# Quantifying local structure effects in network dynamics

Andre S. Ribeiro,* Jason Lloyd-Price, Juha Kesseli, Antti Häkkinen, and Olli Yli-Harja
*Computational Systems Biology Research Group, Tampere University of Technology, Finland*
(Received 14 March 2008; revised manuscript received 4 September 2008; published 24 November 2008)

Mutual information between the time series of two dynamical elements measures how well their activities are coordinated. In a network of interacting elements, the average mutual information over all pairs of elements $I$ is a global measure of the correlation between the elements' dynamics. Local topological features in the network have been shown to affect $I$. Here we define a generalized clustering coefficient $C_p$ and show that this quantity captures the effects of local structures on the global dynamics of networks. Using random Boolean networks (RBNs) as models of networks of interacting elements, we show that the variation of $\langle I\rangle$ ($I$ averaged over an ensemble of RBNs with the number of nodes $N$ and average connectivity $k$) with $N$ and $k$ is caused by the variation of $\langle C_p\rangle$. Also, the variability of $I$ between RBNs with equal $N$ and $k$ is due to their distinct values of $C_p$. Consequently, we propose a rewiring method to generate ensembles of BNs, from ordinary RBNs, with fixed values of $C_p$ up to order 5, while maintaining in- and out-degree distributions. Using this methodology, the dependency of $\langle C_p\rangle$ on $N$ and $k$ and the variability of $I$ for RBNs with equal $N$ and $k$ are shown to disappear in RBNs with $C_p$ set to zero. The $\langle I\rangle$ of ensembles of RBNs with fixed, nonzero $C_p$ values, also becomes almost independent of $N$ and $k$. In addition, it is shown that $\langle C_p\rangle$ exhibits a power-law dependence on $N$ in ordinary RBNs, suggesting that the $C_p$ affects even relatively large networks. The method of generating networks with fixed $C_p$ values is useful to generate networks with small $N$ whose dynamics have the same properties as those of large scale networks, or to generate ensembles of networks with the same $C_p$ as some specific network, and thus comparable dynamics. These results show how a system's dynamics is constrained by its local structure, suggesting that the local topology of biological networks might be shaped by selection, for example, towards optimizing the coordination between its components.

DOI: 10.1103/PhysRevE.78.056108 PACS number(s): 89.75.Fb, 05.45.Tp, 89.70.Cf, 87.16.Yc

## I. INTRODUCTION

The dynamical behavior of large, complex networks of interacting elements is generally quite difficult to understand in detail. The presence of multiple influences on each element can lead to complex patterns of activity even in deterministic models [1]. One example of these systems is the network of interactions between genes in cells, the gene regulatory network (GRN), where the interactions correspond to transcriptional and post-transcriptional regulatory mechanisms. The expression of a gene may be regulated by itself and up to 20 proteins expressed by other genes. The GRN of such interactions has a complicated structure, including positive and negative feedback loops and nontrivial combinatorial logic.

One common approach to investigate the dynamics of networks is the ensemble approach [2], which consists of exploring the behavior of networks with given parameters (such as $N$ and $k$) by studying the properties of many networks generated with the same general parameter values, but each with a unique graph of connections and update rules. One difficulty in applying this method is determining the features of the network that should be accounted for when generating the ensemble.

Predictions of dynamical properties using mean-field approximations only apply to large and sparse networks (locally treelike structures). In Ref. [1], using the average pairwise mutual information ($\langle I\rangle$) as a measure of the overall level of coordination between the elements of models of complex regulatory networks, it was observed that finite size effects play a non-negligible role on the $\langle I\rangle$ of random Boolean networks (RBNs) with a small number of nodes ($N<250$), even with low average connectivity. Also, in Ref. [3], the predictions of avalanche size distributions in random Boolean networks with low connectivity are only valid for large networks ($N>250$) due to non-negligible local structure effects in smaller networks.

Importantly, many studies require numerical simulations which can only be done for small networks (e.g., Ref. [4]) and, in general, real large scale networks are not locally treelike [5–7]. Thus, it is of importance to identify topological features that cause distinct dynamical behavior between treelike and non-tree-like structures.

Some previous studies have focused on the role on network dynamics of specific motifs [8]. These are defined as patterns of interconnections present in GRNs with a higher abundance than in networks with equal $k$ and $N$ and random topology. However, general measures of clustering that could explain global dynamical behaviors have not been proposed, e.g., a measure from whose value one can predict the expected $\langle I\rangle$ between the patterns of activity of all nodes.

In studies of the topology of real networks (initially social networks [9]), it has been identified that a network's clustering coefficient ($C$) [10] (defined as the proportion of links between the nodes connected to a node, averaged over all nodes of the network) should play a relevant role in the dynamics. Subsequent studies have focused mostly on topological effects of $C$, such as the relation between $C$ and average path length [11].

*andre.sanchesribeiro@tut.fi

 

The concept of clustering coefficient has been extended in several works (see, e.g., Refs. [12,13], and references within). We also introduce a generalized measure of clustering to characterize the local topology of networks. However, while the previous works have focused on topological consequences of local structures, e.g., correlating clustering with degree distribution [12], here we study a model that includes dynamics explicitly and focus on the effects of local topological features on the dynamics. Additionally, our extension of the concept of triangular clusters [10] differs from previous approaches in that we account for self connections, bidirectional connections, triangles, squares and higher order clusters. Finally, we note that in Ref. [12], and references within, connections are usually undirected while we assume directed connections.

Some effects of structure on the dynamics have been studied in lattice structures [14]. There, the focus was on the differences in the dynamics of various lattice structures, with the same average connectivity, modeled as cellular automatons with nearest-neighbor interactions and two admissible local states. Four geometries were considered: a chain, a hexagonal lattice, a square lattice, and a cubic lattice. It was shown that the topology affects the distribution of local states, by comparing the dynamics on these lattices. We note that the set of update rules allowed in Ref. [14] was also far more restricted than the one used here.

In Ref. [15], basin entropy was introduced to characterize the dynamics of Boolean networks. Subsequently, in Ref. [16], the effects of increasing connectivity on basin entropy were studied. Basin entropy measures the diversity of the state space occupied by the attractors and their basins. This measure, similar to $\langle I \rangle$, is also maximized for critical networks [1]. Further similarities between these two quantities are currently unknown.

Dependence of the basin entropy [16] or of the mutual information [1] on the degree of clustering of the network has not previously been analyzed directly. Importantly, we show here that networks with equal connectivity have very different dynamical properties if they differ in their local structure. That is, the changes caused in the dynamics of a particular network, by adding connections, cannot be predicted from the number of extra connections alone.

In Ref. [1], $\langle I \rangle$ is studied to understand how global topological features, such as connectivity and network size, affect the flux of information between nodes of dynamical networks. Whereas in that work the networks studied have no local topological structure, and are defined solely in terms of number of connections and number of nodes, in this work we analyze and quantify the effects of non-tree-like local structures on $\langle I \rangle$. Importantly, real gene networks, shaped by evolution, are known to have such local structure [7], and cannot therefore be approximated by treelike structures. This means also that local structures ought to play a very relevant role in the dynamics.

A preliminary observation in Ref. [1] noted that the clustering coefficient (as originally defined in Ref. [10]) could cause distinct behaviors between large and small RBNs with equal connectivity. The local topological structure of finite networks was not pursued further in that work.

Here, we investigate this further. It is shown that it is not sufficient to account for the original $C$ to characterize the local structure of networks. This measure needs to be extended to accommodate other local structures. We introduce here a measure to characterize the local topological structure of finite networks, the "generalized clustering coefficient," $C_p$. This measure allows quantifying the dependence of $I$ on the degree of global clustering. We show that $C_p$ accounts for all local effects by simulating RBNs where the $C_p$ is set to zero, and showing that the mutual information of the network ensemble becomes independent of network size.

In the next section, the Boolean network model is presented, followed by the definitions of temporal pairwise mutual information and generalized clustering coefficient. Next, we present our results, followed by conclusions.

## II. METHODS

### A. Boolean networks

A Boolean network (BN) is a directed graph with $N$ nodes. Nodes represent elements of the system and graph arcs represent interactions between them. Each node is assigned a binary output value and a Boolean function, whose inputs are defined by the graph connections. The network's state is the vector of nodes' values. In a synchronous BN, all nodes are updated simultaneously. By running the network over various time steps starting from an initial state, a trajectory through the network's state space can be observed (referred to as a "time series").

RBNs were introduced as the first model of GRN [17]. Each node is a gene, and is assigned a Boolean function from the set of possible Boolean functions of $k$ variables. In the RBN model used here, time is discrete and all nodes update their activities synchronously. Thus, a state of the network passes to a unique successor state at each moment. Over time, the system follows a trajectory that ends on a state cycle attractor. In general, a RBN has many such attractors.

One way of characterizing the dynamics of a RBN is to measure the mutual information between the time series of all pairs of nodes. This quantity is now introduced.

### B. Pairwise mutual information

The mutual information contained in the time series of two elements gives a measure of how well their activities are coordinated in time. In a large, complex network of interacting elements, $I$ is a global measure of how well the system can coordinate its internal dynamics.

$I$ is defined as follows. Let $s_a$ be a process that generates a 0 with probability $p_0$ and a 1 with probability $p_1$. We define the entropy of $s_a$ as

$$H[s_a] \equiv -p_0 \log_2 p_0 - p_1 \log_2 p_1. \tag{1}$$

For a process $s_{ab}$ that generates pairs $xy$ with probabilities $p_{xy}$, where $x,y \in \{0,1\}$, the joint entropy is defined as

$$H[s_{ab}] \equiv -p_{00} \log_2 p_{00} - p_{01} \log_2 p_{01} - p_{10} \log_2 p_{10} - p_{11} \log_2 p_{11}. \tag{2}$$

In case any of the probabilities in the entropy formulas happens to be zero, its contribution to the entropies should be

zero as well. Thus, in this context we let $0 \log_2 0 = 0$.

Ideally, for a particular RBN, we would run the dynamics starting from all possible initial states and observe the time series for infinitely many time steps. However, the state space of even modestly sized RBNs is prohibitively large for this approach. Instead, the network is started from a random initial state, and is run for a certain number of time steps. The fraction of steps for which the value of node $i$ is $x$ gives $p_x$ for the process $s_i$. The value of $p_{xy}$ for the process $s_{ij}$ is given by the fraction of time steps for which node $i$ has the value $x$ and on the next time step node $j$ has the value $y$. The mutual information between the time series of the pair $ij$ is then defined as [1]

$$I_{ij} = H[s_i] + H[s_j] - H[s_{ij}]. \tag{3}$$

With this definition, $I_{ij}$ measures the extent to which information about node $i$ at time $t$ influences node $j$ one time step later, at $t+1$. Note that the propagation may be indirect; a nonzero $I_{ij}$ can result when $i$ is not an input to $j$ but both are influenced by a common node through previous time steps.

To quantify the efficiency of information propagation through a single RBN, we use the average pairwise mutual information, defined as

$$I = N^{-2} \sum_{i,j} I_{ij}. \tag{4}$$

To characterize the efficiency of information propagation in an ensemble of networks, we use the average pairwise mutual information of the ensemble [1] $\langle I \rangle$, where $\langle \cdots \rangle$ indicates the mean value over the members of the ensemble.

Mutual information of the time series of a network is used here to characterize the network's global dynamics. This is to be compared to the next parameter, the generalized clustering coefficient, which characterizes the structure that gives rise to the dynamics.

When measuring the mutual information of a RBN from $T = 10^3$ time steps, we do not discard transients. The mean-field approach used in Ref. [1] to estimate the mutual information requires either increasing the lengths of discarded transients or increasing the numbers of initial conditions per network. Here, we used a similar approach that consists in averaging over many networks with a random initial state, thus not requiring removal of initial transients to measure $\langle I \rangle$, since from Ref. [1] it is know that there is satisfactory convergence between meanfield and numerical results both for increasing lengths of discarded transients and for increasing numbers of initial conditions per network (or, equivalently, for increasing the number of networks generated to measure $\langle I \rangle$ for some given $N$ and $K$).

### C. Generalized clustering coefficient

Originally, the clustering coefficient ($C$) was introduced to measure the fraction of effective connections between the first nearest neighbors of a node in an undirected graph, out of the total number of possible connections [10]. Let $E_i$ be the number of connections between the $k_i$ nodes connected to a node $i$. The network average $C$ is [10]

$$C = \frac{1}{N} \sum_{i=1}^{N} \frac{2E_i}{k_i(k_i - 1)}. \tag{5}$$

Notably, the $\langle I \rangle$ of small RBNs (i.e., approximately $N < 250$) with an imposed $C = 0$ (no triangles) is still highly dependent on $N$, though not as much as when $C$ is not equal to zero. This can be concluded from results shown later on (Fig. 7).

The results from the numerical simulations shown in the results section (Fig. 7) allow to conclude that $\langle I \rangle$ only becomes independent of the network size $N$ if other higher order topological structures, besides triangles, are not present in the topology of the network. For that reason, we extended the concept of clustering coefficient to include higher order structures (such as squares).

An annealed approximation was able to predict the $\langle I \rangle$ of ensembles of RBN, given $k$ [1], if these networks are large enough. This is because of the network's relatively low connectivity ($k < 5$) resulting in locally treelike topologies.

The reason this approximation failed to capture the dynamics of small RBNs (or highly connected ones) [1] is because these are usually not locally treelike structures when randomly generated, even for small $k$ values. The existence of self-inputs, bidirectional connections, triangles, and other small structures (e.g., squares), destroy the treelikeness of a structure, and hence affect $I$.

Therefore, to generate an ensemble of RBNs whose general dynamical behavior (here characterized by $\langle I \rangle$) does not depend on size of the networks, one must impose an equal value of some generalized version of the clustering coefficient to all the RBNs. Fixing the value of such a generalized clustering coefficient can also be used to study the dynamical behavior of a specific network using an ensemble approach. Namely, the RBNs of the ensemble ought to have the same clustering coefficient as the specific network under study.

Here, we generalize the original clustering coefficient [9] to account for all non-tree-like local structures. The generalized clustering coefficient $C_p$ is, similar to the original $C$, computed for each node and averaged over all the nodes. For simplicity, directionality is not accounted for when computing a network's $C_p$. This did not appear to affect the results on the ensembles of networks studied here. That is, the value of $C_p$, computed assuming all network connections are undirected was sufficient to explain the networks dynamics properties.

In the definition of $C_p$, we denote the coefficient for different orders $p$ as follows: $C_p$ is the clustering coefficient of a network of order $p$ (in the present work, only up to order 5 is considered). $C_{(i,j)}$ is the clustering coefficient of a network from order $i$ up to order $j$ (inclusively).

Let $\alpha$ be the node index. Let $\kappa_{i,j}^{\alpha}$ be the amount of connections between the nodes at distance $i$ and the nodes at distance $j$ from node $\alpha$, and $T_{i,j}^{\alpha}$ be the possible maximum amount of such connections. We define $C_p$ as follows (for order $p > 2$):

$$C_p = \frac{1}{N} \sum_{\alpha=1}^{N} \left( \frac{\sum_{r=1}^{p-2} \kappa_{r,p-r-1}^{\alpha}}{\sum_{r=1}^{p-2} T_{r,p-r-1}^{\alpha}} \right). \tag{6}$$

For $p<3$, we define $C_1$ to be the fraction of nodes with self inputs and $C_2$ to be the mean ratio of connections that are bidirectional connected to each node. That is, if a node has three connections and one of them is bidirectional, then its contribution to $C_2$ is $\frac{1}{3}$. Note that the definition of $C_3$ matches the definition of the original $C$ [10].

The generalized clustering coefficient of a network, from order $i$ up to order $j$, is thus defined as

$$C_{(i,j)} = \sum_{p=i}^{j} C_p. \tag{7}$$

In the numerical simulations presented in the results section, we consider local structures only up to order 5, due to computational limitations in determining higher orders of $C_p$ and because this was sufficient to account for all local structure effects in networks of 25 nodes or more.

We note that the value of $C_p$ for each $p$ value is computed independently, and in fact, the set of $C_p$ values for each $p$ is what characterizes the network topology. For sake of simplicity, so that network topologies can be characterized by a single value, rather than a sequence of values, we opt to sum the values of each $C_p$ into a single quantity, the "network $C_p$." Other ways of combining each of the $C_p$ values into a single quantity could also be considered.

We found (Sec. III) that this quantity is sufficient to characterize, from a global point of view, the effects of the local structures of networks in their dynamics, at the level of detail observed here. However, a more exact procedure to compare the local structure features of two networks is to compare their values of $C_p$, for each order $p$, independently.

Finally, we note that, when computing the $C_p$, our algorithm only recognizes substructures in which each node is unique. This is done so that, for example, self-inputs (order 1) are not accounted for again when counting bidirectional connections (order 2), and so forth. The same principle is followed when imposing a given $C_p$ value to a network.

To test if the $\langle I \rangle$ of RBNs which are generated such that $C_p$ is equal to zero or another imposed value is independent of the network size $N$, it is necessary to generate ensembles of networks with an imposed $C_p$ value. This requires the generation of RBNs with a random topology which is then rewired to impose the target $C_p$.

The rewiring is a procedure that must not change other properties of the topology and logic of the network. Namely, the in- and out-degree distributions must remain the same.

### D. In- and out-degree distributions and the rewiring of random Boolean networks

To compare networks with a different generalized clustering coefficient, we need to generate networks in which other affecting factors are unchanged. Most importantly, we want to compare networks in which the in- and out-degree distributions of the nodes are the same but which differ only in their local clustering structure. To do this, we use an iterative algorithm as follows.

First, a network is generated with the desired in- and out-degree distributions with no local clustering structure. This is done by selecting an in-degree and out-degree for each node to approximate the desired distributions and then drawing the connections randomly so that the constraints created by the degrees are fulfilled. The resulting network is likely to have a low generalized clustering coefficient.

Starting from a network generated as described, we then use an iterative optimization method to increase or decrease the clustering coefficient of the network without changing the degree distributions. At each iteration, we determine if the $C_p$ of the network is higher or lower than the desired value. If it is higher, then we search the network for the loop structures corresponding to the order of the $C_p$ that we are trying to reduce, and one of the connections that forms the loop is removed in such a way that the number of nodes with a given in-or out-degree does not change. This means that if we remove a connection that goes out from a node with out-degree $m$ and connects to a node with in-degree $n$, we will use the connection to link some node with out-degree $m-1$ and a node with in-degree $n-1$. To ensure that the loops are searched for in an unbiased manner, the indices of the nodes are shuffled at each iteration.

Likewise, if the $C_p$ is lower than the desired value, loops must be introduced. The network is searched for a chain of length $p+2$, where the direction of the first connection corresponds to the direction of the last. The first and last connections are then moved so that the first node becomes an input of the last node in the chain, and the second-last becomes an input of the second node. This forms a loop of length $p$ and maintains the in- and out-degree distributions of the nodes. Again, the indices of the nodes are shuffled at each iteration to ensure that the loops are generated uniformly throughout the network.

Both the destruction and construction of loops involves moving connections in the network. This can have unintended side effects such as an increase or decrease in the $C_p$ of other orders than we were intending to modify. For this reason, the algorithm must be rerun for other orders until the target $C_p$ values have been reached.

In some cases, the search for a possible rewiring can be extremely long. For practical purposes, we limit that search. Namely, after a specified number of attempts, if the algorithm is unable to find a rewiring scheme which imposes the desired $C_p$ value, it simply discards this network, and starts all over with a newly created network.

Since the cost of finding loops grows exponentially with $p$, this becomes an expensive operation to perform at higher orders. Here, we limit ourselves to order 5, but show that this is enough to account for almost all local structure effects in the dynamics.

## III. RESULTS

The results consist of measurements of $I$ and $C_p$ from ensembles of RBNs, and show that the variance in $C_p$ accounts for the dependency of $I$ on $N$ and $k$ in RBNs of random topology. First, we quantify the $\langle I \rangle$ and standard deviation $[\sigma(I)]$ over an ensemble of RBNs, each independently simulated, as $N$ and $k$ vary. A network's $I$ is computed from a time series starting at a random state, and corresponds

to the average $I_{i,j}$ between all pairs of nodes. The values of the standard deviation show how variable the $I$ of networks with the same $k$ and $N$ can be, due to the differences in their $C_p$ values.

Next, we show the values of the average and standard deviation of $C_p$, up to order 5, of that ensemble of networks which confirm that these quantities behave similarly to the average and standard deviation of $I$ as $N$ varies, respectively. Also is shown that the differences in $C_p$ between RBNs, with equal $k$ and $N$ values, cause the differences in behavior between them, i.e., in $\langle I \rangle$.

After that, we show that as $C_p$ is set to zero for increasing values of $p$, $\langle I \rangle$ becomes less and less dependent on the size $N$. For $p>3$, no dependency is observed. Importantly, the same is true for the standard deviation of $I$, which indicates that RBNs built according to this new procedure have a far less unpredictable dynamics.

Finally, we show that the $\langle I \rangle$ of an ensemble of RBNs, generated imposing a fixed value of $C_p$ up to order 3, does not depend on network size either. This shows that the ensemble approach can be used to explore the dynamical properties of real networks with a nonzero $C_p$ value, as long as the same value of that quantity is imposed to the networks of the ensemble.

In all following results, each data point corresponds to the average, or standard deviation, of either $I$ or the generalized $C_p$, computed, respectively, from a 1000 time step time series, and topology of 100 independent simulations. Each simulation consists of generating a RBN, with a given $k$ and $N$. In some cases, the network $C_p$ is fixed as well (when explicitly stated). Then, from a random initial state, we extract a time series of 1000 consecutive time steps, from which $I$ is calculated.

We chose not to discount any initial transients in the time series. Although, in 1000 time steps, most of the time the RBN will be in an attractor (representing therefore "long term behavior"), we aimed to observe the effects of $\langle C_p \rangle$ also in the initial transients, thus it was opted to include this transient in the 1000 time steps.

As mentioned, the rewiring of RBNs to attain topologies with specific values of $C_p$ is always done maintaining in- and out-degree distributions unchanged since these two parameters are known to affect the global dynamics [18]. In all cases, the Boolean networks are generated such that their topology is random, i.e., connections are placed following a Poisson distribution of in-degrees with a random selection of inputs resulting in a Poissonian out-degree distribution as well. Strictly speaking, the degree distribution is a binomial distribution, which for large networks, sparsely connected, is a good approximation of the Poisson distribution.

The update rules are random Boolean functions with a $p_{\text{bias}}$ of 0.5. This means that for each combination of input values, the output value is selected randomly (independently with respect to other input combinations) with probability $p_{\text{bias}}$ of obtaining a 1. In the case a given $C_p$ value is imposed on the topology, this is done according to the method described in the previous section.

### A. Mutual information of finite RBNs

A detailed analysis of the value of $I$ as a function of $k$ for RBNs with large $N$ was done in Ref. [1]. These networks were locally treelike structures. Here, to analyze the effects of local structures on $I$, we begin by measuring the variation of mutual information with network size and connectivity. These two parameters are later on shown to affect the generalized $C_p$ of networks.

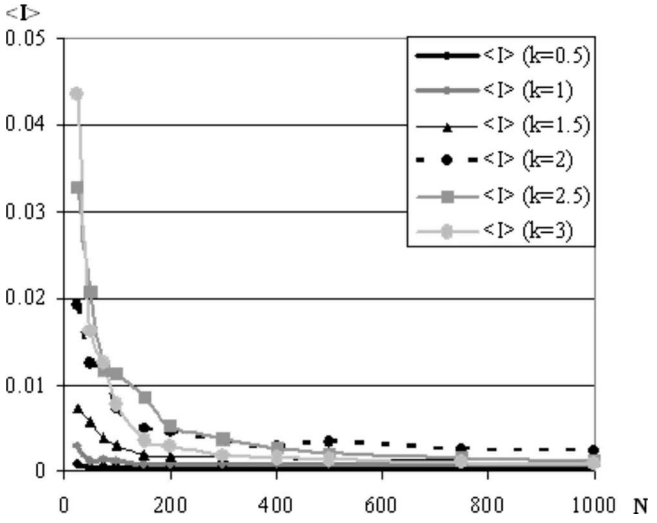


FIG. 1. $\langle I \rangle$ of RBNs computed from time series of length $T=10^3$. Each data point is the average result of 100 independent RBNs, varying $N$ and $k$. Effects of local structures are most prominent for small $N$ and large $k$.

In Fig. 1, the $\langle I \rangle$ of the time series of RBNs is plotted as a function of $k$ and $N$. The most important conclusion from Fig. 1 is that there is some difference in the local structure of the topologies of the networks randomly generated as $N$ is increased, which is the cause for the differences in dynamical behavior (here expressed in the average correlation between all nodes in the time series).

The results show that $\langle I \rangle$ is highly dependent on $k$ and $N$. As $N$ increases, as expected, local structural effects play a less relevant role since maintaining $k$ constant and increasing $N$ creates structures which are locally more treelike, resulting in a significant decrease of $\langle I \rangle$ as $N$ increases. For $N>250$ and low connectivity values ($k<2$), the $\langle I \rangle$ becomes almost invariant for further increases of $N$. Thus, a mean field approximation can be used to predict the $\langle I \rangle$ values for these parameter ranges [1].

All curves in Fig. 1 follow the same trend as $N$ increases, except for $k=2$ which corresponds to the critical regime of RBNs ($k=2$ and $p_{\text{bias}}=0.5$ [18]). It has been proven analytically and numerically confirmed that $\langle I \rangle$ is maximized for critical networks [1] if $N>250$. For $N<250$, RBNs with $k=2$ do not have the highest $\langle I \rangle$, exemplifying the relevance of local structure effects in network dynamics.

A key insight in Ref. [1] is that in infinite size networks, the maximization of average pairwise mutual information is achieved in the critical regime (visible in Fig. 1) due to the formation of long chains of effectively single-input nodes which for other topological constraints do not emerge. This topological feature, unique for critical networks, explains why $\langle I \rangle$ is not zero for large $N$.

As $k$ increases, the effects of local structures increase, i.e., $\langle I \rangle$ increases since $C_p$ increases. Importantly, for $k=3$, e.g.,

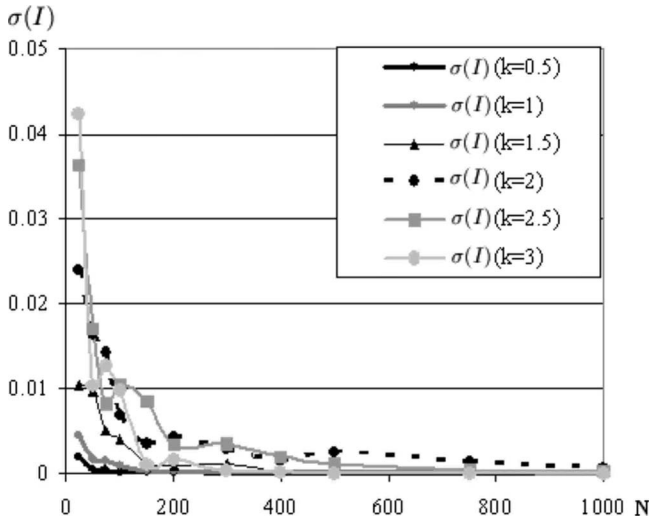


FIG. 2. Standard deviation of $I$ of RBNs from time series of length $T=10^3$. Each data point is the average result of 100 independent RBNs, varying $N$ and $k$. The diversity of dynamical behavior is higher for small $N$ and large $k$.

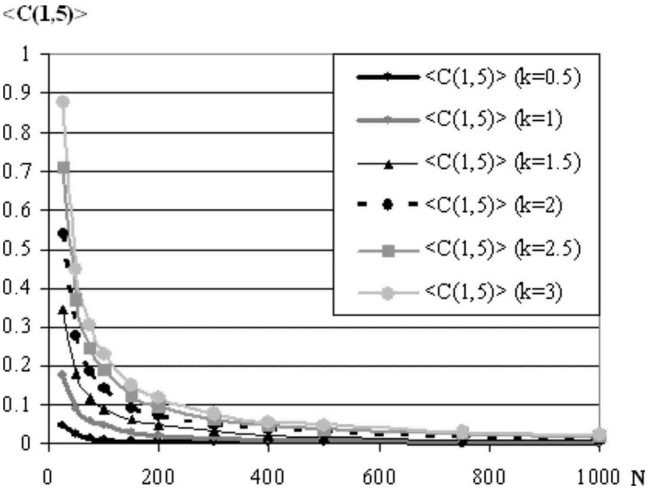


FIG. 3. $C_p$ up to order 5, averaged over 100 networks, for networks with random topology (RBN), varying $k$ and $N$. For lower values of $N$ and higher $k$, more nodes become part of highly interconnected local structures.

only for $N>500$ do local effects become negligible. This is another indication that the main cause for the difference between predictions using the annealed approximation and numerical simulations is the existence of local structures.

We note that the results from the simulations in Fig. 1 are consistent, in the sense that lowering $k$ always decreases $\langle I\rangle$, and increasing $N$ always results in lower $\langle I\rangle$, indicating that numerically sampling 100 networks per data point is sufficient to obtain the average behavior. The only two exceptions to these trends are the RBNs with $N>500$ and $k=2$, and the RBNs with $100<N<300$ and $k=2.5$. Both cases are explained by the fact that these networks are in the critical regime [1]. The maximization, when $100<N<300$, of $\langle I\rangle$ at $k=2.5$ is due to the values of $C_p$ of RBNs. As shown in subsequent results (Fig. 7), $\langle I\rangle$ of RBNs with $100<N<300$ is maximized at $k=2$ if the $C_p$ of these networks is fixed to zero.

In Fig. 2, $\sigma(I)$ is plotted, computed from the same set of RBNs and respective time series used to obtain the $\langle I\rangle$ values in Fig. 1. The standard deviation of $I$ of a set of RBNs $\sigma(I)$ is shown here since it captures how networks with the same values of $N$ and $k$ can differ in their dynamics, if the construction of the topology is not restricted in any other way but fixing $N$ and $k$. The results in Fig. 2 confirm that as $N$ decreases and/or $k$ increases, $\sigma(I)$ increases and that the $\sigma(I)$ is very large, i.e., of the same order of magnitude as $I$, for small $N$, which indicates the high variability of $I$ in RBNs when sampling networks with equal $N$ and $k$. We now investigate if the variation of $C_p$ explains the variability of $I$.

### B. Generalized clustering coefficient of RBNs

Figure 3 shows, up to order 5, the $\langle C_p\rangle$ of the same ensembles of RBNs used to obtain the results in Figs. 1 and 2. Similarly, the standard deviation of the $C_p$ values up to order 5 of these ensembles is plotted in Fig. 4. Given Figs. 1–4, a correlation is clear. Comparing Figs. 1 and 3 one sees that $I$ and $C_p$ vary in a very similar way with $N$ and $k$. Both decrease as $N$ increases. After, approximately $N\simeq 400$, the two quantities decrease very slowly as they converge to the values estimated in Ref. [1], dependent only on the average connectivity. A similar correlation exists between the standard deviation of these two quantities (Figs. 2 and 4).

To confirm this close correlation between $I$ and $C_p$ we computed the Pearson product-moment correlation coefficient values between the mean values of these two quantities for each $K$, as $N$ varies. An identical calculation was done between the standard deviations of these two quantities. The results are shown in Table I, confirming a strong correlation between $\langle I\rangle$ and $\langle C_p\rangle$ and between $\sigma(I)$ and $\sigma(C_p)$ [19].

We now present the results of Figs. 3 and 4 on a log-log scale to observe the behavior of $C_p$ and its standard deviation for large $N$. Within the range of two decades observed, both

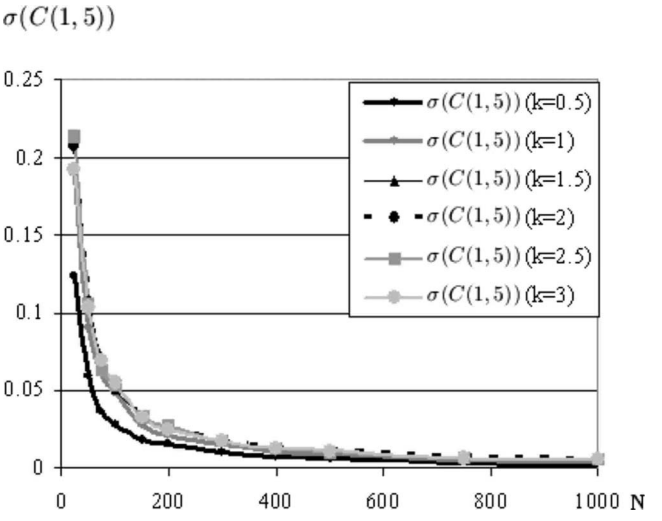


FIG. 4. Standard deviation of $C_p$ up to order 5, averaged over 100 RBNs, varying $k$ and $N$. The diversity of topological features is higher for small $N$ and large $k$.

TABLE I. Pearson product-moment correlation coefficient (PMCC) between the means and between the standard deviations of $I$ and $C(1,5)$ for each $K$ as $N$ varies ($N=25,50,100,150,200,300,400,500,750,1000$).

| $K$ | PMCC $[\langle I\rangle, \langle C(1,5)\rangle]$ | PMCC $\sigma(I), \sigma[C(1,5)]$ |
|---|---|---|
| 0.5 | 0.96168 | 0.96650 |
| 1.0 | 0.95901 | 0.99419 |
| 1.5 | 0.97631 | 0.96645 |
| 2.0 | 0.99531 | 0.98973 |
| 2.5 | 0.99277 | 0.98319 |
| 3.0 | 0.98884 | 0.94816 |

quantities show a power-law dependency on $N$ (Figs. 5 and 6). Further studies might give insight into the reasons and significance of this behavior. It does suggest, however, that $C_p$ does play a role, although small, even in relatively large networks, and thereby should be accounted for even in studies using large scale networks.

### C. Removing the dependence of $\langle I\rangle$ on $\langle C_p\rangle$

To observe if the value of $\langle C_p\rangle$ of the RBNs explains the dependence of $\langle I\rangle$ on $N$, we measured the $\langle I\rangle$ for RBNs with $k=2$, removing various orders of $C_p$. If the dependence of $\langle I\rangle$ on $N$ is completely removed (as the order of the local structures removed by rewiring increases), it indicates that it is sufficient to account for this global measure of local structures to generate ensembles of networks whose dynamics is not dependent on $N$.

Additionally, if this is true, then the variability of $I$ between networks of same $N$ and $k$, ought to be due to the variability in values of $C_p$ between those networks. Thus, the $\langle I\rangle$ of ensembles of networks whose $C_p$ is imposed to a fixed value should become independent of $N$ and its standard deviation diminish significantly, in comparison to the values in Fig. 2.

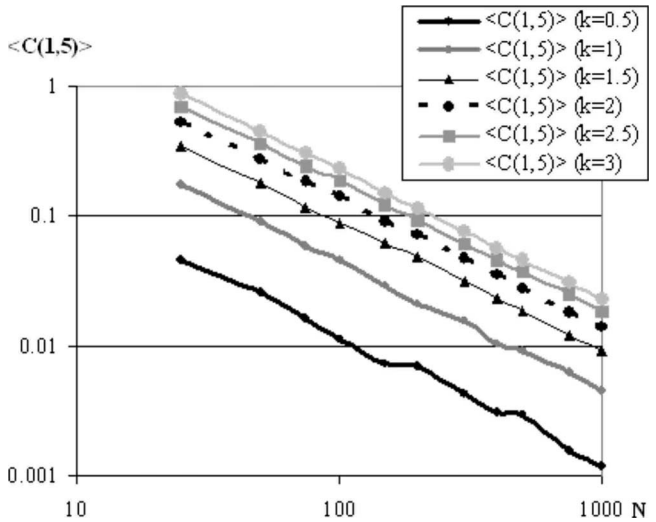


FIG. 5. $C_p$ up to order 5 on a log-log scale, averaged over 100 networks, for RBNs with random topology, varying $k$ and $N$. The linear relationships indicate power-law dependency of $C_p$ on $N$ for all $k$.

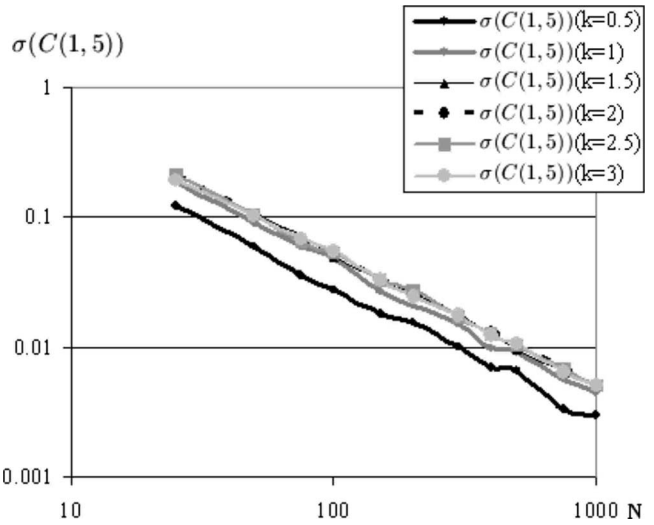


FIG. 6. Standard deviation of $C_p$ up to order 5 on a log-log scale, averaged over 100 networks, for RBNs with random topology, varying $k$ and $N$. The linear relationships indicate power-law dependency of the standard deviation of $C_p$ on $N$ for all $k$.

To test this, we generated ensembles of RBNs with $C_p=0$, of increasing order, up to order 5, using the method described in Sec. II D. The results are shown in Fig. 7. In agreement with our hypothesis, the higher the order of $C_p$ fixed to zero, the more uniform the dynamics of the ensemble is in terms of mutual correlation between temporal patterns of nodes activities.

The results in Fig. 8 agree with our hypothesis as well, by showing that $\sigma(I)$ when $C(1,5)=0$ is much smaller than $\sigma(I)$ when $C_p$ is not imposed, given the same size and connectivity (approximately converging only for $N>800$ where local structures are no longer present in networks of either ensembles).

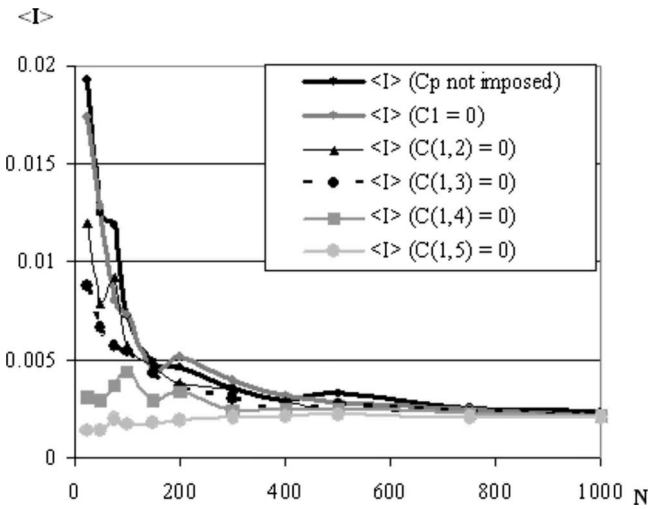


FIG. 7. $\langle I\rangle$ of RBNs computed from time series of length $T=10^3$. The topologies are built such that up to various $p$ the $C_p=0$. Each data point is an average of 100 independent RBNs, varying $N$ and with $k=2$.

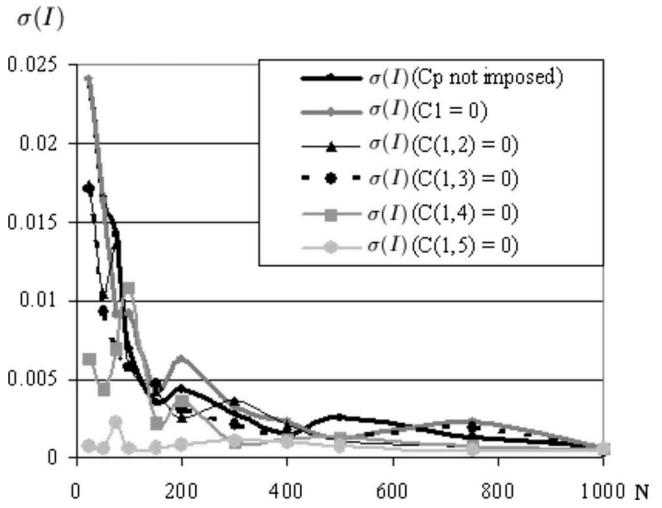


FIG. 8. Standard deviation of $I$ of RBNs computed from time series of length $T=10^3$. The topologies are built such that up to various $p$ the $C_p=0$. Each data point is an average of 100 independent RBNs, varying $N$ and with $k=2$.

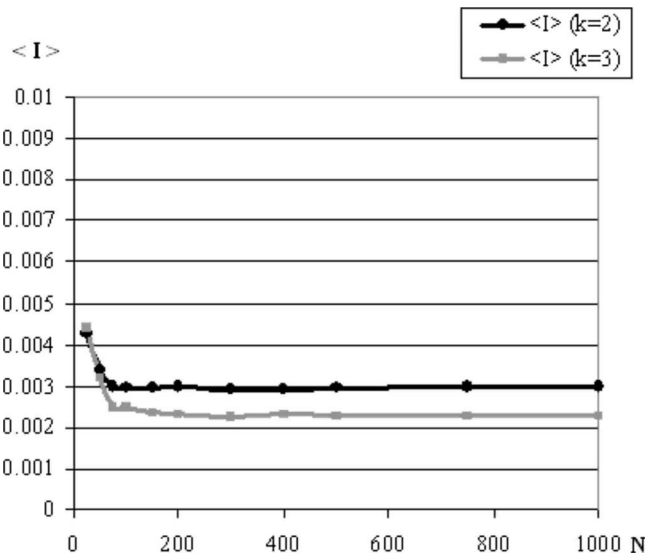


FIG. 9. $\langle I\rangle$ of RBNs computed from time series of length $T=10^3$. The RBN topologies are built such that $C_3$ is imposed to be between 0.09 and 0.11, while other orders of $C_p$ are fixed to zero up to order 5. Each data point is the average result of 100 independent RBNs, varying $N$ and with $k=2$ and $k=3$. The $\langle I\rangle$ of these RBNs shows virtually no dependency on $N$.

We note that based on the simulation results with the ensembles of networks used here, it is necessary to account for $C_p$ only up to order 5. Higher order terms, difficult to account for computationally, did not significantly affect the dynamics of RBNs. For higher values of $k$ than the ones considered here, it might become necessary to account for higher orders of $C_p$. However, it is increasingly difficult to find rewiring schemes to remove increasing orders in high $k$ networks, maintaining in- and out-degree distributions unchanged. We also note the consistency in the results in the sense that as higher orders of $C_p$ are fixed to zero by rewiring, the dependence of $\langle I\rangle$ on $N$ decreases further.

Another observation from Fig. 7, in the case of $C_p$ removed up to order 5, the $\langle I\rangle$ of networks with $N<100$ is slightly smaller than for larger RBN. This difference is of the order of $10^{-4}$, whereas the variation in $\langle I\rangle$ due to $C_p$ variation is on the order of $10^{-2}$. This weak dependency of $\langle I\rangle$ on $N$, only visible when removing $C_p$ up to order 5, is addressed in the conclusions.

It is now important to determine if one can generate RBNs whose dynamics is independent of $N$ by imposing a uniform $C_p$ value up to order 5. If $C_p$ can capture local structure effects on the global dynamics of networks, the $\langle I\rangle$ of networks of an ensemble with an imposed nonzero value of $C_p$ should be independent of $N$.

In Fig. 9 the $\langle I\rangle$ is shown for ensembles of RBNs with increasing size $N$. The topologies of these networks were generated by fixing $C_p$ of orders 1, 2, 4, and 5 to zero, while $C_3$ is imposed such that $0.09\leqslant C_3\leqslant 0.11$. $C_3$ is allowed an interval of values since setting it to a specific unique value makes the generation of RBN topologies that respect such a condition very time consuming.

As shown Fig. 9, the $\langle I\rangle$ is virtually independent of network size, for $N\geqslant 100$. Also, in comparison to the $\langle I\rangle$ of RBNs where $C_p$ is not fixed, for large values of $N$, the networks with fixed values of $C_p$ ought to have a higher $\langle I\rangle$, because, for large values N, the first ones have zero $C_p$. Finally, for small values of $N$, the situation should invert, i.e., the RBNs whose $C_p$ is not fixed, ought to have a higher $\langle I\rangle$ because they will have a higher $C_p$.

In addition, interestingly, networks with fixed $C_p$ and $K=2$ still maintain the property of maximizing the value of $\langle I\rangle$ [1]. This property does not appear to be significantly affected by the procedure of imposing values to the $C_p$ of the RBNs, which explains the difference between the values of $\langle I\rangle$ for $K=2$ and 3. Other factors, such as the average path length are almost negligible, otherwise one would expect the opposite results, namely, $\langle I\rangle$ of networks with $K=3$ is bigger than $\langle I\rangle$ of networks with $K=2$.

In Ref. [1], it was shown that $\langle I\rangle$ is maximized for critical RBNs whose $C_p$ is not fixed. Our results indicate that this holds true for RBNs with an imposed value of $C_p$. Also, for large $N$, both for $k=2$ and $k=3$, networks with fixed $C_p$ have higher $\langle I\rangle$ than otherwise, since if $C_p$ is not fixed it approaches zero for large $N$.

These results suggest that networks with higher $C_p$ will have, given equal values of $k$ and $N$, higher $\langle I\rangle$. Importantly, and in agreement with the goals of applying a constraint in the $C_p$ of the networks of the ensemble, the increase of $\langle I\rangle$ as $N$ decreases was reduced by two orders of magnitude, and only starts increasing at $N<100$, while previously, the increase occurred at $N<250$.

Here, we have mostly focused on the effects of varying $N$ in RBNs where the $C_p$ is imposed up to a certain order. Varying $k$ and maintaining $N$ fixed has the same effects, that is, imposing $C_p$ constant up to order 5 to this ensemble also results in removing the dependency of $\langle I\rangle$ of $k$ (data not shown).

## IV. CONCLUSIONS AND DISCUSSION

It was known that local structure features have a non-negligible effect on the dynamics of networks

[1,8,10,11,14,20]. Here, we propose a global topological measure that allows the prediction of the overall effect of these local structures on the dynamics.

The results of our numerical simulations show that $C_p$ is an appropriate global measure of the "degree" of effects of local structures in network dynamics. Importantly, we show that the original clustering coefficient [9] (here corresponding to $C_3$) is insufficient to explain the effects of local structures. This was shown by the results in Fig. 7, where it is visible that in RBNs with $C_{1,3}=0$, the $\langle I\rangle$ still varies with $N$. Namely, it increases as $N$ decreases, which means that the local structures of these networks still play a role in the dynamics. By removing the fourth and fifth order $C_p$, this dependency of $\langle I\rangle$ on $N$ is removed and even the smallest networks behave identically to the larger "treelike" networks (i.e., have identical $\langle I\rangle$).

Additionally, the values of the standard deviation of $C_p$ over the ensemble explain the diversity in the dynamics of RBNs where only the values of $N$ and $k$ are constrained. The $\langle I\rangle$ of RBNs with zero or some other fixed value of $C_p$, becomes almost independent of $N$. The dynamics of RBNs of equal $N$ and $k$ also become far more uniform. Finally, the higher the order of $C_p$ removed, the more the networks behave similar to "infinite" (treelike) structures.

This means that the method of network construction proposed here allows constructing RBNs with a small number of nodes to which mean field approximations can be applied, since these usually assume treelike structures ($C_p=0$). It also allows the generation of large networks with a given $C_p$, which might be of importance especially when using the ensemble approach [2] to study the dynamics of biological networks such as GRNs, since these rarely have locally treelike structures.

Several observations might deserve future attention. First, we note that not all RBNs are rewirable such that a given $C_p$ value is attainable for a given number of trials. Thus, the search for possible rewiring of a connection was limited in our simulations to a certain number of attempts. If unsuccessful, the process is restarted using a newly generated RBN. This procedure might introduce unknown biases in the choice of networks if, for example, the property of being rewirable is somehow correlated with the network's $I$.

Second, when removing $C_p$ up to order 5, it was observed that the $\langle I\rangle$ of RBNs with $N<100$ is slightly smaller than for RBNs with $N\geqslant 100$ (Fig. 7). This weak effect might emerge from various sources. One possibility is that in such small networks there are so few rewiring schemes, maintaining the degree distributions, that the resulting state space of solutions might have very particular, and currently unknown, properties. For example, limitations in rewiring might cause the directionality of connections to become biased in some way. We point out that, for this range of N values, on average 1000 networks were needed to find a network that can be rewired.

Another possible cause is that, while a time series of length 1000 of a large RBN will likely consist mostly of a long transient, in small RBNs most states of the time series are from attractors. On attractors, more nodes tend to be frozen than on transients, lowering $\langle I\rangle$.

Third, as a first approach, directionality of connections was not accounted for when computing $C_p$. There might be specific networks or methods of generating ensembles, where it is not sufficient to account for the $C_p$ assuming an undirected graph. In those cases, the definition of $C_p$ proposed here might need to be extended to accommodate the directionality of the connections.

Additionally, it remains an open question whether our results apply to other distributions of in-degrees instead of Poisson. Although we believe that, qualitatively, the results would lead to the same conclusions, detailed analysis of how $\langle I\rangle$ varies with $k$ and $N$ in those cases might prove interesting to investigate.

Overall, the results show that "non-tree-like" local structures cause a significant increase in the dynamical correlation between nodes in a RBN. This agrees with one of its topological consequences, the decrease of the average path length [10]. Also, it suggests that networks, such as biological ones, subject to selection, are most likely not treelike structures, if high correlation between the functioning of its elements is required.

If a network is to be selected based on its $\langle I\rangle$, our results indicate that this can be done by selecting a certain $C_p$. Moreover, if the maximization of the expected range of variability of $I$ is what is being selected, then the range of variability of $C_p$ should be maximized. This range is most likely bounded by the necessity of maintaining all nodes connected to a single network, and perhaps also by the necessity of minimizing the network's average path length.

Recently, it was shown that critical networks maximize $\langle I\rangle$ in RBNs [1]. However, in that study, the only topological constraints of the ensemble of networks were in the values of $N$ and $k$. Our results suggest that the value of $C_p$ of the networks in the ensemble cannot be neglected, and perhaps the notion of criticality could be extended to incorporate this topological feature.

To the extent that evolutionary fitness depends on optimal capacity for coordinating the dynamical behavior of the components of an organism, and RBNs models capture essential features of the organization of networks such as gene regulatory networks, networks with a high $C_p$ ought to be naturally favored since these maximize $\langle I\rangle$. Interestingly, studies of the structure of natural networks [21] suggest that their $C$ (as originally defined [10]) is higher than expected by chance. If this holds true, it necessarily implies that the same is true for $C_p$.

Our conclusions suggest that when applying the ensemble approach [2] to study biological networks, such as genetic networks, the networks of these ensembles ought to have similar $C_p$ values as the biological network in question, if they are to properly mimic the global dynamical properties as well. Moreover, our measurements of the standard deviation of $C_p$ and the distribution of $I$ values of networks with equal $N$ and $p$ show that, for those networks to have uniform dynamical properties, it is not sufficient to impose an average $C_p$ over the ensemble. Instead, the $C_p$ of each individual network must be fixed to the same value.

These results are an example of how a system's structure affects its dynamics regardless of the nature of the components and interactions [20]. That is, Boolean network dynamics are highly constrained by their local topological structure. Changing local topology while maintaining update rule dis-

tributions results in significant changes in average global coordination. We expect that these results are extendable to a wide variety of systems, including stochastic delayed genetic networks [22].

Also, we note that the results may be of biological relevance, by suggesting that large networks, including genetic networks, might follow local and global topological constraints, such as criticality and high $C_p$, if they are to maximize the coordination between their components. Namely, the results are another tentative evidence that one should not expect to find biological networks with random topology at local or global level.

## ACKNOWLEDGMENTS

This work was supported by the Academy of Finland, Project No. 213462 (Finnish Centre of Excellence program 2006-11).

[1] A. S. Ribeiro, S. A. Kauffman, J. Lloyd-Price, B. Samuelsson, and J. E. S. Socolar, Phys. Rev. E **77**, 011901 (2008).
[2] S. Kauffman, J. Theor. Biol. **230**, 581 (2004).
[3] P. Rämö, J. Kesseli, and O. Yli-Harja, J. Theor. Biol. **242**, 164 (2006).
[4] A. S. Ribeiro and S. Kauffman, J. Theor. Biol. **247**, 743 (2007).
[5] E. H. Davidson, *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution* (Elsevier, Amsterdam, 2006).
[6] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A.-L. Barabasi, Science **297**, 1551 (2002).
[7] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, Science **306**, 1555 (2004).
[8] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
[9] M. Granovetter, Am. J. Sociol. **78**, 1360 (1973).
[10] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[11] A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. Oltvai, and A.-L. Barabasi, Proc. Natl. Acad. Sci. U.S.A. **101**, 17940 (2004).
[12] M. A. Serrano and M. Boguñá, Phys. Rev. E **74**, 056114 (2006).
[13] A. Pusch, S. Weber, and M. Porto, Phys. Rev. E **77**, 017101 (2008).
[14] A. S. Ribeiro and P. G. Lind, Phys. Scr. **T118**, 165 (2005).
[15] P. Krawitz and I. Shmulevich, Phys. Rev. Lett. **98**, 158701 (2007).
[16] P. Krawitz and I. Shmulevich, Phys. Rev. E **76**, 036115 (2007).
[17] S. Kauffman, J. Theor. Biol. **22**, 437 (1969).
[18] S. Kauffman, *The Origins of Order* (Oxford University Press, New York, 1993).
[19] D. Moore, *Basic Practice of Statistics* (WH Freeman, New York, 2006).
[20] L. von Bertalanffy, *General System Theory: Foundations, Development, Applications* (George Braziller, London, 1976).
[21] A.-L. Barabasi, *Linked: The New Science of Networks* (Perseus Publishing, London, 2002).
[22] A. S. Ribeiro, R. Zhu, and S. Kauffman, J. Comput. Biol. **13**, 1630 (2006).