From topology to dynamics in generic content based networks

Ayşe Erzan
Istanbul Technical University, Feza Gürsey Institute

Duygu Balcan (Istanbul Technical University)
Muhittin Mungan (Bogazici University)
Alkan Kabakçıoğlu (Koç University, Padova U.)

Physics of Risk, Vilnius 2006

"Content based" networks

- long linear codes **spontaneously** generate a **network of self-interactions** between subsequences via sequence matching
- lock-and key type mechanisms
  - **Gene regulatory networks**
  - **Protein homology networks**
  - **Immune networks**
- distribution of amount of shared information

degree of specificity of connection $\Rightarrow$ **topology**

outline

- A **sequence matching model** of content based networks
  - Scaling relations
- A sequence matching analogue for the dynamics of **gene regulatory networks,** with **Random Boolean Functions**
- **modelling the GRN of yeast - comparison with data**

sequence matching ⇒

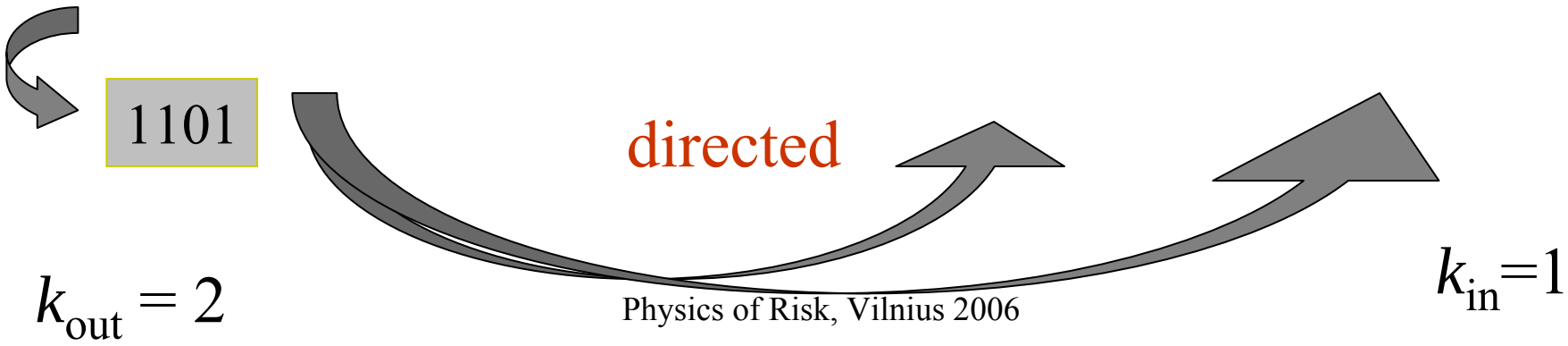
model connectivity matrix

D. Balcan and AE, Eur. Phys. J. B (2004)

$$w_{ij} = \begin{cases} 1 & \text{iff the string } G_i \text{ is embedded inside the string } G_j \\ & (G_i \subset G_j)\,;\ l_i \le l_j \\ 0 & \text{otherwise.} \end{cases}$$

alphabet of length r, string information content $l \ln r$

2 1101 2011000101201000110211 1101 201010 1101 112



Analytical solution - Mungan,Kabakcioglu,Balcan,Erzan, *J. Phys. A* **38**, 9599 (2005)

matching probability

$$p(l, l'; \beta) = 1 - (1 - z^{l})^{l' - l + 1}$$

$$z = \frac{1}{r}[1 + (r-1)e^{-\beta}] \to \frac{1}{r} \quad \text{for } \beta \to \infty$$

then

$$p(l,l';\beta) \to p(l,l') = 1 - \left(1 - \frac{1}{r^{l}}\right)^{l'-l+1}$$

for $l \gg 1$,

$$p(l,l') = \frac{l' - l + 1}{r^{l}}$$

Emergent networks from linear codes: superposition of Erdös-Renyi networks -with an assortment of nodes of length l with connection probabilities $\sim r^{-l}$

Power law out-degree distribution $k^{-\gamma}$

for e**xponential** length distribution of the subsequences

$$n(l) \sim q^{l} \qquad q = 1\text{-}p$$

$$d_l \approx (qz)^l$$

$$\gamma_2 = \frac{1}{2}\frac{\ln z - \ln q}{\ln z + \ln q} = \frac{1}{2}\frac{\ln r + \ln q}{\ln r - \ln q} \approx \frac{1}{2} - \frac{p}{\ln r}$$

$$\gamma_1 = \frac{1}{2} + \gamma_2 \approx 1 - \frac{p}{\ln r}$$

Simulation results: out degree distribution

crossover in the scaling behaviour

Exponential string length distribution

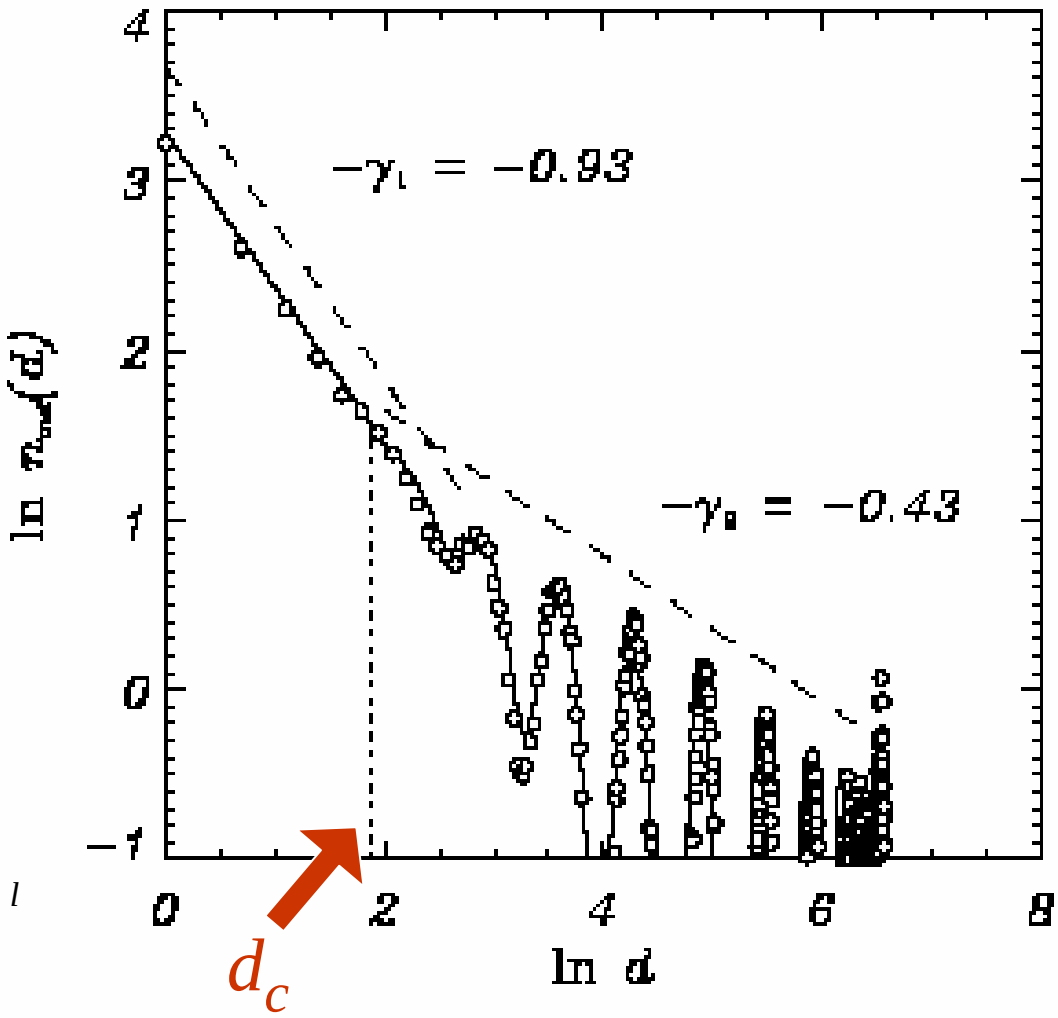
___ analytical

° simulation

L =15000

p = 0.05

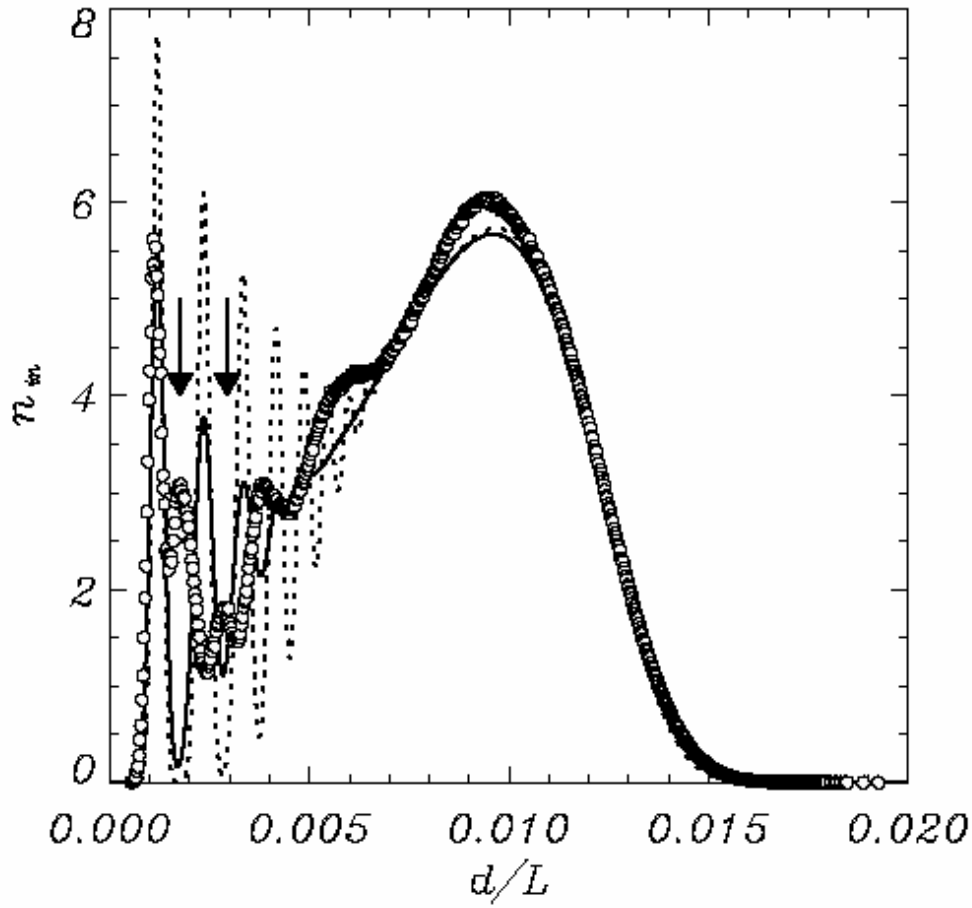
500 realisations



Peak separation $\Delta d_l \sim (qz)^{l}$

Variances $\sim (\Delta d_l)^{1/2}$

Simulation and analytical results: in-degree distribution



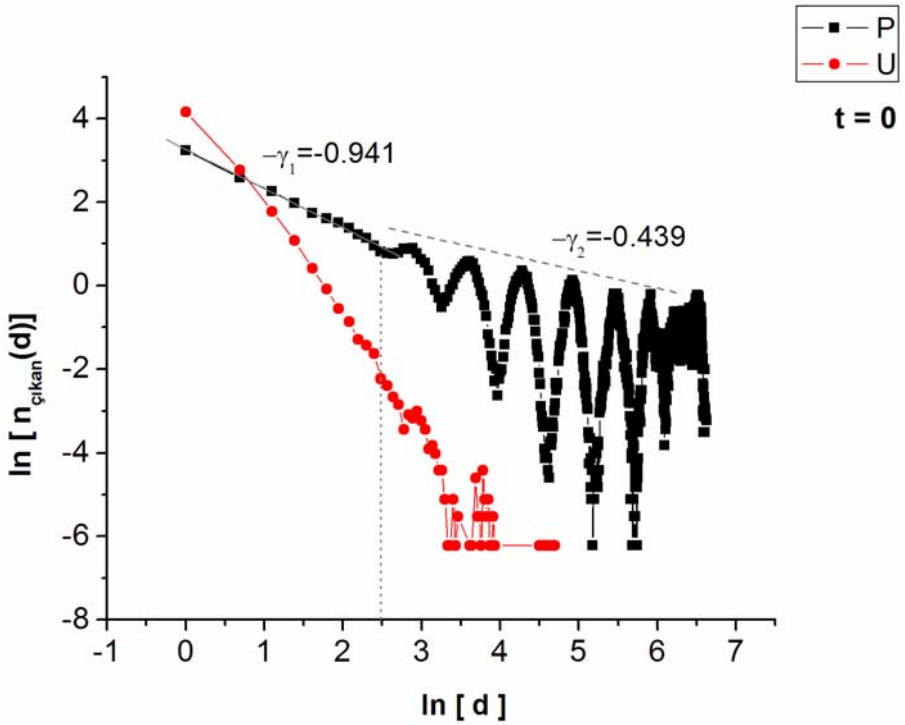
Solid line : finite size effect taken care of by inserting

$$d_l^{\text{in}} = (\sigma_l^{\text{in}})^2$$

Physics of Risk, Vilnius 2006

degree distribution for **Gaussian** and **exponential** length distributions

Out-degree dist



- Exponential

L=15000, N~700

Number of realisations

500

- Gaussian <*l*>= 15 σ = 2

Number of strings *N*=700

Number of realisations 500

Y. Şengün and AE, "Content based networks with duplication and divergence," Physica A, in press

dynamics

N-K models of gene regulation

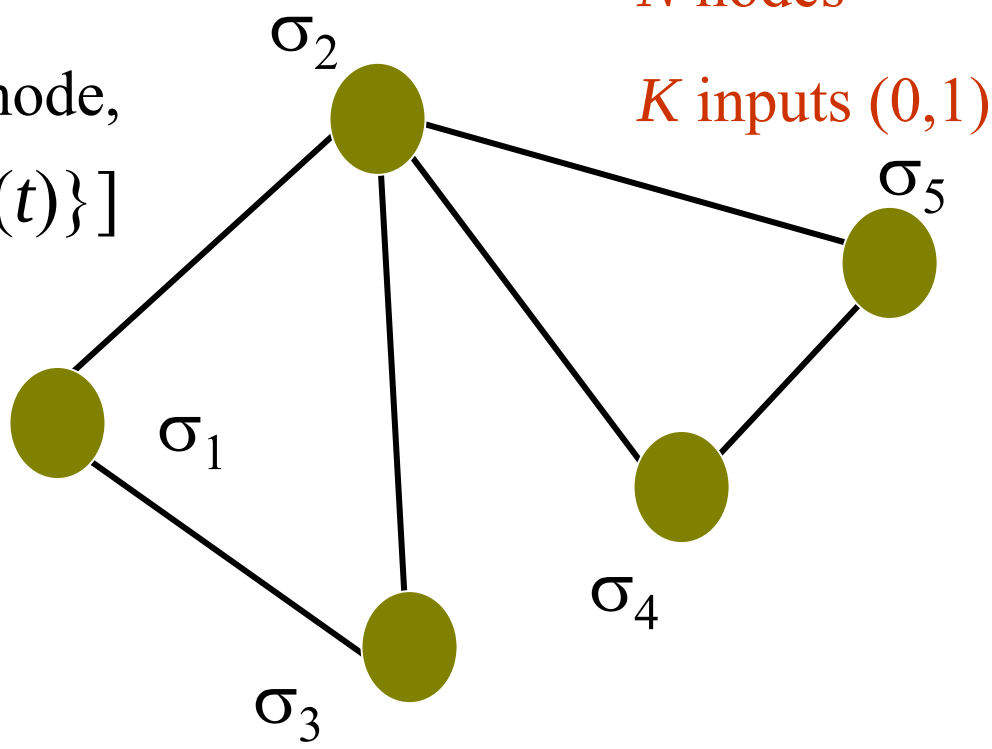
dynamics - N-K models of gene regulation

Random Boolean Functions at each node,

$\sigma_i(t+1) = B_i[\{\sigma_j(t)\}]$

j nn of i

$p = P$ (output =1)

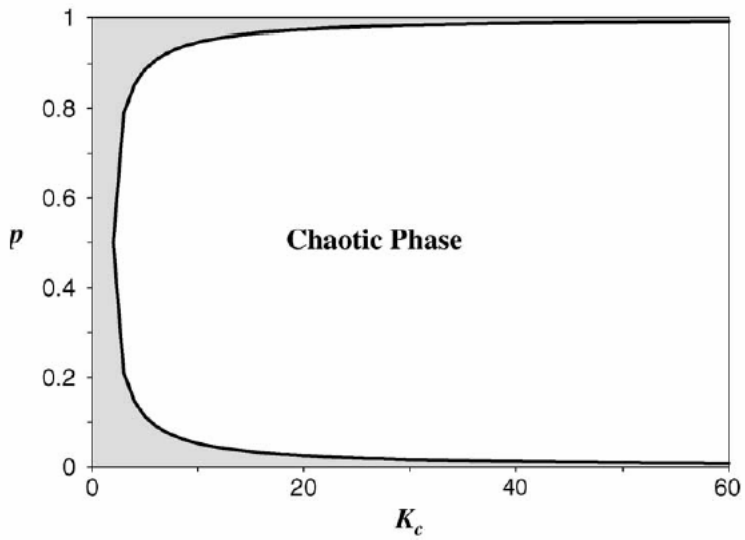


Kauffman 1969, Derrida- Pomeau 1986, Flyvberg 1988, Stauffer 1994, ..

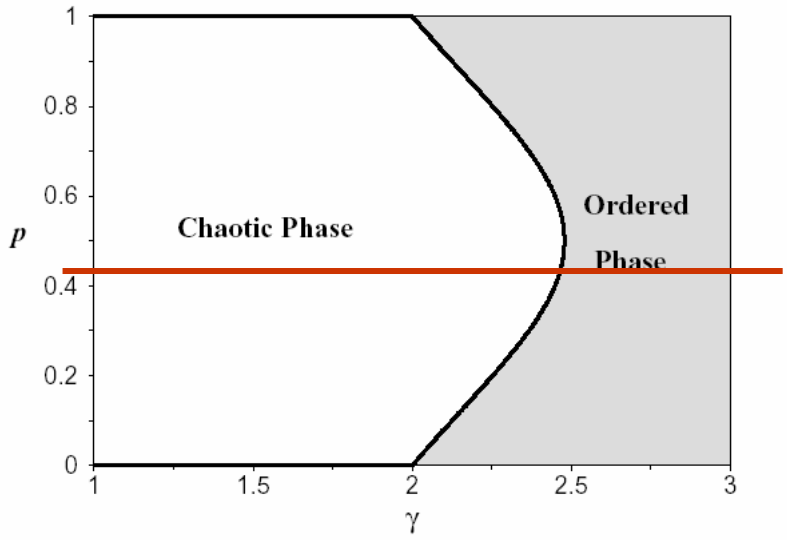
Expect (?) dynamic phase transition at $\langle k \rangle > 2$ or $\gamma < 2.5$

N nodes, *K* inputs (0,1)

Random Boolean gates at each node, $p = P(\text{output} = 1)$



N-K models on random networks



Scale Free NW (Aldana 003)

sequence matching analogue

of N-K models

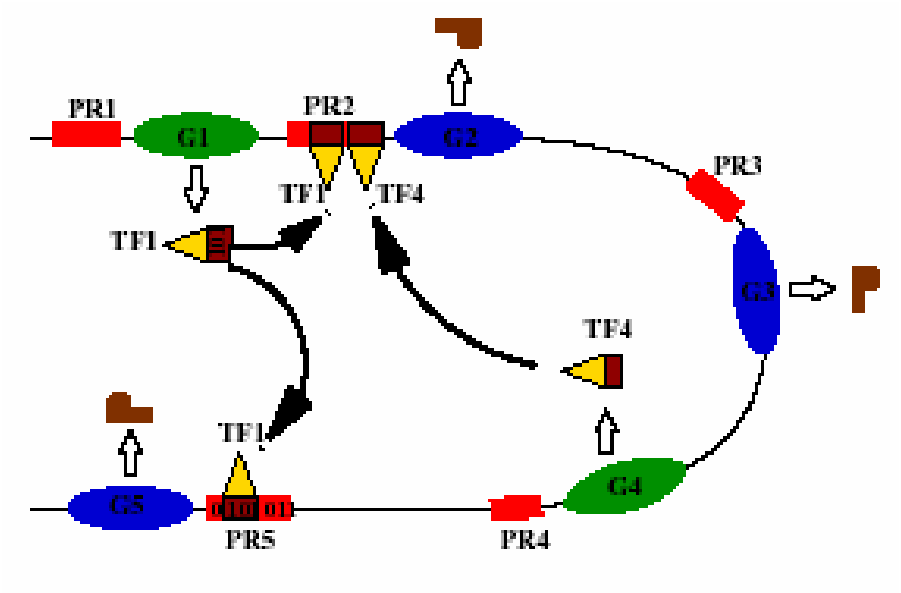
Associate 2 sequences with each node (lengths distributed independently, but identically)

- Promoter region (PR) - Bindings sites for transcription factor
- Coding sequence - a label for a transcription factor (TF) or structural protein

Modelling the gene regulatory network

Two **random** strings per node

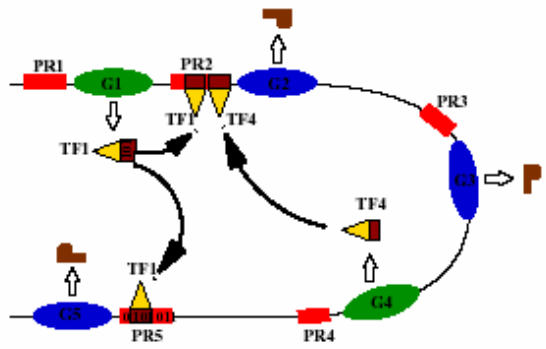
- Transcription Factor **TF**
- Promoter sequence **PR**



Boolean functions at the nodes

Define the states of the nodes $\sigma_i(t) = 0, 1$; $1 \leftrightarrow$ active (ON)
Inputs to the Boolean function at a node i:

$$\sigma_i(t+1) = B_i[\{b_{\Lambda,\nu}\}(t)]$$



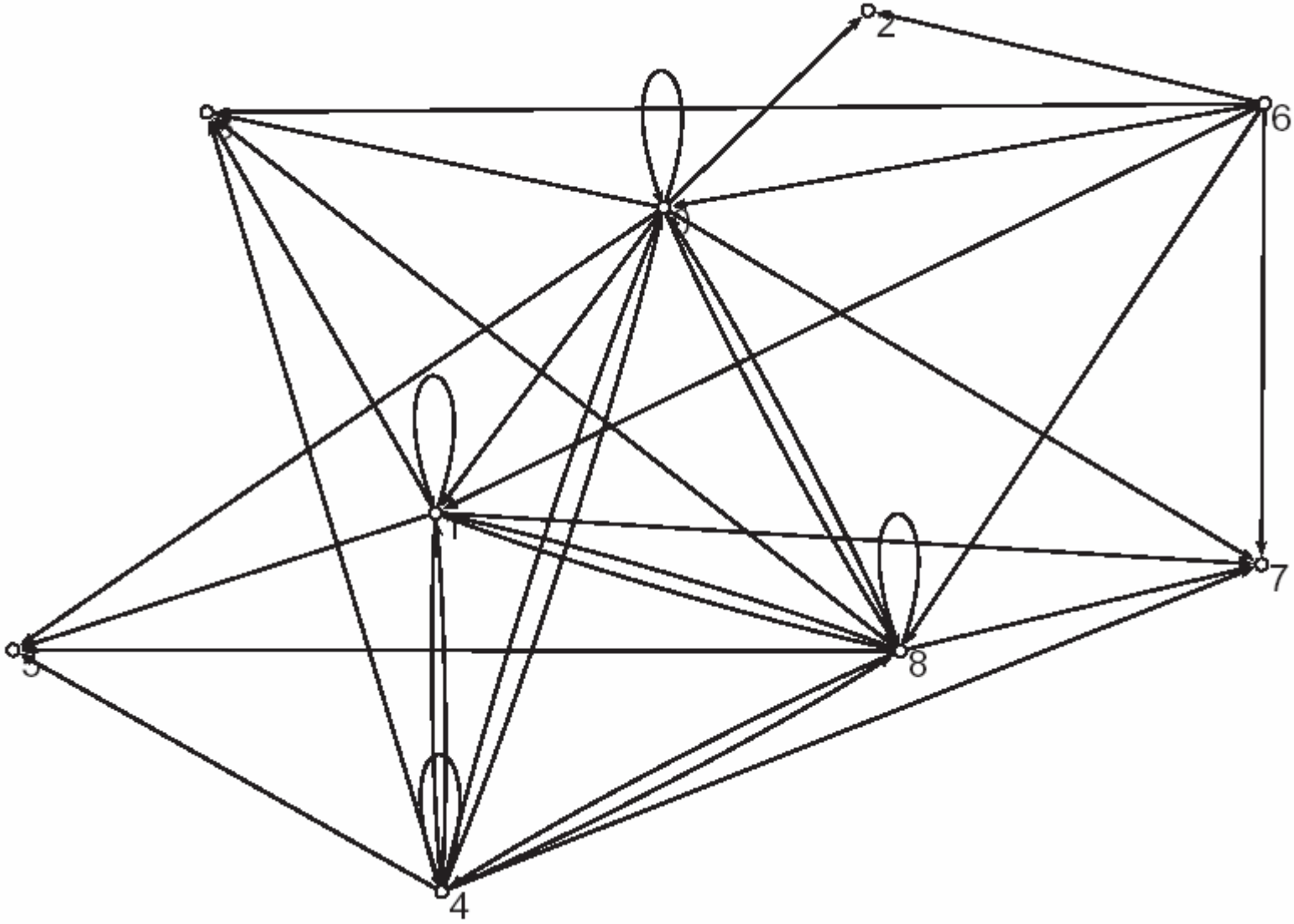
$b_{\Lambda,\nu}$ binding states of all the subsequences $\rho_{\Lambda,\nu}$ of length Λ, shifted by ν, of the promoter region

subsequence $\rho_{\Lambda,\nu}$ will be bound if a TF π_j matching that sequence is being produced at that instant.

Assign outputs $\sigma_i(t+1)$ to all possible $B_i(t)$, with probability p to be 1

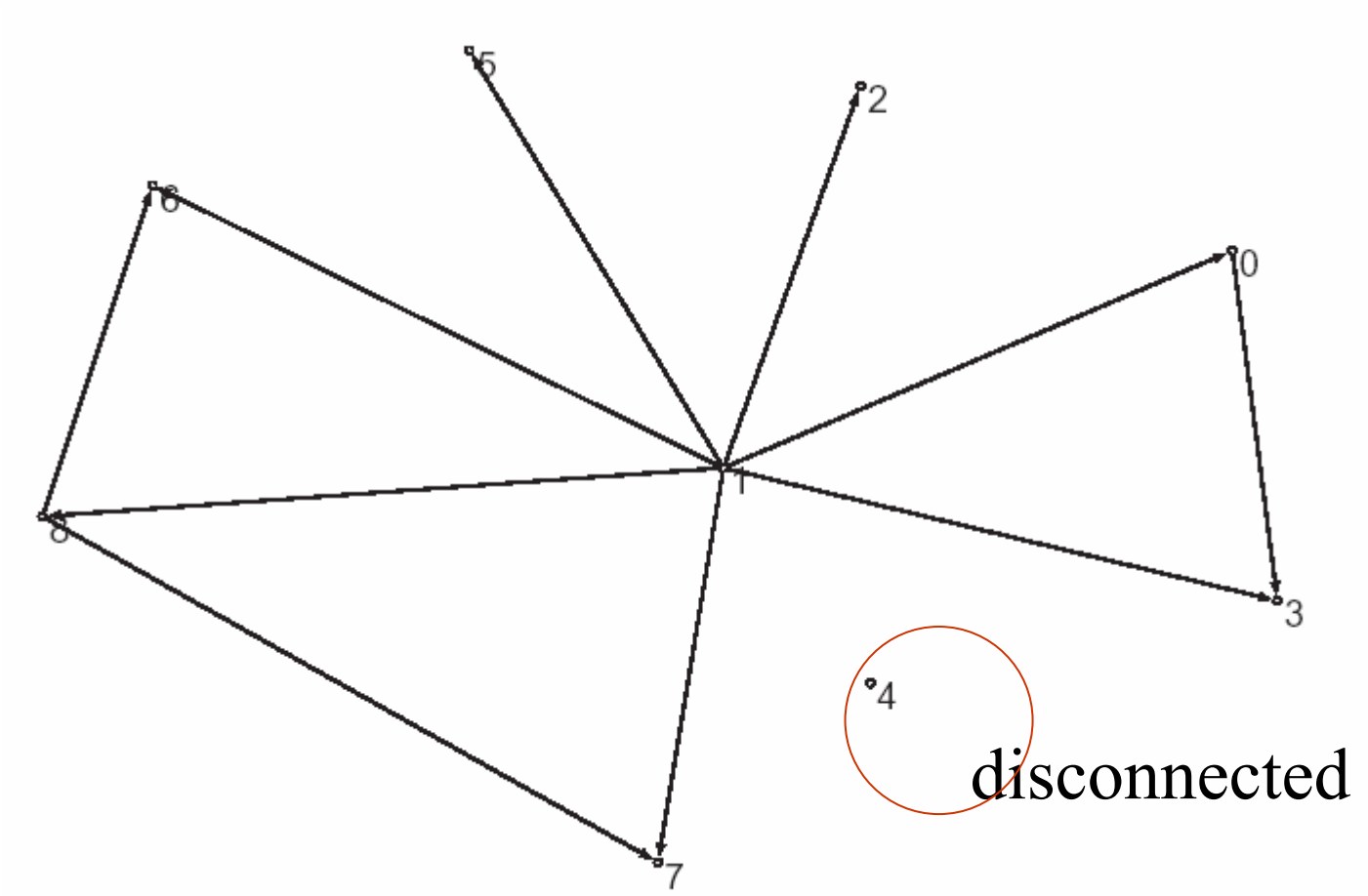
"Exponential" Length Distribution (l_min=1, l_max=25, q=0.9)

Sample 1: Network Topology $N = 9$



Gaussian Length Distribution (l_min=1, l_max=25, σ^2=50, <l>=13)

Sample: Network Topology *N*=9



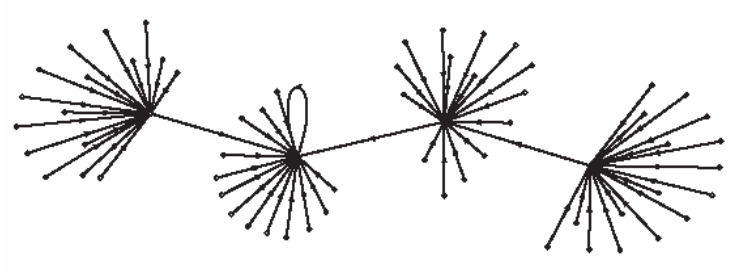
State space

volume $\Omega = 2^N$

a directed graph where each state is represented by a node

and a directed edge is drawn from the node $\Sigma(t)$ to the node $\Sigma(t+1)$

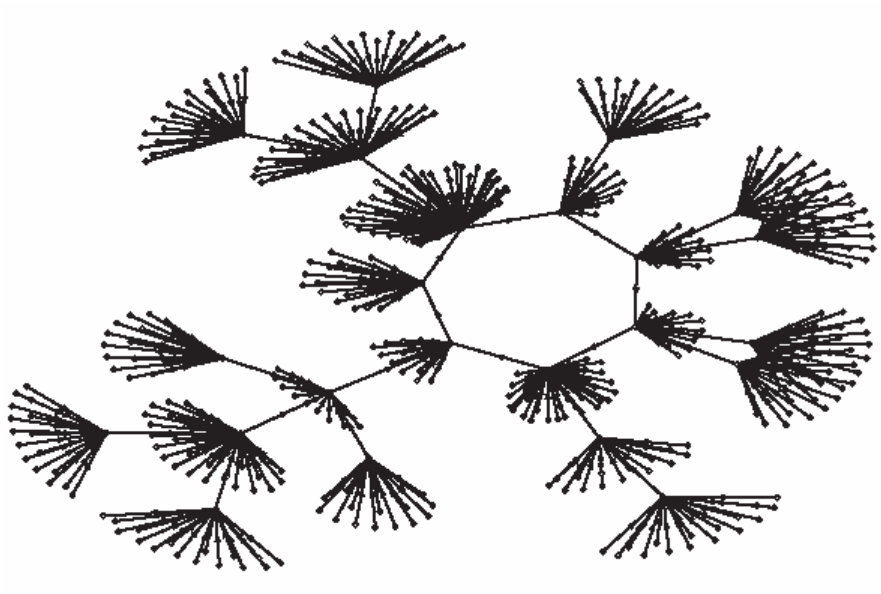
large "surface"



nodes with no precursors

~ "Julia sets" associated

with basins of attraction



Structure of the phase space

- exponential length distribution

similar to critical - ordered

[see, e.g., Aldana, Physica D 185, 45 (2003)]

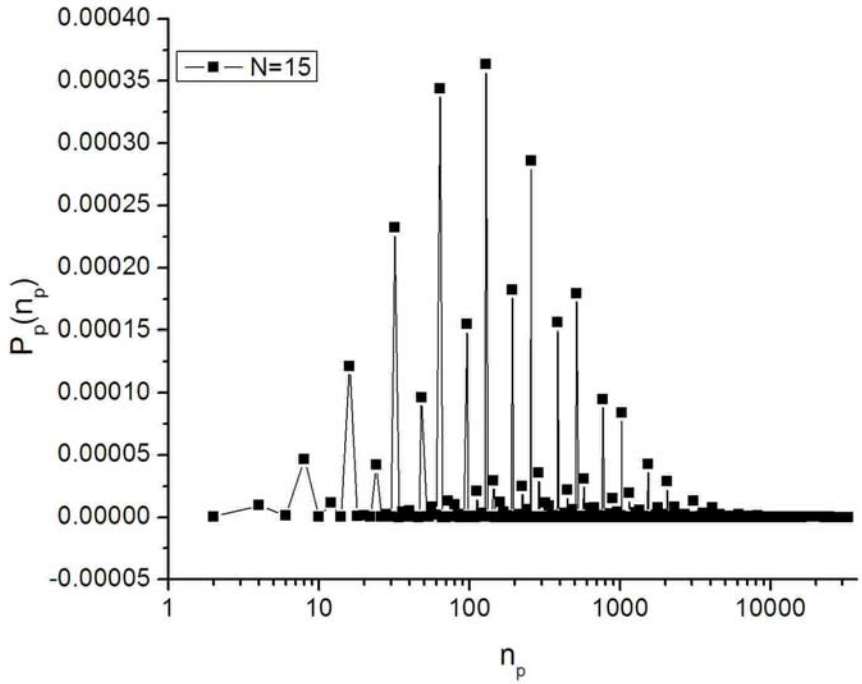
although

- out-deg. distribution $\gamma \sim 1$
- in-deg. dist. non-scaling

n_p = number of precursors

(peak at 0 removed for clarity)

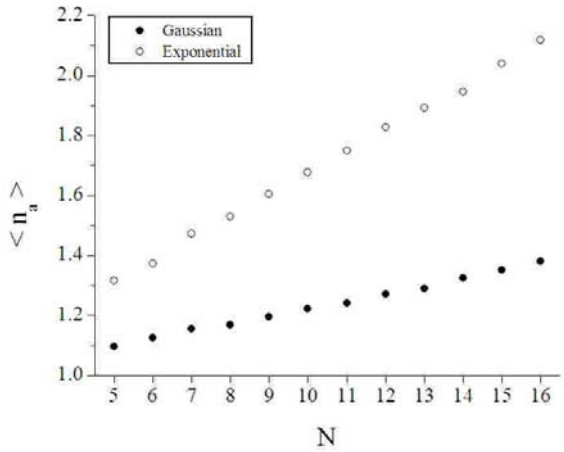
families of powers of 2



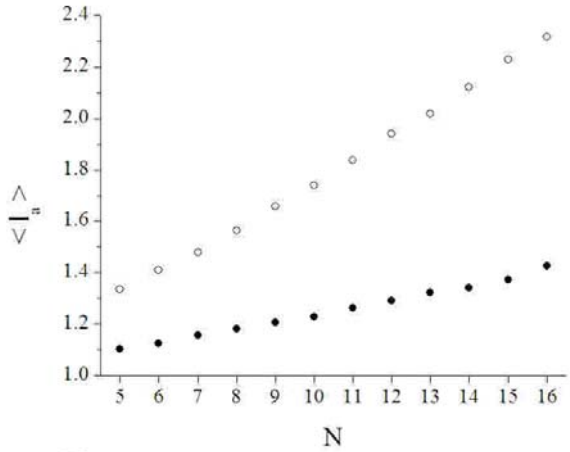
Phase space properties as a function of N

(truncated) Gaussian and exponential length distributions

number of attractors

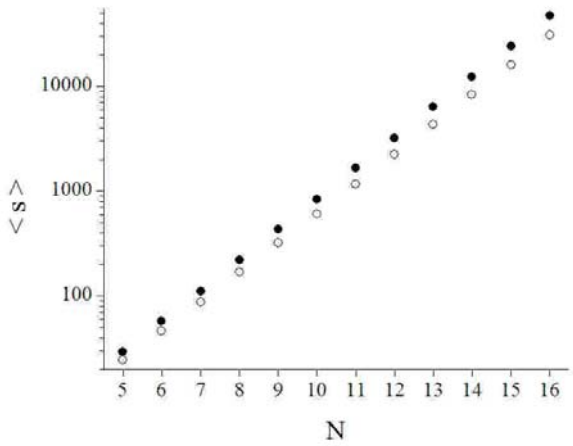


length of the attractors

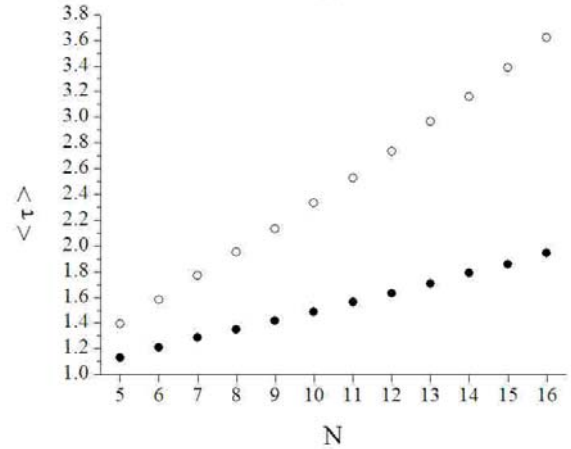


averaged over many realizations

basin size

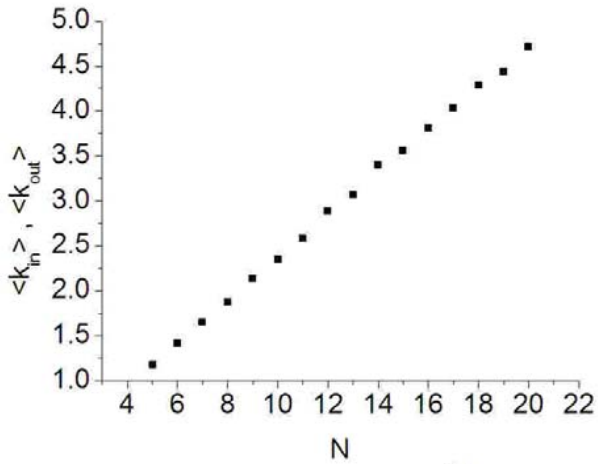


length of transients

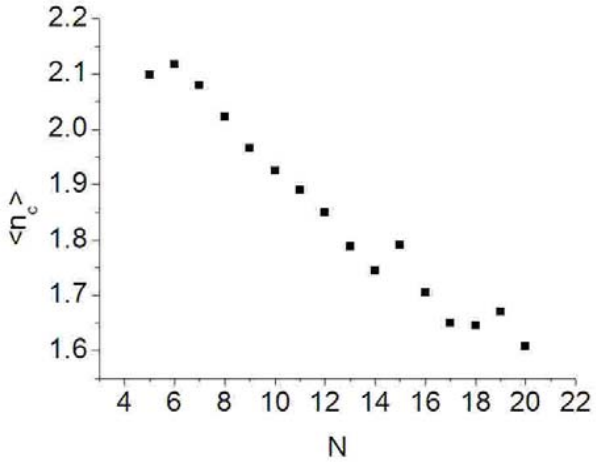


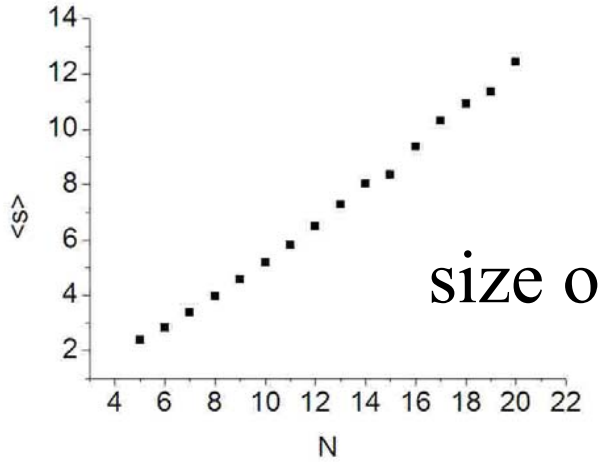
Network properties as a function of N (small)

remains ~ "critical"
although
$\langle k \rangle > 2$ for $N > 8$



number of
Connected
clusters



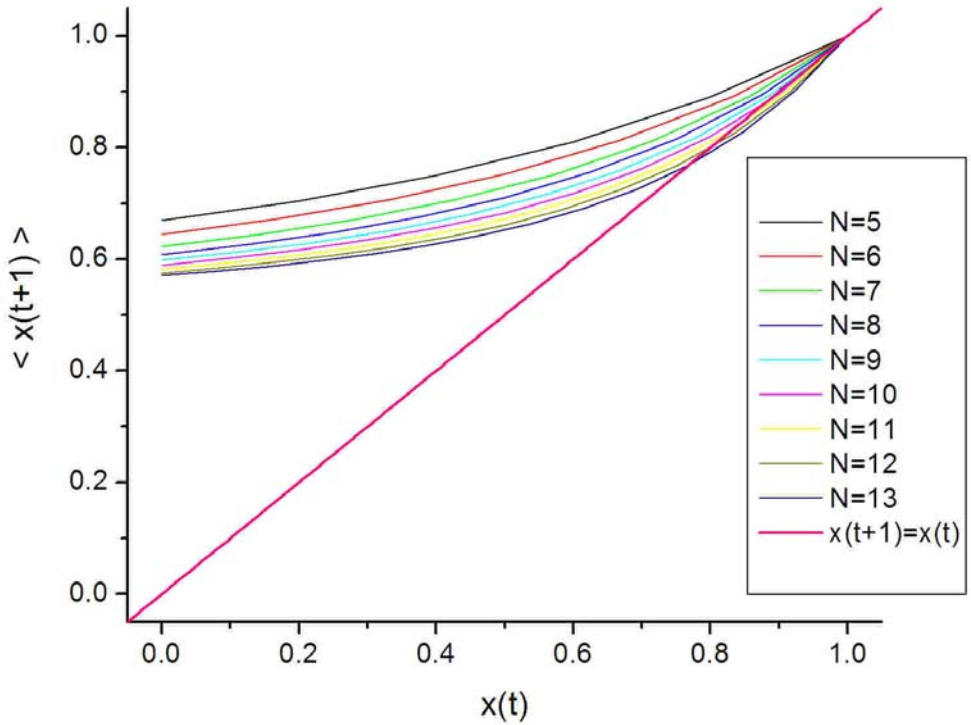


size of clusters

Chaoticity? The average overlap function

between any given pair of points in phase space

$$x(t) = 1 - \frac{1}{N}\sum_{i=1}^{N}\left|\sigma_i(t) - \tilde{\sigma}_i(t)\right|$$



Averaged over all possible pairs and over 10^4 realizations of the network

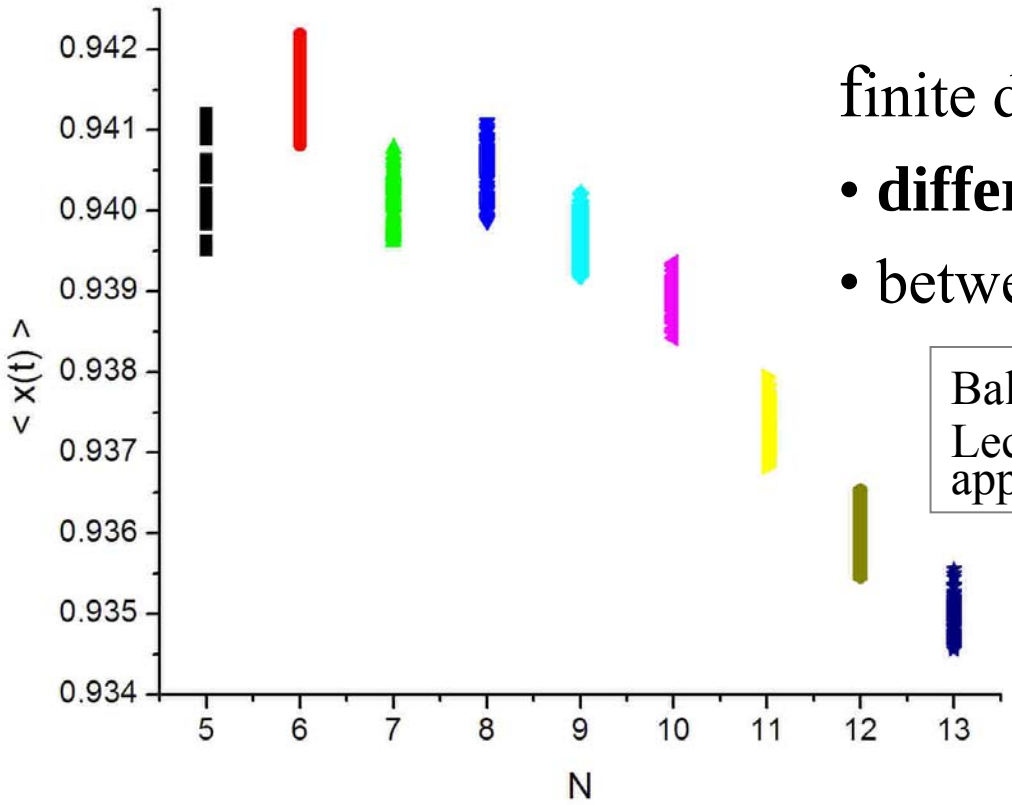
t=0

"fixed point at x=1 becoming unstable?"

1-*d* map misleading!

"bifurcation diagram"

(plotting 100 points after discarding 100 steps - trajectories stabilize after 7-10)

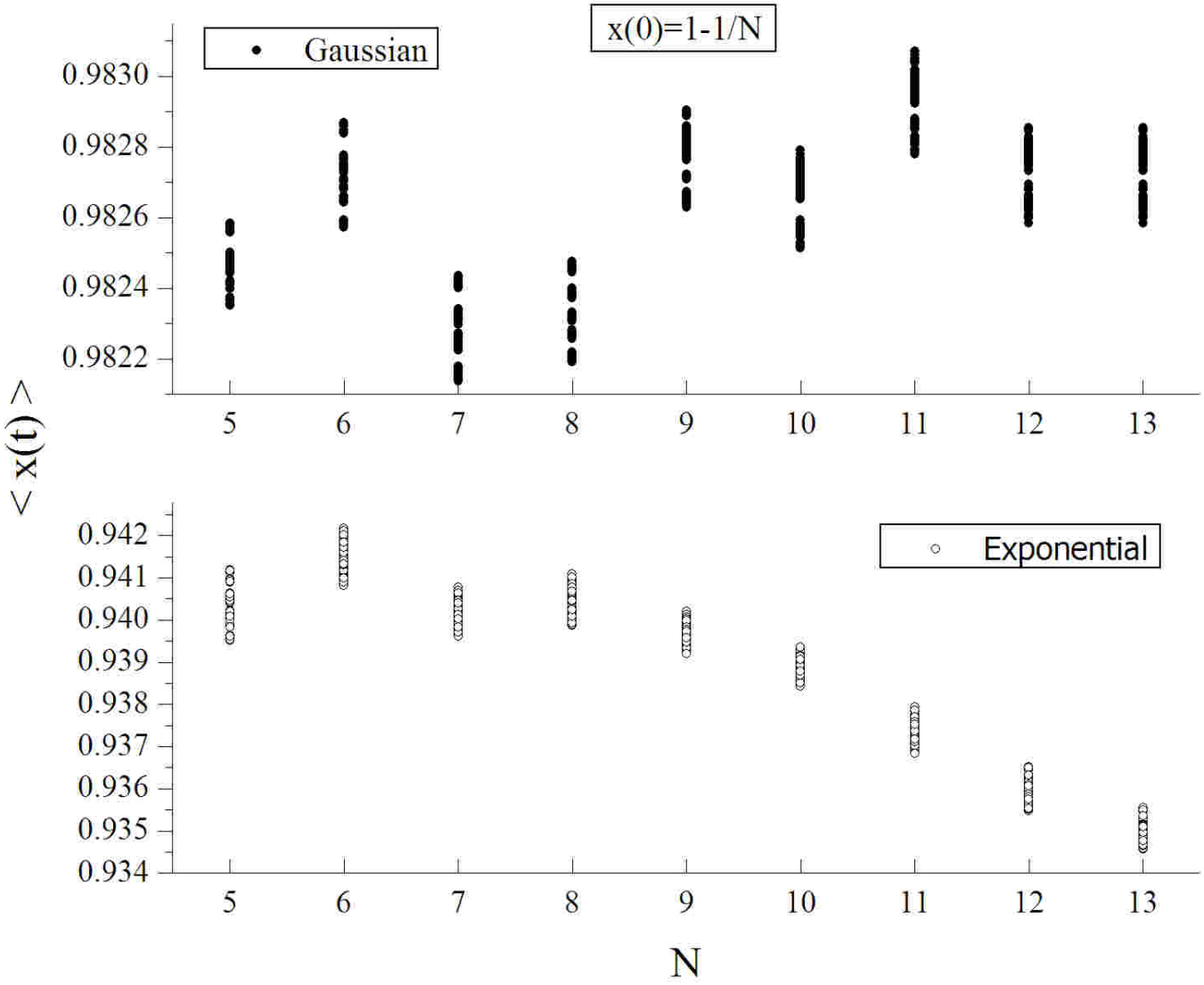


- $\langle x(t) \rangle < 1$ does not mean "chaos" !

finite distances remain between

- **different attractors**;
- between **points on periodic orbits**

Balcan, AE, Proceedings of ICCS06, Lecture Notes series, Springer, to appear


x(0)=1-1/N
Gaussian
0.9830
0.9828
0.9826
0.9824
0.9822
5
6
7
8
9
10
11
12
13
< x(t) >
Exponential
0.942
0.941
0.940
0.939
0.938
0.937
0.936
0.935
0.934
N

Comparison with data : the gene regulatory network of yeast

Modelling the gene regulatory NW of yeast

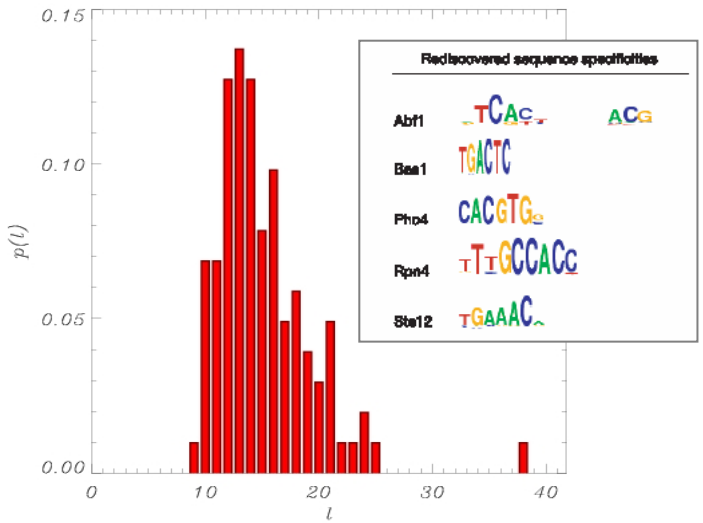
Experimental input:

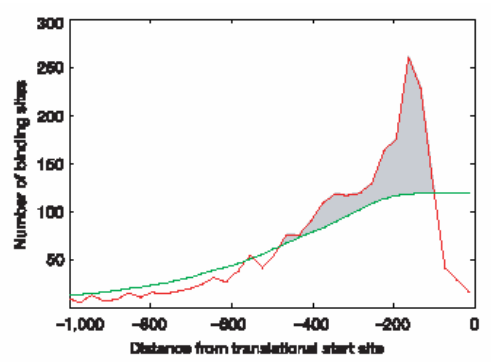
- Length distribution of the TF sequences
- Interpret as the information content (Shannon entropy) in units of base 2

Fitting parameters

- parameters of the PR length distribution
  - Exponential and Gaussian do not work
  - **Power law** suggested by work of Provata et al.

Length distribution (information content) of the TF and PR distributions





TF length distribution computed from

Harbison et al., *Nature* **431,** 99-104 (2004)

High confidence level - 2 bits

Low confidence level - 1 bit

PR length distribution

Oikonomou and Provata q-bio.GN/0510021

Almirantis and Provata, *J. Stat. Phys.* **97**, 233 (1999)

$$p(l) \sim l^{-1-\mu}$$

$$l_{\min} = 13$$

$$l_{\max} = 13+250 \text{ (Harbison)}$$

Topological features of networks

- degree (in-, out-, undirected) distribution
- clustering coefficient

$$c_i = \frac{\Delta_i}{k_i(k_i-1)/2},$$

- *k*-*k*' correlation

$$k_{nn}(k) = \sum_{k'} k' p(k'|k)$$

- rich-club coefficient

$$r(k) = \frac{2e_{>k}}{N_{>k}(N_{>k}-1)}$$

- *k*-core decomposition

decomposing graph into successive *k*-irreducible shells

Alvarez-Hamelin, I., Dall'Asta, L., Barrat, L., Vespignani, cs.NI/0504107

- **most stringent** condition for comparison -

fit the single parameter μ by optimizing the *k*-core decomposition

Generating a model ensemble

Yeast

yeastract data

http://fraenkel.mit.edu/Harbison/release_v24/bound_by_factor

http://www.yeastract.com

http://sandy.topnet.gersteinlab.org/index2.html (Luscombe et al.)

Source	Genes	TFs	Interacting Pairs
Fraenkel Lab*	2884	102	6441
Yeastract†	4252	146	12530
Luscombe et al.‡	3459	142	7071
Kirdar et al.	3763	180	9135

PR length dist: $\mu = 1.2$

best fit for

k-core distributions

ensemble for model NW

100 independent realizations

Ensemble averages

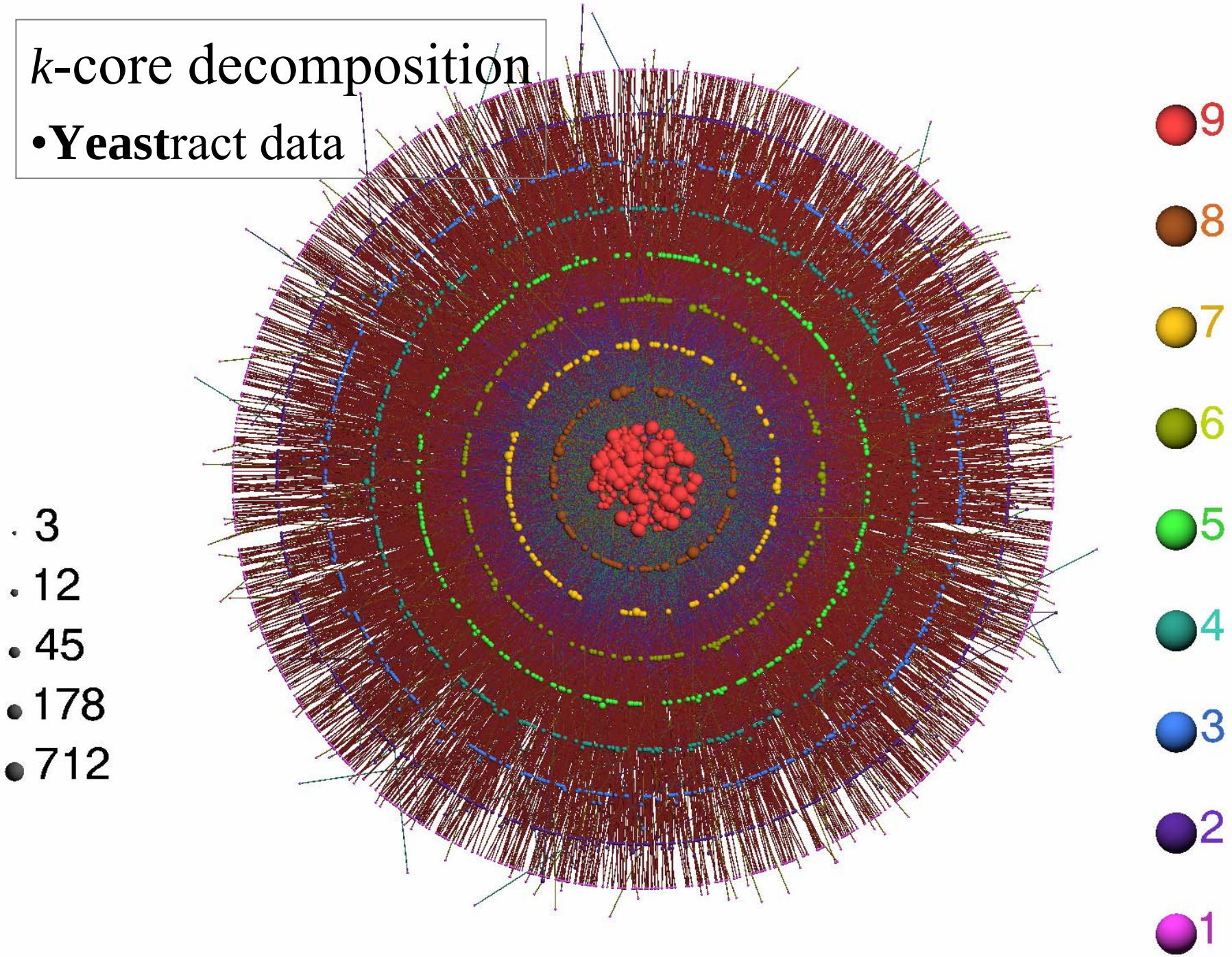
$\langle N \rangle = 6000$

$\langle N_{TF} \rangle = 202$

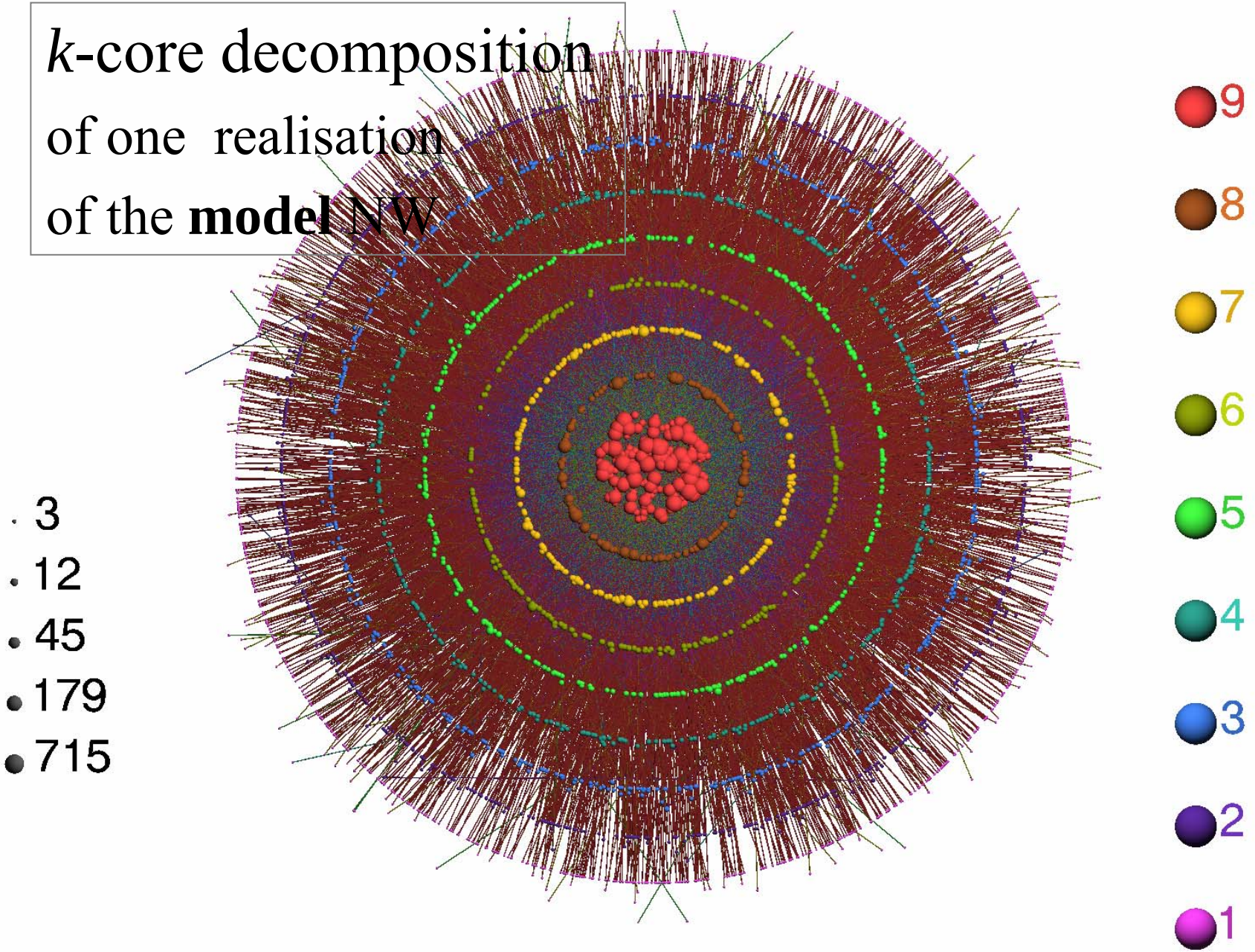
edges = 14 365

k-core decomposition

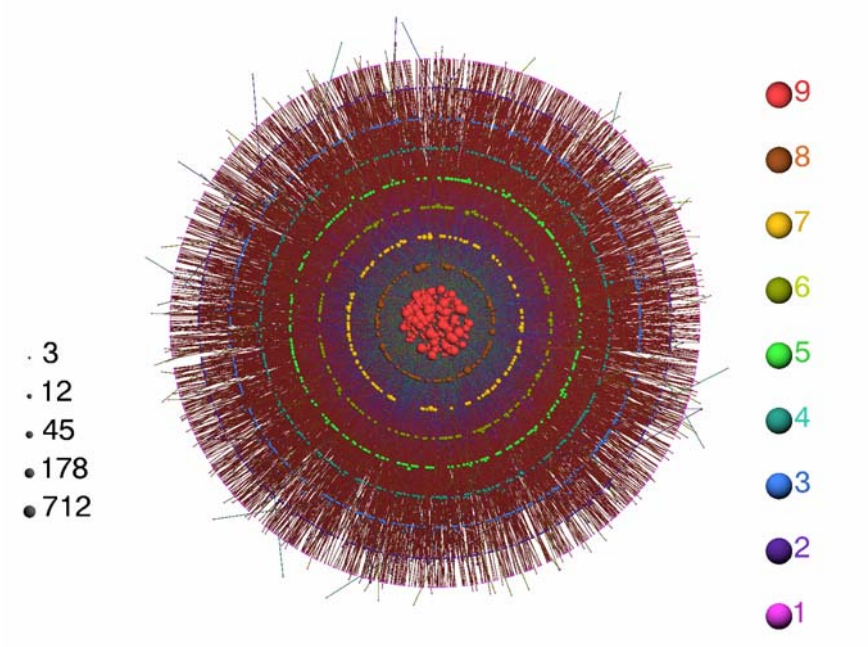
- **Yeast**ract data



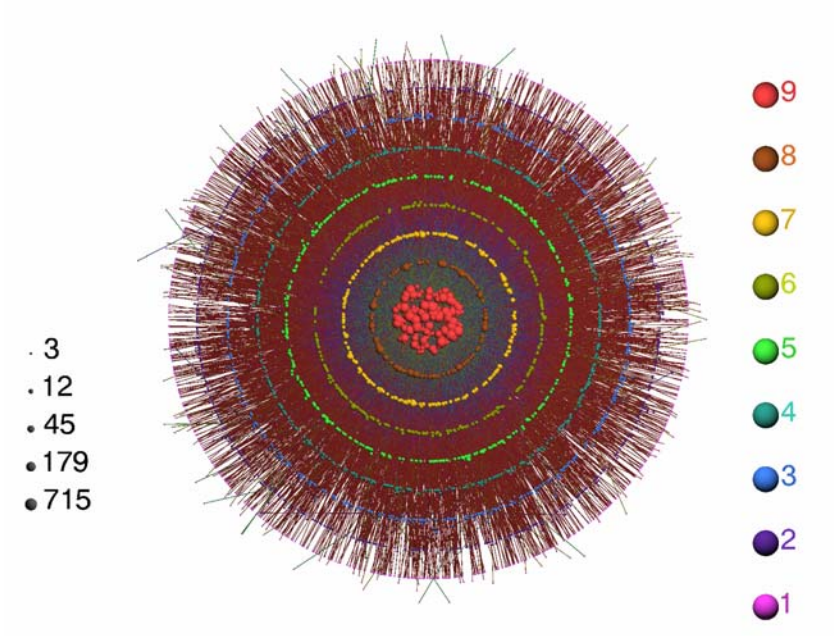
k-core decomposition of one realisation of the **model** NW



•yeast



•model

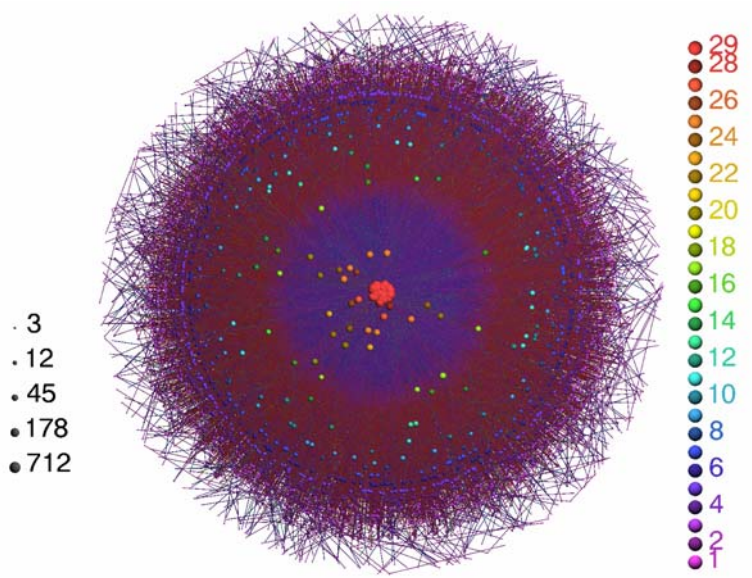


nodes

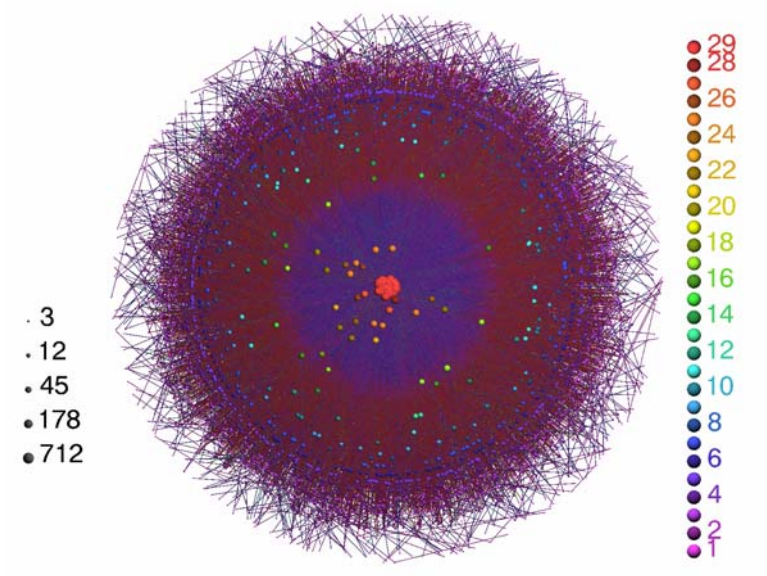
core number

k-core decomposition of **randomized** NW keeping the degrees fixed

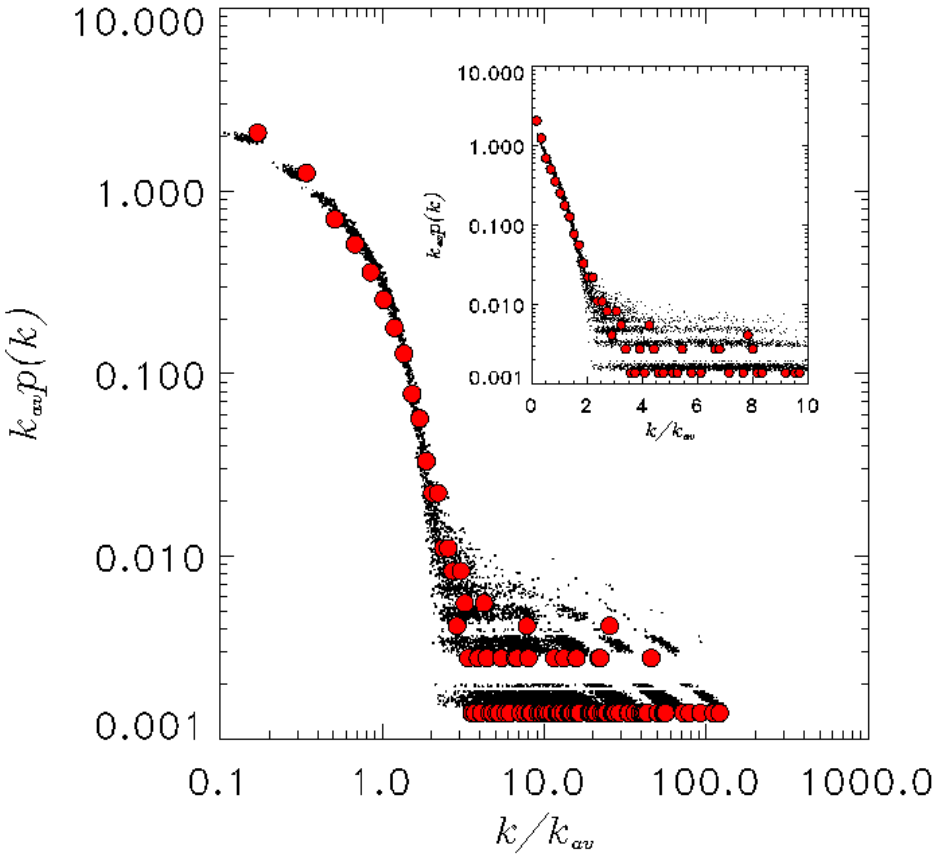
- yeast



- model



degree distribution ● yeast …model



ensemble of model NW

100 independent realizations

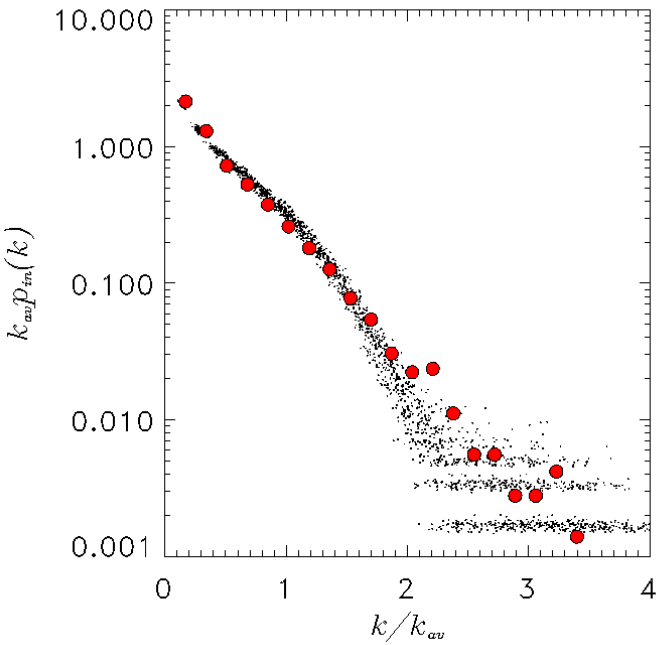
Ensemble averages

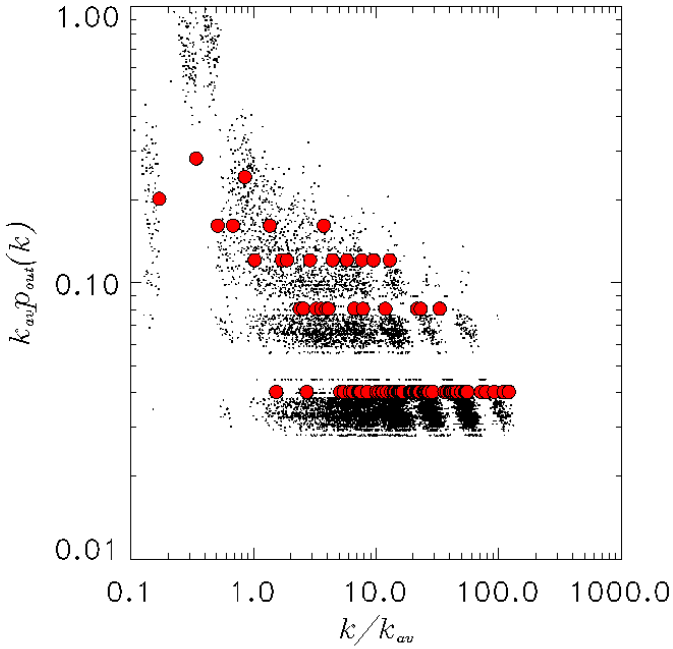
<N> = 6000

$\langle N_{TF} \rangle$ = 202

edges = 14 365

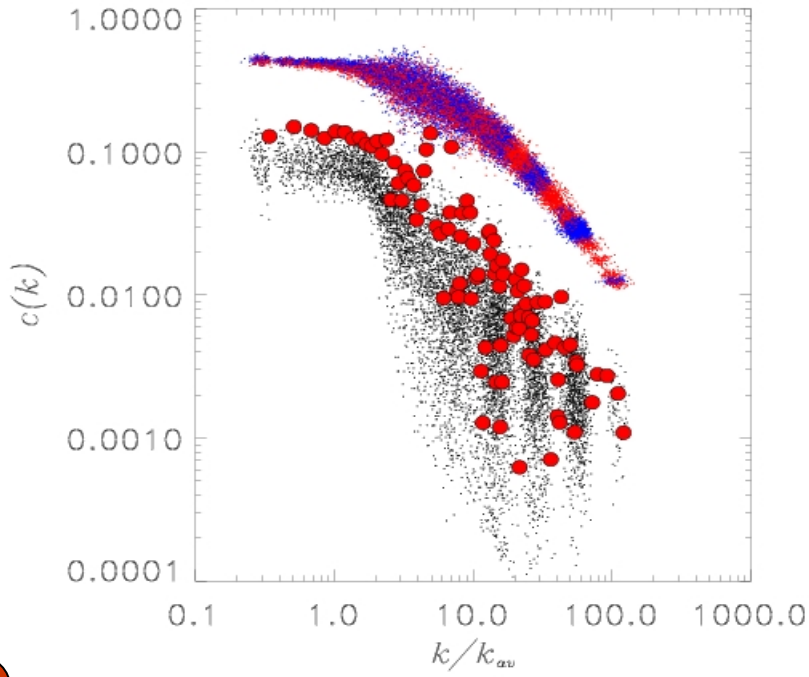
in- and out-degree distributions

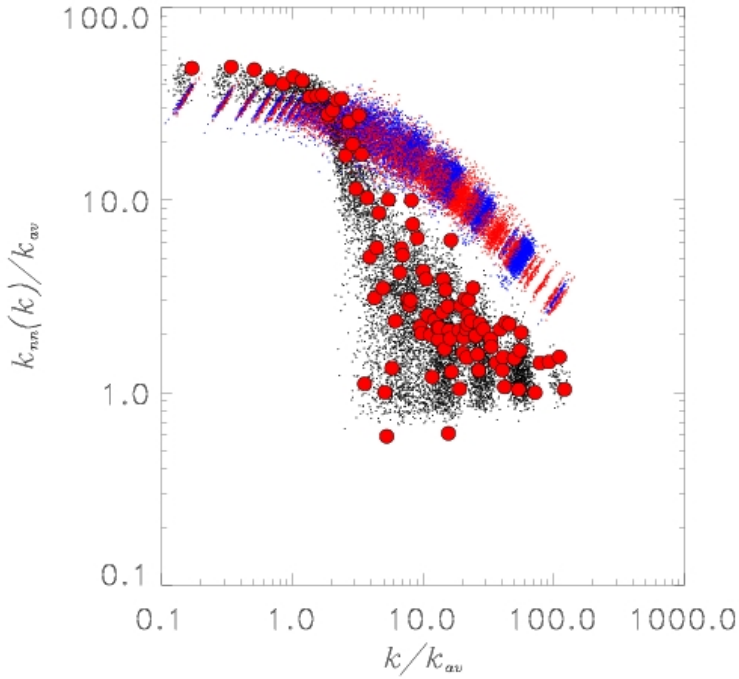




clustering coefficient

degree-degree correlation

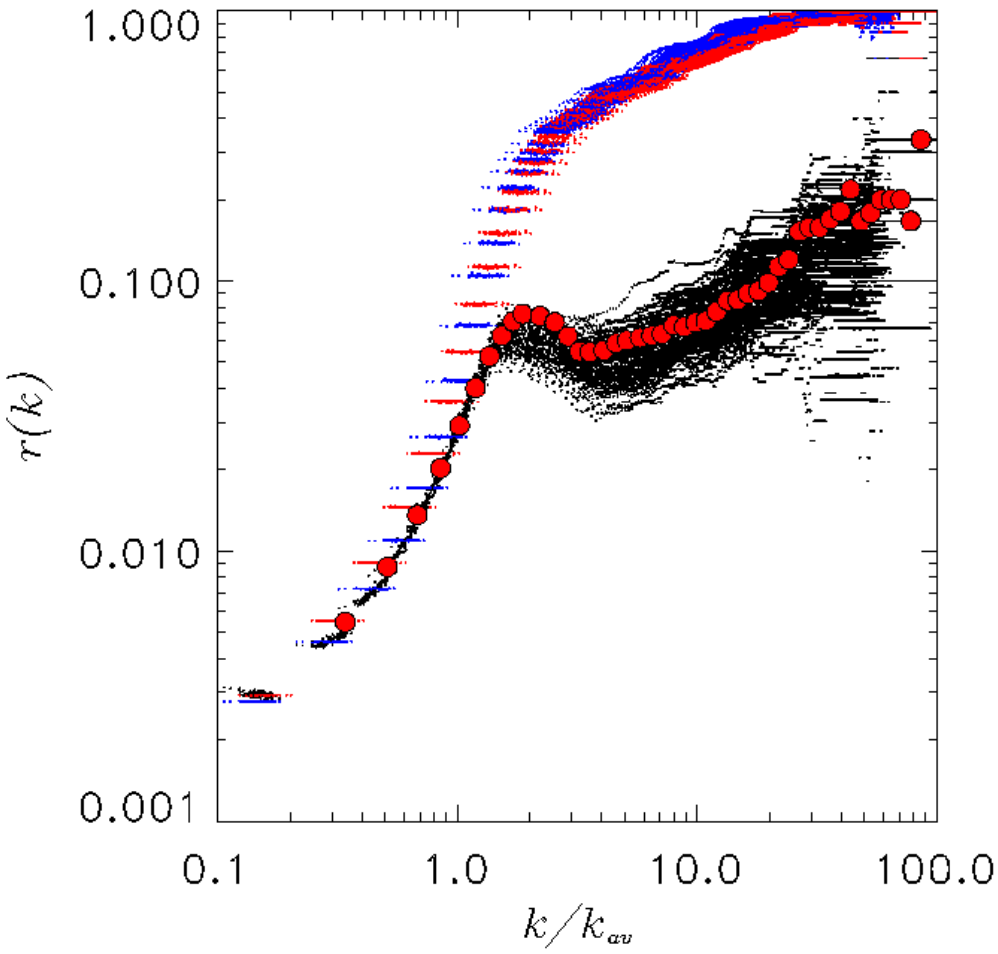




yeast ●●●● model ●●● randomized yeast ●●● randomized model

randomizations performed while keeping degree of each node fixed

rich-club coefficient



"shoulder" common to yeast GRN and to protein-protein interactions

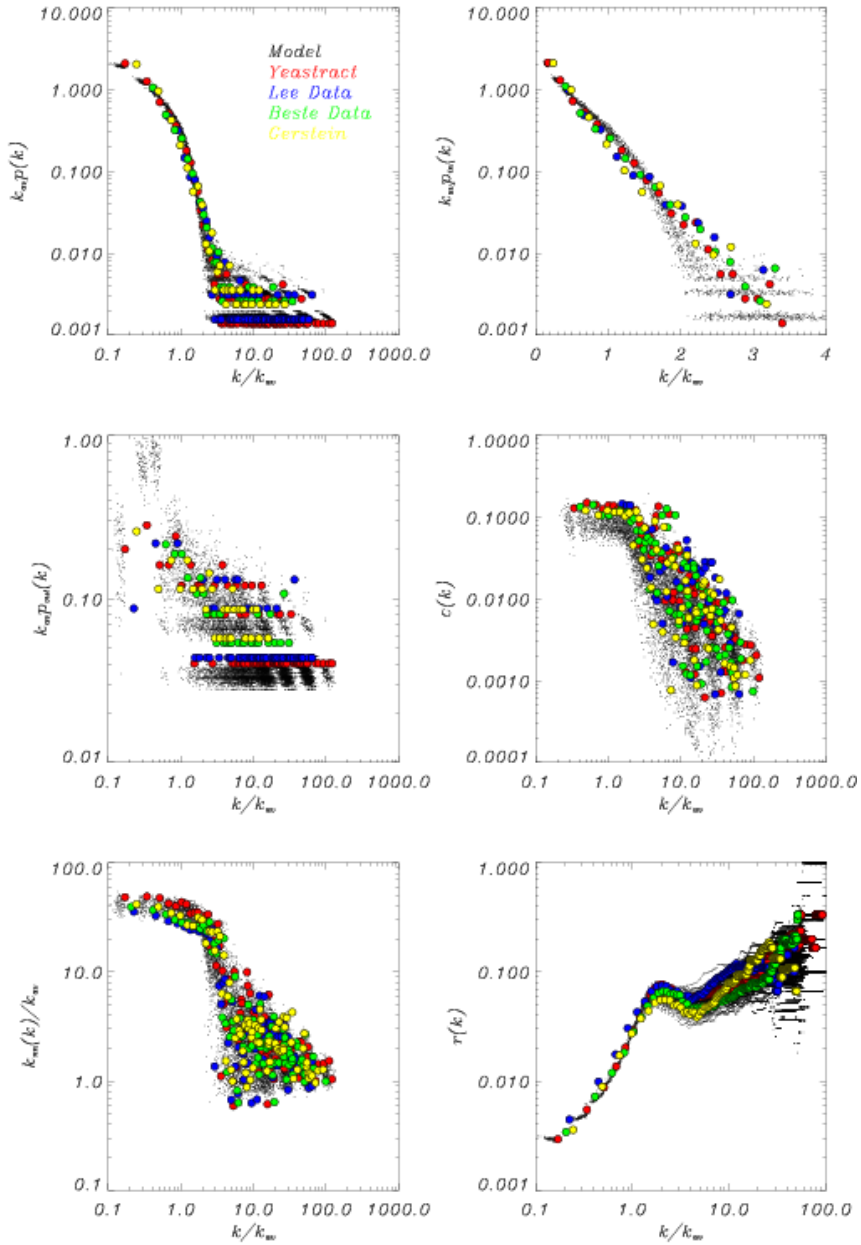
Superposition
of model ensemble
and data from different
sources

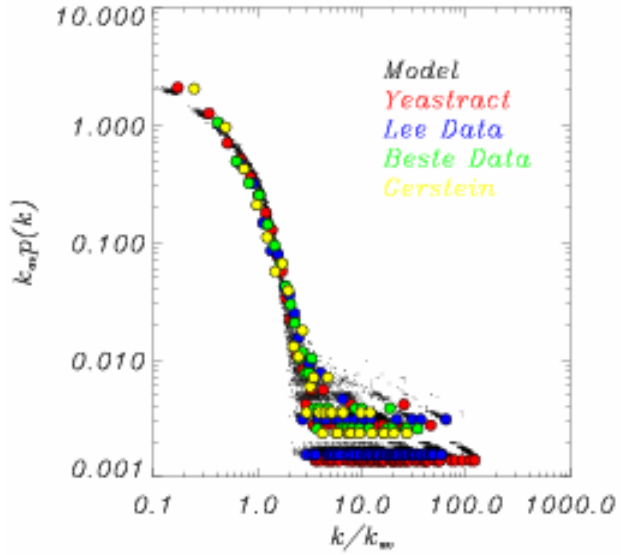
Yeastract

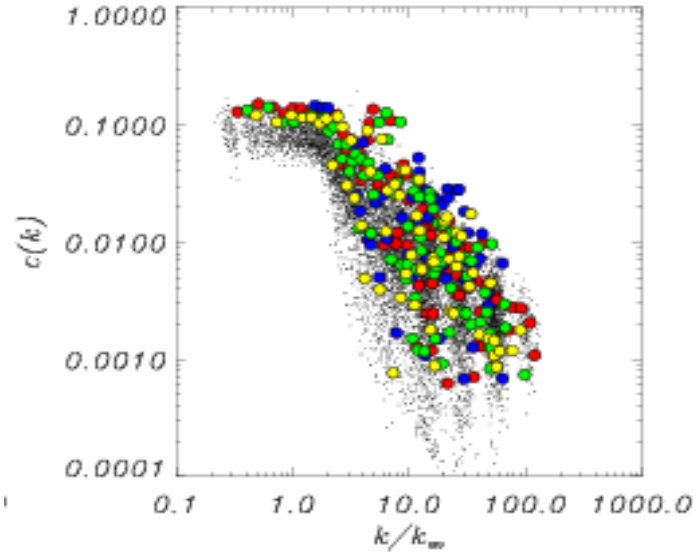
Fraenkel Lab

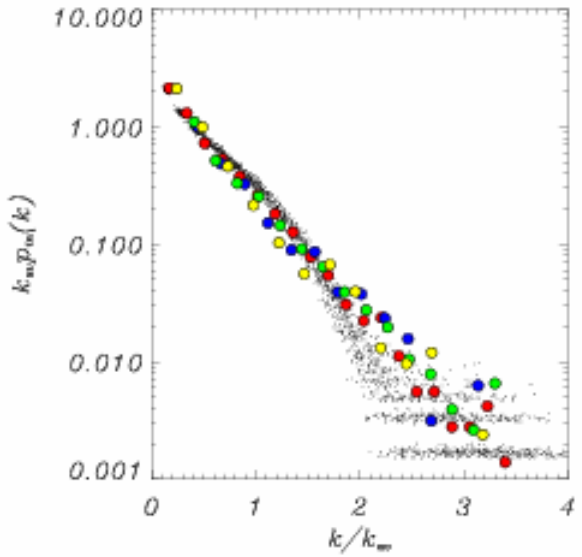
Kirdar and Kinikoglu

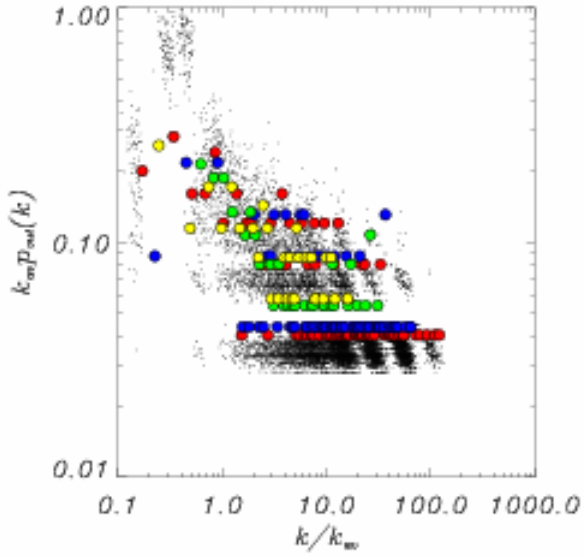
Luschcombe et al.




Model
Yeastract
Lee Data
Beste Data
Gerstein
10.000
1.000
0.100
0.010
0.001
0.1
1.0
10.0
100.0
1000.0
$k_{ss}P(k)$
k/k_{ss}


1.0000
0.1000
0.0100
0.0010
0.0001
0.1
1.0
10.0
100.0
1000.0
$c(k)$
k/k_{ss}


10.000
1.000
0.100
0.010
0.001
0
1
2
3
4
$k_{ss}p_{ss}(k)$
k/k_{ss}


1.00
0.10
0.01
0.1
1.0
10.0
100.0
1000.0
$k_{ss}p_{ss}(k)$
k/k_{ss}

summary

- Completely **capture detailed structure of yeast gene regulatory network** using experimental length distribution and fitting μ by optimizing k-core decomposition
- Content based networks **sustain critical behaviour for k>2**. Also for power law PR length distribution?
- **Statistics suffice** to generate complex network - evolution did not have to assemble it from scratch?!