New Journal of Physics
The open-access journal for physics

Relevant components in critical random Boolean networks

Viktor Kaufman and Barbara Drossel
Institut für Festkörperphysik, TU Darmstadt, Hochschulstraße 6,
64289 Darmstadt, Germany
E-mail: viktor@fkp.tu-darmstadt.de and drossel@fkp.tu-darmstadt.de

New Journal of Physics **8** (2006) 228
Received 19 June 2006
Published 9 October 2006
Online at http://www.njp.org/
doi:10.1088/1367-2630/8/10/228

Abstract. Random Boolean networks (RBNs) were introduced in 1969 by Kauffman as a model for gene regulation. By combining analytical arguments and efficient numerical simulations, we evaluate the properties of relevant components of critical RBNs independently of update scheme. As known from previous study, the number of relevant components grows logarithmically with network size. We find that in most networks all relevant nodes with more than one relevant input sit in the same component, while all other relevant components are simple loops. As the proportion of nonfrozen nodes with two relevant inputs increases, the number of relevant components decreases and the size and complexity of the largest complex component grows. We evaluate the probability distribution of different types of complex components in an ensemble of networks and confirm that it becomes independent of network size in the limit of large network size. In this limit, we determine analytically the frequencies of occurrence of complex components with different topologies.

PII: S1367-2630(06)26891-5

Contents

1. Introduction

Random Boolean networks (RBNs) are often used as generic models for the dynamics of complex systems of interacting entities, such as social and economic networks, neural networks, and gene or protein interaction networks [1]. The simplest and most widely studied of these models was introduced in 1969 by Kauffman [2] as a model for gene regulation. The system consists of N nodes, each of which receives input from K randomly chosen other nodes. The network is updated synchronously, the state of a node at time step t being a Boolean function of the states of the K input nodes at the previous time step, $t-1$. The Boolean update functions are randomly assigned to every node in the network, and together with the connectivity pattern they define the *realization* of the network. For any initial condition, the network eventually settles on a periodic attractor. Of special interest are *critical* networks, which lie at the boundary between a frozen phase and a chaotic phase [3]–[5]. In the frozen phase, a perturbation at one node propagates during one time step on an average to less than one node, and the attractor lengths remain finite in the limit $N \to \infty$. In the chaotic phase, the difference between two almost identical states increases exponentially fast, because a perturbation propagates on an average to more than one node during one time step [6].

During the last few years, great progress has been made in the understanding of critical networks with $K = 2$ inputs per node, see [7]–[15]. They contain three classes of nodes, which behave differently on attractors. Firstly, there are nodes that are frozen on the same value on every attractor. Such nodes give a constant input to other nodes and are otherwise irrelevant. They form the *frozen core* of the network. Secondly, there are nodes without outputs and nodes whose outputs go only to irrelevant nodes. Though they may fluctuate, they are also classified as irrelevant since they do not influence the number and periods of attractors. Thirdly, there are *relevant* nodes, which belong to the nonfrozen nodes and lie on directed loops built of links between relevant nodes, or they connect such loops by chains of relevant nodes. These nodes determine completely the number and period of attractors. A connected set of relevant nodes is called a *relevant component*. The relevant subnetwork consists of the disjoint relevant components. The nonfrozen nodes that are not relevant sit on trees rooted in the relevant components. Numerical simulations and analytical estimation in [8] showed that the number of nonfrozen nodes scales in the limit $N \to \infty$ as $N^{2/3}$ and the number of relevant nodes as $N^{1/3}$. An analytical study of the distribution of attractor lengths in the limit of big network size in [9] is shown in [12] to contain the same result. The corresponding probability distributions for the number of nonfrozen and relevant nodes in terms of scaling functions have been obtained in [13, 14] by means of combination of analytical calculations and numerical simulations. One application of analytical approach developed in [15] extends the findings of [8] and gives slightly better numerical approximations for the scaling functions. In [13], it was additionally shown that the number of nonfrozen nodes with two relevant inputs scales as $N^{1/3}$, and that the number of relevant nodes with two relevant inputs remains finite in the limit $N \to \infty$. Probability distributions for these two quantities were also given. From these results it follows that just a few components are complex and include relevant nodes with two relevant inputs, while most relevant components are simple loops whose number is proportional to $\ln \sqrt{\bar{N}_{\mathrm{nf}}}$ for large networks, $\bar{N}_{\mathrm{nf}}$ being the mean number of nonfrozen nodes. The nonfrozen part of a critical $K = 2$ network is much like a critical network with $K = 1$ inputs per node and only the two nontrivial Boolean functions 'copy' and 'invert'. All N nodes in those networks are nonfrozen, and the relevant nodes are arranged in simple loops. The mean number of relevant nodes is proportional to $\sqrt{N}$ [15, 16], the mean number of loops equals $\ln \sqrt{N}$ for large networks, the number of loops of length l is Poisson distributed with mean $1/l$ for $l \ll \sqrt{N}$ [12], the asymptotic probability distribution for the number of relevant nodes is [13]

$$p_0(N_{\mathrm{rel}}) = \frac{N_{\mathrm{rel}}}{N}\, \mathrm{e}^{-N_{\mathrm{rel}}^2/2N}.$$

The number and period of attractors of a network is given by the number and length of attractors of its relevant subnetwork. These are obtained by combining the attractors of the individual relevant components. If all relevant components were simple loops, it would already follow that the number and length of attractors increases with system size faster than any power law [12, 17]. Complex components make the number of attractors even larger, since they possess themselves exponential numbers and periods of attractors [18]. Therefore, complex components are important for the understanding of dynamics of large RBNs.

Complex components are of interest also for other reasons. As shown in [19], Boolean network models can be appropriate models for the dynamics of genetic regulatory systems. Specialized functional blocks should play an important role in such models, and functional blocks can be modelled as complex components, when dynamics can be treated as Boolean. Since the deterministic parallel update rule used in many Boolean network models is not biologically realistic, other, and in particular stochastic, updating schemes have been investigated [11, 20]. While they affect the number of attractors, they do not influence the classification in frozen, nonfrozen, and relevant nodes. The relevant components and the position of critical line in the phase diagram [5, 6] of Boolean networks are the same for all update rules [10]. Components and motifs (smaller units of nodes) play an important role at determining whether an attractor is robust with respect to stochastic fluctuations [21].

Because of their importance for the dynamics of Boolean networks, we study in this paper complex relevant components of RBNs with $K = 2$. Future study will have to move on to more realistic network structures. In the following, we will first summarize previous results for relevant nodes and the components formed by them, which is the basis of our computer simulations and analytical considerations. Then, in section 3 we will argue that most networks have not more than one complex relevant component. In section 4, we list all possible complex components with one and two nodes with two relevant inputs and calculate their frequencies of occurrences. In section 5, we classify complex components and discuss their probability distribution both analytically and by using numerical simulations. The probability distribution for the total number of relevant components is evaluated in section 6. A summary of our findings is given in section 7.

2. Previous results for relevant nodes and relevant components

The analytical calculations and numerical simulations performed in the subsequent sections are based on methods and results presented in [13]. We summarize here those results of [13] that will be used later.

Critical RBNs with $K = 2$ inputs per node are obtained not only when all possible Boolean functions are chosen with equal probability, but also when different classes of Boolean functions are assigned different weights, provided that the proportion β of constant functions (for which the output value does not depend on the input values) equals the proportion of reversible Boolean function (for which the output changes whenever any input value changes). The remaining Boolean functions are canalizing. By γ, we denote the probability of choosing a canalizing function that yields one value for three different input values combinations and once the other output value. The remaining proportion $1 - 2\beta - \gamma$ of functions are those that respond only to one input. The results cited in the following depend on the parameters β and γ.

The starting point for the determination of the probability distribution for the number of nonfrozen nodes was in [13] to introduce a stochastic process, which finds the frozen core of nodes frozen on all attractors. This process was initiated with nodes having frozen functions, it then identified iteratively nodes which become frozen due to receiving inputs from frozen nodes. Later, only nonfrozen nodes and their connections were considered. The idea for identification of relevant nodes was to first identify nodes without outputs as irrelevant nodes and then to determine iteratively all other nodes whose outputs go only to irrelevant nodes. These nodes are also irrelevant.

The results were the following: if we denote with N_{nf} the number of nonfrozen nodes and define the scaling variable

$$y = \frac{N_{\mathrm{nf}}}{(N/\beta)^{2/3}}, \tag{1}$$

the probability distribution for the number of nonfrozen nodes depends for large N only on y and is very well approximated by the expression

$$G(y) \simeq \mathrm{e}^{-y^3/2}(1 - 0.5\sqrt{y} + 3y)/(4\sqrt{y}). \tag{2}$$

Most of these nonfrozen nodes have only one relevant input, and the number of nonfrozen nodes with two relevant inputs N_2 is distributed according to

$$f(a) = \frac{2}{3a^{1/3}(1+\gamma/\beta)^{2/3}} G\left[\left(\frac{a}{1+\gamma/\beta}\right)^{2/3}\right], \tag{3}$$

with $a = N_2/\sqrt{N_{\mathrm{nf}}}$.

If we denote by N_{rel} the number of relevant nodes and define the scaling variable

$$z = \frac{N_{\mathrm{rel}}}{\left(\frac{N}{\beta+\gamma}\right)^{1/3}}, \tag{4}$$

the probability distribution for the number of relevant nodes depends for large N only on z and is given by

$$P(z) = \int_0^\infty \mathrm{d}a \frac{f(a)}{a^{1/3}} \mathcal{C}_a\left(\frac{z}{a^{1/3}}\right). \tag{5}$$

$\mathcal{C}_a(t)$ is the probability that a random walk that starts at the origin at time $t = 0$ and that steps to the right with a rate t and to the left with a rate a leaves the origin for the last time at t. The function $P(z)$ decays exponentially for large z, and its shape depends on the parameters β and γ through their combination $1 + \gamma/\beta$. $P(z)$ becomes broader as this combination increases.

Most relevant nodes have one relevant input. The full ensemble probability $\tilde{P}(m; z)$ for (4) to be in the range $(z, z + \mathrm{d}z)$, while m of the relevant nodes have two relevant inputs, depends again on β and γ and is given by the expression

$$\tilde{P}(m; z) = \int_0^\infty \mathrm{d}a \frac{f(a)}{a^{1/3}} \mathcal{C}_a\left(\frac{z}{a^{1/3}}\right) \frac{P_{\mathrm{r}}(m|za^{-1/3}) P_{\mathrm{l}}(m|za^{-1/3})}{\sum_k P_{\mathrm{r}}(k|za^{-1/3}) P_{\mathrm{l}}(k|za^{-1/3})}, \tag{6}$$

with

$$P_{\mathrm{r}}(m|t) = \frac{1}{m!} \mathrm{e}^{-t^2/2} \left(\frac{t^2}{2}\right)^m \tag{7}$$

and

$$P_{\mathrm{l}}(m|t) = \frac{1}{m!} \mathrm{e}^{-at} (at)^m \tag{8}$$

being the probability distributions of steps to the right and left of the mentioned random walk. An important observation is that $\tilde{P}(m; z)$ does not depend on N in the limit of large network size. The mean number of relevant nodes with two relevant inputs is therefore finite in this limit.

The underlying stochastic processes can be implemented directly on the computer in order to obtain the number of relevant nodes N_{rel} and the number of relevant nodes with two relevant inputs m, without the need for elaborate direct simulations of network dynamics. By randomly connecting the relevant nodes obtained in a simulation run, we generate a network from the ensemble of relevant subnetworks. There, a finite fraction ϵ of relevant nodes are arranged in simple loops. Simple loops of length l appear [13] with probability

$$P(l) = 1/l \text{ for } l < l_{\mathrm{c}} \quad \text{with} \quad l_{\mathrm{c}} \sim N_{\mathrm{rel}} \tag{9}$$

in the limit $N_{\mathrm{rel}} \to \infty$. The other $(1 - \epsilon)N_{\mathrm{rel}}$ relevant nodes sit on complex components containing nodes with two relevant inputs.

3. Number of complex relevant components

We argue in the following that most relevant nodes with two relevant inputs reside in one component. We calculate the probability that this is not the case and show it to be small and to decrease as $1/m$ for large m.

We neglect the cases where one relevant node has two relevant inputs and two relevant outputs, as well as where a node has more than two relevant outputs. The probability for such nodes vanishes in the limit of large network size. Then, the number of relevant nodes with two relevant outputs is identical to the number of relevant nodes with two relevant inputs m.

In the following, we first fix the value of m. Later, we will perform summations over m using its probability distribution $p(m)$. For a given m, the ensemble of relevant networks can be constructed by connecting the $2m$ nodes with two relevant inputs or two relevant outputs amongst one another, each realization of connections appearing with equal probability, and by subsequently inserting the nodes with one relevant input into the existing connections. The probability that the m nodes are not in the same component is thus the probability that the $2m$ nodes do not form a fully connected directed graph. We denote this probability by G_m, the corresponding number of possible connections of the $2m$ nodes is $N_m = (3m)!G_m$. The strategy for calculating N_m is the following. For some connection of the $2m$ nodes, which does not lead to a fully connected graph, the whole graph consists of two or more fully connected independent subgraphs. We consider all possible partitioning in such connected subgraphs, each partitioning contributing to N_m. The last contribution in the overall sum comes from a partitioning in m connected pairs of nodes, one with two inputs and one output and one with two outputs and one input. For a partitioning, the m nodes with two relevant inputs are arranged in n_i groups, each with i nodes, so that $m = \sum_i in_i$ and $\sum_i n_i > 1$. For a given partitioning, we count the number of ways to construct it, paying attention to avoid double counting. The general formula reads

$$N_m = (m!)^2 \sum_{\text{partitions}\{n_i|i\}} \prod_i \frac{((3i)! - N_i)^{n_i}}{n_i!(i!)^{2n_i}}, \qquad (10)$$
$$\text{with } N_m = (3m)!\, G_m,$$

since there are $((3i)! - N_i)$ ways to obtain a connected subgraph with $2i$ nodes. The sum is over all nontrivial partitions of m nodes in n_i groups of i nodes. We get $G_1 = 0$, $G_2 = 0.1$, $G_3 = 0.1$, $G_4 \approx 0.08$, $G_5 \approx 0.07$.

To estimate the asymptotic behaviour for large m, we note that the largest contribution for large m comes from the partitioning in one connected pair and a fully connected rest. We approximate G_m by the contribution from a partitioning in one connected pair and an arbitrarily connected rest: $6m^2(3(m-1))!/(3m)! \to 2/(9m)$ for large m. We conclude that all nodes with two relevant inputs are usually gathered in one complex component, even for small values of m.

4. Relative frequencies of topologically different components

The complex relevant components have nontrivial dynamics and can be associated with functional blocks in real genetic networks. So far, we characterized relevant complex components by the number of nodes with two relevant inputs found in them. Now, we want to consider the topology of such complex components. That is only possible for small values of m, for larger m one could look at a subset of all complex components, defined on some additional grounds. There seems to be no simple procedure to classify the different complex components according to their dynamics.

We consider two complex components as different if they possess a different topology. Components with the same topology but with different numbers of nodes in the different chains of relevant nodes with one relevant input, have similar dynamics, their main difference being

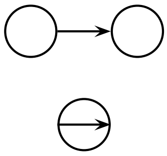

Figure 1. Schematic representation of the two topologically different relevant complex components with one node with two relevant inputs. Arcs stand for one link or a chain of nodes with one relevant input. Each crossing of arcs marks a node with two relevant inputs or outputs. Arrowheads depict the inputs of nodes with two relevant inputs; they appear only where they are important to distinguish the complex component. The first component appears in a network twice as often as the second one.

different time delays along chains of nodes. From our study of simple examples of complex components in [18], we know that changing the length of these chains does not change the types of attractors and the asymptotic dependence of the mean number and length of these attractors on the component size. Therefore, we restrict ourselves to consider all possible different interconnections of relevant nodes with two relevant inputs with relevant nodes with two relevant outputs, i.e., the different topologies, irrespective of the lengths of the chains between them.

We first consider complex components with one node with two relevant inputs, that is the case $m = 1$. The two different possibilities to construct these components were studied in [18] and are schematically shown in figure 1. In our simulations, the relative probability of obtaining two simple loops with a chain of nodes between them in a relevant network is $2/3$, and the relative probability of obtaining a simple loop with an extra link is $1/3$. For small values of m, these relative frequencies of the different possible complex components can easily be determined exactly. One has simply to count the number of different ways to connect the nodes with two relevant inputs or outputs out of the total number $(3m)!$ leading to the considered component. For the case of figure 1, the number of ways to construct the two components are 4 and 2 respectively.

The same counting of ways to construct complex components for $m = 2$ leads to the table presented in figure 2, whereas one has to be aware of misleading equivalent ways to represent the same component. To get the corresponding relative frequencies of occurrence one has to divide the multiplicity in figure 2 by 6!. The first three rows in the figure correspond to distributing nodes with two relevant inputs or outputs over more than one relevant component, compare discussion in section 3. The dynamical behaviour of the components in figure 2 can be studied in detail similarly to [18], if required. It is by analogy clear, though, that one would get many exponentially long attractors with increasing system size.

5. Distribution of complex components

The findings of section 3 together with the results from [13] cited in section 2 enable us to evaluate the mean number of relevant nodes with two relevant inputs, the probability distribution for the number of relevant nodes with two relevant inputs, and the mean number of relevant complex components per network with a given number of nodes with two relevant inputs.

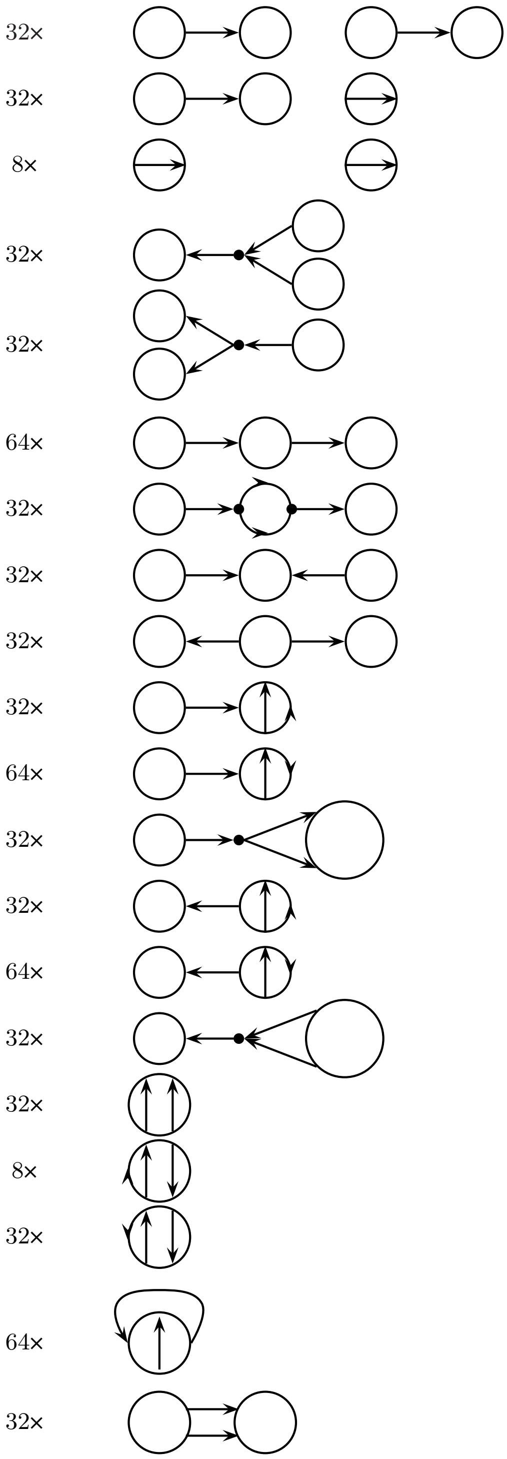

Figure 2. All possible topologically different components with $m = 2$, together with their multiplicities (number of ways to construct them). Arcs stand for one link or a chain of nodes with one relevant input. Each crossing of arcs marks a node with two relevant inputs or outputs. • represent exactly one node. Arrows depict the inputs of nodes with two relevant inputs, they appear only where they are important to distinguish the complex component.

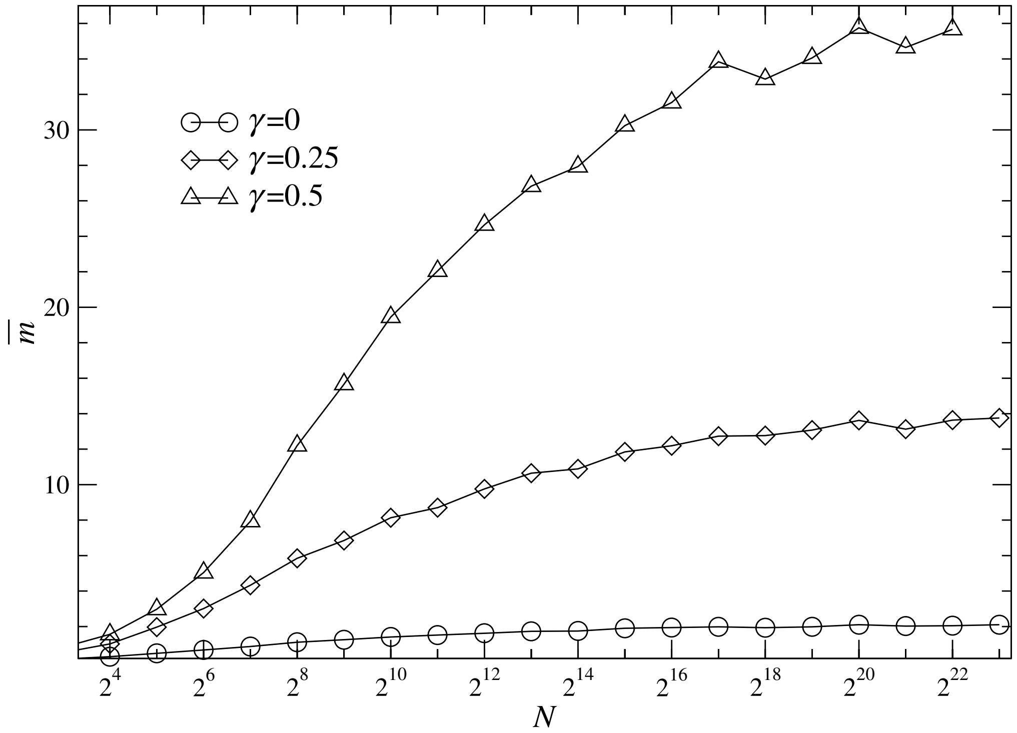

Figure 3. The mean number of nodes with two relevant inputs for $\beta = 1/8$ as function of the network size N. The averaging was performed over 10^4 network realizations.

Let us first evaluate the number of relevant nodes with two relevant inputs. Using the equations cited in section 2, we can write its probability distribution as

$$p(m) = \int_0^\infty \tilde{P}(m; z)\,\mathrm{d}z = \frac{1}{(m!)^2}\int_0^\infty \mathrm{d}z \frac{(z^3/2)^m P(z)}{I_0(\sqrt{2z^3})}, \tag{11}$$

where $I_n(z)$ denotes the modified Bessel function of the first kind and satisfies the relations $I_0(x) = \sum_{m=0}^{\infty}(\frac{x^m}{2^m m!})^2$ and $I_1(x) = 2/x\sum_{m=0}^{\infty} m(\frac{x^m}{2^m m!})^2$.

The mean number of nodes with two relevant inputs is then finite and is given by

$$\bar{m} = \frac{1}{2}\int_0^\infty \mathrm{d}z \frac{\sqrt{2z^3}\; I_1(\sqrt{2z^3})}{I_0(\sqrt{2z^3})} P(z)\,\mathrm{d}z. \tag{12}$$

The integral converges since $P(z)$ decreases exponentially for large z (compare [13]), and the rest of the integrand grows faster than linearly but slower than quadratically with z. Figure 3 shows results of computer simulations for the values of $\bar{m}$. With increasing N, they approach a constant value.

For further convenience, we introduce the notation $\kappa = 1 + \gamma/\beta$. The dependence of $\bar{m}$ and $p(m)$ on the model parameters β and γ is fully determined by the dependence of $P(z)$ on κ. Since $P(z)$ becomes broader for larger κ (therefore assuming smaller values at small z), see [13], $\bar{m}$ increases with κ. This means that increasing κ leads to more relevant nodes with two relevant inputs. We will look at the dependence of $p(m)$ on κ in more detail later.

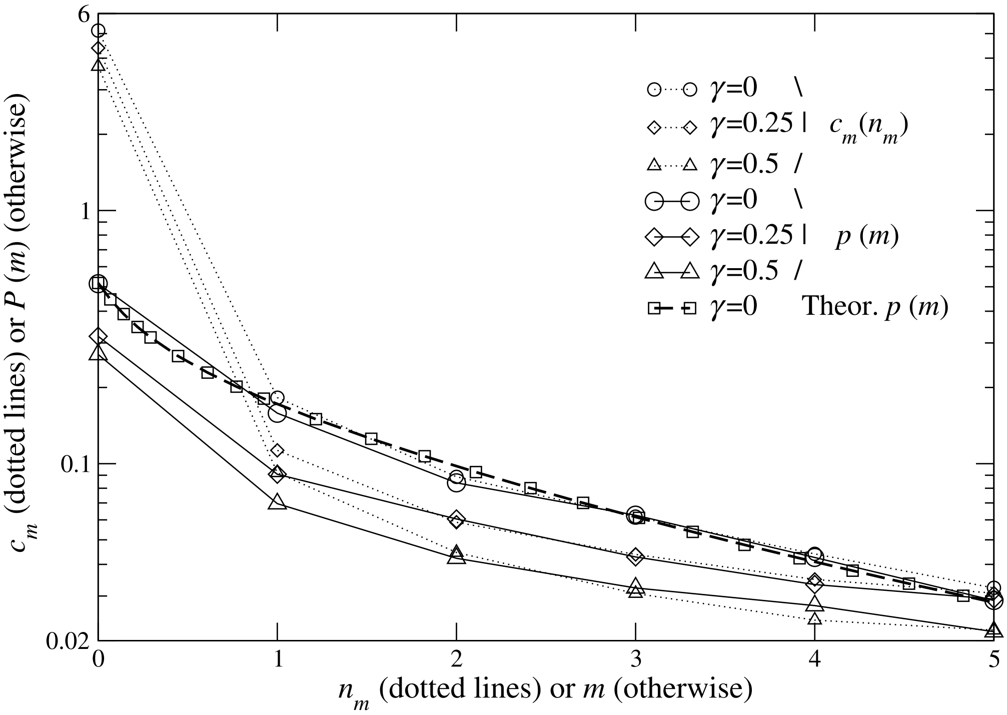

Figure 4. Simulation results for the number $c_m(n_m)$ of components per network with n_m nodes with two relevant inputs (dotted lines). Simulation results for the probability distribution $p(m)$ for the number m of nodes with two relevant inputs per network (solid lines). Theoretical prediction for the probability distribution $p(m)$ (dashed line). The model parameters used are $\beta = 1/8$ and $\gamma = 0$ (○, □) or $\gamma = 1/4$ (◇) or $\gamma = 1/2$ (△). All simulations were run with 8000 different network realizations and results were averaged afterwards. Network size was $N \sim 10^7$ (dotted lines), $N \sim 10^5$ (solid lines). Approximation $P^{\mathrm{fit}}_{\gamma=0}(z) = 0.62\ \mathrm{e}^{-0.65z}$ was used to obtain the dashed curve.

We turn to the discussion of the connection between $p(m)$ and the number of components with a given number of nodes with two relevant inputs. From section 3, we know that, except for a small correction, most nodes with two relevant inputs reside in one complex component. Therefore, the fraction of number of networks, where we find a given value of m, will simultaneously be the fraction of number of networks where we find complex components (one complex component) with this number of nodes with two relevant inputs. In figure 4, the dotted lines represent the mean number c_m of components per network with a given number n_m of nodes with two relevant inputs, obtained in our computer simulations by counting the number of such components in the generated network ensemble for different model parameters. To verify that the relevant nodes with two relevant inputs usually sit in the same component, we also evaluated the probability distribution $p(m)$ numerically (solid lines in figure 4). For $m > 1$ and $n_m > 1$, the difference between the solid and the corresponding dotted lines in figure 4 is mainly due to statistical fluctuations. They become larger for larger networks, we therefore used comparatively small networks with $N \sim 10^5$ to obtain the solid lines. In the rare network realizations with two complex relevant components (appearing more often with increasing $\bar{m}$) the smaller complex

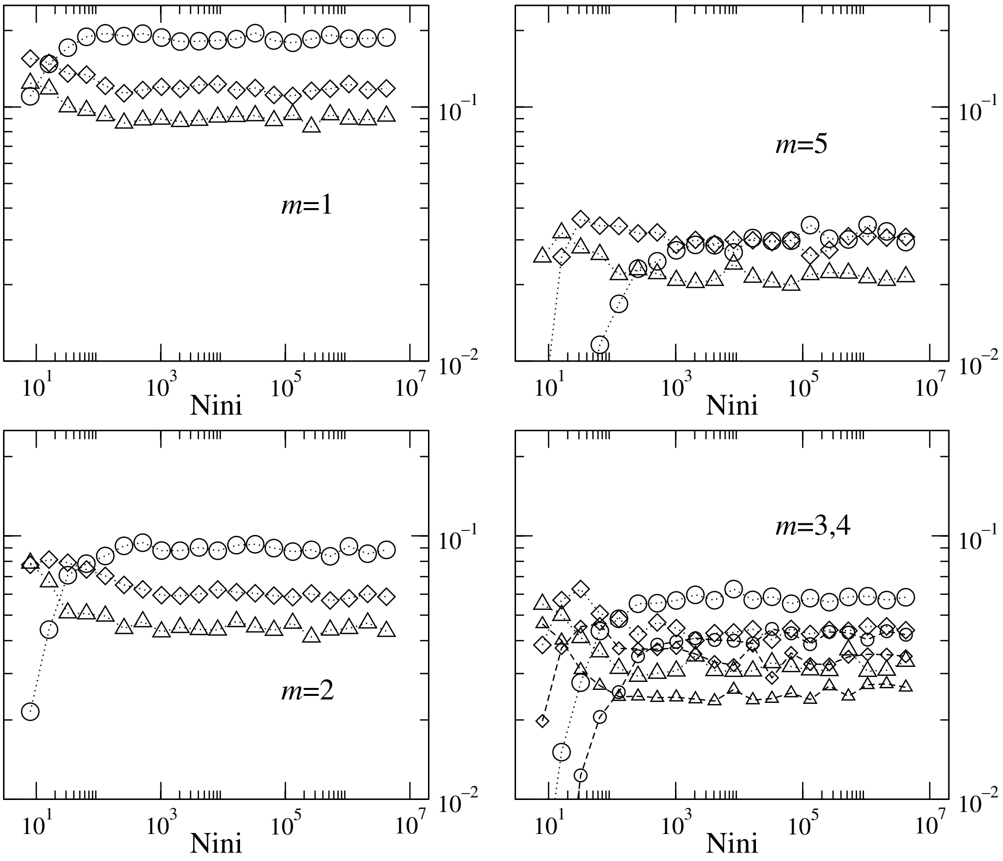

Figure 5. The number of relevant complex components with m nodes with two relevant inputs is constant for large networks. In the figure, ○ , ◇ and △ correspond to model parameters $\beta = 1/8$ and $\gamma = 0$, $1/4$, $1/2$ respectively. All simulations were run with 8000 different network realizations.

component has mostly one relevant node with two relevant inputs ($m = 1$), see the data points for $m = n_m = 1$ in figure 4. Finally, we calculated $p(m)$ analytically with (11). The theoretical dashed curve in figure 4 agrees well with the simulation results. We only calculated analytical results for $\gamma = 0$. That the number of components with a given value of m is indeed independent of the system size is shown in figure 5.

We now complete our understanding of simulation results in figure 4 and extract some data that is needed in the next section 6. For $m = 0$, the solid lines represent the fraction of networks without nodes with two relevant inputs. We will use the results for these fractions 0.51, 0.32, 0.27 in figure 6. Network realizations without relevant nodes with two relevant inputs are indistinguishable from networks in the model with $K = 1$. In this context we want to mention that the fraction of 'frozen' networks without relevant nodes $\sim(\beta/N)^{1/3}$ is negligible in the limit $N \rightarrow \infty$, whereas $\lim_{z\rightarrow 0} P(z) \neq 0$. The dotted curves at $n_m = 0$ represent the number of simple loops per network. For $\gamma = 0$, 0.25, 0.5 this number is 5.1, 4.4, 3.7 respectively. These values will be shown to be in good agreement with those calculated from analytical arguments in the next section, see figure 6.

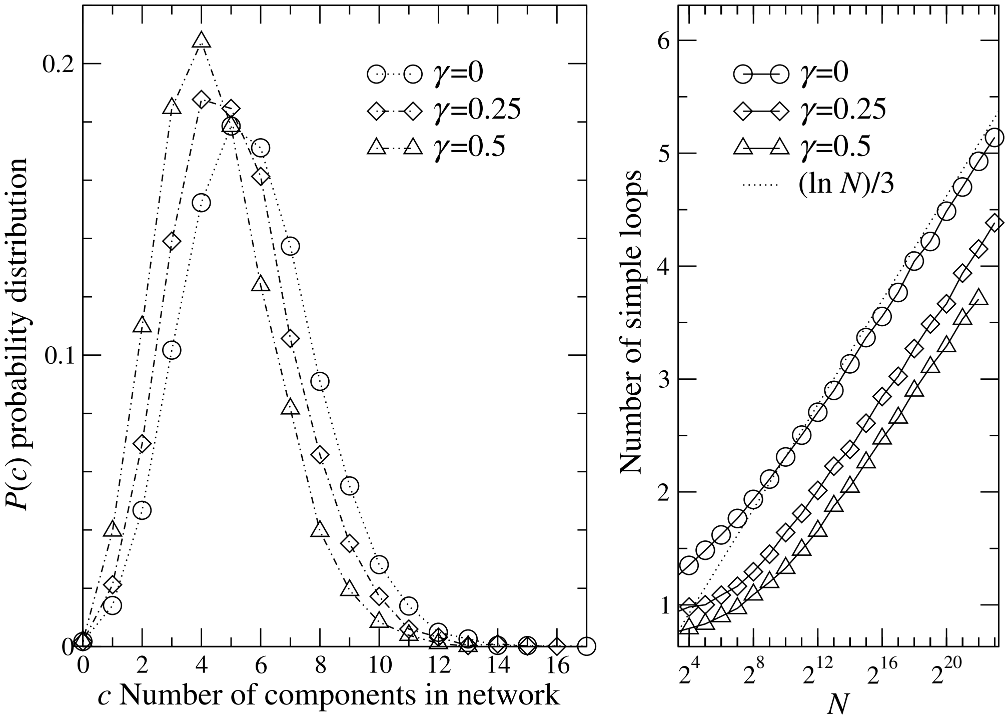

Figure 6. *Left panel*: probability distribution for the number of relevant components per network. Results were averaged over 10^4 network realizations for the following model parameters: $\beta = 1/8$; $\gamma = 0, 1/4, 1/2$; $N = 2^{23} \approx 8 \times 10^6$ for $\gamma \neq 1/2$ and $N = 2^{22}$ for $\gamma = 1/2$. Obtained in simulations results for mean numbers of components for $\gamma = 0, 1/4, 1/2$ are 5.7, 5.1, 4.5 respectively. Using results of simulations or analytical results for $p(m)$, see section 5 for details, and (17) we get 5.5, 4.5, 3.6. The agreement between theoretical estimation and simulations is satisfactory. Half maximal widths of the simulation curves are 5.3, 5.0, 4.6 compared to values 4.43, 4.43, 4.32, determined from (14), whereby $\bar{m}$-dependence has been ignored. *Right panel*: number of simple loop components per network in $K = 2$ critical RBNs for the same sets of model parameters and numerical simulation parameters as in the left panel. Our simulation results for the mean number of simple loop components in the largest considered networks for $\gamma = 0, 1/4, 1/2$ are 5.1, 4.4, 3.7 respectively. From the estimation (17) we get 5.0, 4.2, 3.3, since all nodes with two relevant inputs are assumed to reside in one complex component as explained in the main text.

Let us conclude this section by discussing the dependence of $p(m)$ on the parameter κ. For fixed κ, $p(m)$ is a monotonously decreasing function, which is flatter for larger κ. For fixed m, the value $p(m)$ as function of κ has a maximum, the position of which moves to larger κ with increasing m. We see an indication of this in figures 4 and 5, where the data points for $\kappa = 1$ are not the highest ones when m is larger than 4. Stronger support comes from an analysis of Equation (11). The integrand apart from $P(z)$ has one global maximum, which shifts to larger values of z as m increases. The function $P(z)$ is monotonically decreasing, being flatter and starting at a smaller value for larger κ.

From these properties of the two contributions to the integrand, we find immediately that the value $p(m)$ at a given m as function of κ has a maximum, which moves to larger κ with increasing m. For very large κ, the function $P(z)$ in (11) can be approximated [13] by its value $P(z=0) = \sqrt{2\pi}/(4\kappa^{1/3})$ for not too large m, revealing how $p(m)$ decreases with increasing κ. From these considerations it follows that complex components with a small number of nodes with two relevant inputs will appear with smaller probability if we increase κ, while complex components with larger $m \sim \bar{m}$ occur more often.

6. Total number of relevant components

We build the relevant components by first ignoring one input of each nonfrozen node that has two inputs. The second inputs will be connected later. We thus first build a $K = 1$ critical network from the nonfrozen nodes, many properties of which are known from the literature. In particular, all nonfrozen nodes will at this stage be arranged in

$$C \simeq 1/2[\ln(2N_{\mathrm{nf}}) + \gamma_{\mathrm{E}}] \tag{13}$$

simple loops and trees rooted in loops [22]. $\gamma_{\mathrm{E}} \approx 0.577$ is the Euler–Mascheroni constant. The width of the distribution of the number of loops is asymptotically [22]

$$\sigma^2 = C - \pi^2/8. \tag{14}$$

Equations (13) and (14) are valid for fixed values of N_{nf}. If we insert for N_{nf} the mean value $\bar{N}_{\mathrm{nf}} \approx 0.62(N/\beta)^{2/3}$ (see (2) here or equation (12) in [13]), we get $\bar{C} \approx 6.4$ for $N = 2^{23}$. If we use the full probability distribution for the number of nonfrozen nodes N_{nf} and the number of simple loops C, we obtain the following formula for the mean number of simple loops:

$$\bar{C}(N) \simeq \sum_{\mu=1}^{N} \mu \sum_{L=1}^{N} \frac{1}{L!} \begin{bmatrix} L \\ \mu \end{bmatrix} \sum_{N_{\mathrm{nf}}=1}^{N} (N/\beta)^{-2/3} G\left[N_{\mathrm{nf}}(N/\beta)^{-2/3}\right] \binom{N_{\mathrm{nf}}}{L} \frac{L\,L!}{N_{\mathrm{nf}}^{L+1}}, \tag{15}$$

which is a combination of exact results (II.C16)[1] and (II.C17)[2] for the ensemble of $K = 1$ networks from [22], and of (2). We designed a program to effectively evaluate (15) numerically. For the above-mentioned system size $N = 2^{23}$, we get $\bar{C} = 6.13$. We use this value in the following discussion. Note that using mean values instead of full probability distributions gives good estimations of the order of magnitude in the present context.

Now we estimate the number of components out of the $\bar{C}$ simple loop components on which the relevant nodes with two relevant inputs and the nodes with two relevant outputs sit. By assuming that all of them will end up on the same complex component when the second inputs will be connected, we obtain our desired result for the mean number of relevant components. In the following, we approximate the mean value of some functions by the functions of the mean values, so that the results (16) and (17) are valid approximately.

Because of (9), the mean number of components with loops of lengths from logarithmically constant intervals will be constant. If l_1 and l_2 denote the lower and upper boundary of an interval,

[1] For a fixed N_{nf} the distribution for the number L of nodes on loops: $P(L) = N_{\mathrm{nf}}!L \;/\; N_{\mathrm{nf}}^{L+1} \;/\; (N_{\mathrm{nf}} - L)!$.

[2] For a fixed N_{nf} the joined distribution for L and the number μ of loops: $P(\mu, L) = \left[{L \atop \mu}\right]/L!$.

there is on an average one loop in an interval of logarithmic size $l_2/l_1 = \mathrm{e}$. We consider such intervals with the upper boundaries l_2 equal to e, e^2, and so on. For two neighbouring intervals, the factor between the numbers of nodes on the loops is e on an average. Since for an interval the total number of nodes on the loops and on their trees is proportional to the number of nodes on the loops, the factor between the total numbers of nodes in two neighbouring intervals is also e. With this assumption, the proportion of nodes in the pth largest component is $\lambda \mathrm{e}^{-(p-1)}$, with $\lambda = 1 - 1/\mathrm{e} \approx 0.632$. We assumed that $\mathrm{e}^{-\bar{C}} \ll 1$. The value $\lambda = 1 - 1/\mathrm{e}$ is not far from the numerically determined value $\lambda \approx 0.624$ for the proportion of nodes in the largest component in ensembles of $K = 1$ networks with any large but fixed number of nonfrozen nodes. Let us now turn to the m relevant nodes with two relevant inputs, with one input being ignored. The number of nodes with two relevant inputs or two relevant outputs in the largest nonfrozen component is then of the order $2m\lambda$, and the smaller components will contain $2m\lambda/\mathrm{e}, 2m\lambda/\mathrm{e}^2, \ldots, 1$ nodes with two relevant inputs. For the purpose of estimation, we neglect further the fact that the distribution of the number of nodes with two relevant inputs $p(m)$ depends [13] on the number of relevant nodes and, therefore, correlates with the number of nonfrozen nodes on the loops. The total number $n(m)$ of nonfrozen components with nodes with two relevant inputs (one of them is cut off) is consequently obtained from the condition $2m\lambda\,\mathrm{e}^{-(n(m)-1)} = 1$ for $m \geqslant 1$. For $m = 1$, this gives $n(1) \simeq 1.234$. The exact result $n(1) = 4/3$ is within 10% of this estimate. To explain the exact result for fixed $m = 1$, we consider all possible relevant components containing the one node with two relevant inputs, which result after connecting the cut off input. They are represented in figure 1. Clearly, the two relevant inputs could have been cut off with the same probability. As explained in section 4, relative frequency of occurrence for the second component in figure 1 is $1/3$, and cutting one input off the node with two inputs leads to one nonfrozen component. Such components have therefore resulted from one nonfrozen component. For the first component in figure 1, which appears with relative probability $2/3$, cutting one input leads to two nonfrozen components with probability $1/2$. The mean number of $K = 1$ nonfrozen components connected by the relevant node with two relevant inputs is then $1 * 2/3 + 2 * 1/3 = 4/3$.

Taking all results together, we obtain the following estimate for the mean number of nonfrozen components with relevant nodes with two relevant inputs

$$\bar{n} = \sum_{m=1}^{N} p(m)n(m) \approx \sum_{m=1}^{N} p(m) \ln\left[2m(\mathrm{e}-1)\right]. \tag{16}$$

We recall that (almost) all nodes with two relevant inputs are gathered in one complex component, and with (16) we get an estimation for the mean number of relevant components, which is valid for large networks:

$$C_{\mathrm{rel}}^{K=2}(N) \approx \bar{C}(N) - \sum_{m=1}^{N} p(m)(n(m)-1) = \bar{C}(N) - \overline{\ln m} - [1 - p(0)]\ln(2\lambda). \tag{17}$$

We support (17) by evaluating the number of relevant components in numerical simulations, see figure 6. For comparison with (17), some results from section 5 have been used, see figure 4 there.

The average number of complex components per network can be extracted from results presented in figure 6 as the difference between the number of relevant components and the number of simple loop components. For model parameters $\beta = 1/8$ and $\gamma = 0, 1/4, 1/2$ we get

0.6, 0.7 and 0.8 complex components on average respectively. For the three cases, in section 5 we have seen that the probability to have no nodes with two relevant inputs in a network is 0.51, 0.32 and 0.27, so that the average number of complex components in the networks with complex components (as opposed to all networks) evaluates to 1.2, 1.0 and 1.1 respectively, justifying the assumption that there is usually no more than one complex component. Statistical errors of simulation results consistently allow for errors in the first digit after the comma.

The difference between simulation results in figure 6 and the results from (17) is smaller for smaller values of γ, where the approximations used to obtain (17) are intuitively better. An estimation in the spirit of (16) and (17), which would take into account that the probability for relevant nodes with two relevant inputs to be arranged in more than one relevant complex component is small but greater than zero, would not change (16), but it would lead to slightly smaller values of $C_{\mathrm{rel}}^{K=2}(N, \beta, \gamma)$ in (17). This observation partially explains the difference between the simulation and the estimation results in figure 6.

7. Conclusions

In this paper, we have investigated the properties of relevant components of critical RBNs. We used an efficient numerical method based on [13] to create a sufficiently large ensemble of the relevant parts of sufficiently large networks in order to evaluate the number of relevant components, the number of relevant nodes with two relevant inputs, and the number of components with a given number of nodes with two relevant inputs. The results are in agreement with theoretical predictions made in [13], and they are supplemented by additional analytical results. Our main findings are the following:

1. The number of relevant components increases logarithmically with the system size, and usually only the largest relevant component is complex, i.e., is not a simple loop.
2. For large network sizes, the number of relevant nodes with two relevant inputs and the relative frequencies of different types of complex components become independent of the network size.
3. The relative frequency of topologically different complex components with the same number of nodes with two relevant inputs can be obtained from simple combinatorial considerations.
4. At constant network size, when the mean number of nonfrozen nodes with two relevant inputs grows (that is, for larger $\kappa = 1 + \gamma/\beta$), the number of relevant components decreases roughly by the mean value of the logarithm of the number of relevant nodes with two relevant inputs, while the size of the largest relevant component and the number of nodes with two relevant inputs in this component becomes larger.

Since the relevant part constitutes only a vanishing portion of the network (the fraction $N^{-2/3}$ of all nodes), and since the different relevant components change their state independently of each other, the topology of considered types of networks is most likely very different from the topology of real biological networks, such as genetic regulatory networks, where one would expect that the majority of nodes is relevant or at least not always frozen, and that different parts of the network are not decoupled from each other.

We considered networks with of the order of 10^6 nodes, which is far beyond the number of genes in real regulatory circles. On the other hand, it might be more appropriate to identify the genetic regulatory systems with the relevant nodes, their number being of the order of 100 in our simulations. From this point of view it is interesting to look at specialized complex components. We have not addressed the question of the *function* of a complex component in biological context, which has to be considered together with a suitable definition of the environment and together with an appropriate choice of update rules. A lot of study still needs to be done before real genetic regulatory and other biological networks can be modelled.

References

[1] Kauffman S, Peterson C, Samuelsson B and Troein C 2003 *Proc. Natl Acad. Sci. USA* **100** 14796

[2] Kauffman S A 1969 *J. Theor. Biol.* **22** 437

[3] Derrida B and Pomeau Y 1986 *Europhys. Lett.* **1** 45

[4] Derrida B and Stauffer D 1986 *Europhys. Lett.* **2** 739

[5] Moreira A A and Amaral L A N 2005 *Phys. Rev. Lett.* **94** 218702

[6] Aldana-Gonzalez M, Coppersmith S and Kadanoff L P 2003 *Perspectives and Problems in Nonlinear Science, A Celebratory Volume in Honor of Lawrence Sirovich* ed E Kaplan, J E Mardsen and K R Sreenivasan (*Springer Applied Mathematical Sciences Series* May) (New York: Springer) pp 23–89

[7] Bilke S and Sjunnesson F 2002 *Phys. Rev.* E **65** 016129

[8] Socolar J E S and Kauffman S A 2003 *Phys. Rev. Lett.* **90** 068702

[9] Samuelsson B and Troein C 2003 *Phys. Rev. Lett.* **90** 098701

[10] Gershenson C 2004 *Artificial Life IX, Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems* ed J Pollack, M Bedau, P Husbands, T Ikegami and R A Watson (Cambridge, MA: MIT Press) pp 238–43

[11] Klemm K and Bornholdt S 2005 *Phys. Rev.* E **72** 055101

[12] Drossel B 2005 *Phys. Rev.* E **72** 016110

[13] Kaufman V, Mihaljev T and Drossel B 2005 *Phys. Rev.* E **72** 046124

[14] Paul U, Kaufman V and Drossel B 2006 *Phys. Rev.* E **73** 026118

[15] Samuelsson B and Socolar J E S 2006 *Preprint* at http://arxiv.org/abs/nlin.CG/0605047

[16] Flyvbjerg H and Kjær N J 1988 *J. Phys. A: Math. Gen.* **21** 1695

[17] Drossel B, Mihaljev T and Greil F 2005 *Phys. Rev. Lett.* **94** 088701

[18] Kaufman V and Drossel B 2005 *Eur. Phys. J.* B **43** 115

[19] Li F, Long T, Lu Y, Ouyang Q and Tang C 2004 *Proc. Natl Acad. Sci. USA* **101** 4781 http://www.pnas.org/cgi/content/abstract/101/14/4781

[20] Greil F and Drossel B 2005 *Phys. Rev. Lett.* **95** 048701

[21] Klemm K and Bornholdt S 2005 *Proc. Natl Acad. Sci. USA* **102** 18414 http://www.pnas.org/cgi/content/abstract/102/51/18414

[22] Samuelsson B and Troein C 2005 *Phys. Rev.* E **72** 046112