Entropy of complex relevant components of Boolean networks

Peter Krawitz

Institute for Systems Biology, Seattle, Washington 98103, USA
and Fakultät für Physik, Ludwig Maximilians Universität, 80799 München, Germany

Ilya Shmulevich

Institute for Systems Biology, Seattle, Washington 98103, USA

Boolean network models of strongly connected modules are capable of capturing the high regulatory complexity of many biological gene regulatory circuits. We study numerically the previously introduced basin entropy, a parameter for the dynamical uncertainty or information storage capacity of a network as well as the average transient time in random relevant components as a function of their connectivity. We also demonstrate that basin entropy can be estimated from time-series data and is therefore also applicable to nondeterministic networks models.

PACS number(s): 89.75.Hc, 89.75.Fb, 89.75.Da, 05.45.-a

I. INTRODUCTION

Random Boolean networks are often studied as generic models of gene regulatory networks [1,2]. The ensemble approach to gene regulation, a term coined by Kauffman, aims at studying well-defined ensembles of model networks, the statistical features of which match those of real cells and organisms [3]. Ensembles of special biological interest are critical random Boolean networks, which lie at the boundary between a frozen and a chaotic phase [4–6]. In the frozen phase, a perturbation to one node propagates on average to less than one other node during one time step. In the chaotic phase, the difference between two almost identical states increases exponentially fast, since a perturbation propagates on average to more than one node during one time step [7]. Critical Boolean networks balance robustness against perturbations with complex asymptotic attractor dynamics.

Since Boolean networks are deterministic systems, they eventually settle into periodic attractors. Regarding the behavior in this asymptotic limit, nodes can be classified into three groups. Nodes that are frozen to the same value on every attractor form the frozen core of a network [8]. These nodes give a constant input to other nodes and are otherwise irrelevant. Nodes that are not frozen but have no outputs, or only outputs to other irrelevant nodes, are also classified as irrelevant. Although they might fluctuate, they do not influence the number and periods of attractors. Finally, the relevant nodes are those nonfrozen nodes that influence their own state over directed loops. The recognition of the relevant nodes as the only elements influencing the asymptotic dynamics was an important step in understanding the dynamics of Boolean networks [9,10].

In a biological context, an attractor is associated with a characteristic dynamically stable pattern of gene expression, which may represent a particular fate of the cell. The weight of each such attractor can be defined as the probability for a random state in the state space of the network to flow into this attractor. Based on the state space partition into attractors of different weights, we recently introduced a network parameter, called the basin entropy (hereafter, simply entropy), which measures the uncertainty of the future behavior of a random state. This entropy was shown to increase with

system size in critical network ensembles, whereas for ensembles in the ordered phase and in highly chaotic networks, it approaches a finite value [11]. From an informational processing perspective, this means that the complexity of a network increases only with its system size if it operates near the critical regime.

An intuitive understanding of this growing complexity are networks whose relevant nodes are modularly organized and whose complexity increases if new modules accumulate. In living systems, transcriptional regulation is often highly modular [12,13]. Of special interest are complex relevant components, which consist of relevant nodes containing more than one relevant input or output [14]. Boolean network models for several biological gene regulatory circuits have been constructed and were shown to reproduce experimentally observed results [15–18]. These often highly connected subnetworks can be viewed as biological examples of such complex relevant components.

In this work, we first numerically study the entropy of complex relevant components as a function of their connectivity. We show that the probability of such a component to freeze increases with growing connectivity and that its entropy decreases. Additionally, we also study the average transient time of a random state until it falls into its attractor. The calculation of dynamic network parameters, such as the basin entropy, requires the knowledge of the underlying network functions. This often limits the applicability of such parameters. We demonstrate that the entropy of a network can also be estimated from time series data by the application of clustering techniques. This broadens the applicability of this dynamic network parameter to models whose functions are not necessarily known or that are not deterministic. In order to illustrate our results, we will use a Boolean network model for the mammalian cell cycle as presented in [16].

II. BOOLEAN NETWORKS

In a Boolean network every gene is identified by a node i, whose state $x_i \in \{0,1\}$ indicates whether the gene is switched on or off. Each node i receives input from k_i other nodes and its state at the next time step $t+1$ is a Boolean function f_i of

the states of the nodes it is depending on $x_i(t+1)$ $=f_i(x_{i_1}(t), \ldots, x_{i_{k_i}}(t))$. A network is entirely defined by its directed connectivity graph G and the Boolean functions $\mathbf{F}$ $=\{f_1, \ldots, f_n\}$ assigned to every node. The state of a network $\mathbf{x}(t)=(x_1(t), \ldots, x_n(t))$ contains the values of all n nodes in the network at a given time point t. A Boolean network thus defines a deterministic transition matrix $\mathbf{T}$ on its state space. A random state $\mathbf{x} \in \{0,1\}^n$ that is propagated through the network as

$$\mathbf{x}(t+1) = \mathbf{F}(\mathbf{x}(t)) \qquad (1)$$

generates a time series or trajectory through the state space that finally intersects with itself. The states that are then revisited infinitely often define the attractor ρ, with the number of states on the attractor equal to $l(\rho)$, also called the attractor size. Transient states that lead into an attractor form its basin of attraction. The sum of all attractor states and basin states of a certain attractor normalized by the size of the entire state space (2^n) define the weight w_ρ of that attractor. The weight of an attractor is the probability that a randomly chosen state will flow into this very attractor. The entropy h of the probability distribution defined by w_ρ is called the *basin entropy*:

$$h = - \sum_\rho w_\rho \ln w_\rho. \qquad (2)$$

This measure captures the uncertainty of the future dynamic behavior of the system started in a random state. The scaling of this parameter relative to system size was discussed in [11] for ensembles of different dynamical regimes.

The sensitivity of a node i specifies the number of its input arguments that, when toggled, will result in a flip in the value of that node, averaged over all input combinations. A Boolean function f_i with k_i input variables that takes on the value 1 for any one of its possible 2^{k_i} input vectors with probability p_i has the expected sensitivity [19]:

$$s_i = 2k_i p_i (1 - p_i). \qquad (3)$$

The network sensitivity s is the average sensitivity of all its nodes and indicates to how many nodes a perturbation to a single node is, on average, propagated. The network sensitivity is an order parameter of a network ensemble that divides random Boolean networks with an expected sensitivity of $s < 1$ into the ordered phase and with $s > 1$ into the chaotic phase. Random networks that propagate a perturbation, on average, to one other node ($s=1$), are called critical. Classical Kauffman networks with a fixed in degree $k=2$ and $p=0.5$ are prototypes of critical Boolean network ensembles, though it should be noted that this definition of criticality is independent of the actual in degree distribution.

The attractor dynamics of a Boolean network are entirely determined by its relevant nodes. The scaling of these nodes was first discussed in [20] and derived for a general class of critical network ensembles by Drossel, Kaufman, and Mihaljev [21,22]. They presented a stochastic process that removes frozen nodes and nonfrozen irrelevant nodes from a network and ends when there are only potentially relevant nodes left. We call these nodes "potentially relevant," as some of them
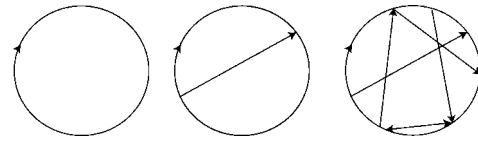

FIG. 1. Random complex component of excess $e=0$, 1, 6 from left to right.

may be part of self-freezing loops. We will discuss this phenomenon below in more detail. The number of these potentially relevant nodes n_r scales in all critical networks of size n and average connectivity $k > 1$ as $n_r \sim n^{1/3}$. In the limit of large n, almost all these relevant nodes depend on only one relevant node; the *proportion* of relevant nodes depending on more than one relevant input approaches zero; and the *number* of nodes depending on more than two relevant inputs is almost surely zero. These results agree with structural findings of random graphs at the point of phase transition, where the number of nodes in complex components scales with $n^{1/3}$ and each such complex component is almost surely topologically equivalent to a three-regular multigraph [23].

III. RANDOM COMPLEX RELEVANT COMPONENTS

A set of relevant nodes that eventually influence each other's state form a relevant component. We define the excess e of the connection digraph of a component in analogy to the graph theoretic terminology as the difference between the number of arrows a between the nodes and the number of nodes n:

$$e = a - n. \qquad (4)$$

The topology of the simplest relevant component is a loop of nodes. The excess of this component is $e=0$. If we randomly add a different link to this loop we have an intersected loop of the same size with excess $e=1$. One node now depends on two nodes and one node influences two nodes.[1]

By randomly adding further edges, we may generate components of arbitrary excess (see Fig. 1).

In a simple relevant loop only the Boolean "copy" ($x_i(t+1)=x_j(t)$) and "negation" ($x_i(t+1)=\overline{x_j(t)}$) functions can be used. A perturbation in a loop node is always propagated to a single successor node and therefore such components have sensitivity $s=1$. If another edge is added, the Boolean function of the node receiving input from two relevant inputs has to be changed. Generally, whenever the in degree of a node increases from k to $k+1$ variables, a new Boolean updating function for $k+1$ variables has to be assigned. By adjusting the bias p [see Eq. (3)], we can randomly generate a Boolean function of $k+1$ variables so that its expected sensitivity is $E(s)=1$. If any of the $k+1$ variables happen to be fictitious (i.e., toggling their value has no effect on the output), we can draw again in order to guarantee that all added edges are

---

[1] In the Boolean network literature only relevant components with an excess $e > 0$ are called complex components. This might cause confusion for readers familiar with the graph theoretic terminology, where a component is called complex if its excess is $e \geq 0$.

irreducible. This process thus generates random components of arbitrary excess, operating in the critical or slightly supracritical regime.

IV. ENTROPY OF RELEVANT COMPONENTS

The only relevant component whose entropy we can discuss analytically is the simple loop. In simple loops all states are attractor states and it is therefore sufficient to know the length $l(\rho)$ of an attractor in order to determine its weight, $w_\rho = l(\rho)/2^n$. Regarding the attractor dynamics, we can substitute an even number of "negation" functions in a loop by "copy" functions, so that it is sufficient to discuss loops with an even number or an odd number of "negations." In even loops, the attractor states that only differ by a rotation of their values belong to the same attractor and attractors can therefore be viewed as an equivalence class under rotation. In combinatorics such an equivalence class is also known as a binary necklace of length n. The number of attractors of a simple even loop is calculated as follows [24]:

$$N_A^{\text{even}}(n) = \frac{1}{n} \sum_{d|n} \phi\left(\frac{n}{d}\right) 2^d, \tag{5}$$

where $\phi(m)$ is Euler's totient function, which is defined as the number of positive integers $\leq m$ that are relatively prime to m (i.e., do not contain any factor in common with m). The sum is taken over all dividers of n. If n is prime, the number of attractors of length n is simply $(2^n - 2)/n$.

In odd loops, a state $\mathbf{x}$ and its negation $\bar{\mathbf{x}}$ are always part of the same attractor and the number of attractors can be calculated with the formula

$$N_A^{\text{odd}}(n) = \begin{cases} \frac{1}{2n} \sum_{d|n} \phi(n/d) 2^d + \frac{3}{4} 2^{n/2}, & \text{if } n \text{ is even,} \\ \frac{1}{2n} \sum_{d|n} \phi(n/d) 2^d + 2^{n-1/2}, & \text{if } n \text{ is odd.} \end{cases} \tag{6}$$

For n prime, the number of attractors of length $2n$ is $(2^n - 1)/2n$. For large even (odd) loops, most attractors are of length n ($2n$) and the entropy can be approximated by considering only those attractors that are of maximal weight: $h(n) \approx n \ln 2 - O(\ln n)$. Therefore, the entropy of simple loops scales linearly with its size under synchronous updating.

In terms of the entropy, the key difference between complex components with excess $e \geq 1$ compared to simple loops is that the attractor length and weight are no longer correlated. In complex components, attractors of length one may have even higher weights than larger attractors in the same network. The mean number and length of attractors of a random component are insufficient in describing its dynamic complexity and we choose to study the entropy as a function of increasing excess.[2]
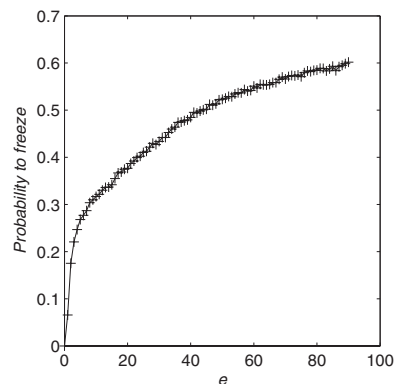

FIG. 2. The probability of a $n=10$ node component to freeze increases roughly logarithmically with increasing excess e.

When we start increasing the excess of our component from $e=0$, two qualitatively different things may happen: Either all nodes stay relevant and only the attractor dynamics change, or parts of the component or perhaps even the entire component freezes. This special case of self-freezing loops was first discussed in [26]. The simplest case of a *self-freezing* component are two nodes that canalize each other to their majority bits, e.g., $f_1 = x_1 \vee x_2$ and $f_2 = x_1 \vee \bar{x}_2$. Such components are clearly not relevant components, but part of the frozen core of a network. In Fig. 2 the probability that a critical component of $n=10$ nodes freezes is shown for increasing excess e. The average was taken over more than 26 000 network instances for every excess e. We obtained the same qualitative progress for component sizes up to $n = 18$:[3] The addition of the first few edges to a loop strongly increases the probability to freeze, whereas the probability to freeze increases slower in components of already high excess.

If a component becomes frozen, or only has a single attractor, it has entropy $h=0$ and we will not consider it as a relevant component. Figure 3 shows the decline in the average entropy $\langle h \rangle$ of the relevant $n=10$ node components as a function of their excess. The entropy drops sharply for the first few additional edges and decreases slower for $e > 10$. For all studied component sizes n up to 18, the average entropy falls below 1 for $e=10$ and continues to decrease slower thereafter.

As soon as additional edges are introduced into the loop, the average number of attractor states drops substantially and the majority of states become transient states. The average number of steps that are required to reach an attractor from a random state in the state space is defined as the average transient time $\langle \chi \rangle$ of a network. In a simple loop, where all states are attractor states, the transient time is zero. If we increase the excess, the transient time first sharply increases, peaks around an excess of $e=2$, $\langle \chi \rangle_{n=10}^{\max}(e=3) = 10$, and begins to decrease thereafter as shown in Fig. 4. We obtain

---

[2]For a thorough and mainly analytical discussion of the mean number and length of attractors in relevant components of excess $e=1$, see [25].

[3]The computational time to calculate the entropy increases exponentially with system size. We therefore chose larger increments for the excess in larger components: $n=12,14,16,18$, $e=1,2,\ldots,10,15,20,\ldots,90$. Over 2000 network instances have been simulated for every ensemble.
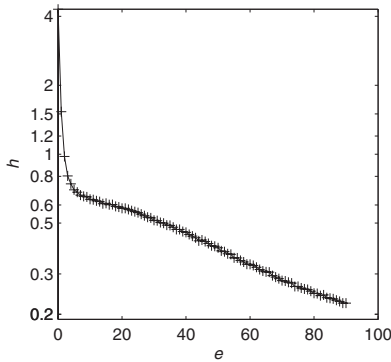
FIG. 3. Average entropy $\langle h \rangle$ of a complex relevant component with increasing excess e.

Gene	Boolean function
CycD	$x_1 = x_1$
Rb	$x_2 = (\overline{x_4 x_5} + x_6)\overline{x_1 x_{10}}$
E2F	$x_3 = (\overline{x_5} + x_6)\overline{x_2 x_{10}}$
CycE	$x_4 = x_3 \overline{x_2}$
CycA	$x_5 = (x_3 + x_5)(\overline{x_5} + x_8)\overline{x_2 x_7}$
p27	$x_6 = (\overline{x_4 x_5} + \overline{x_6 x_4} + x_6 \overline{x_5})\overline{x_1 x_{10}}$
Cdc20	$x_7 = x_{10}$
Cdh1	$x_8 = \overline{x_5 x_{10}} + x_7 + x_6 \overline{x_{10}}$
UbcH10	$x_9 = \overline{x_8} + x_8 x_9(x_7 + x_6 + x_{10})$
CycB	$x_{10} = \overline{x_7 x_8}$

Attractors

G_0	Cell cycle						
0	1	1	1	1	1	1	1
1	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1
0	0	0	0	1	0	1	1
0	0	1	1	1	0	0	1
1	0	0	0	0	0	0	0
0	1	0	1	0	0	0	0
1	1	0	0	0	1	1	1
0	1	1	1	0	1	0	0
0	0	1	1	0	0	0	0

qualitatively the same behavior for the transient time in components of sizes up to $n=18$ [$\langle \chi \rangle_{n=18}^{\max}(e=3) \approx 28$ and $\langle \chi \rangle_{n=18}(e=100) \approx 3$].

We also studied the average transient time as a function of the average sensitivity (Fig. 5). The average transient time starts to grow rapidly, soon after the ensembles enter the chaotic regime ($s > 1$) and scales with the system size in highly chaotic networks.

Compared to the average transient times in networks of higher sensitivity s, even the maximal transient times in the critical relevant components (Fig. 4) are small. The updating functions of nodes of the critical component that depend on more than one relevant input are more likely to be *canalizing*[4] because of the stronger bias p that was used in their generation process [27]. Canalizing functions are found frequently as control rules governing the transcription in eukaryotic genes [28]. Dynamic properties of random Boolean networks with canalizing functions have also been studied in [1,26]. A higher proportion of canalizing functions leads characteristically to short attractors and more robust dynamics.

V. BOOLEAN NETWORKS MODELING BIOLOGICAL CIRCUITS

As a showcase model for a biological gene regulatory circuit, we now analyze the Boolean network of the mammalian cell cycle as presented in [16]. This $n=10$ node network simulates the states of cell cycle genes that regulate the process of cell division and its quiescent G_0 phase. It can be viewed as a biological example of a complex relevant component of highly connected nodes (for a more thorough biological discussion, see [16]). The Boolean functions and attractor states of this network are shown in the tables below. The value on the left-hand side of the equations corresponds to the value at time $t+1$, $x_i(t+1) = f(x_{i_1}(t),\dots,x_{i_{k_i}}(t))$, with the symbols + and · representing the OR and AND operations, respectively:

The attractor of length seven represents the different steps of the cell cycle phases, G_1, S, G_2, and M, whereas the fixed point attractor represents the G_0 phase. Both attractors have the same weight $w=0.5$, which yields an entropy of $h=\ln 2 \approx 0.69$. If we average over the sensitivities of the single nodes, we obtain a network sensitivity of $s=1.04$ which lies in the range of the expected average sensitivity for a relevant component of a critical network. The average transient time of this network is $\langle \chi \rangle = 4.8$, which is on the order of a random relevant component with the same excess ($e=24$).

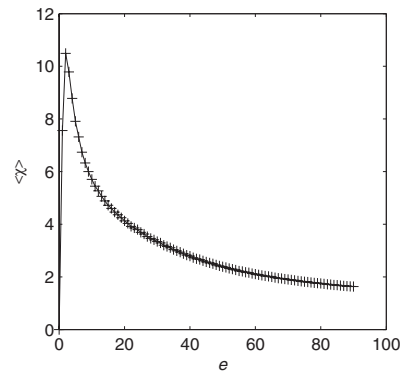So far, other detailed deterministic Boolean models have only been presented for a few other gene regulatory circuits.

FIG. 4. The average transient time decreases after a peak around $e \approx 2$ with increasing excess.
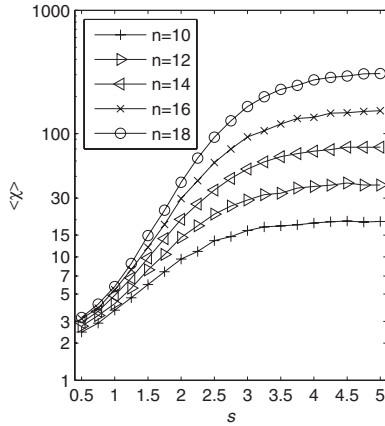
---

[4]A function is canalizing for the value $\sigma_i = \{0,1\}$ of variable i if this value can determine the function value regardless of the values of the other input variables.

FIG. 5. Average transient time $\langle \chi \rangle$ in random network ensembles of increasing sensitivity s.

A prominent example among those is a model for the segment polarity gene network, which governs the embryonic pattern formation during certain stages of the developmental process in the fruit fly *Drosophila melanogaster* [17]. This Boolean network models the interaction between eleven genes and its products and defines certain fixed-point (single-state) attractors that can be identified as stable gene expression patterns in wild-type embryos. For this model, we also find the dynamic network parameters to lie within the range of critical complex components: For the network sensitivity we obtain $s=1.02$, the entropy is $h=2.17$, and the transient time $\langle \chi \rangle = 3.6$.

Generally, the identification of a single deterministic logical function for a gene is often difficult for experimental reasons [18], or determinism might not even be a desired feature of the modeling approach itself. For example, probabilistic Boolean networks consider more than just one Boolean function as possible updating rules for a gene [29]. Also, asynchronous updating schemes lead to nondeterministic dynamics [30]. We therefore conclude with a section in which we sketch an approach that enables us to extend the concept of basin entropy to nondeterministic models.

VI. ENTROPY ESTIMATED FROM TIME-SERIES DATA

In an unperturbed Boolean network a trajectory that started from a random point in the state space will finally be caught in a single attractor. If we allow small random perturbations in the updating procedure, the trajectory will jump out of its attractor from time to time and may settle in another attractor. The deterministic dynamics of the unperturbed network give way to a stochastic (and ergodic) Markov process with transition probabilities $p_{ij}=P(\mathbf{x}_t=j|\mathbf{x}_{t-1}=i)$, such that $\sum_{j=1}^{2^n} p_{ij}=1$ [31]. The transition probabilities can be arranged in an ordered fashion in a stochastic state transition matrix,

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}. \tag{7}$$

Let $\pi(0)$ be the vector of initial state probabilities at time $t=0$. We may calculate the state probability distribution $\pi(m)$ after m steps:

$$\pi(m) = \pi(0)\mathbf{P}^m. \tag{8}$$

We may further sum up the probabilities of states, leading to the same attractor, to get a probability distribution υ_m for the attractors after m time steps. The steady-state probabilities $(m \to \infty)$ for attractors in Boolean networks with perturbations were studied in [32].

Let us consider the following two-node network defined by the Boolean functions $x_1(t+1)=x_1(t)+x_2(t)$ and $x_2(t+1)=x_1(t)\overline{x_2(t)}+\overline{x_1(t)}x_2(t)$ to illustrate the difference between the weight distribution of the deterministic case and the attractor probability distribution in the perturbed case. When we use a perturbation probability of $q=\mathrm{Prob}(x_i \to \overline{x_i})=0.1$ for switching the value of a node *after* calculating the successor of a state, the deterministic transition matrix

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

is replaced by the stochastic transition matrix

$$\mathbf{P} = \begin{pmatrix} 0.81 & 0.09 & 0.09 & 0.01 \\ 0.01 & 0.09 & 0.09 & 0.81 \\ 0.01 & 0.09 & 0.09 & 0.81 \\ 0.09 & 0.01 & 0.81 & 0.09 \end{pmatrix},$$

where the states along the rows and columns are arranged in the usual lexicographic order (00, 01, 10, 11).

This Boolean network divides its state space into two different basins of attraction: The first one consists of its fixed point attractor (00) while the second one contains the transient state (01) that is flowing into the attractor $(10) \rightleftarrows (11)$. Thus the weight distribution is $w_1=0.25$ and $w_2=0.75$. If we solve the steady state equation $\pi = \pi\mathbf{P}$, we obtain $\pi_{(00)}=0.201$, $\pi_{(01)}=0.0598$, $\pi_{(10)}=0.3618$, and $\pi_{(11)}=0.3775$. We may sum up the probabilities of states contributing to the same attractor basin to get a probability distribution ν of the basins, $\upsilon_1=0.201$ and $\upsilon_2=0.799$. If the history of a trajectory is unknown, this distribution describes the probability of the network operating in a certain basin. It can be estimated with arbitrary precision by sampling the states of a time series. The perturbation probability q clearly affects ν; for large q, the network rules (connections, updating functions) lose their importance, and the time series become random.

For our described two-node example network, the basin weight distribution $\mathbf{w}$ and the basin probability distribution ν differ. We therefore studied how increasing network size affects the average deviation,

$$\sigma = \frac{1}{|\rho|}\sqrt{\sum_\rho (w_\rho - \nu_\rho)^2} \tag{9}$$

between the basin probability distribution ν and the weight distribution $\mathbf{w}$ in different network ensembles. In Fig. 6(a) the behavior of σ is shown when increasingly long time series with $q=0.01$ are used to estimate the basin probability distribution of critical $n=10$ network ensembles. Figure 6(b)
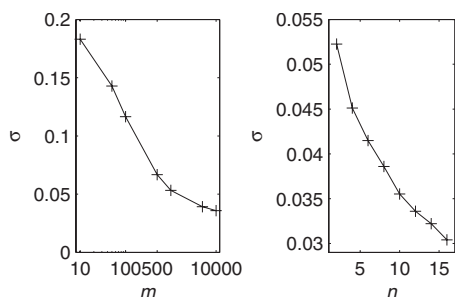
FIG. 6. (a) The deviation σ between the weight distribution **w** and the estimated basin probability distribution ν decreases for longer time series m ($n=10$ and $q=0.01$ fixed). (b) The deviation σ becomes smaller with increasing network size n ($m=10\,000$ and $q=0.01$ fixed).

shows that the mean deviation σ decreases for increasing network sizes ($q=0.01$ and $m=10\,000$). Therefore, especially in larger networks, the basin weight distribution may be well estimated by the basin probability distribution. For all network ensembles the average was taken over more than 10 000 network instances.

The assignment of a state in the time series to its attractor, if the underlying network rules are unknown, is a challenging classification problem. Two states occurring in the time series with comparable frequency may either belong to the same attractor or to two different attractors that just happen to have a similar weight or probability. In order to solve this problem, we have to make a second assumption: States that belong to the same attractor are more likely to occur in temporal proximity. Instead of looking at a single state, we consider all states in a time window of a certain length τ. The size of the time window τ generally has to be estimated from the time series data [33]. In a perturbed trajectory, generated from a Boolean network, τ should be on the order of the expected attractor lengths. In a model based on a biological circuit, the choice for the expected length may also be motivated by "biological intuition."

The classification of these time profiles into several attractors is a clustering problem, where the number of clusters is not known. Many algorithms to estimate the optimal number of clusters in a data set have been developed and extensively studied. Generally, more free parameters (clusters) enable one to further minimize the error function on which the cluster algorithm is based. This better "fit" is then penalized by a term growing with the number of free parameters. Based on this tradeoff criterion, an "optimal" clustering model can be found. When dealing with biological data, a range for the number of expected clusters (e.g., different cell states) can also be provided by the experimentalist. It is not our intention here to discuss the challenges of clustering and we refer the interested reader to the extensive literature in this field [34–36]. However, for illustrational purposes, we sketch the analysis of the already introduced network of the mammalian cell cycle by a perturbed time series and clustering.

We generated a time series of $m=10\,000$ successor states from the Boolean model of the mammalian cell cycle. The value of every node was flipped with probability $q=0.01$ after calculating the successor state. Profiles were generated

by adding up the values of a node during the time window τ:

$$c_i(t) = \sum_{t'=t-\tau}^{t} x_i(t'), \tag{10}$$

with $c_i \in \{0, \ldots, \tau+1\}$. Different values of τ have been tested: $\tau=4,\ldots,10$. The distance between two profiles **c** $=(c_1,\ldots,c_n)$ and $\mathbf{c}'=(c_1',\ldots,c_n')$ was measured by the city block (L_1) distance:

$$d(\mathbf{c},\mathbf{c}') = \sum_{i=1}^{n} |c_i - c_i'|. \tag{11}$$

The profiles were then clustered by the k-means algorithm [35]. In order to determine the optimal number of clusters, the Bayesian information criterion (BIC) and Akaike information criterion (AIC) have been used and yielded an optimal number of two clusters for all used τ. This correctly corresponds to the two attractors of the underlying Boolean network: The fixed-point attractor of the G_0 phase and the attractor of length seven of the cell cycle. The classification of the time series into two attractors yields a probability distribution ν, whose entropy $h=0.691$ is close to the "true" network's entropy, $h \approx \ln 2$. This example demonstrates how the attractors and their weight distribution, a dynamical property of the network, can be derived from a time series using a straightforward clustering approach that does not require knowledge about the underlying network rules.

VII. DISCUSSION

We studied the average entropy and transient time of random complex relevant components near criticality as a function of their excess. This was motivated in part by new analytical results on the degree distribution of relevant nodes in critical network ensembles [22]. In random graphs of such given degree distributions, most nodes are organized with high probability as a single giant component [37,38]. The regulatory elements in gene networks, on the other hand, seem to be organized in a more modular manner [12,13]. This raises the question of whether (ordinary) critical random Boolean networks that have been successfully used as models for the study of gene regulatory dynamics still lack characteristic properties of their biological archetypes.

When we consider, for example, the excess of a network as a fixed constraint, a modular organization of the interacting nodes will yield a higher basin entropy and a shorter average transient time.[5] One might ask if a maximization of the basin entropy or a minimization of the transient time are desirable features in biological networks. A short transient time might, for instance, speed up the process of settling down to a certain cellular state (corresponding to an attrac-

---

[5]A critical relevant component of $n=10$ nodes and excess $e=10$ has an average entropy of $\langle h \rangle \approx 0.6$. If two such components with a combined entropy of $\langle h \rangle \approx 1.2$ merge into an $n=20$ node component of excess $e=20$ and are randomly rewired, the entropy decreases to an average value of $\langle h \rangle \approx 0.7$. The average transient time, on the other hand, increases from $\chi \approx 5.5$ to $\chi \approx 12$.

tor) whereas a high entropy would minimize the number of genes required to perform a classification or decision task. Thus the ability to rapidly respond to environmental cues by switching to a particular functional cellular state as well as the economy with which cellular information processing and decision-making can be carried out may be evolutionarily selected features of biological networks.

With the declining costs of microarray and other high-throughput technologies, time series data will be increasingly available from biological experiments, enabling network parameters such as basin entropy and transient time to be studied in a biological context. Our approach to estimate entropy from time series data sketches one possible way how that might be accomplished.

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

[2] S. Kauffman, J. Theor. Biol. **22**, 437 (1969).

[3] S. Kauffman, Physica A **340**, 733 (2004).

[4] B. Derrida and Y. Pomeau, Europhys. Lett. **1**, 45 (1986).

[5] B. Derrida and D. Stauffer, Europhys. Lett. **2**, 739 (1986).

[6] I. Shmulevich, S. Kauffman, and M. Aldana, Proc. Natl. Acad. Sci. U.S.A. **102**, 13439 (2005).

[7] M. Aldana, S. Coppersmith, and L. P. Kadanoff, in *Perspectives and Problems in Nonlinear Science*, edited by E. Kaplan, J. E. Marsden, and K. R. Sreenivasan (Springer, New York, 2002), pp. 23–89.

[8] H. Flyvbjerg, J. Phys. A **21**, L955 (1988).

[9] U. Bastolla and G. Parisi, Physica D **115**, 203 (1998).

[10] U. Bastolla and G. Parisi, Physica D **115**, 219 (1998).

[11] P. Krawitz and I. Shmulevich, Phys. Rev. Lett. **98**, 158701 (2007).

[12] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Nat. Genet. **31**, 64 (2002).

[13] T. I. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph *et al.* Science **298**, 799 (2002).

[14] V. Kaufman and B. Drossel, New J. Phys. **8**, 228 (2006).

[15] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang, Proc. Natl. Acad. Sci. U.S.A. **101**, 4781 (2004).

[16] A. Faure, A. Naldi, C. Chaouiya, and D. Thieffry, Bioinformatics **22**, 214 (2006).

[17] R. Albert and H. Othmer, J. Theor. Biol. **223**, 1 (2003).

[18] C. Espinosa-Soto, P. Padilla-Longoria, and E. Alvarez-Buylla, Plant Cell **16**, 2923 (2004).

[19] I. Shmulevich and S. A. Kauffman, Phys. Rev. Lett. **93**, 048701 (2004).

[20] J. E. S. Socolar and S. A. Kauffman, Phys. Rev. Lett. **90**, 068702 (2003).

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

[22] T. Mihaljev and B. Drossel, Phys. Rev. E **74**, 046101 (2006).

[23] T. Luczak, B. Pittel, and J. Wierman, Trans. Am. Math. Soc. **341**, 721 (1994).

[24] N. de Bruijn, Proc. K. Ned. Akad. Wet. **49**, 578 (1946).

[25] V. Kaufman and B. Drossel, Eur. Phys. J. B **43**, 115 (2005).

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

[27] W. Just, I. Shmulevich, and J. Konvalina, Physica D **197**, 211 (2004).

[28] S. Harris, B. Sawhill, A. Wuensche, and S. Kauffman, Complexity **7**, 23 (2002).

[29] I. Shmulevich, E. Dougherty, S. Kim, and W. Zhang, Bioinformatics **18**, 261 (2002).

[30] F. Greil and B. Drossel, Phys. Rev. Lett. **95**, 048701 (2005).

[31] I. Shmulevich, E. R. Dougherty, and W. Zhang, Bioinformatics **18**, 1319 (2002).

[32] M. Brun, E. R. Dougherty, and I. Shmulevich, Signal Process. **85**, 1993 (2005).

[33] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2004).

[34] G. Schwartz, Ann. Stat. **6**, 461 (1978).

[35] R. Duda and P. Hart, *Pattern Classification and Scene Analysis* (John Wiley & Sons, New York, 1973).

[36] A. K. Jain, M. N. Murty, and P. J. Flynn, ACM Comput. Surv. **31**, 264 (1999).

[37] C. Cooper and A. Frieze, Combinatorics, Probab. Comput. **13**, 319 (2004).

[38] M. Molloy and B. Reed, Combinatorics, Probab. Comput. **7**, 295 (1998).