Structural transitions in scale-free networks

Gábor Szabó,[1,2] Mikko Alava,[2,3] and János Kertész[1,4]
[1]*Department of Theoretical Physics, Institute of Physics, Budapest University of Technology, 8 Budafoki út, H-1111 Budapest, Hungary*
[2]*Laboratory of Physics, Helsinki University of Technology, P.O. Box 1100, FIN-02015 HUT, Finland*
[3]*NORDITA, Blegdamsvej 17, DK-2100 Copenhagen, Denmark*
[4]*Laboratory of Computational Engineering, Helsinki University of Technology, FIN-02015 HUT, Finland*
(Received 27 August 2002; published 6 May 2003)

Real growing networks such as the World Wide Web or personal connection based networks are characterized by a high degree of clustering, in addition to the small-world property and the absence of a characteristic scale. Appropriate modifications of the (Barabási-Albert) preferential attachment network growth capture all these aspects. We present a scaling theory to describe the behavior of the generalized models and the mean-field rate equation for clustering. This is solved for a specific case with the result $C(k) \sim 1/k$ for the clustering of a node of degree k. This mean-field exponent agrees with simulations, and reproduces the clustering of many real networks.

DOI: 10.1103/PhysRevE.67.056102 PACS number(s): 89.75.Hc, 05.70.Ln, 87.23.Ge, 89.75.Da

I. INTRODUCTION

In diverse fields of scientific interest underlying network structures can be recognized, which provide a unifying concept of investigation [1]. Examples range from biology (metabolic networks [2], protein nets in the cell [3]) through sociology (movie actor relationships [4], co-author networks [5], sexual nets [6]) to informatics (Internet [7], World Wide Web (WWW) [8]). In all these examples it is easy to identify the constituents of the problem with the nodes of a graph and their relationships with directed or undirected links. During the past few years a great deal of information has accumulated about such structures. Three apparent features seem to characterize them rather robustly: (i) a high degree of clustering, i.e., if nodes A and B are linked to node C then there is a good chance that A and B are also linked; (ii) the ''small-world'' property, i.e., the expected number of links needed to reach from one arbitrarily selected node to another one is low; (iii) the absence of a characteristic scale, which often appears so that the distribution $P(k)$ of the degrees k of nodes follows a power law.

Clustering in real networks is an essential and an almost ubiquitous feature [9]. It measures the deviation from a structure with vanishing correlations, and it has been used to describe the tendency of networks to form cliques or tightly connected neighborhoods. As an organizing principle, this is most obvious in social networks, where connections are usually created by personal acquaintances, such as in the scientific collaboration network. Considerable clustering has also been found in networks of more diverse nature. Prime examples are the WWW, metabolic and protein interaction networks, the actor network, the power grid of the United States, the semantic web of english words [9], and the backbone of the Internet on both the autonomous system and the router level [10,11]. The number of entries in this list is on the rise as new disciplines are being taken under consideration and raw data are made available. A comprehensive examination of a variety of examples of clustering can be found in Ref. [9]. For a particular node, the *clustering coefficient* is defined as $C = n/[k(k-1)/2] \in [0;1]$, where n is the number of links between the neighbors of the node and k is its degree. In real networks, as a combination of the properties (i) and (iii), the clustering coefficient as a function of the degree of the nodes often follows a power law: $C(k) \propto k^{-\alpha}$. The value of α is in many networks close to 1.

In 1998, Watts and Strogatz created an interesting family of models: introducing a rather low proportion of random links between arbitrarily selected pairs of nodes in a regular lattice has the consequence that property (ii) gets fulfilled while clustering does not decrease considerably, assuring (i) [12,13]. However, the distribution of the degrees of nodes shows a characteristic peak instead of the required power law. Barabási and Albert (BA) realized that in the examples mentioned at the beginning an important aspect is that the networks are created by growth. BA proposed preferential attachment (PA) as a growth rule: the new nodes are linked to the old ones with a probability proportional to their actual degree [4]. The structures obtained this way are scale-free and have the small-world property. In spite of capturing important aspects of growing networks, the clustering tends rapidly to a constant as a function of the degree k and vanishes in the thermodynamic limit.

Recently, attempts have been undertaken to modify the PA network growth models so as to increase clustering. In these models a mechanism, controlled by a new parameter, is introduced to take into account the effect that ''friends of friends get friends.'' Indeed, it has been possible to create models which have all the three properties (i)–(iii) [9,14,15].

The aim of this paper is to present a general framework, applicable to the transition from a PA graph with zero clustering to still scale-free graphs with $C(k) \propto k^{-\alpha}$. For this purpose we consider a corresponding mean-field (MF) and a rate-equation theory. We propose these as a combined approach to study structural correlations (here clustering, i.e., triangle formation or three-point correlations, but loops in general could be discussed). As an example we will take the Holme-Kim model [14] (a modified BA one), for which the MF rate equations can be solved, leading to $\alpha = 1$. This is also shown to describe the simulations very well, and the mechanism involved, though very simple, suggests why

many real networks have such an α as well. At the end, we discuss further possibilities.

II. GENERAL SCALING THEORY

We start from the simplest undirected BA model: a new node j with m links is added to the system at (discrete) time t. A link from node j to node i is drawn with probability $k_i/\Sigma k_i$. It is known that the average clustering at node i is independent of the degree k_i [15]:

$$C(i|k_i=k)=\frac{m-1}{8}\frac{(\ln N)^2}{N}, \tag{1}$$

i.e., it is inversely proportional to the number N of nodes (with a logarithmic correction) [16]. For the generalization of the BA model with enhanced clustering, we have a parameter p representing an imposed tendency to form triangles on the graph. It is chosen such that at $p=0$ the original BA model is recovered.

We propose as a scaling ansatz to describe the clustering coefficient C as a function of the degree k, the number of nodes N, and the parameter p:

$$C(k,N,p)=N^{-1}f\left(\frac{k}{k^*(N,p)}\right), \tag{2}$$

where $f(x)$ is a scaling function with $f(x)\to$const for $x\gg 1$ and $f(x)\to x^{-\alpha}$ for $x\ll 1$ and the behavior in Eq. (1) is already taken into account by fixing the exponent of the prefactor of f. The characteristic degree k^* is a monotonically increasing function of N for fixed p and it should decrease as p goes to zero. A natural assumption is then

$$k^*(N,p)\sim N^{\gamma}p^{\delta}. \tag{3}$$

As for small k the clustering C in Eq. (2) should go like $k^{-\alpha}$ and become independent of N, we have $\gamma=1/\alpha$. The exponent $\delta\alpha$ describes, how for $N\to\infty$ the clustering C approaches its limiting value zero as p goes to zero. If we accept that in most cases $\alpha=1$, there is one exponent to be determined, say δ. We now clarify the origin of $\alpha=1$ and $\delta=1$ for the model employed.

For this purpose we write down the rate equations for the clustering in a general form. We thus need to consider the rate of change averaged over many realizations,

$$\frac{\partial n_i}{\partial t}=R(k_i,p)\sum_{n\in\Omega}R(k_n,p), \tag{4}$$

where n_i is the average number of connected neighbors of site i, and $C_i=n_i/[k_i(k_i-1)/2]$. Here R is the rate at which i gets new links (or even loses them, if applied to processes with reattachment or deletion of links). We allow, in analogy with the scaling ansatz presented above, the rate to depend on both the degrees of the node in question and the parameter p. This can be ''annealed'' or ''quenched,'' depending on whether the parameter describes stochastic rules (as in the example below) or a fixed property of each node i, e.g., R can simply follow from the preferential attachment rule. Ω is the set of neighbors of node i and the sum accounts for the probability that a new node linked to i also links to one of the neighbors of i. This increases n_i and enhances clustering. In order to make Eq. (4) more concrete, we discuss the triad formation model [14] as an example.

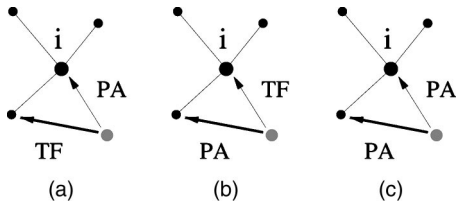

FIG. 1. Three different options to connect to node i with $m\geqslant 2$. In (a), a PA step is performed first linking to i and then a TF step creates a link between neighbors of i. In (b), the same happens, in a different order. (c) shows how two PA steps may contribute to n_i. Bold edges increase n_i.

III. THE TRIAD FORMATION MODEL

The complications in solving a rate equation like Eq. (4) arise from the correlations that are embedded between the degree of node i and the properties of its neighborhood. For the triad formation model, the rules consist of a BA model extended by a triad formation step. Initially, the network contains m_0 vertices and no edges, and in every time step a new vertex is added with m undirected edges. The m edges are then one by one subsequently linked to m different nodes in the network. One performs a preferential attachment step for the first edge as defined in the BA model. With probability p, the second and further edges are joined to a randomly chosen neighbor of the node selected in the previous PA step. Alternatively, with probability $1-p$, a PA step is performed again.

In the limit when p approaches zero, one recovers the original BA model, and by setting p to a value between 0 and 1 the average clustering can be adjusted continuously and grows monotonically with an increasing p. The microscopic mechanisms that increase n_i are illustrated in Fig. 1 and are the following: (a) the new node connects to node i in a PA step, which is potentially followed by several TF steps; (b) the new node connects to one of the neighbors of i in a PA step and then i conversely gets linked to the new node in one of the subsequent TF steps; (c) the new node connects to node i in a PA step and a neighbor of i is also selected for connection to the new node in another PA step.

IV. SOLUTION OF THE RATE EQUATIONS

Using the above for $R(k_i,p)$, the rate equation for n_i reads

$$\frac{\partial n_i}{\partial t} = m_{PA}\frac{k_i}{2mt}m_{TF} + m_{PA}\sum_{n\in\Omega}\frac{k_n}{2mt}\frac{1}{k_n}m_{TF} + m_{PA}\frac{k_i}{2mt} \times (m_{PA}-1)\sum_{n\in\Omega}\frac{k_n}{2mt}. \quad (5)$$

The first term in the sum gives the increase in n_i by mechanism (a). m_{PA} is the number of PA steps attempted for each new node (recall that per time unit one new node is added). $k_i/(2mt)$ is the preferential attachment probability to node i; m_{TF} is the expected number of triad formation steps that take place on the average after a single PA step. Given this, we have that $m_{PA}+m_{PA}m_{TF}=m$. Again, it should be noted that n_i and all quantities are expectation values, and can only be compared to simulations if an ensemble average is performed.

The second term describes mechanism (b); in this term, the sum goes over all neighbors Ω of i, and their degrees are denoted by k_n. $1/k_n$ comes from the fact that the neighboring node where a TF step links is chosen uniformly from the neighbors. We exclude here secondary triangle formation that takes place if two TF steps from the new node form a triangle with i and one of i's neighbors, which becomes more relevant for large p's. The term for (b) gives the same expression as (a) after simplification.

The last term belongs to (c) and it is the only one that would remain if we considered the simple BA model. It is the product of the probabilities of linking to node i and to one of the neighbors of i, respectively, using only PA steps. The term contains the sum of the degrees of neighboring nodes; this is k_i times the average degree of the neighbors. It has been shown that for uncorrelated random BA networks [17]

$$\langle k_n\rangle = \frac{\sum_{n\in\Omega}k_n}{k_i} = \frac{\langle k\rangle}{4}\ln t = \frac{m}{2}\ln t. \quad (6)$$

In this model the numerical result follows the same scaling not only for $p\ll 1$ but for p general.

Finally, we approximate n_i at the end of the network growth by going over from discrete to continuous variables and integrating both sides in Eq. (5). The integral for term (a) or (b) is simply

$$\int_1^N m_{PA}\frac{k_i}{2mt}m_{TF}dt = \frac{m_{PA}m_{TF}}{m}\int_1^N\frac{dk_i}{dt}dt = \frac{m_{PA}m_{TF}}{m}[k_i(N)-m] \approx \frac{m_{PA}m_{TF}}{m}k_i(N), \quad (7)$$

where we made use of the fact that $\partial_t k_i = k_i/(2t)$ [14]. From this, it also follows that $k_i(t)=m(t/t_i)^{1/2}$, where t_i is the time at which node i was introduced [4]. Thus integrating (c) gives

$$\int_1^N m_{PA}\frac{k_i}{2mt}(m_{PA}-1)\sum_{n\in\Omega}\frac{k_n}{2mt}dt = \frac{m_{PA}(m_{PA}-1)}{4m^2}\int_1^N\frac{k_i^2}{t^2}\frac{m}{2}\ln t dt = \frac{mm_{PA}(m_{PA}-1)}{8t_i}\left[\frac{(\ln t)^2}{2}\right]_1^N = \frac{m_{PA}(m_{PA}-1)}{16m}\frac{(\ln N)^2}{N}k_i^2(N), \quad (8)$$

with $k_i(t)$ being substituted where needed. Combining this with Eq. (7) yields

$$n_i = n_{i,0} + \frac{2m_{PA}m_{TF}}{m}k_i + \frac{m_{PA}(m_{PA}-1)}{16m}\frac{(\ln N)^2}{N}k_i^2. \quad (9)$$

The clustering coefficient for nodes with degree k becomes

$$C(k) = \frac{n}{k(k-1)/2} \approx \frac{4m_{TF}}{k} + \frac{m-1}{8}\frac{(\ln N)^2}{N}, \quad (10)$$

after neglecting $n_{i,0}$ and approximating m_{PA} by m, which is reasonable when the triad formation probability is small. It is not surprising that the constant offset in the expression of C is for $p\to 0$ exactly the constant clustering coefficient of pure BA graphs. The first term, more importantly, can be attributed to the triad formation induced clustering, and shows the $1/k$ behavior typical of many real networks and other models [9,15,18]. $C(k)$ is composed of a power law and a constant, so perfect power-law behavior follows only when the former one dominates. In the opposite case an effective exponent will be less than 1. Furthermore, since $n_{i,0}$ has been neglected, Eq. (10) and the inverse proportionality apply to nodes with k_i large enough, only.

For further progress, m_{TF}, the expected number of links created in the several possible TF steps after a PA step for a particular node, needs to be approximated. Take $m-1$ edges to be available for successive TF steps (this is an upper limit) and assume that node i is not saturated yet as far as the connections to the neighbors are concerned. This gives $m_{TF} = \sum_{z=1}^{m-2} z p^z(1-p) + (m-1)p^{m-1} \approx p$ for p small.

The fact that the local clustering coefficient contains a constant term means that there is a crossover at a certain k^*. At this point, a power law turns over to a constant clustering coefficient. k^* can be estimated by taking the two terms in Eq. (10) to be equal:

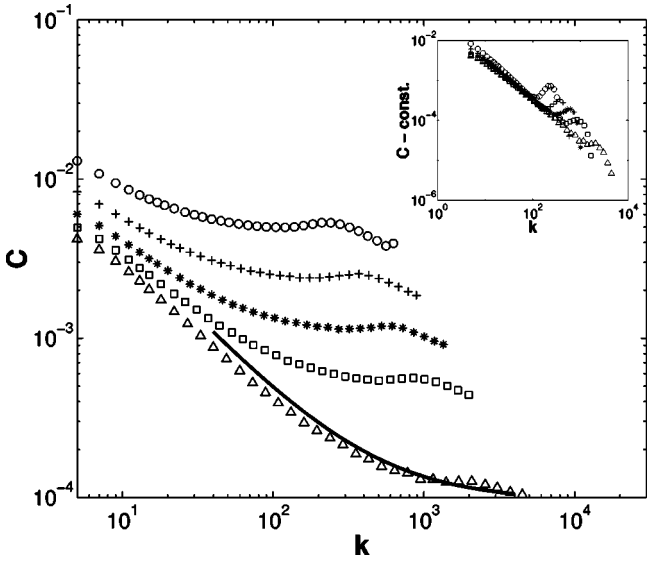

FIG. 2. Clustering coefficient as a function of the node degree for $m=5$ and different sizes (10^4 for ○, 25 119 for +, 63 096 for *, 158 489 for □, and 10^6 for △). The triad formation probability is uniformly $p=0.01$. The bold line is the prediction given for the largest system, $C(k)\approx 0.04k^{-1}+9.5\times 10^{-5}$. The crossover degree from Eq. (11) is $k^*\approx 400$. The inset shows the data collapse of the power-law part of $C(k)$.

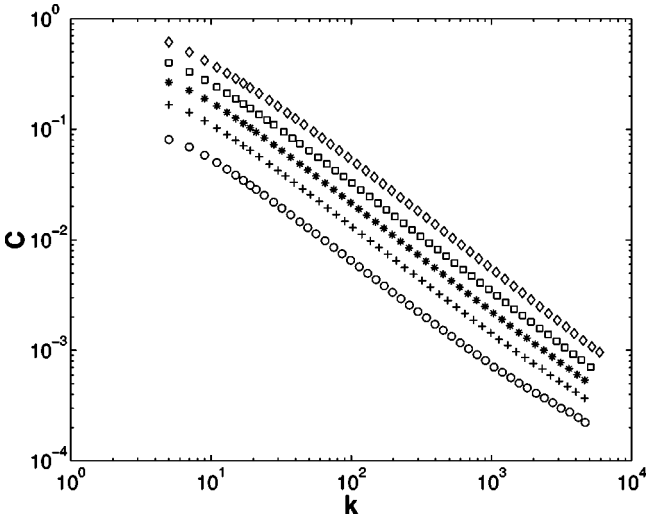

FIG. 3. Clustering coefficient for networks of 10^6 nodes and $m=5$; the triad formation probability is $p=0.2$, 0.4, 0.6, 0.8, and 1, for ○, +, *, □, and ◇, respectively. The curves descend with an exponent of -1, invariably, thus ensuring a good qualitative match to Eq. (10). The data have been logarithmically binned and the lack of fluctuations indicates a uniform behavior even at large degrees.

$$k^*\approx \frac{32}{m(\ln N)^2}pN. \tag{11}$$

Thus we can conclude that the exponents of Eq. (3) are $\gamma=1/\alpha=1$ and $\delta=1$ for the triad formation model, and from above, $\alpha=1$.

V. SIMULATIONS

Simulations of the model consistently confirm the analytical results obtained from the rate equation. In Fig. 2 networks of different sizes are shown to undergo such a transition to constant clustering by tuning p so that k^* is smaller than the maximum degree in the networks. The peaks that are visible in the inset at large degrees, especially when the systems are small, come from the initial network core that is chosen to be a fully connected graph of size $m+1$. This has a large clustering coefficient for each node that remains highly connected even after a long time. The inset of Fig. 2 has been obtained by subtracting the expected value of the k-independent term of Eq. (10) from the data, thus revealing how the $1/k$ behavior universally emerges.

A similar phenomenon to the transition described above can be observed in the case of the actor network of the Internet Movie Database [9], where the tail of a decreasing power law becomes constant, although large fluctuations naturally affect this part of the statistics. Figure 3 shows networks well below the transition and thus almost only the power-law part is conceivable.

It is not unusual in the physics of scale-free networks that mean-field approaches work well [1]. This fact is related to the strongly hierarchical nature of the networks grown by preferential attachment and our study demonstrates that this situation remains unaltered even when considering a mechanism that enhances clustering. The agreement between the $1/k^\alpha$ dependence with $\alpha=1$ obtained in Eq. (10) and that found in real networks indicates that the same ''mean-field'' mechanisms of clustering are operative. For PA growth with enhanced clustering the simplest interpretation is that for each new link a node i gains from a new node introduced to the network, its neighbors (''friends'') have also a constant probability to be linked to the same new one. This is in fact exactly the Holme-Kim model, and just expresses the fact that as $C_i\approx n_i/k_i^2$, to get $\alpha=1$ one needs $n_i\sim k_i$.

VI. SUMMARY

It is interesting to ask how robust the mean-field exponent is and what are the limits of the above approach, especially in the light of the recently discovered networks with $\alpha\neq 1$ [19]. The rate equations allow to discuss the ways how exponents like such can emerge. Equation (4) implies that the clustering is crucially dependent on the properties of the nodes in the neighborhood, Ω. If, say, correlations from ''assortative'' or ''disassortative'' mixing arise between k_i and the average degree $\langle k_n\rangle$ ($n\in\Omega$) [20], this may either enhance ($\alpha<1$) or inhibit ($\alpha>1$) clustering from the mean-field result. On the level of models, one can envision changing the k and the p dependence of the rates. The second possibility is fluctuation effects that limit the validity of the rate-equation theory. It would seem interesting to explore both these issues.

In conclusion, we have formulated a scaling assumption and a mean-field theory of the clustering of scale-free net-

works. A specific example, the triad formation model [14] has been solved and comparisons to the simulations indicate both good agreement and yield the MF value of the exponent α. This approach should be amenable to many of the models in the literature, and it should help to understand the origins of clustering, in particular, for $\alpha \neq 1$ and with respect to other statistical aspects than the $C(k)$ distribution, only. In particular, it might be possible to compute, e.g., the probability distribution of C with k fixed, and not only the average. We have here considered only growing networks, but obviously the rate equations can be written down also in the case the structural dynamics allows for deleting edges, as well [21,22].

ACKNOWLEDGMENTS

J.K. and G.S. are grateful for the warm hospitality at HUT and for partial support by OTKA under Grant No. T029985. They also thank the Center for Applied Mathematics and Computational Physics of the BUT. This work has been supported by the Academy of Finland's Center of Excellence Program and by the Center for International Mobility (CIMO).

[1] A.-L. Barabási, *Linked, The New Science of Networks* (Perseus Publishing, Cambridge, MA, 2002); D.J. Watts, *Small Worlds* (Princeton University Press, Princeton, NJ, 2000); S.N. Dorogovtsev and F.F. Mendes, Adv. Phys. **51**, 1079 (2002); R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[2] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabási, Nature (London) **407**, 651 (2000).

[3] H. Jeong, S. Mason, A.-L. Barabási, and Z.N. Oltvai, Nature (London) **411**, 41 (2001).

[4] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[5] M.E.J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).

[6] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, and Y. Åberg, Nature (London) **411**, 907 (2001).

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).

[8] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).

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

[10] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001); A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).

[11] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **88**, 108701 (2002).

[12] Of course, the regular lattice has to have high clustering, e.g., by higher neighbor connections.

[13] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).

[14] P. Holme and B.J. Kim, Phys. Rev. E **65**, 026107 (2002).

[15] K. Klemm and V.M. Eguíluz, Phys. Rev. E **65**, 057102 (2002).

[16] Note that there is a factor of $m-1$, in contrast to the original result of m in Ref. [15].

[17] V.M. Eguíluz and K. Klemm, Phys. Rev. Lett. **89**, 108701 (2002).

[18] J.-P. Eckmann and E. Moses, Proc. Natl. Acad. Sci. U.S.A. **99**, 5825 (2002).

[19] A. Capocci, G. Caldarelli, and P. De Los Rios, e-print cond-mat/0206336.

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

[21] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz, e-print cond-mat/0207366.

[22] This kind of dynamics is met in models that describe protein interaction networks. See, e.g., J. Berg, M. Lässig, and A. Wagner, e-print cond-mat/0207711; R.V. Solé, R. Pastor-Satorras, E.D. Smith, and T.B. Kepler, Adv. Complex Syst. **5**, 43 (2002), and the references therein.