Journal of Algorithms and Computation

journal homepage: http://jac.ut.ac.ir

Relative Clustering Coefficient

Elena Farahbakhsh Toulii*[1] and Oscar Lindberg†[2]

[1,2]Department of Mathematics, Stockholm University

ABSTRACT

In this paper, we relatively extend the definition of the global clustering coefficient to another clustering, which we call it *relative clustering coefficient*. The idea of this definition is to ignore the edges in the network that the probability of having an edge is 0. Here, we also consider a model as an example that using the relative clustering coefficient is better than the global clustering coefficient for comparing networks and also checking the properties of the networks.

Keyword: Global clustering coefficient, local clustering coefficient, relative clustering coefficient, networks, graphs.

ARTICLE INFO

Article history:
Research paper
Received 01, April 2022
Received in revised form 11, May 2022
Accepted 28 May 2022
Available online 01, June 2022

AMS subject Classification: 05C78.

1 Introduction

Recently in the field of physics and statistics, networks and graphs are two interesting topics for example Internet, biological networks, email, media, and social network, citation network, and so on [10, 9, 5, 7, 6, 8]. One of the properties of graphs is clustering and one of the most important characteristics of networks is that they are highly clustered. It is easy to see that the probability that a person in Germany and a person in Iran make a friendship is so low, but the probability that in a small city in Iran two friends of a person become friends is so high. And this is one of the important characteristics of networks in

*Corresponding author: E. Farahbakhsh Touli. Email: elena.touli@math.su.se
†oscar.lindberg97@gmail.com

the real life. In the topological view of the graph, a highly clustered network contains a lot of triangles or cycles of length three. [3]

Networks or graphs contain a set of vertices and a relation between the vertices [3, 1, 2]. The relations between two vertices are defined by edges between the vertices and an edge between two vertices is shown by a line s.t. connects the two vertices.

If two vertices have a relation, we add an edge between them otherwise, we do not add any edge between them. If the relationship is one-sided we have a directed graph otherwise, if the relationship is two-sided we have undiracted graph. For example, friendship on Facebook is a two-sided relation, therefore if someone sends a request on Facebook to us we both become friends, and we can see what our friends share with us. But, friendship on Instagram is one-sided. When someone sends a friendship request to us, until we do not follow him or her, they will not be our friends. Edges in directed graphs are shown by using a line with an arrow indicating in which direction we have the relationship [3, 1, 2]. Examples of directed and undirected graphs are shown in Figure1.

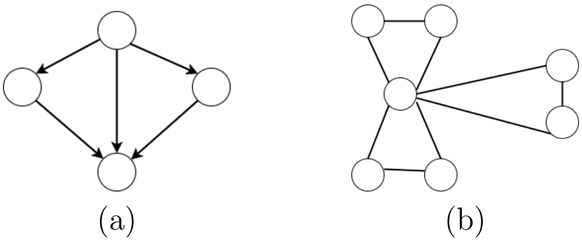


Figure 1: (a) A directed graph. (b) An undirected graph

The local clustering coefficient for v is the number of edges that exist in the neighborhood, divided by $k_v(k_v - 1)/2$, i.e. the proportion of friendships that exist. However, the global clustering coefficient is calculated by the number of triangles divided by the number of triples that could make a triangle. But, after looking and thinking about some networks we see that these clustering coefficients are not sufficient for comparing all the networks and we need to establish and extend the clustering coefficients that we already have read about. And this was the idea of writing this paper.

New work. In this paper, we extend the definition of clustering coefficient to a clustering that can indicate characteristics of networks better and we call it Relative Clustering Coefficient. At this definition, we consider only the edges in the network that we are allowed to add to the network. For example, if two people are in two different prisons they are not connected, although they may have the same lawyer. And if two people are in two different hospitals they are not connected, even though they may have the same doctor who works in those two hospitals.

In this paper, we consider just simple undirected graphs. We will use and talk about some properties of graphs that we illustrate more here. They are bipartite graphs and cliques.

Bipartite Graphs [2]:In graph theory, a bipartite graph is a graph that we can divide vertices into two groups of vertices, such that there is no edge between the vertices in each group. (For more illustration look at Figure 2 part (a))

Cliques [2]: A complete graph is a simple graph (undirected graph without loops nor

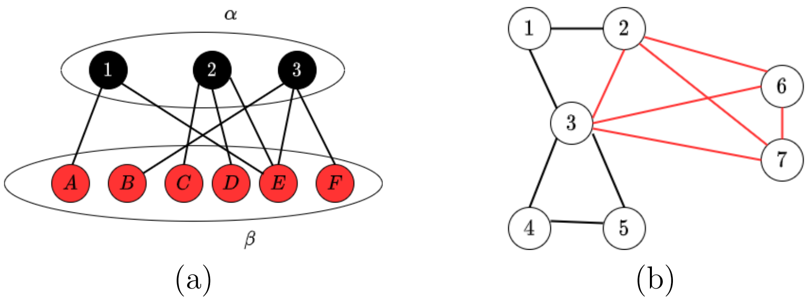


Figure 2: (a) A bipartite graph with two parts α and β. (b) A clique with vertices $2, 3, 6, 7$ and red edges in the graph.

multiple edges) that includes all the possible edges between vertices or in other words, there is an edge between any two vertices of the graph. A subset of a graph is a graph that its vertices are the subset of the vertices of the main graph and edges are the subset of the edges of the main graph that connects vertices of the sub-graph. A clique is a sub-graph of a graph that is complete. (For more illustration look at Figure 2 part (b))
Here the outline of this paper is as follows: in Section 2 we present the definition of clustering coefficients that we already had; global clustering coefficient and local clustering coefficient. We extend the definition of clustering coefficient at Section 3. We also present an example of a model and we use relative clustering coefficient instead of clustering coefficient. Last section which is Section 4 is conclusion.

2 Clustering Coefficient

If we consider different networks in the real life we see that most of them are highly clustered, i.e. we can see that a friend of a friend of a person is a friend of the person as well. In another word, two friends of a person are with high probability friends in a way. From the topological view, we can see that there are lots of triangles in a network [7, 11]. There are two definitions to measure the clustering in the network; local clustering coefficient and global clustering coefficient. Here we want to illustrate these two definitions.

2.1 Local Clustering Coefficient

As it is defined in [11], the local clustering coefficient for a vertex v in a graph G is defined as the number of triangles in the graph such that one vertex of the triangle is v divided by the number of paths π in G with the length of 2 such that the vertex v is the middle vertex of π.
In other words, for an undirected graph, if we consider ν_i as a set of vertices in the neighborhood of a vertex v_i, that means the set of vertices that there is an edge between v_i and every vertex in the set, so we have that

$$\nu_i = \{v_i : e_{ij} \in E\}.$$

Therefore, for a vertex v_i we can define the local clustering coefficient C_i as the number of edges between vertices in the set of ν_i divided by the number of edges that can exist between the vertices in the set ν_i.
Therefore, we can define the local clustering coefficient for a vertex v_i as follows:

Definition 1. *[7] Local Clustering Coefficient*
For a vertex v_i, if ν_i is the set of vertices in the neighbourhood of v_i and $|\nu_i|$ is the size of this set, the local clustering C_i for the vertex v_i is defined as follows:

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in \nu_i \ \& \ e_{jk} \in E\}|}{\binom{|\nu_i|}{2}}$$

We can define the local clustering coefficient for a directed graph in a similar way, but it is beyond the scope of this paper.

2.2 Global Clustering Coefficient

If we consider $N_\triangle$ as the number of triangles in the graph and N_3 as the number of sub-graphs containing three vertices that are connected at least by two edges (It means there is two edges between them or three edges, look at the Figure 3), the definition of clustering in a graph or network is as follows:

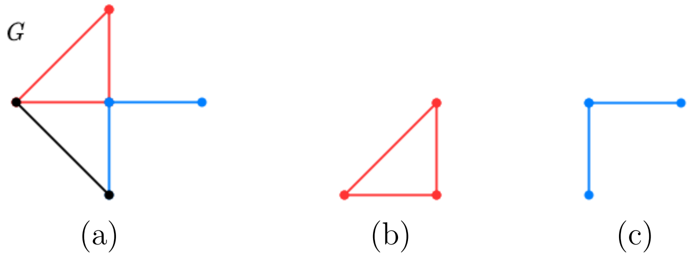


Figure 3: (a) Graph G and two of its sub-graphs with three vertices. (b) The red sub-graph is connected by three edges. (c) The blue sub-graph is connected by two edges.

Definition 2. *[7] Global Clustering Coefficient*
By using the notations $N_\triangle$ and N_3 that we defined earlier, the global Clustering Coefficient is defined as follows:

$$C = \frac{3N_\triangle}{N_3}$$

After looking and considering other examples of networks we see that these definitions are not sufficient for considering high clustering in the network, so we define relative clustering coefficient. We consider the next section to illustrate and talk about the relative clustering coefficient. Later we give an example that we see that considering the global clustering coefficient is not good enough.

3 Relative Clustering Coefficient (RCC)

The idea of the definition of RCC is as follows that we just consider pairs of vertices that we can have an edge between them in the network. In other words, the probability of having an edge between them is larger than 0. Here, for each pair of vertices, we define a capacity (a number, 0 or 1). If we can have an edge between two vertices (Or if the probability of having an edge between two vertices is larger than 0) the capacity is 1, otherwise, the capacity is 0. $N^{\triangle_1}$ is the number of all the triangles in the network that all the edges in the triangles have a capacity of 1, and $N_3^{\triangle_1}$ is the number of triangles that the capacity of all the edges in it is 1, such that all the edges in the triangle are in the network, and $N_2^{\triangle_1}$ is the number of triangles that the capacity of all edges in it is 1, but just two of the edges in the triangle are in the network. Now, we define the relative clustering coefficient in a network as follows:

Definition 3. *Relative Clustering Coefficient*
By using the notation that we have illustrated earlier, we defined RCC as follows:

$$C_R = \frac{3N_3^{\triangle_1}}{3N_3^{\triangle_1} + N_2^{\triangle_1}}.$$

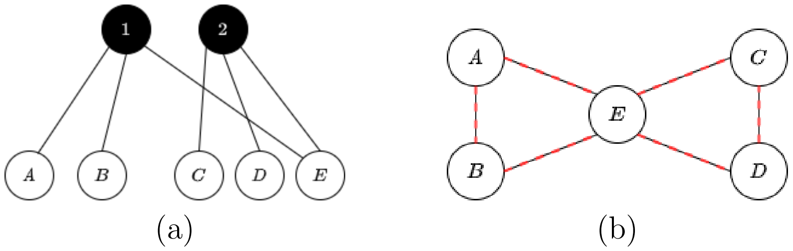
Example 1. *If we consider a group of people (A and B) in hospital number 1 and two other people (C and D) that are hospitalized in hospital number 2, and the person E a doctor that works with patients in both hospitals (For more information look at Figure 4) we have RCC and CC as follows:*

$$C_R = 1,$$

while

$$C < 1.$$

But, in this case, we cannot add any edge to this network, because these two groups of people are separated and there is no physical contact between them. Therefore, the clustering coefficient should be 1.



(a) (b)

Figure 4: (a) A bipartite graph indicating that person A and B are hospitalized in hospital number 1 and C and D are hospitalized in hospital number 2, and E is working in both hospitals. (b) Full graph of the model, that contains all the possible edges. Edges with the capacity of 1 are black. Dashed red lines indicate the edges in the network.

3.1 RCC instead of CC

In [7] M. E. J. Newman illustrated a model for highly clustered networks. The model is as follows:

Model 1. *[7]*
→ *We have N individuals in total*
→ *These individuals are divided into M different groups.*
→ *Individuals can belong to more than one group*
→ *Individuals belong to groups randomly*
→ *If two individuals belong to one group with the probability of p they are connected otherwise they are not connected.*

To illustrate this model better, we use an example:

Example 2. *We have some professors (A,B,C,D,E,F,G,H,I) who work at the University of Tehran. At the University of Tehran, we have some different departments (1, 2, 3, 4, 5) that each professor belongs to. Some professors work in different departments. Therefore, in meetings that are held in different departments that they participate in, they can meet other professors in the department. But, two professors who do not work in the same department, do not have any information from each other. (For more information look at Figure 5)*

Definition 4. *Full Graph*
If we construct a network that contains all the edges with the probability of $p > 0$, we have a graph that we call it full graph. (See the lower figure in Figure 5)

For this model, M. E. J. Newman used the clustering coefficient and showed that

$$C = pC'$$

such that C' is the clustering of the full graph of the network.

Reason for using RCC. Here, if we look at the full graph, we cannot add any edge to the network, so it is better if we use RCC instead of CC. Because it (the full graph) has all the edges between any pair of vertices and we are not allowed to add any other edges to the full graph. Therefore, by using RCC we have the following theorem which is more reasonable to use for this example and model.

Theorem 1. *For the Model 1 for large n, where n is the number of vertex in the network, we have that*

$$C_R = p,$$

which p is the probability of having an edge between two vertices in the network.

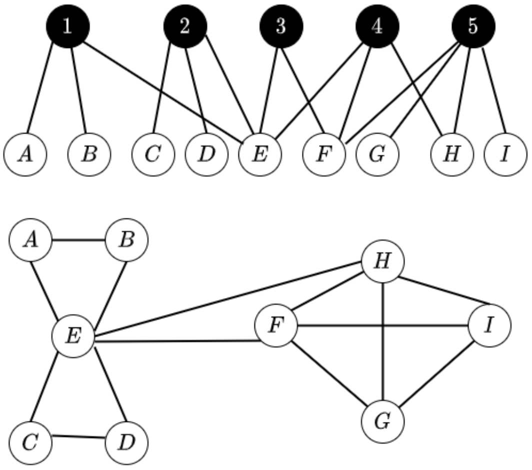


Figure 5: The upper figure is a bipartite graph that indicates which professor belongs to which departments. Numbers indicate the departments and letters indicate the name of the professors. The lower figure is the full graph using this model. That is for example professor E works for departments 1, 2, 3, and 4, therefore E can know each people who work at those departments. But, there is no edge connecting E to people who work at the department of 5 that do not work in any of 1 nor 2 nor 3 nor 4.

Proof. For each of the $N^{\triangle_1}$ cliques of size three (that also have capacity 1 for each edge), we define E_i as the number of edges within clique number $i = 1, 2, ..., N^{\triangle_1}$. The number of given one of this cliques in binomially distributed with three numbers of trials and probability of success p, therefore $E_i \sim bin(p, 3), i = 1, ..., N^{\triangle_1}$.
The probability of all three edges existing within one of these cliques is

$$P(E_i = 3) = p^3$$

while the probability of exactly two edges existing with one of these cliques is

$$P(E_i = 2) = 3p^2(1 - p).$$

We define indicator variables as follows:

$$I(E_i = 3) = \begin{cases} 1, & \text{if } E_i = 3 \\ 0, & \text{otherwise} \end{cases}$$

$$I(E_i = 2) = \begin{cases} 1, & \text{if } E_i = 2 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, these indicator variables will have Bernoulli distributions and expectations, so we have:

$I(E_i = 3) \sim be(p^3)$ has expectation $E[I(E_i = 3)] = p^3$
$I(E_i = 2) \sim be(3p^2(1 - p))$ has expectation $E[I(E_i = 2)] = 3p^2(1 - p)$.

Asymptotically we have that for large enough n

$$N_3^{\triangle_1} = \sum_{i=1}^{N^{\triangle_1}} I(E_i = 3) = \theta(N^{\triangle_1}.E[I(E_i = 3)]) = \theta(N^{\triangle_1}.p^3)$$
$$N_2^{\triangle_1} = \sum_{i=1}^{N^{\triangle_1}} I(E_i = 2) = \theta(N^{\triangle_1}.E[I(E_i = 2)]) = \theta(N^{\triangle_1}.3p^2(1-p))$$

where the notation θ is defined in Definition 5.
Therefore, for large enough n the following relative clustering coefficient will asymptotcally go towards

$$C_R = \frac{3N_3^{\triangle_1}}{3N_3^{\triangle_1} + N_2^{\triangle_1}} = \frac{3p^3 N^{\triangle_1}}{3p^3 N^{\triangle_1} + 3p^2(1-p)N^{\triangle_1}} = \frac{p}{p + (1-p)} = p.$$

□

We can use a similar definition for defining relative local clustering coefficient as well.

Definition 5. *[4] For two functions $f(n)$ and $g(n)$ we say that $f(n) = \theta g(n)$, if there exist two constant numbers c_1 and c_2 and an integer numbers n' such that for all $n > n'$, we can write:*

$$c_1 g(n) \leq f(n) \leq c_2 g(n).$$

3.2 Applications

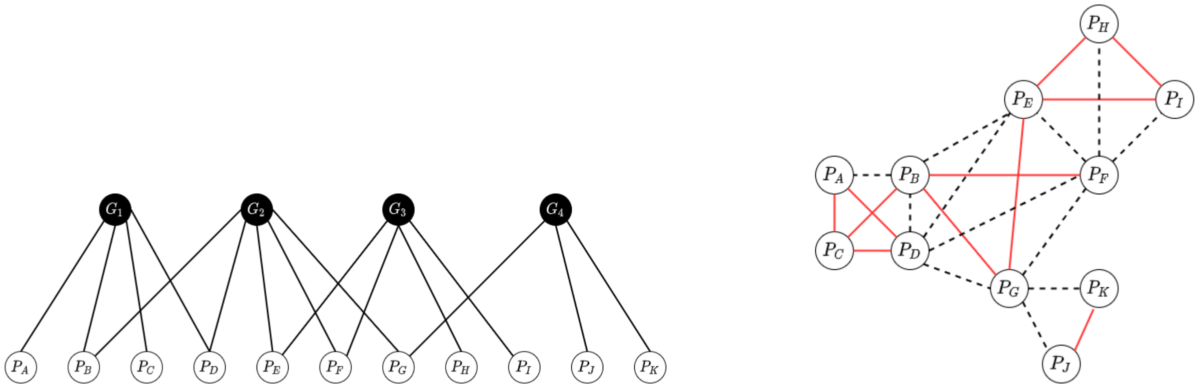
At this section, we consider an example from [7]. M. E. J. Newman considered 11 individuals that they belong to some groups randomly. Then, by using the bond percolation method they connect individuals to each other (For more illustration look at figure 6). After a simple calculation we reach that

$$C = \frac{3}{7} \quad \text{while} \quad C_R = \frac{3}{5},$$

which Relative Clustering coefficient is more natural to use for this example.

4 Conclusion

In this paper, we extend the definition of clustering coefficient to another definition which we called it relative clustering coefficient. This coefficient can measure the properties of networks better. In section 2 we defined two clustering coefficients, one local clustering, and another global clustering coefficient. We proposed the definition of relative clustering coefficient in section 3 and we used the clustering for measuring the property of a model as an example.



(a) (b)

Figure 6: (a) Individuals P_A to P_K are divided in 4 groups G_1 to G_4 randomly. (b) The dashed black lines indicates the full graph and red lines indicate the graph after bond percolation.

References

[1] Bollobás, B., Modern Graph Theory. Graduate Texts in Mathematics 184. Springer-Verlag New York, 1st edition, (1998).

[2] Bondy, J. A., and Murty, U. S. R., Graph Theory with Applications. Elsevier, New York, (1976).

[3] Caldarelli, G., Pastor-Satorras, R., and Vespignani, A., Structure of cycles and local ordering in complex networks. Eur. Phys. J. B, 38 (2004) 183–186.

[4] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C., Introduction to Algorithms. The MIT Press, 2nd edition, (2001).

[5] Dorogovtsev, S. N., and Mendes, J. F. F., Evolution of networks with aging of sites. Phys. Rev. E, 62 (2000) 1842–1845.

[6] Girvan, M., and Newman, M. E. J., Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99 (2002) 7821–7826.

[7] Newman, M. E. J., Properties of highly clustered networks. Phys. Rev. E, 68 (2003) 026121+.

[8] Newman, M. E. J., Strogatz, S. H., and Watts, D. J., Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E, 64 (2001) 026118+.

[9] Réka, A., and Albert-László, B., Statistical mechanics of complex networks. Reviews of Modern Physics, 74 (2002) 47–97.

[10] Strogatz, S. H., Exploring complex networks. Nature, 410 (2001) 268–276.

[11] Watts, D. J., and Strogatz, S. H., Collective dynamics of ‘small-world’ networks. Nature, 393 (1998) 440–442.