Elucidating the Role of Network Structure in Gene Regulation: Connecting Mathematical Models and Empirical Data

Michelle Girvan

DEPARTMENT OF
PHYSICS
UNIVERSITY OF MARYLAND

Sponsors:

Collaborators:

- Shane Squires, UMD Physics PhD, 2014
- Andrew Pomerance, UMD Physics PhD 2010
- Ed Ott, UMD, Prof of Physics
- Sanya Pushkar, UMD Applied Math grad student
- John Platig, UMD Physics PhD 2013
- Wolfgang Losert, UMD, Prof of Physics
- Lou Staudt, National Cancer Institute
- Dan Larson, National Cancer Institute
- Hector Bravo, UMD, Prof of Computer Science

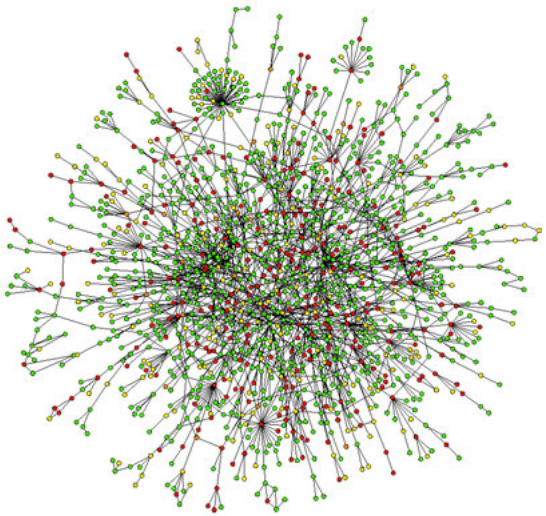
Overview

- **The goal:** To gain insights into the complex process of gene regulation and make meaningful connections between biological data and mathematical models.
- **The approach:** Starting from simple mathematical models of genetic control, we explore the joint effects of network topology and interaction rules
- **Application #1:** We hypothesize that a dynamical instability in the gene network may be a causal mechanism contributing to the occurrence of some cancers.
- **Application #2:** These models may help to explain how the network structures observed in empirical data reflects tradeoffs between diversity of function and system robustness.

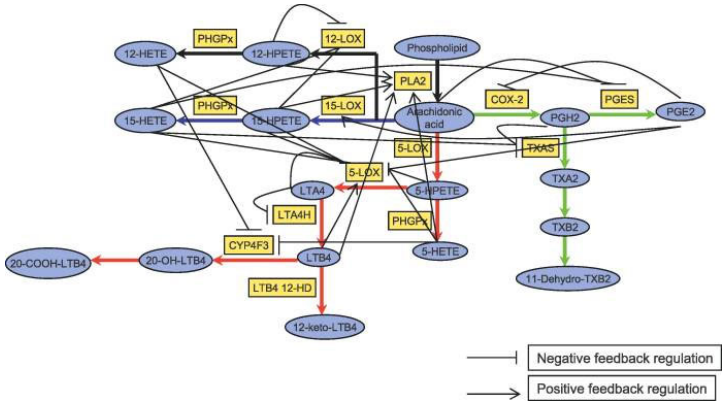
Why think about gene *networks*?

- The pattern of interactions between genes (i.e. network effects) can play a significant and complex role in gene regulation
- The network approach helps us to identify groups of genes that are working in concert to produce undesirable outcomes

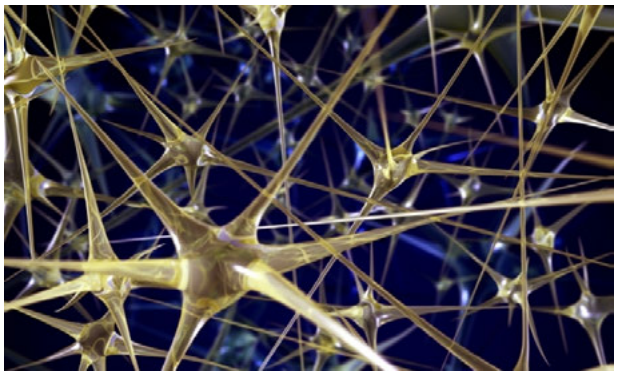
Network Science and Systems Biology



Protein Interaction Network



Metabolic Network

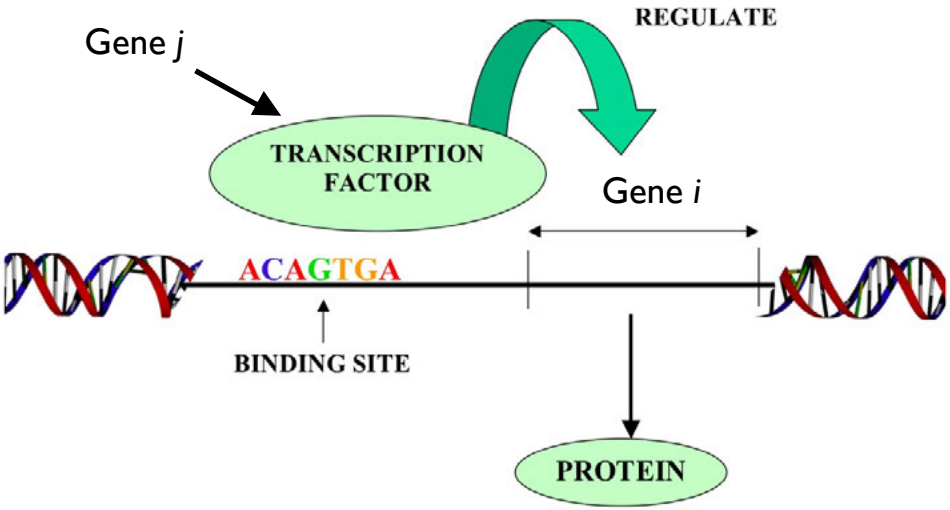


Neural Network

- Network methods are most useful when links represent real interactions and not merely correlations
- Network connections can tell us how information flows in a system
- Network centrality measured can be used to identify drivers of activity and develop optimized attack/intervention strategies
- Network community detection methods can be used to identify functional related groups of nodes

A Network Approach to Gene Regulation

From individual regulatory relationships between genes...



Interpreting link *ij*

...to a complex web of interactions

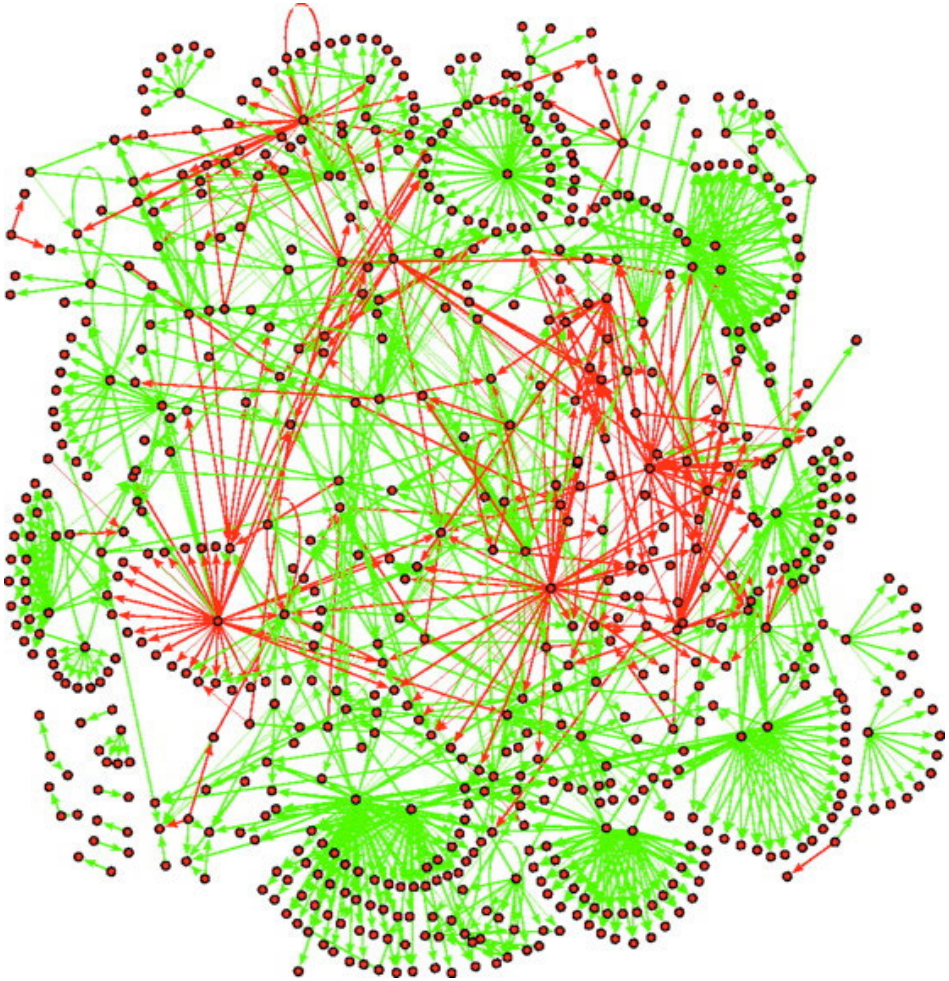
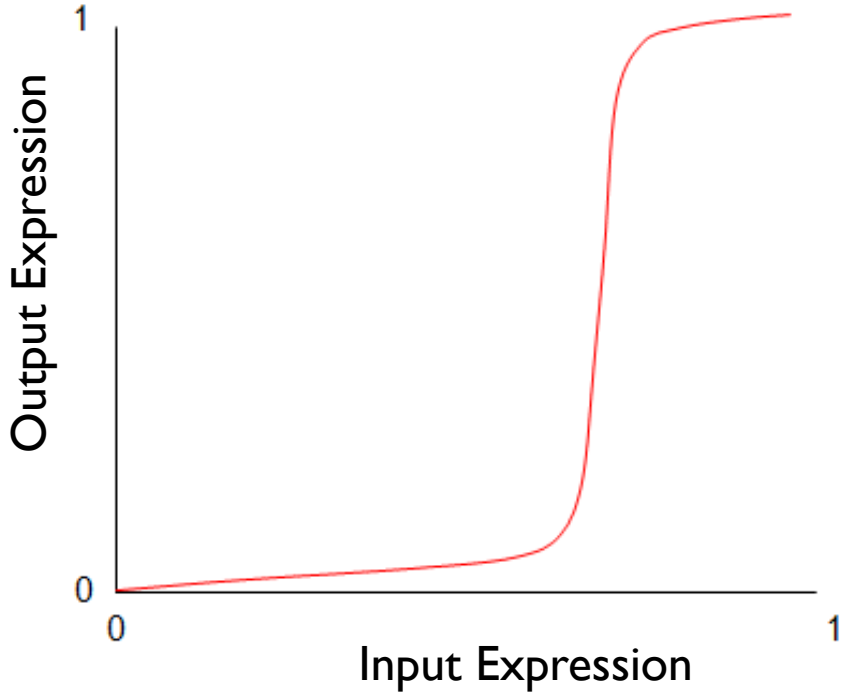


Figure taken from
http://rsif.royalsocietypublishing.org/content/5/Suppl_1/S85.full

Connecting Models with Data: Goals and Challenges

- Goal: Use mathematical models to understand the role of network structure in gene regulation ***and*** to make connections to empirical data
- Challenge #1: Identifying the network structure from data
- Challenge #2: Identifying the dynamical parameters of the model from data
- Challenge #3: Appropriately accounting for inference errors when developing intervention strategies

Building a simple model for gene regulation: Why Boolean?



Motivation: Input/output regulatory relationships between genes are often observed to be strongly sigmoidal and well approximated by step functions.

Caution: When the expression levels of multiple inputs are varied, we often see more than two output expression levels

Modeling Genetic Control with Boolean Networks

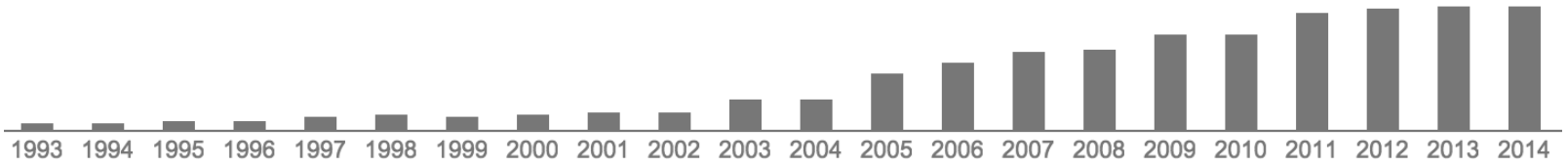
Kauffman's *N-K* model:

- *N* Genes on or off
- Each gene has exactly *K* inputs, which are randomly chosen
- Discrete updates
- Evolves by a random update function at each node

Our work:

- Focuses on stability in response to small perturbations
- Explores the effect of network topology on stability
- Also explores more biologically realistic Boolean update rules

Citations to Kauffman's original paper:

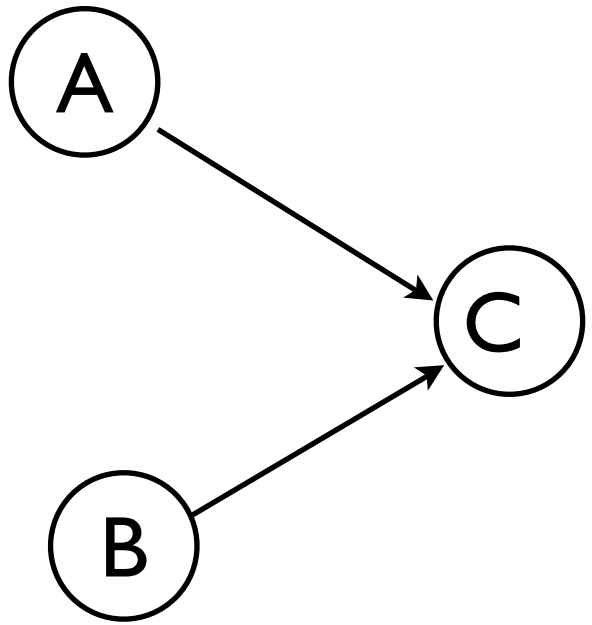


"Metabolic stability and epigenesis in randomly constructed genetic nets," JTB 1969

→ Dramatic increase in citations driven by advances in high throughput biological data collection and the growth of network science.

Local update rules: An example

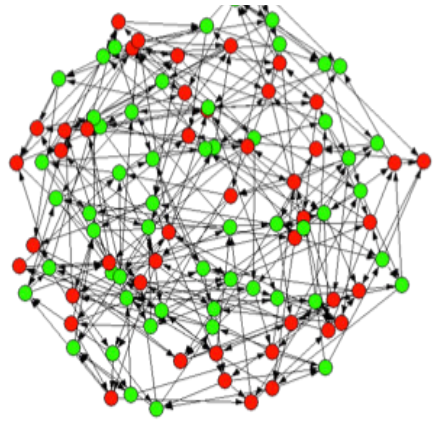
current state time t		State of gene C at t+1
Gene A	Gene B	
0	0	0
0	1	0
1	0	1
1	1	0

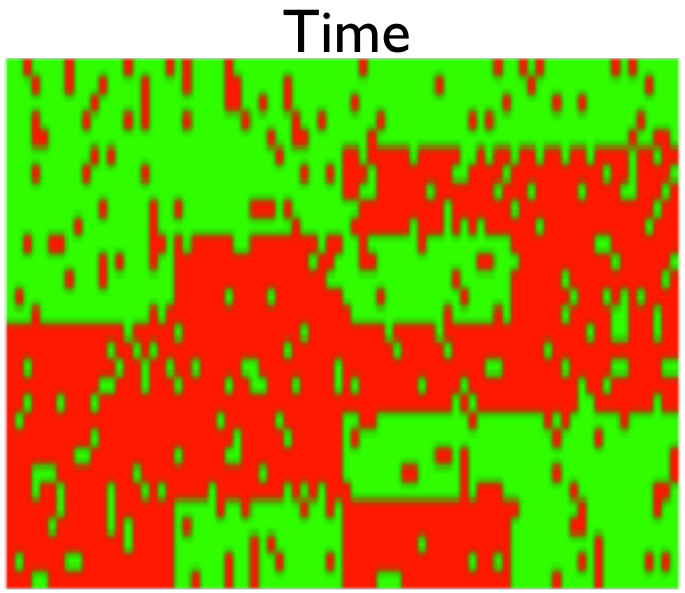


Node with 2 inputs

Random update rules: Output column filled in randomly at the start of the simulation with bias (probability of 0), p, and fixed forever after.

Local Rules Lead to Global Patterns

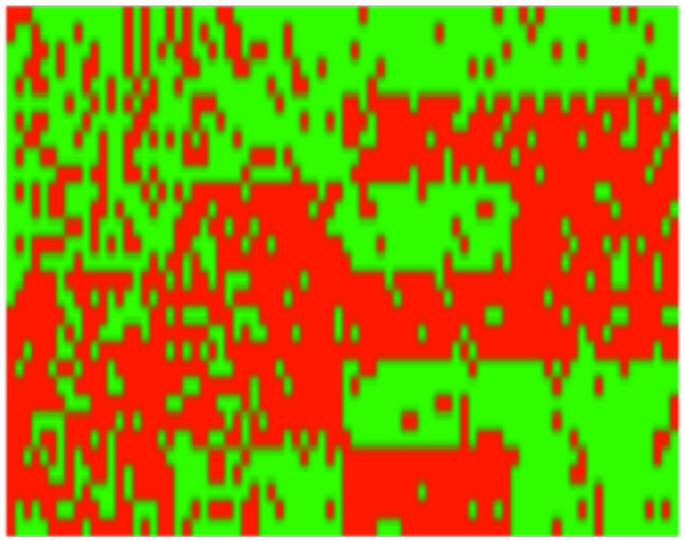




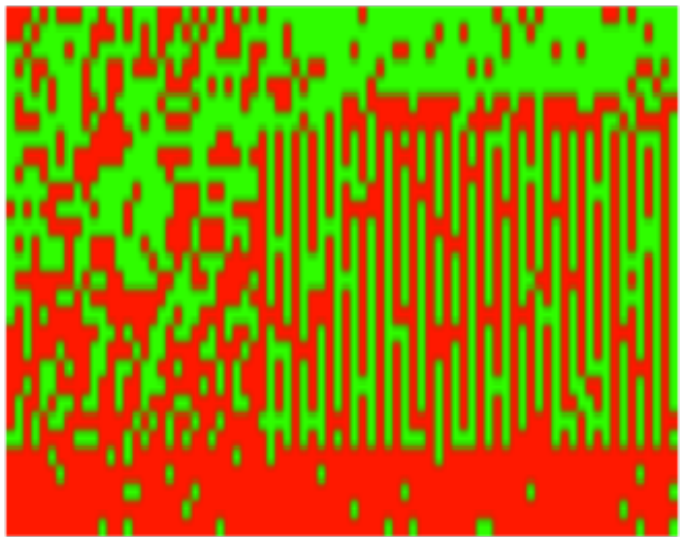
original pattern

Is the network stable or chaotic?

Flip the states of a few genes.
Do we see the same pattern as before?



pattern in stable network

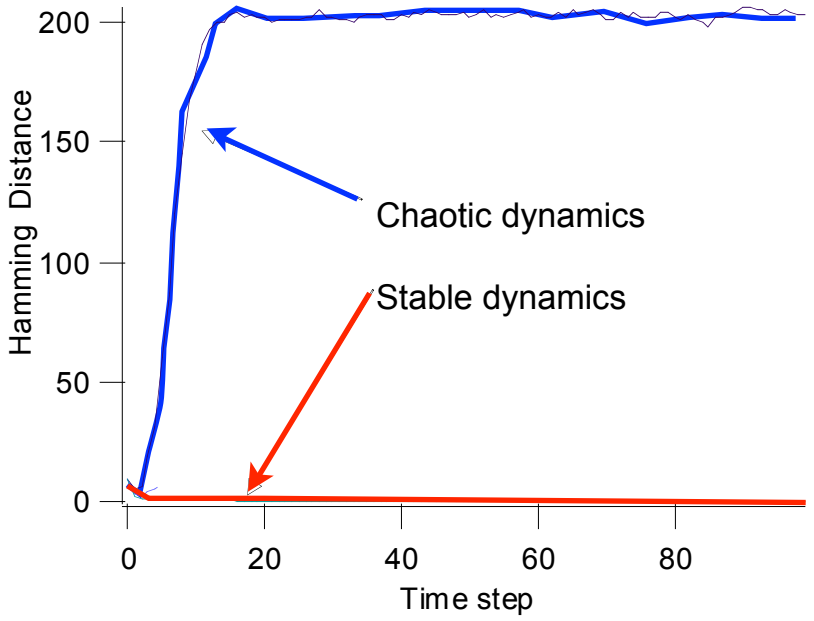


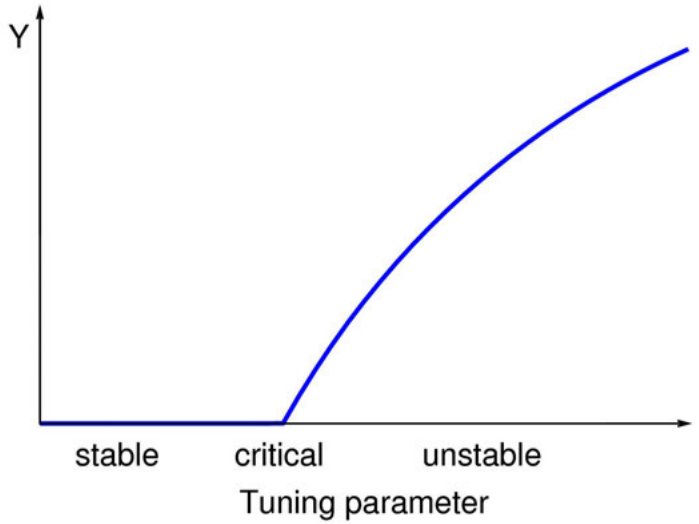
pattern in chaotic network

Illustrative examples (not from actual simulations)

System Dynamics

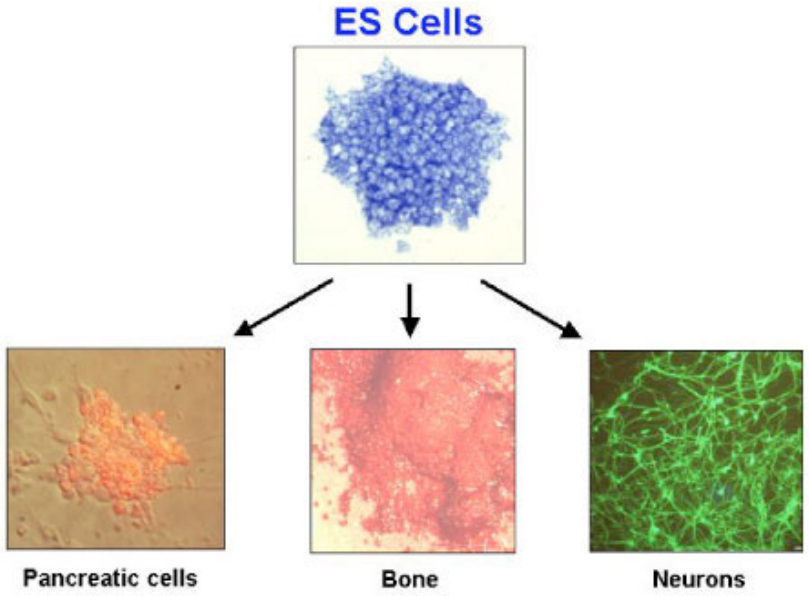
- Chaotic dynamics: The system can exhibit what are termed "chaotic" dynamics in which the distance between initially close states grows in time until some saturation level is reached
- As the system is tuned by varying either K (number of inputs) or *p* (the update function bias), we see a second order phase transition in Y, the normalized average saturated Hamming distance.
- Kauffman hypothesized that cells exist at the critical point between stable and unstable regimes.





Significance of the patterns

- The patterns of activity may define a cell's character
- In single celled organisms this could correspond to different cell states: growing, dividing, starving, etc.
- In multicellular organisms these could correspond to different cell types.



Motivation for our work

- Early results for these systems relied on annealed networks and annealed truth tables (network connections and truth tables randomized at each time step) to find the transition between stability and instability. (Derrida & Pomeau, 1986; Aldana & Cluzel, 2003)
- Since real networks are far from the idealized models studied previously, the aim of our initial work was to be able to handle almost any specified network topology for the case of random update functions.
- Our more recent work is focused on the joint effects of network topology and update rules and considers more biologically realistic gene interactions
- We also have extended our analysis to the case of non-synchronous update.

References

A. Pomerance, E. Ott, M. Girvan, and W. Losert, "The effect of network topology on the stability of discrete state models of genetic control," Proc. Natl. Acad. Sci. USA 106, 8209-8214 (2009).

A. Pomerance, M. Girvan, and E. Ott, "Stability of Boolean networks with generalized canalizing rules," Phys. Rev. E 85, 046106 (2012).

S. Squires, E. Ott, and M. Girvan, "Dynamical instability in Boolean networks as a percolation problem," Phys. Rev. Lett. 109, 085701 (2012).

S. Squires, E. Ott, and M. Girvan, "Stability of Boolean networks: The joint effects of topology and update rules," Phys. Rev. E 90, 049905 (2014).

Describing the Boolean Network Mathematically

- Network topology:

$$A_{ij} = \begin{cases} 1 \text{ if link from } j \rightarrow i \\ 0 \text{ otherwise} \end{cases}$$

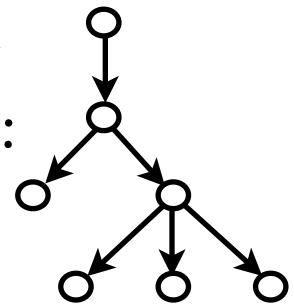
- Random update functions (a simple model of genetic control):
  - Output column randomly filled in
  - Bias p - probability of a 0 appearing in the output column

A starting point: Handling arbitrary network topologies and sensitivity distributions

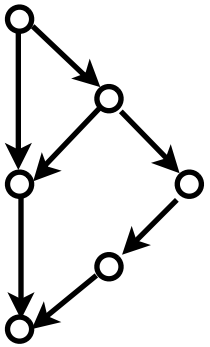
(for otherwise random update functions)

- We consider a *semi-annealed approximation* in which the *network is fixed* and the output entries of the *truth tables are randomized* at every step, subject to a bias p_i that depends on i.
- We perform *numerical tests with frozen truth tables* to test the applicability of semi-annealing
- *Locally "asynchronous" requirement*: we can handle any network structure in which pairs of nodes are rarely connected by multiple short paths of the same length (weaker version of locally tree-like)

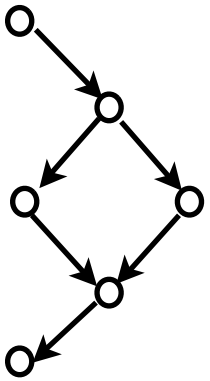
Locally treelike:



Locally asynchronous:



Not locally asynchronous:



Semi-annealed analysis

- Consider two state vectors, $\underline{\sigma}(t)$ and $\underline{\tilde{\sigma}}(t)$, that have evolved from slightly different initial conditions
- Let $y_i(t) =$ the probability that $\sigma_i(t)$ and $\tilde{\sigma}_i(t)$ differ
- Let $q_i =$ the probability that $\sigma_i(t)$ and $\tilde{\sigma}_i(t)$ differ, given a difference in the states of the inputs to i at time $t-1$

$$q_i = 1 - [p_i^2 + (1-p_i)^2] = 2p_i(1-p_i)$$

- Using semi-annealing we can write an update equation for $y_i(t)$

Probability that the inputs at t-1 to i are not all the same

$$y_i(t) = q_i \left\{ 1 - \prod_{j, A_{ij}=1} \left[1 - y_j(t-1) \right] \right\}$$

Probability that the input from node j is the same

- Perturb around $\underline{\sigma} = \underline{\tilde{\sigma}}$ $(y_i \ll 1)$, linearization gives:

$$\underline{y}(t) = \underline{\underline{Q}}\,\underline{y}(t-1)$$

Stability Conditions:

If $\lambda_Q < 1$: stable
If $\lambda_Q > 1$: unstable
If $\lambda_Q = 1$: "edge of chaos"

where the $Q_{ij} = q_i A_{ij}$ are the elements of a modified adjacency matrix

λ_Q is the largest eigenvalue of Q, which, according to the Perron-Frobenius theorem is real and positive ($Q_{ij} \geq 0$).

Numerical tests

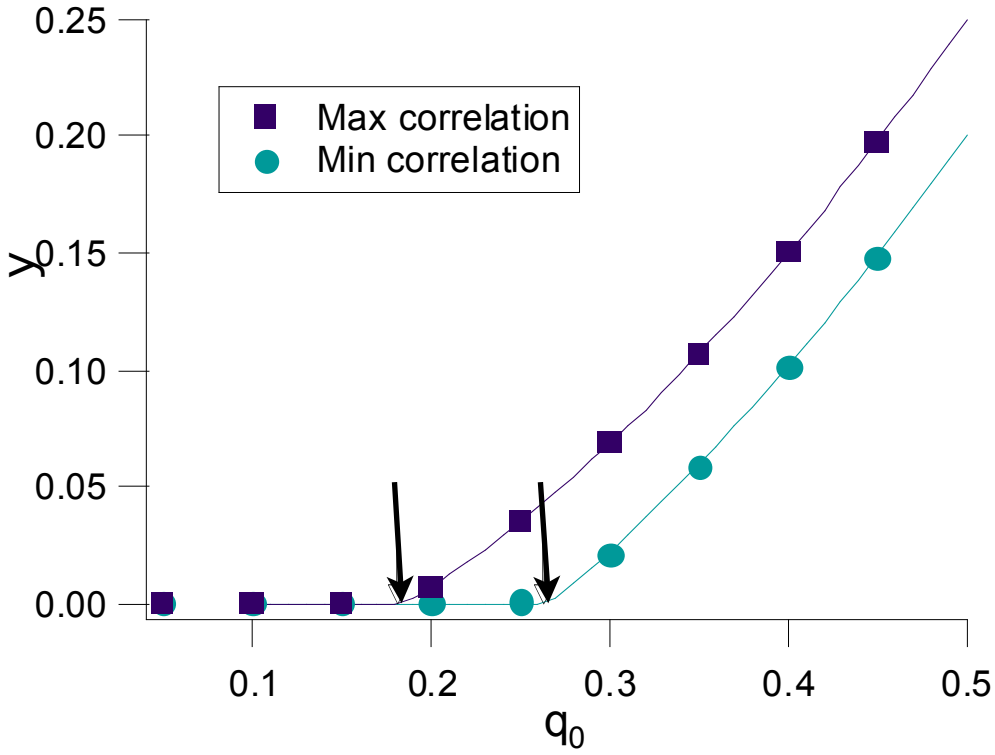
We numerically test the predictions of

- λ_Q stability criterion
- Saturated normalized Hamming distance between $\underline{\sigma}$ and $\underline{\tilde{\sigma}}$:

$$\bar{y} = \lim_{t \to \infty} \frac{1}{N} \sum_i y_i(t)$$

Sensitivity-Degree correlations

Nodes have sensitivity drawn from distribution centered around q_0



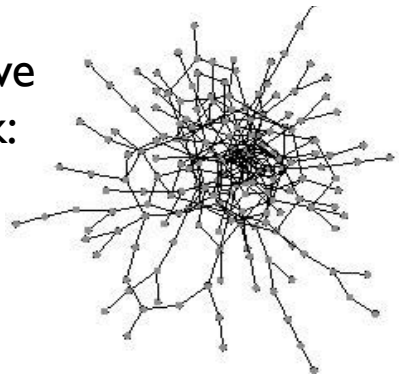
$$q_i = 2p_i(1 - p_i)$$

y is the average saturated distance between two initially close states that have been evolved

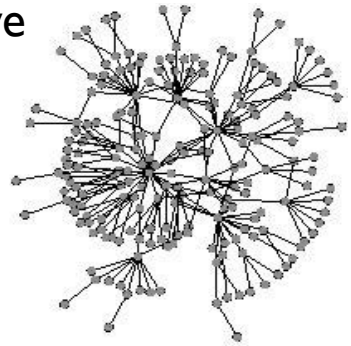
An example of how network topology can affect stability: Assortativity

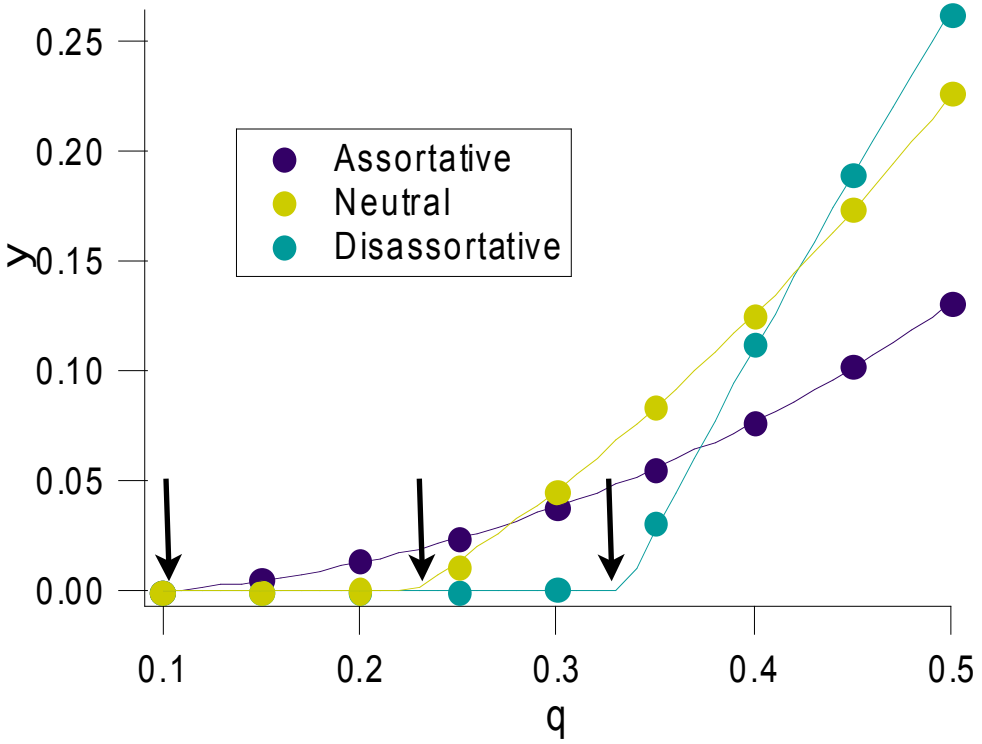
Assortativity: highly connected nodes tend to connect preferentially to other highly connected nodes, tends to increase eigenvalue

assortative network:



disassortative network:





Markers are reflect results from frozen truth tables. Arrows indicate the location of the analytically derived stability transition. Solid lines calculated by iterating the non-linear semi-annealed update equations.

The problem with random Boolean functions

- In reality, regulatory links can generally be classified as either activating (the input gene being on increases the probability that the target gene is on) or repressing (the input gene being on increases the probability that the target gene is off)
- Random Boolean functions do not reflect this biological feature.

Sample update function:

Inputs at t-1			Output
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	0

The threshold Boolean model

The state of node *i* at time *t* depends on the states of its inputs at *t-1* in the following way:

$$\sigma_i(t) = U\left(\sum_{j=1}^{N} w_{ij}\sigma_j(t-1) - \theta_i\right)$$

- *U* is the unit step function
- w_{ij} is the weight of the link from node *j* to *i*
- ϑ_i is the threshold of node *i*

Extending the semi-annealed approach to more biologically realistic update functions

- In approximating the random Boolean model, we focused on the $y_i(t)$ (the probability that the state of node *i* differs at time *t* in two initially close state vectors):

  $y(t) \approx Q y(t-1)$, where $Q_{ij} = q_i A_{ij}$

- Consider the more general scenario:

  $y(t) \approx R y(t-1)$

  where R_{ij} represents the probability that node *i* changes its state given a change in node *j*'s state, considering all other inputs as random.

- Constructing the appropriate matrix *R* whose largest eigenvalue determines the stability of the update functions of interest involves two important steps
  - By iterating a set of self-consistency equations, we find the *dynamical biases* of the nodes which reflect the fraction of time each node spends in the 0 state
  - Starting from the update rule of interest, each node must is assigned an appropriate update rule ensemble for semi-annealing.

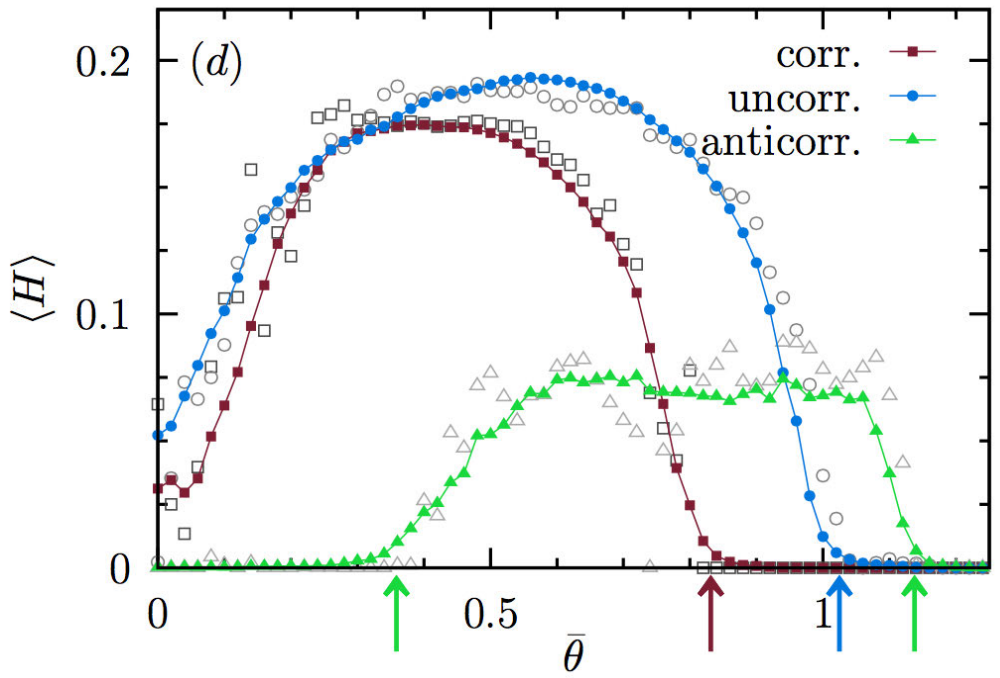
Some Results for Threshold Networks

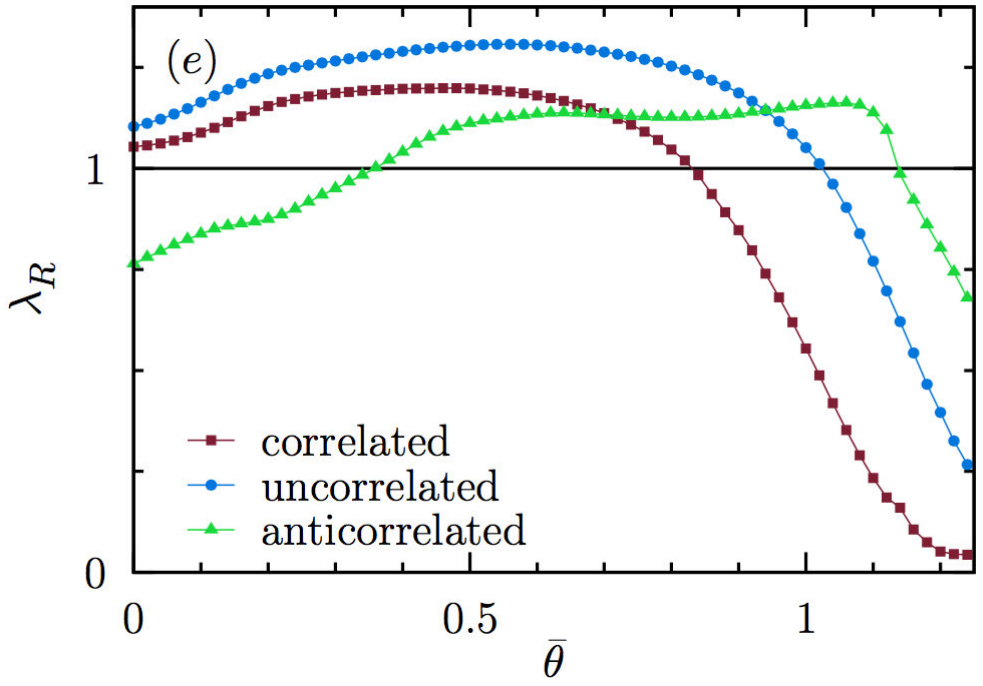
$$\sigma_i(t) = U\left(\sum_{j=1}^{N} w_{ij}\sigma_j(t-1) - \theta_i\right)$$

Network Construction: Scale free out- and poisson in- degree dist. Weights drawn from N(±1,1/4). Correlated case: positive correlation between incoming weights and degree product (in×out) of nodes. Fixed thresholds θ_i from normal dist. with σ_θ.

Semi-annealing: At each time step, the thresholds are randomly drawn from a normal dist. with specified mean and σ_θ.

Results: Solid markers represent averages over 50 realizations of frozen truth tables. Open markers reflect a single realization. Arrows indicate the transition locations from our semi-annealed approximation.





Connecting with data:
Stability and Cancer

- Data from tumor dissections show that nearby cells have vastly different gene expression profiles.
- Could these fluctuations imply a breakdown of genetic control due to dynamical instability?
- What additional data do we need to answer these questions?

Evidence for dynamical instability in cancer

Taken from:

Bravo, Pihur, McCall, Irizarry, Leek, "Gene expression anti-profiles as a basis for accurate universal cancer signatures," BMC Bioinformatics (2012).

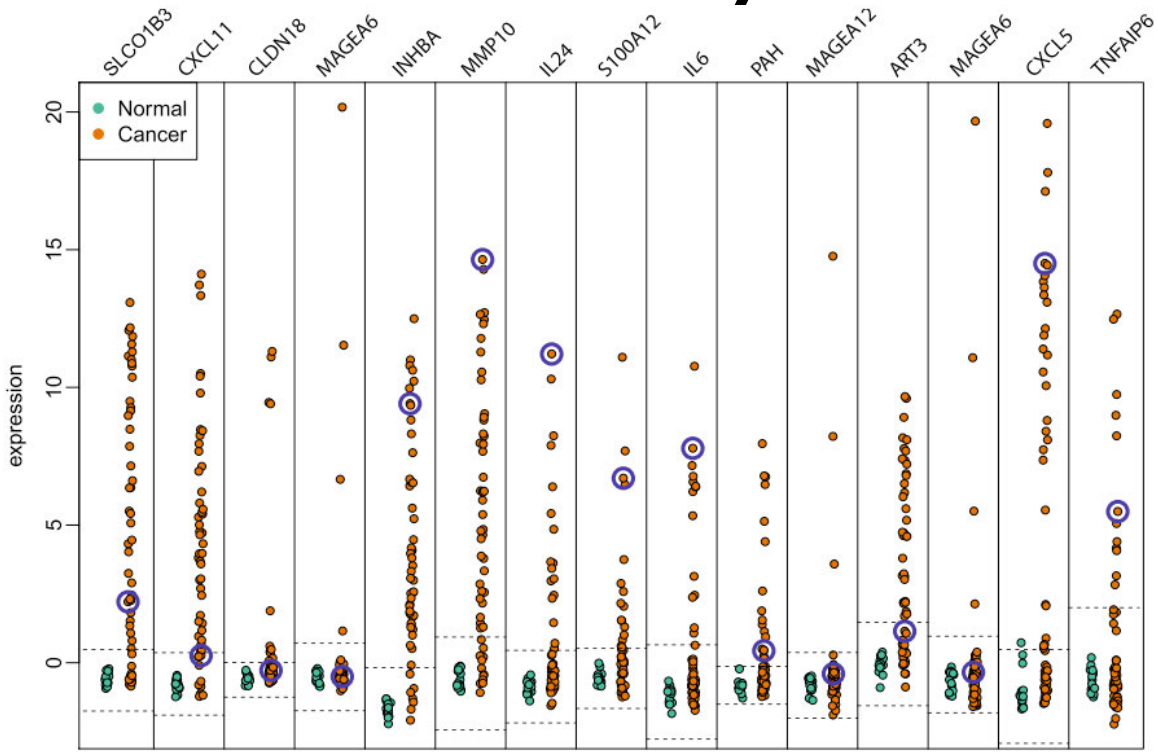


Figure shows normalized gene expression for 15 hyper-variable genes in cancer from two independent colon cancer datasets. Normal samples are shown in green, cancer samples are shown in orange. The anti-profile as the set of genes and a corresponding range of normal expression values for each gene (indicated by dotted lines). The anti-profile score for each sample is the number of genes in the signature that are outside their defined range of expression. Blue circles highlight expression for one specific cancer sample.

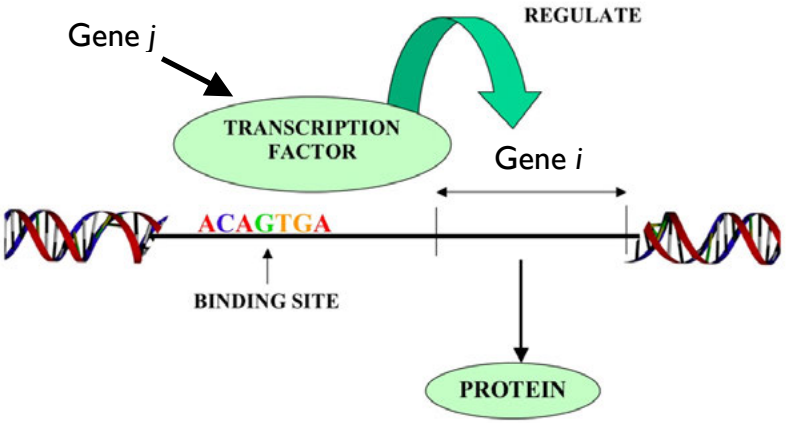
→ A possible analogy to Tolstoy's famous opening to Anna Karenina?

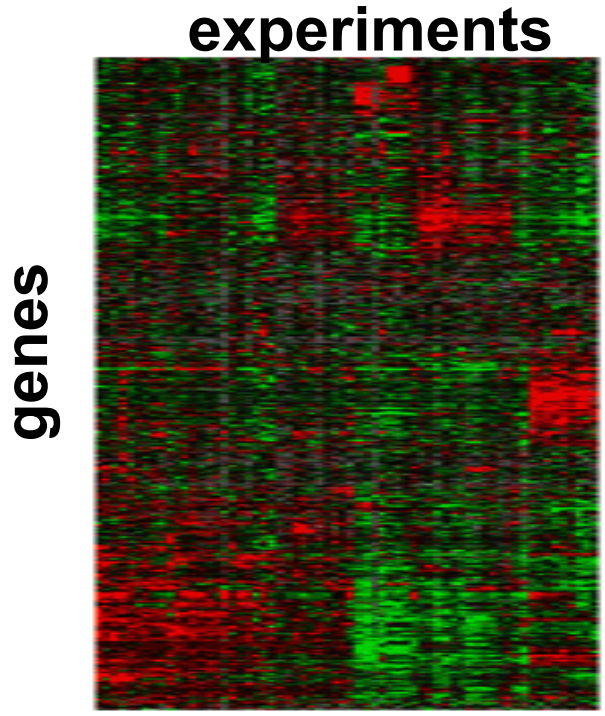
All happy families are alike; each unhappy family is unhappy in its own way.

All healthy cells are alike; each unhealthy cell is unhealthy in its own way.

Elucidating the network from data

- **Network:** Weighted links can be inferred from data by combining different types of data, e.g.:
  - Sequence motif data to indicate the potential for a transcription factor to regulate a specific gene.
  - Gene expression data to determine the weight and sign of the interactions.
- Network links are easier to infer than dynamical parameters
- Caveat: networks and dynamical parameters may be harder to infer in the presence of hypervariability.



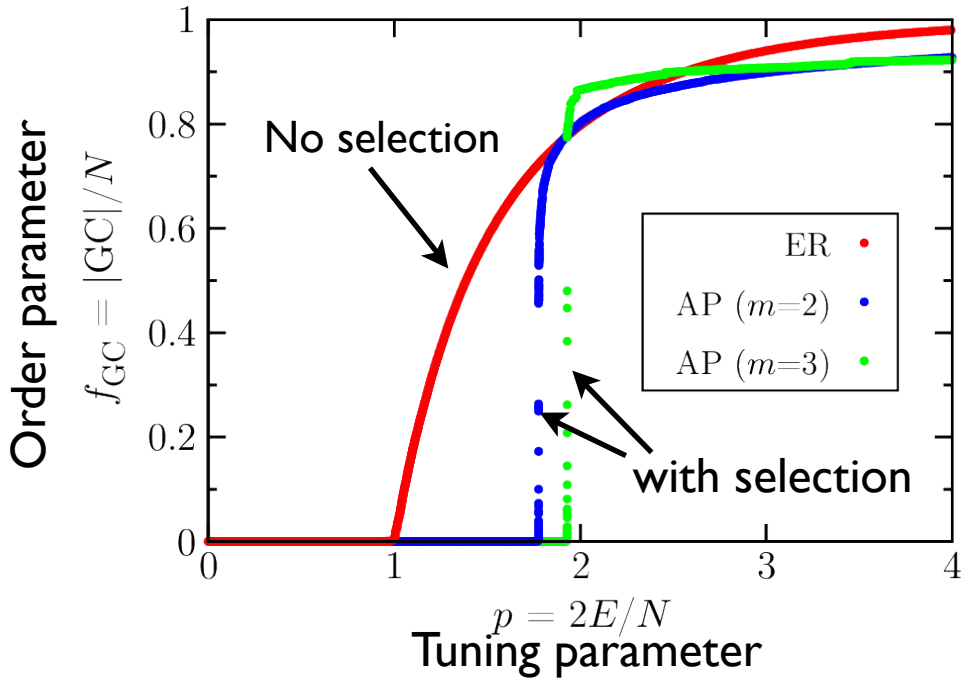


Limitations of the attractor hypothesis

- The attractor hypothesis assumes that different cell types represent different attractors of a *single* underlying network.
- A more common view is to think of different cell types as having distinct regulatory networks with differences due to epigenetic changes.
- Using features from both views, we might imagine that epigenetic changes associated with a given cell type serve to enlarge the basin of attraction of a specific attractor of a pluripotent underlying network. From this perspective, the pluripotent network would not have to be stable, since it doesn't correspond to a realized biological condition.

Modeling the Evolution of Gene Regulatory Networks

- Our modeling framework can be used to explore the evolution of gene regulatory networks
- Q: What kind of network structures optimize fitness?
- Q: Can we represent fitness as a tradeoff between functional diversity and system robustness?
- Connecting we data: if we assume this form of fitness, do the network structures evolved in our model reflect those in data?
- Preliminary work: Models of edge competition in directed networks provide a picture of how selection might shape the structure of gene networks over evolutionary time.



S Squires, K Sytwu, D Alcala, TM Antonsen, E Ott, and M Girvan, "Weakly explosive percolation in directed networks," Physical Review E 87 (5), 052127 (2013).

Summary

- Simple Boolean models of genetic control, starting with random Boolean models and progressing to the more realistic threshold Boolean models, can be used to gain insights into the effects of network structure in the process of gene regulation.
- A major challenge is to connect the model predictions with real data in meaningful ways.
- Future directions: This kind of modeling approach may also be useful for studying the evolution of gene regulatory networks.

References

A. Pomerance†, E. Ott*, M. Girvan*, and W. Losert*, "The effect of network topology on the stability of discrete state models of genetic control," Proc. Natl. Acad. Sci. USA 106, 8209-8214 (2009).

A. Pomerance*, M. Girvan*, and E. Ott*, "Stability of Boolean networks with generalized canalizing rules," Phys. Rev. E 85, 046106 (2012).

S. Squires, E. Ott, and M. Girvan, "Dynamical instability in Boolean networks as a percolation problem," Phys. Rev. Lett. 109, 085701 (2012)

S. Squires, E. Ott, and M. Girvan, "Stability of Boolean networks: The joint effects of topology and update rules," Phys. Rev. E 90, 049905 (2014).