

Correlated Random Networks

Johannes Berg and Michael Lässig

Institut für Theoretische Physik, Universität zu Köln, Zùlpicher Straße 77, 50937 Köln, Germany
(Received 16 May 2002; published 11 November 2002)

We develop a statistical theory of networks. A network is a set of vertices and links given by its adjacency matrix \mathbf{c} , and the relevant statistical ensembles are defined in terms of a partition function $Z = \sum_{\mathbf{c}} \exp[-\beta \mathcal{H}(\mathbf{c})]$. The simplest cases are uncorrelated random networks such as the well-known Erdős-Rényi graphs. Here we study more general interactions $\mathcal{H}(\mathbf{c})$ which lead to *correlations*, for example, between the connectivities of adjacent vertices. In particular, such correlations occur in *optimized* networks described by partition functions in the limit $\beta \rightarrow \infty$. They are argued to be a crucial signature of evolutionary design in biological networks.

DOI: 10.1103/PhysRevLett.89.228701

PACS numbers: 89.75.Hc, 05.20.-y

Networks describe structures as diverse as the interaction links between proteins in a cell, the wiring of the brain, or the connections of the internet. Recent theoretical work [1,2] has focused on communication networks, and a wealth of quantitative data is now becoming available on networks in molecular biology. Examples include control networks in gene transcription [3], the interaction map of proteins [4], and the pathways of cell metabolism [5]. All these systems consist of many different kinds of molecules linked by complex interactions. Network models are a simplified description, which neglects quantitative aspects of these interactions and focuses solely on their pathways.

A network is a set of vertices $i = 1, \dots, N$ connected by links. It is uniquely defined by the adjacency matrix \mathbf{c} , whose entries are $c_{ij} = 1$ if there is a link from i to j and $c_{ij} = 0$ otherwise. We consider here networks with undirected links, where \mathbf{c} is symmetric. The *connectivity* or *degree* of a vertex is then defined as the total number of links connected to it, $k_i \equiv \sum_j c_{ij}$. The *distance* d_{ij} between two vertices i and j is the number of links along the shortest path connecting them [6]. We assume the vertices are labeled (for example, by their biochemical identity), so that the correspondence between adjacency matrix \mathbf{c} and its graph is one-to-one [7].

Networks with an irregular wiring naturally lend themselves to a statistical description [6]. We discuss here the equilibrium statistics of networks. The partition function Z can be defined as a sum over all graphs with a fixed number N of vertices and a fixed number $M = \sum_{i<j} c_{ij} = \text{Tr } \mathbf{c}^2/2$ of links

$$Z = \prod_{i<j} \sum_{c_{ij}=0}^1 \delta(M - \text{Tr } \mathbf{c}^2/2) \exp[-\beta \mathcal{H}(\mathbf{c})]. \quad (1)$$

Averages over this ensemble are denoted by $\langle \dots \rangle$. Alternatively, one can define Z with an arbitrary number of links adjusted by a suitable chemical potential. The ensembles of relevance here have a finite average connectivity $\kappa \equiv 2M/N$. For fixed κ , the distribution of connectivities,

$$p(k) \equiv \frac{1}{N} \sum_{i=1}^N \langle \delta(k_i - k) \rangle, \quad (2)$$

becomes asymptotically independent of N , implying that typical adjacency matrices \mathbf{c} become sparse for large N . This is the case of interest for applications.

A satisfactory mathematical theory exists to date only for what we call *uncorrelated random networks* [9,10]. In this case, the Hamiltonian depends only on single-point connectivities,

$$\mathcal{H}_1(\mathbf{c}) = \sum_{i=1}^N f(k_i), \quad (3)$$

leading to $p(k) \sim \exp[-\beta f(k) - \mu k]/k!$ where the constant of proportionality is fixed by normalization and μ is adjusted to give the correct average connectivity. Since all graphs with the same connectivities k_1, \dots, k_N have the same statistical weight, this ensemble ensures the maximally random wiring compatible with the distribution $p(k)$. The simplest example is the well-known Erdős-Rényi graphs, where $\beta \mathcal{H} = 0$ and $p(k)$ is a Poissonian.

Many natural networks are, however, not of this type. The simplest kind of correlations occur if the joint distribution of connectivities for neighboring vertices,

$$q(k, k') \equiv \frac{1}{\kappa N} \sum_{i,j=1}^N \langle \delta(k_i - k) c_{ij} \delta(k_j - k') \rangle, \quad (4)$$

differs from its form for uncorrelated random networks, $q_0(k, k') = (kk'/\kappa^2)p(k)p(k')$. Higher correlations can be defined in a similar way [11]. Connectivity correlations have been found in growth models of communication networks [12,13] as well as in data of genetic and protein networks [14,15].

These observations call for a statistical theory of more general ensembles called *correlated random networks*, which is the subject of this Letter. The ensembles of interest are characterized by finite distributions $p(k)$ and $q(k, k')$ in the limit of large N . One then expects a universal logarithmic scaling of the average distance d_{ij} between vertices, $\sum_{i,j} \langle d_{ij} \rangle / N_\Omega^2 \sim \log N_\Omega$, in any

connected component Ω with N_Ω nodes [16]. This is consistent with our numerical findings. Hence, correlated random networks maintain a sparse connectivity matrix and are locally treelike. The “inverse temperature” β in (1) measures the deviation from Erdős-Rényi graphs. Quite remarkably, these structural properties are preserved in the limit $\beta \rightarrow \infty$, where we obtain nontrivial *optimized* networks. Ensembles of this kind generically have strong correlations.

The simplest type of Hamiltonian producing correlations has nearest-neighbor connectivity interactions,

$$\mathcal{H}_2(\mathbf{c}) = \sum_{i < j} c_{ij} g(k_i, k_j), \quad (5)$$

where $g(k, k')$ is some function of the connectivities. The resulting class of graph ensembles can be seen as a showcase for correlated random networks where analytic expressions can be derived. Higher order correlations are generated by Hamiltonians with next-nearest neighbor interactions, etc., We also study a Hamiltonian $\mathcal{H} = \mathcal{H}_1 + \lambda \mathcal{H}_d$ with a nonlocal part,

$$\mathcal{H}_d(\mathbf{c}) = \sum_{i < j}^N d_{ij}, \quad (6)$$

often called the *diameter* of the graph. For $\lambda > 0$ and a suitable scaling $\lambda \sim 1/(N \log N)$, this Hamiltonian is found to generate *compactified networks* with finite $p(k)$ and $q(k, k')$, provided the extra term \mathcal{H}_1 stabilizes the network against collapse to a star.

Compactified networks occur in communication and transport [17], and may play a role in biology [18,19]. For example, distance-optimized networks obtained from (6) show an abundance of high-connectivity vertices (hubs)

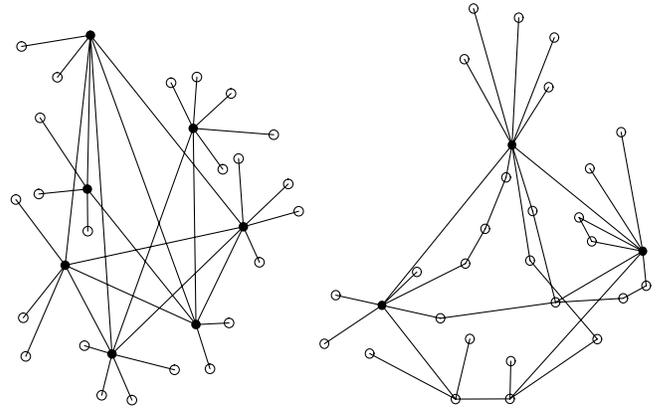


FIG. 1. Optimized networks, obtained from local interactions (left) and nonlocal interactions for large β . Hubs of high connectivity (solid circles) are preferentially connected to peripheral vertices of low connectivity (open circles).

which are preferentially connected to peripheral vertices of low connectivity. A very similar structure (with even more pronounced hubs) can also be obtained from local interactions of the form (5). Typical graphs are shown in Fig. 1.

To establish these results, we first discuss the analytical treatment of local interactions, choosing a generic Hamiltonian of the form $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$ given by (3) and (5). From the partition function (1), the free energy is derived by using an integral representation for the constraints due to the connectivity k_i at each vertex and making a Hubbard-Stratonovich transformation. The resulting integral can be evaluated in a saddle-point approximation, yielding the reduced free energy per vertex in the thermodynamic limit,

$$-\beta f \equiv \lim_{N \rightarrow \infty} \frac{1}{N} \log Z = \frac{\kappa}{2} \log N + \frac{\kappa}{2} (\log \kappa - 1) - \frac{\kappa}{2} \log \left(\sum_{k, k'} Q_k \mathcal{G}_{kk'}^{-1} Q_{k'} + e^{-2\mu} \right) + \log \left[\sum_k \frac{(Q_k + e^{-\mu})^k}{k!} \exp[-\beta f(k)] \right]. \quad (7)$$

The “order parameters” Q_k and the chemical potential μ have to be determined self-consistently from the saddle-point condition. This form is closely related to a field-theoretic approach [10,20,21], where networks appear as the Feynman diagrams of a Gaussian integral with a propagator matrix $\mathcal{G}_{kk'} = \exp[-\beta g(k, k')] - 1$ and interactions as specified by the last term in (7). Notice the superextensive scaling of the entropy, $(\kappa/2) \log N$, which reflects that unlike in a regular lattice, each vertex can be connected to all $N - 1$ other vertices. The last term in (7) is directly related to the degree distribution,

$$p(k) = C \frac{(Q_k + e^{-\mu})^k}{k!} \exp[-\beta f(k)], \quad (8)$$

where C is a normalization constant. For example, a power-law tail in $p(k)$ may be generated by a suitable choice of the weights $f(k)$ but it is not generic.

A more detailed account of networks with generic local interactions will be published elsewhere [8]. Here we turn to the simplest Hamiltonian with local interactions producing nontrivial optimized networks, see Fig. 1. It has the form $\mathcal{H}_L = \mathcal{H}_1 + \mathcal{H}_2$ with $\mathcal{H}_1(\mathbf{c}) = -(1/2) \sum_i k_i^2 + \eta \sum_i k_i^3$ and $\mathcal{H}_2(\mathbf{c}) = \zeta \sum_{i < j} \delta_{k_i, 1} c_{ij} \delta_{k_j, 1}$. The first term $(1/2) \sum_i k_i^2 = (1/2) \sum_{ijk} c_{ik} c_{kj}$ gives the number of paths of length two on the graph. It rewards the formation of hubs, i.e., highly connected vertices, which in turn lead to short distances. In fact this term has the maximally compact, starlike configuration as its ground state. The collapse to a star, where the connectivity of the central vertex scales with the size of the graph, however, is prevented by the regularization term $\eta \sum_i k_i^3$. The correlation term \mathcal{H}_2 with $\zeta \rightarrow \infty$ suppresses single, isolated links connecting two vertices of connectivity 1. Without this term an extensive

number of isolated links remains even in the limit $\beta \rightarrow \infty$ leading to graphs with a large number of disconnected parts. A minimal connectivity of 1 of each vertex is enforced. For this Hamiltonian, the free energy (7) contains only one nonzero order parameter Q_1 given by $p(1) = \kappa[Q_1/(Q_1 - e^{-\mu})]$. The chemical potential μ is determined by $\sum_k k p(k) = \kappa$. Remarkably, the connectivity correlation can be obtained from \mathcal{H}_2 and single-vertex quantities,

$$q(k, k') \sim k p(k) t_k k' p(k') t_{k'} \exp[-\beta g(k, k')]. \quad (9)$$

The constant of proportionality is fixed by normalization and the t_k are determined by the marginal distribution $\sum_{k'} q(k, k') = k p(k)/\kappa$, giving $t_1 = \kappa/\{T[\kappa - p(1)]\}$ and $t_k = T = \sqrt{1 - p(1)^2/[\kappa - p(1)]^2}$ for $k > 1$.

The properties of optimized networks resulting from the Hamiltonian \mathcal{H}_L are readily inferred from Eqs. (8) and (9). At finite values β one finds that the degree distribution (8) has an exponentially decaying tail. In order to analyze the limit $\beta \rightarrow \infty$, we replace the sum over k in (7) by an integral. One finds that the vertices arrange themselves into *hubs* of connectivity

$$k^* = (1 - 2\eta)/4\eta \quad (10)$$

and peripheral vertices of connectivity 1. The peripheral vertices are connected only to hubs, while the hubs form an uncorrelated random network.

A remarkably similar structure is found for compactified networks generated by the Hamiltonian with nonlocal interactions $\mathcal{H} = H_1 + \lambda H_d$ with H_d given by (6) and $\mathcal{H}_1 = \eta \sum_i k_i^3$. For $\lambda > 0$, the nonlocal part \mathcal{H}_d favors networks with short distances, while \mathcal{H}_1 prevents the collapse to a star as before. Hence, by choosing $\lambda = 2/(N \log N)$, one obtains a well-defined thermodynamic limit with the average distance between vertices scaling as $\log N$. We have studied this ensemble, as well as the case of local interactions \mathcal{H}_L , by a Monte-Carlo link dynamics. Starting, for example, from an Erdős-Renyi graph, randomly chosen links are moved to previously unlinked vertex pairs with probability $p = \min(1, \exp[-\beta \Delta \mathcal{H}])$ where $\Delta \mathcal{H}$ denotes the corresponding change in the Hamiltonian. The minimum degree of 1 is enforced throughout, self-links are excluded [7]. No dependence on the initial conditions has been found. For the local Hamiltonian we use a network with $N = 200$, $\kappa = 2.4$, $\eta = 0.03$, for the nonlocal Hamiltonian we use $N = 100$, $\kappa = 2.4$, $\eta = 0.001$. We averaged over 100 samples. Figure 2 juxtaposes analytical and numerical results for local interactions on the left with numerical results for nonlocal interactions on the right. The connectivity distribution $p(k)$ shows the formation of high-connectivity hubs in both cases.

For local interactions the hub connectivity is given by (10), for nonlocal interaction the distribution remains broad even in the limit $\beta \rightarrow \infty$ [Fig. 2(a)] [22].

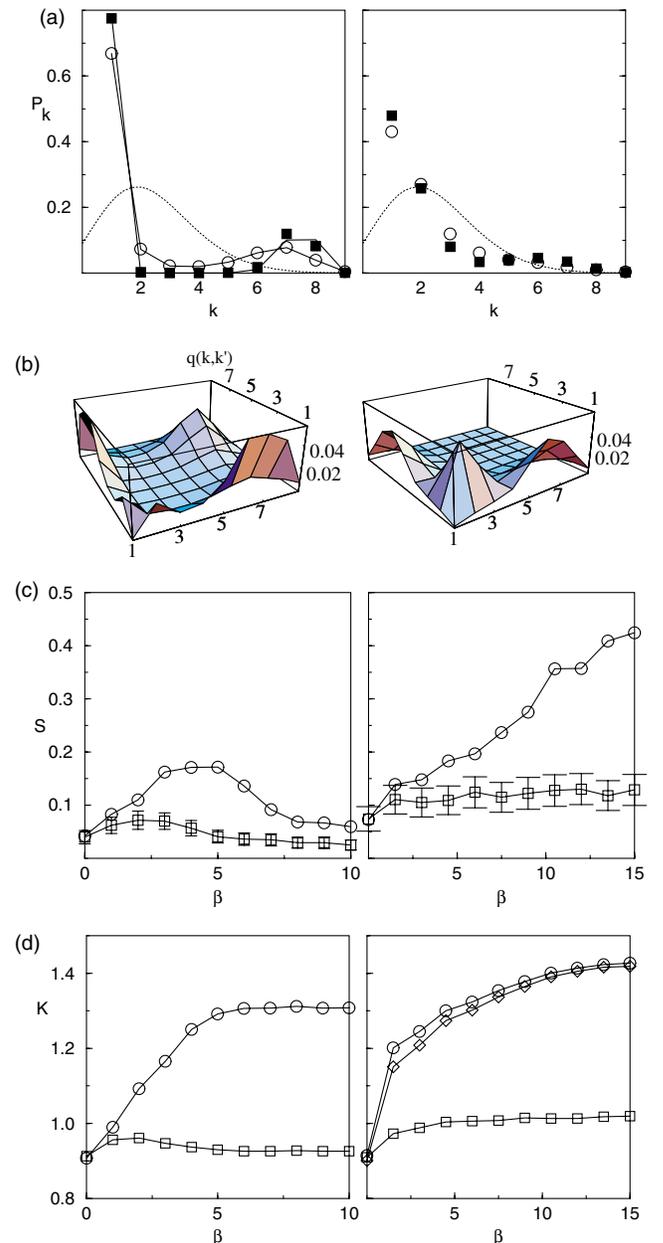


FIG. 2 (color online). Statistical features of correlated random networks with local interactions (left) and nonlocal interactions (right). (a) Single-point connectivity distribution $p(k)$ for various values of β . The Poisson form (dotted line), data for intermediate β (open circles) and large β (solid squares), and analytical values (8) for the case of local interactions (solid lines). (b) Neighbor connectivity distribution $q(k, k')$. Left: Local interactions, analytical form (9) for $\beta = 3$. Right: Nonlocal interactions, numerical results for $\beta = 15$. (c) Average relative entropy $\langle S(\hat{q}|\hat{q}_0) \rangle$ (circles), compared to the average sampling entropy $\langle S_0(\hat{q}_0) \rangle$ (squares) and its standard deviation $\langle \Delta S_0(\hat{q}_0) \rangle$, see text. (d) Average inverse distance K as a function of β (open circles), compared to the same quantity for the equivalent uncorrelated network (open squares) and for the equivalent locally correlated network (right figure, open diamonds); see text.

Low-connectivity vertices are preferentially attached to hubs as indicated by the peaks at $q(1, k^*)$ for local interactions and the corresponding peaks of $q(k, k')$ for non-local interactions [Fig. 2(b)]. The deviation from uncorrelated random networks is measured by the *relative entropy* $S(q|q_0) = \sum_{k,k'} q(k, k') \log[q(k, k')/q_0(k, k')]$. In a *single* network of size N , we obtain the estimate $S(\hat{q}|\hat{q}_0)$ from the observed frequencies $\hat{q}(k, k')$ and $\hat{p}(k)$, with $\hat{q}_0(k, k') = (kk'/\kappa^2)\hat{p}(k)\hat{p}(k')$. We then generate a sufficient number of uncorrelated random networks [23] labeled by α with frequencies $\hat{q}^\alpha(k, k')$ and sampling entropies $S(\hat{q}^\alpha|\hat{q}_0)$; their average and standard deviation are denoted by $S_0(\hat{q}_0)$ and $\Delta S_0(\hat{q}_0)$, respectively. Connectivity correlations in the original network are significant if $S(\hat{q}|\hat{q}_0) - S_0(\hat{q}_0) \geq \Delta S_0(\hat{q}_0)$. This is typically the case from a certain optimization degree β on, as shown by the ensemble averages over ten samples $\langle S(\hat{q}|\hat{q}_0) \rangle$, $\langle S_0(\hat{q}_0) \rangle$, and $\langle \Delta S_0(\hat{q}_0) \rangle$ of Fig. 2(c).

Both kinds of networks become more compactified with increasing β , as shown by the average inverse distance $K \equiv [2/N(N-1)] \sum_{i<j} d_{ij}^{-1}$ [Fig. 2(d)] [24]. We also plot K for the equivalent uncorrelated random networks; no such compactification is seen. Hence, the one-point distribution $p(k)$ may miss important functional properties. On the other hand, the nonlocally interacting networks and their equivalent locally interacting counterparts [constructed to have the same $p(k)$ and $q(k, k')$] have a very similar degree of compactification [23]. This illustrates how *optimization induces correlations*.

In summary, we have shown how interactions shape the structure of a network. Hamiltonians beyond the “single-vertex” form (3) generate correlations such as a neighbor connectivity distribution $q(k, k')$ which differs from that in uncorrelated networks. Higher correlations can be defined in a similar way [11]. These correlations provide a more detailed fingerprint of the interactions present than the single-point connectivity distribution $p(k)$. This observation should carry over to the dynamical rules for nonequilibrium ensembles such as the well-known growth models [1,12].

In transcription control networks, structural motifs have been identified that can be expressed in terms of connectivity correlations [14]. Such correlations have also been observed in protein networks [15]. In view of our findings for optimized networks, they appear to be a natural consequence of the underlying dynamics and functional optimization. We expect the data to give important information on the underlying design principles of networks and on the selective forces governing their evolution. Reverse engineering seems feasible, with the aim of inferring the relevant dynamics from the data. The *nonequilibrium* theory of correlated random networks will thus be an important avenue for future research.

Many thanks to Sergei Maslov for fruitful discussions and for making [15] available prior to publication.

- [1] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
- [2] S. N. Dorogovtsev and J. F. F. Mendes, *Adv. Phys.* **51**, 1079 (2002).
- [3] See, e.g., www.mgs.bionet.nsc.ru/systems/mgl/genenet
- [4] P. Uetz *et al.*, *Nature (London)* **403**, 623 (2000).
- [5] See, e.g., <http://igweb.integratedgenomics.com/IGwit/>
- [6] B. Bollobas, *Random Graphs* (Academic Press, London, 1985).
- [7] The graphs can be defined with or without allowing self-links (given by matrix elements $c_{ii} = 1$). The corresponding ensembles differ only by corrections of order $1/N$. Similarly, networks with unlabeled vertices are rather simply related to the ensembles discussed here in the large- N limit; see [8].
- [8] J. Berg and M. Lässig (to be published).
- [9] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Phys. Rev. E* **64**, 026118 (2001).
- [10] Z. Burda, J. D. Correia, and A. Krzywicki, *Phys. Rev. E* **64**, 046118 (2001).
- [11] The simplest higher term is $\text{Tr } c^3$, which counts the number of triangles.
- [12] D. S. Callaway *et al.*, *Phys. Rev. E* **64**, 041902 (2001).
- [13] R. Pastor-Satorras, A. Vazquez, and A. Vespignani, *Phys. Rev. Lett.* **87**, 258701 (2001).
- [14] S. S. Shen-Orr *et al.*, *Nature Genetics* **31**, 64–68 (2002).
- [15] S. Maslov and K. Sneppen, *Science*, **296**, 910–913 (2002).
- [16] The same scaling is known for uncorrelated random networks; see, e.g., G. Szabo, M. Alava, and J. Kertesz, *cond-mat/0203278*.
- [17] A Hamiltonian of the form (6) has been used to generate small-world networks, see N. Mathias and V. Gopal, *Phys. Rev. E* **63**, 21117 (2001).
- [18] H. Jeong, B. Tomber, R. Albert, Z. N. Oltvai, and A. L. Barabasi, *Nature (London)* **407**, 651 (2000).
- [19] R. Ferrer i Cancho and R. V. Solé, *cond-mat/0111222*.
- [20] A. Krzywicki, *cond-mat/0110574*.
- [21] This correspondence is strictly valid only in the thermodynamic limit, where tadpole diagrams and multiple contractions between the same pair of vertices can be neglected.
- [22] The parameters η were chosen such that the third moment of the degree distribution $\langle k^3 \rangle$ is the same for the two ensembles in the limit of large β .
- [23] There is an efficient algorithm to produce an equivalent uncorrelated random network from a given network: Choose two edges connecting the pairs of vertices i, j and k, l at random with uniform probability and rewire them as i, k and j, l . Repeating this procedure sufficiently many times leaves $p(k)$ invariant but destroys all higher correlations. It has been used in Ref. [15] for protein networks. Similarly, the equivalent locally correlated network is obtained by swapping only links with $k_i = k_l$ and $k_j = k_l$, which leaves $p(k)$ and $q(k, k')$ invariant but destroys all higher correlations.
- [24] This form is chosen since it remains well defined even for disconnected components.