Correlations in connected random graphs

Piotr Bialas[1,2,*] and Andrzej K. Oleś[1,†]
[1]*Marian Smoluchowski Institute of Physics, Jagellonian University, Reymonta 4, 30-059 Krakow, Poland*
[2]*Mark Kac Complex Systems Research Centre, Faculty of Physics, Astronomy and Applied Computer Science, Jagellonian University, Reymonta 4, 30-059 Krakow, Poland*
(Received 28 October 2007; published 27 March 2008)

We study the properties of the giant connected component in random graphs with arbitrary degree distribution. We concentrate on the degree-degree correlations. We show that the adjoining nodes in the giant connected component are correlated and derive analytic formulas for the joint nearest-neighbor degree probability distribution. Using those results we describe correlations in maximal entropy connected random graphs. We show that connected graphs are disassortative and that correlations are strongly related to the presence of one-degree nodes (leaves). We propose an efficient algorithm for generating connected random graphs. We illustrate our results with several examples.

DOI: 10.1103/PhysRevE.77.036124 PACS number(s): 89.75.Hc, 05.10.−a, 05.90.+m

I. INTRODUCTION

In the last decade or so, there has been a great increase of interest in the theory of random graphs and networks (in the following we will use those two terms interchangeably). While in principle this is a branch of mathematics, much of this effort was fueled by the availability of "experimental" data on real graphs (see [1] for review). These data are compared to the predictions of various random graphs models. Probably the best known and simplest example of such reference models is the ensemble of all labeled graphs with V vertices and L links (without multiple- and self-links), chosen with uniform probability. We will call this model Erdös-Rényi (ER) graphs after the authors, who were the first to introduce and study them [2].

The ER ensemble is the simplest example of the so-called "maximally random" graphs. Intuitively those are the ensembles where the distributions of vertices and links joining them are "as random as possible" for a given set of constraints. In the case of ER graphs the only constraints are the fixed number of links and vertices. The "maximal randomness" can be formalized using the notion of *entropy* (see Sec. II). The maximally random ensembles serve as null hypothesis. For example, it was the deviation of data collected on the World Wide Web (WWW) graph from the predictions of the ER model that triggered the interest in random networks [3], because it implied that those graphs were not created just by joining vertices at random, but required the existence of another mechanism [4].

A popular generalization of the ER ensemble are graphs with a given degree distribution (degree of a node is the number of links attached to it) [5–10]. One feature of those ensembles is the absence of correlations between neighboring nodes' degrees, at least for degree distributions without heavy tails (see the discussion in Sec. IV C). The object of our study was to find what happens when we constrain to connected graphs only. A simple argument indicated that correlations would appear: a neighbor of a node with degree one (leaf) must have its degree greater than 1; otherwise, they would form a separate connected component. Similarly, all neighbors of a node cannot have their degree equal to 1, as such a "hedgehog" would also form a separate connected component [11,12]. This obviously leads to correlations. It is not clear, however, how strong they are and if they survive the large-V limit. We have already studied those correlations numerically in Ref. [12] and found that they also appear in large graphs. In this paper we derive the analytic formulas describing them. We also found a strong indication that the described mechanism is the only one responsible for the correlation in maximally random connected graphs: when we forbid vertices with degree 1 correlations disappear.

Connectivity is a nonlocal constraint hard to deal with. To study the properties of connected graphs we use another feature of maximally random graphs with a given degree distribution: the appearance of a connected component that includes a finite fraction of all the vertices (and links). From the properties of this giant connected component we can infer the properties of connected graphs.

The paper is organized as follows: Section II introduces some basic definitions concerning random graphs. In Sec. III we present the method of generating functions used to study the properties of the giant connected component in random graphs with arbitrary degree distribution [6]. Then we calculate degree-degree correlations in the giant component. Section IV contains some examples where we compare our predictions with the results of Monte Carlo (MC) simulations. Finally, we show in Sec. V how to relate connected random graphs to giant connected components in other ensembles. In Sec. VI we address the situation when correlations in random graphs are suppressed by the absence of vertices with degree one (leaves). The paper is summarized in Sec. VII.

II. RANDOM GRAPHS

A. Average degree

Formally we consider random graphs as an ensemble of graphs $\mathcal{G}$ with probability $P(G)$ assigned to every graph

*pbialas@th.if.uj.edu.pl
†oles@th.if.uj.edu.pl

$G \in \mathcal{G}$. Using this definition we introduce the entropy of the ensemble:

$$S = -\sum_{G\in\mathcal{G}} P(G)\ln P(G). \tag{1}$$

The maximally random ensembles described in the previous section are those which for given constraints have maximal entropy.

Denoting by $O(G)$ some property of graph G we can calculate its average over the whole ensemble:

$$\langle O\rangle_G = \sum_{G\in\mathcal{G}} O(G)P(G). \tag{2}$$

The most widely studied example is the probability distribution of node degrees:

$$p_k = \left\langle \frac{n_k(G)}{V(G)} \right\rangle_G, \tag{3}$$

where $n_k(G)$ is the number of vertices with degree k and $V(G)$ is the total number of vertices in graph G (in the following we will often omit the argument G). The mean of this distribution is the "link density,"

$$\langle k\rangle = \sum_k k p_k = \left\langle \frac{2L(G)}{V(G)} \right\rangle_G \equiv z, \tag{4}$$

because $\Sigma_k k n_k = 2L(G)$; by $L(G)$, we denote the number of links in graph G.

However, what is frequently observed is not an average (2), but the properties of a single graph (e.g., WWW). That is why we are actually interested in the probability that our model will produce a graph with those properties. It is described by the distribution

$$P(O) = \sum_{G\in\mathcal{G}} \delta(O - O(G))P(G). \tag{5}$$

In many cases this distribution is sufficiently well characterized by its mean (2) with relative fluctuations disappearing in the large-V limit. In this situation we will say the O is self-averaging. In such a case one can infer the properties of the whole ensemble from the properties of just one large graph. We want to emphasize, however, that this is only an assumption that has to be checked for each particular model (see [13] for a discussion of self-averaging in real graphs).

In Appendix A we show for illustration a definition of a non-self-averaging ensemble. Although this is an artificial example, let it serve as a warning. In this paper we assume that our models are self-averaging without any further formal proofs.

We end with the following comment: As in the self-averaging ensemble fluctuations do not matter, in the large-volume limit we have

$$p_k = \left\langle \frac{n_k(G)}{V(G)} \right\rangle_G \sim \frac{\langle n_k(G)\rangle_G}{\langle V(G)\rangle_G}. \tag{6}$$

We will use this kind of approximations in the following sections.

B. Correlations

The distribution p_k does not give any information about the correlations between vertices. An obvious generalization is the joint distribution $p_{q,r}$ which describes the probability that a pair of nearest neighbors (NNs) has degrees q and r (we assume that we pick a pair of NNs with uniform probability):

$$p_{q,r} = \left\langle \frac{n_{q,r}}{2L} \right\rangle, \tag{7}$$

where $n_{q,r}$ is the number of links with their start point having degree q and end point having degree r. Note that we treat each undirected link as two directed links. On an undirected graph,

$$n_{q,r} = n_{r,q}, \quad \sum_{q,r} n_{q,r} = 2L \quad \text{and} \quad \sum_r n_{q,r} = q n_q. \tag{8}$$

If vertex degrees are independent, the probability (7) should factorize:

$$p_{q,r} = \tilde{p}_q \tilde{p}_r, \quad \tilde{p}_q = \sum_r p_{q,r}, \tag{9}$$

leading to the relation

$$\left\langle \frac{n_{q,r}}{2L} \right\rangle = qr \left\langle \frac{n_q}{2L} \right\rangle \left\langle \frac{n_r}{2L} \right\rangle. \tag{10}$$

One should, however, keep in mind that this defines the absence of correlations in the *ensemble* of graphs. A more appropriate question could be, are the vertices on individual graphs uncorrelated (see previous section)? The condition for absence of correlations between vertices in each individual graph G is

$$\frac{n_{q,r}(G)}{2L(G)} = qr \frac{n_q(G)}{2L(G)} \frac{n_r(G)}{2L(G)} \tag{11}$$

or, after averaging,

$$\left\langle \frac{n_{q,r}}{2L} \right\rangle = qr \left\langle \frac{n_q}{2L} \frac{n_r}{2L} \right\rangle. \tag{12}$$

As already pointed out, for a large class of ensembles conditions (10) and (12) are equivalent in the large-volume limit. However, it is easy to check that for the non-self-averaging ensemble in Appendix A vertices on each individual graph are uncorrelated according to the condition (12), but correlated according to (10). Again, we leave this as a warning and proceed further with the assumption that our models are self-averaging and that those two conditions are equivalent.

In practice, checking the condition (9) is difficult as it entails measuring a two dimensional distribution with good accuracy. Therefore we introduce another quantity [14]

$$\bar{k}(k) = \sum_q \left\langle \frac{q n_{k,q}}{k n_k} \right\rangle. \tag{13}$$

It describes the average degree of nearest neighbors of a vertex with degree k. Obviously $\bar{k}(k)$ is defined for a given k

only if $n_k>0$. $\bar{k}(k)$ can be interpreted as the first moment of the conditional probability:

$$p(q|k)=\frac{p_{q,k}}{\tilde{p}_k}. \tag{14}$$

Assuming self-averaging,

$$\bar{k}(k)\approx\sum_q qp(q|k). \tag{15}$$

If the degrees are independent, $\bar{k}(k)$ should not depend on k and (12) implies

$$\bar{k}(k)=\sum_q q^2\left\langle\frac{n_q}{2L}\right\rangle\approx\frac{\langle k^2\rangle}{\langle k\rangle}. \tag{16}$$

When $\bar{k}(k)$ grows with k the graph is called *assortative* and when it shrinks *disassortative*.

III. CONNECTED COMPONENTS

In general, maximally random graphs with a given degree distribution do not need to be connected. However, if

$$\sum_k k(k-2)p_k>0 \tag{17}$$

(which translates into $z>1$ in the case of ER graphs), one of the connected components (called the giant connected component) will gather a finite fraction of all links and vertices [6]. This is a phenomenon akin to percolation. In Ref. [6] the size of the giant component and the size distribution of finite components were calculated. The degree distribution in the giant component $p_k^{(g)}$ was calculated in Ref. [8]. Here we generalize those results and calculate the two-point distributions $p_{q,r}^{(g)}$ and $\bar{k}^{(g)}(k)$ for the giant component.

We will use the method of generating functions introduced in [6]. The crucial observation is that the finite connected components are essentially trees. That is because a link emerging from one of the vertices in the component has the probability $\propto s/V$ of connecting back to a node from this component, where s is the size of the component. So for finite s this becomes negligible in the large-V limit.

Now let us pick a link from the graph at random. It belongs to some connected component. We will call $P_1(s)$ the probability that cutting this link will split the component into two parts, one of them finite and having size s. Stated differently, $P_1(s)$ is the probability that a randomly chosen link will lead into a finite part of size s. By the argument above this finite "half" will be a tree. Because of that, one can write down the equation for the generating function $H_1(x)=\Sigma_s P_1(s)x^s$ [6]:

$$H_1(x)=xG_1(H_1(x)), \tag{18}$$

where

$$G_1(x)=\frac{G_0'(x)}{G_0'(1)}=\frac{1}{z}G_0'(x),\quad G_0(x)=\sum_{k=0}^{\infty}p_kx^k. \tag{19}$$

We denote by u the value of $H_1(1)$:

$$u\equiv H_1(1)=\sum_s P_1(s). \tag{20}$$

When there is no giant component in the graph, all connected components are finite and are trees. This means that cutting each link will result in two finite parts; thus, $u=1$. However, when the giant component appears, then there is a nonzero probability that the chosen link will belong to this component and either cutting it will split the component into two infinite parts, or will not split it at all. As this probability is missing from $P_1(s)$ the sum (20) will be smaller the one. u is to be interpreted as the probability that a randomly chosen link is connected to a finite part on at least one side of the graph [10]. It follows that u^2 is the probability that a random link belongs to a finite component of arbitrary size.

That can be derived in a more explicit way. Let us denote by $P_{1,1}(s)$ the probability that a randomly chosen link belongs to a component of size s. Then,

$$P_{1,1}(s)=\sum_{t=0}^{s}P_1(t)P_1(s-t). \tag{21}$$

It is a *convolution* of the probability distribution $P_1(s)$ with itself, so its generating function is just $H_1^2(x)$. Then $u^2=H_1^2(1)=\Sigma_s P_{1,1}(s)$ is the probability that a link belongs to a finite connected component of arbitrary size and $1-u^2$ is the probability that it is inside the giant component.

Finally, if we denote by $P_0(s)$ the probability that a randomly chosen *vertex* belongs to a finite component of size s, we can obtain its generating function $H_0(x)$ from $H_1(x)$ [6]:

$$H_0(x)\equiv\sum_s P_0(s)x^s=xG_0(H_1(x)). \tag{22}$$

By the same arguments as above,

$$h\equiv H_0(1)=G_0(u) \tag{23}$$

is the probability that a randomly chosen vertex belongs to a finite connected component and $1-h$ is the probability that it belongs to the giant component.

It follows from (18) and (20) that u is the solution of the equation

$$u=G_1(u). \tag{24}$$

From the definition (19) it is easy to note that $u=1$ is always a solution, but when condition (17) is fulfilled the above equation has a solution smaller than 1 as well [6]. As argued, this signals the appearance of a giant component.

A. Average degree

Using the results of the previous section it is easy to derive formulas for the average degree in the giant component $z^{(g)}$ and in the rest of the graph $z^{(f)}$:

$$z^{(g)}=\left\langle\frac{2L^{(g)}}{V^{(g)}}\right\rangle=z\frac{1-u^2}{1-h}, \tag{25a}$$

$$z^{(f)}=\left\langle\frac{2L^{(f)}}{V^{(f)}}\right\rangle=z\frac{u^2}{h}. \tag{25b}$$

As we have already pointed out, the giant connected component is not a tree. The number of independent loops that it contains equals

$$L^{(g)} - V^{(g)} + 1 \approx V\left(\frac{z}{2}(1-u^2) - 1 + h\right), \tag{26}$$

and as all the remaining connected components are trees, this is also the number of loops in the whole graph.

We can also easily calculate the number of finite connected components n_{cn} knowing that they form a forest. The number of links in the forest is $L^{(f)}=V^{(f)}-n_{cn}$ which gives

$$\langle n_{cn}\rangle = \left(h - u^2\frac{z}{2}\right)V. \tag{27}$$

From that we can derive the formula for the average size of the finite connected component:

$$\langle s\rangle^{(f)} = \left\langle \frac{V^{(f)}}{n_{cn}} \right\rangle = \frac{2h}{2h - u^2 z}. \tag{28}$$

B. Degree distribution

In this section we will calculate the degree distribution $p_k^{(f)}$ in the nongiant component part of the graph. From the relation

$$p_k = (1-h)p_k^{(g)} + hp_k^{(f)}, \tag{29}$$

we automatically get the distribution $p_k^{(g)}$ in the giant component. This has already been done in [8], but we find it instructive to use the same method of generating functions as described in Sec. III. The idea is to apply it only to the graph with the giant component excluded—i.e., to the finite connected components. We will use a tilde to denote the generating functions of the sought probability:

$$\tilde{G}_0(x) = \sum_{k=0}^{\infty} p_k^{(f)} x^k, \quad \tilde{G}_1(x) = \frac{\tilde{G}_0'(x)}{\tilde{G}_0'(1)}. \tag{30}$$

Using the argument from Ref. [6] we obtain the same equations

$$\tilde{H}_1(x) = x\tilde{G}_1(\tilde{H}_1(x)), \tag{31a}$$

$$\tilde{H}_0(x) = x\tilde{G}_0(\tilde{H}_1(x)), \tag{31b}$$

for the generating functions of the probabilities $\tilde{P}_1(s)$ and $\tilde{P}_0(s)$. Here $\tilde{P}_0(s)$ is the probability that a vertex belongs to a finite component of size s provided that it belongs to a finite component and $\tilde{P}_1(s)$ is the probability that a link leads into a finite component of size s provided that it leads into a finite component. From this we can write the relations

$$P_0(s) = h\tilde{P}_0(s), \quad P_1(s) = u\tilde{P}_1(s), \tag{32}$$

which leads to

$$H_0(x) = h\tilde{H}_0(x), \tag{33a}$$

$$H_1(x) = u\tilde{H}_1(x). \tag{33b}$$

To solve Eqs. (30), (31a), (31b), (33a), and (33b) for $p_k^{(f)}$ we make an ansatz

$$p_k^{(f)} = \frac{p_k a^k}{G(a)}. \tag{34}$$

Then,

$$\tilde{G}_0(x) = \frac{G_0(xa)}{G_0(a)}, \quad \tilde{G}_1(x) = \frac{G_1(xa)}{G_1(a)}, \tag{35}$$

so that Eq. (31a) can be rewritten as

$$a\tilde{H}_1(x) = \frac{a}{G_1(a)} x G_1(\tilde{H}_1(x)a). \tag{36}$$

Comparing with (18) we see that it will be fulfilled if

$$a\tilde{H}_1(x) = H_1\left(\frac{a}{G_1(a)}x\right). \tag{37}$$

Inserting this into (33b) we get

$$aH_1(x) = uH_1\left(\frac{a}{G_1(a)}x\right), \tag{38}$$

because of Eq. (24), which can be solved by putting $a=u$.

Now we must check Eq. (33a). Using Eqs. (22), (31b), and (33b) we get

$$h\tilde{H}_0(x) = hx\tilde{G}_0(\tilde{H}_1(x)) = hx\frac{G_0(u\tilde{H}_1(x))}{G_0(u)}$$

$$= xh\frac{G_0(H_1(x))}{h} = H_0(x). \tag{39}$$

So finally,

$$p_k^{(f)} = \frac{p_k u^k}{h}. \tag{40}$$

From that and relation (29) we get the formula for the degree distribution in the giant component:

$$p_k^{(g)} = p_k\frac{(1-u^k)}{1-h}. \tag{41}$$

In the limit $u\to 1$ and $h\to 1$ this reduces to

$$p_k^{(g)} = \frac{k}{z}p_k. \tag{42}$$

In this limit the connected giant cluster is a tree. Indeed, one can check that

$$\sum_k k p_k^{(g)} = \sum_k \frac{k^2}{z} p_k = 2. \tag{43}$$

To see this we must first note that Eq. (24) has always the solution $u=1$. It becomes the only one when $G_1'(1)=1$, which is equivalent to the condition (43).

C. Correlations

To calculate $p^{(g)}_{q,r}$ we use the relation

$$n_{q,r}(G) = n^{(g)}_{q,r}(G) + n^{(f)}_{q,r}(G). \tag{44}$$

We have already assumed that vertex degrees are uncorrelated; we further assume that this is also true for the finite connected components (nongiant) part of the graph. Assuming self-averaging and using Eq. (10) for $n_{q,r}$ and $n^{(f)}_{q,r}$ we obtain

$$p^{(g)}_{q,r} = \frac{qp_q rp_r}{z^2} \frac{1}{1-u^2}\left(1 - \frac{u^q u^r}{u^2}\right) \tag{45}$$

and

$$\bar{k}^{(g)}(k) = \frac{\langle k^2\rangle}{z} \frac{1}{1-u^k}\left(1 - \frac{\langle k^2\rangle^{(f)}}{z^{(f)}} \frac{z}{\langle k^2\rangle} u^k\right). \tag{46}$$

In the derivation we have used the relation $\langle \frac{A}{B}\rangle = \frac{\langle A\rangle}{\langle B\rangle}$, which should be valid for self-averaging quantities in the large-V limit. Comparing this with formulas (10) and (16) we note that the correlations disappear in the limit $u \to 0$. In the tree limit $u \to 1$ the formulas above take the form

$$\lim_{u,h\to 1} p^{(g)}_{q,r} = (q + r - 2)\frac{1}{2}\frac{qp_q rp_r}{z^2} \tag{47}$$

and

$$\lim_{u,h\to 1} \bar{k}^{(g)}(k) = \frac{1}{zk}((k-2)\langle k^2\rangle + \langle k^3\rangle). \tag{48}$$

IV. EXAMPLES

While deriving our formulas we have made several assumptions: (i) the vertex orders are uncorrelated, (ii) the measured quantities are self-averaging, and of course (iii) all the derivations are only valid in the large-V limit. To check to what extent those assumptions are satisfied and, more importantly, to check the magnitude of the finite size effects, we have compared our predictions to the results of MC simulations of moderate-sized graphs (5000 vertices). To simulate ER graphs we used a straightforward algorithm which connects vertices at random. To generate maximally random graphs with a given distribution we used the method described in Refs. [9,15] and implemented in Ref. [16]. This method consists of generating graphs with suitably chosen one-point weights using a Metropolis-type algorithm.

A. Erdös-Rényi graphs

For ER graphs the distribution p_k is Poissonian, $p_k = e^{-z}\frac{z^k}{k!}$ and

$$G_0(x) = G_1(x) = e^{z(x-1)}. \tag{49}$$

It follows that $H_1(x) = H_0(x) \equiv H(x)$, so $h = u$ with h being the closest to one (from below) positive solution of the equation

$$h = e^{z(h-1)}. \tag{50}$$

The results for $z^{(f)}$ and $z^{(g)}$ are shown in Fig. 1. They are compared with the results of the MC simulations of ER graphs. The agreement is perfect, and there are no visible finite-size effects (error bars are smaller than the size of the points). The degree distribution can be now easily obtained from (41). The results are presented in Fig. 2. Again, the agreement is very good without any noticeable finite-size effects.

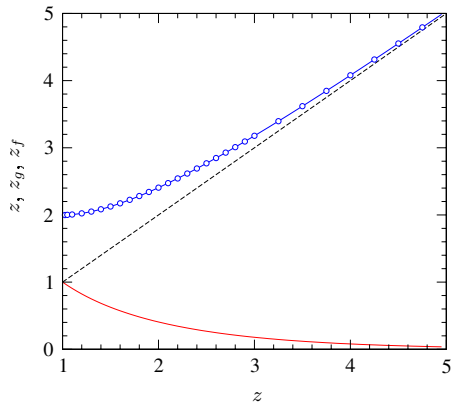

FIG. 1. (Color online) Average degree z (dashed line), average degree z_g of the connected component (upper solid line), and average degree of the rest z_f (lower solid line) as a function of z for ER graphs. Circles mark the results of MC simulations.

In this case it may be instructive to derive those results in a simpler way: when we omit the giant component from our considerations we are left with a graph with hN vertices and h^2L links on average. As there are no further restrictions, we can assume that this graph is an Erdös-Rényi graph as well. This means that its degree distribution is again Poissonian with mean $z^{(f)}$:

$$p^{(f)}_k = e^{-z_f}\frac{(z^{(f)})^k}{k!} = e^{-hz}\frac{z^k h^k}{k!}. \tag{51}$$

From the relation $h = u$ we obtain formula (40).

Finally, for $\bar{k}(k)$ we get

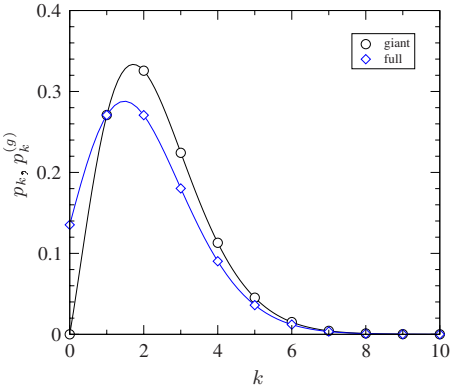

FIG. 2. (Color online) Degree distribution for ER graphs with $z = 2$. Circles mark the results of MC simulations for the giant component and diamonds for the full graph. Solid lines denote analytical solutions.

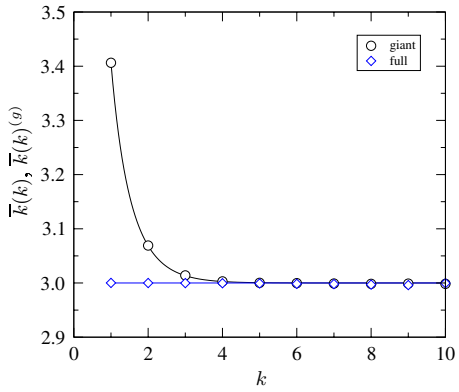

FIG. 3. (Color online) $\bar{k}(k)$ for ER graphs with $z=2$. Circles mark the results of MC simulations for the giant component and diamonds for the full graph; solid lines stand for analytical solutions.

$$\bar{k}^{(g)}(k) = \frac{z+1}{1-h^k}\left(1 - \frac{zh+1}{z+1}h^k\right). \tag{52}$$

The results are presented in Fig. 3. One can see clearly the appearance of correlations in the giant connected component as advocated in the Introduction. The agreement with the predicted values is again very good.

B. Exponential degree distribution

As the second example we take graphs with exponential degree distribution

$$p_k = (1 - e^{-1/\kappa})e^{-k/\kappa}. \tag{53}$$

The average degree in this case is

$$z = \frac{e^{-1/\kappa}}{1 - e^{-1/\kappa}} \approx \kappa - \frac{1}{2}, \quad \kappa \gg 1, \tag{54}$$

and [6]

$$G_0(x) = \frac{1 - e^{-1/\kappa}}{1 - xe^{-1/\kappa}}, \quad G_1(x) = G_0^2(x). \tag{55}$$

This implies $u=h^2$. The giant component appears for $\kappa > 1/\ln 3 \approx 0.91$. The results for $z^{(g)}$ and $z^{(f)}$ are presented in Fig. 4. As in the previous example, there are no visible deviations from the theoretical predictions.

In Figs. 5 and 6 results for $p_k^{(g)}$ and $\bar{k}^{(g)}(k)$ are presented for $\kappa=1.5$. We observe the same kind of correlations in the giant component as in the case of ER graphs.

C. Scale-free graphs

Probably the most interesting case are scale-free graphs with distribution $p_k \sim k^{-\beta}$. While studying them we have to consider two scenarios $2 < \beta \leq 3$ and $\beta > 3$. In the first case we expect correlations between node degrees, as pointed out in Refs. [9,17–19]. This invalidates both the derivation of Eqs. (18) and (45). Additionally the quantity $\langle k^2 \rangle$ diverges and so $\bar{k}(k)$ is not defined. Because our aim was to investigate the correlations appearing solely as an effect of the connectedness of graphs, we have decided not to study the $\beta \leq 3$ case in this paper. This is, however, an interesting issue and merits further investigation. One line of pursuit is to use the algorithm proposed in [19] to generate uncorrelated graphs with heavy tails. Then one should obtain predictions at least for the joint probability $p_{q,r}$ which does not contain any divergences. One could also use the V-dependent "cutoff" distribution as proposed in [19] instead of the "full" distribution $p_k \sim k^{-\beta}$. This would yield the V depending results, but may not be feasible analytically. In the case of $\beta < 2$ already the first moment of the distribution p_k is not defined and the generating function approach fails completely.

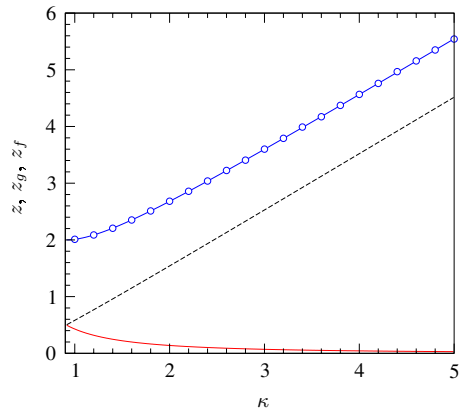

FIG. 4. (Color online) Average degree z (dashed line), average degree z_g of the connected component (upper solid line), and average degree of the rest z_f (lower solid line) as a function of κ for graphs with exponential degree distribution. Circles mark the results of MC simulations.

When $\beta > 3$ the $\langle k^2 \rangle$ is finite and there are no correlations, at least in the infinite-size limit [18,19]. However, for finite V we expect strong finite-size effects for β close to 3. To see this let us estimate the asymptotic behavior of $\langle k^2 \rangle$:

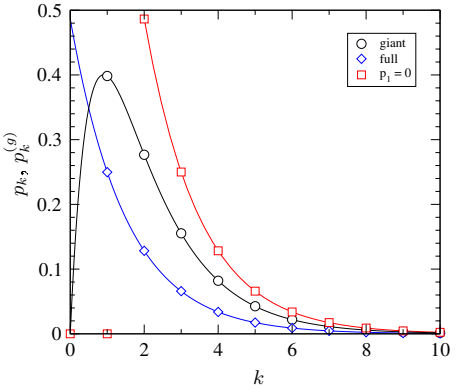

FIG. 5. (Color online) Degree distribution for graphs with exponential degree distribution with $\kappa=1.5$. Circles mark the results of MC simulations for the giant component and diamonds for the full graph; squares stand for the special case of connected graphs without leaves described in Sec. VI. Solid lines denote analytical solutions.

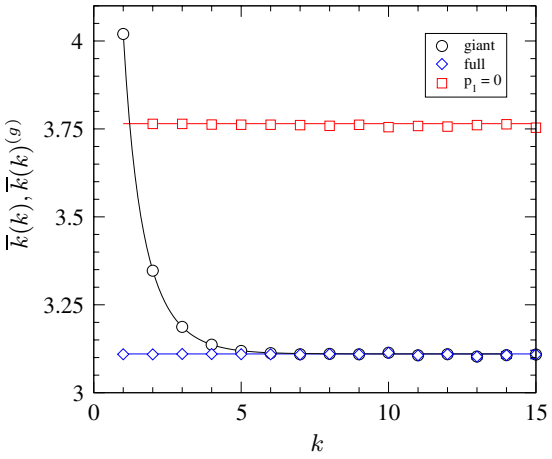

FIG. 6. (Color online) $\bar{k}(k)$ for graphs with exponential degree distribution with $\kappa=1.5$. Circles mark the results of MC simulations for the giant component and diamonds for the full graph; squares stand for the special case of connected graphs without leaves described in Sec. VI. Solid lines denote analytical solutions.

$$\langle k^2\rangle \approx \sum_k k^2 p_k - \int_{k_c(V)}^{\infty} k^2 p_k \approx \langle k^2\rangle_\infty - cV^{-(\beta-3/\beta-1)}. \tag{56}$$

In the above we have assumed the natural cutoff $k_c(V) \sim V^{1/\beta-1}$ [9,17–19]. For β close to 3, this converges very slowly. To observe those effects we have simulated our system at $\beta=13/4$, when $\langle k^2\rangle$ approaches its asymptotic value as $V^{-1/9}$. The results of our simulations of graphs with 5000 vertices are presented in Figs. 7 and 8. As expected the data for p_k and $p_k^{(g)}$ distributions show strong cutoff effects around $k=40$, but for smaller values of k the agreement with theoretical predictions is rather good. Looking at the results

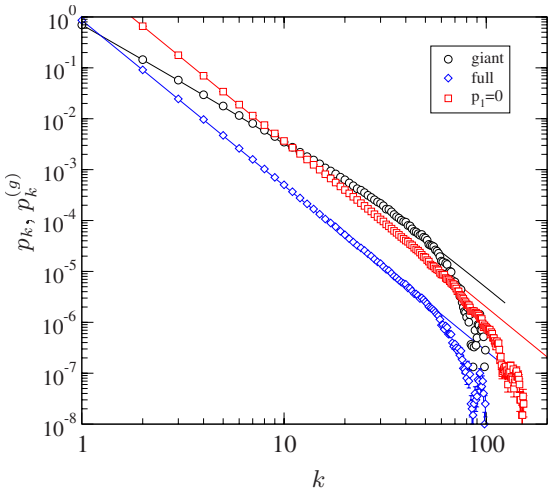

FIG. 7. (Color online) Degree distribution for scale-free graphs with $\beta=3.25$. Circles mark the results for the giant component and diamonds for the full graph; squares stand for the special case of connected graphs without leaves described in Sec. VI. Solid lines denote analytical solutions.

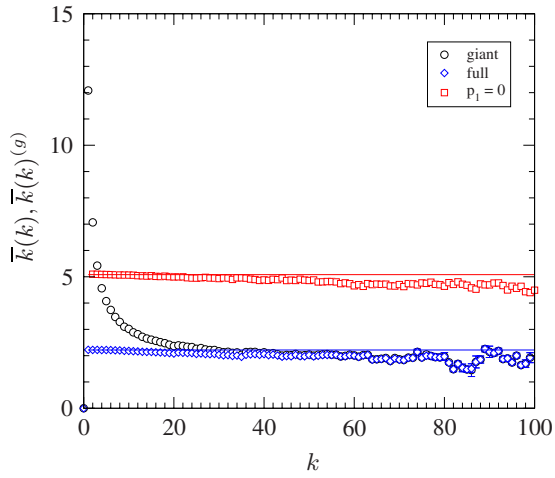

FIG. 8. (Color online) $\bar{k}(k)$ for scale-free graphs with $\beta=3.25$. Circles mark the results for the giant component and diamonds for the full graph; squares stand for the special case of connected graphs without leaves described in Sec. VI. Solid lines denote analytical solutions.

for $\bar{k}(k)$ we notice two things: (i) Data for the full graph show a deviation from a straight line, indicating the presence of some correlations due to heavy tails. (ii) Data for the giant connected component show a very strong effect of correlations. The agreement with theoretical values is very poor, so we have not included them in the picture. This is due to the described cutoff effect on $\langle k^2\rangle$. We can obtain a better agreement if we use in Eq. (46) the actual value of $\langle k^2\rangle$ measured in simulations instead of its infinite-volume limit.

V. CONNECTED GRAPHS

Finally, we would like to calculate the properties of the maximally random connected graphs. To this end we assume that the ensemble of giant connected components of the maximal entropy graphs with distribution p_k is a maximal entropy ensemble of connected graphs with distribution $p_k^{(g)}$ (we neglect the fluctuations in the number of vertices and links of the giant component). This is a plausible assumption as we do not put any additional constraints except connectivity. In Appendix B we provide a more detailed argumentation. With this assumption the properties of the maximal entropy connected random graphs with distribution $p_k^{(g)}$ and/or average degree $z^{(g)}$ are the same as that of the maximal entropy random graphs with distribution p_k and/or average degree z given by Eqs. (41) and (25a).

A. Connected ER graphs

By connected ER graphs we mean maximal entropy connected graphs with a given average degree $z^{(g)}$. According to the arguments from the previous section this ensemble corresponds to the ensemble of giant components in ER graphs with average degree z related by Eq. (25a). For a given $z^{(g)}$ we solve this equation for z (numerically) and use formulas (41) and (52) for degree distribution and for $\bar{k}(k)$ respectively. The results are presented in Figs. 9 and 10 and com-

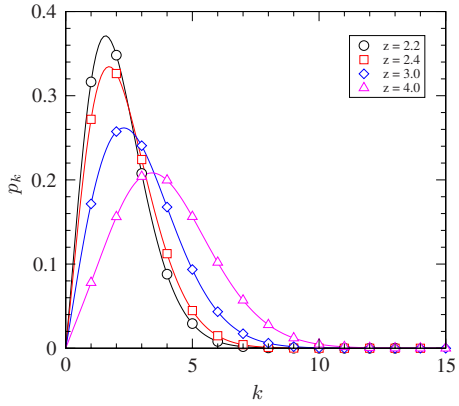

FIG. 9. (Color online) Degree distribution $p_k(k)$ in connected ER graphs with various average degrees. Points mark the results of MC simulations, while solid lines denote analytical solutions. The size of each graph is 5000 vertices.

pared with the MC data for connected graphs taken from [12]. The agreement is very good which confirms the validity of the assumption made in the previous section.

B. Connected random graphs with arbitrary degree distribution

To calculate the properties of connected random graphs with arbitrary degree distribution we need to invert Eq. (41). This can be done by rewriting it as

$$p_k = (1-h)\frac{p_k^{(g)}}{1-u^k}, \quad p_0 > 0, \tag{57}$$

where u satisfies Eq. (24):

$$u = \frac{\sum_{k=1}^{\infty} p_k^{(g)} \frac{k u^{k-1}}{1-u^k}}{\sum_{k=1}^{\infty} p_k^{(g)} \frac{k}{1-u^k}}. \tag{58}$$

The above equation can be solved by the simple iteration procedure. To prove that it has a solution we rewrite it as

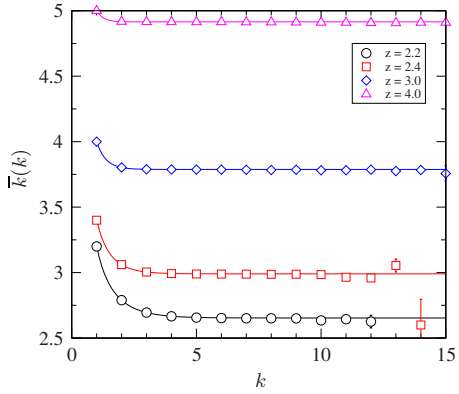

FIG. 10. (Color online) $\bar{k}(k)$ for connected ER graphs with various average degrees. Points mark the results of MC simulations, while solid lines denote analytical solutions. The size of each graph is 5000 vertices.

$$\sum_{k=1}^{\infty} p_k^{(g)} k u \frac{1-u^{k-2}}{1-u^k} \equiv g(u) = 0. \tag{59}$$

It is easy to check that

$$g(0) = -p_1^{(g)}, \quad \lim_{u\to 1} g(u) = \sum_{k=1}^{\infty} p_k^{(g)} k - 2. \tag{60}$$

So for connected graphs $g(1)$ is positive ($z^{(g)} \geq 2$) and $g(0)$ negative ($p_1^{(g)} \geq 0$).

Once we know u we can calculate h and p_0 from the normalization of the distribution p_k and Eq. (23):

$$1 = p_0 + (1-h)\sum_{k=1}^{\infty} \frac{p_k^{(g)}}{1-u^k}, \quad h = p_0 + (1-h)\sum_{k=1}^{\infty} \frac{u^k p_k^{(g)}}{1-u^k}. \tag{61}$$

Because $\Sigma_{k=1}^{\infty} \frac{p g_k}{1-u^k} - \Sigma_{k=1}^{\infty} \frac{u^k p g_k}{1-u^k} = 1$, those two equations are not independent and we can set $p_0 = 0$. Then,

$$h = 1 - \left(\sum_{k=1}^{\infty} \frac{p_k^{(g)}}{1-u^k} \right)^{-1}. \tag{62}$$

C. Simulating connected graphs

This procedure may be actually used to generate connected random graphs in an efficient way. Instead of generating connected graphs with degree distribution $p_k^{(g)}$ and checking the connectivity after every move, we can generate graphs with distribution p_k given by (57) and use the giant connected component. This still requires calculating the connected parts, but it need to be done only once before each measurement.

As an example, we have generated connected maximally random graphs with Poissonian degree distribution

$$p_k^{(g)} = e^{-z}\frac{z^k}{k!}, \quad k > 0, \quad p_0 = 0, \tag{63}$$

with $z^{(g)} \approx 2.7236$. For this distribution $u \approx 0.1209$, $h \approx 0.0341$, and $z \approx 2.6696$. Using the program [16] we have simulated a maximally random graph with $5000/(1-h) \approx 5177$ vertices and 6910 links with degree distribution (57). We generated 10 000 independent graphs. The average size of the giant component was 5000.24 ± 0.25 with standard deviation ≈ 20. The degree distribution in the connected component agrees very well with the desired one, as can be seen in Fig. 11.

VI. UNCORRELATED CONNECTED GRAPHS

An interesting situation arises when $p_1 = 0$; i.e., vertices with degree 1 (leaves) are forbidden. Then $u = 0$ and $h = p_0$. This means that the resulting graph consists of one giant connected component and $p_0 V$ isolated vertices only. It is easy to understand—finite connected components are trees, but there are no trees without leaves, except the degenerated ones made of a single vertex. If we additionally set $p_0 = 0$

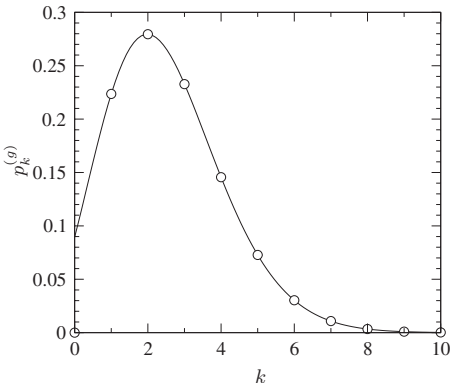

FIG. 11. Degree distribution $p_k^{(g)}$ in the connected giant component. Circles mark the results of MC simulation, while the solid line denotes the desired distribution (63).

then we will obtain a graph containing only the giant component—i.e., a connected graph.

But as observed in Sec. III C, $u=0$ implies the absence of correlations. That would support our argument made in the Introduction about the role of the one-degree vertices in the appearance of correlations in a connected graph. Using the results of the previous section we can state that vertex degrees in the maximal entropy random graphs are uncorrelated if and only if $p_1=0$; i.e., there are no leaves in the graph.

As a check, we have carried out simulations with the exponential degree distribution and no leaves:

$$p_k = \frac{1-e^{-1/\kappa}}{e^{-2/\kappa}} e^{-k/\kappa}, \quad k>1, \quad p_0 = p_1 = 0, \tag{64}$$

for $\kappa=1.5$ ($z\approx 3.055$). The results for the giant component which consisted on average of more the 99.9% of the whole graph are presented in Figs. 5 and 6 (squares). As predicted, vertices are uncorrelated in the stark contrast to the $p_1>0$ case plotted in the same figures.

We have also performed simulations for the scale-free distribution $1/k^{13/4}$ and no leaves. The results are presented in Figs. 7 and 8 (squares). We see that correlations are very much suppressed compared to the case when we admit leaves (presented in the same figures). The slight remaining correlation is due to long tails as explained in Sec. IV C.

VII. SUMMARY

In this paper we have studied the correlations in connected random graphs. We have extended the results of Refs. [6,8,10] and calculated correlations in the giant connected components of random graphs. We argue that those correlations are related to the presence of nodes with degree 1, suggesting that the only cause of correlations is the absence of "hedgehogs." This has been already stated in [11] where it has been shown that in the grand-canonical ensemble of arbitrary-sized trees, where "hedgehogs" appear, correlations vanish. We find this to be a very interesting issue that merits further studies.

The correlations observed in connected random graphs are an example of the so-called "structural" or "kinematic" correlations, as they appear in consequence of some global constraint. This should be contrasted with "dynamic" correlations which are the result of local two-point interactions between vertices. Such correlations may be generated by two-point weights [20]. This distinction can be important in simplicial quantum gravity where degree-degree correlations are interpreted as curvature-curvature correlations (see, for example, [21]). However, as the simplicial manifolds are connected by definition those correlations are due to the above described mechanism rather than to some kind of gravitational interaction [11,22]. We believe that our results may help in clarifying such issues and in the interpretation of data obtained from MC simulations.

Finally, we have shown how to relate the giant connected components to the maximal entropy connected graphs ensemble. This allowed us to propose an efficient method for generating connected random graphs based on the Metropolis algorithm.

ACKNOWLEDGMENTS

We would like to thank Zdzislaw Burda, Jerzy Jurkiewicz, Andrzej Krzywicki, and Bartłomiej Wacław for valuable discussions. This work was supported by KBN Grant No. 1P03B-04029 and EU Grants Nos. MTKD-CT-2004-517186 (COCOS) and MRNT-CT-2004-005616 (ENRAGE).

APPENDIX A: NON-SELF-AVERAGING ENSEMBLE

Denoting by $\mathcal{G}(V;k)$ the ensemble of all simple regular graphs with V vertices and degree k (in a regular graph all vertices have the same degree), we define

$$\mathcal{G}(V) = \bigcup_k \mathcal{G}(V;k), \quad P(G) = \frac{w_k}{\#\mathcal{G}(V;k)}, \tag{A1}$$

where $\#\mathcal{G}(V;k)$ denotes the number of graphs in the ensemble $\mathcal{G}(V;k)$ and w_k is an arbitrary probability distribution. With this definition we find

$$p_q = \sum_{G\in\mathcal{G}} \frac{n_q}{V} P(G) = \sum_k \sum_{G\in\mathcal{G}(V;k)} \frac{w_k \delta_{k,q}}{\#\mathcal{G}(V;k)} = \sum_k w_k \delta_{k,q} = w_q. \tag{A2}$$

It is easy to note that this poorly describes the distributions of single graphs which are just δ's. The variance of p_k is

$$\delta^2 p_q = \sum_{G\in\mathcal{G}} \left(\frac{n_q}{V} - w_q\right)^2 P(G) = \sum_k \sum_{G\in\mathcal{G}(V;k)} \frac{w_k(\delta_{k,q}-w_q)^2}{\#\mathcal{G}(V;k)}$$
$$= \sum_k w_k(\delta_{k,q}-w_q)^2 = w_q - 2w_q^2 + w_q^2 \sum_k w_k \tag{A3}$$

and indeed does not disappear in the large-V limit.

For correlations we obtain

$$\left\langle \frac{n_{q,r}}{2L} \right\rangle = qr \left\langle \frac{n_q}{2L}\frac{n_r}{2L} \right\rangle = \sum_k w_k \delta_{q,k}\delta_{r,k} = w_q \delta_{q,r} \tag{A4}$$

and

$$qr\left\langle \frac{n_q}{2L} \right\rangle \left\langle \frac{n_r}{2L} \right\rangle = \sum_k w_k \delta_{k,q} \sum_{k'} w_{k'} \delta_{k',r} = w_q w_r. \quad \text{(A5)}$$

So the condition (10) is not satisfied. It means that vertices on each particular graph are uncorrelated, but correlated if the whole ensemble is considered. This is easy to explain: if we pick a link from a graph with a given k, then the information about the first vertex does not provide any additional information; however, if we do not know k, then the degree of the first vertex will give us immediately the value of its neighbor.

APPENDIX B: ENTROPY OF THE GIANT CONNECTED COMPONENTS

Let $\mathcal{G}$ and $P(G)$ define a maximal entropy ensemble with V vertices, L links, and vertex degree distribution p_k. We assume that the probability $P(G)$ factorizes:

$$P(G) = \prod_{C \in G} P_c(C), \quad \text{(B1)}$$

where C are the connected components of the graph G.

Let $\mathcal{G}_c$ denote the ensemble of all giant connected components. We assume that we can neglect the fluctuations, so all the graphs in this ensemble have $V^{(g)}$ vertices and $L^{(g)}$ links. The degree distribution in this ensemble is $p_k^{(g)}$. Because of the property (B1), the entropy (1) of the whole ensemble $(\mathcal{G},P)$ is the sum of the entropy of the giant connected component ensemble and the rest:

$$S = S^{(g)} + S^{(f)}. \quad \text{(B2)}$$

Now we assume that there exists a probability P_c' defined on the ensemble $\mathcal{G}_c$ such that the entropy

$$-\sum_{G \in \mathcal{G}_c} P_c'(G) \ln P_c'(G) \quad \text{(B3)}$$

is greater than $S^{(g)}$, but the vertex degree probability distribution remains unchanged. Then we can define a new probability on the ensemble $\mathcal{G}$:

$$P'(G) = P_c'(C^{(g)}) \prod_{C \neq C^{(g)}} P_c(C), \quad \text{(B4)}$$

where $C^{(g)}$ is the giant connected component of graph G. The degree distribution of the ensemble $(\mathcal{G},P')$ would be the same as that of $(\mathcal{G},P)$ ensemble, but according to (B2), its entropy would be greater. This contradicts the assumption that $(\mathcal{G},P)$ is the maximal entropy ensemble and proves that the ensemble of giant connected components is a maximal entropy ensemble.

[1] R. Albert and A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).
[2] P. Erdös and A. Rényi, Publ. Math. **6**, 290 (1959); Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1961).
[3] R. Albert, H. Yeong, and A.-L. Barábasi, Nature (London) **401**, 130 (1999).
[4] A.-L. Barábasi and R. Albert, Science **286**, 509 (1999).
[5] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995); Combinatorics, Probab. Comput. **7**, 295 (1998).
[6] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[7] Z. Burda, J. D. Correia, and A. Krzywicki, Phys. Rev. E **64**, 046118 (2001).
[8] M. Bauer and D. Bernard, e-print arXiv:cond-mat/0206150.
[9] Z. Burda and A. Krzywicki, Phys. Rev. E **67**, 046118 (2003).
[10] A. Fronczak, P. Fronczak, and J. Hołyst, in *Science of Complex Networks: From Biology to the Internet and WWW; CNET 2004*, edited by J. F. F. Mendes *et al.*, AIP Conf. Proc. No. 776 (AIP, Melville, NY, 2005), p. 52. In this reference, u is denoted by $1-R$.
[11] P. Bialas, Phys. Lett. B **373**, 289 (1996).
[12] A. K. Oleś, Master's thesis (in Polish), Jagellonian University, 2006.
[13] M. Serrano, A. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani, ACM Trans. Web **1**, 10 (2007).
[14] R. Pastor-Satorras, A. Vazquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[15] L. Bogacz, Z. Burda, and B. Wacław, Physica A **366**, 587 (2006).
[16] L. Bogacz, Z. Burda, W. Janke, and B. Wacław, Comput. Phys. Commun. **173**, 162 (2005).
[17] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. E **63**, 062101 (2001).
[18] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **38**, 205 (2004).
[19] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).
[20] P. Bialas, Nucl. Phys. B **575**, 645 (2000).
[21] B. V. de Bakker and J. Smit, Nucl. Phys. B **454**, 343 (1995).
[22] P. Bialas, Z. Burda, B. Petersson, and J. Tabaczek, Nucl. Phys. B **495**, 463 (1997).