Network clustering coefficient without degree-correlation biases

Sara Nadiv Soffer[1] and Alexei Vázquez[2]
[1]*Department of Mathematics, Rutgers University Piscataway, New Jersey 08854, USA*
[2]*Department of Physics and Center for Complex Network Research, University of Notre Dame, Notre Dame, Indiana 46556, USA*
(Received 26 September 2004; published 13 May 2005)

The clustering coefficient quantifies how well connected are the neighbors of a vertex in a graph. In real networks it decreases with the vertex degree, which has been taken as a signature of the network hierarchical structure. Here we show that this signature of hierarchical structure is a consequence of degree-correlation biases in the clustering coefficient definition. We introduce a definition in which the degree-correlation biases are filtered out, and provide evidence that in real networks the clustering coefficient is constant or decays logarithmically with vertex degree.

DOI: 10.1103/PhysRevE.71.057101 PACS number(s): 89.75.Hc, 89.75.Fb, 89.20.Hh, 89.65.−s

The increasing availability of network data representing many real systems have motivated the development of statistical measures to characterize large networks [1–5]. These measures revealed that, as a difference with the classical Erdös-Rényi [6] random graph model, real networks are characterized by a power law distribution of vertex degrees [1,7,8], a high clustering coefficient or transitivity [1,9], and degree correlations between connected vertices [10–12]. Yet, it is important to characterize up to which extent the measures provide information about the studied networks. For instance, it has been shown that in some networks the degree correlations are a consequence of the existence of large degree vertices and, therefore, the sequence of vertex degrees is sufficient to characterize those networks [12–14].

In this work we study the influence of degree correlations on the clustering coefficient. We show that most of the observed variations of the clustering coefficient with the vertex degrees [15–18] are determined by the degree correlations among connected vertices. Based on this fact, we introduce a new definition of clustering coefficient, filtering out the effect of degree correlations. The similarities and differences between the two definitions are analyzed through the study of different real networks.

Consider undirected simple graphs on $i=1,\ldots,N$ vertices. Let k_i be the degree of a vertex and t_i the number of edges among its neighbors. The standard definition of local clustering coefficient is

$$c_i = \frac{t_i}{\binom{k_i}{2}}, \tag{1}$$

where $\binom{k_i}{2}$ is the number of pairs that can be made using k_i neighbors. Furthermore, to characterize the global clustering coefficient two different measures have been introduced. The first is just the average of c_i over all vertices with degree larger than one

$$\langle c\rangle = \frac{\sum_{i|k_i>1} c_i}{\sum_{i|k_i>1} 1}. \tag{2}$$

The second is obtained computing first the average of t_i and $\binom{k_i}{2}$ and then their ratio

$$C = \frac{\sum_i t_i}{\sum_i \binom{k_i}{2}}. \tag{3}$$

As noticed in Ref. [19], the two definitions of global clustering coefficient may give different values. Consider, for instance, a double star of N vertices (Fig. 1). In this case $\langle c\rangle \approx 1$ while $C=O(1/N)$, the two global clustering coefficients dramatically differing for $N\gg 1$. This discrepancy makes the comparison between analytical results obtained for different graph models and different definitions of global clustering coefficient difficult. At the local level of a single vertex the clustering coefficient (1) may also give counterintuitive results. For instance, the local clustering coefficient of the two central vertices of the double star is $c_1=c_2=O(1/N)$, approaching zero for $N\gg 1$. We cannot, however, increase the number of connections among the neighbors of vertex 1

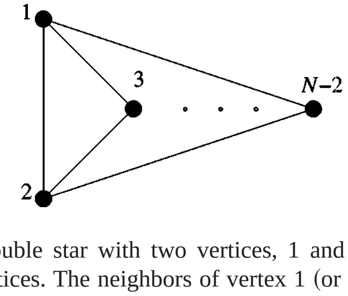

FIG. 1. Double star with two vertices, 1 and 2, connected to $N-2$ other vertices. The neighbors of vertex 1 (or 2) are connected as most as their degrees allow. Yet, with the usual definition of clustering coefficient we obtain $c_1=O(1/N)$, approaching zero in the limit $N\gg 1$.

without increasing the degree of its neighbors. In this sense, the neighbors of vertex 1 are as clustered as they can be.

This example shows that the local clustering coefficient of a large degree vertex connected to vertices with much smaller degrees will be always small, no matter how its neighbors are interconnected. We would like instead a measure of clustering coefficient that allows us to quantify the connectivity among the neighbors of a vertex, independently of its degree and the degree of its neighbors. The clustering coefficient is a three vertex correlation measure and, as it is the general case in statistics, to define a three point correlation measure we should filter out two point correlations, represented here by the degree correlations between connected vertices. We tackle this problem defining the clustering coefficient relative to the maximum possible number of edges between the neighbors of a vertex, given their degree sequence. Let ω_i be the maximum number of edges that can be drawn among the k_i neighbors of a vertex i, given the degree sequence of its neighbors. A neighbor j can have at most $\min(k_i-1,k_j-1)$ edges with the other neighbors, therefore

$$\omega_i \leqslant \Omega_i = \left\lfloor \frac{1}{2} \sum_{\text{neighbors}} [\min(k_i,k_j) - 1] \right\rfloor \leqslant \binom{k_i}{2}. \tag{4}$$

While $\binom{k_i}{2}$ takes into account only the degree of the vertex, Ω_i considers that occasionally, not all the k_i-1 excess edges are available at the neighbors of i. ω_i considers, in addition, the possibility of the excess edges to actually form triangles. ω_i can be computed using the following algorithm: (1) Starting from the neighbor's degree sequence $\{k_1,\ldots,k_n\}(n=k_i)$, construct the list $\{\min(k_i,k_1)-1,\ldots,\min(k_i,k_n)-1\}$, arranged in a decreasing order. (2) Draw an edge from the first element to as many as possible other elements in the list, always going from largest to smaller. Each time an edge is drawn, one is subtracted from the remaining degree of the connected vertices. (3) Remove the first element and any zero from the list and sort the list in decreasing order. (4) Repeat the process and stop when the list is empty. The number of maximum possible connections ω_i is the total number of edges drawn (see Fig. 2).

A proper definition of local clustering coefficient, removing the effects of degree correlations, is

$$\tilde{c}_i = \frac{t_i}{\omega_i} \tag{5}$$

and the two different measures of global clustering coefficient read

$$\langle \tilde{c} \rangle = \frac{\sum_{i|\omega_i>0} c_i}{\sum_{i|\omega_i>0} 1}, \quad \tilde{C} = \frac{\sum_i t_i}{\sum_i \omega_i}. \tag{6}$$

Some general properties of our definition of clustering coefficient are the following. (i) If all the neighbors of a vertex has degree one (star) then its clustering coefficient is undefined. Indeed, the concept of clustering is meaningless for the central vertex of a star, as it is meaningless for degree one vertices. (ii) $\tilde{c}_i \geqslant c_i$, as follows from (4). Therefore, when the clustering is one by the usual definition it is one by our definition. Notice that the opposite is not necessarily true [see Fig. 2(c)]. (iii) When all the k_i neighbors of a vertex i have degrees larger than or equal to the degree of the vertex itself (a regular graph, for instance) $\tilde{c}_i = c_i$.

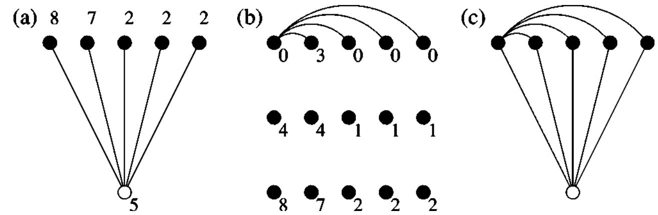

FIG. 2. Algorithm to compute ω_i. (a) A vertex i (open circle) is connected to five neighbors (filled circles) with degree sequence {8,7,2,2,2}. (b) Since each neighbor can be connected at most with four other neighbors, we replace the neighbors degree sequence (lowest row) by {4,4,1,1,1} (middle row). It is easy to see that after connecting the first neighbor to all others, we get four triangles and three extra edges that cannot be used anymore (upper row). Summarizing, for this example, $\omega_i=4$, $\Omega_i=5$ and $\binom{5}{2}=10$. (c) Subgraph with maximum number of edges among the neighbors, with $c_i=0.4$ and $\tilde{c}_i=1$.

The example in Fig. 2 shows how the usual definition underestimates the clustering around a given vertex i. In this case, while $c_i=0.4$, the number of edges between neighbors is as large as it can be given for their degree sequence, as it is correctly quantified using our definition ($\tilde{c}_i=1$). In the following we compare the usual and our clustering coefficient definitions using the graph representation of four real systems. The degree of correlations present on these graphs is quantified by the assortativity coefficient r [11], taking values between −1 (highly disassortative) to 1 (highly assortative). The systems considered are, in increasing order of assortativity, (1) the autonomous system representation of the Internet, as for April 2001 [21], (2) the protein-protein interaction network of the yeast *Saccharomyces cerevisiae* [22], (3) the semantic web of English synonyms [17], and (4) the co-authorship network of mathematical publications between 1991 and 1999 [23]. In Table I we show the two global clustering coefficients as computed with the usual and our definitions. For the two disassortative graphs ($r<0$), there is an order of magnitude difference between the global cluster-

TABLE I. Average clustering coefficient as computed with the usual and our definitions. The graphs are listed in increasing order of their degree of assortativity, quantified by the degree-correlation coefficient r [11], taking values from −1 (fully disassortative) to 1 (fully assortative).

Network	r	$\langle c \rangle$	C	$\langle \tilde{c} \rangle$	$\tilde{C}$
Internet	−0.19	0.45	0.0090	0.49	0.45
Protein interaction	−0.13	0.12	0.055	0.16	0.19
Semantic	0.085	0.75	0.31	0.83	0.59
Co-authorship	0.67	0.65	0.56	0.78	0.85

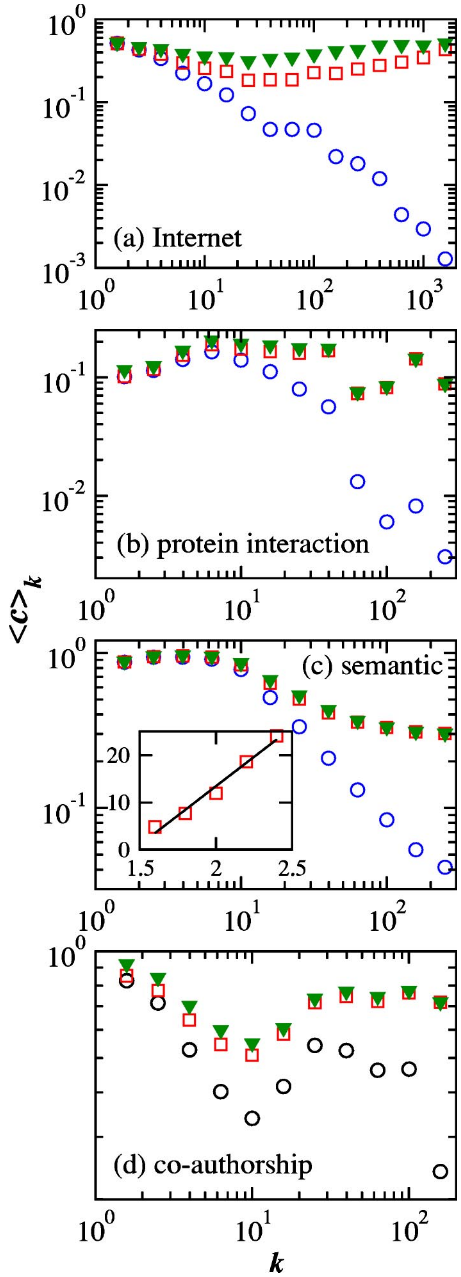

FIG. 3. Average clustering as a function of the vertex degree, as computed using the usual definition (circles), our definition approximating ω_i by Ω_i (squares), and our definition using ω_i (triangles). The graphs are shown in increasing order of their assortativity, with the most disassortative graph on the top, and the more assortative graph on the bottom.

ing coefficients $\langle c \rangle$ and C computed with the usual definition. With our definition, however, both global measures of clustering coefficient (6) give values of the same order, independently of the degree correlations.

Another characteristic feature of the usual definition of clustering coefficient is that, when the average is restricted to vertices with the same degree $\langle c \rangle_k$, it decays as $\langle c \rangle_k \sim k^{-\alpha}$ with vertex degree [15–18]. This decay can be observed in Fig. 3 for the four graphs considered here, being more pronounced for the two disassortative graphs in Figs. 3(a) and 3(b), and almost absent for the highly assortative co-authorship graph in Fig. 3(d). In contrast, when computed with our definition (5), $\langle \tilde{c} \rangle_k$ does not exhibit a strong variation with increasing vertex degree (see Fig. 3).

In particular, the decreasing trend is completely absent for the Internet [Fig. 3(a)], indicating that the variations previously observed with the standard definition [15] are reflecting degree correlations. The large variations of $\langle c \rangle_k$ with the vertex degree k have been interpreted as the existence of a hierarchical structure, with high degree vertices interconnecting highly connected subgraphs made of smaller degree vertices, but with no or few connections among vertices in different subgraphs [15,16]. The existence of this hierarchical structure, however, was already predicted from the analysis of the degree correlations [5,10]. The present work makes the bridge between these two different approaches to quantify the hierarchical structure of the Internet, showing that the variations in the clustering coefficient with the vertex degrees, as measured with the usual definition, are just reflecting the existence of degree correlations. These conclusions are also applicable for the protein-protein interaction graph, with a degree of disassorative close to that of the Internet graph.

In the case of the Internet we can also follow changes in the clustering coefficient as the network evolves, with around 3000 vertices in 1997 to 10 000 vertices in 2001. $\langle \tilde{c} \rangle_k$ remains essentially stationary within this period (data not shown), as does $\langle c \rangle_k$ [15]. In contrast, in random graphs with fixed degree distribution and degree correlations the local clustering coefficient approaches zero with increasing graph size, independently of the vertex degree [24]. Therefore, the Internet is more clustered than expected from the degree distribution and degree correlations alone.

In the case of the semantic web [Fig. 3(c)], although the clustering coefficient variations are reduced after filtering out the degree correlations, there is still a logarithmic decrease with increasing the vertex degree [see inset of Fig. 3(c)]. Using a deterministic growing graph model introduced in Ref. [25], we show that this logarithmic decay may be the general case for graphs where $\langle c \rangle_k \sim 1/k$. In the deterministic model, we start with one edge at time $t=-1$. At each time step we create a new triangle on each existing edge by connecting its two endpoints to a new vertex. At time $t=0$ we get one triangle and at time $t=1$, we will have the triangle from the previous step and three new ones, each is using one edge from the existing usual triangle and two new edges with a new vertex between. Since this model is built recursively, we can find by induction the degree of a vertex $k_i(\tau)=2^{\tau+1}$ and the number of triangles passing through it $t_i=k_i-1$, where τ is the time elapsed from the introduction of the vertex, resulting in the clustering coefficient $c_i=2/k_i$ [25]. To compute the clustering coefficient according to our definition (5) we need to determine the scaling of ω_i with the vertex degree k_i. From the Ω_i definition (4) and the evolution rules of the model we obtain the following recursive relation $\Omega_i(\tau+1)=2\Omega_i(\tau)+2^{\tau+1}$. From this recursive relation and the initial condition $\Omega_i(0)=1$ we obtain by induction $\Omega_i(\tau)=(\tau+1)2^{\tau}$. We have also obtained an exact expression for ω_i [20], which in the $\tau \gg 1$ limit results in $\omega_i \approx \Omega_i$ and

$$\tilde{c}_i \approx \frac{2}{\log_2 k_i}. \quad (7)$$

The analysis of the deterministic model indicates that in graphs where the usual definition of clustering coefficient is characterized by an inverse proportionality with the vertex degree, our clustering coefficient will exhibit a logarithmic decrease with increasing the vertex degree. This observation is in agreement with the semantic web data as well [Fig. 3(c)], where $\langle c \rangle_k \sim 1/k$ and $\langle \tilde{c} \rangle_k \sim 1/\log k$.

Finally, for the most assortative graph in Fig. 3(d), we do not observe a substantial difference between the two definitions of clustering coefficient. This observation is explained by the fact that in a highly assortative graph the degree of connected vertices is quite similar, $\omega_i \approx \Omega_i \approx \binom{k_i}{2}$ and the two clustering coefficient definitions give similar results.

The dependence of the usual clustering coefficient with the vertex degree gives information about the degree correlations present in the corresponding graph. These degree correlations, however, can be already characterized using measures that target this topological property, like the degree-correlation coefficient r [11] and the degree dependency of the average degree of the neighbors of a vertex as a function of its degree [5]. Therefore, a definition of clustering coefficient containing the effect of degree correlations is giving redundant information, information which is better characterized using the proper degree correlation measures. In contrast, our definition filters out the degree correlations providing a more specialized topological measure that targets the intrinsic clustering properties alone.

After removing these biases the local clustering coefficient does not depend strongly on the vertex degrees, being of the same order for small and large degree vertices. More precisely, we observe two different scenarios, either the local clustering coefficient is approximately constant or it decays logarithmically with increasing the vertex degree. These results will eventually force us to reevaluate the clustering based analysis of complex networks, and other approaches [16,26–28] based on this magnitude.

The authors thank A.-L. Barabási and A. Vespignani for helpful comments and discussion.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2001).
[2] S. N. Dorogovtsev and J. F. F. Mendes., Adv. Phys. **51**, 1079 (2002).
[3] M. E. Newman, SIAM Rev. **45**, 167 (2003).
[4] *Handbook of Graphs and Networks: From the Genome to the Internet*, edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, Weinheim, 2003).
[5] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, 2004).
[6] P. Erdös and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).
[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[8] A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi, Science **287**, 2115a (2000).
[9] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[10] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[11] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[12] S. Maslov and K. Sneppen, Science **296**, 910 (2002).
[13] J. Park and M. E. J. Newman, Phys. Rev. E **68**, 026112 (2003).
[14] M. Catanzaro, M. Boguña, and R. Pastor-Satorras, e-print cond-mat/0408110.
[15] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).
[16] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002).
[17] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[18] A. Vázquez, Phys. Rev. E **67**, 056104 (2003).
[19] B. Bollobás and O. M. Riordan, in *Handbook of Graphs and Networks: From the Genome to the Internet*, edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, Weinheim, 2003), pp. 1–34.
[20] S. N. Soffer and A. Vázquez (unpublished).
[21] The National Laboratory for Applied Network Research (NLANR), National Science Foundation, http://moat.nlanr.net/
[22] The Database of Interacting Proteins (DIP), http://dip.doe-mbi.ucla.edu/
[23] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **311**, 590 (2002).
[24] S. N. Dorogovtsev, e-print cond-mat/0308444.
[25] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. E **65**, 066122 (2002).
[26] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
[27] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, Proc. Natl. Acad. Sci. U.S.A. **101**, 3747 (2004).
[28] A. Vázquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. Oltvai, and A.-L. Barabási, Proc. Natl. Acad. Sci. U.S.A. **101**, 17940 (2004).