly significant in food webs, protein-protein interaction networks and road networks (more than 2% in all eight networks of these three types, and more than 5% in five networks when both quadrangle coefficients are added). The common characteristic of these types of networks is that they tend to have larger quadrangle coefficients compared to the clustering and closure coefficients. In other words, the extra information brought by the proposed coefficients is particularly useful in networks that are rich in quadrangles.

To give more statistical insight into these results, we adopt the non-parametric Wilcoxon Signed-Rank Test [54] to quantify the difference between classifiers with different feature sets, reporting the p-value where applicable. Note that this method is rank-based and essentially tests the null hypothesis that two paired samples come from the same distribution. In our setting, paired samples are paired columns from the result table, and rejected null hypothesis means that we would expect one approach to outperform another in a new dataset.

We find that adding the i-quad coefficient, adding the o-quad coefficient, and adding both of them to the baseline features all provide statistically significant gains over only using the baseline feature set (p-values are far less than 0.001 for all three). Moreover, the gains of adding the o-quad coefficient and adding both quadrangle coefficients to baseline features over adding the i-quad coefficient to baseline features are also critically different ($p = 0.005$, comparing adding the o-quad coefficient with adding the i-quad coefficient; $p = 0.003$ comparing adding both quadrangle coefficients with adding the i-quad coefficient). However, there is no significant difference between adding the o-quad coefficient and adding both quadrangle coefficients ($p = 0.35$). Accordingly, we create the critical difference diagram in Figure 11.

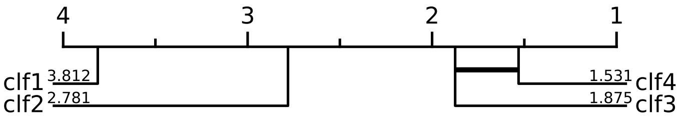


Fig. 11: Critical difference diagram of four classifiers with different feature sets. Classifier 1 (clf1) uses baseline features; classifier 2 (clf2) uses baseline features plus i-quad coefficient; classifier 3 (clf3) uses baseline features plus o-quad coefficient; classifier 4 (clf4) uses baseline features plus i-quad and o-quad coefficients.

4.5 Limitations and Future Directions

Now, we describe several limitations of our work and outline how these might be overcome in future studies.

Directed edges. Our work currently is limited to undirected networks (unweighted or weighted). A natural extension is to further propose the directed quadrangle coefficients in a similar approach as in extending the clustering coefficient and closure coefficient to directed networks [13], [55]. The complexity of this approach comes from the 16 different directed quadrangles. Another possible direction is to focus on one or two directed quadrangles that are proved to be more important in many types of networks, such as the bi-fan or the bi-parallel structures [56], [57].

Network dynamics. Both the i-quad coefficient and the o-quad coefficient are motivated by the view of network evolution — a closing edge appears between the two endpoints of an existing open quadriad and forms a quadrangle. Their definitions, however, do not take into consideration the dynamics of the network. An interesting future direction is to develop the notion of temporal open quadriad, meaning that an open quadriad is present at a certain timestamp while its two endpoints are not connected by a closing edge. Then we can define the temporal quadrangle coefficients as the fraction of temporal open quadriads that are closed at a later time point. With extra temporal information, these counterparts could therefore be more powerful in predicting future links.

Potential applications. Being new metrics of measuring quadrangle formation, the proposed coefficients could be promising in studying networks that are rich in quadrangles — discovering similarities among protein-protein interaction networks [58], detecting compartments in food webs [59], and exploring how robust ecological systems are in the face of species loss [60]. More generally, the quadrangle coefficients also have the potential to be applied in community detection, as shown by the clustering and closure coefficients [17], [61]. Plus, although Graph Neural Networks have achieved state-of-the-art results in various applications, a recent study has exposed their shortcomings in capturing network structures [62]. Therefore, an interesting avenue is to incorporate the structural information brought by the proposed coefficients in the message passing scheme.

5 Related Work

We here recapitulate some related works that proposed other metrics to measure quadrangle formations in networks. Fronczak et al. [63] proposed a higher order clustering coefficient for random networks. It is defined as $C_i(x) = \frac{2E_i(x)}{k_i(k_i-1)}$, where i is the focal node and x is the length of path. $E_i(x)$ denotes the number of x-length paths between the neighbours of i. When x equals 2, this definition deals with the formation of quadrangles. The limitation of this definition is that the normalisation only takes the degree of the focal node i into account while neglects the degree of i's neighbours. Since each pair of neighbours could have multiple length-2 paths between them, the clustering value can be larger than one.

Aiming to measure the formation of 4-cycles, Caldarelli et al. [64] proposed two grid coefficients, i.e., the primary grid coefficient and the secondary grid coefficient. The former is defined as: $G^p(i) = \frac{Q^p(i)}{Z^p(i)}$, where $Q^p(i)$ is the number of actual "primary quadrilaterals" containing node i, and $Z^p(i)$ is calculated by: $Z^p(i) = \frac{k_i(k_i-1)(k_i-2)(k_i-1)}{2}$. With this definition, however, it actually deals with the formation of 4-cycle with an extra diagonal edge. The secondary grid coefficient is defined as: $G^s(i) = Q^s(i)/Z^s(i)$, where $Q^s(i)$ is the number of actual "secondary quadrilaterals" containing node i, and $Z^s(i)$ is calculated by: $Z^s(i) = \frac{k_{i,2nd}k_i(k_i-1)}{2}$. A potential problem within this definition is that it does not rule out the possibility that the 2-hop neighbour connects to two other 1-hop neighbours, making the formed structure containing five nodes.

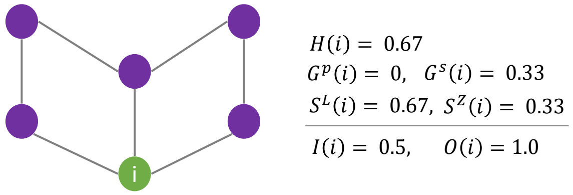

i
$H(i) = 0.67$
$G^p(i) = 0, \; G^s(i) = 0.33$
$S^L(i) = 0.67, \; S^Z(i) = 0.33$
$I(i) = 0.5, \quad O(i) = 1.0$

Fig. 12: An example of the coefficients proposed in related works, compared with our proposed quadrangle coefficients. $H(i)$ is the higher order clustering coefficient proposed by Fronczak et al. [63]; $G^p(i)$ and $G^s(i)$ are the primary grid coefficient and the secondary grid coefficient proposed by Caldarelli et al. [64]; $S^L(i)$ is the square clustering coefficient proposed by Lind et al. [65]; $S^Z(i)$ is another square clustering coefficient proposed by Zhang et al. [66]; $I(i)$ and $O(i)$ are the two quadrangle coefficients proposed by us.

Lind et al. [65] later proposed a square clustering coefficient in the context of bipartite networks by taking into consideration the degree of the neighbours, in other words, the length-2 paths starting from the focal node. It is defined as $C_{4,mn}(i) = \frac{q_{imn}}{(k_m-\eta_{imn})(k_n-\eta_{imn})+q_{imn}}$, where m and n are a pair of neighbours of the focal node i, and q_{imn} denotes the number of squares containing the three nodes. What is uncommon about this definition is that it deems squares are formed via node overlapping, which is not a standard approach. Zhang et al. [66] then modified the equation and proposed another more standard square clustering coefficient for bipartite networks. Their definition is: $C_{4,mn}(i) = \frac{q_{imn}}{(k_m-\eta_{imn})+(k_n-\eta_{imn})+q_{imn}}$. However, in both of these definitions, there is no notion of open quadriad introduced, and the scope is limited within 2-hop distance from the focal node.

The proposed i-quad and o-quad coefficients are different from previous works in that 1) the scope of the o-quad coefficient is larger since it takes into account length-3 paths emanating from the focal node, whereas the square clustering coefficients or the grid coefficients only calculates length-2 paths in the normalisation; 2) the quadrangle coefficients proposed by us view a formed quadrangle as being built from open quadriads via connecting two endpoints with one edge, which conform with the classic clustering and closure coefficients (in their definitions a formed triangle is viewed as being built from open triads). In contrast, two edges are required to form a quadrangle in the grid coefficients; 3) the quadrangle coefficients are proposed for the general unipartite networks on which multiple experiments are conducted. In Figure 12, we provide a simple example to illustrate the five coefficients proposed by previous works and the two quadrangle coefficients proposed by us.

6 Conclusion

In this paper, we introduced the i-quad coefficient and the o-quad coefficient to measure quadrangle formation in networks, according to the different location of the focal node in an open quadriad. We also extended them to weighed networks. Through experiments on 16 real-world networks from six domains, we revealed that 1) in most types of networks, the average o-quad coefficient is smaller than the average i-quad coefficient; 2) in food webs, protein-protein interaction networks and road networks, the i-quad and o-quad coefficients are larger than the clustering and closure coefficients respectively; 3) the o-quad coefficient tends to increase with node degree while the i-quad coefficient does not change too much as the node degree increases.

We also demonstrated that including the two coefficients leads to improvement in both network-level and node-level analysis tasks, such as network classification and link prediction. The improvement is especially significant in food webs, protein-protein interaction networks and road networks in link prediction task. Additionally, we plan to further consider the dynamics of time-varying networks and link directions of directed networks when measuring quadrangle formation in the future. Due to the simplicity and interpretability in the definitions, we anticipate that the i-quad and o-quad coefficients will become standard descriptive features and be incorporated in other network mining tasks.

Acknowledgments

The authors thank the editors and anonymous reviewers for their excellent comments and suggestions. The authors would also thank Joakim Skarding, Xiaohan Zhang and Yu-Xuan Qiu for their helpful comments and discussions. This work was supported by the Australian Research Council, Grant No. DP190101087: “Dynamics and Control of Complex Social Networks”.

References

[1] A.-L. Barabási *et al.*, *Network science*. Cambridge university press, 2016.
[2] M. Newman, *Networks*. Oxford university press, 2018.
[3] K. Musiał and P. Kazienko, “Social networks on the internet,” *World Wide Web*, 2013.
[4] W. L. Hamilton, R. Ying, and J. Leskovec, “Representation learning on graphs: Methods and applications,” *arXiv preprint arXiv:1709.05584*, 2017.
[5] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *KDD*, 2016.
[6] S. Bhagat, G. Cormode, and S. Muthukrishnan, “Node classification in social networks,” in *Social network data analytics*. Springer, 2011.
[7] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
[8] F. Gao, K. Musial, C. Cooper, and S. Tsoka, “Link prediction methods and their accuracy for different social networks and network metrics,” *Scientific programming*, 2015.
[9] I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao *et al.*, “Network-based prediction of protein interactions,” *Nature communications*, 2019.
[10] C. C. Noble and D. J. Cook, “Graph-based anomaly detection,” in *KDD*, 2003.
[11] L. Akoglu, H. Tong, and D. Koutra, “Graph based anomaly detection and description: a survey,” *Data mining and knowledge discovery*, 2015.
[12] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’networks,” *nature*, 1998.
[13] G. Fagiolo, “Clustering in complex directed networks,” *Physical Review E*, 2007.
[14] T. Lee, B. Choi, Y. Shin, and J. Kwak, “Automatic malware mutant detection and group classification based on the n-gram and clustering coefficient,” *The Journal of Supercomputing*, 2018.
[15] R. Goldstein and M. S. Vitevitch, “The influence of clustering coefficient on word-learning: how groups of similar sounding words facilitate acquisition,” *Frontiers in psychology*, 2014.

[16] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "Rolx: structural role extraction & mining in large graphs," in *KDD*, 2012.

[17] H. Yin, A. R. Benson, and J. Leskovec, "The local closure coefficient: A new perspective on network clustering," in *WSDM*, 2019.

[18] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, 2002.

[19] E. Rainone, "The network nature of over-the-counter interest rates," *Journal of Financial Markets*, 2020.

[20] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, 1992.

[21] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, 2009.

[22] M. E. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical review E*, 2001.

[23] S. Qasim, "Some problems related to the food chain in a tropical estuary," *Marine food chains*, 1970.

[24] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *PNAS*, 2004.

[25] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, "Intensity and coherence of motifs in weighted complex networks," *Physical Review E*, 2005.

[26] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, 2005.

[27] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertesz, "Generalizations of the clustering coefficient to weighted complex networks," *Physical Review E*, 2007.

[28] V. Colizza, R. Pastor-Satorras, and A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks," *Nature Physics*, 2007.

[29] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social networks*, 2009.

[30] J. Kunegis, "Konect: the koblenz network collection," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[31] J. Leskovec and R. Sosič, "Snap: A general-purpose network analysis and graph-mining library," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2016.

[32] R. E. Ulanowicz and D. L. DeAngelis, "Network analysis of trophic dynamics in south florida ecosystems," *US Geological Survey Program on the South Florida Ecosystem*, 1999.

[33] N. D. Martinez, "Artifacts or attributes? effects of resolution on the little rock lake food web," *Ecological monographs*, 1991.

[34] A. Paranjape, A. R. Benson, and J. Leskovec, "Motifs in temporal networks," in *WSDM*, 2017.

[35] P. Panzarasa, T. Opsahl, and K. M. Carley, "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community," *Journal of the American Society for Information Science and Technology*, 2009.

[36] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "Rev2: Fraudulent user prediction in rating platforms," in *WSDM*, 2018.

[37] B. Rozemberczki, C. Allen, and R. Sarkar, "Multi-scale attributed node embedding," *arXiv preprint arXiv:1909.13021*, 2019.

[38] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen *et al.*, "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, 2005.

[39] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li *et al.*, "Large-scale mapping of human protein–protein interactions by mass spectrometry," *Molecular systems biology*, 2007.

[40] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou *et al.*, "Towards a proteome-scale map of the human protein–protein interaction network," *Nature*, 2005.

[41] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. Del-Toro *et al.*, "The mintact project—intact as a common curation platform for 11 molecular interaction databases," *Nucleic acids research*, 2014.

[42] M. Ley, "The dblp computer science bibliography: Evolution, research issues, perspectives," in *International symposium on string processing and information retrieval*. Springer, 2002.

[43] L. Šubelj and M. Bajec, "Model of complex networks based on citation dynamics," in *Proceedings of the 22nd international conference on World Wide Web*, 2013.

[44] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, "Configuring random graph models with fixed degree sequences," *SIAM Review*, 2018.

[45] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, 2004.

[46] S. E. Ahnert and T. M. Fink, "Clustering signatures classify directed networks," *Physical Review E*, 2008.

[47] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *EMNLP-CoNLL*, 2007.

[48] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, 2007.

[49] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, 2003.

[50] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, 2009.

[51] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *SDM06: workshop on link analysis, counter-terrorism and security*, 2006.

[52] C. Doerr and N. Blenn, "Metric convergence in social network sampling," in *Proceedings of the 5th ACM workshop on HotPlanet*, 2013.

[53] T. Zhou, Y.-L. Lee, and G. Wang, "Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 564, p. 125532, 2021.

[54] S. Siegel, "Nonparametric statistics for the behavioral sciences." 1956.

[55] M. Jia, B. Gabrys, and K. Musial, "Directed closure coefficient and its patterns," *PLOS ONE*, vol. 16, pp. 1–23, 06 2021.

[56] Q.-M. Zhang, L. Lü, W.-Q. Wang, T. Zhou *et al.*, "Potential theory for directed networks," *PloS one*, vol. 8, no. 2, p. e55437, 2013.

[57] X. Hu, S. Liu, S. Chang, and H. Li, "A quad motifs index for directed link prediction," *IEEE Access*, vol. 7, pp. 159 527–159 534, 2019.

[58] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, "Topological network alignment uncovers biological function and phylogeny," *Journal of the Royal Society Interface*, vol. 7, no. 50, pp. 1341–1354, 2010.

[59] A. E. Krause, K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor, "Compartments revealed in food-web structure," *Nature*, vol. 426, no. 6964, pp. 282–285, 2003.

[60] J. A. Dunne, R. J. Williams, and N. D. Martinez, "Network structure and biodiversity loss in food webs: robustness increases with connectance," *Ecology letters*, vol. 5, no. 4, pp. 558–567, 2002.

[61] Q. Ji, D. Li, and Z. Jin, "Divisive algorithm based on node clustering coefficient for community detection," *IEEE Access*, vol. 8, pp. 142 337–142 347, 2020.

[62] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

[63] A. Fronczak, J. A. Hołyst, M. Jedynak, and J. Sienkiewicz, "Higher order clustering coefficients in barabási–albert networks," *Physica A: Statistical Mechanics and its Applications*, 2002.

[64] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani, "Structure of cycles and local ordering in complex networks," *The European Physical Journal B*, vol. 38, no. 2, pp. 183–186, 2004.

[65] P. G. Lind, M. C. Gonzalez, and H. J. Herrmann, "Cycles and clustering in bipartite networks," *Physical review E*, 2005.

[66] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan, "Clustering coefficient and community structure of bipartite networks," *Physica A: Statistical Mechanics and its Applications*, 2008.

Mingshan Jia received the B.E. degree in Information Engineering from Xi'an Jiaotong University, Xi'an, China, in 2008, and the M.E. degree in Information and Telecommunication System from the University of Technology of Troyes, Troyes, France, in 2011. He had worked at China Electronics Technology Group, Huawei Technologies and Xijing University for seven years before he studied a PhD. Currently, he is working towards the PhD degree at the School of Computer Science, University of Technology Sydney, Sydney, Australia. His general research area is applied machine learning and data science for large complex systems.

Bogdan Gabrys received the M.Sc. degree in electronics and telecommunication from Silesian Technical University, Gliwice, Poland, in 1994, and the Ph.D. degree in computer science from Nottingham Trent University, Nottingham, U.K., in 1998. Over the last 27 years, he has been working at various universities and research and development departments of commercial institutions. He is currently a Professor of Data Science and a Co-Director of the Complex Adaptive Systems Laboratory at the University of Technology Sydney, Sydney, Australia. His research activities have concentrated on the areas of data science, complex adaptive systems, computational intelligence, machine learning, predictive analytics, and their diverse applications. He has published over 200 research papers, chaired conferences, workshops, and special sessions, and been on program committees of a large number of international conferences with the data science, computational intelligence, machine learning, and data mining themes. He is also a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE), a Memebr of IEEE Computational Intelligence Society and a Fellow of the Higher Education Academy (HEA) in the UK. He is frequently invited to give keynote and plenary talks at international conferences and lectures at internationally leading research centres and commercial research labs. More details can be found at: http://bogdan-gabrys.com

Katarzyna Musial received her MSc in Computer Science from Wroclaw University of Science and Technology, Poland, and an MSc in Software Engineering from the Blekinge Institute of Technology, Sweden, both in 2006. She was awarded her PhD in November 2009 from WrUST, and in the same year she was appointed a Senior Visiting Research Fellow at Bournemouth University, where from 2010 She was a Lecturer in Informatics. She joined King's in November 2011 as a Lecturer in Computer Science. In September 2015 she returned to Bournemouth University where she was an Associate Professor in Computing as well as a Head of SMART Technology Research Group and a member of Data Science Initiative. In September 2017 she moved to Australia and started working as Associate Professor in Network Science in the School of Computer Science at University of Technology Sydney.