Mean-field theory for clustering coefficients in Barabási-Albert networks

Agata Fronczak, Piotr Fronczak, and Janusz A. Hołyst
Faculty of Physics and Center of Excellence for Complex Systems Research, Warsaw University of Technology, Koszykowa 75, PL-00-662 Warsaw, Poland

(Received 11 June 2003; published 23 October 2003)

We applied a mean-field approach to study clustering coefficients in Barabási-Albert (BA) networks. We found that the local clustering in BA networks depends on the node degree. Analytic results have been compared to extensive numerical simulations finding a very good agreement for nodes with low degrees. Clustering coefficient of a whole network calculated from our approach perfectly fits numerical data.

DOI: 10.1103/PhysRevE.68.046126 PACS number(s): 89.75.−k, 02.50.−r, 05.50.+q

INTRODUCTION

During the last decade networks became a very popular research domain among physicists (for a review see Refs. [1–3]). It is not surprising, since networks are everywhere. They surround us. In our daily life we participate in dozens of them. A number of social institutions, communication, and biological systems may be represented as networks, i.e., sets of nodes connected by links. It was observed that despite functional diversity most of real web-like systems share similar structural properties. The properties are: fat-tailed degree distribution (that allows for hubs, i.e., nodes with high degree), small average distance between any two nodes (the so-called *small world* effect) and a large penchant for creating cliques (i.e., highly interconnected groups of nodes).

A number of network construction procedures have been proposed to incorporate the characteristics. The Barabási-Albert (BA) [4,5] growing network model is probably the best known. Two important ingredients of the model are continuous network growth and preferential attachment. The network starts to grow from an initial cluster of m fully connected sites. Each new node that is added to the network creates m links that connect it to the previously added nodes. The preferential attachment means that the probability of a new link to end up in a vertex i is proportional to the connectivity k_i of this vertex

$$\Pi_i = m\frac{k_i}{\sum_j k_j}. \tag{1}$$

Taking into account that $\Sigma_j k_j = 2mt$ the last formula may be rewritten as $\Pi_i = k_i/(2t)$. By means of mean-field approximation [5] one can find that the average degree of a node i that entered the network at time t_i increases with time as a power law

$$k_i(t) = m\sqrt{\frac{t}{t_i}}. \tag{2}$$

Taking advantage of the above formula one can calculate the probability that two randomly selected nodes i and j are nearest neighbors. It is given by

$$p_{ij} = \frac{m}{2}\frac{1}{\sqrt{t_i t_j}}. \tag{3}$$

It was shown that the degree distribution in BA network follows a power law

$$P(k) = \frac{2m^2}{k^3}, \tag{4}$$

where $k = m, m+1, \ldots, m\sqrt{t}$. The power-law degree distribution is characteristic of many real-world networks and the scaling exponent $\alpha_{BA} = 3$ is close to those observed in real systems [α_{real} is roughly limited to the range (1–3)]. It was also shown that the BA model is a small world. The mean distance between sites in the network having t nodes behaves as $l \sim \ln t/\ln\ln t$ [6,7]. The only shortcoming of the model is that it does not incorporate a high degree of cliqueness observed in real networks.

In this paper we study cliqueness effects in BA networks. The cliqueness is measured by the clustering coefficient C [8,9]. The clustering coefficient C_i of a single node i describes the density of connections in the neighborhood of this node. It is given by the ratio of the number E_i of links between the nearest neighbors of i and the potential number of such links $E_{max} = k_i(k_i-1)/2$,

$$C_i = \frac{E_i}{E_{max}} = \frac{2E_i}{k_i(k_i-1)}. \tag{5}$$

The clustering coefficient C of the whole network is the average of all individual C_i's. It is known, from numerical calculations, that the clustering coefficient in BA networks rapidly decreases with the network size t. In this article we apply a mean-field approach to study the parameter. Our calculations confirm that in the limit of large ($t \gg 1$) and dense ($m \gg 1$) BA networks the clustering coefficient scales as the clustering coefficient in random graphs [10–12] with an appropriate scale-free degree distribution (4)

$$C = \frac{(m-1)}{8}\frac{(\ln t)^2}{t}. \tag{6}$$

We also show that the individual clustering coefficient C_i in the BA network weakly depends on the node's degree k_i. The dependence is almost invisible when one looks at numerical data presented by other authors [13].

MEAN-FIELD APPROACH

Let us concentrate on a certain node i in a BA network of size t. We assume that $m \geqslant 2$. The case of $m=1$ is trivial. BA networks with $m=1$ are trees thus the clustering coefficient in these networks is equal to zero. By the definition (5) the clustering coefficient C_i depends on two variables E_i and k_i. Since in the BA model only new nodes may create links, the coefficient C_i changes only when its degree k_i changes, i.e., when new nodes create connections to i and $x=0,\dots,m-1$ of its nearest neighbors. The appropriate equation for changes of C_i is then

$$\frac{dC_i}{dt}=\sum_{x=0}^{m-1} \tilde{p}_{ix}\Delta C_{ix}, \qquad (7)$$

where ΔC_{ix} denotes the change of the clustering coefficient when a new node connects to the node i and to x of the first neighbors of i, whereas $\tilde{p}_{ix}$ describes the probability of this event. ΔC_{ix} is simply the difference between clustering coefficients of the same node i calculated after and before a new node attachment

$$\Delta C_{ix}=\frac{2(E_i+x)}{k_i(k_i+1)}-\frac{2E_i}{k_i(k_i-1)}=-\frac{2C_i}{k_i+1}+\frac{2x}{k_i(k_i+1)}. \qquad (8)$$

The probability $\tilde{p}_{ix}$ is a product of two factors. The first factor is the probability of a new link to end up in i. The probability is given by Eq. (1). The second one is the probability that among the rest of $(m-1)$ new links x links connect to nearest neighbors of i. It is equivalent to the probability that $(m-1)$ Bernoulli trials with the probability for success equal to $\Sigma_{j*}k_j/\Sigma_v k_v=\Sigma_{j*}k_j/(2mt)$ result in x successes (Σ_{j*} runs over the nearest neighbors of the node i). Replacing the sum Σ_{j*} by an integral one obtains

$$\sum_{j*} k_j=\int_1^t k_j p_{ij} dt_j=\frac{m}{2}k_i \ln t. \qquad (9)$$

Summarizing the above discussion one yields the relation

$$\tilde{p}_{ix}=\frac{k_i}{2t}\binom{m-1}{x}\left(\frac{k_i \ln t}{4t}\right)^x\left(1-\frac{k_i \ln t}{4t}\right)^{m-1-x}. \qquad (10)$$

Now, inserting Eqs. (2), (8), and (10) into Eq. (7) one obtains after some algebra

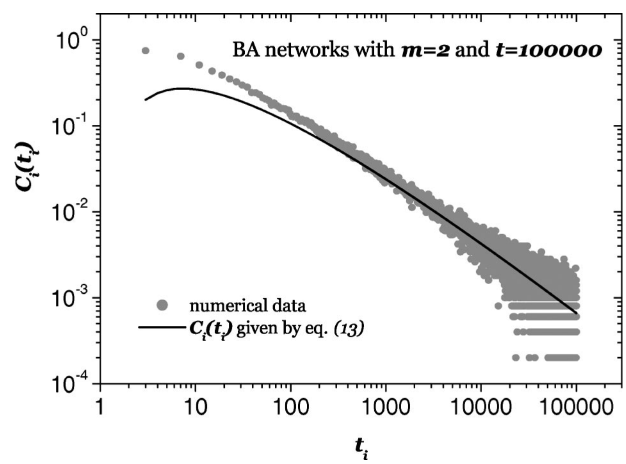

FIG. 1. The initial value of the local clustering coefficient $C_i(t_i)$ (averaged over 1000 BA networks).

$$\frac{dC_i}{dt}=-\frac{m}{(mt+\sqrt{tt_i})}C_i+\frac{m(m-1)\ln t}{4(mt^2+t\sqrt{tt_i})}. \qquad (11)$$

Solving the equation for C_i one gets

$$C_i(t)=\frac{(m-1)}{8(\sqrt{t}+\sqrt{t_i}/m)^2}\left((\ln t)^2-\frac{4}{m}\sqrt{\frac{t_t}{t}}\ln t-\frac{8}{m}\sqrt{\frac{t_i}{t}}+B\right), \qquad (12)$$

where B is an integration constant and may be determined from the initial condition $C_i(t_i)$ that describes the clustering coefficient of the node i exactly at the moment of its attachment t_i

$$C_i(t_i)=\frac{1}{2}\sum_j \sum_v p_{ij}p_{iv}p_{jv}/\binom{m}{2}=\frac{m^2}{8(m-1)}\frac{(\ln t_i)^2}{t_i}. \qquad (13)$$

Following the notation introduced by Bianconi and Capocci [14], the initial clustering coefficient $C_i(t_i)$ may be written as

$$C_i(t_i)=\frac{1}{\binom{m}{2}}\left[\frac{\partial\langle N_3(t)\rangle}{\partial t}\right]_{t=t_i}, \qquad (14)$$

where $\partial\langle N_3(t)\rangle/\partial t$ describes how the number of triangular loops increases in time. Figure 1 shows the prediction of Eq. (13) in comparison with numerical results. For small values of t_i, the numerical data differ from the theory in a significant way. This can be explained by the fact that the formula for the probability of a connection p_{ij} (3), that we use three times in Eq. (13), holds only in the asymptotic region $t_i \to \infty$.

Taking into account the initial condition $C_i(t_i)$ and neglecting mutually compensating terms that occur in Eq. (12) after putting B calculated from Eq. (13) one obtains the formula for time evolution of the clustering coefficient of a given node i

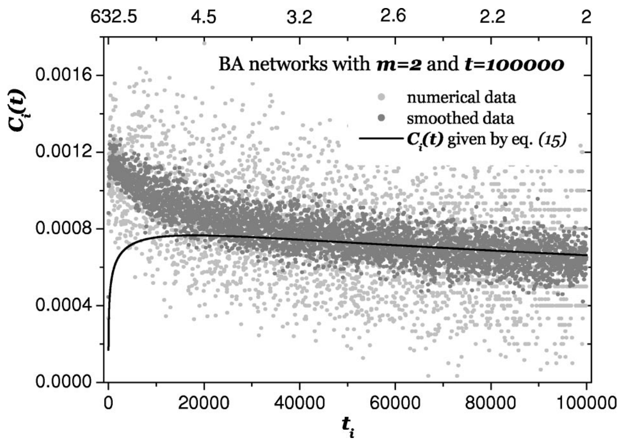

FIG. 2. The local clustering coefficient $C_i(t)$ as a function t_i (averaged over 10^4 networks). Note that the k_i axis is nonlinear.

$$C_i(t)=\frac{(m-1)}{8(\sqrt{t}+\sqrt{t_i}/m)^2}\left[(\ln t)^2+\frac{4m}{(m-1)^2}(\ln t_i)^2\right]. \tag{15}$$

Let us note that if $t_i \ll t$ or $m \gg 1$, the local clustering coefficient does not depend on the node under consideration and approaches $C_i(t) \simeq (m-1)(\ln t)^2/(8t)$, i.e., the formula (6) that gives the clustering coefficient of a random graph with a power-law degree distribution (4). Since one knows how the node's degree evolves in time (2) one can also calculate the formula for $C_i(k_i)$. At Fig. 2 we present the formula (15) (solid line) and corresponding numerical data (scatter plots). The two kinds of scatter plots represent respectively: real data (light gray circles) and the same data subjected to adjacent averaging smoothing (dark gray circles). As before (see Fig. 1), we observe a significant disagreement between the numerical data and the theory for small t_i. We suspect that the effect has the same origin, i.e., the relations (2) and (3) that we use in our derivation work well only in the asymptotic region $t_i < t \to \infty$.

To obtain the clustering coefficient C of the whole network the expression (15) has to be averaged over all nodes within a network $C=\int_1^t C_i(t)dt_i/t$. We were not able to find an exact analytic form of this integral but we found its asymptotic form in the limit of large $t\to\infty$ and dense $m \gg 1$ BA networks. Taking advantage of the second mean value theorem for integration [15,16] the clustering coefficient C may be written in the form

$$C=\frac{m^2(m+1)^2}{4(m-1)}\left[\frac{m}{(m+1)}+\ln((m+1)\sqrt{t})-\frac{m\sqrt{t}}{\sqrt{\xi}+m\sqrt{t}}-\ln(\sqrt{\xi}+m\sqrt{t})\right]\frac{[\ln(t)]^2}{t}, \tag{16}$$

where $1<\xi<t$ is unknown parameter. Note that for large and dense networks the term $\sqrt{\xi}$ in formula (16) may be neglected in comparison with $m\sqrt{t}$ and the expression for C may be rewritten as

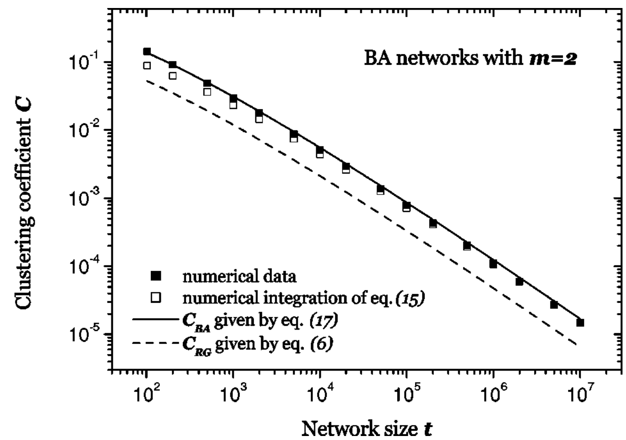

FIG. 3. The clustering coefficient C of a whole BA network as a function of the network size t (averaged over 100 networks).

$$C=\frac{m^2(m+1)^2}{4(m-1)}\left[\ln\left(\frac{m+1}{m}\right)-\frac{1}{m+1}\right]\frac{[\ln(t)]^2}{t}. \tag{17}$$

For large $(t\to\infty)$ and dense $(m\gg 1)$ networks the above formula approaches Eq. (6). The effect lets us deduce that the structural correlations [17] characteristic for growing BA networks become less important in larger and denser networks. The same was suggested in Ref. [6]. Figure 3 presents the average clustering coefficient in BA networks as a function of the network size t compared with the analytical formula (17) (solid line) and numerical integration of Eq. (15) (open squares). Paradoxically, the approximate integration of Eq. (15) given by Eq. (17) better fits the real data then the numerical integration of Eq. (15).

CONCLUSIONS

In summary, we applied a mean-field approach to study clustering effects in Barabási-Albert networks. We found that local clustering coefficients $C_i(t)$ in BA networks are not completely homogeneous as suggested in Refs. [10,13]. The observed small deviations of $C_i(t)$ from the global network parameter $C(t)$ are especially visible for old nodes $(t_i \ll t)$. We derived a general formula for the clustering coefficient C characterizing the whole BA network. We found that in the limit of large $(t\to\infty)$ and dense $(m\gg 1)$ networks both the local (C_i) and the global (C) clustering coefficients approach clustering coefficient derived for a random graph with a power-law degree distribution (4). Our derivations were checked against numerical simulation of BA networks finding a very good agreement.

ACKNOWLEDGMENTS

We would like to thank Jarosław Suszek and Daniel Kikoła for their help in computer simulations. A.F. acknowledges The State Committee for Scientific Research in Poland for support under Grant No. 2P03B01323. The work of J.A.H. was supported by the special program of the Warsaw University of Technology, *Dynamics of Complex Systems*.

[1] S. Bornholdt and H. G. Schuster, *Handbook of Graphs and Networks* (Wiley, New York, 2002).
[2] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks* (Oxford University Press, Oxford, 2003).
[3] R. Albert and A.L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[4] A.L. Barabási and R. Albert, Science **286**, 509 (1999).
[5] A.L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).
[6] A. Fronczak, P. Fronczak, and J.A. Hołyst, e-print cond-mat/0212230.
[7] R. Cohen and S. Havlin, Phys. Rev. Lett. **90**, 058701 (2003).
[8] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).
[9] A. Fronczak, J.A. Hołyst, M. Jedynak, and J. Sienkiewicz, Physica A **316**, 688 (2002).
[10] K. Klemm and V.M. Eguíluz, Phys. Rev. E **65**, 057102 (2002).
[11] G. Szabó, M. Alava, and J. Kertész, Phys. Rev. E **67**, 056102 (2003).
[12] M.E.J. Newman, e-print cond-mat/0202208.
[13] E. Ravasz and A.L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[14] G. Bianconi and A. Capocci, Phys. Rev. Lett. **90**, 078701 (2003).
[15] The second mean value theorem for integration [16] states: If $f(x):[a,b]\to R$ is a positive and monotone increasing function and $g(x):[a,b]\to R$ is an integrable function, then there exists a number ξ in (a,b) such that $\int_a^b f(x)g(x)dx = f(b)\int_\xi^b g(x)dx$. To estimate the integral we assigned the numerator of the formula (15) to $f(x)$ and the denominator of the same formula to $g(x)$.
[16] G. M. Fichtenholz, *Differential and Integral Calculus* (PWN, Warsaw, 1995), Vol. 2.
[17] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, e-print cond-mat/0206467.