Clustering of correlated networks

S. N. Dorogovtsev*
Departamento de Física and Centro de Física do Porto, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
and A.F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia

(Received 22 August 2003; published 27 February 2004)

We obtain the clustering coefficient, the degree-dependent local clustering, and the mean clustering of networks with arbitrary correlations between the degrees of the nearest-neighbor vertices. The resulting formulas allow one to determine the nature of the clustering of a network.

DOI: 10.1103/PhysRevE.69.027104 PACS number(s): 05.50.+q, 05.10.−a, 05.40.−a, 87.18.Sn

In principle, loops and, in particular, loops of length three which lead to the clustering of networks, are a specific kind of correlations. Usually, real-world networks are strongly clustered structures, and many efforts were made to invent special mechanisms producing strong clustering even in small nets [1,2]. The number of the proposed mechanisms is rapidly growing, but the recent development of the field [3–6] shows that in very many real networks the high clustering is only a finite-size effect. So, in this case, no additional mechanism of strong clustering is needed. The problem is to reliably conclude whether or not the clustering of a real network is a finite-size effect which can be explained by using basic random graph constructions [7]. Evidently, comparison with results obtained in the framework of specific models with many adjusting parameters cannot lead to any convincing conclusion.

Another basic though particular kind of correlations in networks are correlations between the numbers of connections (degrees) of the nearest-neighbor vertices [8–20]. Networks with these specific correlations are being extensively studied these days, and the term ‘‘correlated networks’’ often implies just this type of correlations. These pair correlations were measured in a number of real networks [9–14,19], so the joint distribution of the degrees of the nearest-neighbor vertices, $P(k,k')$ is considered as one of metrics of a network. Note that as a rule, these correlations do not vanish in the large network limit.

The classical random graphs [21,22] with their Poisson degree distribution provide a nonadequate image of a real complex network and a very weak clustering, $C=\overline{k}/N$. Here $\overline{k}$ is the mean degree of a graph and N is its size (the total number of vertices). Random graphs with given degree-distribution $P(k)$ (the configuration model of mathematical graph theory [23]) are much closer to real complex networks. It is the values of the clustering coefficient of this model $C\propto N^{-1}$ [7,24] that were compared with empirical data for real-world networks.

The configuration model and its variations provide (uncorrelated) random graphs which are maximally random (i.e., with the maximum entropy) under the constraint that their degree distribution is equal to a given one, $P(k)$. These graphs are closer to reality than the classical random graphs, but the absence of correlations is a very restrictive factor. If we wish to make a step toward real networks, we have to introduce a network with degree–degree correlations, $P(k,k')$. The simplest formal way to do this is in the spirit of the configuration model. That is, consider random graphs which are maximally random under the constraint that their degree–degree distribution is equal to a given one, $P(k,k')$. This is the minimal construction of a random graph with these correlations. In this construction, as the size of a graph approaches infinity, loops become insignificant, and the clustering vanishes [25].

In the present report we obtain analytical expressions for the complete list of the clustering characteristics of the random graphs with these important degree–degree correlations [see Eqs. (8)–(10)]. These formulas, after the substitution of a measured distribution $P(k,k')$, allow one to conclude whether or not the clustering of a real-world or generated network is simply a finite-size effect, the same as in a maximally random graph with this degree–degree distribution. Furthermore, the resulting clustering characteristics are qualitatively different from those of uncorrelated networks.

The graphs in this report are completely described by the joint distribution $P(k,k')$ of the degrees of end vertices of an edge of the graph, $\Sigma_{k,k'}P(k,k')=1$, $P(k,k')=P(k',k)$. The degree distribution $P(k)$ is determined by $P(k,k')$

$$P(k)=\frac{\overline{k}}{k}\sum_{k'} P(k,k'), \tag{1}$$

where the mean degree $\overline{k}\equiv\langle k\rangle\equiv\Sigma_k kP(k)$ is

$$\overline{k}=\left[\sum_{k,k'}\frac{P(k,k')}{k}\right]^{-1}. \tag{2}$$

In the following, we assume that the total number of vertices of the graph, N, is large and consider only the main contribution to the clustering.

$P(k,k')$ can be obtained by using empirical data as follows. If $k\neq k'$, $P(k,k')=P(k',k)$ is one half of the ratio of the number of edges connecting vertices of degrees k and k' to the total number of edges, $L=\overline{k}N/2$. If $k=k'$, $P(k,k)$ is the ratio of the number of edges connecting vertices of de-

*Electronic address: sdorogov@fc.up.pt

grees k and k to L. As is natural, the case $k=k'$ is perfectly adjacent to the case $k\neq k'$. These cases are presented separately for the sake of clarity.

The set of clustering characteristics of networks, considered up to now, includes the following.

(i) The degree-dependent local clustering $C(k)$. This is the mean relative number of connections (less than 1) between two nearest neighbors of a vertex of degree k

$$C(k)\equiv\frac{\langle m_{\mathrm{nn}}(k)\rangle}{k(k-1)/2}, \tag{3}$$

where $\langle m_{\mathrm{nn}}(k)\rangle$ is the average number of connections between the nearest neighbors of a vertex of degree k.

(ii) The mean clustering (mean clustering coefficient), which is defined as

$$\overline{C}\equiv\sum_k P(k)C(k). \tag{4}$$

(iii) The clustering coefficient, which is defined as

$$C\equiv\frac{\Sigma_k P(k)\langle m_{\mathrm{nn}}(k)\rangle}{\Sigma_k P(k)k(k-1)/2}=\frac{\Sigma_k k(k-1)P(k)C(k)}{\langle k^2\rangle-\overline{k}}. \tag{5}$$

This coincides with the traditional definition: the clustering coefficient is three times the ratio of the total number of loops of length three in a graph to the total number of connected vertex triples. In simple terms, this is the ‘‘concentration’’ of loops of length three.

In the network literature, C and $\overline{C}$ are often mistakenly mixed, and both are called ‘‘the clustering coefficient.’’ Nonetheless, the difference may be great. In real networks, up to a tenfold difference was observed [6]. One can even find examples of (infinite) nets where $C=0$ while $\overline{C}$ is finite.

We shall obtain the clustering characteristics $C(k)$, $\overline{C}$, and C of correlated graphs, but let us first introduce the conditional probability $P(k|k')$ that if one end vertex of an edge is of degree k', then its other end vertex is of degree k:

$$P(k|k')=\frac{P(k,k')}{\Sigma_k P(k,k')}=\overline{k}\frac{P(k,k')}{k'P(k')}. \tag{6}$$

Then the local clustering, that is, the probability that two nearest neighbors of a vertex of degree $k>1$ are connected is

$$C(k)=\sum_{q,q'>1}P(q'|k)P(q|k)\cdot P(q'|q)\frac{(q'-1)}{Nq'P(q')}\cdot(q-1). \tag{7}$$

One can easily understand this formula.

(i) The first two factors $P(q'|k)P(q|k)$ on the right-hand side, which should be accounted for before the summation over q and q', are evident: these are the probabilities that the vertices are of degrees q and q'.

(ii) In fact, we must calculate the probability that the nearest neighbors with degrees q and q' of a vertex of degree k are connected to each other. We have two vertices with $q-1$ and $q'-1$ ‘‘free connections’’ (apart of the connections to the mother vertex). Let us select one of the free connections of the q vertex. The probability that this edge will ‘‘choose’’ one of the $q'-1$ free connections of the q' vertex is given by the product between two central dots on the right-hand part of the formula. The factor $F(q'|q)$ is evident: the second end of the edge must be of degree q'. So our edge must choose (‘‘grasp’’) one of the $q'-1$ free connections of the q' vertex among almost $Nq'P(q')$ possibilities in the network. (All these possibilities are equiprobable in the construction which is considered here.) This is the total number of free connections provided by $NP(q')$ vertices of degree q' in the network. This gives $(q'-1)/[Nq'P(q')]$.

(iii) Finally, we must multiply this probability by the number $q-1$ of the free connections of the q-vertex.

The result is Eq. (7). Note that we used the fact that N is large and the probability that the edge between the nearest neighbor present is small, so our formulas are asymptotic. Substituting Eq. (6) into Eq. (7) gives the degree-dependent local clustering

$$C(k)=\frac{\overline{k}^3}{Nk^2P^2(k)}\times\sum_{q,q'>1}\frac{(q'-1)(q-1)P(q',q)P(q',k)P(q,k)}{qq'P(q)P(q')}, \tag{8}$$

the mean clustering

$$\overline{C}=\frac{\overline{k}^3}{N}\sum_{k,q,q'>1}\frac{(q'-1)(q-1)P(q',q)P(q',k)P(q,k)}{k^2qq'P(q)P(q')P(k)}, \tag{9}$$

and the clustering coefficient

$$C=\frac{\overline{k}^3}{N(\langle k^2\rangle-\overline{k})}\times\sum_{k,q,q'>1}\frac{(k-1)(q'-1)(q-1)P(q',q)P(q',k)P(q,k)}{kqq'P(q)P(q')P(k)} \tag{10}$$

of the correlated network with given correlations $P(k,k')$. The degree distribution $P(k)$ in these formulas may be expressed in terms of $P(k,k')$ by using the relations (1) and (2). The results (8)–(10) may be written in a more compact form in terms of conditional probabilities, see Eq. (6), but the present form is more convenient for empirical researchers.

In uncorrelated networks, $P(k,k')=kP(k)k'P(k')/\overline{k}^2$ and the probability that the nearest neighbor of a vertex is of degree k is $kP(k)/\overline{k}$. In this case, Eqs. (8)–(10) reduce to the known result [7,24]

$$C(k)=\overline{C}=C=\frac{(\langle k^2\rangle-\overline{k})^2}{N\overline{k}^3}. \tag{11}$$

The formulas (8)–(11) are asymptotically exact.

Note that in uncorrelated networks, $C(k)$ is independent of k and so all the three characteristics are equal. Contrastingly, degree-degree correlations lead to a degree-dependent local clustering [see Eq. (8)]. Previously, this feature was observed in a number of model and real networks [26–31]. Here we demonstrate that this dependence is a direct consequence of degree-degree correlations. The degree-dependent local clustering leads to the difference between $\overline{C}$ and C, which were found in many real-world networks [28–30].

One should note that the formula (8) for the degree-dependent local clustering resembles the expression (60) for the local clustering of a correlated network with hidden variables in the recent paper of Mariań Boguñá and Romualdo Pastor-Satorras, Ref. [32]. However, there is an essential difference between these two results. The result of Ref. [32] is $C(k)$, expressed in terms of the correlations of hidden variables (‘‘fitnesses’’) which were used to generate a correlated network. It is impossible to find the exact form of these hidden variable correlations from empirical data. Contrastingly, Eq. (8) in the present work is obtained for a random network, which is completely described by $P(k,k')$, and expresses $C(k)$ directly in terms of the observable degree–degree distribution $P(k,k')$. It is the latter circumstance that allows one to use Eqs. (8)–(10) for the structural analysis of networks.

The number of edges connecting vertices of degrees k and k' can be easily measured in any real-world or generated network [11–13,19]. Substituting these numbers together with the numbers of vertices of degree k into Eqs. (8)–(10) will provide one with the clustering characteristics of a maximally random graph with the same degree–degree correlations as the real network. These clustering characteristics may be compared with those of the real net. If the results are close enough, then the clustering of a net is explained by the basic correlated random graph construction and so is a simple finite-size effect. Only if the calculated characteristics differ strongly from the measured ones, the clustering has nontrivial nature.

Note that in sparse networks, measured degree–degree distributions strongly fluctuate due to poor statistics. This factor cannot spoil the results (8)–(10), since even strong fluctuations are summed out.

One should indicate two restrictions. (i) The formulas (8)–(11) are asymptotic (large N, sufficiently ‘‘weak’’ clustering). So, one may hope that they are good if C is less than, say, 0.1, but only qualitative comparison is possible if, e.g., $C \sim 0.3$. (ii) The growth of real-world networks produces a wide spectrum of correlations, and the correlations between the degrees of the nearest-neighbor vertices are only one specific type of correlations. The construction that is considered in this report ignores the long-range and multivertex correlations. The empirical data on such correlations is absent.

In summary, we obtained the clustering characteristics of networks with correlations between degrees of the nearest-neighbor vertices. These correlations are a common feature of real networks. Our formulas allow one to easily conclude whether or not the clustering of a network is determined by the form of its degree–degree distribution and so is a simple finite-size effect. So, Eqs. (8)–(10) can shed light on the nature of the clustering of networks. We hope that these simple expressions will be a useful tool for the analysis of real-world and generated networks.

This work was partially supported by Project No. POCTI/1999/FIS/33141. The author thanks A.V. Goltsev, M. Boguñá, and A.N. Samukhin for useful discussions. Special thanks to the Centro de Física do Porto and J.F.F. Mendes.

[1] D.J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton University Press, Princeton, NJ, 1999).

[2] S.H. Strogatz, Nature (London) **401**, 268 (2001).

[3] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[4] S.N. Dorogovtsev and J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002).

[5] S.N. Dorogovtsev and J.F.F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

[6] M.E.J. Newman, SIAM Rev. **45**, 167 (2003).

[7] M.E.J. Newman, in *Handbook of Graphs and Networks: From the Genome to the Internet*, edited by S. Bornholdt and H.G. Schuster (Wiley-VCH, Weinheim, 2002), pp. 35–68.

[8] P.L. Krapivsky and S. Redner, Phys. Rev. E **63**, 066123 (2001).

[9] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).

[10] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).

[11] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[12] S. Maslov, K. Sneppen, and A. Zaliznyak, e-print cond-mat/0205379.

[13] S. Maslov, K. Sneppen, and U. Alon, in *Handbook of Graphs and Networks: From the Genome to the Internet*, edited by S. Bornholdt and H.G. Schuster (Wiley-VCH, Weinheim, 2002), pp. 168–198.

[14] M.E.J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).

[15] J. Berg and M. Lässig, Phys. Rev. Lett. **89**, 228701 (2002).

[16] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).

[17] J. Park and M.E.J. Newman, Phys. Rev. E **68**, 026112 (2003).

[18] A. Vázquez, M. Boguñá, Y. Moreno, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **67**, 046111 (2003).

[19] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen, e-print cond-mat/0308339.

[20] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, e-print cond-mat/0206467.

[21] P. Erdős and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959); Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).

[22] R. Solomonoff and A. Rapoport, Bull. Math. Biophys. **13**, 107 (1951); R. Solomonoff, *ibid.* **14**, 153 (1952); A. Rapoport, *ibid.* **10**, 145 (1957).

[23] A. Bekessy, P. Bekessy, and J. Komlos, Stud. Sci. Math. Hung. **7**, 343 (1972); E.A. Bender and E.R. Canfield, J. Comb. Theory, Ser. A **24**, 296 (1978); B. Bollobás, Eur. J. Comb. **1**, 311 (1980); N.C. Wormald, J. Comb. Theory, Ser. B **31**, 156 (1981).

[24] H. Ebel, L.-I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 035103 (2002).

[25] There is a hierarchy of minimal constructions of complex networks: (i) the maximally random network with a given degree distribution $P(k)$; (ii) the maximally random network with a given distribution $P(k,k')$ of the degrees of the nearest-neighbor vertices; (iii) the maximally random network with a given distribution $P(k,k',k'')$ of the degrees of a triple of connected vertices; and so on. All these are equilibrium networks with a treelike local structure. In this report we discuss only random graphs with a given degree-degree distribution $P(k,k')$, since any empirical data on $P(k,k',k'')$ and higher-order multivertex correlations is absent.

[26] S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes, Phys. Rev. E **65**, 066122 (2002).

[27] G. Szabó, M. Alava, and J. Kertesz, Phys. Rev. E **67**, 056102 (2003).

[28] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).

[29] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002).

[30] A. Vázquez, Phys. Rev. E **67**, 056104 (2003).

[31] A. Fronczak, P. Fronczak, and J.A. Holyst, Phys. Rev. E **68**, 046126 (2003).

[32] M. Boguñá and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).