Growing scale-free networks with tunable clustering

Petter Holme* and Beom Jun Kim†
Department of Theoretical Physics, Umeå University, 901 87 Umeå, Sweden
(Received 21 September 2001; published 11 January 2002)

We extend the standard scale-free network model to include a "triad formation step." We analyze the geometric properties of networks generated by this algorithm both analytically and by numerical calculations, and find that our model possesses the same characteristics as the standard scale-free networks such as the power-law degree distribution and the small average geodesic length, but with the high clustering at the same time. In our model, the clustering coefficient is also shown to be tunable simply by changing a control parameter—the average number of triad formation trials per time step.

DOI: 10.1103/PhysRevE.65.026107 PACS number(s): 89.75.Fb, 89.75.Hc, 89.65.−s

A great number of systems in many branches of science can be modeled as large sparse graphs, sharing many geometrical properties [1]. For example: social networks, computer networks, and metabolic networks of certain organisms all have a logarithmically growing average geodesic (shortest path) length ℓ and an approximately algebraically decaying distribution of vertex degree. In addition to this, social networks typically show a high clustering, or local transitivity: If person *A* knows *B* and *C*, then *B* and *C* are likely to know each other.

Works on the geometry of social networks, which is the main focus of the present paper, have originated from Rapoport's studies of disease spreading [2], and have been further developed in Refs. [3,4]. General mathematical models for random graphs with a structural bias are called the Markov graphs and were studied in Ref. [5]. In the physics literature, networks with high clustering are commonly modeled by the small-world network model of Watts and Strogatz (WS) [6], while networks with the power-law degree distribution by the scale-free network model of Barabási and Albert (BA) [7]. Although both models have a logarithmically increasing ℓ with the network size, each model lacks the property of the other model: the WS model shows a high clustering but without the power-law degree distribution, while the BA model with the scale-free nature does not possess the high clustering. In this work, we propose a network model that has *both* the perfect power-law degree distribution *and* the high clustering. Furthermore, in our model, the degree of the clustering, measured by the clustering coefficient (see below), is shown to be tunable and thus controllable by adjusting a parameter of the model.

We start from the definition of a network as a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges [8]. An edge connects pairs of vertices in $\mathcal{V}$ and not more than one edge may connect a specific pair of vertices. To quantify the clustering, Watts and Strogatz introduced the clustering coefficient $\gamma\equiv\langle\gamma_v\rangle$ with the average $\langle\cdots\rangle$ for all vertices in $\mathcal{V}$. The local clustering coefficient γ_v for the vertex v is defined as follows: Suppose that the vertex v has k_v neighbors (k_v is called the degree of the vertex v, a neighbor is a vertex separated by exactly one edge). For those k_v neighbors, there can exist at most

$$\binom{k_v}{2}=k_v(k_v-1)/2$$

edges connecting two of k_v vertices. If one defines $|\mathcal{E}(\Gamma_v)|$ as the number of actual edges existing in the network connecting those neighbors, the local clustering coefficient is written as [6]

$$\gamma_v\equiv\frac{|\mathcal{E}(\Gamma_v)|}{\binom{k_v}{2}}. \tag{1}$$

From the above definition, it is clear that γ is a measure of the relative number of triads (fully connected subgraphs of three vertices). Note also that γ is strictly in the interval [0, 1] with the upper limit attained only for a fully connected graph. In a social acquaintance network, for example, $\gamma=1$ if everyone in the network knows each other. It should be noted that even though the BA model successfully explains the scale-free nature of many networks, it has $\gamma\approx 0$ and thus fails to describe correctly networks with the high clustering, such as social networks.

We below review briefly the BA model of the scale-free network and present our model for the scale-free network with the high clustering. The BA model [7] is defined as follows:

(i) Initial condition: To start with, the network consists of m_0 vertices and no edges.

(ii) Growth: One vertex v with m edges is added at every time step. Time t is identified as the number of time steps.

(iii) Preferential attachment (PA): Each edge of v is then attached to an existing vertex with the probability proportional to its degree, i.e., the probability for a vertex w to be attached to v is [15]

$$P_w=\frac{k_w}{\sum_{v\in\mathcal{V}}k_v}. \tag{2}$$

*Electronic address: holme@tp.umu.se
†Electronic address: kim@tp.umu.se

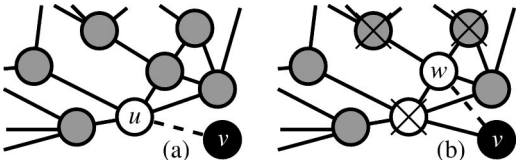

FIG. 1. *Preferential attachment and triad formation.* In the preferential attachment step (a) the new vertex v chooses a vertex u to attach to with a probability proportional to its degree. In the triad formation step (b) the new vertex v chooses a vertex w in the neighborhood of the one linked to in the previous preferential attachment step. $\times$ symbolizes "not allowed to attach to" (either since no triad would be formed, or that an edge already exists).

In the BA model, the growth step is then iterated $N=|\mathcal{V}|$ times, and for each growth step the PA step is iterated m times for m edges of the newly added vertex v.

In order to incorporate the high clustering we modify the above BA algorithm by adding an additional step: Triad formation (TF): If an edge between v and w was added in the previous PA step, then add one more edge from v to a randomly chosen neighbor of w. If there remains no pair to connect, i.e., if all neighbors of w were already connected to v, do a PA step instead.

When a vertex v with m edges is added to the existing network, we first perform one PA step, and then perform a TF step with the probability P_t or a PA step with the probability $1-P_t$. The average number m_t of the TF trials per added vertex is then given by $m_t=(m-1)P_t$, which we take as the control parameter in our model (see Fig. 1). It should be noted that our model reduces to the original BA model when $m_t=0$.

The standard scale-free network model not only generates networks with certain geometrical properties, it suggests a mechanism for the emergence of power-law degree distributions in evolving networks: New actors (vertices) in a social context prefers to attach to more connected ("well known") actors. The sociological interpretation for the triad formation step is that after being acquainted with (linked to) w an actor v is likely to be acquainted to w's acquaintances as well. This mechanism of the emergence of clustering is well known, and was discussed under the name "sibling bias" already in Ref. [4]. Recently, Ref. [9] provided empirical evidence for both the mechanisms of triad formation and preferential attachment used in our construction algorithm.

The clustered scale-free network algorithm defined above gives the same degree distribution as the standard scale-free network, at least if every TF step follows a PA step. To see this, first observe that in a PA step an arbitrary vertex v increases its degree with the rate

$$\frac{\Delta k_v}{\Delta t}=A\frac{k_v}{\sum_{w\in\mathcal{V}}k_w} \quad \text{for a PA step,} \tag{3}$$

where the normalization factor A for one edge is determined to be unity following Ref. [7]. For a TF step the average increase of k_v is proportional to the probability that a vertex

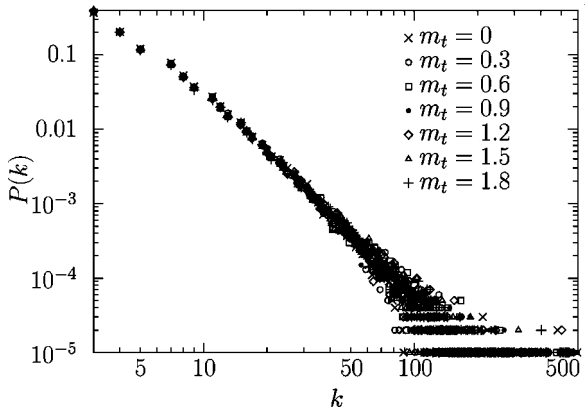

FIG. 2. Degree distribution for the scale-free network model with tunable clustering with parameter values $m=m_0=3$, $N=10^5$ at various values of m_t: At any value of m_t, which determines the average number of triad formations, $P(k)$ exhibits a power-law behavior like the BA model corresponding to $m_t=0$.

in the neighborhood w is linked in the PA step before, times the inverse of that vertex's degree (the probability that v is linked from w):

$$\frac{\Delta k_v}{\Delta t}=\frac{\sum_{w\in}\Gamma_v k_w(1/k_w)}{\sum_{w\in\mathcal{V}}k_w}=\frac{k_v}{\sum_{w\in\mathcal{V}}k_w} \quad \text{for a TF step,} \tag{4}$$

where we have used the same normalization as in Eq. (3) and Γ_v is the neighborhood of v (we use that the number of vertices in Γ_v is k_v). From Eqs. (3) and (4) the total rate for one time step, composed of m_t TF steps and $m-m_t$ PA steps, is expressed as

$$\frac{\Delta k_v}{\Delta t}=m_t\left(\frac{k_v}{\sum_{w\in\mathcal{V}}k_w}\right)+(m-m_t)\left(\frac{k_v}{\sum_{w\in\mathcal{V}}k_w}\right)=\frac{k_v}{2t}, \tag{5}$$

which has the same form as the original BA model and thus results in

$$k_v\propto t^{1/2}. \tag{6}$$

Consequently, the degree of an arbitrary vertex increases as the square root of the time, which then yield the power-law degree distribution: $P(k)\sim k^{-3}$ [7].

In the above discussion, we have assumed that a TF step always follows a PA step. If a TF step would be proceeded by another TF step the factor $k_w(1/k_w)$ in Eq. (4) would be replaced by $k_w[1/(k_w-1)]$ that is a small correction when k_w is large (which it is likely to be by the definition of the PA step). And thus the resulting degree distribution would not differ much from a power law. In Fig. 2, the degree distributions $P(k)$ at various values of m_t are displayed and we find that at any value of m_t, the distribution is well described by the power law with the exponent $a\approx 3$ in $P(k)\sim k^{-a}$, as is expected from the above analytic consideration.

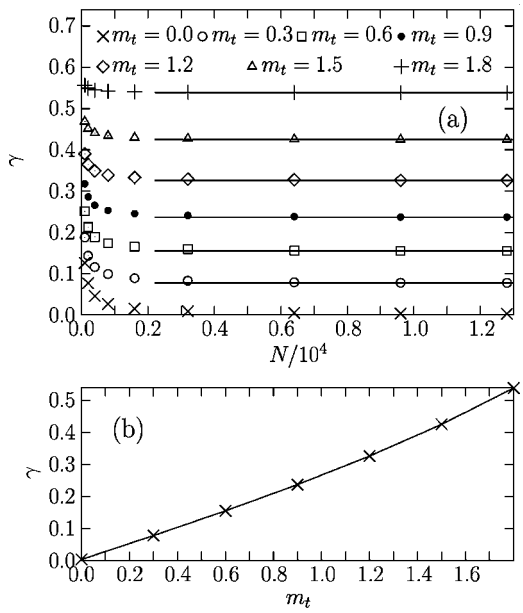

FIG. 3. (a) Clustering coefficient γ vs the network size N at various values of the average number m_t of triads per time step. Straight lines show asymptotic values of γ at each m_t. For $m_t \neq 0$, γ approaches a nonzero value as N is increased. (b) $\gamma(N \rightarrow \infty)$ vs m_t: The clustering coefficient can be varied systematically by changing m_t.

The parameter m_t in our model introduces the clustering effect into the system by allowing the formation of triads. We only focus on the case of $m=3$ with expectation that other values of m should give qualitatively the same behavior. One expects then that for any m_0 a finite m_t gives a finite clustering coefficient γ in the thermodynamic limit of $N \rightarrow \infty$, whereas for $m_t=0$ (the BA scale-free network model) γ goes to zero as N becomes larger. In Fig. 3(a), γ at various values of m_t is shown as a function of system size N. As expected, we find that γ approaches to a finite nonzero value as N is increased at nonzero m_t, whereas the BA model, which corresponds to the limiting case of $m_t=0$ in our model, is confirmed to have $\gamma=0$. Furthermore, we also observe that the relation between m_t and γ is almost linear, as depicted in Fig. 3(b).

From the above observations, we conclude that our model exhibits both the scale-free nature and the high clustering at the same time, while the WS model (the BA model) lacks the former (the latter) property. We note that in many real networks, both properties usually coexist, and thus believe that our model is more realistic. The triad formation step in our model, which inevitably gives a high clustering coefficient, is expected to make the average geodesic length larger than in a BA network, since the edge for the triad could have been used to connect two vertices separated by a large distance if only the preferential attachment step was allowed. However, the characteristic path length, defined as the average of the geodesic length, ℓ, is found to behave logarithmically with the size N, the same behavior as the WS model and the BA model. In Fig. 4, we present ℓ vs N at various values of m_t.

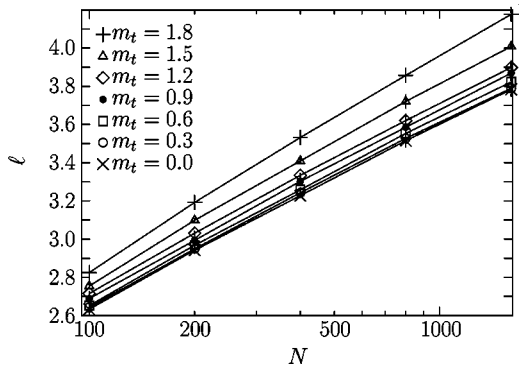

FIG. 4. The characteristic path length for the arbitrary clustered scale-free network model with the parameters $m=m_0=3$ and at various values of m_t. Although ℓ becomes larger with m_t, ℓ is found to behave logarithmically as a function of N.

It is shown that ℓ becomes larger as m_t is increased, as expected. Furthermore Fig. 4 shows that the increase of ℓ is logarithmic for all m_t.

By mimicking principles in network formation, a generation algorithm can construct graphs with certain topological statistics, such as a degree distribution, clustering coefficient, and so on. However, it should be emphasized that these kinds of algorithms cannot claim to uniformly sample the ensemble of networks with specific statistical properties. This drawback exists even in more general classes of random graphs where structural biases, such as clustering, are imposed [5,10].

Recently, Klemm and Equíluz [11] have proposed a network model based on a finite memory of vertices, i.e., vertices become inactive and do not get new edges after a finite number of time steps, and have shown that their growth and deactivation model exhibits both the high clustering and the scale-free nature. Our model provides an alternative possibility to achieve the same feature, the clustered scale-free nature, based on our frequent everyday experience on how we are acquainted by newcomers: *B* becomes *A*'s new friend since *B* is introduced by one of *A*'s friends. Even in the network of scientific citations, it is likely that authors of paper *A* refer paper *B* since they have found *B* when they read a famous review paper *C* [12]. This then has close resemblance to our model, the TF step accompanied by the PA step. In Ref. [13], a model with both the high clustering and the scale-free distribution has also been suggested. However, the power-law degree distribution was assigned to the network to start with, and the next following steps were devised not to change the degree at each vertex. In other words, the power-law distribution in Ref. [13] was not an emerging property in the model, which is different from the BA model as well as our model in this work. Very recently, we have learned about the work by Davidsen, Ebel, and Bornholdt [14], which is based on the same observation of triad formation as ours and has been shown to possess similar network properties, i.e., the high clustering, small average geodesic length, and a scale-free distribution. We believe, however, that our model has some advantage in describing networks that grow in time, whereas the network model in Ref. [14] has fixed network size.

In conclusion, we have proposed an algorithm for generation of growing networks with power-law degree distribution, a logarithmic increase of the average geodesic length, and a finite clustering. The last two properties make the generated graphs qualify as a small-world network in the Watts and Strogatz sense, in addition to their scale freeness. The simple relation between the coefficient m_t and γ further increases the usefulness of the suggested algorithm, making it possible to tune the clustering coefficient in a systematic way.

Note added in proof: We have become aware of another model generating a clustered scale-free network [16]. This model is similar to the special case $m=m_0=2$, $m_t=1$ of our model, with the initial PA step replaced by a purely random selection of vertices.

This work was supported in part by the Swedish Natural Research Council through Contract Nos. F 5102-659/2001 and E 5106-1643/1999.

[1] For reviews, see e.g., D. J. Watts, *Small Worlds* (Princeton University Press, Princeton, 1999); M. E. J. Newman, J. Stat. Phys. **101**, 819 (2000); S. H. Strogatz, Nature (London) **410**, 268 (2001); S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. (to be published).

[2] A. Rapoport, Bull. Math. Biophys. **15**, 523 (1953); 535 (2001).

[3] A. Rapoport, Bull. Math. Biophys. **19**, 257 (1957); T. J. Fararo and M. Sunshine, *A Study of a Biased Friendship Net* (Syracuse University Press, Syracuse, 1964); J. A. Davis, Human Relations **20**, 181 (1967); I. Pool and M. Kochen, Soc. Networks **1**, 1 (1978); J. Skvoretz, *ibid.* **7**, 225 (1985).

[4] C. C. Foster, A. Rapoport, and C. J. Orwant, Behav. Sci. **8**, 56 (1963).

[5] O. Frank and D. Strauss, J. Am. Stat. Assoc. **81**, 832 (1986).

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

[7] A.-L. Barabási and R. Albert, Science **286**, 509 (1999); A.-L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999); R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).

[8] For an introduction to graph theory, see for example: R. J. Wilson, *Introduction to Graph Theory* (Oliver & Boyd, Edinburgh, 1972); G. Chartrand and L. Lesniak, *Graphs and Digraphs* (Weber and Smith, Boston, 1986).

[9] M. E. J. Newman, Phys. Rev. E **64**, 025102 (2001).

[10] D. Strauss, SIAM Rev. **28**, 513 (1986).

[11] K. Klemm and V. M. Equíluz, e-print cond-mat/0107606; e-print cond-mat/0107607.

[12] See the following recent studies, and references therein: S. Redner, Eur. Phys. J. B **4**, 131 (1998); S. Bilke and C. Peterson, Phys. Rev. E **64**, 036106 (2001); Y. Fang and R. Rosseau, Scientometrics **50**, 273 (2001).

[13] A. R. Puniyani, R. M. Lukose, and B. A. Huberman, e-print cond-mat/0107212.

[14] J. Davidsen, H. Ebel, and S. Bornholdt, cond-mat/0108302.

[15] In practice one may use $P_w=(k_w+1)/\Sigma_{v\in\nu}(k_v+1)$, to make it possible for disconnected vertices to be connected. This does not change the resultant network geometry significantly.

[16] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukin, Phys. Rev. E **63**, 062101 (2001).