The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

World Year of PHYSICS 2005

SMR.1656 - 25

School and Workshop on
Structure and Function of Complex Networks

16 - 28 May 2005

Large-scale topological patterns in protein networks

Sergei Maslov
Brookhaven National Laboratory

Strada Costiera 11, 34014 Trieste, Italy - Tel. +39 040 2240 111; Fax +39 040 224 163 - sci_info@ictp.it, www.ictp.it

Large-scale topological patterns in protein networks

Sergei Maslov

Brookhaven National Laboratory

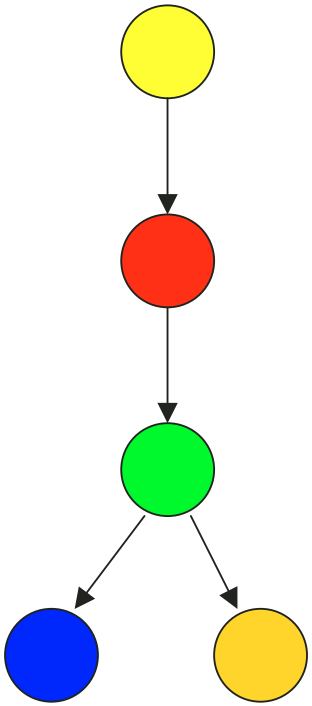
Life strives on interaction

Complex biological processes use the coordinated activity of many interacting molecules

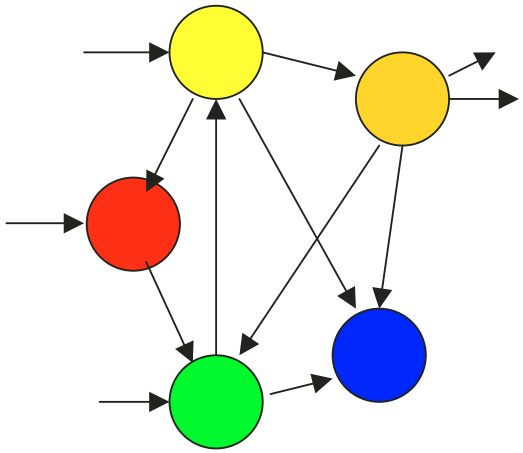
Interactions between molecules serve to:

- Turn genes on and off in response to environmental stimuli and (more rarely) maintain complex dynamical patterns (e.g. cell cycle, circadian cycle, etc.)
- Propagate signals e.g. from outside the cell, through the membrane and the cytoplasm to the nucleus
- Make structural elements of the cell and multi-protein complexes (yeast ribosome ~32+46=78 proteins encoded by 137 genes +4 rRNA)

Pathway → network paradigm shift

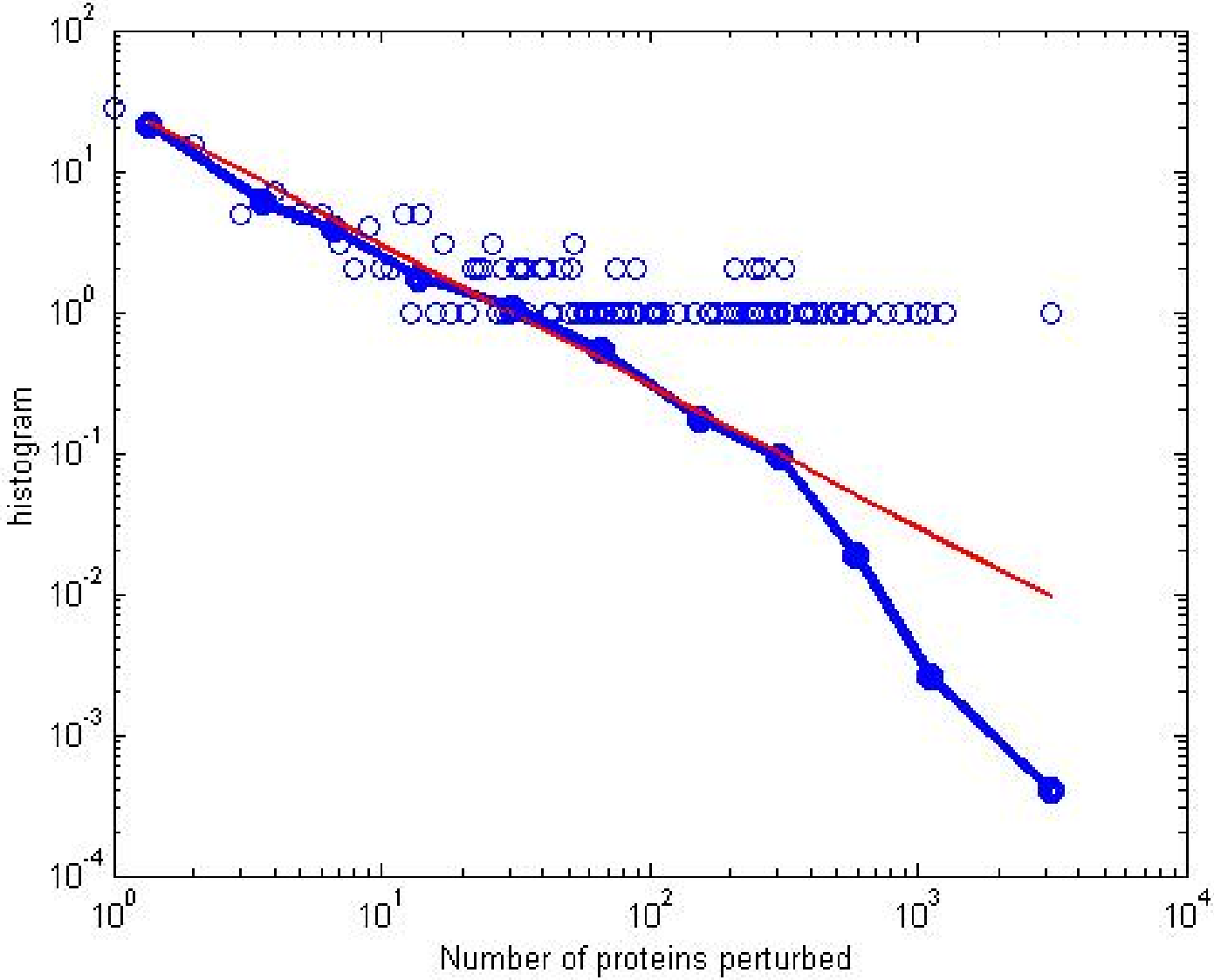


Pathway



Network

Gene disruptions in yeast



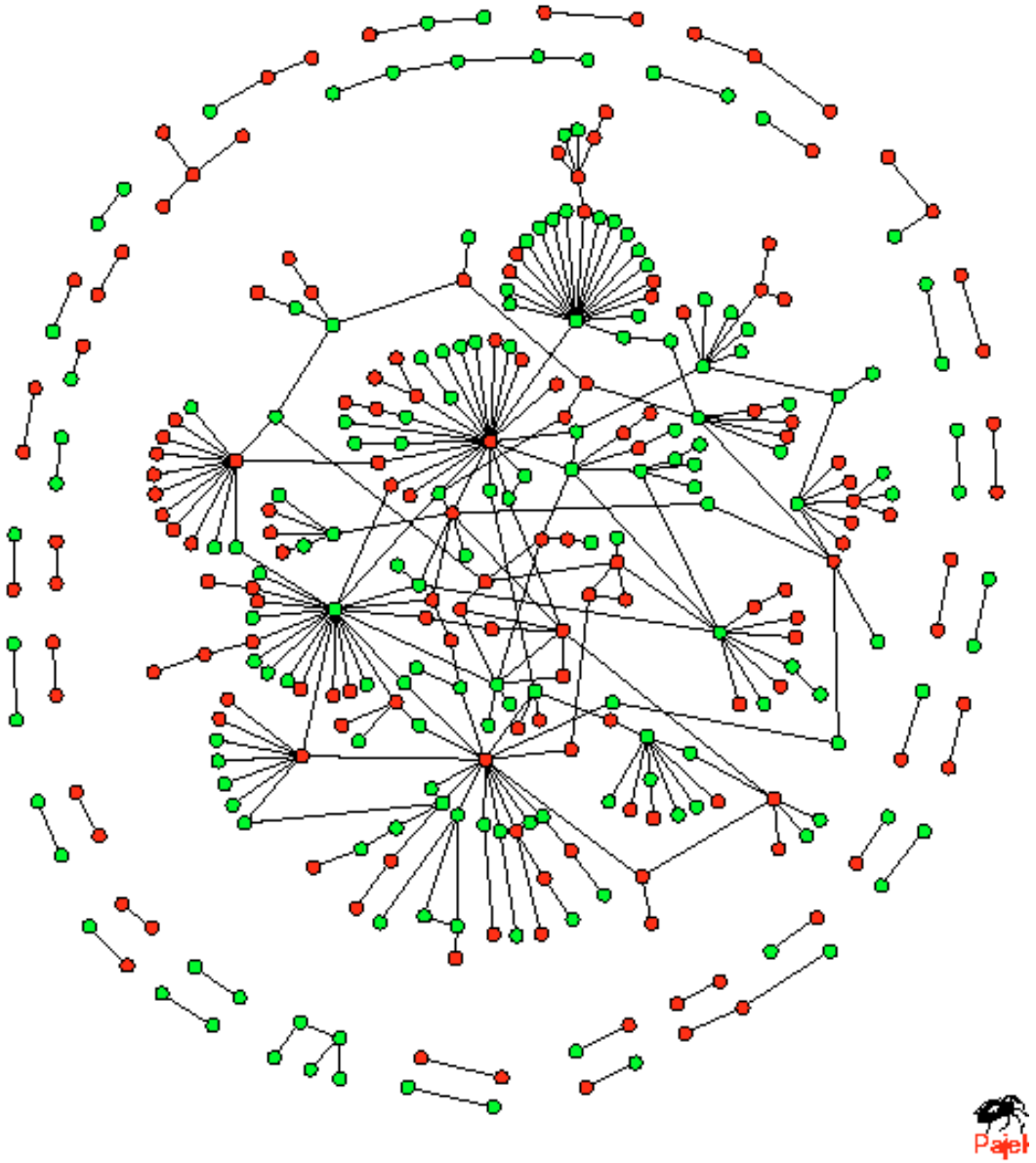
Genome-wide protein networks

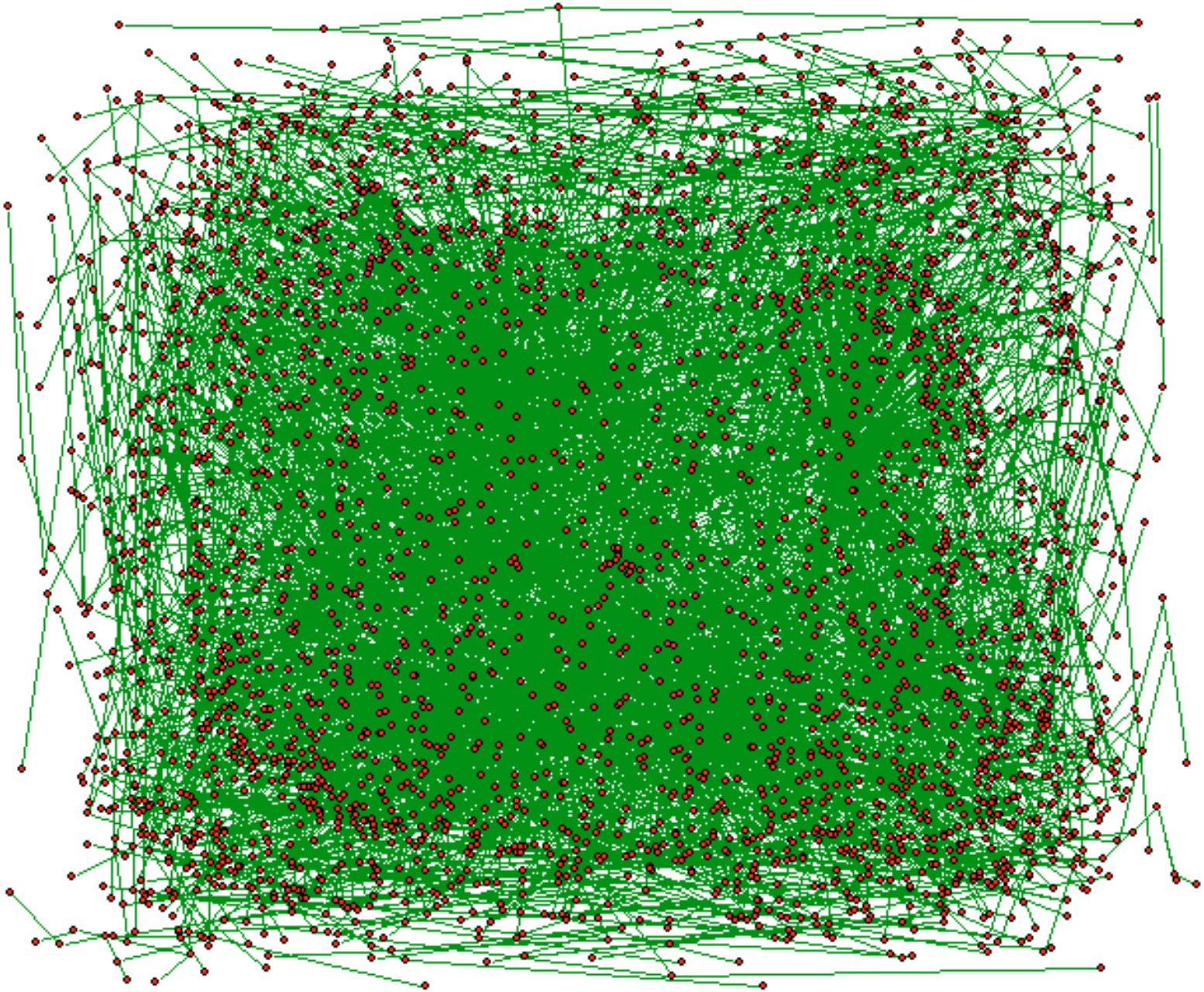
Nodes - proteins

Edges – interactions between proteins

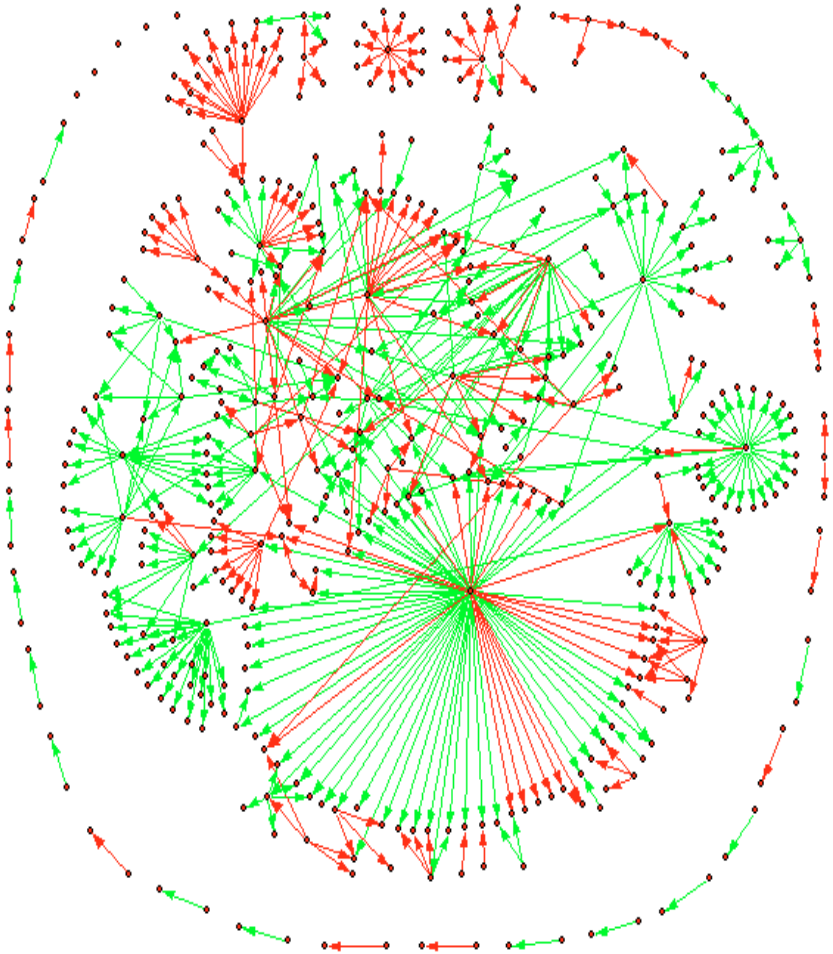
- Direct:
  - Bindings (physical interactions)
  - Regulations
    - Transcriptional (specialized proteins binding DNA)
    - protein modifications (e.g. phosphorylations by kinases)
    - etc.
- Indirect:
  - Disruptions in expression (mRNA production from genes)
  - Co-expression
  - Involvement in consecutive metabolic reactions
  - Etc, etc, etc.

Protein binding network

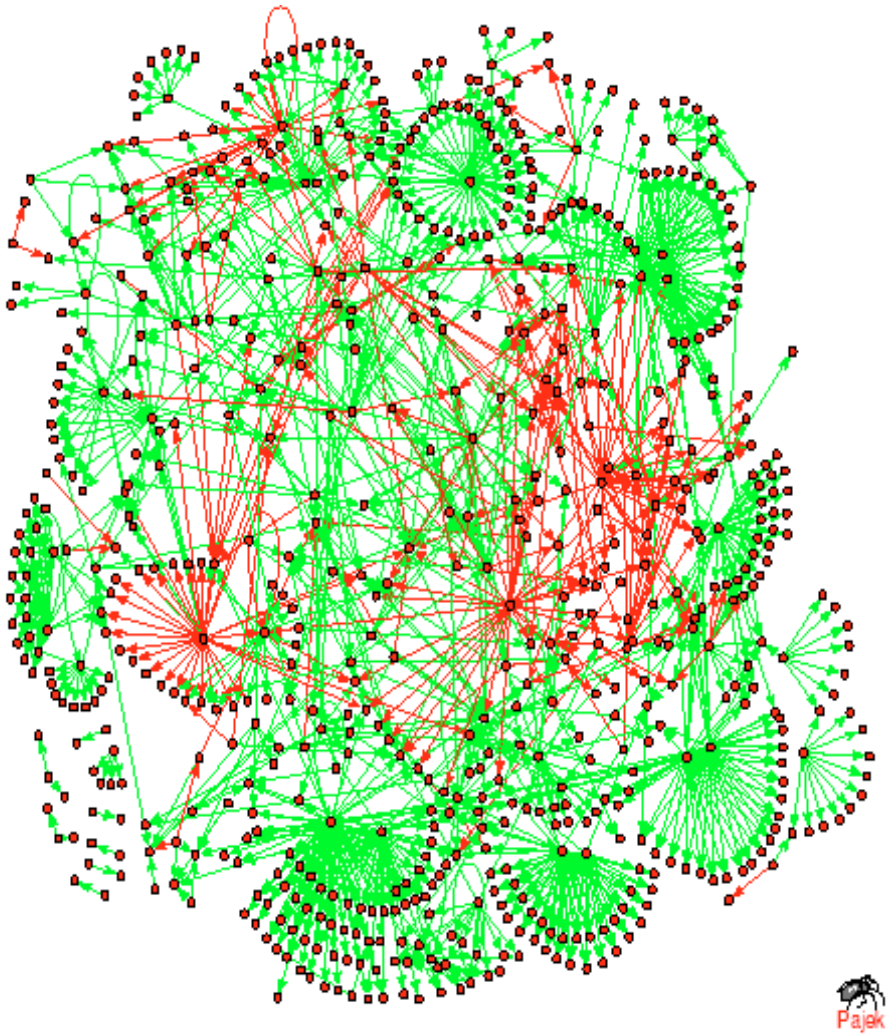




Transcription regulatory networks

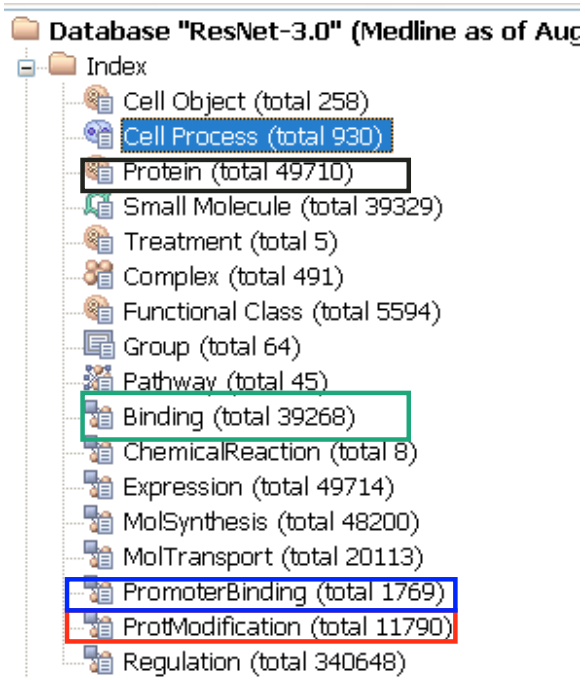


Prokaryotic bacterium:
E. coli

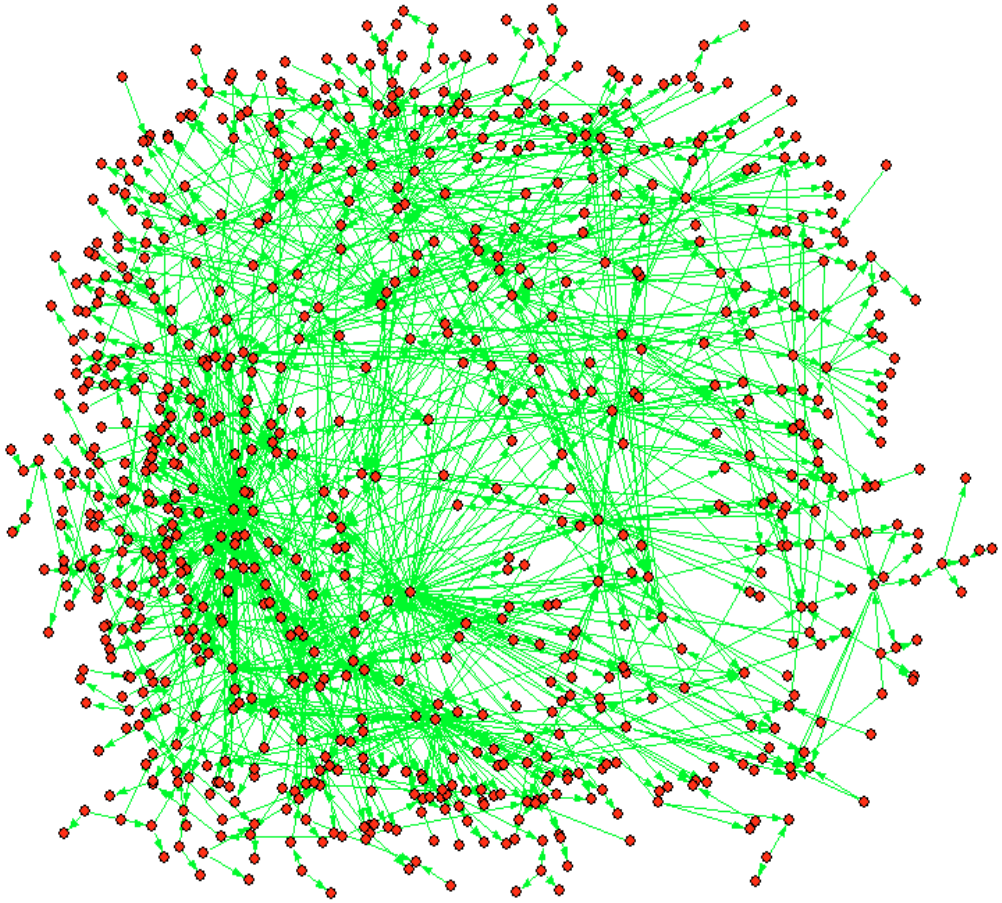


Single-celled eukaryote:
S. cerevisiae

Homo sapiens



Total: 120,000 interacting protein pairs extracted from PubMed as of 8/2004



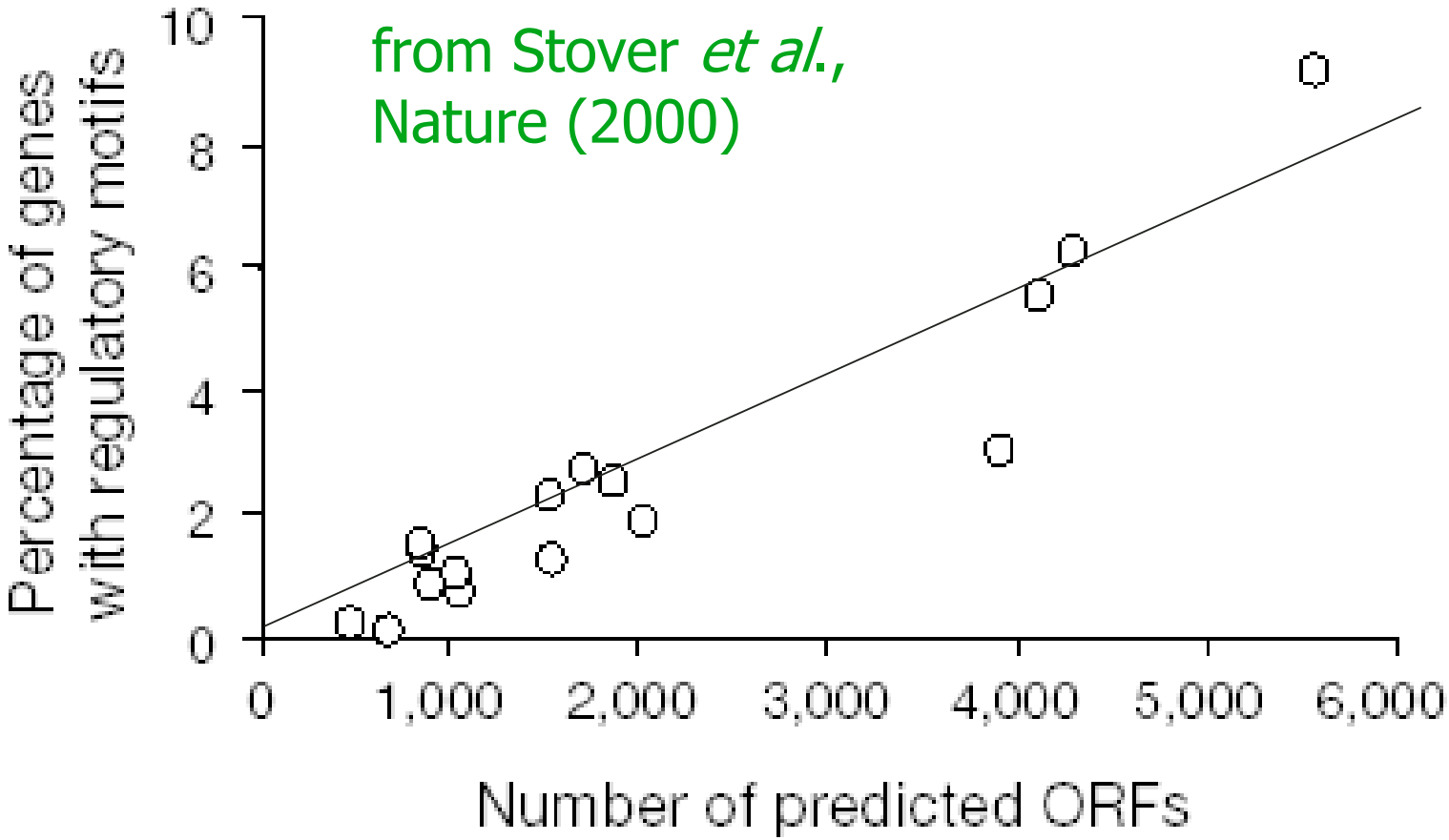
Giant component of Transcription Regulatory network: 1271 regulations 801 proteins
Data from Ariadne Genomics

General properties

- Densely interconnected: most nodes are in giant component
- Not very modular: functional units talk to each other
- Have many random features
- Few proteins (hubs) interact with a lot of neighbors: but most – with just one

How many transcriptional regulators are out there?

Fraction of transcriptional regulators in bacteria

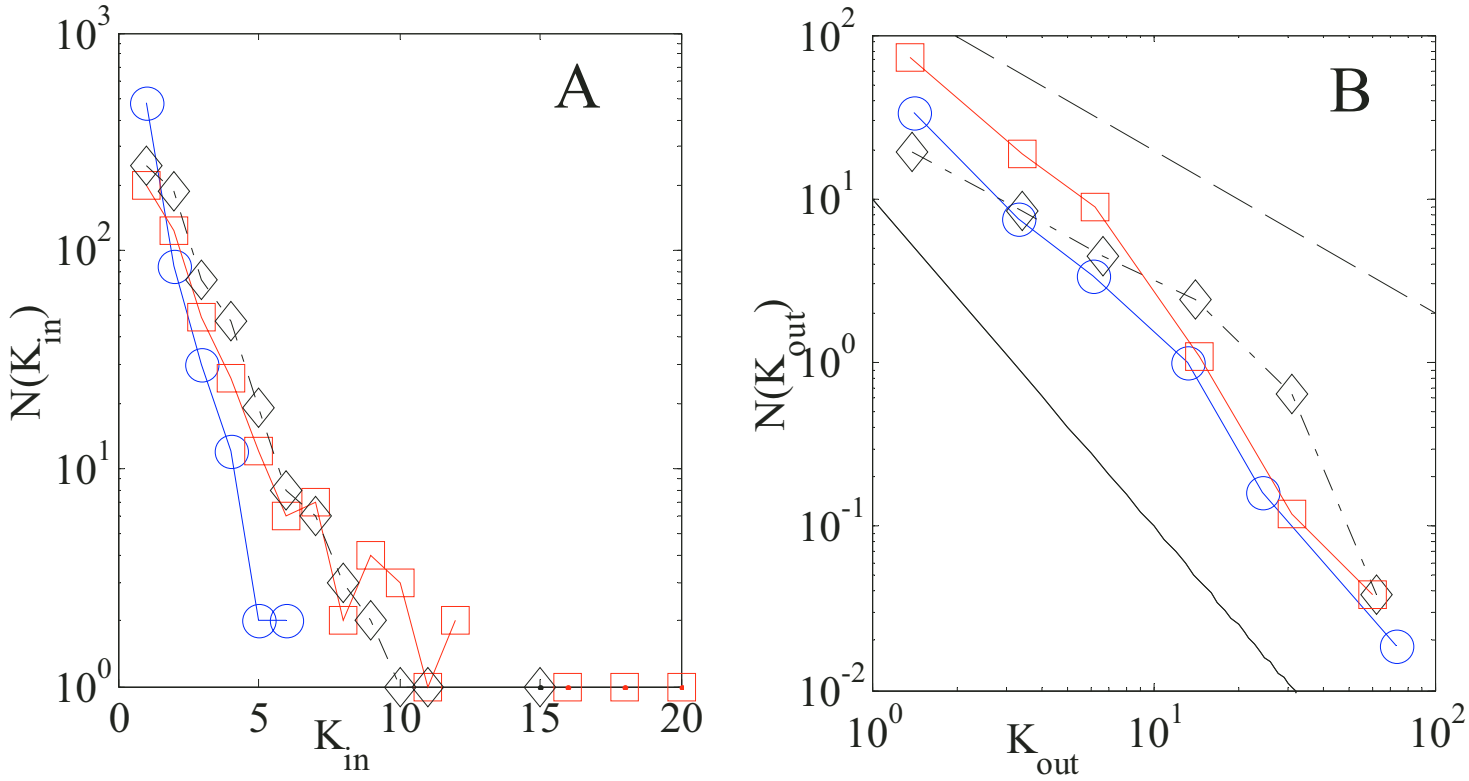


Complexity of regulation grows with complexity of organism

- $N_R<K_{out}>=N<K_{in}>$=number of edges
- $N_R/N= <K_{in}>/<K_{out}>$ increases with N
- $<K_{in}>$ grows with N
  - In bacteria $N_R \sim N^2$ (Stover, et al. 2000)
  - In eucaryots $N_R \sim N^{1.3}$ (van Nimwengen, 2002)
- Networks in more complex organisms are more interconnected then in simpler ones
- Life is not just a bunch of independent modules!

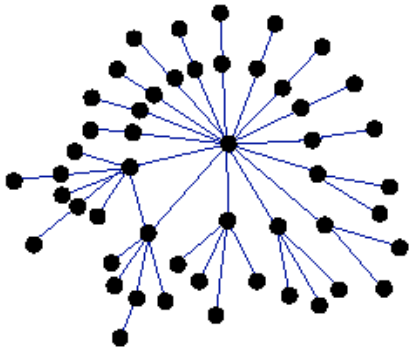
Complexity is manifested in K_{in} distribution

E. coli vs. S. cerevisiae vs. H. sapiens

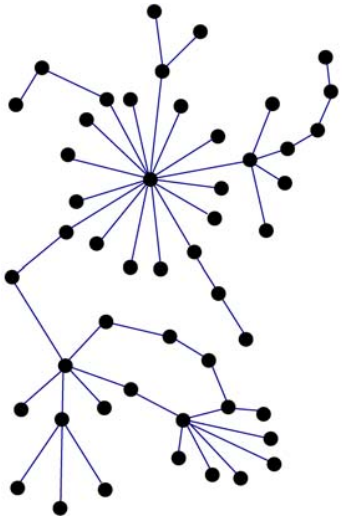


Beyond degree distributions: How is it all wired together?

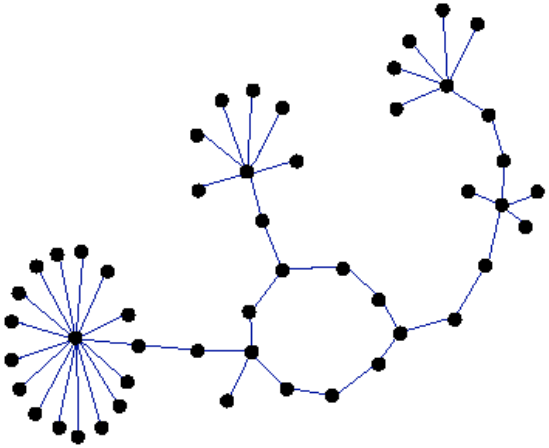
Central vs peripheral network architecture



Largest hub is in the center (very hierarchical) "assortative"



Random

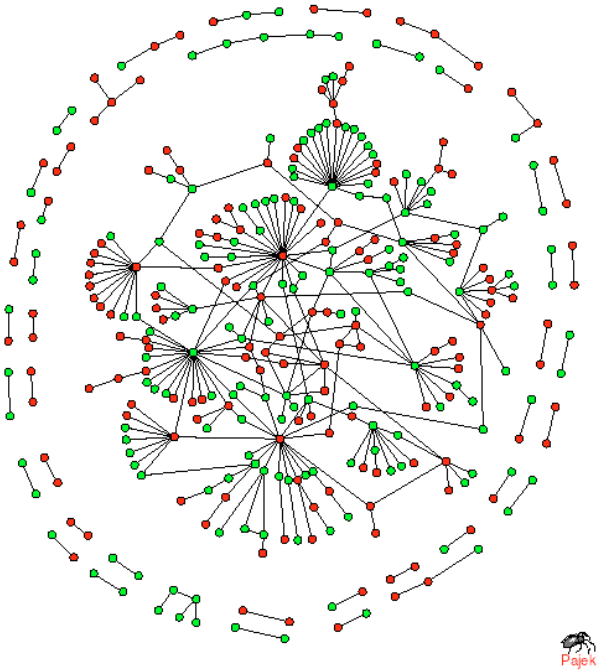


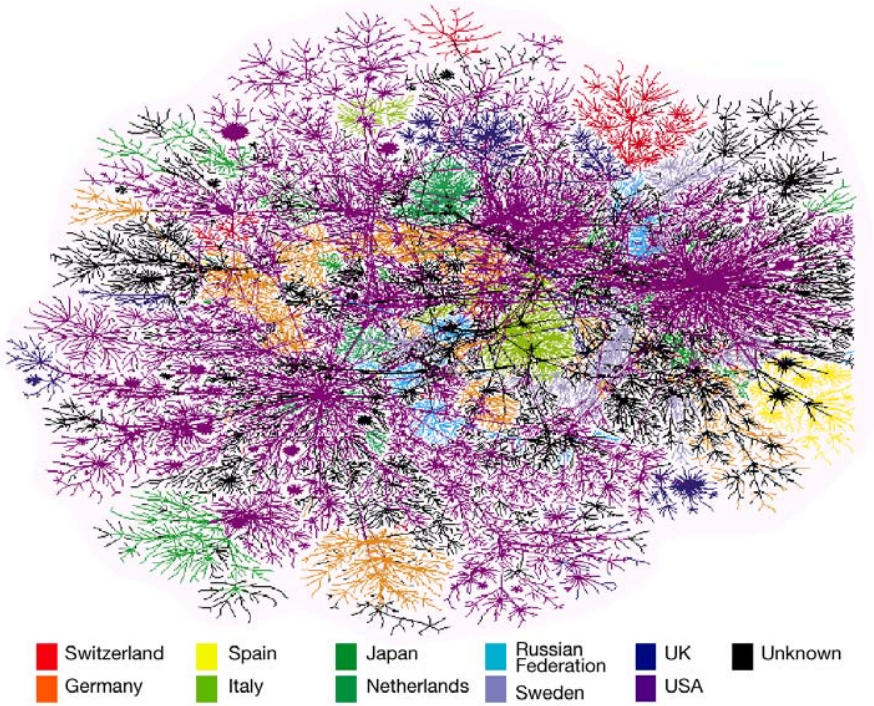
Hubs are peripheral (very anti-hierarchical) "disassortative"

Correlation profile

- Count $N(k_0,k_1)$ – the number of links between nodes with connectivities k_0 and k_1
- Compare it to $N_r(k_0,k_1)$ – the same property in a random network
- Qualitative features are very noise-tolerant with respect to both false positives and false negatives

Some scale-free networks may appear similar

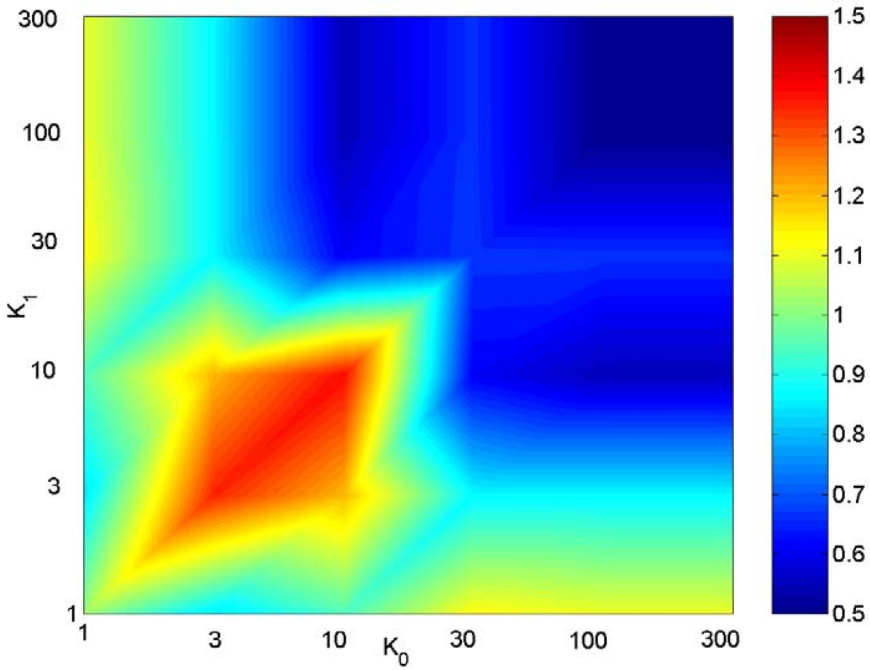




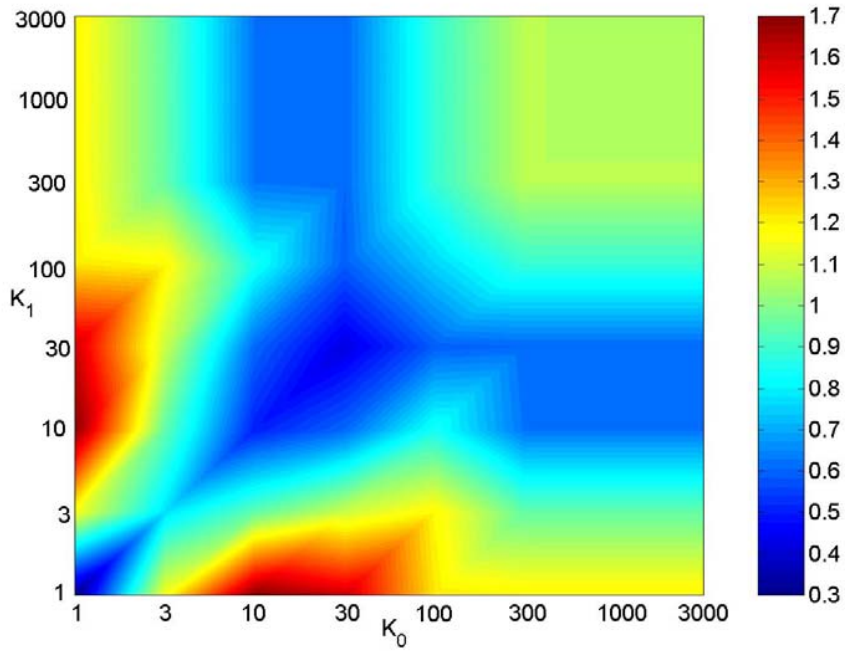
In both networks the degree distribution is scale-free $P(k) \sim k^{-\gamma}$ with $\gamma \sim 2.2\text{-}2.5$

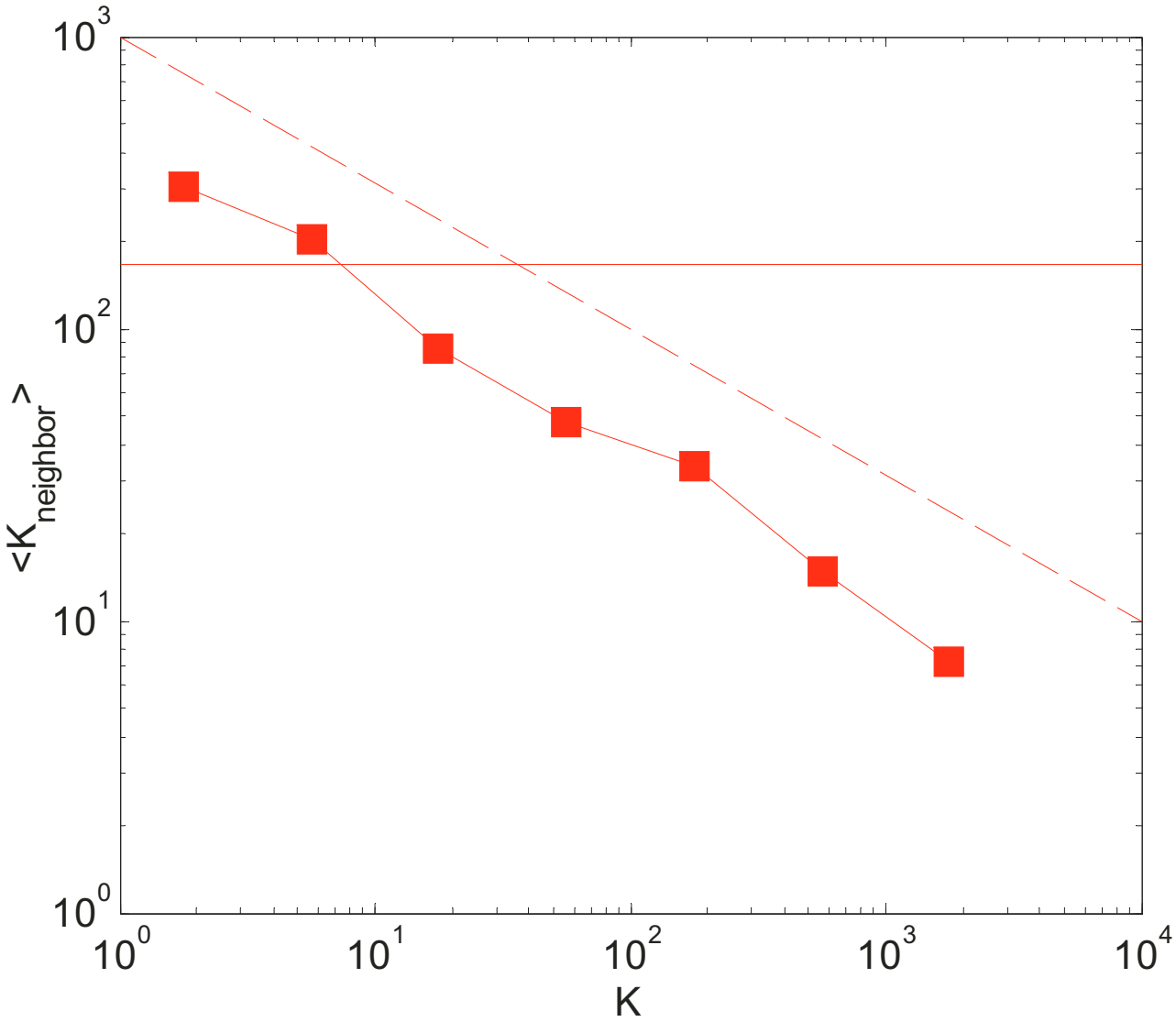
But: correlation profiles give them unique identities

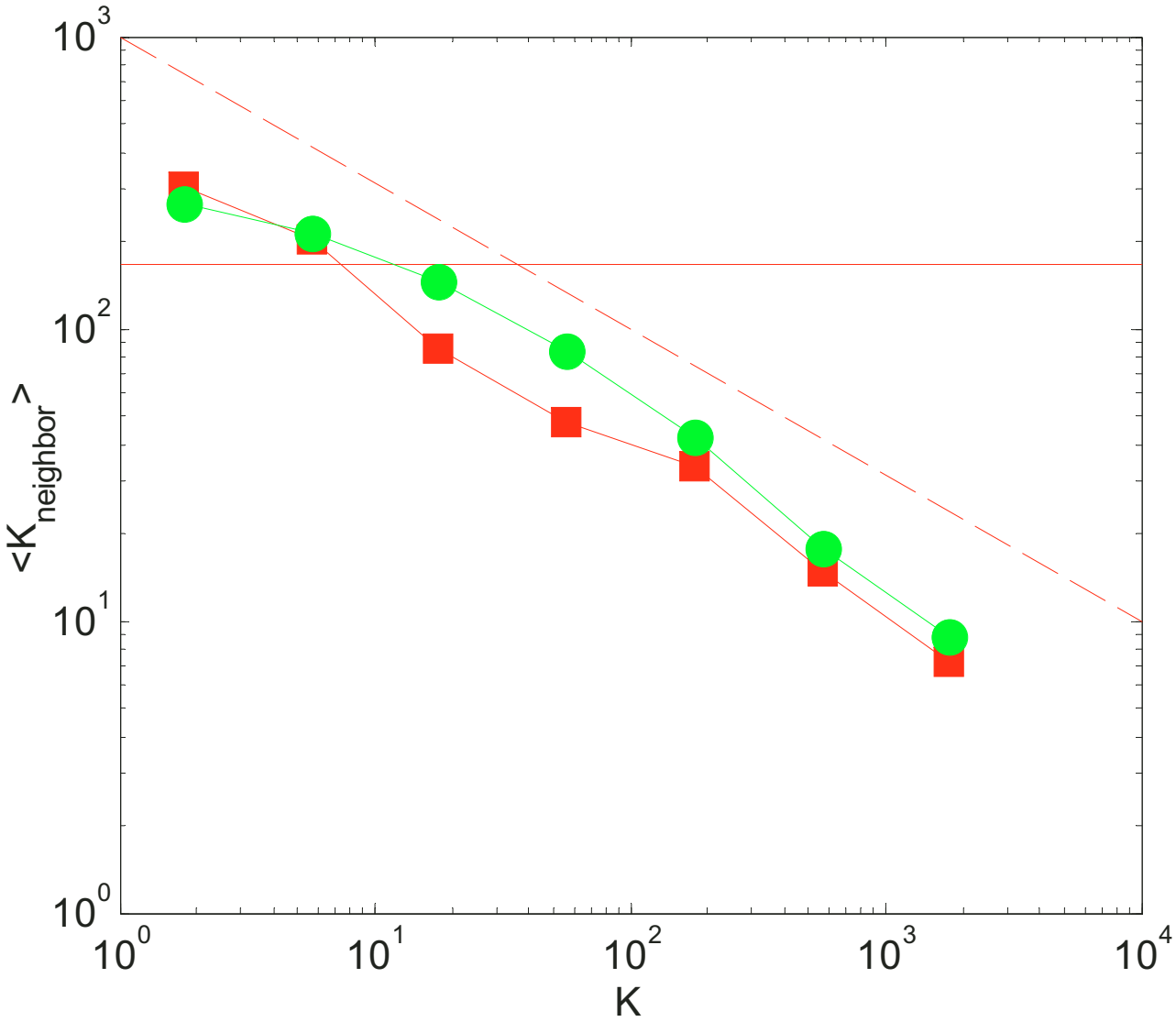
Protein interactions



Internet




10^3
10^2
10^1
10^0
$<K_{neighbor}>$
10^0
10^1
10^2
10^3
10^4
K


10^3
10^2
10^1
10^0
$<K_{neighbor}>$
10^0
10^1
10^2
10^3
10^4
K

No multiple edges

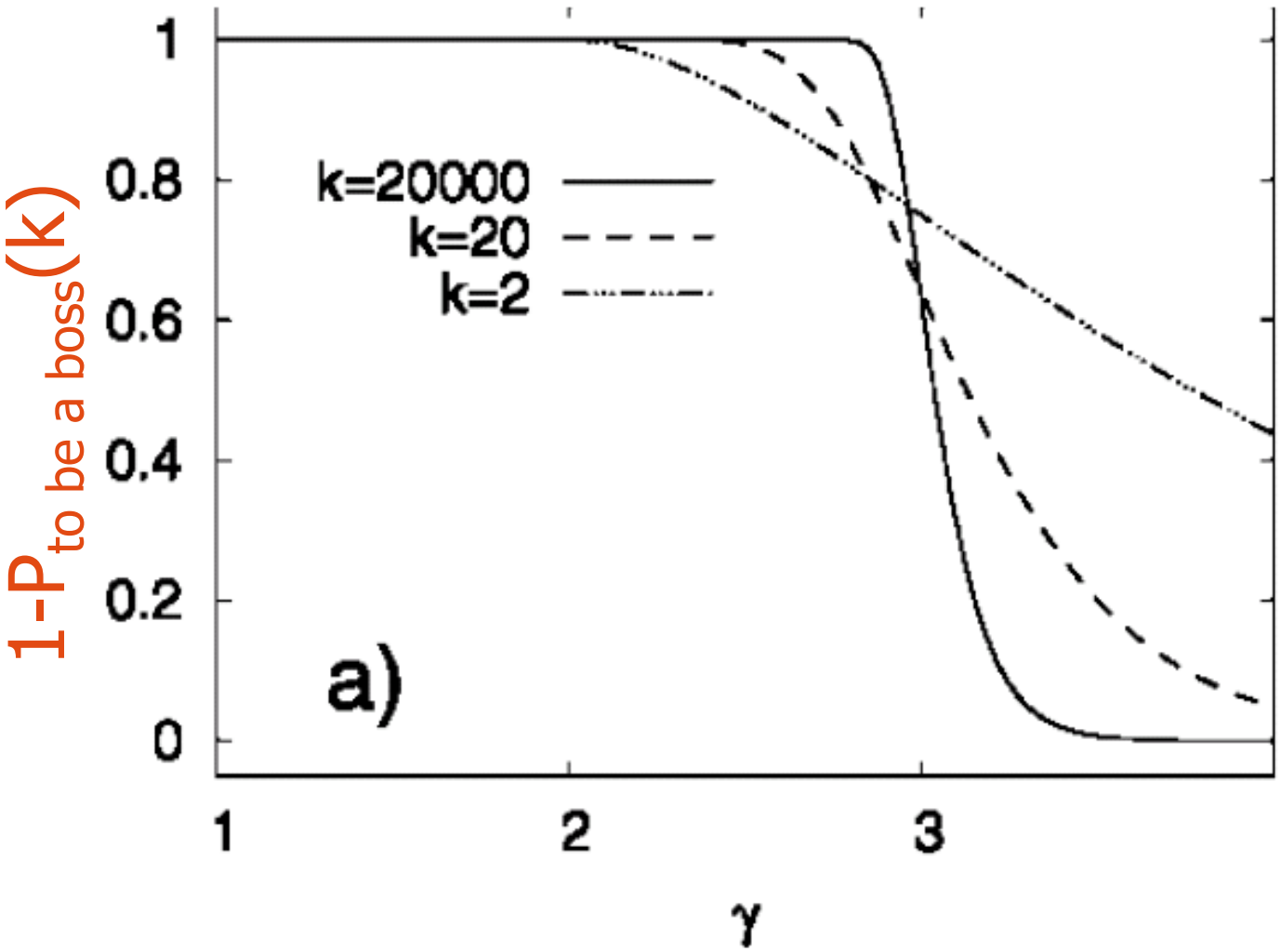
- Expected number of edges between two highest connected hubs is 1458 * 750/ (2 * 12,573)=43.5 edges!
- When constructing a random network – allow no multiple edges
- Dangerous for $\gamma<3$ (especially $\gamma\approx2$) as (# of hub-hub edges) ~ $N^{(3-\gamma)/(\gamma-1)}$

Which networks are truly hierarchical ?

- "Importance" of a node is approximated by its' degree
- If all neighbors of a node have a degree lower than itself – the node is at the top of some local hierarchy
- How many local hierarchies are out there for a random SF network with the exponent γ?

Probability to be a local boss

- $P_{\text{to be a boss}}(k)=(1- c\, k^{2-\gamma})^k \rightarrow \exp(-c\, k^{3-\gamma})$
- Some limiting cases:
  - $\gamma<3$: $P_{\text{to be a boss}}(k) \rightarrow 0$ for $k\rightarrow \infty$
  - $\gamma=3$: $P_{\text{to be boss}}(k) \rightarrow$ const for $k\rightarrow \infty$
  - $\gamma>3$: $P_{\text{to be a boss}}(k) \rightarrow 1$ for $k\rightarrow \infty$
- Thus for $\gamma>3$ – many local hierarchies (at least one per hub) for $\gamma\approx 2+\varepsilon$ - few



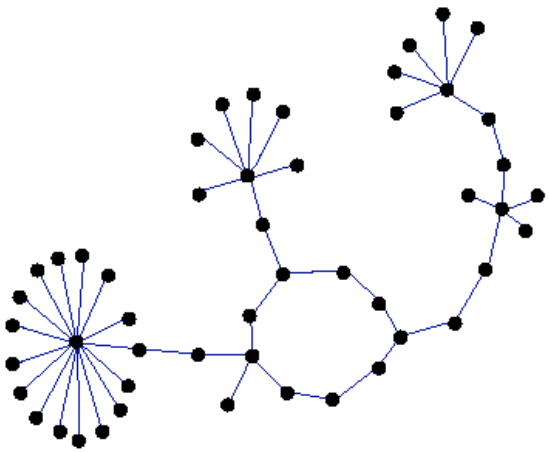
From A. Trusina, P. Minnhagen, SM, K. Sneppen, *Phys. Rev. Lett. (2004*)

How to construct a proper random network?

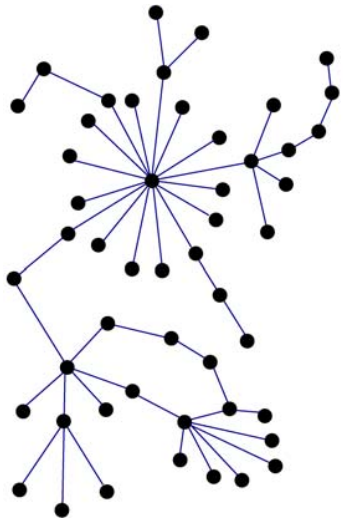
Null-model of a network

- Distribution of degrees is non-random: the degree of every node has to be conserved in a random network
- Other topological properties may be also conserved as well:
  - The extent of modularity (by function, sub-cellular localization, etc.)
  - Small motifs (e.g feed-forward loops)

Randomization

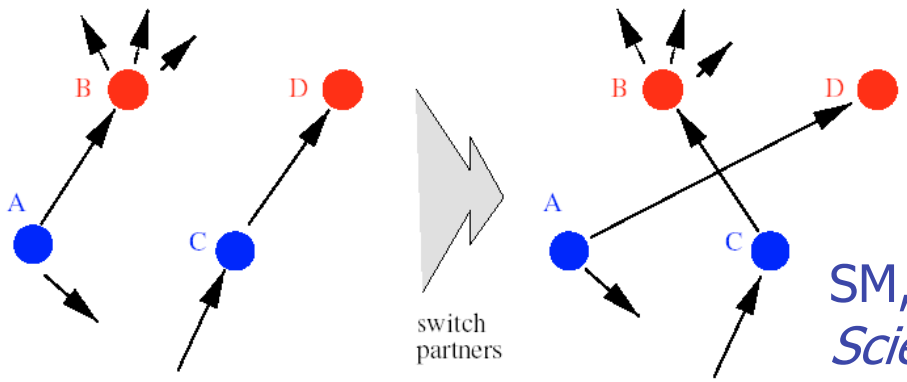


given complex network



random

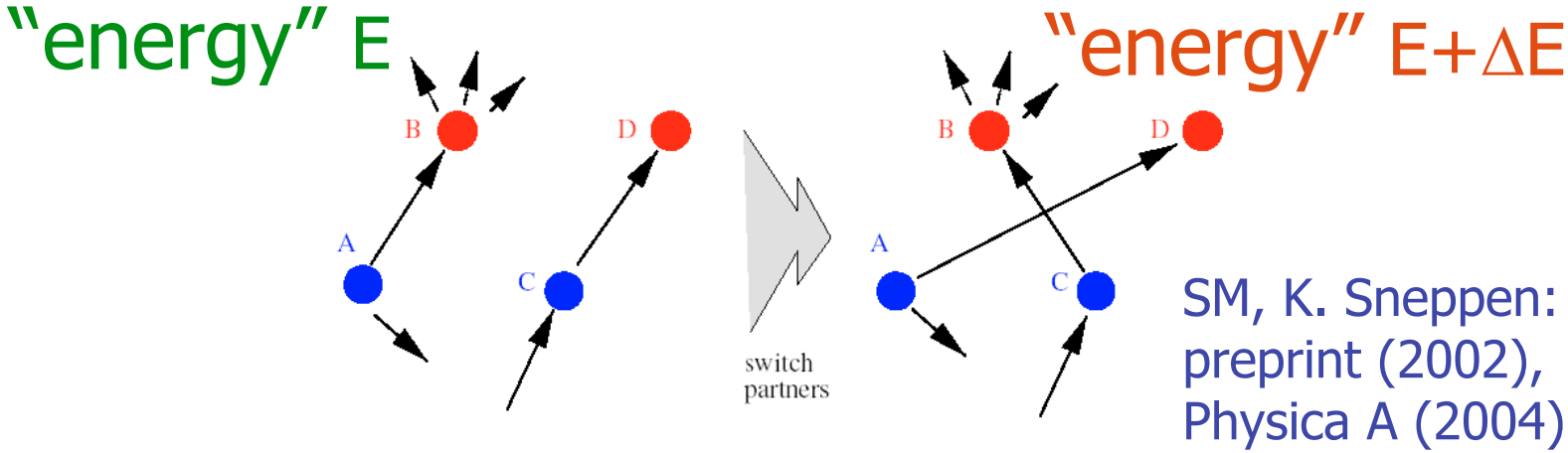
Edge swapping (rewiring) algorithm



SM, K. Sneppen, *Science* (2002)

- Randomly select and rewire two edges
- Repeat many times

Metropolis rewiring algorithm



SM, K. Sneppen: preprint (2002), Physica A (2004)

- Randomly select two edges
- Calculate change ΔE in "energy function" $E=(N_{actual}-N_{desired})^2/N_{desired}$
- Rewire with probability $p=\exp(-\Delta E/T)$

Network properties of self-binding proteins

AKA homodimers

I. Ispolatov, A. Yuryev, I. Mazo, SM
q-bio.GN/0501004.

There are just TOO MANY homodimers

	N_{dimer}	$\langle k \rangle$
yeast	179	6.6 ± 0.2
worm	89	3.3 ± 0.1
fly	160	5.9 ± 0.1
human	1045	5.7 ± 0.1

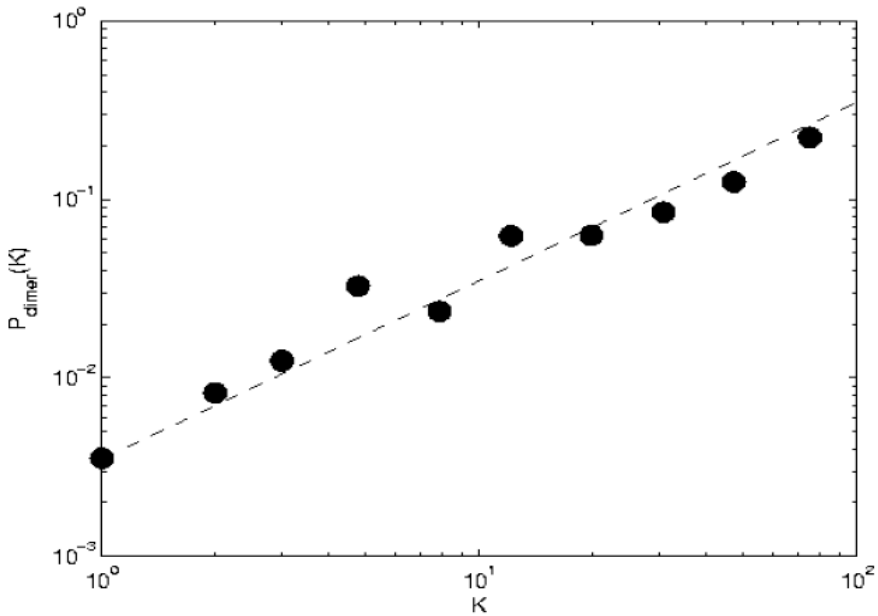
- Null-model
- $P_{self} \sim \langle k \rangle / N$
- $N_{dimer} = N \bullet P_{self} = \langle k \rangle$
- Not surprising as homodimers have many functional roles

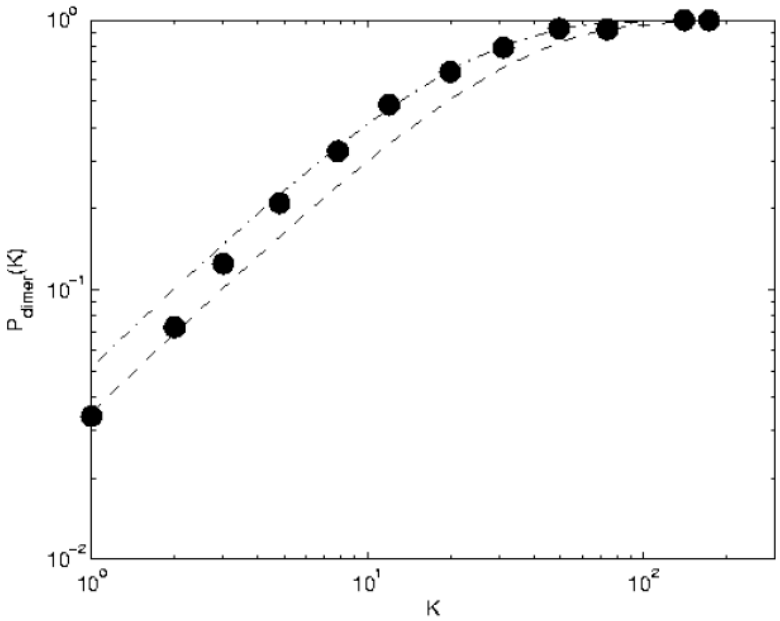
Network properties around homodimers

	$\langle k \rangle$	$\langle k \rangle_{\text{dimer}}$
yeast	6.6 ± 0.2	12.4 ± 1.2
worm	3.3 ± 0.1	13.1 ± 2.2
fly	5.9 ± 0.1	14.2 ± 1.2
human	5.7 ± 0.1	14.0 ± 0.6

Likelihood to self-interact vs. K

$$P_{dimer}(k) = 1 - (1 - p_{self})^k$$





Fly: two-hybrid data
P_{self}~0.003, P_{others}~0.0002

Human: database data
P_{self}~0.05, P_{others}~0.0002

What we think it means?

- In random networks $p_{dimer}(K) \sim K^2$ not $\sim K$ like our empirical observation
- K is proportional to the "stickiness" of the protein which in its turn scales with
  - the area of hydrophobic residues on the surface
  - # copies/cell
  - its' popularity (in datasets taken from databases)
  - etc.
- Real interacting pair consists of an "active" and "passive" protein and binding probability scales only with the "stickiness" of the active protein
- "Stickiness" fully accounts for higher than average connectivity of homodimers

Postdoc position

- Looking for a postdoc to work in my group at Brookhaven National Laboratory in New York starting Fall 2005
- Topic - large-scale properties of (mostly) bionetworks (partially supported by a NIH/NSF grant with Ariadne Genomics)
- E-mail CV and 3 letters of recommendation to: maslov@bnl.gov
- Talk to me while I am here!

Collaborators:

- Kim Sneppen – U. of Copenhagen
- Hierarchy:
  - Ala Trusina – Nordita and U. of Umea
  - Petter Minnhagen – Nordita
- Evolution:
  - Koon-Kiu Yan – Stony Brook
  - Kasper Eriksen – U. of Lund
- Homodimers:
  - Slava Ispolatov, Ilya Mazo, Anton Yuriev – Ariadne Genomics