Generation of arbitrarily two-point-correlated random networks

Sebastian Weber and Markus Porto
Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstrasse 8, 64289 Darmstadt, Germany
(Received 14 February 2007; published 18 October 2007)

Random networks are intensively used as null models to investigate properties of complex networks. We describe an efficient and accurate algorithm to generate arbitrarily two-point degree-degree correlated undirected random networks without self-edges or multiple edges among vertices. With the goal to systematically investigate the influence of two-point correlations, we furthermore develop a formalism to construct a joint degree distribution $P(j,k)$, which allows one to fix an arbitrary degree distribution $P(k)$ and an arbitrary average nearest neighbor function $k_{\mathrm{nn}}(k)$ simultaneously. Using the presented algorithm, this formalism is demonstrated with scale-free networks $[P(k) \propto k^{-\gamma}]$ and empirical complex networks [$P(k)$ taken from network] as examples. Finally, we generalize our algorithm to annealed networks which allows networks to be represented in a mean-field-like manner.

DOI: 10.1103/PhysRevE.76.046111 PACS number(s): 89.75.Hc, 05.40.−a

I. INTRODUCTION

The fast developing research field of complex networks [1,2] focuses on the three main aspects of (i) measuring network topology, (ii) investigating dynamics on networks, and (iii) studying the interplay between dynamical processes on networks and the network topology. Surprisingly, empirical networks from a vast variety of scientific fields share a lot of characteristical features. Prominent examples are the small-world property [3], high clustering [4], and the scale-free degree distribution [5]. One possibility to unravel the properties of empirical networks is to compare them to null models. Appropriate null models are random networks with some of the statistical features preserved being present in the empirical network under investigation. This idea gave birth to the well-known configuration model (CM) algorithm [6–10], which is capable of generating random networks with an *a priori* given degree distribution. Some extensions to this model have been proposed to even conserve some further statistical properties than the plain degree distribution, for instance, the degree dependent clustering coefficient [11].

A fundamental way to categorize and distinguish empirical networks beyond the degree distribution and clustering has been proposed by Newman [12,13] who introduced the Newman factor r. This number is basically the Pearson correlation coefficient of degrees (the number of edges emanating from a vertex) from connected vertices in a network and is therefore fully defined by two-point degree-degree correlations in a network. The range of the Newman factor is in the interval $[-1,1]$ where positive (negative) values indicate that vertices with the same (different) degree tend to be connected, while a value of 0 means no correlation. Practically all empirical networks show a nontrivial two-point correlation structure. An astonishing observation is, for example, the fact that biological networks show negative Newman factors, while technological networks display rather small values of the Newman factor close to zero, whereas social networks tend to have rather large positive values [14]. The evident importance of correlations within the degree distribution has led to lots of efforts, for example, a hidden variable approach has been developed in Ref. [15] and so-called dK-series networks, which systematically describe the full correlation structure of a network, have been introduced in Ref. [16] together with an algorithm for the lowest dK classes. Thus, an efficient random network generator which constructs null model networks at the basis of an *a priori* prescribed two-point correlation structure is very important. Such a generator is presented below and allows one to construct undirected random networks with a prescribed two-point correlation structure and hence much more realistic null models. The major advantage of our generator in comparison with similar algorithms previously introduced [15–17] is its high accuracy and the generality of the approach, which allows one to construct networks with an arbitrary two-point correlation structure. As an application of this scheme and in order to investigate the influence of two-point correlations within empirical networks, we address the question of how one can model two-point correlations while preserving the degree distribution of a network. This is fundamental, for instance, in order to shed light on the interplay between dynamical processes on networks on the underlying network topology with respect to two-point correlations.

The modeling of two-point correlations is especially interesting for the verification of theoretical predictions from theories describing dynamical processes on networks which do incorporate two-point correlations. Due to the small-world effect present in networks, it is common use to utilize a mean-field (MF) ansatz. Hence, within these theories the network is modeled using a probabilistic approach and vertices are only connected with a certain probability to each other. The idea to represent a network by probabilities has already been brought up in the context of Kauffman's model of random complex automata [18,19]. This so-called annealed network changes in every time step such that all edges are redistributed. A similar approach has recently been applied by Stauffer and Sahimi [20] to scale-free networks to study the effect of "annealed disorder" on a diffusion process [20]. Such annealed networks are ideally suited to test the validity of MF theories of dynamics on networks. We extend this approach below by generalizing our algorithm to allow for the construction of two-point-correlated annealed networks.

This paper is organized as follows. Section II introduces the network correlation measures used in this paper. Section III describes the algorithm to construct arbitrarily two-point-correlated networks. Section IV develops a formalism which allows one to fix a degree distribution and to arbitrarily choose the two-point correlations at the same time. The formalism is demonstrated with scale-free networks and empirical networks as examples. Section VI introduces the notion of a two-point-correlated annealed network. We conclude and give an outlook in Sec. VII.

II. CORRELATION MEASURES

The following is a short summary of common definitions adapted to our purposes which will be used frequently within this paper. Two-point correlations are statistically described by the joint degree distribution $P(j,k)$, which is the probability that a randomly chosen edge of the network has vertices with degrees j and k at its ends. This distribution is a symmetric function in the case of undirected networks, $P(j,k)=P(k,j)$. By summation over either parameters of $P(j,k)$, one obtains the distribution over edge ends,

$$P_e(k) = \sum_j P(j,k), \tag{1}$$

which is related to the degree distribution by

$$P(k) = \frac{\bar{k}}{k} P_e(k). \tag{2}$$

This last relation (2) between the edge end distribution $P_e(k)$ and the degree distribution $P(k)$ can easily be understood by the fact that every vertex with degree k has probability $P(k)$ of being drawn at random from the network. Therefore, the probability to draw an edge end connected to a vertex of degree k is proportional to $kP(k)$. Normalizing this last expression yields the edge end distribution $P_e(k)=kP(k)/\bar{k}$. Here, $\bar{k}=\Sigma_k kP(k)$ denotes the mean with respect to the degree distribution $P(k)$. This mean has to be carefully distinguished from the mean with respect to the edge end distribution $P_e(k)$, which we denote by $\langle k\rangle=\Sigma_k kP_e(k)=\overline{k^2}/\bar{k}$. It is convenient [21] to extract the actual correlations from $P(j,k)$ by relating it to the uncorrelated case $P_{\mathrm{uc}}(j,k)$, which has the special product form

$$P_{\mathrm{uc}}(j,k) = P_e(j)P_e(k). \tag{3}$$

By taking the ratio between $P(j,k)$ and $P_{\mathrm{uc}}(j,k)$, this defines

$$f(j,k) = \frac{P(j,k)}{P_{\mathrm{uc}}(j,k)} \tag{4}$$

as a correlation function.

However, the joint degree distribution $P(j,k)$ and the correlation function $f(j,k)$ are complex functional objects which are hard to imagine. A way to quantify the overall correlation present in a network was introduced by Newman [12]. He defined the Newman factor r to be the Pearson correlation coefficient of the remaining degrees of two vertices at either end of a randomly chosen edge. The use of the remaining degree, which is the actual degree of a vertex minus one, is only an arithmetic trick to suppress some terms in calculations performed by Newman. In this paper, we directly use the degrees of the vertices, which is equivalent to Newman's definition in the limit of large networks,

$$r = \frac{1}{\sigma_e^2} \sum_{j,k} jk[P(j,k) - P_e(j)P_e(k)]. \tag{5}$$

The Newman factor r is normalized by $\sigma_e^2=\langle k^2\rangle-\langle k\rangle^2$ to fall into the range $[-1,1]$. A positive (negative) value means that vertices with a degree k preferentially attach to vertices with a degree of the same (different) order, which is referred to as (dis-)assortative mixing. The special case of $r=0$ is achieved in the case of no correlation, which can be seen by substituting $P_{\mathrm{uc}}(j,k)$ of Eq. (3) into Eq. (5). It is clear that the Newman factor r quantifies the correlations present in a network only on a global scale. An intermediate approach, being on the level of degrees, has been introduced in Ref. [22] with the average nearest neighbor function $k_{\mathrm{nn}}(k)$. Using the conditional probability

$$P(j|k) = \frac{P(j,k)}{P_e(k)}, \tag{6}$$

which is the probability that a randomly chosen neighbor of any vertex with degree k has the degree j, one defines $k_{\mathrm{nn}}(k)$ to be

$$k_{\mathrm{nn}}(k) = \sum_j jP(j|k). \tag{7}$$

In the case of an (dis-)assortative network the average nearest neighbor $k_{\mathrm{nn}}(k)$ has to be an (de-)increasing function, while it has the constant value $\langle k\rangle$ for uncorrelated networks. It is interesting to note that

$$\langle k_{\mathrm{nn}}(k)\rangle = \langle k\rangle \tag{8}$$

is generally valid, which can be seen by plugging Eq. (6) into Eq. (7) and averaging the resulting equality over k with respect to the edge end distribution $P_e(k)$.

III. ALGORITHM

The well-known CM algorithm [6–9] fixes *a priori* a degree sequence which is usually drawn from a given degree distribution $P(k)$. Each element of this degree sequence is the number of desired edges emanating of a vertex. These may be thought of as half-edges which still need to be joined with half-edges of other vertices. To construct the network, the CM algorithm may be implemented by placing all half-edges of all vertices into a single list, which is a discrete representation of the edge end distribution $P_e(k)$. An edge is formed by selecting two random members of that list. If the constraint of neither self-edges nor multiple edges is met, the edge is created and the two half-edges are removed from the list. As the first and the second draw is done from the same list or, equivalently, each draw is done independently with the edge end distribution $P_e(k)$, the resulting network is al-

ways uncorrelated. Only the constraint of self-edges and multiple edge prevention induces some intrinsic correlations, which can be avoided if the maximal degree $k_{\max}$ is limited (cf. Sec. IV A). The CM algorithm paired with the correct choice of the maximal degree $k_{\max}$ is as well known as the uncorrelated CM (UCM) algorithm [10]. However, almost all empirical networks do display two-point correlations in their topology. The algorithm discussed below allows one to fix *a priori* an arbitrary joint degree distribution $P(j,k)$ and generates a network which is completely random under all other topological aspects, just as the CM algorithm does with respect to the degree distribution $P(k)$.

A major computational complication arises from the fact that probabilities in the $P(j,k)$ matrix may become very small as the probability for one edge is of the order $1/\bar{k}N$ and computationally hard to handle for large N. Due to this problem, we sample in a first step a half-edge with the usual edge end distribution $P_e(k)$; in a second step, we sample a half-edge from the conditional probability distribution $P(j|k)$. The former two objects are much easier to sample as those are the result of integrals over $P(j,k)$ and therefore contain probabilities of greater order.

The overall scheme of the algorithm to construct a network with N vertices and a given joint degree distribution $P(j,k)$ is the following:

(1) As in the CM algorithm, one first has to draw a degree sequence by calculating the theoretical (continuous) edge end distribution $P_e(k)$ from the joint degree distribution $P(j,k)$ and transform that into a degree distribution $P(k)$. From this distribution, a degree sequence of length N is drawn.

(2) Each element of the degree sequence represents a vertex. All vertices with the same degree k are then sorted into degree classes, each containing only vertices of the same degree k.

(3) To compensate for discretization effects caused by the finiteness of the sampled network, one has to calculate the discrete edge end distribution $P_e^{(d)}(k)$ from the generated degree sequence. To do so, one acquires, by estimating the size of each degree class, the discrete degree distribution $P^{(d)}(k)$, which corresponds to a discrete edge end distribution by $P_e^{(d)}(k)=kP^{(d)}(k)/\bar{k}$.

(4) Next, the discrete conditional probability $P^{(d)}(j|k)$ is set up. To obtain a matrix which accommodates the discretization effects, one replaces the continuous edge end distributions $P_e(k)$ in the definition of the conditional probability distribution of Eq. (6) by the discrete edge end distributions $P_e^{(d)}(k)$ and obtains therefore

$$P(j|k) = \frac{P(j,k)}{P_e(k)} = P_e(j)f(j,k) \approx P_e^{(d)}(j)f(j,k) = P_e^{(d)}(j)\frac{P(j,k)}{P_e(j)P_e(k)}. \quad (9)$$

Since we mix the discrete edge end distribution $P_e^{(d)}(j)$ and the continuous correlation function $f(j,k)$, the resulting conditional degree distribution $P^{(d)}(j|k)$ is only approximately normalized for a given degree class k. To obtain a conditional probability distribution suitable for sampling degree classes, we normalize each degree class separately, leading to the final form

$$P^{(d)}(j|k) = \frac{P_e^{(d)}(j)}{P_e(j)}P(j,k)\left(\sum_j \frac{P_e^{(d)}(j)}{P_e(j)}P(j,k)\right)^{-1}. \quad (10)$$

This definition is consistent with the limes $N\to\infty$, as the discrete edge end distribution $P_e^{(d)}(j)$ becomes equal in this limit to the continuous edge end distribution $P_e(j)$ and the ratios $P_e^{(d)}(j)/P_e(j)$ become exactly 1, respectively.

(5) After all base data structures have been initialized, the algorithm starts to draw edges by drawing edge ends. The first edge end is selected by first drawing a degree class k from the edge end distribution $P_e^{(d)}(k)$ and then randomly choose a vertex from that degree class.

(6) The second end of the edge is chosen in the same two step manner. However, the first draw of a degree class j is done with the appropriate conditional probability distribution $P^{(d)}(j|k)$ instead of the edge end distribution $P_e^{(d)}(k)$. This construction scheme yields correctly correlated graphs, since we have

$$\underbrace{P_e(k)}_{1.\,\text{draw}}\underbrace{P(j|k)}_{2.\,\text{draw}} = P(j,k). \quad (11)$$

An edge is created whenever the constraints of neither self-edges nor multiple edges is met. Otherwise the drawn edge is rejected and the algorithm continues with step (5).

(7) If the edge is created, the probability weights of the two edge ends are removed from the corresponding degree classes in the edge end distribution $P_e^{(d)}(k)$ and the conditional probability distribution matrix $P^{(d)}(j|k)$. The removal of the probability weight is equivalent to the removal of the two half-edges from the list of eligible half-edges in the CM algorithm.

(8) Steps (5)–(7) are repeated until no edge ends are left and all edges are formed.

The principal numerical costs of the algorithm arises from the continuous sampling of degree classes in steps (5) and (6) above. Since the algorithm has to sample only the degree classes actually realized, which is a significantly lower number than the system size N, the numerical costs are of the order $\mathcal{O}(N^\alpha)$ with $\alpha<1$. Furthermore, due to the removal of probability weight of used half-edges throughout the construction procedure, the algorithm samples only the possible configuration space which remains valid in each iteration step just as in the CM algorithm. The memory usage of the algorithm scales with the square of the number of realized degree classes. This can become a significant advantage over the CM procedure as described above, since the memory usage of the CM procedure scales with the number of half-edges needed to construct the network.

To validate our algorithm, we use three empirical networks as test cases: (i) a social network where the 392 340 vertices are actors and the edges between those are assigned if they performed in at least one movie together [5]; (ii) a subset of the World Wide Web (WWW) containing 325 759

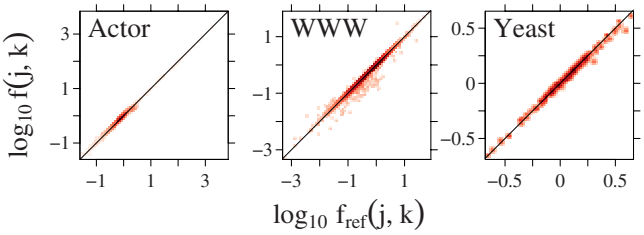

FIG. 1. (Color online) Density plot of the correlation function $f_{\mathrm{ref}}(j,k)$ of the empirical network versus the correlation function $f(j,k)$ of the corresponding random network as generated by the algorithm for all indices j and k. Darker red regions contain a higher density of data points, while lighter red indicates a lower density. The reference line $y=x$ is drawn as a guide to the eye.

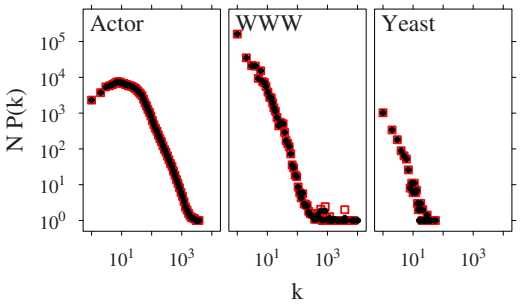

FIG. 2. (Color online) Degree distribution $P(k)$ of empirical networks and their corresponding degree distribution as generated by the algorithm. The red squares denote the reference points as measured from the empirical networks and the black circles mark values measured from the randomized networks.

web pages which are connected if there exists a link among them [23]; and (iii) the yeast protein-interaction network constituent of 1846 proteins [24]. The data has been downloaded from Barabási's web site http://www.nd.edu/~networks. All self-edges and multiple edges were removed from each network. The actor network is assortatively ($r=0.27$), the WWW network weakly ($r=-0.053$) and the yeast protein-interaction network disassortatively ($r=-0.16$) correlated. To test the correctness of the algorithm, one measures the joint degree distribution $P_{\mathrm{ref}}(j,k)$ of the base networks and uses this function as input for the construction algorithm. The resulting random network has to display the same degree distribution $P(k)$ and joint degree distribution $P(j,k)$ as the empirical one. A very sensitive test to validate if the correlation structure of the reference and the random network indeed match is on the level of the correlation function $f(j,k)$, which varies on a much smaller scale than the joint degree distribution $P(j,k)$. Thus, comparing the reference correlation function $f_{\mathrm{ref}}(j,k)$, which one obtains from the empirical network, with the correlation function $f(j,k)$ of the network as generated by the algorithm by means of a correlation coefficient (1 means total agreement, -1 indicates that the two functions are of opposite sign, and 0 means no correlation among the two functions in comparison) reveals almost complete agreement of (i) 0.99(6), (ii) 0.9(9), and (iii) 0.99(8). A density plot of the reference correlation function versus the resulting correlation function in Fig. 1 verifies the excellent agreement of the correlation functions $f(j,k)$ and $f_{\mathrm{ref}}(j,k)$. The plot shows the corresponding values of $f(j,k)$ versus $f_{\mathrm{ref}}(j,k)$ for all indices j and k at either axis. Ideally, all data points would be on the diagonal, which would be the case if the two functions were identical and the density plot would show a delta-shaped line along the diagonal. As one can see from the plots, the highest density of points, which is indicated by darker red, is almost solely centered at the diagonal. Just as the correlation functions coincide, the degree distributions show the same very good agreement, which is illustrated in Fig. 2. The statistics per curve are 10^2 randomized realizations for the actor network, 10^3 for the WWW network, and 10^4 for the yeast network in both figures.

IV. CONTROLLING CORRELATIONS IN NETWORKS

The algorithm described in this paper constructs undirected random networks with an arbitrary two-point-correlation structure. This allows us to test explicitly the influence of two-point correlations present in a network on its properties. For example, being able to control the two-point-correlation structure of a network allows one to directly test their influence on dynamical processes taking place on the networks. We therefore aim at developing a formalism which allows one to control the two-point correlations of a whole network in terms of the average nearest neighbor degree $k_{\mathrm{nn}}(k)$ and the Newman factor r, given a fixed degree distribution $P(k)$.

As we want to preserve a given degree distribution $P(k)$, which translates into a given edge end distribution $P_e(k)$, while varying the joint degree distribution $P(j,k)$, some restrictions apply to the joint degree distribution. We begin with an ansatz by writing the joint degree distribution $P(j,k)$ in product form as in Eq. (4),

$$P(j,k) = P_e(j)P_e(k)f(j,k). \tag{12}$$

It is clear that the correlations in the network are encoded by this ansatz within the correlation function $f(j,k)$. The relation to the Newman factor r from the definition Eq. (5) is

$$r\sigma_e^2 = \langle jk[f(j,k)-1]\rangle_{j,k} = \langle jkf(j,k)\rangle_{j,k} - \langle k\rangle^2. \tag{13}$$

By the notation $\langle\cdot\rangle_{j,k}$, we indicate that the average with respect to $P_e(k)$ is to be taken simultaneously over the indices j and k, similarly as $\langle\cdot\rangle$ denotes the average with respect to $P_e(k)$. The correlation function $f(j,k)$ is as well tightly connected to the average nearest neighbor degree function $k_{\mathrm{nn}}(k)$. Using that the conditional probability $P(j|k) = P(j,k)/P_e(k) = P_e(j)f(j,k)$, the definition of Eq. (7) turns into

$$k_{\mathrm{nn}}(k) = \langle jf(j,k)\rangle_j. \tag{14}$$

Multiplying the average nearest neighbor function $k_{\mathrm{nn}}(k)$ with $kP_e(k)$ and summing over all k, we are led to

$$\langle kk_{\mathrm{nn}}(k)\rangle = \langle jkf(j,k)\rangle_{j,k}, \tag{15}$$

which we can substitute into Eq. (13), leading us finally to

$$r\sigma_e^2 = \langle k k_{\mathrm{nn}}(k)\rangle - \langle k\rangle^2. \tag{16}$$

From the constraint of a given degree distribution $P(k)$ it follows that an integration over either argument of the joint degree distribution $P(j,k)$ has to be equal to the corresponding edge end distribution $P_e(j)$ [or $P_e(k)$]. Thus, the correlation function $f(j,k)$ has to fulfill the condition

$$P_e(k) = \sum_j P(j,k) = P_e(k)\langle f(j,k)\rangle_j, \tag{17}$$

which means

$$\langle f(j,k)\rangle_j = 1. \tag{18}$$

The considerations so far are general. However, as we want to control correlations within the network, we seek for an explicit correlation function $f(j,k)$ which has the property of Eq. (18) and produces a joint degree distribution which yields a given average nearest neighbor degree $k_{\mathrm{nn}}(k)$ function. To do so, we make a simple ansatz for the correlation function

$$f(j,k) = 1 + h(j)h(k). \tag{19}$$

This functional form may be understood as a series expansion of first order, fulfilling the necessary symmetry property that the correlation function has to be constant under exchange of indices j and k. Plugging this ansatz into Eq. (14) takes us to

$$k_{\mathrm{nn}}(k) = \langle k\rangle + \langle jh(j)\rangle h(k), \tag{20}$$

which means that

$$h(k) = \frac{k_{\mathrm{nn}}(k) - \langle k\rangle}{\langle jh(j)\rangle}. \tag{21}$$

The constant $\langle jh(j)\rangle$ can easily be calculated by multiplying Eq. (21) with $kP_e(k)$ and summing over all k. Rearranging the terms then yields

$$\langle kh(k)\rangle = \sqrt{\langle k k_{\mathrm{nn}}(k)\rangle - \langle k\rangle^2} = \sqrt{r\sigma_e^2}. \tag{22}$$

Finally, the correlation function $f(j,k)$ has the form

$$f(j,k) = 1 + \frac{1}{r}\frac{(k_{\mathrm{nn}}(j) - \langle k\rangle)(k_{\mathrm{nn}}(k) - \langle k\rangle)}{\sigma_e^2}. \tag{23}$$

Employing condition (18) to the ansatz in Eq. (19) yields

$$\langle h(j)\rangle = 0. \tag{24}$$

This property is consistent with the functional form of $h(k)$ in Eq. (21), since the average of $h(k)$ over k with respect to the edge end distribution $P_e(k)$ yields zero by usage of Eq. (8) ($\langle k_{\mathrm{nn}}(k)\rangle = \langle k\rangle$). Equation (8) helps furthermore to construct valid average nearest neighbor functions $k_{\mathrm{nn}}(k)$ with an arbitrary functional dependence upon the degree k. Taking a sufficiently smooth and positive weighting function $g(k)$, the corresponding $k_{\mathrm{nn}}(k)$ compatible with Eq. (8) is then

$$k_{\mathrm{nn}}(k) = \frac{\langle k\rangle}{\langle g(k)\rangle} g(k). \tag{25}$$

However, the resulting correlation function $f(j,k)$ is still constrained by even further conditions [25–27]. For example, the ratio $r_{j,k}$ as introduced in Ref. [27], is defined as the actual number of connections $E_{j,k}[=P(j,k)\bar{k}N]$ divided by the maximal number of connections $m_{j,k}$ among the degree classes j and k. For networks without multiple edges this ratio is given by

$$r_{j,k} = \frac{E_{j,k}}{m_{j,k}} = \frac{P(j,k)}{\min\{P_e(j), P_e(k), \bar{k}NP_e(j)P_e(k)/jk\}}. \tag{26}$$

It is clear that this ratio must always be in the range between 0 and 1 for all valid degree classes j and k present in the network,

$$0 \leqslant r_{j,k} \leqslant 1 \quad \forall \quad j,k \in [k_{\min}, k_{\max}]. \tag{27}$$

From this condition the admissible degree range $[k_{\min}, k_{\max}]$ becomes dependent upon the details of the correlation function $f(j,k)$. To proceed, we choose as an example the average nearest neighbor function to be a power law $k_{\mathrm{nn}}(k) \propto k^{\alpha}$, as this functional form roughly approximates the measured average nearest neighbor function of various empirical networks. Using this ansatz, one obtains the final form of the correlation function as

$$f(j,k) = 1 + \frac{1}{\langle k^{\alpha+1}\rangle/\langle k\rangle - \langle k^{\alpha}\rangle}\frac{1}{\langle k^{\alpha}\rangle}(j^{\alpha} - \langle k^{\alpha}\rangle)(k^{\alpha} - \langle k^{\alpha}\rangle). \tag{28}$$

Up to this point the degree distribution $P(k)$, or equivalently, the edge end distribution $P_e(k)$ is still arbitrary as the former does only enter Eq. (28) via the averages $\langle\cdot\rangle$ used in the definition of the correlation function $f(j,k)$. Nevertheless, the range of the exponent α is limited, since the condition of Eq. (27) has to be fulfilled. A further complication arises from intrinsic correlations caused by the constraint of the absence from self-edges and multiple edges. In the following we discuss these issues for scale-free networks and empirical networks in detail.

A. Scale-free networks

The degree distribution $P(k)$ of a scale-free network is defined by

$$P(k) \propto k^{-\gamma}, \tag{29}$$

where γ is the scale parameter. The edge end distribution is therefore given by

$$P_e(k) \propto k^{-\gamma+1}. \tag{30}$$

As we only discuss finite networks, the range of admissible degrees k is limited by various conditions. First, the rapidly decreasing probability for increasing degrees k requires one to cut off the degree range at a maximal degree $k_{\max}$ above which the accumulated probability weight is equal to $1/N$. This yields the so-called natural cutoff [28],

$$k_{\max}^{\text{natural}} = N^{1/(\gamma-1)}. \tag{31}$$

This natural cutoff is necessary to prevent large fluctuations in a finite random network ensemble and is an upper limit for the maximal degree $k_{\max}$. It is important to emphasize that this cutoff is by no means induced by the topology of the complex network.

However, it turns out that the natural cutoff is not always compatible with the condition of Eq. (27), which can easily be used to determine the so-called structural cutoff. In the case of scale-free networks, Eq. (26) reduces for sufficiently large degrees j and k to $r_{j,k}=jkf(j,k)/\bar{k}N$ and defines therefore a maximal degree $k_{\max}$ at the upper bound for the ratio $(r_{k_{\max},k_{\max}}=1)$. With this criteria, one obtains, in the case of uncorrelated networks having a constant correlation function $f(j,k)=1$, the scale parameter independent cutoff $k_{\max}^{\text{structural}} \propto N^{1/2}$. This is smaller than the natural cutoff for values of the scale parameter in the range $2<\gamma\leqslant 3$. Nevertheless, newer calculations by Dorogovtsev *et al.* [26] reveal that this structural cutoff is still too large in that particular range of the scale parameter γ and causes intrinsic correlations to arise within otherwise uncorrelated networks without self-edges or multiple edges. Due to the maximal degree $k_{\max}$ being too large and the required constraints, the vertices with large degrees k do have a tendency to connect preferably with low degree vertices, which effectively yields disassortativity. The reason for the failure of condition (27) in the case of scale-free networks with a scale parameter γ in the range (2,3] can be seen in the diverging fluctuations in the degree distribution as only the first moment of the degree distribution $P(k)$ is finite. The approach taken by Dorogovtsev *et al.* is based upon a statistical ensemble ansatz. A canonical network ensemble is defined as the set of networks with a fixed set of vertices and a fixed number of edges. The final networks are then the outcome of an evolution process where randomly chosen edges are removed and simultaneously added to a pair of vertices in the network. A pair of vertices is chosen at random with weights given by the product of a preferential function $f(j)f(k)$ where j and k are the degrees of the respective vertices. With the preferential function $f(k)=k+1-\gamma$ and beneath the critical temperature, the authors observe that the degree distribution becomes scale-free. However, depending upon the finiteness of the second moment of the degree distribution, Dorogovtsev find different cutoffs of the degree range

$$k_{\max}^{\text{ensemble}} = \begin{cases} N^{1/2} & \text{if } \gamma > 3 \\ N^{1/(5-\gamma)} & \text{if } 2 < \gamma \leqslant 3. \end{cases} \tag{32}$$

The evolution process driving a network into this equilibrium network is, of course, neither the same as constructing a network with the CM algorithm nor with the algorithm developed in this paper. The CM algorithm and the algorithm presented in this paper, however, fix *a priori* the number of vertices and edges as well, just as in the canonical network ensemble. Thus, both algorithms can be interpreted to produce graphs which are members of the canonical network ensemble below the critical temperature, since both approaches evidently yield random networks with the correct degree distribution.

Up to this point, we have only treated the uncorrelated case, which corresponds to $\alpha=0$ in Eq. (28). Numerical experiments indicated a strong deviation from the expected power law for the measured average nearest neighbor $k_{\text{nn}}(k)$ function in the case of assortative networks which have $\alpha>0$, if one naively uses a cutoff as it is applicable for uncorrelated networks. The average nearest neighbor function shows that the vertices with the largest degree fall below their expected average nearest neighbor value and tend therefore to cause some degree of disassortivity. This effect roots in the constraint of the prevention of self-edges and multiple edges and becomes stronger for larger values of the exponent α. To compensate for this effect, we incorporated the exponent α in the exponents of the maximal degrees identified so far in a simple way (an analytically exact derivation is beyond the scope of this paper) and always use the minimal resulting maximal degree,

$$k_{\max} = \min\{N^{(1-\alpha)/(5-\gamma)}, N^{(1-\alpha)/(\gamma-1)}, N^{1/(\gamma-1)}\}. \tag{33}$$

Using a maximal degree of this form lowers (raises) the cutoff degree for assortative (disassortative) correlations with increasing (decreasing) exponent α. Having fixed the maximal degree $k_{\max}$, we set the minimal degree $k_{\min}$ to be 2 in all simulations. This ensures that we always obtain a largest giant component in the network having almost the size of the entire network, which in turn guarantees that the largest giant component has the same two-point correlation structure as the entire network. This is favorable, since in most applications only the largest component of the generated random networks is of interest.

As already pointed out, it is crucial to note that only the first moment of the degree distribution is finite for values of the scale parameter γ in the range (2,3] while all higher moments diverge. However, already the first moment of the edge end distribution $P_e(k)$ is diverging in this range of the scale parameter γ. This has the important consequence that the average nearest neighbor function $k_{\text{nn}}(k)$ becomes system size dependent, as $\langle k_{\text{nn}}(k)\rangle=\langle k\rangle$ by Eq. (8). To validate the predicted power-law behavior of the average nearest neighbor function $k_{\text{nn}}(k)$, we employ a dimensionless data collapse of the function,

$$k_{\text{nn}}(k)k^{-\alpha}\langle k^{\alpha}\rangle/\langle k\rangle = 1. \tag{34}$$

This type of plot is extremely sensitive even against the smallest deviations from the predicted power law in the average nearest neighbor function $k_{\text{nn}}(k)$. The numerical results for various values of the scale parameters γ and the exponent α are shown in Fig. 3 for networks of size $N=10^6$. Each data point is calculated over an ensemble of 10^3 random networks. The curves run quite nicely along the predicted constant line of 1. Especially the $\alpha=0$ curves coincide with the constant line of 1, which is a further, very important validation of the algorithm, since in this case the algorithm has to coincide with the well-known UCM algorithm [10]. Three details are interesting to note: (i) with decreasing α the curves become longer as the maximal degree $k_{\max}$ increases;

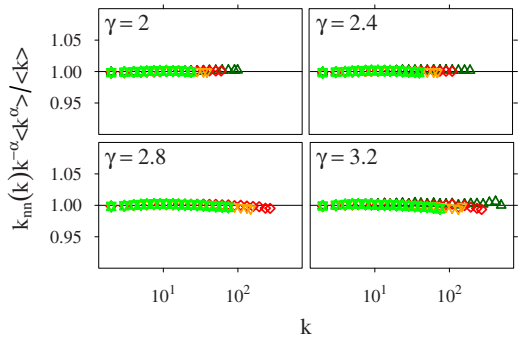

FIG. 3. (Color online) Data collapse for average nearest neighbor function $k_{nn}(k) \propto k^{\alpha}$ with various values of the power parameter α for networks with a scale-free degree distribution with varying values of the scale parameter γ. The symbols used for the different values for the scale parameter α are blue circle, −0.2; pink square, −0.1; dark green up triangle, 0.0; red diamond, 0.1; yellow down triangle, 0.2; and light green star, 0.3.

(ii) not all values of the exponent α can be realized for a given value of the scale parameter γ as condition $r_{j,k} \geqslant 0$ is violated for some curves and would require a further adjustment of k_{max} or even k_{min}; and (iii) with increasing scale parameter γ the curves for larger values of the exponent α show a trend to slightly bend below the constant line of 1, which is an indication that the cutoff as of Eq. (33) still gives slightly too large values for the maximal degree k_{max}. Another test of our formalism can be accomplished by comparing the Newman factor r of the resulting networks to the values of the analytically predicted ones by Eq. (13). Figure 4 shows that numerical simulations (points) and theoretical predictions (lines) coincide very well.

The diverging moments $\langle k \rangle$ and $\langle k^{\alpha+1} \rangle$ of the edge end distribution $P_e(k)$ for values of the scale parameter γ within the range (2,3] make a careful inspection of finite-size effects necessary. One can easily see that the ratio $\langle k^{\alpha+1} \rangle / \langle k \rangle$, appearing in the denominator of the correlation function $f(j,k)$ in Eq. (28), diverges, as the ratio becomes proportional to k_{max}^{α}. Nevertheless, a detailed calculation reveals certain restrictions on the maximal range of admissible degrees k if α is chosen to be different than 0. In this case, the criterion $r_{j,k} \geqslant 0$ leads to a relation between the minimal degree k_{min} and the maximal degree k_{max}. Thus, the range of admissible degrees is limited and the moments $\langle k \rangle$ and $\langle k^{\alpha+1} \rangle$, which would otherwise diverge, remain finite. Figure 5 shows the finite-size effects on the Newman factor r as a function of the exponent α. The plot shows only a marginal effect of the system size N on the curves. However, for smaller sizes, a broader range in the exponent α can be used. This is due to a violation of the $r_{j,k} \geqslant 0$ criterion, which requires for larger networks either a smaller maximal degree k_{max} than the one used from Eq. (33) or a greater minimal degree k_{min}. Despite the restrictions which apply to the ansatz made, the range of correlations span very well the range of correlations found in empirical networks.

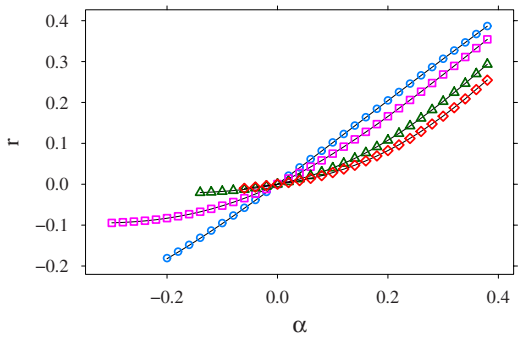

FIG. 4. (Color online) Newman factor r as a function of the scale parameter α for different values of the scale parameter γ. The straight line denotes the theoretic values of the Newman factor r as of Eq. (16). The symbols denote the value of the scale parameter γ: blue circle, 2.0; pink square, 2.4; dark green triangle, 2.8; and red diamond, 3.2.

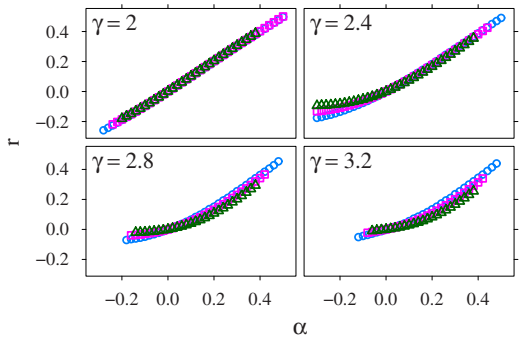

FIG. 5. (Color online) Network size dependence of the Newman factor r as a function of the exponent α for different values of the scale parameter γ. The network size N is marked by the symbols: blue circle, 10^4; pink square, 10^5; and dark green triangle, 10^6.

B. Empirical networks

A very interesting aspect of our formalism is its applicability to empirical networks. By extracting a degree sequence from an empirical network and employing the formalism developed in the last section, it is possible to create random networks which have the same degree sequence as the empirical network and an arbitrarily chosen average nearest neighbor function $k_{nn}(k)$, for instance, following a power law with tunable exponent α. Thus, given a degree sequence from a network, one constructs from this the corresponding edge end distribution $P_e(k)$ and calculates then via Eq. (28) a joint degree distribution $P(j,k)$ with which one builds a randomized network. As a result, one obtains randomized versions of the empirical network with freely tunable two-point-correlation strength, depending upon the choice of the exponent α. However, the range of the exponent α is limited by condition (27). In Figs. 6(a)–6(c) the numerical results are shown for the actor, the WWW, and the yeast networks. The plot uses the same type of data collapse as already presented in Fig. 3. The deviations from the expected constant value of 1 for the data collapse are due to intrinsic correlations which arise in networks with neither self-edges nor multiple edges and are caused by the maximal degree k_{max} in the degree sequence (see Sec. IV A). Especially the WWW network is

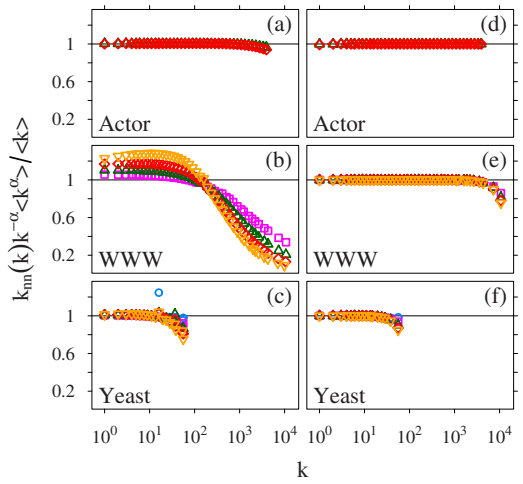

FIG. 6. (Color online) Data collapse for average nearest neighbor function $k_{nn}(k)$ for the three empirical networks actor, WWW, and yeast protein-interaction network. The left column (a), (b), and (c) show the data collapse for networks generated by the algorithm, while the right column (d), (e), and (f) show the same data collapse for networks simulated in an annealed manner. The statistics for each curve is 10^2, 10^3, and 10^4 realizations, respectively. The different symbols indicate different values for the exponent α: blue circle, −0.2; pink square, −0.1; dark green up triangle, 0.0; red diamond, 0.1; and yellow down triangle, 0.2.

strongly affected by this as it has a maximal degree k_{max} of the order 10^4, while the network size is 10^5 and hence only one order of magnitude greater.

V. ANNEALED NETWORKS

To investigate, for example, a dynamical process on random networks, one typically performs the dynamical process on a whole ensemble of networks and computes averages of the observables one is interested in. The algorithm presented so far is suitable to generate such random network ensemble. The network itself always stays constant during one dynamical process and one refers to this type of network typically as a static or a quenched network. A different approach is to change the network on a certain time scale during a dynamical process and then calculate averages over time of the observables one is interested in. In an extreme case, the vertices of the network are reshuffled before every microscopic step of the dynamic. Such changing networks are referred to as annealed networks (see Refs. [20,29–31]). If the dynamic is local in each microscopic step (for instance, a diffusion step from one vertex to another along an edge), it is sufficient to draw edges on demand only and to generate solely the local connections around the vertex considered. Here, we propose a scheme which efficiently simulates such annealed networks. The idea is to treat vertices of a network discretely while the edges are solely represented by an arbitrary joint degree distribution $P(j,k)$ such that the connectivity structure of the network is only defined on average. Hence, this scheme effectively simulates the network's connectivity structure in a mean-field (MF)-like manner.

This is a very convenient tool as theoretical approaches to complex network topics are frequently based on MF theories. Successful examples are reaction-diffusion systems [32,33], epidemic disease spreading [22], and phase transitions in ferromagnetic magnets [34], to mention just a few examples. These theories usually describe the network topology via a statistical approach. Thus, it is desirable to numerically represent networks in a probabilistic manner as well. This allows an even better test of MF based theories since the network is represented as it is done within the theory. Furthermore, by comparison of quenched with annealed simulations, one can analyze in detail which aspects of such a MF theory are an overapproximation due to the MF assumption. We define such an annealed network to consist of a degree sequence $\{k_i\}$ of size N and a corresponding joint degree distribution $P(j,k)$. Each element i of the degree sequence represents a vertex with k_i connections. Thus, the set of edges is not fixed, only the total number of edges ($N_e = \Sigma_i k_i$) is held constant. Whenever, for example, a dynamical process requests an adjacent vertex of a given vertex, the neighbor vertex is instantly determined by sampling one edge which emanates from the given vertex. This edge is drawn from the joint degree distribution $P(j,k)$ and will instantly be removed after usage.

This simulates a continuously rewired network which is only locally defined by means of one edge at a time. The first four steps to set up such an annealed network are basically the same as done for the initialization of the algorithm of Sec. III: (i) Draw a degree sequence from the joint degree distribution $P(j,k)$ or take the degree sequence from a real network. That degree sequence is (ii) sorted according to degree classes and (iii) mapped into a discrete edge end distribution $P_e^{(d)}(k)$. In the same manner as done previously, (iv) one calculates the discrete conditional degree distribution $P^{(d)}(j|k)$ from the theoretical joint degree distribution $P(j,k)$. Now, instead of constructing the network, one only redefines how neighbors of vertices and hence how edges have to be understood:

(1) The neighbor vertices of a vertex with degree k are always drawn by the conditional probability distribution $P^{(d)}(j|k)$.

(2) An edge is sampled by first drawing a vertex via the edge end distribution $P_e^{(d)}(k)$ and secondly, the vertex neighbor is found by sampling the conditional probability distribution $P^{(d)}(j|k)$.

As we want the network to be free of self-connections, we assure that the sampled vertices at both ends of the sampled edges are not the same. However, the constraint of preventing multiple edges among vertices is not possible to be enforced within this local definition of the network. Therefore, these annealed networks are free of the intrinsic degree correlations which arise due to this particular constraint. This becomes apparent in Figs. 6(d)–6(f) where numerical results of annealed networks are shown as a data collapse for the average nearest neighbor function $k_{nn}(k)$, aside with the corresponding curves in the case where the network is actually constructed (Figs. 6(a)–6(c)). Only the curve for the WWW network, Fig. 6(e), deviates from the expected value of 1 for very large degrees. This has to be attributed to the prevention

of self-connections, which is still enforced. Since these vertices with a very large degree are not allowed to connect to themselves, they have to connect on average with vertices which have a degree below the preassigned average nearest neighbor function $k_{nn}(k)$, causing some slight trend towards disassortativity.

VI. CONCLUSIONS

In summary, we have presented an efficient and accurate algorithm which generates networks with an *a priori* defined two-point degree-degree correlation structure defined by an arbitrary joint degree distribution $P(j,k)$. This provides much better null models for the investigation of empirical networks, as these are usually two-point correlated. Besides the applicability to reconstruct the two-point correlations of empirical networks, we developed a formalism which allows one to systematically tune the strength of two-point correlations in a network while preserving the degree distribution $P(k)$ of a network. The two-point correlations are specified in our ansatz via the average nearest neighbor function $k_{nn}(k)$, which we exemplified by a power-law ansatz $k_{nn}(k) \propto k^{\alpha}$ with the tunable exponent α. As two important examples, we employed this formalism in the cases of scale-free networks and empirical networks. However, as intrinsic degree correlations arise from the constraint of the prevention of self-edges and multiple edges, these cause inevitable deviations from the theoretically preassigned two-point correlations. Furthermore, we found that the maximal cutoff degree k_{max} in the case of artificial scale-free networks to prevent these intrinsic correlations is substantially lower than it was believed.

At last, we introduced the notion of two-point-correlated annealed networks which are ideally suited to test the validity of mean-field theories, since the edges of these networks are solely represented in a probabilistic manner.

Using this algorithm and the new formalism developed, one can investigate the effects of two-point correlations in empirical and artificial networks. Such a scheme is expected to be an important tool to better understand, for example, how the topology of a network influences dynamical processes on it.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] S. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[3] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[4] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
[5] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[6] E. A. Bender and E. R. Canfield, J. Comb. Theory, Ser. A **24**, 296 (1978).
[7] B. Bollobas, Eur. J. Comb. **1**, 311 (1980).
[8] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).
[9] M. Molloy and B. Reed, Combinatorics, Probab. Comput. **7**, 295 (1998).
[10] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).
[11] M. A. Serrano and M. Boguñá, Phys. Rev. E **72**, 036133 (2005).
[12] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[13] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).
[14] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[15] M. Boguñá and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).
[16] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, in *SIGCOMM* (2006), e-print arXiv:cs.NI/0605007.
[17] A. Vázquez and M. Weigt, Phys. Rev. E **67**, 027101 (2003).
[18] B. Derrida and Y. Pomeau, Europhys. Lett. **1**, 45 (1986).
[19] U. Bastolla and G. Parisi, Physica D **98**, 1 (1996).
[20] D. Stauffer and M. Sahimi, Phys. Rev. E **72**, 046128 (2005).
[21] C. Song, S. Havlin, and H. A. Makse, Nat. Phys. **2**, 275 (2006).
[22] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, in *Statistical Mechanics of Complex Networks*, edited by R. Pastor-Satorras, M. Rubi, and A. Diaz-Guilera, Lecture Notes in Physics, Vol. 625 (Springer Verlag, Berlin, 2003).
[23] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[24] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).
[25] J.-S. Lee, K.-I. Goh, B. Kahng, and D. Kim, Eur. Phys. J. B **49**, 231 (2006).
[26] S. N. Dorogovtsev, J. F. F. Mendes, A. M. Povolotsky, and A. N. Samukhin, Phys. Rev. Lett. **95**, 195701 (2005).
[27] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **38**, 205 (2004).
[28] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).
[29] Z. Burda, J. D. Correia, and A. Krzywicki, Phys. Rev. E **64**, 046118 (2001).
[30] S. Dorogovtsev and J. Mendes, *Evolution of Networks* (Oxford University Press, New York, 2003).
[31] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Nucl. Phys. B **653**, 307 (2003).
[32] M. Catanzaro, M. Boguna, and R. Pastor-Satorras, Phys. Rev. E **71**, 056104 (2005).
[33] S. Weber and M. Porto, Phys. Rev. E **74**, 046108 (2006).
[34] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. E **66**, 016104 (2002).