Social triangles and generalized clustering coefficient for weighted networks

Roy Cerqueti[1] Giovanna Ferraro[2] Antonio Iovanella[2]

[1] Department of Economics and Law
University of Macerata
Via Crescimbeni, 20 - 62100 Macerata, Italy
roy.cerqueti@unimc.it

[2] Department of Enterprise Engineering
University of Rome Tor Vergata
Via del Politecnico, 1 - 00133 Rome, Italy.
giovanna.ferraro@uniroma2.it
antonio.iovanella@uniroma2.it

Abstract

A way to measure the community structure of a network is the clustering coefficient. Such a quantity is based on the number of existing triangles around the nodes over the theoretical ones. To the best of our knowledge, scarce attention has been paid to the fictitious triangles due to the presence of indirect connections among the nodes of the network. This paper fills this gap by providing a new definition of the clustering coefficient for weighted networks when missing links might be also considered. Specifically, a novel concept of triangles is here introduced by assuming that a strong enough aggregate weight of two arcs sharing a node induces a link between the not common nodes. Beyond the intuitive meaning of such social triangles, we also explore the usefulness of them for gaining insights on the topological structure of the underline network. Empirical experiments on the standard networks of 500 commercial US airports and on the nervous system of the Caenorhabditis elegans support the theoretical framework.

Keywords: networks; clustering coefficient; weighted arcs; social triangles.

1 Introduction

The network approach to complex systems has revealed several general and unexpected findings applicable to a large number of systems, such as the ubiquity of scale freeness, the frequent appearance of high clustering, and the relationship between functionality and the presence of specific motifs ([2], [27]).

In recent years, it has become clear that it is relevant to consider the heterogeneity of the interactions and their correlations with the network structure in order to understand the characteristics of the system. In particular, a survey of measures of complex networks is reported in [11] while an extended approach related to some centrality measures is presented in [17]. Such quantities are applied in different real networks, for example, in [25] are considered measures of importance and power in terrorist networks, in [13] are analyzed corporate systems while in [22] are examined social networks.

We consider, in this work, the tendency for clustering, i.e. the link formation between neighboring vertices [40] that reveals the clustering of edges in tightly connected neighborhood and identifies the local group cohesiveness.

The measure for assessing the tendency of vertices to cluster is the *local cluster coefficient* [40]. Such quantity has been extensively studied by several authors and applied in different networks ([30], [39], [40], [44]). It captures the degree of social embeddedness of the nodes in a network and is based on local density [37]. Indeed, especially in social networks, vertices tend to create tightly knit groups that are characterized by a relatively high density of links [36].

The clustering coefficient assesses the connectivity in a node's neighborhood; a node has a high value of clustering coefficient of its neighbors tends to be directly connected with each other [12]. This quantity is relevant to determine the small-world property of a network [19] and can be considered as an index of the redundancy of a node ([9], [21]). In the contribution of Benati et al. [8], it is proposed a new combinatorial model to detect clusters that takes into consideration the standard individual data pertinent to a single population unit and the data describing the connections among units. More in general, a review of the dominant set clustering framework is presented in [34]. Regarding the clustering coefficient in weighted networks, it has been analyzed in ([6], [18], [28],

[29], [31], [43]) as reported in Section 3. Related to the clustering coefficient in terms of link prediction, the estimation of the likelihood of new link creation is presented in ([23], [42]). The connections between nodes using suitable definitions of the matrix governing the connections has been considered by [3].

The analysis of the weights along the edges and their correlations is able to change the view of the hierarchical and structural organization of the systems. This is evident if we consider, as an example, a network in which the weights of all links forming triangles of interconnected vertices are extremely small. In this case, even for a large clustering coefficient, these triangles play a minimal role in the network dynamics and organization, and the clustering features are certainly overestimated by a simple structural analysis [7]. Also, vertices with high degree can be attached to a majority of low-degree nodes whilst concentrating the largest portion of their strength only on the vertices with high degree. In this situation, the topology reveals a disassortative characteristic of the network, whereas the system could be considered assortative since the more relevant edges in terms of weights are linked to the high-degree vertices.

In this paper, we aim at elaborating on the presence of large weights on the arcs of the network, which are able to induce indirect links among the nodes.

Specifically, we provide a new definition of the concept of triangles of a network of social type, by allowing the presence of indirect links. One has triangles even in the case of one missing side, but under the constraint that the weights of the remaining sides – suitably aggregated – are *large enough*. These new triangles admit the introduction of social intensity as a potential mechanism for links formation.

In particular, we define two different new kinds of triangle around a node i: the first one is composed by two not linked nodes which are adjacent to i; the second one is composed by a trajectory of length two from i such that i is not linked with the extreme node of the trajectory.

The constraints on the weights of the arcs are driven by two thresholds, one for each type of triangle. As we will see below in the formalization of our setting, null thresholds mean no constraints – and all the two-sided figures are triangles – while a large value of the thresholds is associated to very restrictive constraints – and a small number of two-sided

figures will be accepted as triangles.

It is very important to notice that the case of zero thresholds gives also insights on the topological structure of the unweighted graph associated to the network. We address the reader to the empirical analysis section for an intuitive explanation of this point.

Starting from the considered definition of triangles, a definition of a new generalized weighted clustering coefficient is provided. This measure is able to give useful hints on the community structure of the network, when weights play a relevant role in inducing missing links among the node. Moreover, this measure is also useful for predicting the fictitious links that may appear in the future of evolving networks. Link prediction is a relevant problem that attempt to estimate the likelihood of the existence of a link between two vertices based on observed links and the attributes of nodes [1]. Such prediction can be used to analyze a network to suggest promising interactions or collaborations that have not yet identified or is related to the problem of inferring missing or additional links that, while not directly visible, are likely to exist ([24], [26]). Our generalized clustering coefficient has a further very relevant property: it assumes unitary value in several situations and not only when the graph is a clique. Specifically, the community structure of the network is intended to include also the realistic cases of the presence of indirect connections among two agents induced by their strong links with a third node.

To gain more information, we have also implemented a sensitivity analysis of the clustering coefficient with respect to the exogenous thresholds defining the triangles. Moreover, a discussion on the way to aggregate the weights for identifying the triangles has been provided, in order to specify different concepts of sociality. The proposed clustering coefficient is tested on two well-established empirical settings.

The paper is structured as follows; after preliminaries and notations about the graph theory (Section 2), we review the literature on the clustering coefficient in weighted networks (Section 3), we propose a generalized clustering coefficient and generalized triangles with the relative interpretation (Section 4), the computational experience of two empirical networks: the nervous system of the nematode Caenorhabditis elegans and the network among the 500 commercial airports in the United States is presented in Section 5. The paper ends with some conclusions and remarks on future research directions.

2 Preliminaries and notations about graph theory

Networks can be considered as a valuable representation of many complex systems found in the real world. The abstract representation of a network is a graph. It consists of a set of vertices (or nodes) and a set of edges (or links). The presence of an edge between two vertices signifies the presence of some kind of interactions or connections between the vertices. The classical mathematical abstraction of a network is a graph $G = (V, E)$, where V is the set of N nodes (or vertices) and E is the set of M links (or edges) stating the relationships among the nodes. We refer to a node by an index i, meaning that we allow a one-to-one correspondence between an index in $\{1, \ldots, N\}$ and a node in V. Generally, the existence of a link between two given nodes is captured by a binary variable, whose value is 1 if the link exists and 0 otherwise. In so doing, the set M can be conceptualized through a squared matrix of order N – the so-called *adjacent matrix* $\mathbf{A} = (a_{ij})_{i,j=1,\ldots,N}$ – which is filled by 0s and 1s and whose general element a_{ij} is 1 if the link between i and j exists and 0 otherwise. The graph is symmetric when $a_{ij} = a_{ji}$, for each $i, j = 1, \ldots, N$, and asymmetric otherwise. In this latter case, links are (and the graph is) oriented. In this paper, we examine weighted networks that can be denoted mathematically by an adjacency matrix $\mathbf{W}$ with elements $w_{ij} \geq 0$ which represent the weights on the link connecting nodes i and j, with $i, j = 1, \ldots, N$. Thus, w_{ij} denotes the intensity of the interactions between two nodes i and j and allows for modeling the ties' strength of the observed system.

The degree d_i of the node i is a nonnegative integer representing the number of links incident upon i. The degree gives a measure of how well a node is connected to the other elements of the graph, and provides also information on which nodes tend to cluster together.

3 Literature review on the clustering coefficient for unweighted and weighted networks

3.1 Unweighted networks

The *local clustering coefficient* is defined for any vertex $i = 1, \ldots, N$ as the fraction of its connected neighbors and captures the capacity of edge creations among neighbors, i.e. the

tendency in the network to create stable groups [40]. Thus, the clustering around a vertex i is quantified by the (unweighted) local clustering coefficient C_i defined as the number of triangles in which vertex i participates normalized by the maximum possible number of such triangles:

$$C_i = \frac{2t_i}{d_i(d_i - 1)} \tag{1}$$

where t_i represents the number of triangles around i. The local clustering coefficient quantifies how well a node is connected to the other elements of the graph together with how nodes tend to cluster together. Therefore, $C_i = 0$ if none of the neighbors of a node are connected and $C_i = 1$ if all of the neighbors are linked.

The value of the local clustering coefficient is influenced by the nodes degrees, a node with several neighbours is likely to be embedded in rather fewer closed triangles. Hence it has a smaller local clustering coefficient respect to a node linked to fewer neighbors where they are more likely to be clustered in triangles [4].

The clustering coefficient for a given graph is computed in two classical modes [27]. The first one is the *averaged clustering coefficient* $\overline{C}$ given as the average of all the local clustering coefficients, while the second, called the *global clustering coefficient* and denoted by C_G, is defined as the ratio among three times the number of closed triangles in the graph and the number of its triplets, i.e. the number of 2-paths among three nodes.

Note that both $\overline{C}$ and C_G assume values from 0 to 1 and are equal to 1 in case of a clique, i.e. a fully coupled network. In real networks, the evidence shows that nodes are inclined to cluster into densely connected groups ([15], [38]) and the difficulty to compare the values of clustering of nodes with different degrees makes the average value of local clustering sensitive with respect to how degrees are distributed across the whole network.

The quantities $\overline{C}$ and C_G are specifically talored to unweighted networks, and they cannot be satisfactorily employed for describing the community structure of the network in presence of weights on the links and when arcs are of direct type.

Next section is devoted to the analysis of the more general weighted case.

3.2 Weighted networks

In many real networks, connections are relevant not only in terms of the classical binary state – whenever they exist or do not exist – but also with regards to their strength which, for any node $i = 1, \ldots, N$, is defined as:

$$s_i = \sum_{j=1}^{N} w_{ij}. \tag{2}$$

The introduction of weights and strengths extends the study of the macroscopic properties of the network, by adding to the mere interactions some forms of entity of connections and capability. In particular, the strength integrates information about the vertex connectivity and the weights of its links [7]. It is considered a natural measure of the importance or centrality of a vertex i. Indeed, the identification of the most central nodes represents a major issue in network characterization [16].

Barrat et al. [6] combine the topological information of the network with the distribution of weights along links and define the weighted clustering coefficient for a node $i = 1, \ldots, N$ as follows:

$$\tilde{C}_{i,B} = \frac{1}{s_i(d_i - 1)} \sum_{j,k \in V} \frac{w_{ij} + w_{ik}}{2} a_{ij} a_{jk} a_{ik}. \tag{3}$$

This coefficient is a quantity of the local cohesiveness that considers the importance of the clustered structure taking into account the intensity of the interactions found on the local triangles. This measure counts for each triangle created in the neighborhood of the node i the weight of the two related edges. The authors refer not just to the number of the closed triangles in the neighborhood of a node but also to their total relative weight with respect to the strength of the nodes.

The normalization factor $s_i(d_i - 1)$ accounts for the weights of each edge times the maximum possible number of triangles in which it may participate, and it ensures that $0 \leq \tilde{C}_{i,B} \leq 1$. The definition of $\tilde{C}_{i,B}$ recovers the topological clustering coefficient in the case where w_{ij} is constant, for each j.

Therefore, the authors introduce the weighted clustering coefficient averaged over all nodes of the network, say C^W, and over all nodes with degree d, say $C^W(d)$. These

measures offer global information on the correlation between weights and topology by comparing them with their topological analogs.

Note that $s_i = d_i(s_i/d_i) = d_i\langle w_i\rangle$, so $\tilde{C}_{i,B}$ can be written as:

$$\tilde{C}_{i,B} = \frac{1}{d_i(d_i-1)} \sum_{j,k\in V} \frac{w_{ij}+w_{kj}}{2\langle w_i\rangle} a_{ij}a_{jk}a_{ik} \tag{4}$$

where $\langle w_i\rangle = \sum_j w_{ij}/d_i$. In such equation the contribution of each triangle is weighted by a ratio of the average weight of the two adjacent links of the triangle to the average weight $\langle w_i\rangle$.

Thus, $\tilde{C}_{i,B}$ compares the weights related with triangles to the average weight of edges connected to the local node.

Zhang and Horvath [43] describe the weighted clustering coefficient in the context of gene co-expression networks. Unlike the unweighted clustering coefficient, the weighted clustering coefficient is not inversely related to the connectivity. Authors show a model that reveals how an inverse relationship between clustering coefficient and connectivity occurs from hard thresholding. In formula:

$$\tilde{C}_{i,Z} = \frac{\sum_{j,k\in V} \hat{w}_{ij}\hat{w}_{jk}\hat{w}_{ik}}{(\sum_{k\in V}\hat{w}_{ik})^2 - \sum_k \hat{w}_{ik}^2} \tag{5}$$

where the weights have been normalized by $\max(w)$. The number of triangles around the node i can be written in terms of the adjacency matrix elements as $t_i = 1/2\sum_{i,k\in V} a_{ij}a_{jk}a_{ik}$ and the numerator of the above equation is a weighted generalization of the formula. The denominator has been selected by considering the upper bound of the numerator, ensuring $\tilde{C}_{i,Z} \in [0,1]$. The equation can be written as:

$$\tilde{C}_{i,Z} = \frac{\sum_{j,k\in V} \hat{w}_{ij}\hat{w}_{jk}\hat{w}_{ik}}{\sum_{j,k\in V:j\neq k} \hat{w}_{ij}\hat{w}_{ik}} \tag{6}$$

In [18] a similar definition has been shown, indeed the edge weights are considered as probabilities such that in an ensemble of networks, i and j are linked with probability $\hat{w}_{ij}$. Finally, Holme et al [20] discuss the definition of weights and express a redefined weighted clustering coefficient as:

$$\tilde{C}_{i,H} = \frac{\sum_{j,k\in V} w_{ij}w_{jk}w_{ik}}{\max(w)\sum_{j,k\in V} w_{ij}w_{ik}} = \frac{\mathbf{W}^3}{(\mathbf{W}\mathbf{W}_{max}\mathbf{W})_{ii}} \tag{7}$$

where $\mathbf{W}_{max}$ indicates a matrix where each entry equals $\max(w)$. This equation seems similar to those of [43] though, $j \neq k$ is not required in the denominator sum.

Onnela et al. ([28], [29]) refer to the notion of motif defining it as a set (ensemble) of topologically equivalent subgraphs of a network. In case of weighted systems, it is relevant to deal with intensities instead of numbers of occurrence. Moreover, the latter concept is considered as a special case of the former one. For the authors the triangles are among the simplest nontrivial motifs and have a crucial role as one of the classic quantities of network characterization in defining the clustering coefficient of a node i. They propose a weighted clustering coefficient taking into consideration the subgraph intensity that is defined as the geometric average of subgraph edge weights. In formula:

$$\tilde{C}_{i,O} = \frac{2}{d_i(d_i-1)} \sum_{j,k \in V} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3} \tag{8}$$

where $\hat{w}_{ij} = w_{ij}/\max_{j \in V}(w_{ij})$ are the edge weights normalized by the maximum weight in the network of the edges linking i to the other nodes of V.

Formula (8) shows that triangles contribute to the creation of $\tilde{C}_{i,O}$ according to the weights associated to their three edges. More specifically, $\tilde{C}_{i,O}$ disregards the strength of the local node and measure triangle weights only in relation to the maximum edge weight.

Moreover, $\tilde{C}_{i,O}$ collapses to C_i when, for each $i, j \in V$, one has $w_{ij} = a_{ij}$, hence in the unweighted case.

4 The generalized clustering coefficient

This section contains our proposal for a new definition of the clustering coefficient of weighted networks. The ground of the definition is a novel concept of triangles, to include also the presence of real indirect connections among individuals. To our purpose, we first provide the definition of the triangles and discuss it, and then we introduce the clustering coefficient.

4.1 Generalized triangles

Here, we propose a generalization of the concept of triangle, and rewrite accordingly the indices C_i in (1) for the case of weighted networks.

Definition 4.1 *Let us consider a weighted non-oriented graph $G = (V, E)$ with vertices $V = \{1, \ldots, N\}$, symmetric adjacent matrix $\mathbf{A} = (a_{ij})_{i,j=1,\ldots,N}$ and weight matrix $\mathbf{W} = (w_{ij})_{i,j=1,\ldots,N}$, with nonnegative weights. Moreover, let us take $\alpha, \beta \in [0, \infty)$ and a function $F : [0, +\infty)^2 \to [0, +\infty)$ which is not decreasing in its arguments.*

For each triple of distinct vertices $(i, j, k) \in V^3$, a subgraph $t = (\{i, j, k\}, E_T)$ is a triangle around i *if one of the following conditions are satisfied:*

$T1$ $a_{ij} = a_{ik} = a_{jk} = 1$*;*

$T2$ $a_{ij} = a_{ik} = 1$, $a_{jk} = 0$ *and* $F(w_{ij}, w_{ik}) \geq \alpha$*;*

$T3$ $a_{ij} = a_{jk} = 1$, $a_{ik} = 0$ *and* $F(w_{ij}, w_{jk}) \geq \beta$.

We denote the set of triangles associated to case T_h as $\mathcal{T}_h^{(i)}$, for $h = 1, 2, 3$. By definition, $\mathcal{T}_1^{(i)} \cap \mathcal{T}_2^{(i)} = \mathcal{T}_1^{(i)} \cap \mathcal{T}_3^{(i)} = \mathcal{T}_2^{(i)} \cap \mathcal{T}_3^{(i)} = \emptyset$. We denote the set collecting all the triangles by $\mathcal{T}^{(i)} = \mathcal{T}_1^{(i)} \cup \mathcal{T}_2^{(i)} \cup \mathcal{T}_3^{(i)}$.

Figure 1 reports the three different type of triangles, respectively $T1$, $T2$ and $T3$. Clearly, in case $T1$, the concept of triangle given in Definition 4.1 coincides with the standard one.

Note that with N nodes, the maximum number of possible triangles is $|\mathcal{T}_1|^* = \max |\mathcal{T}_1| = \binom{N}{3}$. This is the case of a clique with $C_i = 1$, for each $i \in V$.

When considering the maximum number of candidates triangles for a node i to belong to $T2$, it is $|\mathcal{T}_2^{(i)}|^* = \max |\mathcal{T}_2^{(i)}| = \binom{d_i}{2}$. Then, in this case for the node i the number of triangles is $|\mathcal{T}_2^{(i)}| = |\mathcal{T}_2^{(i)}|^* - |\mathcal{T}_1^{(i)}|$

Triangles in $T3$ for node i are path of length 2 (i.e. triplets), that can be computed considering the square of the adjacency matrix. Indeed. the number of different path of length 2 from i to k equals the entry a_{ik} of A^2 [32]. For a given row i of A^2, the sum of the element (excluding the element a_{ii}) equals the maximum potential number of pseudo-triangles of type $T3$.

Herein we define the elements in $T2$ and $T3$ as *pseudo-triangles* since they do not are contained in G but can be seen as triangles under conditions on the weights of the two edges.

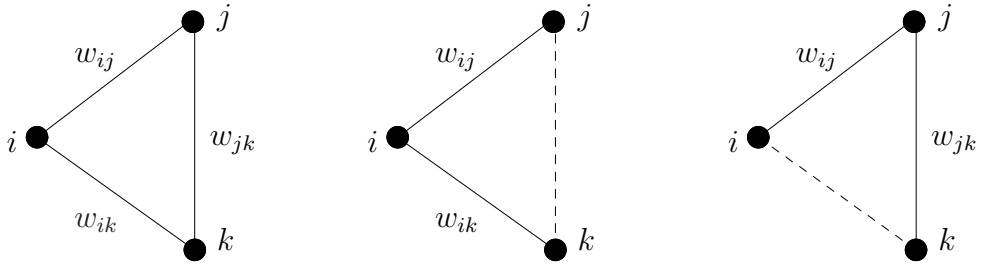

Figure 1: Types of triangles: $T1$ (left), $T2$ (center), $T3$ (right).

Figure 1 shows the types of triangles, without emphasis on the conditions on the weights.

4.1.1 Interpretation of the triangles and equivalent graphs

The pseudo-triangles $T2$ and $T3$ have an interpretation whose ground is the theory of social networks.

The former case describes a situation in which agent i has a direct relationship with agents j and k. One can say that there exists a triangle among the three if the strength of the connections of i with the others is sufficiently high – in the sense described by function F. The idea is that the cooperation between i and the other agents is so effective and fruitful that the presence of a direct link between j and k is not required.

The latter case is associated to the presence of a strong link between i and j and between j and k, always in terms of the entities of the weights – in the sense described by function F. In this peculiar situation, the node j represents the intermediate agent letting also the (indirect) collaboration between i and k be possible.

In our empirical experiments (see below), we have considered four cases of function F:

F_1 sum of the weights is greater than the correspondent coefficient: $w_{ij} + w_{ik} \geq \alpha$ and $w_{ij} + w_{jk} \geq \beta$;

F_2 average of the weights is greater than the correspondent coefficient: $(w_{ij}+w_{ik})/2 \geq \alpha$ and $(w_{ij} + w_{jk})/2 \geq \beta$;

F_3 minimum of the weights is greater than the correspondent coefficient: $\min\{w_{ij}, w_{ik}\} \geq \beta$ and $\min\{w_{ij}, w_{jk}\} \geq \beta$;

F_4 maximum of the weights is greater than the correspondent coefficient: $\max\{w_{ij}, w_{ik}\} \geq \alpha$ and $\max\{w_{ij}, w_{jk}\} \geq \beta$.

The selection of the specific function F – to be implemented among $F_1, \ldots, F_4$ defined above – provides further insights on the interpretation of the triangles of type $T2$ and $T3$. Indeed, once α and β are kept fixed, then F_1 and F_2 state that both weights of the considered edges should be taken into account in an identical way by considering their mere aggregation in the former case or their mean in the latter one. When considering functions F_3 and F_4, only one of the weights is relevant for the measurement of the strength of the connections – the minimum weight and the maximum one, respectively. Naturally, the former case is more restrictive than the latter one, since it implicitly assumes that both weights should be greater than α or β for having a triangle of type $T2$ or $T3$.

Social sciences can suggest other functions F's to be considered in Definition (4.1) in order to capture some peculiarities of the system under observation.

Notice also that $|\mathcal{T}_2^{(i)}|$ and $|\mathcal{T}_3^{(i)}|$ are decreasing functions of α and β, respectively, as Definition 4.1 immediately gives.

Triangles $T1$, $T2$ and $T3$ serve also for deriving topological information on the graph. In particular, assume that $\alpha = \beta = 0$, so that the number of $T2$ and $T3$ around each node does not depend on the specific selection of function F. In this case, we know that $|\mathcal{T}_2^{(i)}| = \binom{d_i}{2}$, so that we are able to infer the degree of the node i by the knowledge of the number of triangles of type $T2$ around it. Differently, $|\mathcal{T}_3^{(i)}|$ represents the number of existing paths of length two having i as one of the extreme nodes. By collecting the number of the triangles $T1$, $T2$ and $T3$ for each node of the graph, we are able to identify a class of graphs.

Formally, consider a $3 \times N$ matrix collecting $|\mathcal{T}_1^{(i)}|$, $|\mathcal{T}_2^{(i)}|$ and $|\mathcal{T}_3^{(i)}|$, for each node $i \in V$. Denote by $\mathcal{M}^{3,N}(\mathbb{N})$ the set of all the matrices with dimension $3 \times N$ and filled by integer nonnegative numbers.

Thus, each matrix $\mathbf{M} \in \mathcal{M}^{3,N}(\mathbb{N})$ identifies a not unique graph having N nodes and with edges described by $\mathbf{M}$. We call such matrix as the *triangles matrix*. In this sense, $\mathbf{M}$ can be viewed as an equivalent class in the set of the graph with N nodes, where two graphs G_1 and G_2 are said to be equivalent when they share the same matrix $\mathbf{M}$.

Figure 2: Two equivalent graphs, according to the equivalence defined through the triangles matrix. In this case, the triangles matrix $\mathbf{M}$ associated to the graphs is given in Table 1.

node i	$\lvert\mathcal{T}_1^{(i)}\rvert$	$\lvert\mathcal{T}_2^{(i)}\rvert$	$\lvert\mathcal{T}_3^{(i)}\rvert$
1	0	0	3
2	0	0	3
3	0	0	3
4	0	5	1
5	0	1	6
6	0	5	1
7	0	0	3
8	0	0	3
9	0	0	3

Table 1: Triangles matrix $\mathbf{M}$ associated to the graphs in Figure 2.

Figure 2 and matrix in (1) provides an example of two equivalent class along with their common triangles matrix $\mathbf{M}$. In particular, notice that matrix $\mathbf{M}$ is symmetric, hence suggesting that the equivalent class identified by $\mathbf{M}$ contains more than one graph.

4.2 Conceptualization of the generalized clustering coefficient

Under Definition (4.1), we can introduce a generalization of the clustering coefficients presented in Formula (1) for weighted networks.

Definition 4.2 *Given a graph $G = (V, E)$ and a node $i \in V$, the* generalized unweighted clustering coefficient of i *is*

$$C_i^{(g)} = \frac{|\mathcal{T}^{(i)}|}{D_i} \tag{9}$$

where $D_i = \frac{d_i(d_i-1)}{2} + |\{j \in V : \Delta_{min}(i,j) = 2\}|$, where $\Delta_{min}(i,j)$ is the minimum distance between the nodes i and j.

The term *unweighted* in Definition 4.2 points to absence of w's in the coefficient in (9). However, weights intervene in the identification of the triangles, according to Definition 4.1.

In particular, formula (9) extend (1). Indeed, notice that $C_i^{(g)} = C_i$ when $\mathcal{T}_2 \cup \mathcal{T}_3 = \emptyset$. This happens in the case of α and β *large enough*, so that there are no pseudo-triangles around a given node of the network.

Importantly, $C_i^{(g)}$ assumes unitary value not only in the clique case, but also when any missing link is compensated by the high weights of the other two links. This property of the generalized clustering coefficient is very relevant, since it allows to extend the sense of community given by the clustering coefficient to the case of presence of indirect links, according to the definition of pseudo-triangles $T2$ and $T3$.

5 Test on real instances

Herein we considered the analysis of the generalized clustering coefficient on two empirical networks: the nervous system of the nematode Caenorhabditis elegans ([40]; [41]) and the network among the 500 busiest US commercial airports [10]. The data processing, the network analysis and all simulations are conducted using the software R ([35]) with the *igraph* package ([14]). The dataset are obtained from the R packege *tnet* authored by Tore Opsahl (http://toreopsahl.com).

The network of nematode Caenorhabditis elegans (C.elegans) has $n = 296$ nodes representing neurons and $m = 1370$ edges occurring when two neurons are connected by either a synapse or a gap junction; for each edge, weights are equal to the number of junctions between node i and j. The network has a scale-free organization with $\gamma \simeq 3.14$ ([5]; [33]).

The US commercial airport network has $n = 500$ nodes denoting airports and $m = 2980$ edges representing flight connections. In this network, weights are the number of seats available on that connections in 2010. The network has both small-world and scale-free organization with $\gamma \simeq 1.8$ [7].

In Figure 3 we show the network visualization while in Table 2 are reported some basic measures as the density δ, the averaged clustering coefficient $\overline{C}$, the global clustering coefficient C_G and the minimum, the maximum and the average degree, weight and strength.

Network	δ	$\overline{C}$	C_G	k_{min}	k_{max}	$\overline{k}$
C.elegans	0.0314	0.228	0.121	1	134	9.26
US airport	0.0239	0.617	0.351	1	145	11.92
Network	w_{min}	w_{max}	$\overline{w}$	s_{min}	s_{max}	$\overline{s}$
C.elegans	1	61	4.198	1	1700	38.86
US airport	9	2253992	152320.19	9416	49316361	1815656.66

Table 2: Basic measures for the networks under analysis

Note that for the C.elegans network we considered the giant component of 296 nodes while the complete network is composed of 306 nodes.

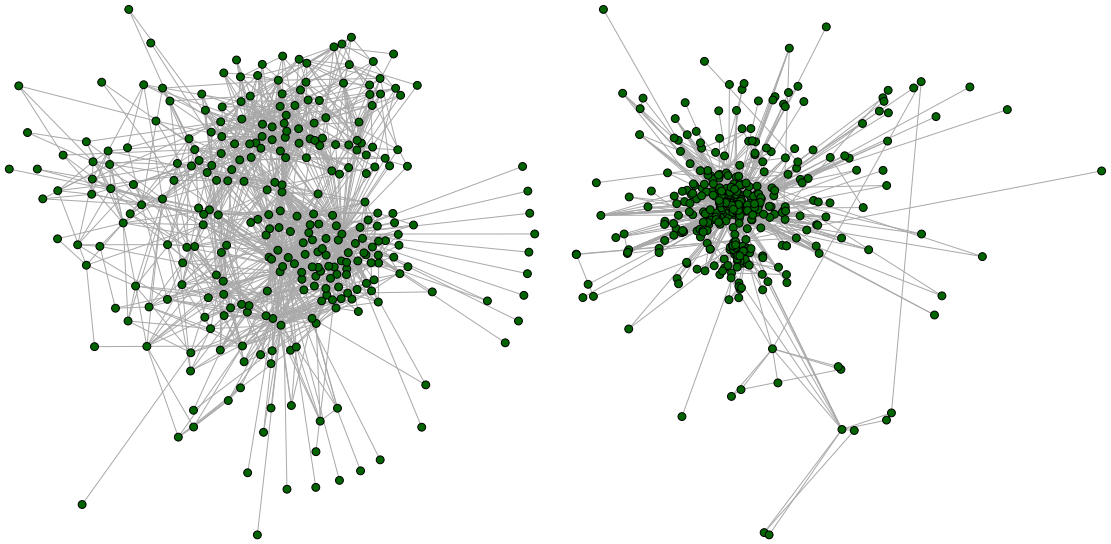

Figure 3: Networks visualization for the C.elegans (left) and US airport (right).

In Figure 4 we report the strength distributions for the two networks, with the strength (s_i) as the sum of the weights of the links incident on i. The two networks are very different, especially in the distribution of low and high values of strength. The weight profiles in Figure 5 confirm such differences, mostly caused by a difference of scale in the values.

According to Definition (4.1), $\mathcal{T}_2^{(i)}$ and $\mathcal{T}_3^{(i)}$, i.e. the pseudo-triangles for every nodes in a network, can be computed considering $\alpha = 0$ and $\beta = 0$. Since in this case a generic function F_i is always satisfied we obtain every pseudo-triangles in $\mathcal{T}_2^{(i)}$ and $\mathcal{T}_3^{(i)}$. Concerning the sets $\mathcal{T}_1^{(i)}$, such triangles can be easily computed by a built-in function in *igraph*.

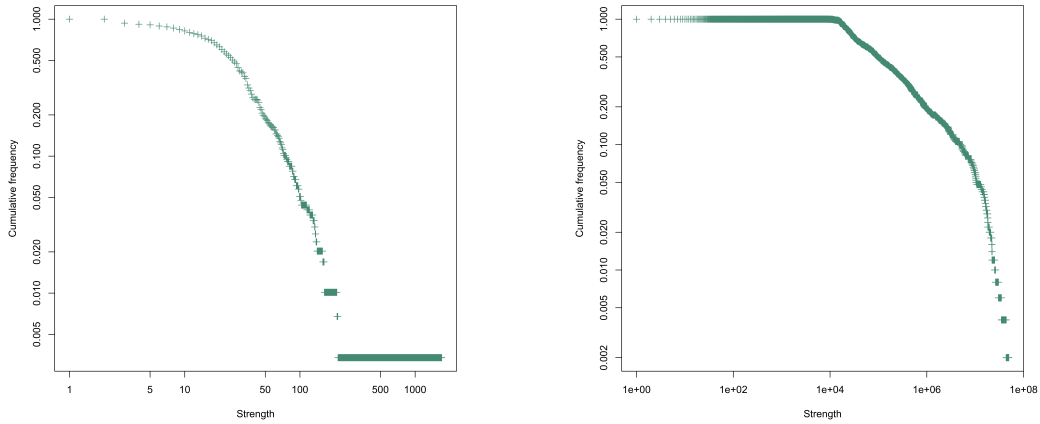

Figure 4: Strength distributions for the C.elegans (left) and US airport (right) networks.

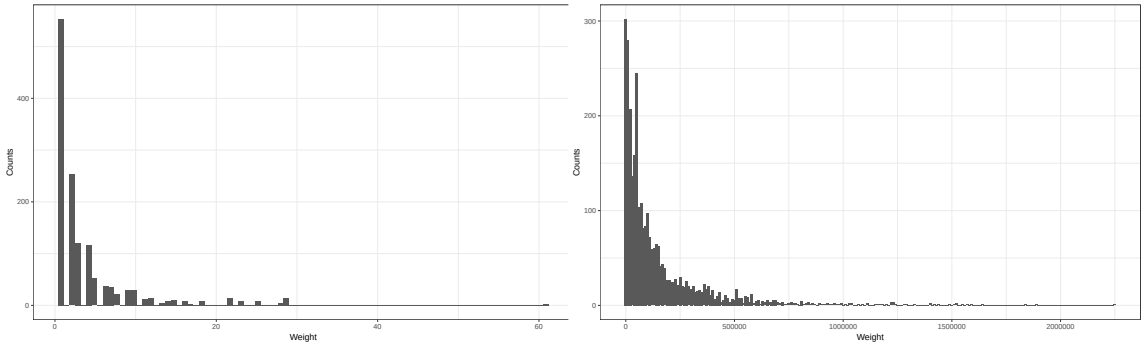

Figure 5: Weights histogram for C.elegans (left) and US airport (right) networks.

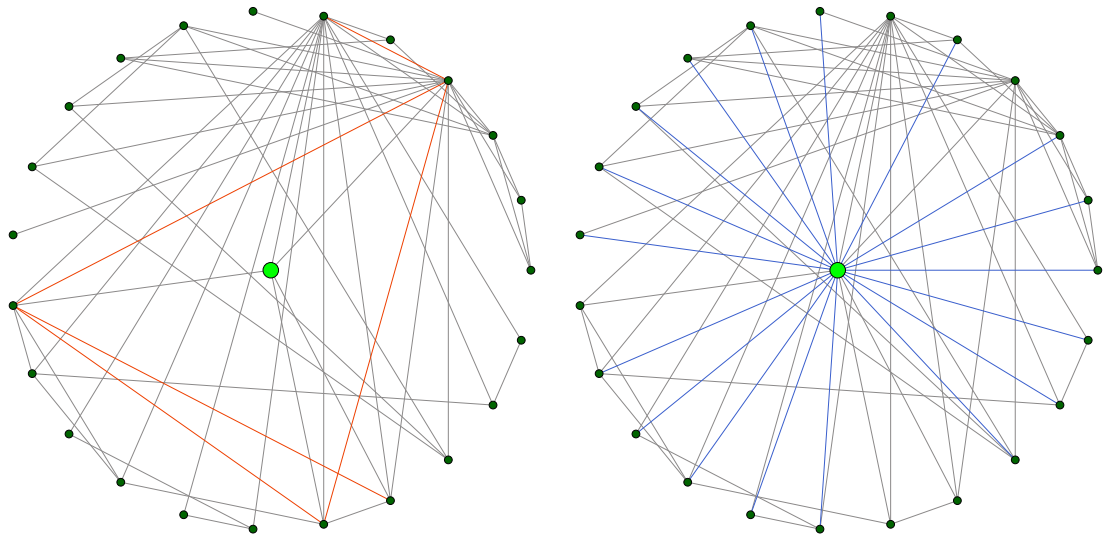

Figure 6: 2^{nd} order neighborhood of node $n = 488$ of US Airport network and triangles $\mathcal{T}_2^{(488)}$ (left) and $\mathcal{T}_3^{(488)}$ (right).

As example, in Figure 6 we show the arcs composing the pseudo-triangles in $\mathcal{T}_2^{(488)}$ and in $\mathcal{T}_3^{(488)}$ for the neighborhood of order 2 of node $n = 488$ in the US airport network. Such node has a degree $d_{488} = 5$, a second order neighborhood of cardinality 18 and a local clustering coefficient $C_{488} = 0.5$, since it close 5 triangles up to the theoretical 10. Thus, $|\mathcal{T}_2^{(488)}| = 5$ while triangles in $\mathcal{T}_3^{(488)}$ are computed obtaining $|\mathcal{T}_3^{(488)}| = 22$. Note that the blue arcs in the right panel of Figure 6 are $18(< 22)$ because some arcs can be mentioned twice in the set, since arc (i, k) can derive from $i \to j \to k$ as well as from $i \to l \to k$.

The generalized clustering coefficieng has value $C_{488}^{(g)} = 0.0632$, much lower than C_{488} since the proportion of closed triangles and pseudo-triangles when $\alpha = 0$ and $\beta = 0$ is smaller than the original network.

Figures 7 for C.elegans network and 9 for US airports network report three curves, respectively and for each node, the total number of triangles $|\mathcal{T}_1^{(i)}|$, the number of potential pseudo-triangles of type $|\mathcal{T}_2^{(i)}|$ and the number of potential pseudo-triangles of type $|\mathcal{T}_3^{(i)}|$. Figures 8 and 10 compare the degree d_i and the local clustering coefficient C_i for each node i. Note that in these benchmark instances nodes in C.elegans network are enumerated without a particular rule, while the nodes in US airport network are enumerated in non-increasing order of their degree.

The nodes in C.elegans network have a relatively small values of local clustering coefficient for nodes with high degree while nodes with small degree have in general, higher values of local clustering coefficient. This means that small degree nodes tends to form dense local neighborhoods, while the neighborhood of hubs is much sparser. Such observations motivate the limited number of pseudo-triangles in $T2$ because, for each node i, they are in number of $\binom{d_i}{2} - |\mathcal{T}_1^{(i)}|$, thus to dense neighborhood corresponds a small number of possible pseudo-triangles.

Note in Figure 7 that node $i = 295$ has a peak because $|\mathcal{T}_2^{(i)}| = 8658$ when the thresholds α and β are null (this is the case of potential pseudo-triangles). This is motivated by its particular neighborhood composed of a limited number of triangles in which it is embedded ($|\mathcal{T}_1^{(295)}| = 253$ and $C_{295} = 0.028$) despite its degree ($d_{295} = 134$). Choosing two edges on 134 lead to 8911 potential pseudo triangle of type $T2$ and when subtract 253

it results 8658. Such a remarkable presence of triangles of type $T2$ for a single node for the case of $\alpha = \beta = 0$ suggests that the C.elegans network is star-shaped.

Similar arguments can be considering for $T3$, indeed, we have a small number of potential pseudo-triangles for both small degree nodes and hubs since low values of degree allow for a small number of transitive closure.

The analysis of Figures 9 and 10 depict a very different picture for the US airport network. The nodes with indices until $i \simeq 100$ have values of degree and clustering coefficient which allow for a large number of pseudo-triangles $T2$ and a significant number of pseudo-triangles $T3$. Then, when the degree decreases and the local clustering coefficient increases the local neighborhoods preclude the formation of pseudo-triangles.

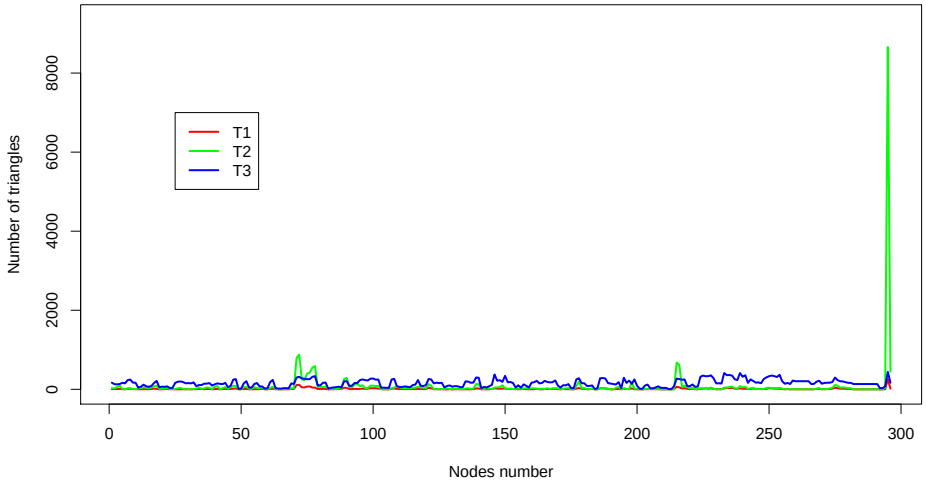

Figure 7: C.elegans. Comparison between the number of triangles $T1$, the number of potential pseudo-triangles $T2$ and the number of potential pseudo-triangles $T3$.

Figure 11 shows the global values of the generalized clustering coefficient $C_i^{(g)}$ for the C.elegans network when considering the four different functions F_1, F_2, F_3 and F_4. In each figure, the values are presented for every combination of α and β while the horizontal axis report the values of $C_i^{(g)}$ as averaged over every nodes in the network. For the simulations, the value of α and β are $\alpha, \beta = \{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70\}$ for the C.elegans network and $\alpha, \beta = \{0, 250000, 500000, 750000, 1000000, 1250000, 1500000,$

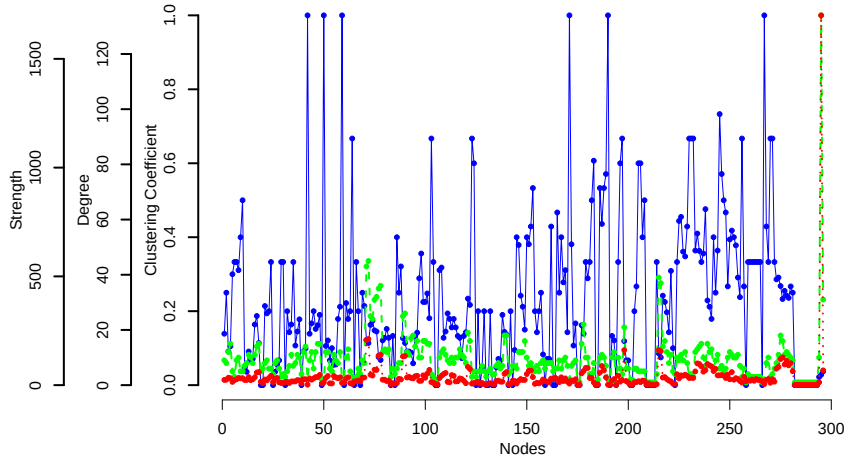

Figure 8: C.elegans. Comparison between local clustering coefficient (blue points), degree (red points) and strength (green points).

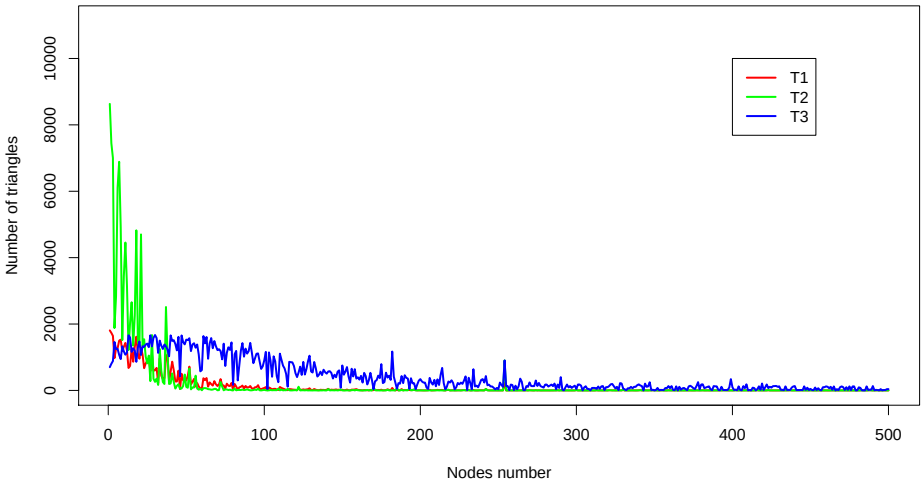

Figure 9: US airport. Comparison between the number of triangles $T1$, the number of potential pseudo-triangles $T2$ and the number of potential pseudo-triangles $T3$.

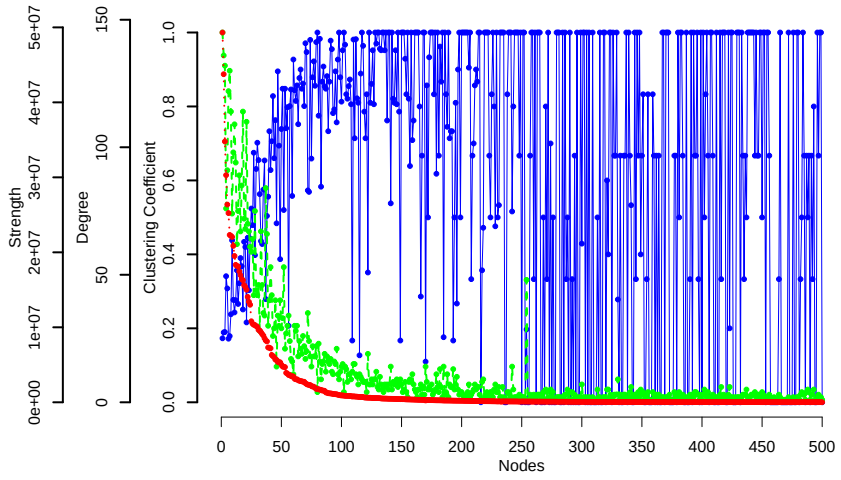

Figure 10: US airport. Comparison between local clustering coefficient (blue points), degree (red points) and strength (green points).

1750000, 2000000, 2225000} for the US airport network. Max values are chosen considering that the function F_1 could possibly be true also when considering arcs with the higher weights and the step is set as to have 10 runs for each coefficient. Thus, we performed 100 computations for every network.

As expected, higher values of $C_i^{(g)}$ are obtained for lower values of α and β and, globally, we have a non increasing trend with a higher slope for functions F_2 and F_3 since the average function smooths the values and as well as for the min function, they are true only for small values of weights. Regarding F_1 and F_4, they are more prone to be true for higher values of arcs weight and the slope declines slower.

A common behavior for all four cases is that the magnitude of $C_i^{(g)}$ is more dependent from pseudo-triangles $T2$ than those in $T3$. This is due to the tendency of high degree nodes to have also high strength. Therefore, the functions are more prone to be true for pseudo-triangles $T2$ than for pseudo-triangles in $T3$ since the adjacent links could possibly lie in a low degree node with a low value of strength.

Figure 12 reports the same plots for the US airport network and same comments on the general behavior can be repeated as for the C.Elegans network. The main differences

are in the steepest slope since the profile of weights distribution is more concentrated on lower values (see Figure 5).

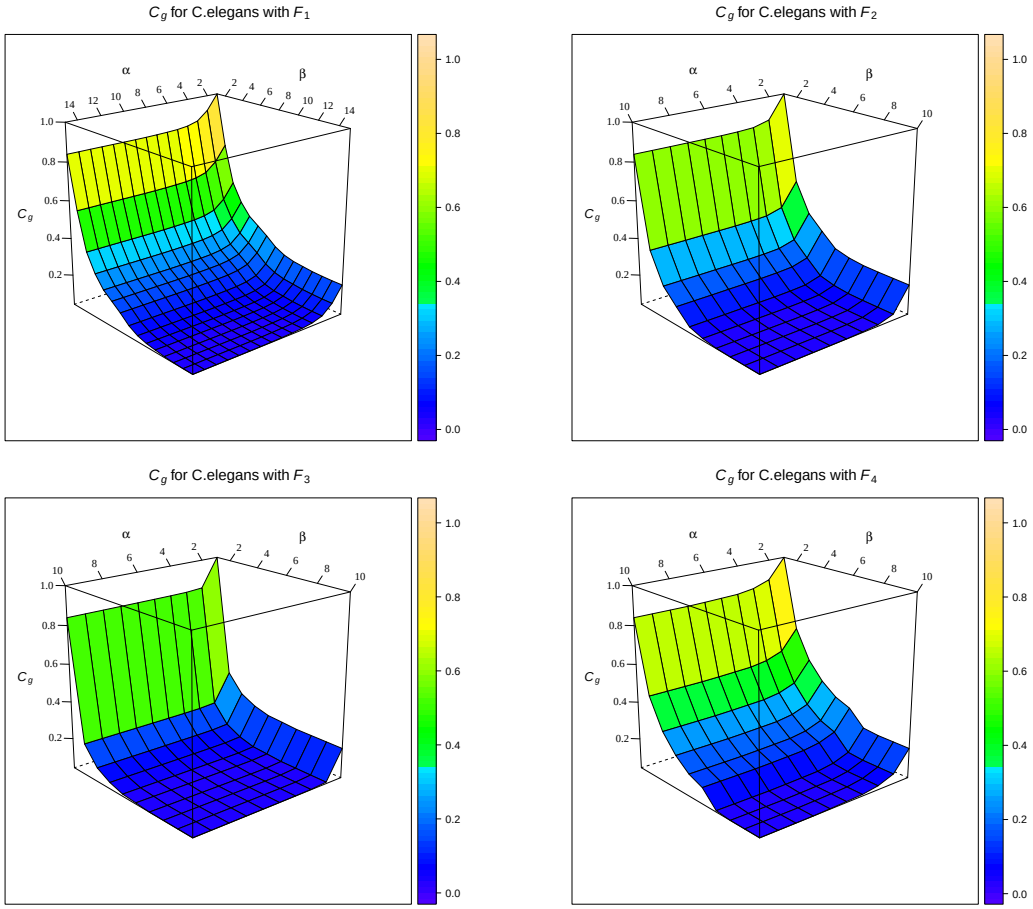

Figure 11: C.elegans. Average values of $C_i^{(g)}$ for cases F_1 (upper left), F_2 (upper right), F_3 (lower left) and F_4 (lower right).

In order to study the evolution of the generalized clustering coefficient $C_i^{(g)}$ when varying α and β, we provide a series of diagrams in which, for the two networks under examination, the density of the $C_i^{(g)}$ values are reported when considering fixed values of $\alpha = 0$ or $\beta = 0$ and varying the other coefficient.

In particular, for the C.Elegans network Figure 13 shows different density values for each α when $\beta = 0$ and Figure 14 for each β when $\alpha = 0$. Same setting is proposed in Figure 15 and Figure 16 for the US Airport network. In all the figures are also reported the density values of the local clustering coefficient C_i.

When $\beta = 0$ (see Figures 13 and 15) we can observe the contribution of pseudo-triangles

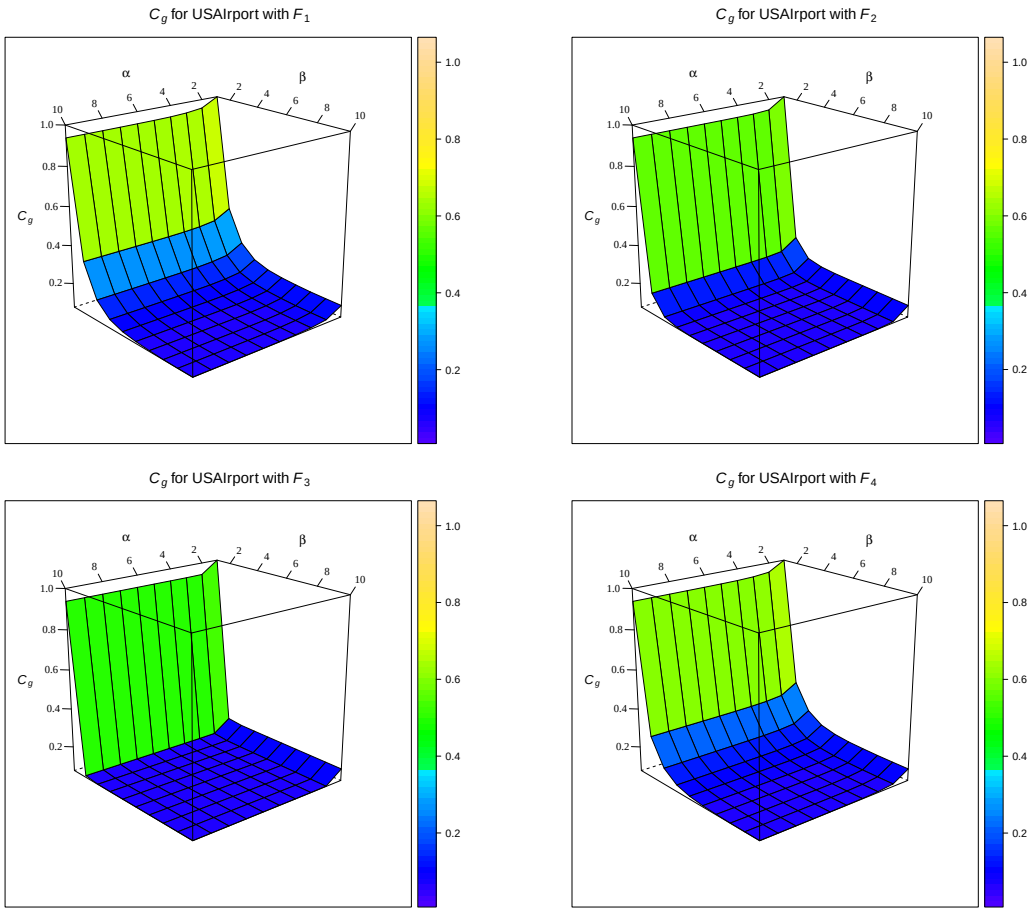

Figure 12: US Airport: Average values of $C_i^{(g)}$ for cases F_1 (upper left), F_2 (upper right), F_3 (lower left) and F_4 (lower right).

in $T2$ to $C_i^{(g)}$. For both network we observe the density of $C_i^{(g)}$ more concentrated around the max value 1 when $\alpha = 0$, while when α starts to growth the values shift to be close to 0. As in the previous figure, the effect is more evident for the C.elegans network.

For $\alpha = 0$, Figures 14 and 16 highlight that $C_i^{(g)}$ receives a small contribute from pseudo-triangles in $T3$ and the values lay around 0 as soon β grows.

The different figures confirm that for the two networks under observation, the main contribution to $C_i^{(g)}$ is provided by the pseudo-triangles in $T2$, i.e. their structures and weight profiles cause the networks to be more prone to close pseudo-triangles in $T2$ rather than pseudo-triangles in $T3$.

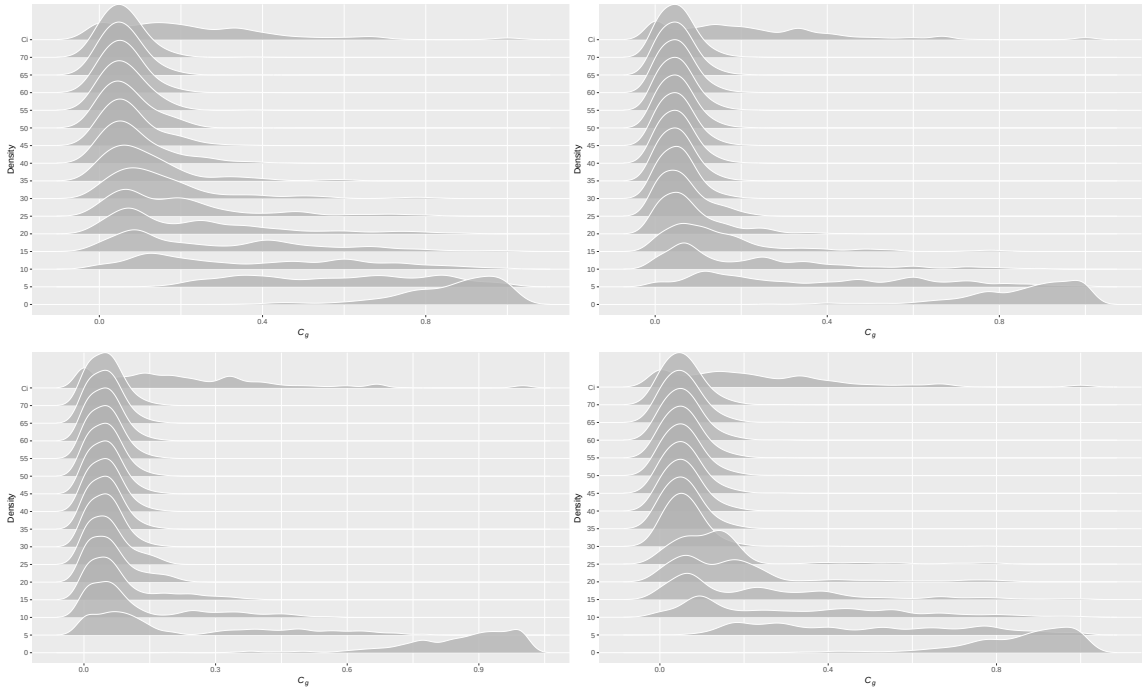

Figure 13: C.elegans. Density of $C_i^{(g)}$ for different values of α when $\beta = 0$ for cases F_1 (upper left), F_2 (upper right), F_3 (lower left) and F_4 (lower right).

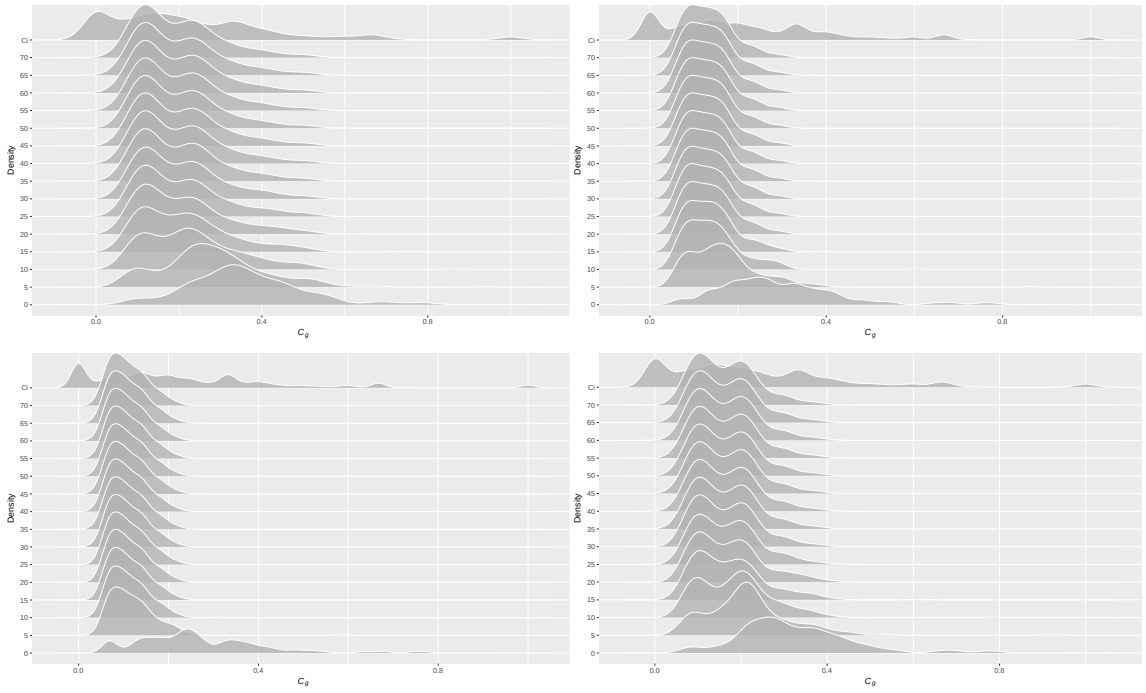

Figure 14: C.elegans. Density of $C_i^{(g)}$ for different values of β when $\alpha = 0$ for cases F_1 (upper left), F_2 (upper right), F_3 (lower left) and F_4 (lower right).

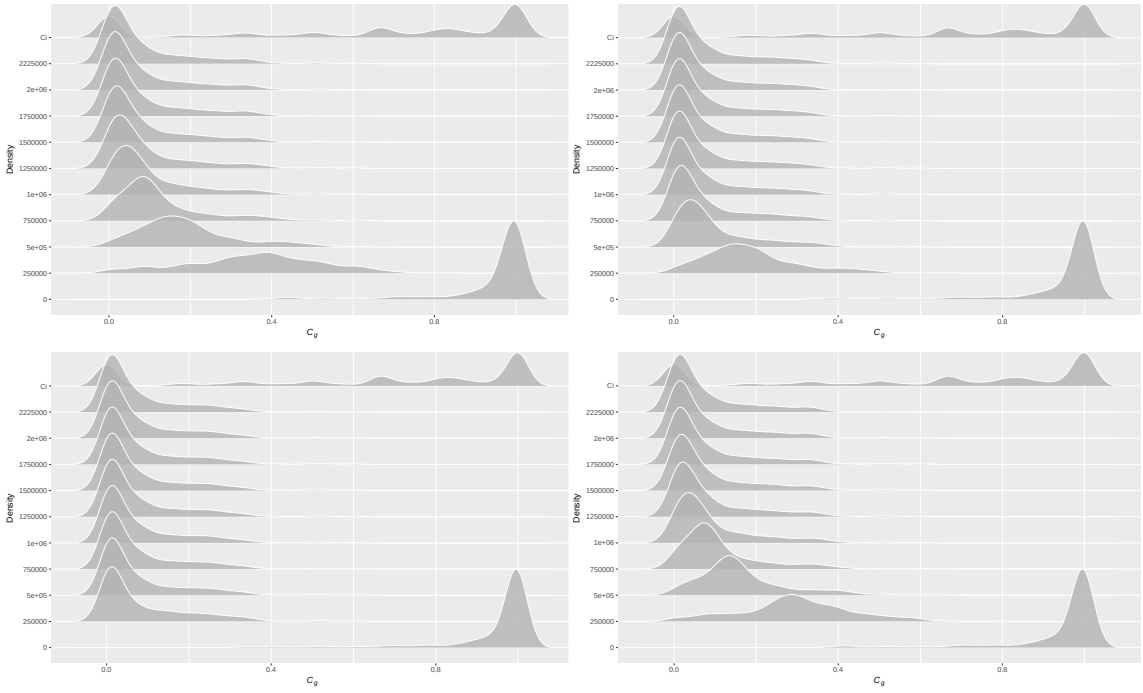

Figure 15: US airports. Density of C_g for different values of α when $\beta = 0$ for cases F_1 (upper left), F_2 (upper right), F_3 (lower left) and F_4 (lower right).

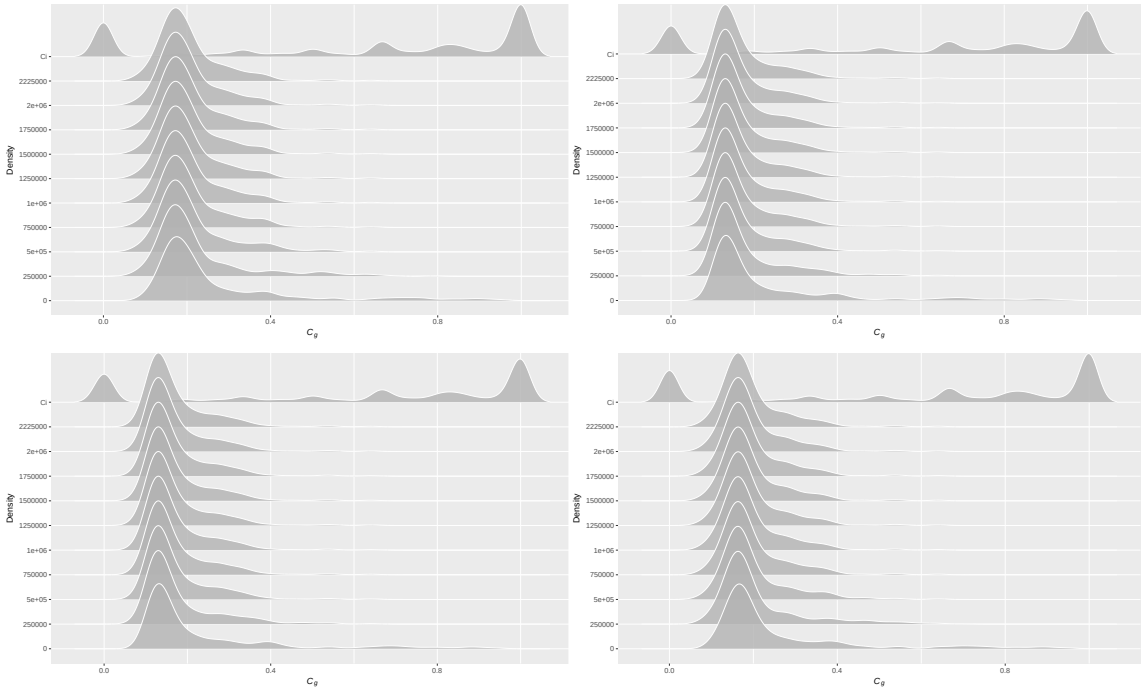

Figure 16: US airports. Density of C_g for different values of β when $\alpha = 0$ for cases F_1 (upper left), F_2 (upper right), F_3 (lower left) and F_4 (lower right).

6 Conclusions

This paper deals with a novel definition of the clustering coefficient for weighted networks when triangles are viewed under a social perspective, even when one of the sides is missing. The thresholds α and β provides a key information on the strength of the links able to induce missing arcs.

Triangles $T2$ and $T3$ serve for two scopes: by one side, they model the evidence that transitive relations among the nodes appear when the direct links are strong enough; by the other side, the knowledge of number and types of the triangles around the nodes when $\alpha = \beta = 0$ identify equivalent classes of networks on the basis of their topological structures.

Interestingly, our setting leaves some unsolved problems.

First, one can deal with the definition of more complex ways to connect indirectly nodes, beyond the mere triangles. In particular, the concept of social polygon with more than three sides can be introduced and explored. In so doing, a wider concept of community structure of the network can be effectively provided, with a novel definition of clustering coefficient where triangles are replaced by polygons. Of course, this generalization offers a high degree of complexity in implementing the empirical experiments.

Second, the analysis of the topological structure of the network can be discussed in more details. In this respect, notice that one can introduce a novel formulation of the concepts of hubs and centrality measures on the basis of the social connections among the nodes, according to our definition of induced indirect links. In this context, one is able to generalize the exploration to the case of α and β not necessarily null.

References

[1] Adamic L.A., and Adar E. (2003). Friends and neighbors on the web, *Social Networks*. vol. 25, no. 3, pp. 211-230.

[2] Albert R., and Barabási A.-L. (2002). Statistical mechanics of complex networks, *Rev. Mod. Phys.* 74, pp. 47-98.

[3] Arcagni A., Grassi R., Stefani S., and Torriero A.. (2017). Higher order assortativity in complex networks, *European Journal of Operation Research*. vol. 262, no 2, 16, pp 708-719.

[4] Barabási AL, *Network Science*, Cambridge University Press, UK, 2016.

[5] Barabási AL, Albert R (1999). Emergence of scaling in random networks, *Science*, vol. 286, pp. 509-512.

[6] Barrat A., Barthélemy M., Pastor-Satorass R., Vespignani A. (2004). Weighted Evolving Networks: Coupling Topology and Weight Dynamics, *Physical Review Letters*, vol. 92, no. 22, pp. 228701-4.

[7] Barrat A., Barthélemy M., Pastor-Satorass R., Vespignani A. (2004). The architecture of complex weighted networks, *PNAS*, vol. 101, no. 11, pp 3747-3752.

[8] Benati S., Puerto J., and Rodriguez-Chia AM. (2017). Clustering data that are graph connected, *European Journal of Operational Research*, vol. 261, no. 1, pp 43-53.

[9] Borgatti SP., (1997). Structural holes: unpacking Burt's redundancy measures, *Connections*, vol. 20, no. 1, pp 35-38.

[10] Colizza V., Pastor-Satorras R. and Vespignani A. (2007). Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, vol. 3, no. 4, pp. 276-282.

[11] Costa L.d., Rodrigues F.A., Travieso G., and Villas Boas P.R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, vol. 56, no. 1, pp. 167-242.

[12] Costantini G. and Perugini M., (2014). Generalization of clustering coefficients to signed correlation networks, *Plos One*, vol. 9 (2): e88669.

[13] Crama Y., and Leruth I., (2007). Control and voting power in corporate networks: Concepts and computational aspects, *European Journal of Operational Research*, vol. 178, no. 3, pp 879-893.

[14] Csardi G. and Nepusz T., (2006). The igraph software package for complex network research, *InterJournal Complex System*, vol. 1695, http://igraph.org.

[15] Ferraro G. and Iovanella A. (2016) Revealing correlations between structure and innovation attitude in inter-organizational innovation networks, *International Journal of Computational Economics and Econometrics*, vol. 6, n. 1, pp. 93113.

[16] Freeman L.C., (1977), A set of measures of centrality based on betweenness, *Sociometry*, vol. 40, no. 1, pp. 35-41.

[17] Gomez D., Figueira J.R., and Eusbio A. (2013). Modeling centrality measures in social network analysis using bi-criteria network flow optimization problems, *European Journal of Operational Research*, vol. 226, no. 2, pp 354-365.

[18] Grindrod P. (2002). Range-dependent random graphs and their application to modelling large small-world Proteome dataset, *Physical Review E*, 66, 066702.

[19] Humphries M.D. and Gurney K. (2008). Network "Small-World-Ness": A quantitative method for determining canonical network equivalence, *Plos One*, 3(4): e0002051.

[20] Holme P., Park S.M., Kim B.J and Edling C.R. (2007). Korean university life in a network perspective: Dynamics of a large affiliation network, *Physica A: Statistical Mechanics and Its Applications*, vol. 373, pp. 821-830.

[21] Latora V., Nicosia V., and Panzarasa P. (2013). Social cohesion, structural holes, and a tale of two measures, *J Stat Phys*, vol. 151, pp 745-764.

[22] Levine S.S., and Kurzban R. (2006). Explaining clustering in social networks: toward an evolutionary theory of cascading benefits, *Managerial and Decision Economics*, vol. 27, pp 173-187.

[23] Li Y., Luo P., Fan Z-p., Chen K., and Liu J. (2017). A utility-based link prediction method in social networks, *European Journal of Operational Research*, vol. 260, no. 2, pp 693-705.

[24] Liben-Nowell D., and Kleinberg J. (2007). The link-prediction problem for social networks, *Journal of the American Society for Information Science and Technology*. vol. 58, no. 7, pp. 1019-1031.

[25] Lindelauf R.H., Hamers H.J.M., and Husslage B.G.M. (2013). Cooperative game theoretic centrality analysis of terroristic networks: The case of Jemaah Islamiyah and Al Qaeda, *European Journal of Operational Research*, vol. 229, no. 1, pp 230-238.

[26] Lü L., and Zhou T., (2010). Link prediction in complex networks: A survey, *Physica A* 390, pp. 1150-1170.

[27] Newman M., (2003). The structure and function of complex networks., *SIAM Review* 45, pp. 167-256.

[28] Onnela J.-P., Chakraborti A., Kaski K., Kertsz J. and Kanto A., (2003). Dynamics of market correlations: Taxonomy and portfolio analysis, *Physical Review E* 68, 065103-4.

[29] Onnela J.-P., Saramaki J., Kertsz J., and Kaski K., (2005). Intensity and coherence of motifs in weighted complex networks, *Physical Review E* 71, 056110-12.

[30] Opsahl T., (2013). Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks* 35, doi: 10.1016/j.socnet.2011.07.001.

[31] Opsahl T., Panzarasa P., (2009). Clustering in weighted networks, *Social Networks* 31, pp. 155-163.

[32] Rosen K. H. (2012) Discrete Mathematics and its Application, McGraw-Hill Education.

[33] Varshney L. R., Chen B. L., Paniagua E., Hall D. H. and Chklovskii D. B. (2011). Structural properties of the Caenorhabditis elegans neuronal network. *PLoS Comput Biology*, 7(2), e1001066.

[34] Rota Bul S., Pellillo M. (2017). Dominant-set clustering: A review, *European Journal of Operational Research*, vol. 262, no. 1, pp 1-13.

[35] R Core Team (2014) R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*, Vienna, Austria, http://www.R-project.org.

[36] Scott J. (2000) Social Network Analysis: A Handbook, Sage Publications, London, UK.

[37] Soffer S.N. and Vázquez A. (2005). Network clustering coefficient without degree-correlation biases, *Physical Review E*, 71, 057101.

[38] Wang X.F. and Chen G., (2003). Complex networks: small-world, scale-free and beyond, *Circuits and Systems Magazine, IEEE Circuit and System Magazine*, 3 (1), pp. 6-20.

[39] Wasserman S. and Faust K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press, New York, NY.

[40] Watts D.J. and Strogatz S.H. (1998). Collective dynamics of small world networks, *Nature*, 393, pp. 440-442.

[41] White J., Southgate E., Thomson J. and Brenner S., (1986) The structure of the nervous system of the nematode Caenorhabditis elegans, *Philosophical Transactions of the Royal Society of London B, Biological Sciences*, 314, pp. 1-340.

[42] Wu J., Zhang G., and Ren Y. (2017). A balanced modularity maximization link prediction model in social networks, *Information Processing & Management*, vol. 53, no. 1, pp 295-307.

[43] Zhang B. and Horvath S. (2005). A general framework for weighted gene co-expression network analysis, *Statistical Applications in Genetics and Molecular Biology*, vol. 4, issue 1.

[44] Zhang P., Wang J., Xiaojia L., Menghui L., Di Z., and Fan Y. (2008). Clustering coefficient and community structure of bipartite networks, *Physica A.*, vol. 387, pp. 6869-6875