

## Research



**Cite this article:** Alexander B, Pushkar A, Girvan M. 2021 Phase transitions and assortativity in models of gene regulatory networks evolved under different selection processes. *J. R. Soc. Interface* **18**: 20200790. <https://doi.org/10.1098/rsif.2020.0790>

Received: 29 September 2020

Accepted: 18 March 2021

### Subject Category:

Life Sciences—Physics interface

### Subject Areas:

biocomplexity, computational biology

### Keywords:

gene regulation, evolution, explosive percolation, complex networks, assortativity

### Author for correspondence:

Brandon Alexander

e-mail: [bralex1@umd.edu](mailto:bralex1@umd.edu)

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.5359388>.

# Phase transitions and assortativity in models of gene regulatory networks evolved under different selection processes

Brandon Alexander<sup>1,2,3</sup>, Alexandra Pushkar<sup>5</sup> and Michelle Girvan<sup>2,3,4,6</sup>

<sup>1</sup>Department of Mathematics, <sup>2</sup>Institute for Physical Science and Technology, <sup>3</sup>Program in Applied Mathematics & Statistics and Scientific Computation, and <sup>4</sup>Department of Physics, University of Maryland, College Park, MD 20740, USA

<sup>5</sup>Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA

<sup>6</sup>Santa Fe Institute, Santa Fe, NM 87501, USA

BA, 0000-0002-3488-0558

We study a simplified model of gene regulatory network evolution in which links (regulatory interactions) are added via various selection rules that are based on the structural and dynamical features of the network nodes (genes). Similar to well-studied models of ‘explosive’ percolation, in our approach, links are selectively added so as to delay the transition to large-scale damage propagation, i.e. to make the network robust to small perturbations of gene states. We find that when selection depends only on structure, evolved networks are resistant to widespread damage propagation, even without knowledge of individual gene propensities for becoming ‘damaged’. We also observe that networks evolved to avoid damage propagation tend towards disassortativity (i.e. directed links preferentially connect high degree ‘source’ genes to low degree ‘target’ genes and vice versa). We compare our simulations to reconstructed gene regulatory networks for several different species, with genes and links added over evolutionary time, and we find a similar bias towards disassortativity in the reconstructed networks.

## 1. Introduction

Mathematical models of gene regulatory networks provide a powerful tool for understanding the complex features of genetic control. While various modelling efforts have been successful at explaining gene expression patterns, much less is known about how evolution shapes the structure of these networks. Recent studies suggest that evolutionary ‘tinkering’ plays a large role in the organization of biological networks [1]. Furthermore, these networks are often thought to exist near some critical point [2–4], where dynamic variability is maximized without reaching widespread network failure/breakdown. The phase transition between stability and instability in networks has been widely investigated, with studies focusing on the effects of topological features [5–8], dynamical features [9–12] or both [13] and their contributions to the location and behaviour of the transition. Of particular interest is the evolutionary process that leads to this critical point and how this process depends on both the topological and dynamical properties of the network and its nodes. Some previous studies have considered the evolution of these networks but begin with already-established networks that are then allowed to change over time [14–16]. By contrast, we consider the process from the very beginning, starting with an ‘empty’ network.

In order to gain insight into this process, we study simple models of gene network evolution in which links are added according to various competitive selection rules. Similar rules were explored by Achlioptas *et al.* [17], who demonstrated that undirected networks grown following competitive link selection can have drastically delayed percolation transitions, leading to a seemingly

discontinuous ‘explosive’ transition, a topic which has since become popular in the network science literature [8,18–23]. Squires *et al.* [8] extended the idea of competitive selection to directed networks, where similar, but less drastic ‘weakly explosive’ transitions occurred. Incrementally growing networks according to competitive selection rules is a natural choice for modelling biological networks, which are the result of generations of improvements through the complex process of natural selection. For the same reason, biological networks share commonalities with systems that are said to exhibit highly optimized tolerance, in which a yield or fitness is to be maximized [24,25]. In such systems, competitive selection rules can be used heuristically to maximize a yield dependent on the size of a giant component. Besides evolution through natural selection, biological systems develop over time through neutral evolution, or evolution through mutation without selection [26–29]. While we primarily focus on networks evolved via selection, we will also consider a simple case of neutral evolution for comparison. Other approaches have used a combination of mutation and selection to evolve networks that maximize the topological entropy [30] or minimize the distance to a time-dependent network output function [15]. By comparison, our modelling efforts focus on growth starting from a blank state, serving to highlight how simple differences in selection rules can give rise to very different network structures and change the nature of phase transitions in the network.

Our models of network growth build upon various studies of Boolean models of gene regulation. From a dynamical perspective, Boolean networks have been widely used as mathematical models of gene regulation since their introduction by Kauffman [5]. In these models, at any given time, nodes (genes) exist in one of two states: ‘on’ (1), meaning that the gene is expressed, or ‘off’ (0), meaning that the gene is unexpressed. In Kauffman’s original model, gene states are synchronously updated at each time step according to a truth table, in which the state of a gene at time  $t$  is determined by the states of all of its input genes at time  $t-1$ . (While we connect our evolutionary modelling approach to the case of synchronous update, we also discuss how the approach may be connected to cases of asynchronous update.) By tuning system parameters, the models can exhibit a transition from stable dynamics to ‘chaotic’ (unstable) dynamics [5,6,10]. Stability here is measured by the saturated average Hamming distance, or fraction of genes that differ in state after a long time, between the trajectories of two close initial conditions. Since a finite network of size  $N$  has at most  $2^N$  possible states and the dynamics are deterministic, the trajectories will eventually reach a steady state in the form of a fixed point or periodic orbit. If the state of a gene in the steady state differs between these two trajectories at time  $t$ , the gene is said to be ‘damaged’ at time  $t$  [31].

As demonstrated by Squires *et al.* [31], the stability of Boolean networks can be mapped directly to a percolation problem under certain assumptions. In a percolation problem, the size of a ‘giant component’ is tracked as a certain tuning parameter is varied, typically passing through a critical value where the giant component transitions from being vanishingly small to being a non-trivial fraction of the size of the network. Following Squires *et al.*, by setting a gene’s sensitivity, i.e. the probability that the gene changes state if one of its inputs changes state, equal to its probability of being ‘damaged’, we can map the saturated average Hamming distance, which

provides a measure of instability in the gene network dynamics, directly onto the size of the largest out-component in the subnetwork of ‘damaged’ nodes. As either the node sensitivities or network connections are tuned, the phase transition from dynamic stability to instability then corresponds to the percolation transition of this giant component.

We extend this idea by allowing the topology of our network to change over time following different competitive selection rules, which works as a primitive model of biological networks evolving via natural selection. As new links are added, at some critical point, percolation will occur in the damaged subnetwork, corresponding to the network’s transition from stability to instability.

To summarize, we are motivated by a model of gene regulatory dynamics in which there are benefits for the system to exist on the stable side of the transition between stability and instability. We utilize the mapping between a model of this kind and an appropriately defined percolation problem to explore simple models of evolution via growth. We note that the percolation problem can be considered as its own simple abstraction of the propagation of failure in gene regulatory networks. Using the percolation framework, we investigate the following growth processes: (i) random link addition, as a control; (ii) sensitivity-based competitive selection; (iii) structure-based competitive selection; and (iv) hybrid competitive selection. We find that incorporating network structure into the selection rule is key to delaying percolation of the damaged subnetwork, even while having no knowledge of each gene’s sensitivity, or probability of becoming damaged. In fact, incorporating information about sensitivity can backfire, leading to highly exclusive networks that would rather form large components than connect to high-risk genes. Finally, we compare the results of our simulations to data for several empirically derived gene regulatory networks and find similarities in structural properties.

## 2. Methods

### 2.1. Modelling failure propagation in gene networks as a percolation problem

We construct a model of gene regulation as follows: we first consider damage propagation on a network with a fixed topology, which will later be allowed to evolve. We are motivated by the traditional Boolean models of gene regulation in which each gene in the network has a fixed ‘truth table’, or output rule, that determines its state (0 or 1) at time  $t$  based on the states of its inputs at time  $t-1$ . To simplify the problem for our purposes but still capture the effects of network topology, we apply the semi-annealed approximation [10], in which the network remains fixed, but the truth table for gene  $i$  is no longer fixed but rather randomly filled in at each time step according to some bias  $b_i$ , the probability that its state is 1. We can think of  $b_i$  as the expression bias of gene  $i$ , with genes having a wide distribution of expression biases. Working within the semi-annealed approximation, from the gene’s bias, we can then calculate its sensitivity  $q_i = 2b_i(1-b_i)$ , which represents the probability that a gene changes state if the state of one or more of its inputs change. This can also be thought of as the probability that a gene spreads ‘damage’. Damaged genes are those whose states differ in the steady states of two initially close trajectories of the network’s dynamics, meaning a small perturbation in the state of the network can lead to unpredictable dynamics of such genes, likely causing improper regulatory behaviour.

Following [31], for a locally treelike network under the semi-annealed approximation, dynamic instability of the network, measured in terms of the average saturated Hamming distance between trajectories of two close initial conditions, is equivalently measured by the expected size of the giant out-component of the damaged subnetwork, which takes into account both the network topology and the genes' propensities for spreading damage, i.e. their sensitivities. In particular, the phase transition from stability to instability occurs simultaneously as the onset of a non-trivial out-component of the damaged subnetwork (i.e. a component whose size does not go to zero as the number of nodes gets larger). In the electronic supplementary material, we show that our simulated networks display the relevant locally treelike properties.

While we are motivated by a model of synchronously updated genes with locally treelike connectivity, we note, importantly, that the semi-annealed approximation works well even for an asynchronous update scheme and for the case of networks with many small loops. Although there are important differences between synchronous and asynchronous update models (e.g. the occurrence of limit cycles [32]), Pomerance *et al.* [10] and Squires *et al.* [13] demonstrate that the semi-annealed approach for determining the transition between global stability and instability (and the growth of perturbations) also applies for asynchronous update. Further, Pomerance *et al.* [10] show that the approximation holds even for networks that deviate substantially from the locally treelike assumption. This is consistent with studies that show that calculations involving the locally treelike approximation provide accurate results even for networks that are very far from treelike [33,34].

By using the semi-annealed approximation, we are able to work entirely in the realm of the percolation problem, without needing to directly simulate the complex dynamics of the Boolean network. In fact, we can think of the percolation problem as its own simple abstraction of gene regulatory networks, in which genes become damaged with different propensities and they can propagate that damage to their target genes.

To incorporate evolutionary changes into our percolation framework, we add new links (edges) individually to the network according to a given selection rule, described in §2.3. Each new link represents a new regulatory interaction that has evolved in the network, potentially changing the stability of the dynamics. We then repeat this procedure until a sufficient number of links have been added to the network.

For our simulations, we initialize the network to consist of  $N$  genes and no links. We let  $E$  denote the number of links (edges) in the network at later steps. The gene biases are drawn from a uniform distribution  $b_i \sim U[0, 1]$ . Note that for a single simulation, bias values do not change, meaning we are fixing an inherent dynamical property of each gene. This is a restrictive assumption, but it provides a nice starting point for comparing networks evolved under different selection rules. Before the simulated evolution process begins, each gene is assigned damaged status with probability  $q_i$ , leading to an average of one-third of all genes being damaged (due to how biases were chosen). By averaging over multiple simulations, we found that keeping the same damage status constant throughout all steps of a single simulation produces very similar results to re-randomizing the statuses at each individual step, so we have chosen to keep the damaged statuses constant. Finally, at each step, we track the size of the giant components in both the full network and damaged subnetwork, which also allows us to efficiently simulate each growth process using the algorithm suggested in [8].

## 2.2. Tracking network components

For a gene  $i$ , define the *in-component*  $IN(i)$  to be the set of all genes that can reach  $i$  via a directed path in the regulatory network. Similarly, define the *out-component*  $OUT(i)$  to be the set

of all genes that  $i$  can reach via a directed path. The *strongly connected component*  $SCC(i)$  is the intersection of these two sets and the *bow tie*  $BT(i)$  their union.

For a directed network, we say the *giant strongly connected component* (GSCC) is the largest strongly connected component. We then define the *giant out-component*  $GOUT = OUT(GSCC)$ , the *giant in-component*  $GIN = IN(GSCC)$ , and the *giant bow tie*  $GBT = BT(GSCC)$ . In systems where percolation occurs, all four giant components form simultaneously at the same critical point [35].

Within our networks, we will also have a subset of genes that are 'damaged'. The damaged subnetwork is the directed network consisting of these genes and any links that exist between them. We use an asterisk (\*) to distinguish the components of the damaged subnetwork from those of the full network.

## 2.3. Network growth via selection

To decide which link to add at each step, we use a class of competitive selection rules based on those introduced by Achlioptas *et al.* [17] and extended to directed networks by Squires *et al.* [8]. The form of these rules is as follows:

1. Consider  $m$  potential 'source' genes  $\{i_1, \dots, i_m\}$  and  $m$  potential 'target' genes  $\{j_1, \dots, j_m\}$ . All genes are sampled uniformly from the network.
2. Select the directed link  $i \rightarrow j$  such that

$$i = \arg \min_{i \in \{i_1, \dots, i_m\}} f_s(i), \quad j = \arg \min_{j \in \{j_1, \dots, j_m\}} f_t(j). \quad (2.1)$$

3. If the link  $i \rightarrow j$  already exists in the network or if  $i = j$ , then a different link is selected starting back at step 1.

The choices of  $f_s(i)$  and  $f_t(j)$  are made so as to delay the percolation transition of  $GOUT^*$ . We consider three forms of competitive selection (sensitivity-based, structure-based, and hybrid) and the case of random link addition as a control. As in the original Achlioptas study, the selection rules we consider rely only on local information about the candidate links. In the electronic supplementary material, we study the extreme case of  $m = N$  for structure-based selection and demonstrate that, as expected, having access to global information when adding links serves to delay the transition for a much, much longer period of time.

### 2.3.1. Random selection

In the case of  $m = 1$ , we select a random link uniformly from all possible links. This is equivalent to the directed Erdős-Rényi process [8,36]. Random selection serves as a simple baseline to which we can compare the performance and behaviour of other selection rules. In addition, one can consider random selection as a model for neutral evolution.

### 2.3.2. Sensitivity-based selection

To consider the effect of sensitivity, we define

$$f_s(i) = q_i, \quad f_t(j) = q_j. \quad (2.2)$$

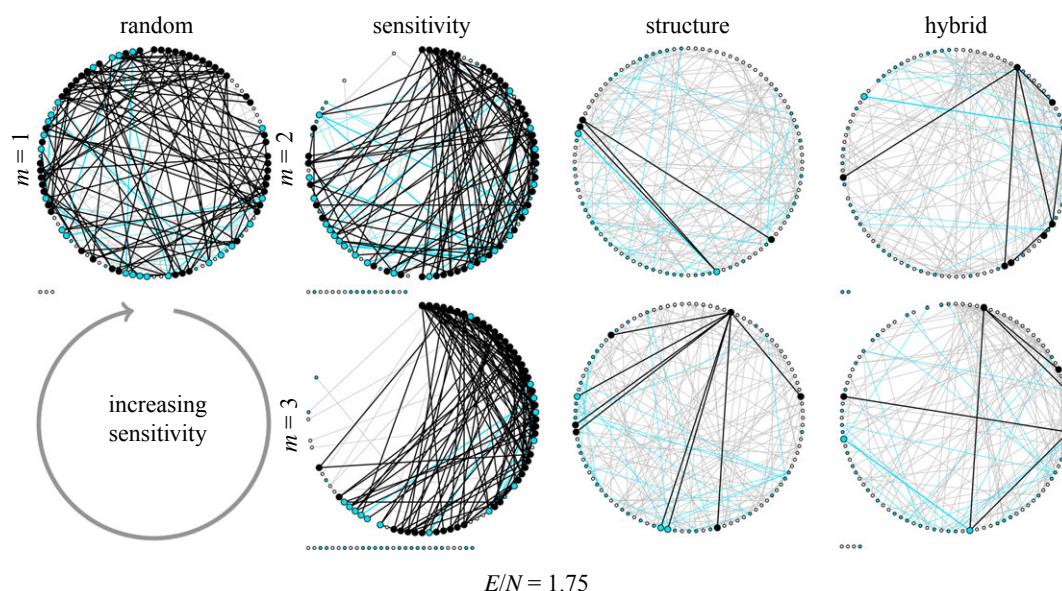
Genes with higher sensitivity values are more likely to be damaged. The strategy used by sensitivity-based competition is then to prevent growth of the damaged subnetwork by preventing 'risky' genes from connecting.

### 2.3.3. Structure-based selection

To consider the effect of network structure in delaying the onset of percolation, we define

$$f_s(i) = |IN(i)|, \quad f_t(j) = |OUT(j)|. \quad (2.3)$$

This rule is a generalization of the dCDGM rule applied to directed networks [37]. The strategy imposed by structure-based



**Figure 1.** Structure of networks formed under our four different selection rules on a network of size  $N = 100$ , stopped after  $E/N = 1.75$ . Nodes are sorted according to sensitivity with least sensitive at the top of the circle and increasing in sensitivity clockwise. If a node is damaged, it is filled in blue. Nodes in GOUT are enlarged and filled in black (if undamaged) or blue (if damaged); all other nodes are filled in white. Isolated nodes with no links are placed below each network with their corresponding locations in the circular layout left empty. Thick links connect nodes in GOUT. Links are blue if they connect two damaged nodes, black if they connect nodes in GOUT that are not both damaged, or grey otherwise. We see that GOUT is much larger for the random and sensitivity-based processes as compared with the structure-based and hybrid processes. Both the sensitivity-based and hybrid processes show a higher density of links between low sensitivity nodes, which is much more pronounced in the former.

competition is to prevent percolation in the damaged subnetwork by preventing large components from forming in the full network.

### 2.3.4. Hybrid selection

The final rule is a combination of structure-based competition and sensitivity-based competition. We set

$$f_s(i) = q_i \times |\text{IN}(i)|, \quad f_i(j) = q_j \times |\text{OUT}(j)|. \quad (2.4)$$

The hybrid strategy will prevent components from becoming large in the full network, while also showing a preference towards connecting low risk (low sensitivity) genes. High sensitivity genes will only be accepted if they belong to small components.

## 3. Results

### 3.1. Visualizing networks evolved via different selection rules

For each of our network measures, we use the average node degree  $E/N$  as the tuning parameter. We can first gain intuition about the structure of our networks by considering smaller networks of size  $N = 100$ , as is done in figures 1 and 2. We begin by considering the networks when  $E/N = 1.75$ , a point after which percolation of GOUT occurs for both the random and sensitivity-based selection rules, but not the others. Compared to random selection, sensitivity-based competition has a preference towards connecting low sensitivity genes, even going so far as to exclude a large number of high sensitivity genes from the network, an effect that is more exaggerated for higher values of  $m$ . By comparison, both the structure-based and hybrid competition rules take much longer to form a GOUT component. These two methods are nearly indistinguishable, except the hybrid method mimics some properties of sensitivity-based competition, namely having a higher link

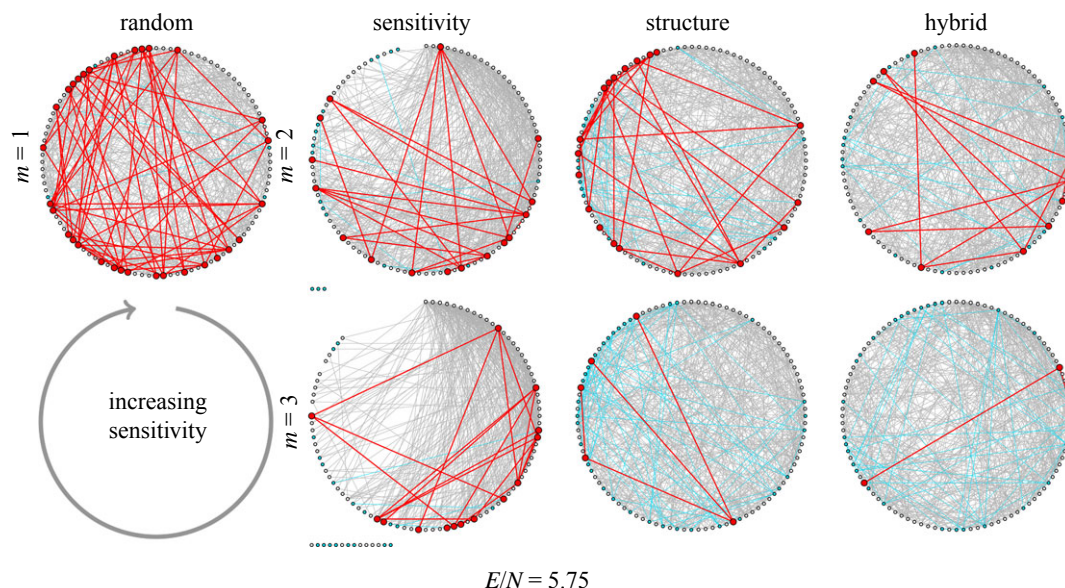
density between low sensitivity genes and exclusion of high sensitivity genes.

We next consider a higher parameter value of  $E/N = 5.75$ , where GOUT\* has now formed for random and sensitivity-based selection and is beginning to form for structure-based and hybrid selection. Because sensitivity-based selection avoids connecting to high sensitivity genes overall, the GOUT\* component mainly consists of lower sensitivity genes. Since most damaged genes have higher sensitivities, this also leads to fewer links overall between damaged genes and a smaller GOUT\* component than in the case of random selection. We also observe that in the random selection and sensitivity-based selection, nearly every link between damaged genes belongs to GOUT\*. By comparison, structure-based and hybrid selection have a large number of links between damaged genes, but only a few belong to GOUT\*, since the individual out-components are kept small and avoid connecting to each other.

### 3.2. Growth of giant components

Now equipped with better intuition about these selection methods, we turn our attention to growth of the connected components in the  $N \rightarrow \infty$  limit. We can approximately achieve this goal by simulating on a network of size  $N = 10^5$ , shown in figure 3.

As before, we begin by considering the growth of the full-network components, ignoring damage status. As expected, the formation of GOUT is delayed much further in structure-based and hybrid selection methods, due to their nature of selecting links to maintain smaller component sizes overall. Meanwhile, sensitivity-based selection instead tends towards earlier formation of the giant component, with higher values of  $m$  pushing this transition earlier. By increasing  $m$  in structure-based selection, we are more likely to find a smaller component, and thus are able to spread



**Figure 2.** Network structure for a network of size  $N = 100$  grown under the four selection rules stopped after  $E/N = 5.75$ . As in figure 1, nodes are sorted according to sensitivity, with least sensitive at the top of the circle and increasing in sensitivity clockwise and with isolated nodes left out of the circular layout and placed below each network. We now focus on GOUT\* rather than GOUT. Nodes belonging to GOUT\* are enlarged and red with thick red links connecting between them. Damaged nodes that do not belong to GOUT\* are blue, with blue links connecting pairs of damaged nodes. All other nodes are coloured white, and all other links are coloured grey. Even though the size of GOUT for the sensitivity-based process is similar to the random process, the former has a much smaller GOUT\*. The structure-based and hybrid processes show even smaller GOUT\*, despite a large number of connections between damaged nodes.

out links more efficiently to avoid forming larger components. By comparison, in sensitivity-based selection, increasing  $m$  makes us more likely to find low sensitivity genes (and avoid high sensitivity genes altogether), resulting in a higher link density among the lower sensitivity genes and a faster formation of a giant component. The exclusion of highly sensitive genes leading to faster emergence of a giant component also explains why hybrid selection transitions slightly earlier than structure-based selection.

Of greater interest is the formation of GOUT\*. We again have structure-based and hybrid selection leading to later percolation of GOUT\*, but unlike before, sensitivity-based selection leads to formation of GOUT\* later than random selection. The sensitivity-based selection rule is designed to prevent formation of GOUT\* by preventing connections between damaged genes, which leads to slower formation of GOUT\* but not GOUT. Comparing structure-based and hybrid selection, we see that the critical point at which GOUT\* forms is approximately the same when  $m = 2$ . But when  $m$  is increased, structure-based selection can delay the transition much more easily. Again, the additional effect introduced by considering sensitivities seems to be at play. Hybrid selection is willing to form larger components, as long as the genes involved have relatively low sensitivity. Eventually, however, the high sensitivity genes will have smaller components and be more favourable to gain a new link. It appears that this shift is enough to lead to faster growth of GOUT\* than if all components were to be kept smaller, ignoring sensitivity.

Other than the location of the percolation transition for GOUT\*, we can also look at the growth rate of this component post-transition. One advantage of sensitivity-based and hybrid selection rules is their slow growth of GOUT\*, despite the earlier transition. Since these rules will tend to prevent connections between damaged genes, a certain number of highly sensitive genes will have trouble joining with GOUT\*, and the rate at which GOUT\* can achieve its

maximum size is reduced. In particular, increasing  $m$  will slow down the growth even further in sensitivity-based and hybrid selection, since larger values of  $m$  lead to increased exclusion of genes.

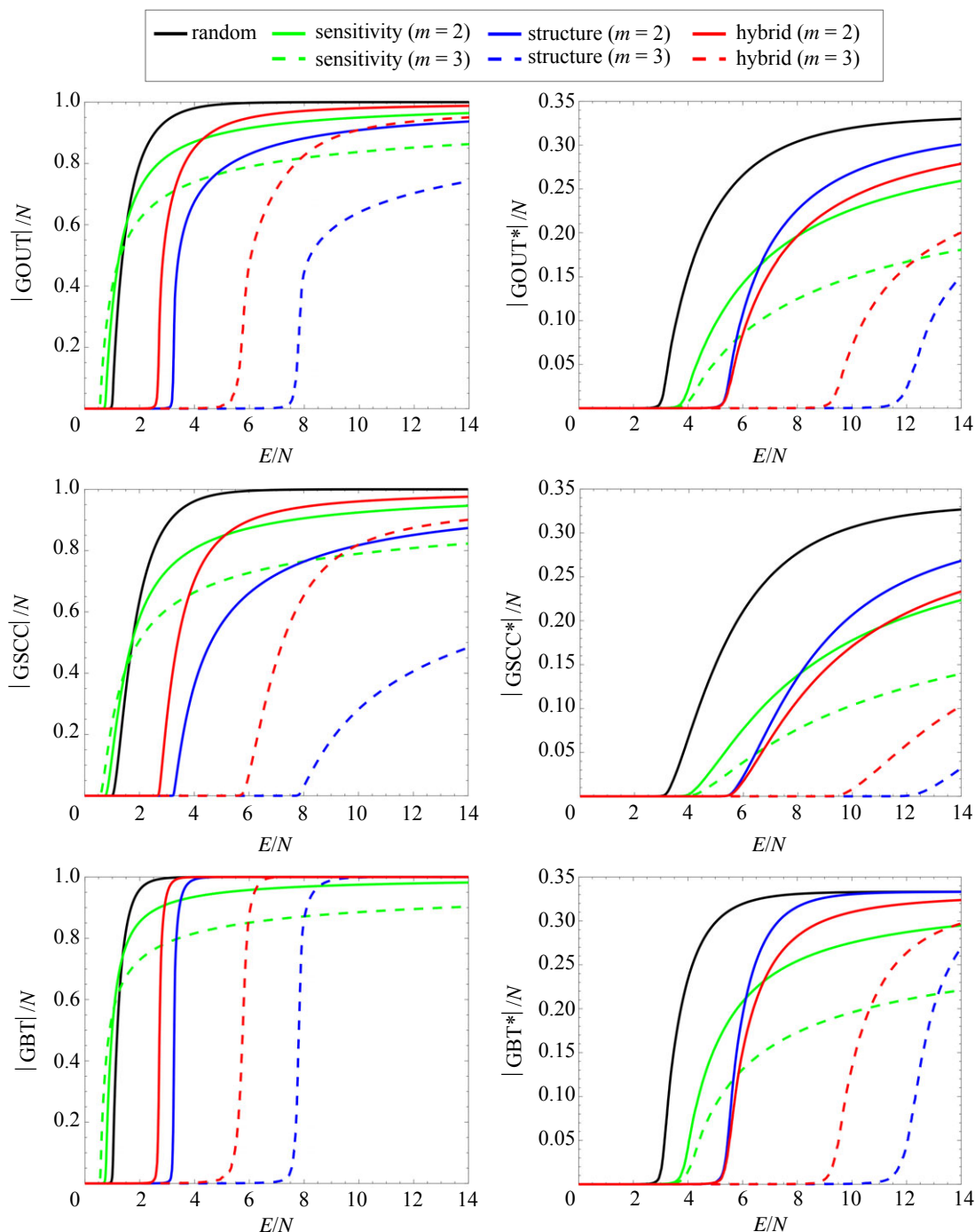
### 3.3. Maximizing yield

Our model exhibits certain features common to systems developed through *highly optimized tolerance* [24,25], namely incremental improvements in response to maximizing some yield or fitness function, leading to specialized structures. While we do not directly maximize any yield function, our selection methods can be thought of as heuristics for maximizing

$$Y = \frac{1}{N}(n_{\text{active}} - |\text{GOUT}^*|), \quad (3.1)$$

where  $n_{\text{active}}$  is the number of genes with at least one link. Only genes that interact with others will contribute to the dynamics in a meaningful way. Therefore, we should maximize the number of active genes to increase the dynamical variation of the system. At the same time, we do not want GOUT\* to grow too big, since the size of GOUT\* directly gives the number of genes that are damaged in the steady state. We consider this yield for each of our selection methods, as shown in figure 4.

Although avoiding connection to high-risk genes is a potentially useful feature in preventing connections between damaged genes, we do not want these genes to be kept entirely isolated from the network. In the case of each selection rule, the yield quickly grows as the number of active genes increases, but reaches a maximum when either all genes have been assimilated into the network or the percolation transition of GOUT\* has been reached. The yield then begins to decrease and approach  $1 - N^*/N$ , the fraction of undamaged genes in the network.



**Figure 3.** Growth of the connected components on the full network (left) and the damaged subnetwork (right). We compare growth under random selection to our three competitive rules. Results shown are the averages of 100 runs on networks of size  $N = 10^5$ . We omit GIN and GIN\* due to their symmetry with GOUT and GOUT\*, respectively. Of the selection rules considered, we see that structure-based selection (blue curves) does best at delaying the phase transition for the full network, while the structure-based and hybrid selection (red curves) exhibit simultaneous transitions when we consider only the damaged subnetwork, with the hybrid selection showing greater suppression of the component size after the transition.

Sensitivity-based selection is the only of our rules that does not reach a maximum of  $Y=1$  before decreasing, due to a large number of inactive genes. One should note that after peaking, sensitivity-based selection does continue to assimilate genes into the network, but at a slower rate than the growth of GOUT\*.

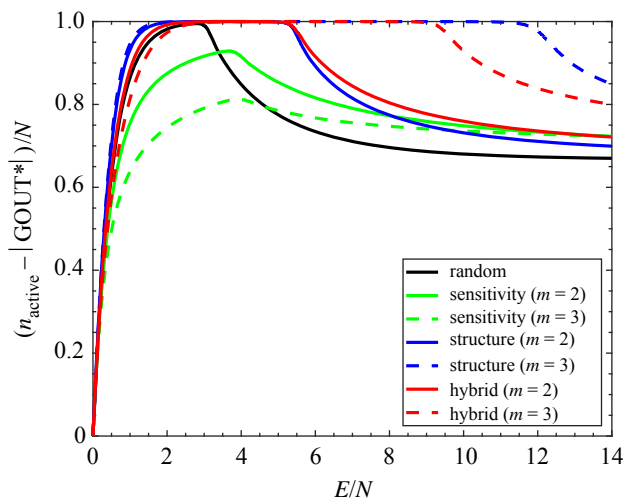
For structure-based and hybrid selection, there is a long period during which the yield is maximized. Both rules are able to quickly incorporate genes into the main network, since small components of size one are much more preferable. Hybrid selection tends to grow slower at first, due to its slight exclusion of high sensitivity genes. For  $m=2$ , when both of these rules form GOUT\* at approximately the same time,

hybrid selection performs slightly better, since the slower growth of GOUT\* leads to a higher yield after the transition. Increasing to  $m=3$  and beyond, however, the longer delay in percolation makes structure-based selection significantly better.

### 3.4. Correlations

Finally, we consider two forms of correlations in our networks. The first is the in-out degree assortativity, given by

$$r = \frac{\sum_{i \rightarrow j} (d_{in}^i - \bar{d}_{in}^s)(d_{out}^j - \bar{d}_{out}^t)}{\sqrt{\sum_{i \rightarrow j} (d_{in}^i - \bar{d}_{in}^s)^2} \sqrt{\sum_{i \rightarrow j} (d_{out}^j - \bar{d}_{out}^t)^2}}, \quad (3.2)$$

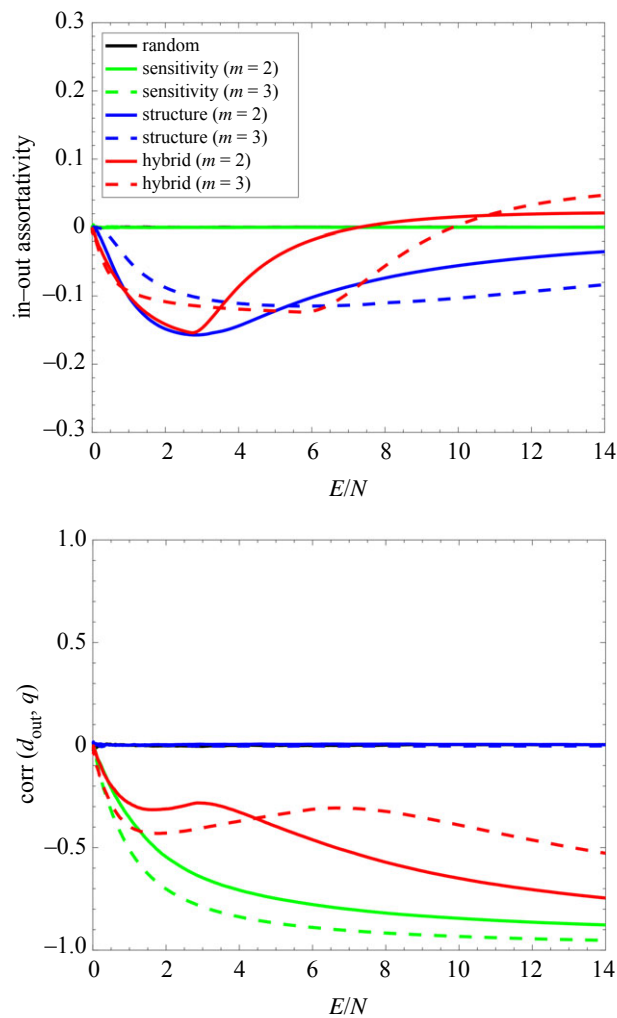


**Figure 4.** The average yield of networks grown under our four selection rules, as described by equation (3.1). Results are averaged over 100 runs on networks of size  $N = 10^5$ . Hybrid and structure-based selection processes show similar yield profiles, consistently outperforming both random and sensitivity-based yields.

where sums are taken over all directed links  $i \rightarrow j$ ,  $d_{in}^i$  ( $d_{out}^j$ ) is the in- (out-) degree of node  $i$  ( $j$ ), and  $\bar{d}_{in}^s$  ( $\bar{d}_{out}^t$ ) is the average in- (out-) degree of all source (target) nodes averaged over links. The in-out degree assortativity is equivalent to the Pearson correlation coefficient between in-degree of source nodes and out-degree of target nodes, with all averages taken over links. Very positive assortativity indicates a preference of large degree nodes to connect to other large degree nodes, while very negative assortativity, also called high disassortativity, indicates large degree nodes tend to connect to small degree nodes. When the assortativity is near zero (no preference), the network is said to be assortatively neutral.

The second relationship we consider is the correlation between node out-degree and node sensitivity, calculated by the Pearson correlation coefficient between out-degree and sensitivity averaged only over active nodes. Due to the symmetry of our selection rules, the correlation between out-degree and sensitivity is equal to the correlation between in-degree and sensitivity. We will refer to either case as the degree-sensitivity correlation, since they are interchangeable. Both the in-out assortativity and degree-sensitivity correlation are shown in figure 5.

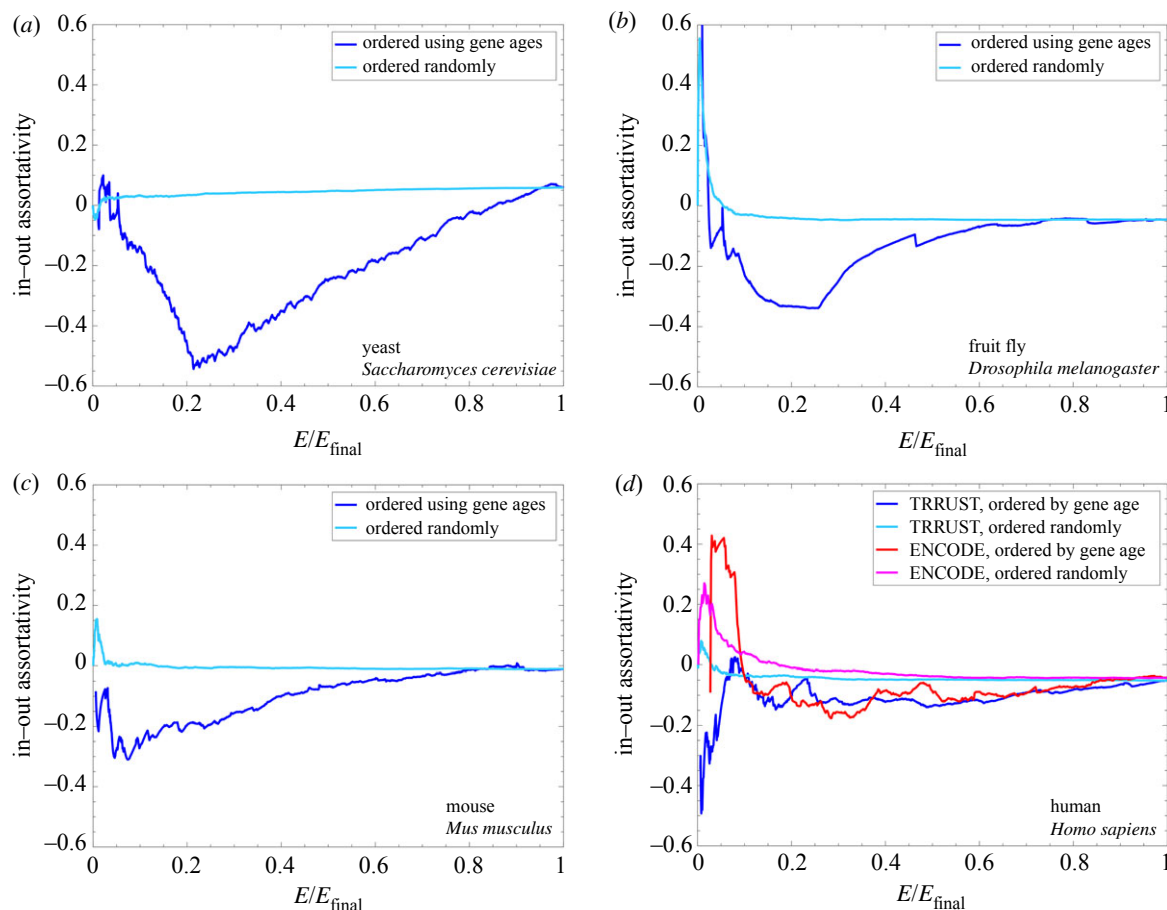
As one should expect, both random selection and sensitivity-based selection are assortatively neutral, since they do not directly consider component sizes or node degree in the decision process. Structure-based selection and hybrid selection are highly disassortative processes, since both will tend to prevent large in-components from connecting to large out-components. Nodes with large in-components typically have large in-degree and similarly with out-components and out-degree. In both of these rules, the assortativity continues to decrease until the critical point at which GSCC percolates, after which the assortativity tends toward zero. Once the giant components are large enough, there is a high probability to select genes from the GSCC. If all  $m$  genes are from the GSCC, then they will each have the same component sizes, at which point a link will be chosen either randomly, in the case of structure-based selection, or based on sensitivity, in the case of hybrid selection. Either approach will tend towards neutral assortativity, since the decisions are independent of



**Figure 5.** In-out assortativity (top) and the degree-sensitivity correlation (bottom) for our four selection rules. Results shown are averages over 100 runs on networks of size  $N = 10^5$ . While the random and sensitivity-based processes are assortatively neutral (i.e. have near-zero in-out assortativity), we see that both the structure-based and hybrid selection processes result in disassortative networks before the phase transition, with the former continuing to show weak disassortativity after the transition, while the latter is nearly neutral after the transition. The random and structure-based processes result in no correlation between degree and sensitivity while, as expected, the sensitivity-based selection process creates a negative correlation between a node's out-degree and its sensitivity. The hybrid process shows weaker and non-monotonic negative correlations.

component size, which relates to degree. Since hybrid selection leads to a faster-growing GSCC, its assortativity tends to zero much faster. For the case of  $m = 3$ , the trend is similar, though the assortativities of structure-based and hybrid selection decrease at a slower rate to a higher (less negative) minimum value. Since higher values of  $m$  are better at keeping components small, more components exist near the average component size.

This trend is clarified by considering the degree-sensitivity correlation. First, we note that both random and structure-based selection have uncorrelated degrees and sensitivities, since they ignore sensitivity information. Sensitivity-based selection has a strong negative degree-sensitivity correlation, which is also to be expected since low sensitivity genes have a higher probability of gaining links. The most interesting case is hybrid selection. Initially, when all components are small, sensitivity is a larger influence in the hybrid selection,



**Figure 6.** In–out assortativity for the transcriptional gene regulatory networks listed in table 1, as a function of the number of links added. Links are ordered using gene age information [43], and a randomized ordering of the links in the final network is shown for comparison. Both orderings yield the same final network. *E. coli* is omitted, due to lack of gene age information from [43]. Note that the horizontal axis is scaled by the total number of final links ( $E_{\text{final}}$ ), rather than the number of genes ( $N$ ), to allow for comparison across the different networks.

causing a negative degree–sensitivity correlation. Once components are of moderate size, component size begins to take over until the formation of GSCC. As discussed previously, once GSCC forms there is a greater chance that all  $m$  potential genes will come from GSCC, in which case all  $m$  genes have equal component sizes and sensitivity again takes over.

### 3.5. Connecting with data

We compare our simulated networks to networks reconstructed from data. We consider several reconstructed transcriptional gene regulatory networks for different species, listed in table 1. Using [43], we have estimates on the ages of most genes found in these networks, which allows us to construct an ‘ancestral ordering’ of links. We do this by ordering each gene by age and adding a link once both genes appear. Admittedly, this way of capturing the network through evolutionary time provides only a very coarse view, since it relies on the strong assumption that a regulatory link between two genes appears as soon as the ‘younger’ gene appears. While this assumption is almost certainly violated, we believe it provides a good starting point for connecting with data. For comparison, we also consider a random ordering of the links. In both the ancestral and random orderings, the final networks are identical, with the only difference being the order in which the links were added.

Due to the small size of the networks, which results because the empirical data are incomplete, the connected components

**Table 1.** Number of nodes, edges, and the in–out assortativity for the *E. coli*, yeast, fruit fly, mouse, and human transcriptional networks. These in–out assortativity values are also the final values in figure 6.

species	nodes	edges	in–out assort.
<i>E. coli</i> [38]	201	460	0.0475
<i>S. cerevisiae</i> [39]	409	5068	0.0590
<i>D. melanogaster</i> [40]	329	3407	−0.0447
<i>M. musculus</i> [41]	827	2213	−0.010
<i>H. sapiens</i> (TRRUST) [41]	795	1882	−0.0516
<i>H. sapiens</i> (ENCODE) [42]	122	624	−0.0430

do not have clear transitions. In addition, we do not have known values for the bias or sensitivity of each gene. The sensitivity of a gene may be approximated using expression data, but calculating accurate sensitivities is a challenging problem and beyond the scope of our investigation.

We can, however, analyse the in–out assortativity of the network as a function of the number of links added, as shown in figure 6. The final values of the in–out assortativity are given in table 1. As we observed in our simulated structure-based and hybrid networks, when we add links using the ancestral ordering that incorporates gene age, we see a general trend in in–out assortativity: a significant early

drop to a moderately negative minimum, followed by a slow increase back toward zero. This pattern is clear for three of the four species for which [43] provides extensive gene age data: yeast, drosophila, and mouse. We see a similar, yet weaker, trend for human in figure 6d, but only after an initial peak with elevated assortativity, which is especially pronounced for the ENCODE data. We believe that the difference in the pattern may be due to the small network size (the ENCODE network includes only 122 genes) and the biases involved in subsampling from the full human gene regulatory network. Nonetheless, the difference between the ancestral ordering and the random ordering for the human data is similar to the other species: the assortativity for the ancestral ordering dips far lower than for the random ordering and stays significantly negative for a much larger range of links added. Perhaps interestingly, the two species with positively assortative final networks, yeast and *Escherichia coli*, are also the only single-celled organisms considered. Unfortunately, incomplete gene age data from [43] means we cannot verify if the assortativity of the *E. coli* network follows the same trend as the others.

## 4. Discussion

We have seen that using simple competitive rules, a network can be evolved to resist large-scale damage propagation significantly better than the random case. Often this leads to specialized structure such as higher link density between low sensitivity genes, high disassortativity, and strong negative correlations between degree and sensitivity. Perhaps most surprising is that knowing a gene's sensitivity is not necessary to effectively delay the onset of a system-wide damaged component, although it does help to ensure that the components remain small. In fact, when we include sensitivity but not structure in the selection rule, more links form between a smaller subset of genes. Not only does this feature result in faster growth of giant components, but it also tends to exclude some high sensitivity genes from contributing to the network at all. Further, the in-out assortativity of our

simulated networks evolved using structure-based selection mimics the in-out assortativity of our reconstructed real networks, further corroborating the importance of structure information in the evolution of actual regulatory networks. Future investigations into the connection between this kind of model and real-world data would benefit tremendously from accurate estimates of gene biases.

Surprisingly, using gene sensitivity information along with structure does not provide a benefit, at least when applied through a simple multiplicative rule, when compared to using structure alone. This hybrid rule results in a three-phase growth process that switches between emphasizing low sensitivity genes, then small component genes (which at such a point are high sensitivity), and then back to low sensitivity genes. The drawbacks found in the purely sensitivity-based selection take effect, leading to earlier percolation than if sensitivity were ignored. One can construct more complicated selection rules that incorporate sensitivity and structure to perform better than structure alone, but such rules tend to be overly engineered, and hence we suspect less biological.

One missing aspect of this process is the ability for gene sensitivities to evolve over time. The sensitivity of a gene is an innate dynamical property, and should not be expected to stay constant throughout every generation. In fact, as the structure of the network evolves over time, the dynamical behaviour of individual genes would likely change in response. Allowing for the coevolution of structure and sensitivity may lead to a different story, and represents a natural next step.

**Data accessibility.** All relevant code and data are publicly available at <https://github.com/bralex1/EvoNets>.

**Authors' contributions.** B.A. performed the simulations and analysis. M.G. and B.A. designed the study, with M.G. directing the study. A.P. developed an early prototype of the numerical experiments. B.A. and M.G. wrote the manuscript.

**Competing interests.** We declare we have no competing interests.

**Funding.** This work was supported in part by UMD's COMBINE programme through NSF award DGE-1632976.

## References

- Solé R, Valverde S. 2020 Evolving complexity: how tinkering shapes cells, software and ecological networks. *Phil. Trans. R. Soc. B* **375**, 20190325. (doi:10.1098/rstb.2019.0325)
- Mora T, Bialek W. 2011 Are biological systems poised at criticality? *J. Stat. Phys.* **144**, 268–302. (doi:10.1007/s10955-011-0229-4)
- Daniels BC, Kim H, Moore D, Zhou S, Smith HB, Karas B, Kauffman SA, Walker SI. 2018 Criticality distinguishes the ensemble of biological regulatory networks. *Phys. Rev. Lett.* **121**, 138102. (doi:10.1103/PhysRevLett.121.138102)
- Martínez E, Villani M, La Rocca L, Kauffman SA, Serra R. 2018 Dynamical criticality in gene regulatory networks. *Complexity* **2018**, 5980636. (doi:10.1155/2018/5980636)
- Kauffman S. 1969 Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467. (doi:10.1016/0022-5193(69)90015-0)
- Derrida B, Pomeau Y. 1986 Random networks of automata: a simple annealed approximation. *Europhys. Lett.* **A 1**, 45. (doi:10.1209/0295-5075/1/2/001)
- Aldana M, Cluzel P. 2003 A natural class of robust networks. *Proc. Natl Acad. Sci. USA* **100**, 8710–8714. (doi:10.1073/pnas.1536783100)
- Squires S, Sytzu K, Alcalá D, Antonsen TM, Ott E, Girvan M. 2013 Weakly explosive percolation in directed networks. *Phys. Rev. E* **87**, 052127. (doi:10.1103/PhysRevE.87.052127)
- Kauffman S, Peterson C, Samuelsson B, Troein C. 2004 Genetic networks with canalizing Boolean rules are always stable. *Proc. Natl Acad. Sci. USA* **101**, 17 102–17 107. (doi:10.1073/pnas.0407783101)
- Pomerance A, Ott E, Girvan M, Losert W. 2009 The effect of network topology on the stability of discrete state models of genetic control. *Proc. Natl Acad. Sci. USA* **106**, 8209–8214. (doi:10.1073/pnas.0900142106)
- Pomerance A, Girvan M, Ott E. 2012 Stability of Boolean networks with generalized canalizing rules. *Phys. Rev. E* **85**, 046106. (doi:10.1103/PhysRevE.85.046106)
- Seshadhri C, Smith AM, Vorobeychik Y, Mayo JR, Armstrong RC. 2016 Characterizing short-term stability for Boolean networks over any distribution of transfer functions. *Phys. Rev. E* **94**, 012301. (doi:10.1103/PhysRevE.94.012301)
- Squires S, Pomerance A, Girvan M, Ott E. 2014 Stability of Boolean networks: the joint effects of topology and update rules. *Phys. Rev. E* **90**, 022814. (doi:10.1103/PhysRevE.90.022814)
- Bornholdt S, Rohlf T. 2000 Topological evolution of dynamical networks: global criticality from local dynamics. *Phys. Rev. Lett.* **84**, 6114–6117. (doi:10.1103/PhysRevLett.84.6114)

15. Oikonomou P, Cluzel P. 2006 Effects of topology on network evolution. *Nat. Phys.* **2**, 532–536. (doi:10.1038/nphys359)
16. Aldana M, Balleza E, Kauffman S, Resendiz O. 2007 Robustness and evolvability in genetic regulatory networks. *J. Theor. Biol.* **245**, 433–448. (doi:10.1016/j.jtbi.2006.10.027)
17. Achlioptas D, D'Souza RM, Spencer J. 2009 Explosive percolation in random networks. *Science* **323**, 1453–1455. (doi:10.1126/science.1167782)
18. Grassberger P, Christensen C, Bizhani G, Son S-W, Paczuski M. 2011 Explosive percolation is continuous, but with unusual finite size behavior. *Phys. Rev. Lett.* **106**, 225701. (doi:10.1103/PhysRevLett.106.225701)
19. Riordan O, Warnke L. 2011 Explosive percolation is continuous. *Science* **333**, 322–324. (doi:10.1126/science.1206241)
20. D'Souza RM, Nagler J. 2015 Anomalous critical and supercritical phenomena in explosive percolation. *Nat. Phys.* **11**, 531–538. (doi:10.1038/nphys3378)
21. Viles W, Ginestet CE, Tang A, Kramer MA, Kolaczyk ED. 2016 Percolation under noise: detecting explosive percolation using the second-largest component. *Phys. Rev. E* **93**, 052301. (doi:10.1103/PhysRevE.93.052301)
22. Waagen A, D'Souza RM, Lu T-C. 2017 Explosive percolation on directed networks due to monotonic flow of activity. *Phys. Rev. E* **96**, 012317. (doi:10.1103/PhysRevE.96.012317)
23. D'Souza RM, Gmez-Gardees J, Nagler J, Arenas A. 2019 Explosive phenomena in complex networks. *Adv. Phys.* **68**, 123–223. (doi:10.1080/00018732.2019.1650450)
24. Carlson JM, Doyle J. 1999 Highly optimized tolerance: a mechanism for power laws in designed systems. *Phys. Rev. E* **60**, 1412–1427. (doi:10.1103/PhysRevE.60.1412)
25. Carlson JM, Doyle J. 2000 Highly optimized tolerance: robustness and design in complex systems. *Phys. Rev. Lett.* **84**, 2529–2532. (doi:10.1103/PhysRevLett.84.2529)
26. van Noort V, Snel B, Huynen MA. 2003 Predicting gene function by conserved co-expression. *Trends Genet.* **19**, 238–242. (doi:10.1016/S0168-9525(03)00056-8)
27. Teichmann SA, Babu MM. 2004 Gene regulatory network growth by duplication. *Nat. Genet.* **36**, 492–496. (doi:10.1038/ng1340)
28. Kuo PD, Banzhaf W, Leier A. 2006 Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *BioSystems* **85**, 177–200. (doi:10.1016/j.biosystems.2006.01.004)
29. Manrubia S, Cuesta JA. 2015 Evolution on neutral networks accelerates the ticking rate of the molecular clock. *J. R. Soc. Interface* **12**, 20141010. (doi:10.1098/rsif.2014.1010)
30. Demetrius L, Manke T. 2005 Robustness and network evolution—an entropic principle. *Physica A* **346**, 682–696. (doi:10.1016/j.physa.2004.07.011)
31. Squires S, Ott E, Girvan M. 2012 Dynamical instability in Boolean networks as a percolation problem. *Phys. Rev. Lett.* **109**, 085701. (doi:10.1103/PhysRevLett.109.085701)
32. Demongeot J, Elena A, Sené S. 2008 Robustness in regulatory networks: a multi-disciplinary approach. *Acta Biotheor.* **56**, 27–49. (doi:10.1007/s10441-008-9029-x)
33. Melnik S, Hackett A, Porter MA, Mucha PJ, Gleeson JP. 2011 The unreasonable effectiveness of tree-based theory for networks with clustering. *Phys. Rev. E* **83**, 036112. (doi:10.1103/PhysRevE.83.036112)
34. Chandra S, Ott E, Girvan M. 2020 Critical network cascades with re-excitable nodes: why treelike approximations usually work, when they break down, and how to correct them. *Phys. Rev. E* **101**, 062304. (doi:10.1103/PhysRevE.101.062304)
35. Newman MEJ, Strogatz SH, Watts DJ. 2001 Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118. (doi:10.1103/PhysRevE.64.026118)
36. Erdős P, Rényi A. 1960 On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61.
37. da Costa RA, Dorogovtsev SN, Goltsev AV, Mendes JFF. 2010 Explosive percolation transition is actually continuous. *Phys. Rev. Lett.* **105**, 255701. (doi:10.1103/PhysRevLett.105.255701)
38. Santos-Zavaleta A *et al.* 2018 RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Res.* **47**, D212–D220. (doi:10.1093/nar/gky1077)
39. Chen C, Zhang D, Hazbun TR, Zhang M. 2019 Inferring gene regulatory networks from a population of yeast segregants. *Sci. Rep.* **9**, 1197. (doi:10.1038/s41598-018-37667-4)
40. Rhee D *et al.* 2014 Transcription factor networks in *drosophila melanogaster*. *Cell Rep.* **8**, 2031–2043. (doi:10.1016/j.celrep.2014.08.038)
41. Han H *et al.* 2017 TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386. (doi:10.1093/nar/gkx1013)
42. Gerstein MB *et al.* 2012 Architecture of the human regulatory network derived from encode data. *Nature* **489**, 91–100. (doi:10.1038/nature11245)
43. Liebeskind BJ, McWhite CD, Marcotte EM. 2016 Towards consensus gene ages. *Genome Biol. Evol.* **8**, 1812–1823. (doi:10.1093/gbe/evw113)