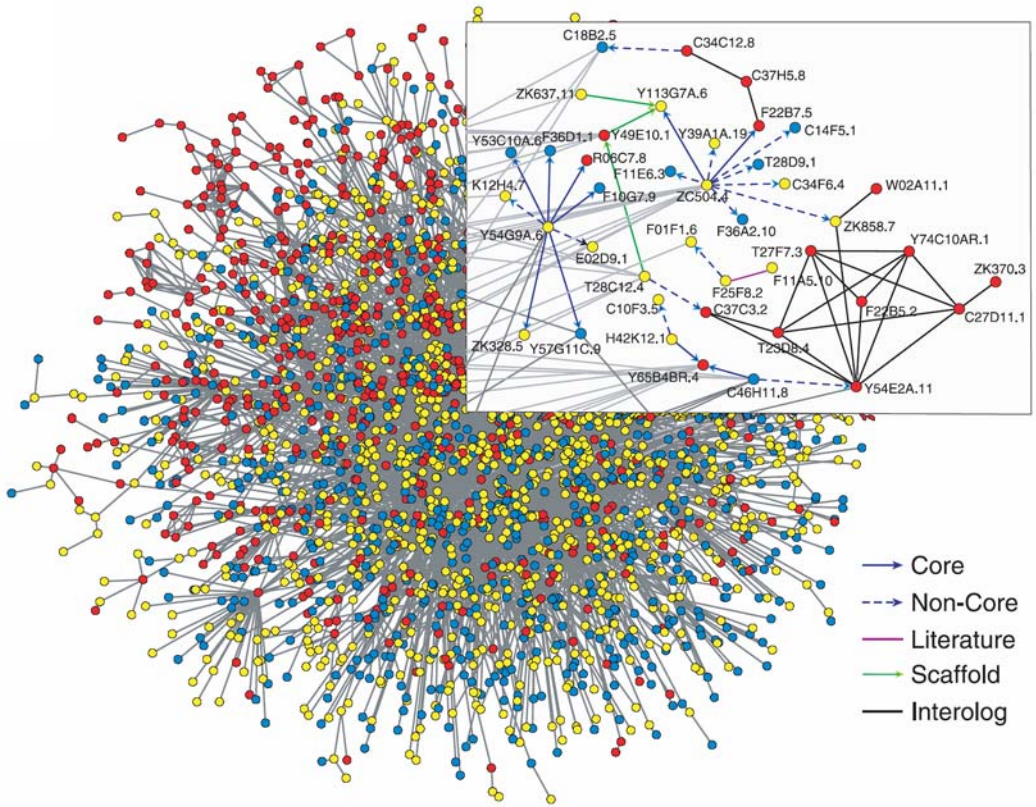
Protein Binding Networks: from Topology to Kinetics

Sergei Maslov

Brookhaven National Laboratory

Genome-wide protein binding networks



- Nodes - proteins
- Edges - protein-protein binding interactions
- Functions
  - structural
  - complexes/dimers
  - regulation/signaling
  - unknown?
  - etc

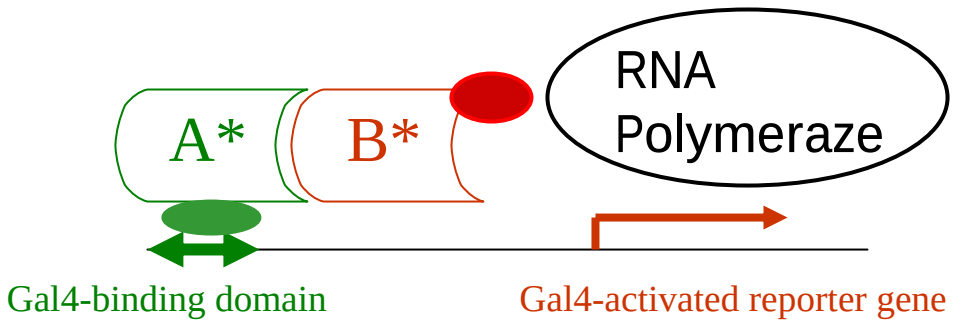
C. elegans PPI from
Li et al. (Vidal's lab), Science (2004)

How much data is out there?

Species	Set	nodes	edges	# of sources
S.cerevisiae	HTP-PI	4,500	13,000	5
	LC-PI	3,100	20,000	3,100
D.melanogaster	HTP-PI	6,800	22,000	2
C.elegans	HTP-PI	2,800	4,500	1
H.sapiens	LC-PI	6,400	31,000	12,000
	HTP-PI	1,800	3,500	2
H. pylori	HTP-PI	700	1,500	1
P. falciparum	HTP-PI	1,300	2,800	1

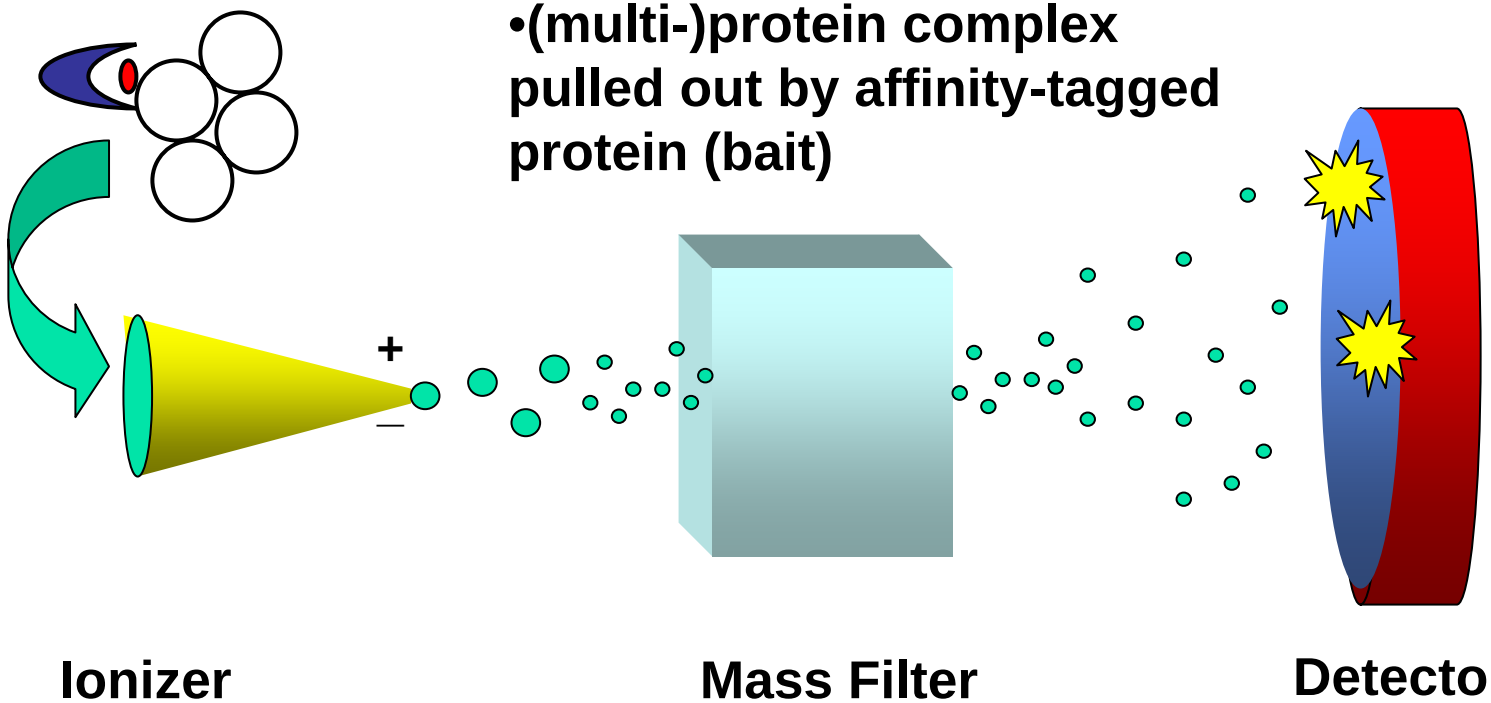
Yeast two-hybrid technique

uses two “hybrid proteins”: bait A* (A fused with Gal4p DNA-binding domain) and prey B* (B fused with Gal4p activation domain)



- Cons: wrong (very high) concentrations, localization (unless both proteins are nuclear), and even host organism (unless done in yeast)
- Pros: direct binding events
- Main source of noise: self-activating baits

Affinity capture + Mass Spectrometry



- Pros: in vivo concentrations and localizations
- Cons: binding interactions are often indirect
- Main source of noise: highly abundant and sticky proteins

Breakup by experimental technique in yeast

BIOGRID database	*S. cerevisiae*
Affinity Capture-Mass Spec	28172
Affinity Capture-RNA	55
Affinity Capture-Western	5710
Co-crystal Structure	107
FRET	43
Far Western	41
Two-hybrid	11935
Total	46063

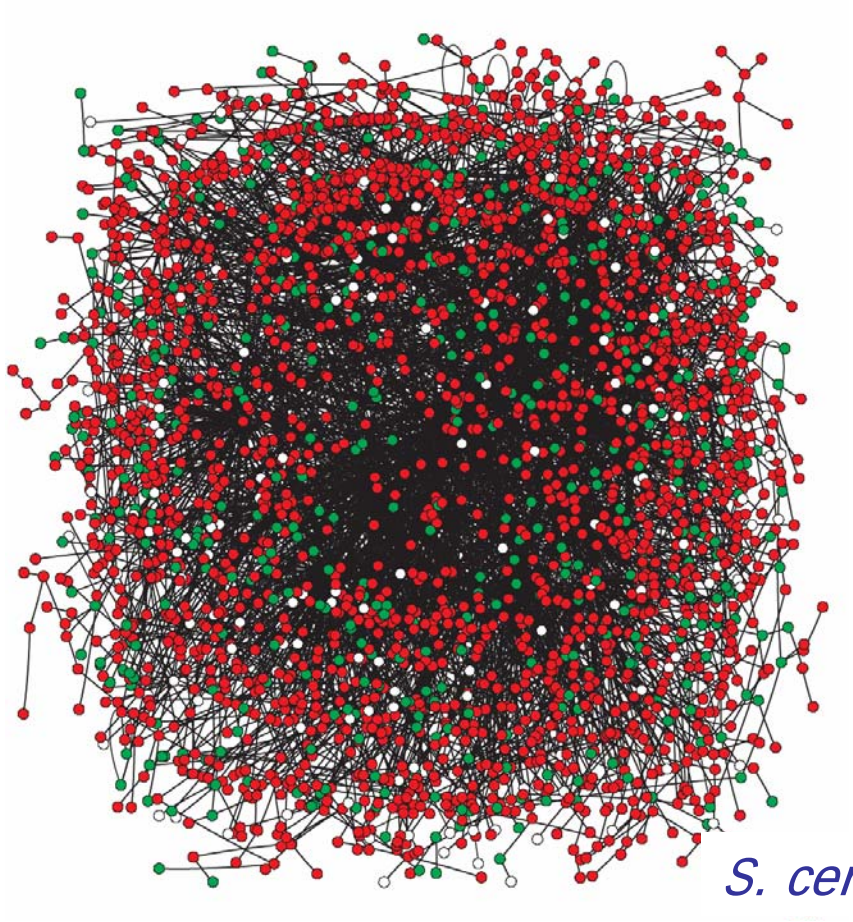
What are the common topological features?

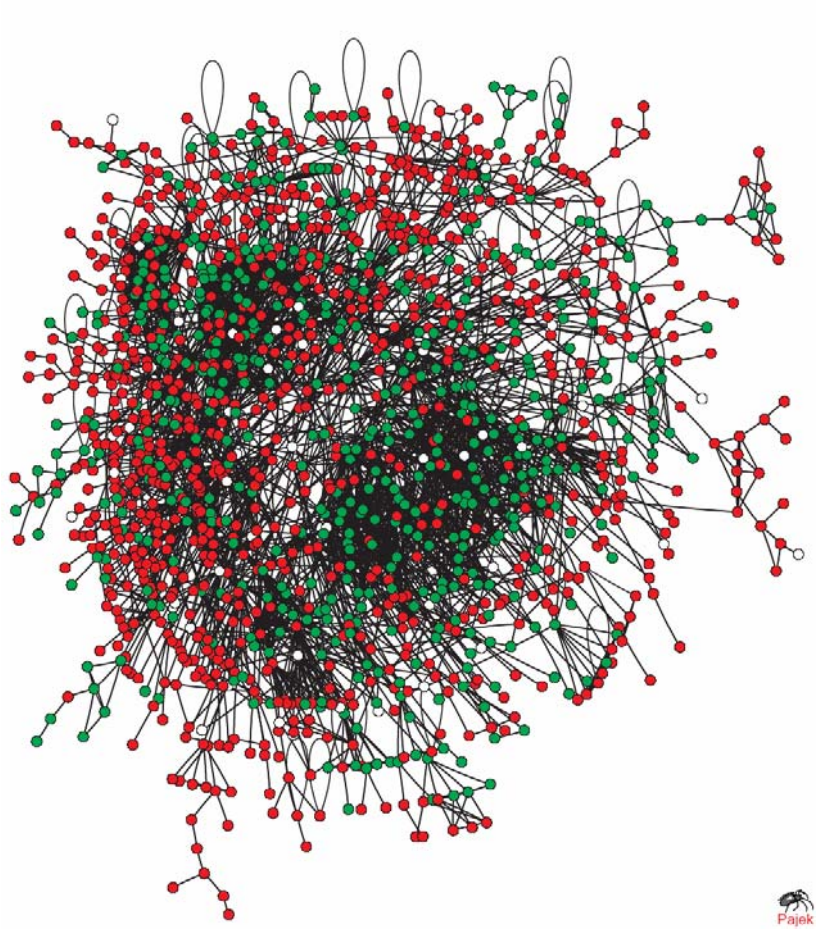
1. Broad distribution of the number of interaction partners (degree K) of individual proteins
2. Anti-correlation of degrees of interacting proteins
3. Small-world-property
   (follows from 1. for $<K^2>/<K>>2$)

Protein binding networks have small-world property

86% of proteins could be connected

83% in this plot





Large-scale Y2H experiment

Curated dataset from our study

Why small-world matters?

- Claims of "robustness" of this network architecture come from studies of the Internet where breaking up the network is a disaster
- For PPI networks it is the OPPOSITE: interconnected networks present a problem
- In a small-world network equilibrium concentrations of all proteins are coupled to each other
- Danger of undesirable cross-talk

Going beyond topology and modeling the equilibrium and kinetics

SM, K. Sneppen, I. Ispolatov, q-bio/0611026; SM, I. Ispolatov, subm. (2007)

What is needed to model?

- A reliable network of reversible (non-catalytic) protein-protein binding interactions
  - √ CHECK! e.g. physical interactions between yeast proteins in the BIOGRID database with 2 or more citations
- Total concentrations and sub-cellular localizations of all proteins
  - √ CHECK! genome-wide data for yeast in 3 Nature papers (2003, 2003, 2006) by the group of J. Weissman @ UCSF
  - Left us with 1700 yeast proteins and ~5000 interactions
- *in vivo* dissociation constants K_{ij}
  - OOPS! ☹. High throughput experimental techniques are not there yet

Let's hope it doesn't matter

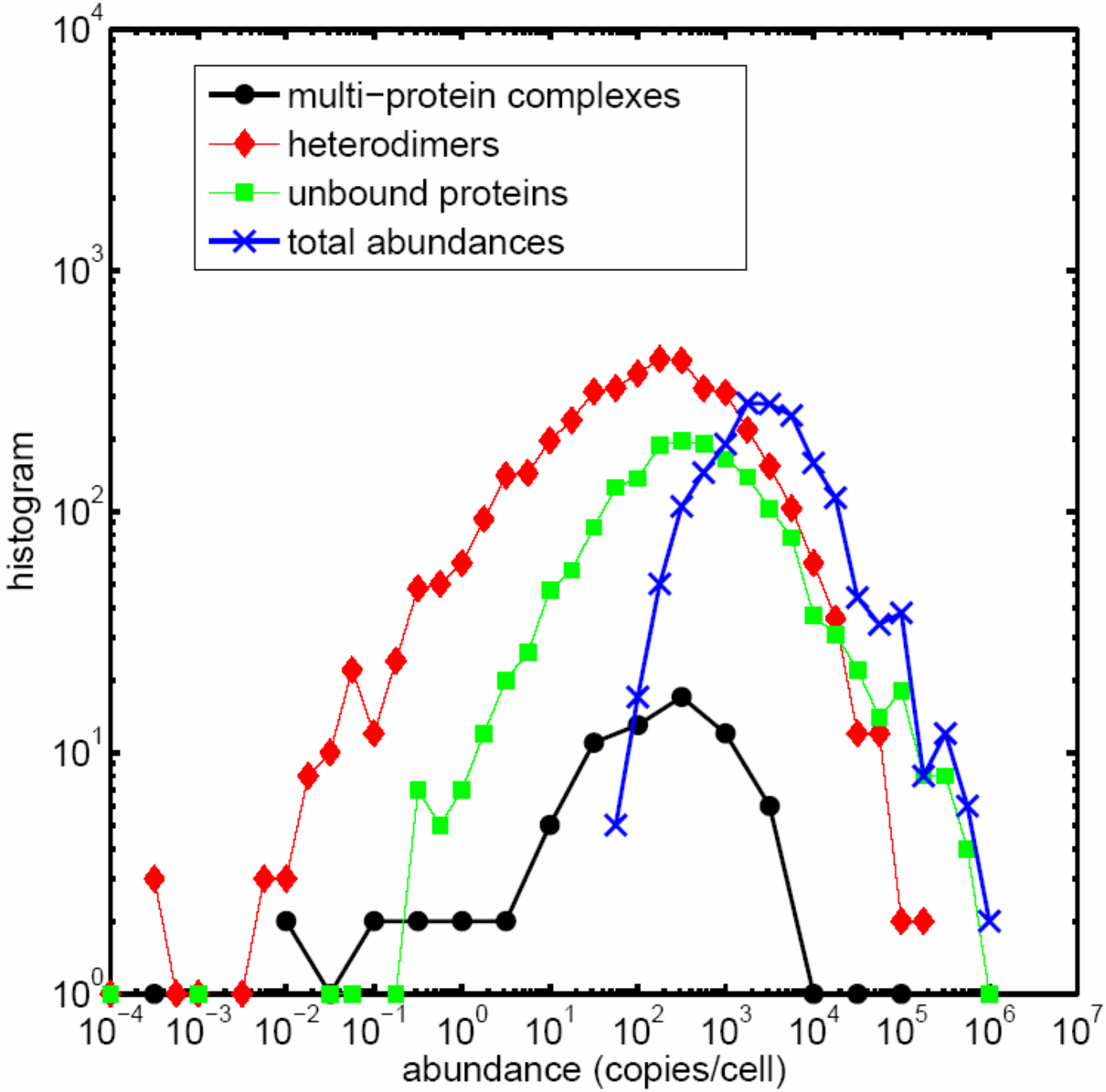
- The overall binding strength from the PINT database: $<1/K_{ij}>=1/(5nM)$. In yeast: 1nM ~ 34 molecules/cell
- Simple-minded assignment $K_{ij}=const=10nM$ (also tried 1nM, 100nM and 1000nM)
- Evolutionary-motivated assignment: $K_{ij}=max(C_i,C_j)/20$: K_{ij} is only as small as needed to ensure binding
- All assignments of a given average strength give ROUGHLY THE SAME RESULTS

Law of Mass Action

- $dD_{AB}/dt = r_{on} F_A F_B - r_{off} D_{AB}$
- In equilibrium $D_{AB}=F_A F_B/K_{AB}$ where the dissociation constant $K_{AB}= r_{off}/r_{on}$ has units of concentration
- Total concentration = free concentration + bound concentration→ $C_A= F_A+F_A F_B/K_{AB}$; $C_B= F_B+F_A F_B/K_{AB}$
- $F_A=C_A/(1+F_B/K_{AB})$; $F_B=C_B/(1+F_B/K_{AB})$

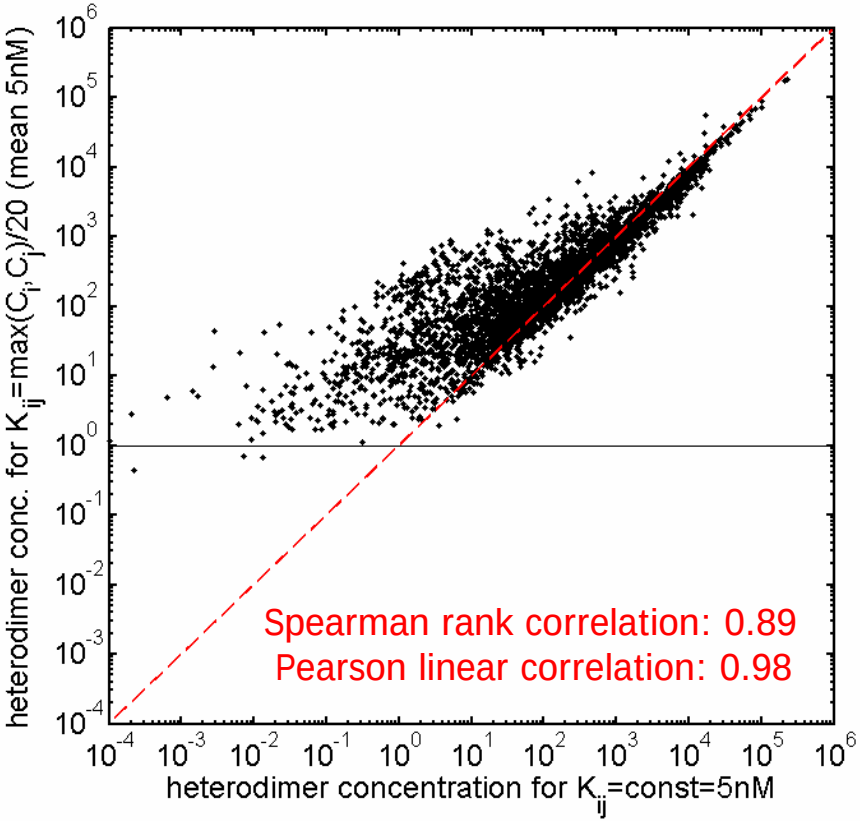
Law of Mass Action equilibrium of a PPI network

- In a network $F_i = C_i / (1 + \sum_{\text{neighbors } j} F_j / K_{ij})$
- Even though it cannot be solved analytically it is easily solved numerically e.g. by iterations
- We use experimentally measured total concentrations C_i to calculate all unbound (free) F_i and all bound $D_{ij} = F_i F_j / K_{ij}$ concentrations

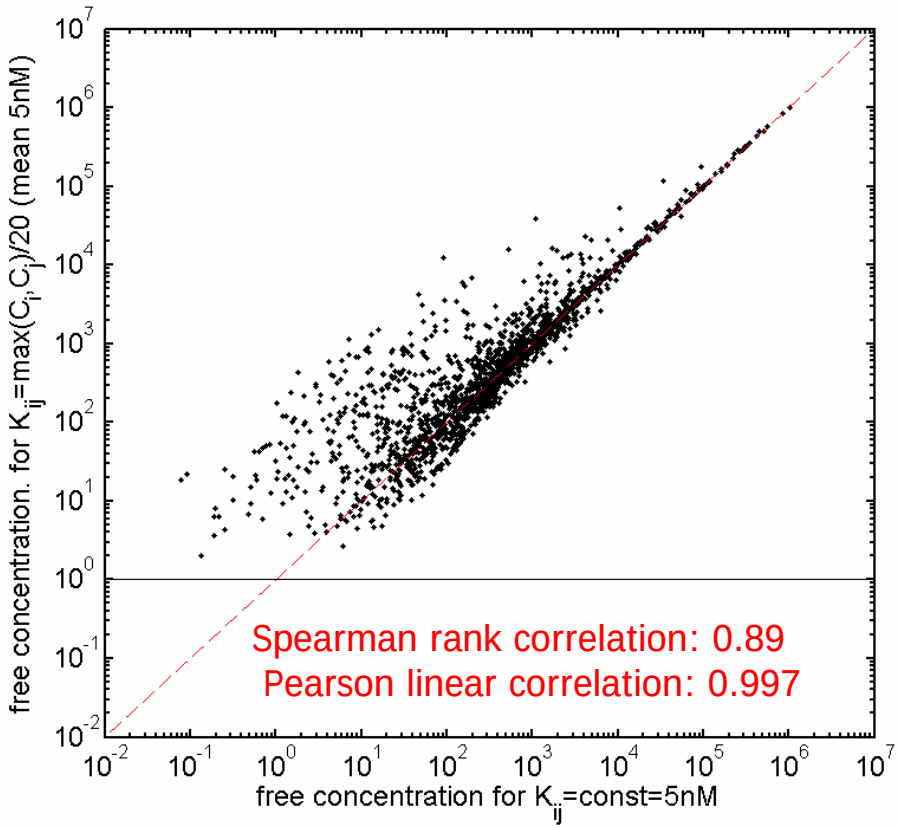

multi−protein complexes
heterodimers
unbound proteins
total abundances
histogram
abundance (copies/cell)
10^0
10^1
10^2
10^3
10^4
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7

Robustness with respect to assignment of K_{ij}

Bound concentrations: D_{ij}



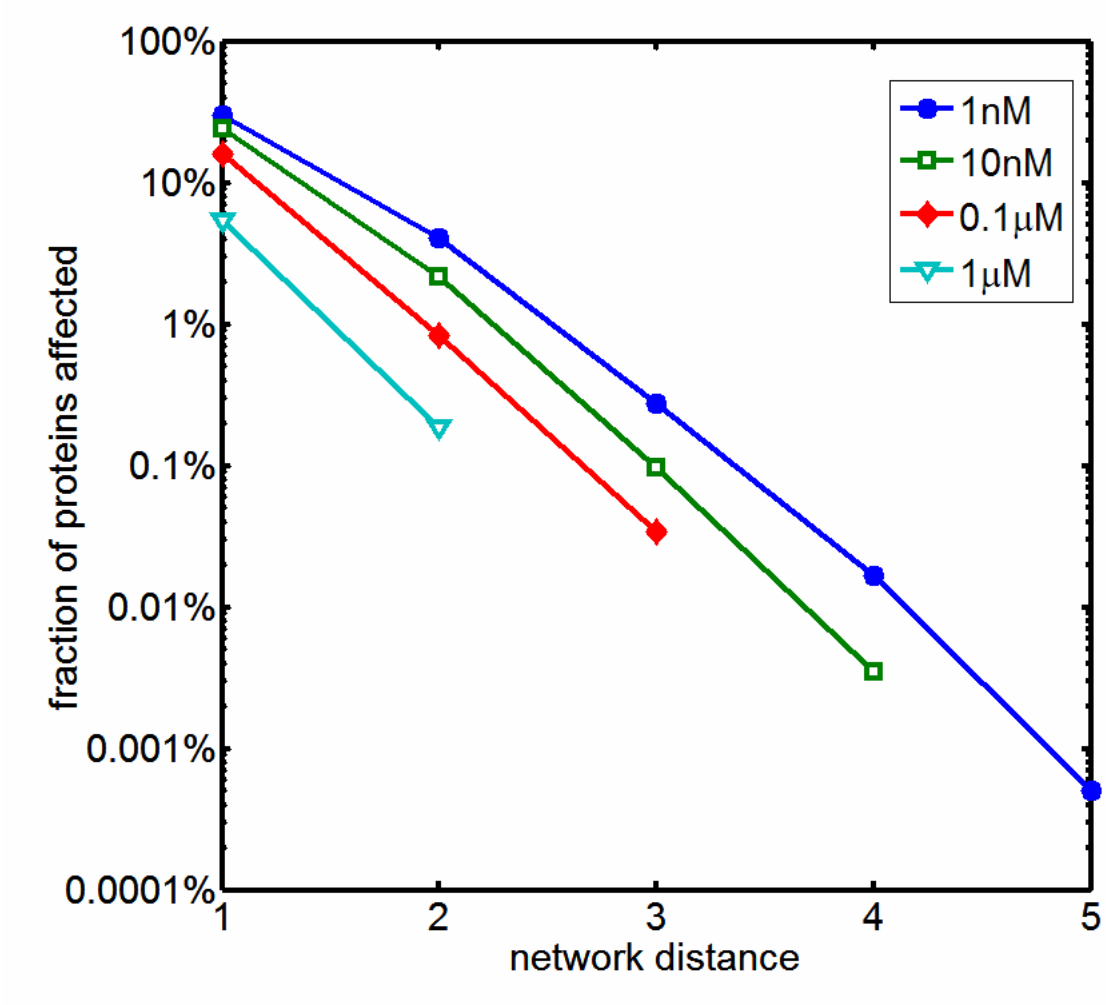
Free concentrations: F_i

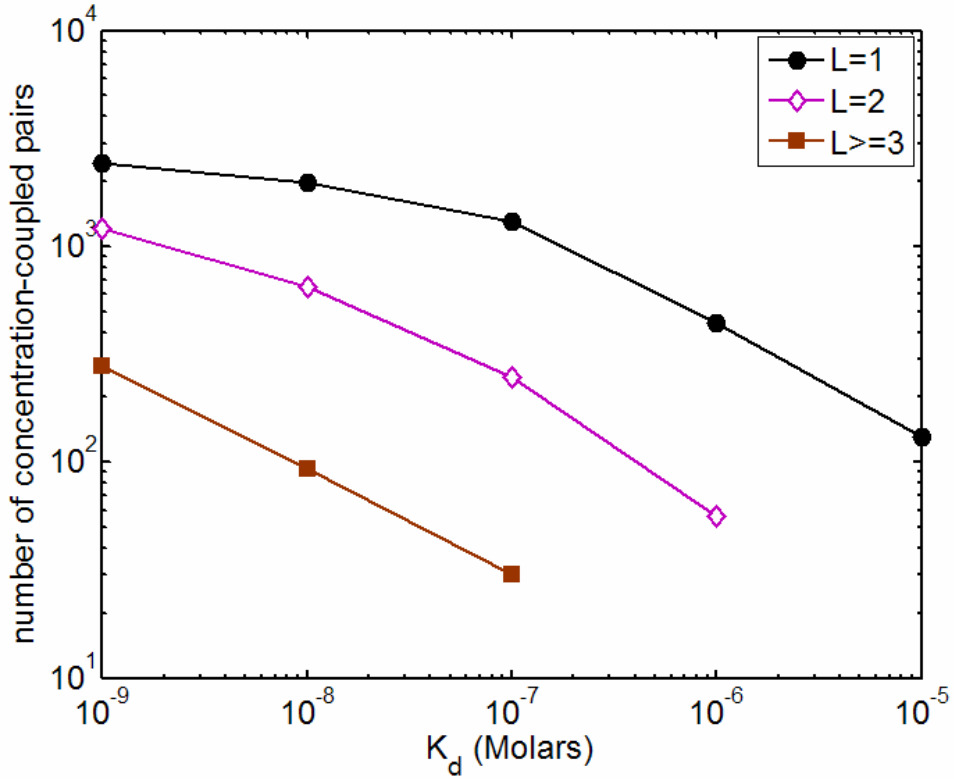


Numerical study of propagation of perturbations

- We simulate a twofold increase of the abundance C_0 of just one protein
- Proteins whose free concentration F_i changes by >20% are considered to be significantly perturbed.
- We refer to such proteins i as concentration-coupled to the protein 0
- Look for cascading perturbations: changes in the total concentration C_0 of P_0 affects F_1 of its binding partner P_1, which in turn affects F_2 of its partner P_2, etc.

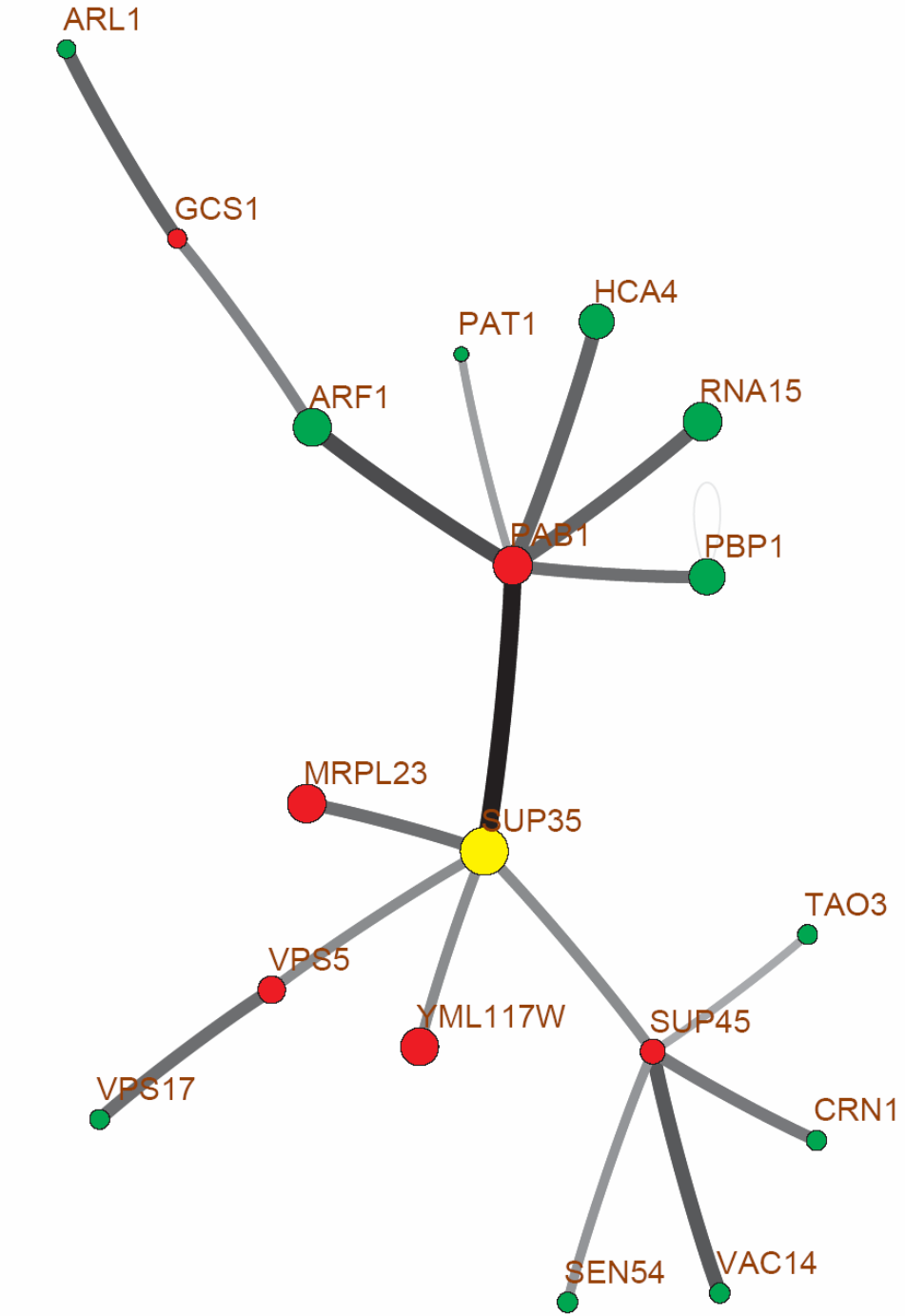
Indiscriminate cross-talk is suppressed

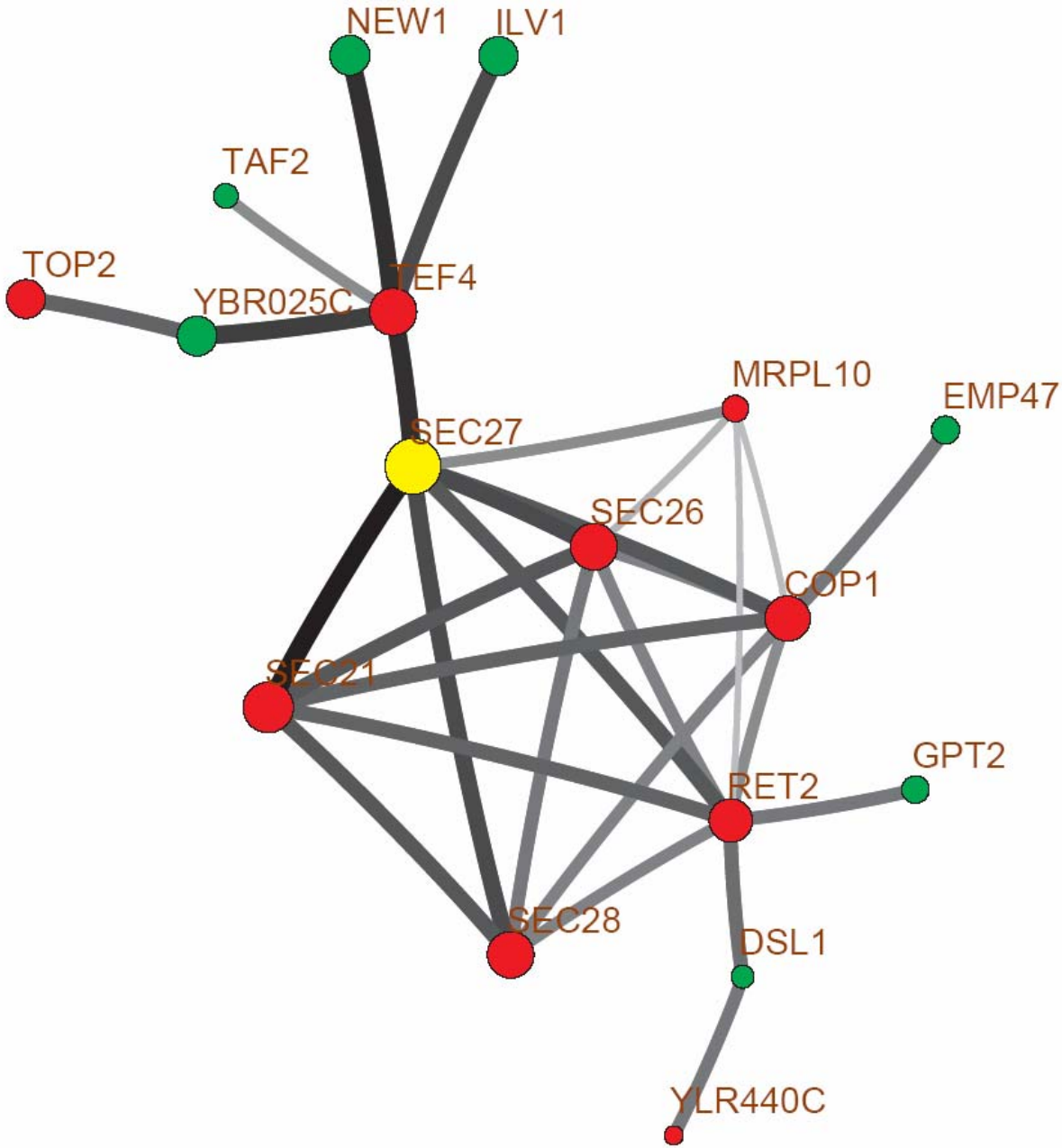




L	variable K_{ij}, mean= 5nM	constant $K_{ij} = 1$nM	constant $K_{ij} = 10$nM	constant $K_{ij} = 0.1\mu$M	constant $K_{ij} = 1\mu$M	all pairs at distance L
1	2003	2469	1915	1184	387	8168
2	415	1195	653	206	71	29880
3	15	159	49	8	0	87772
4	2	60	19	0	0	228026
5	0	3	0	0	0	396608

SM, I. Ispolatov, submitted (2007)


ARL1
GCS1
HCA4
PAT1
ARF1
RNA15
PAB1
PBP1
MRPL23
SUP35
TAO3
VPS5
YML117W
SUP45
VPS17
CRN1
SEN54
VAC14


NEW1
ILV1
TAF2
TOP2
YBR025C
TEF4
MRPL10
EMP47
SEC27
SEC26
COP1
SEC21
GPT2
RET2
SEC28
DSL1
YLR440C

What conditions
make some

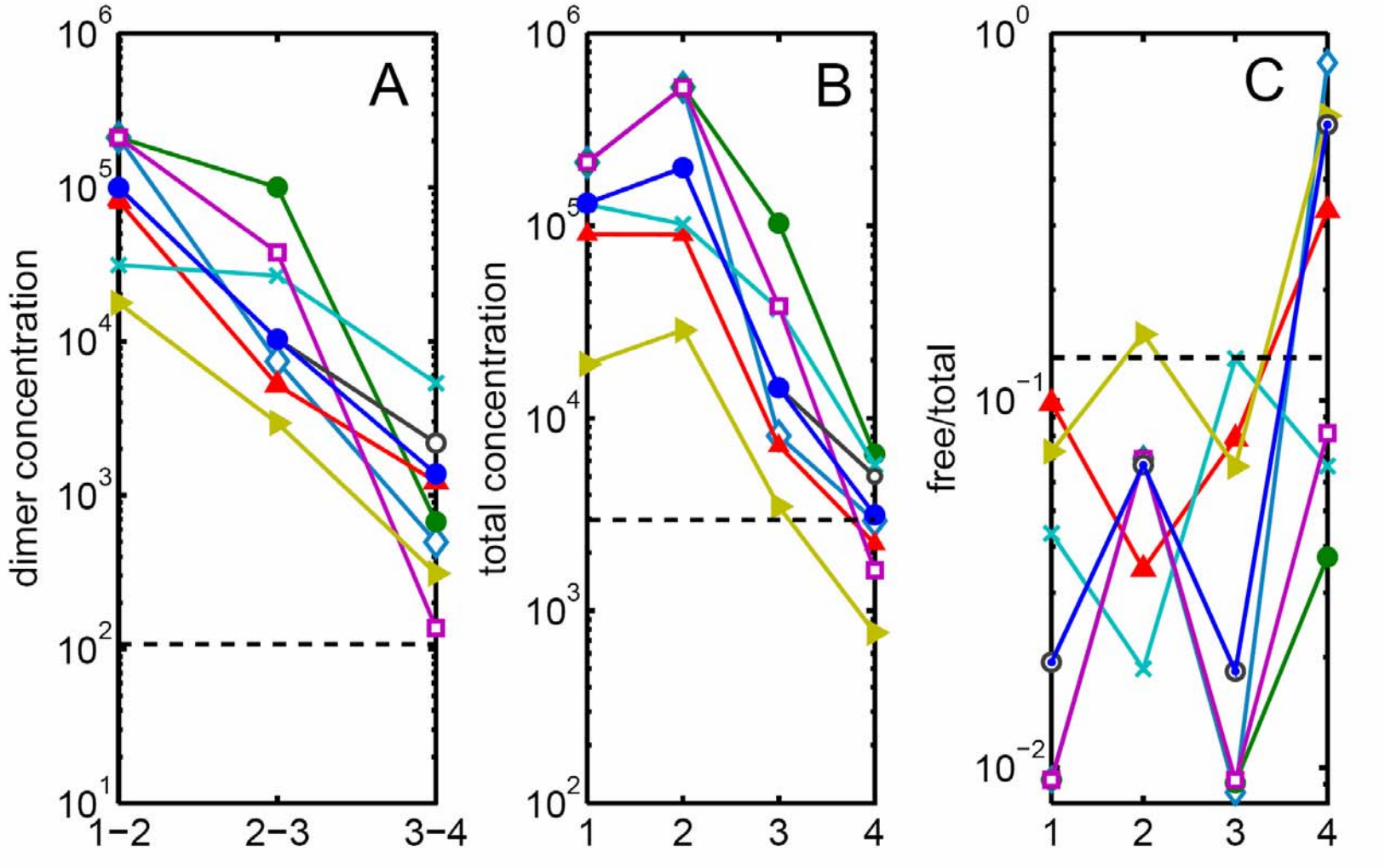
long chains
good conduits

for propagation of
concentration perturbations
while suppressing it
along the rest ?

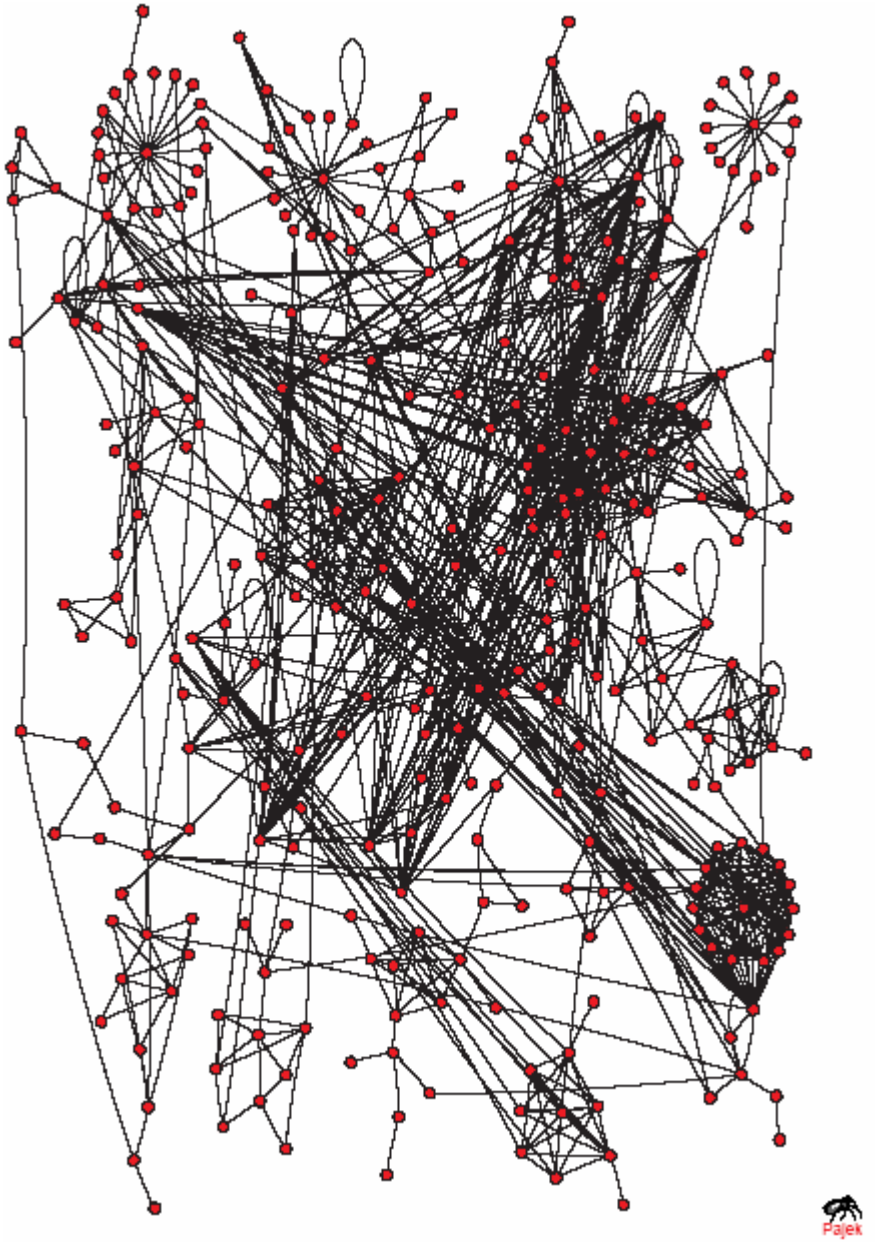
Resistor network analogy

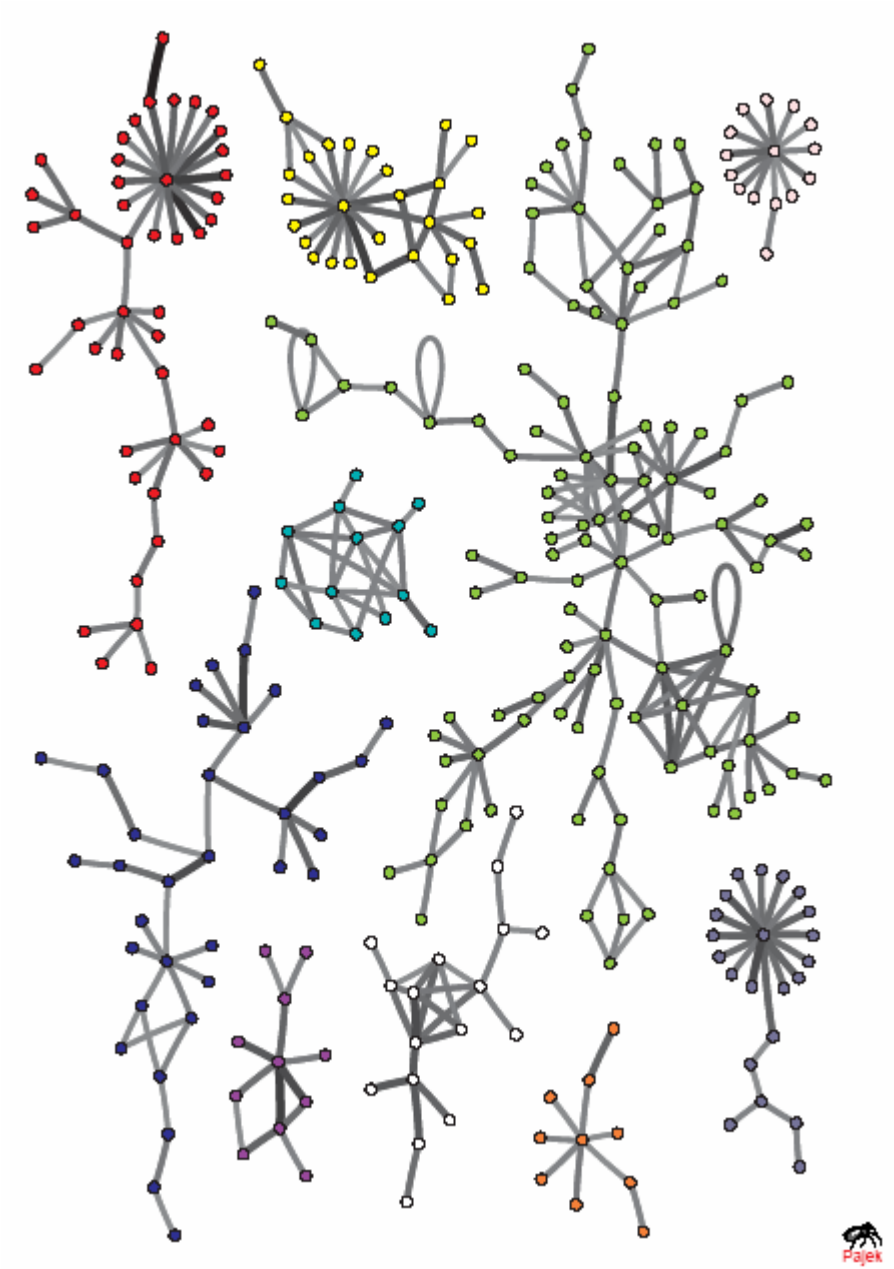
- Conductivities σ_{ij} – dimer (bound) concentrations D_{ij}
- Losses to the ground σ_{iG} – free (unbound) concentrations F_i
- Electric potentials – relative changes in free concentrations $(-1)^L \delta F_i/F_i$
- Injected current – initial perturbation δC_0

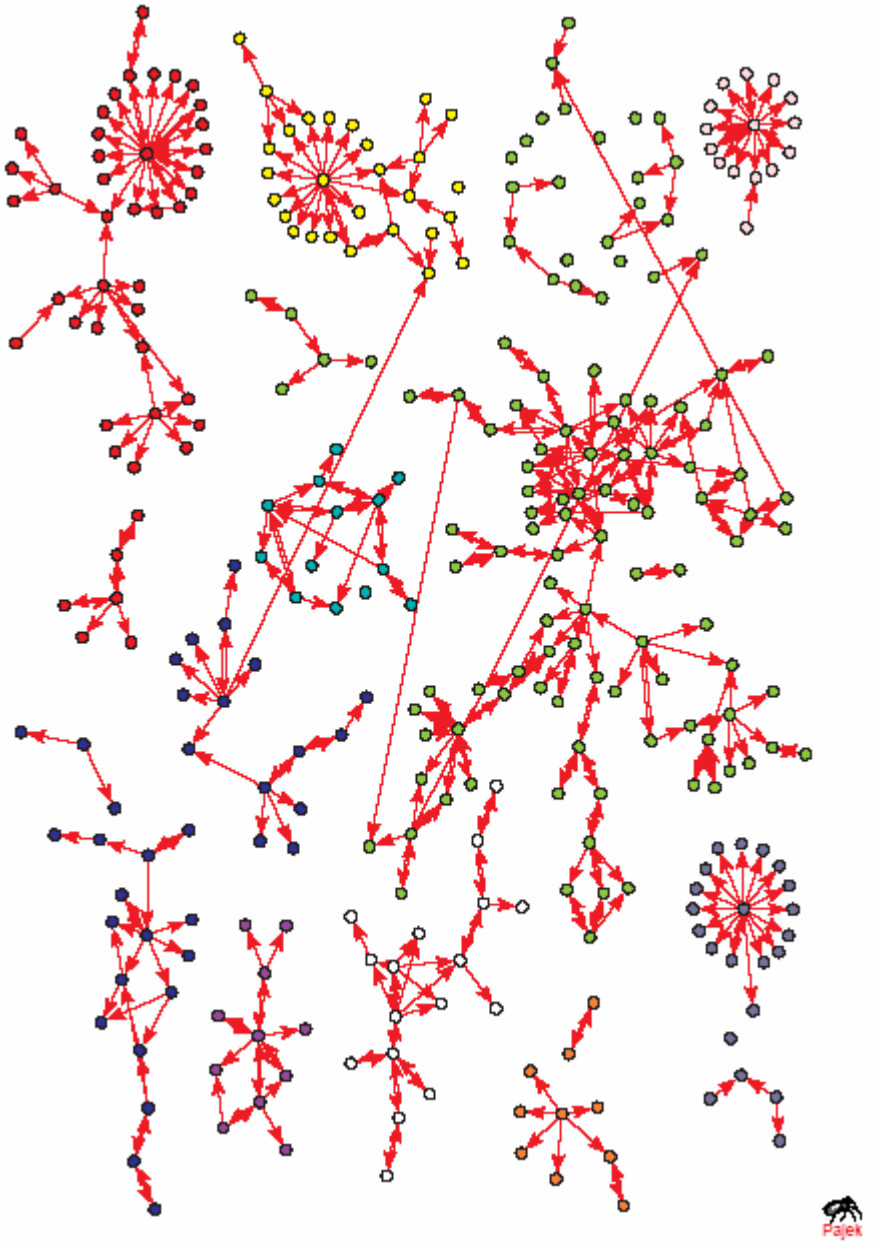
SM, K. Sneppen, I. Ispolatov, q-bio/0611026



- Perturbations propagate along dimers with large concentrations
- They cascade down the concentration gradient and thus directional
- Free concentrations of intermediate proteins are low


Pajek


Pajek


Pajek

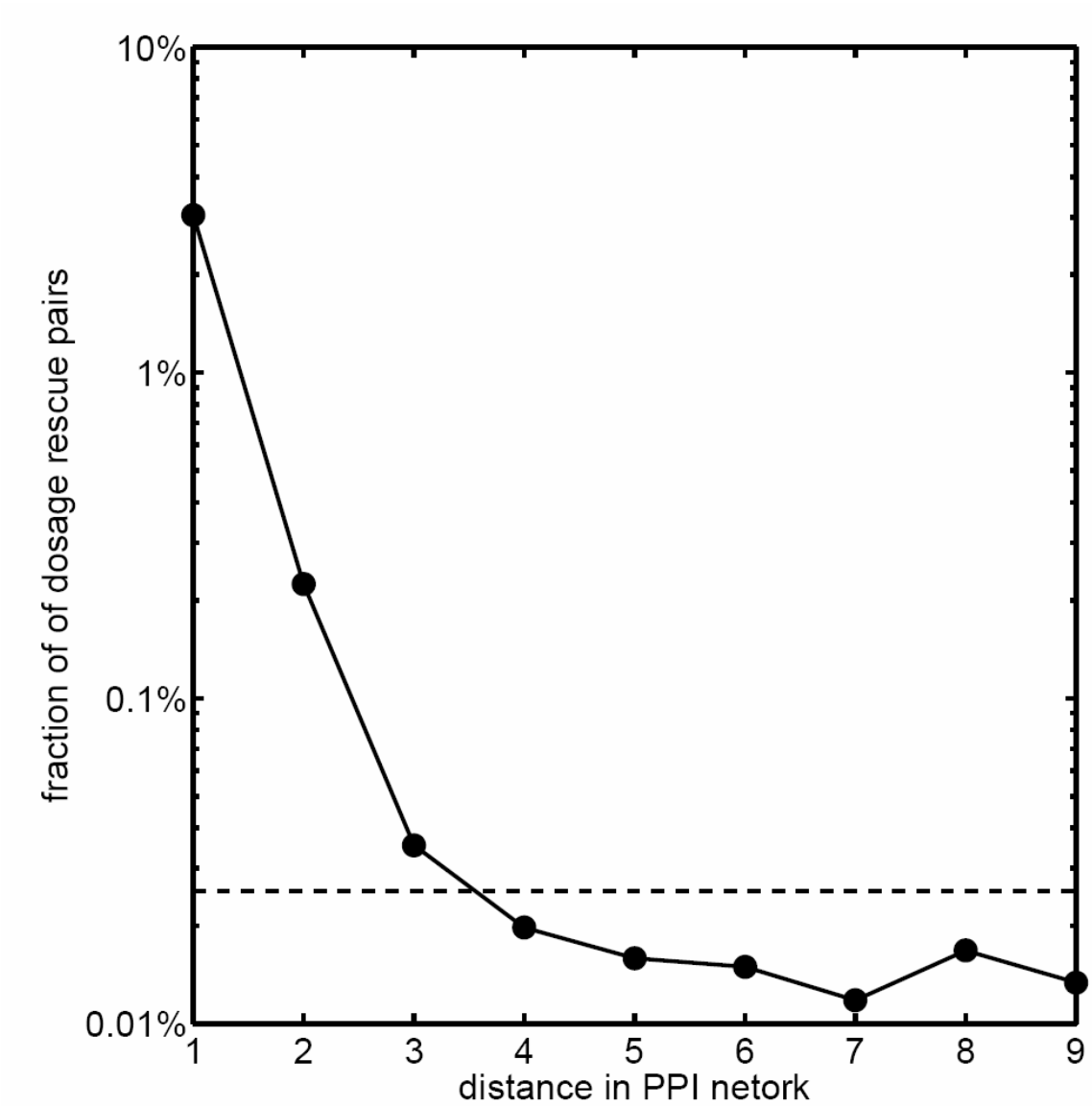
Implications of our results

Cross-talk via small-world topology is suppressed, but...

- Good news: on average perturbations via reversible binding rapidly decay
- Still, the absolute number of concentration-coupled proteins is large
- In response to external stimuli levels of several proteins could be shifted. Cascading changes from these perturbations could either cancel or magnify each other.
- Our results could be used to extend the list of perturbed proteins measured e.g. in microarray experiments

Genetic interactions

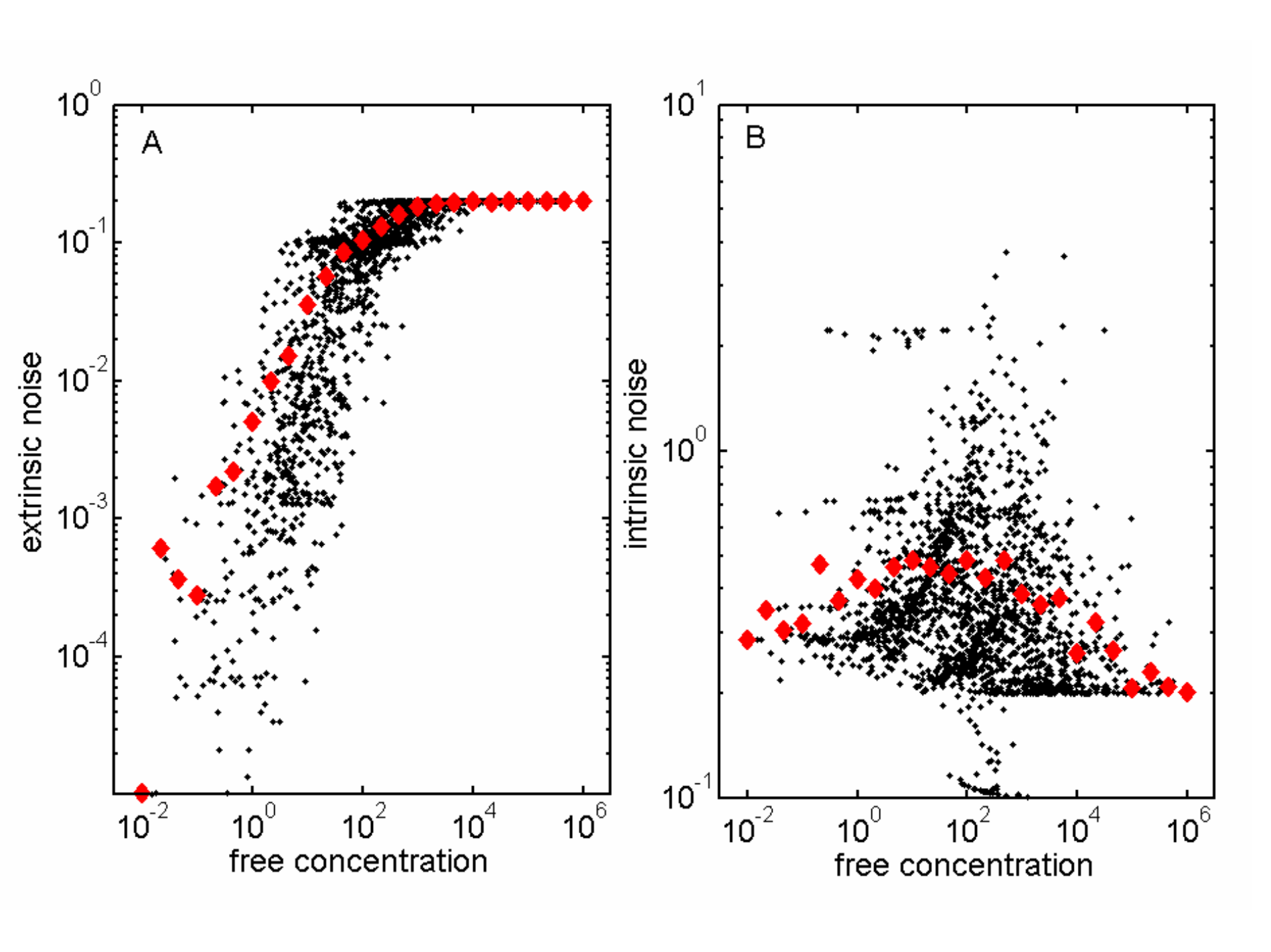
- Propagation of concentration perturbations is behind many genetic interactions e.g. of the “dosage rescue” type
- We found putative “rescued” proteins for 136 out of 772 such pairs (18% of the total, P-value 10^{-216})



SM, K. Sneppen, I. Ispolatov, q-bio/0611026; SM, I. Ispolatov, subm. (2007)

Intra-cellular noise

- Noise is measured for total concentrations C_i (Newman et al. Nature (2006))
- Needs to be converted in biologically relevant bound (D_{ij}) or free (F_i) concentrations
- Different results for intrinsic and extrinsic noise
- Intrinsic noise could be amplified (sometimes as much as 30 times!)


A
B
extrinsic noise
intrinsic noise
free concentration
free concentration
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{1}
10^{-1}
10^{-2}
10^{0}
10^{2}
10^{4}
10^{6}

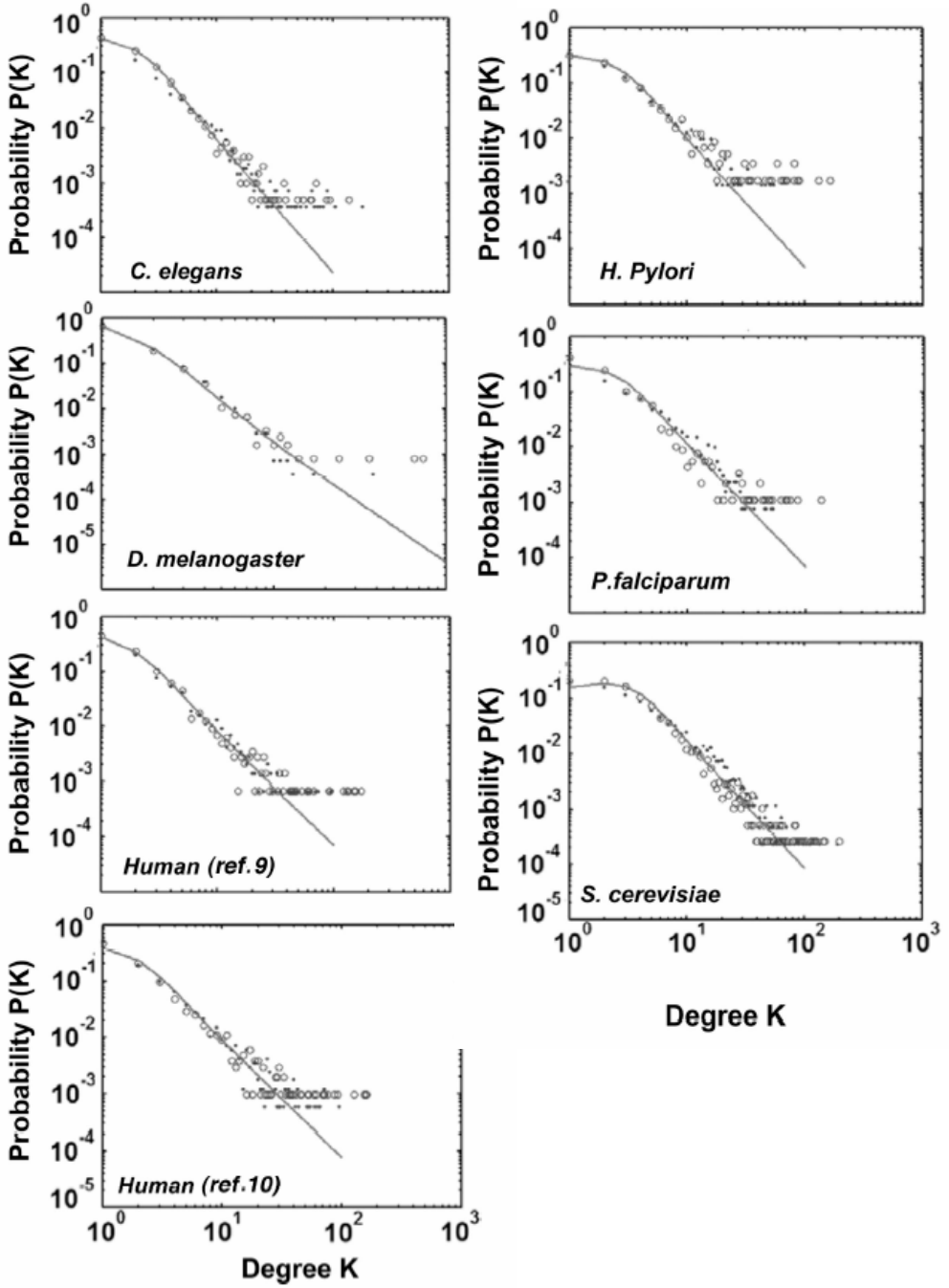
Could it be used for regulation and signaling?

- 3-step chains exist in bacteria: anti-anti-sigma-factors → anti-sigma-factors → sigma-factors → RNA polymerase
- Many proteins we find at the receiving end of our long chains are global regulators (protein degradation by ubiquitination, global transcriptional control, RNA degradation, etc.)
  - Other (catalytic) mechanisms spread perturbations even further
  - Feedback control of global protein abundance?

NOW BACK TO TOPOLOGY

What are the common topological features?

1. Broad distribution of the number of interaction partners of individual proteins



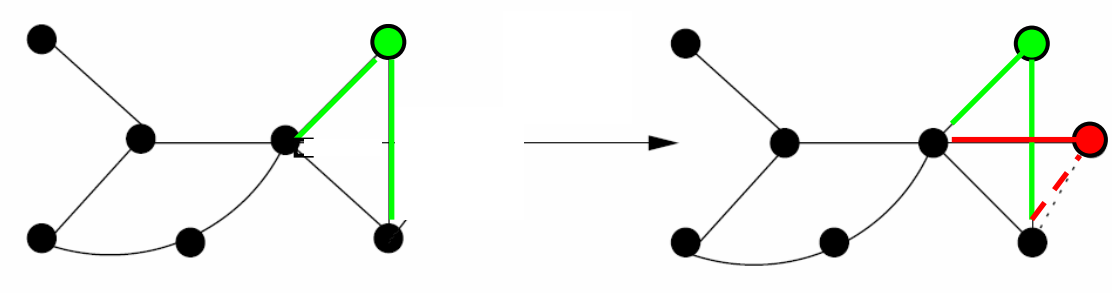
- What's behind this broad distribution?

- Three explanations were proposed:
  - EVOLUTIONARY (duplication-divergence models)
  - BIOPHYSICAL (stickiness due to surface hydrophobicity)
  - FUNCTIONAL (tasks of vastly different complexity)

From YY. Shi, GA. Miller., H. Qian., and K. Bomsztyk, PNAS **103**, 11527 (2006)

Evolutionary explanation: duplication-divergence models

- A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modelling of protein interaction networks. cond-mat/0108043, (2001) published in ComPlexUs 1, 38 (2003)
- Followed by R. V. Sole, R. Pastor-Satorras, E. Smith, T. B. Kepler, A model of large-scale proteome evolution, cond-mat/0207311 (2002) published in Advances in Complex Systems 5, 43 (2002)
- Then many others including I.Ispolatov, I., Krapivsky, P.L., Yuryev, A., Duplication-divergence model of protein interaction network, Physical Review, E 71, 061911, 2005.



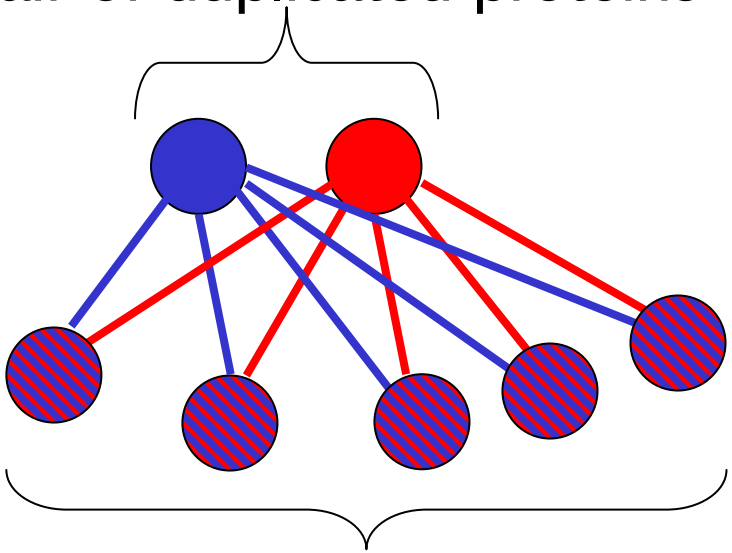
- Network has to grow
- Divergence has to be asymmetric

(K Evlampiev, H Isambert, q-bio.MN/0611070)

Gene duplication

Right after duplication

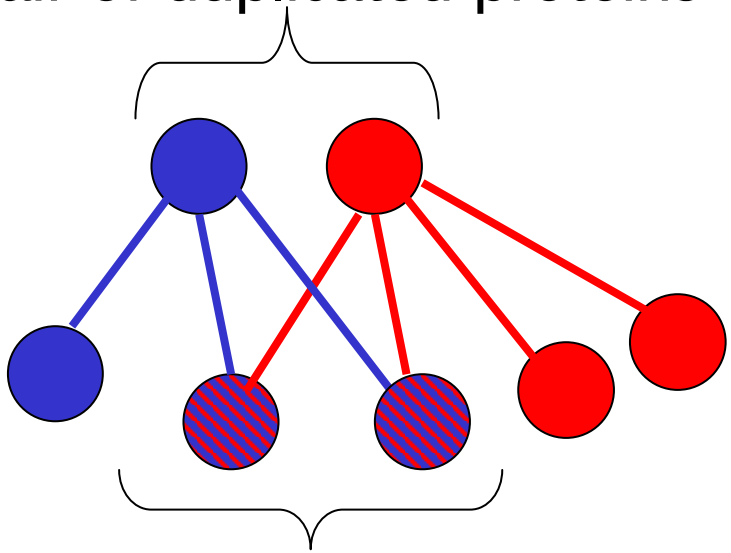
Pair of duplicated proteins



Shared interactions

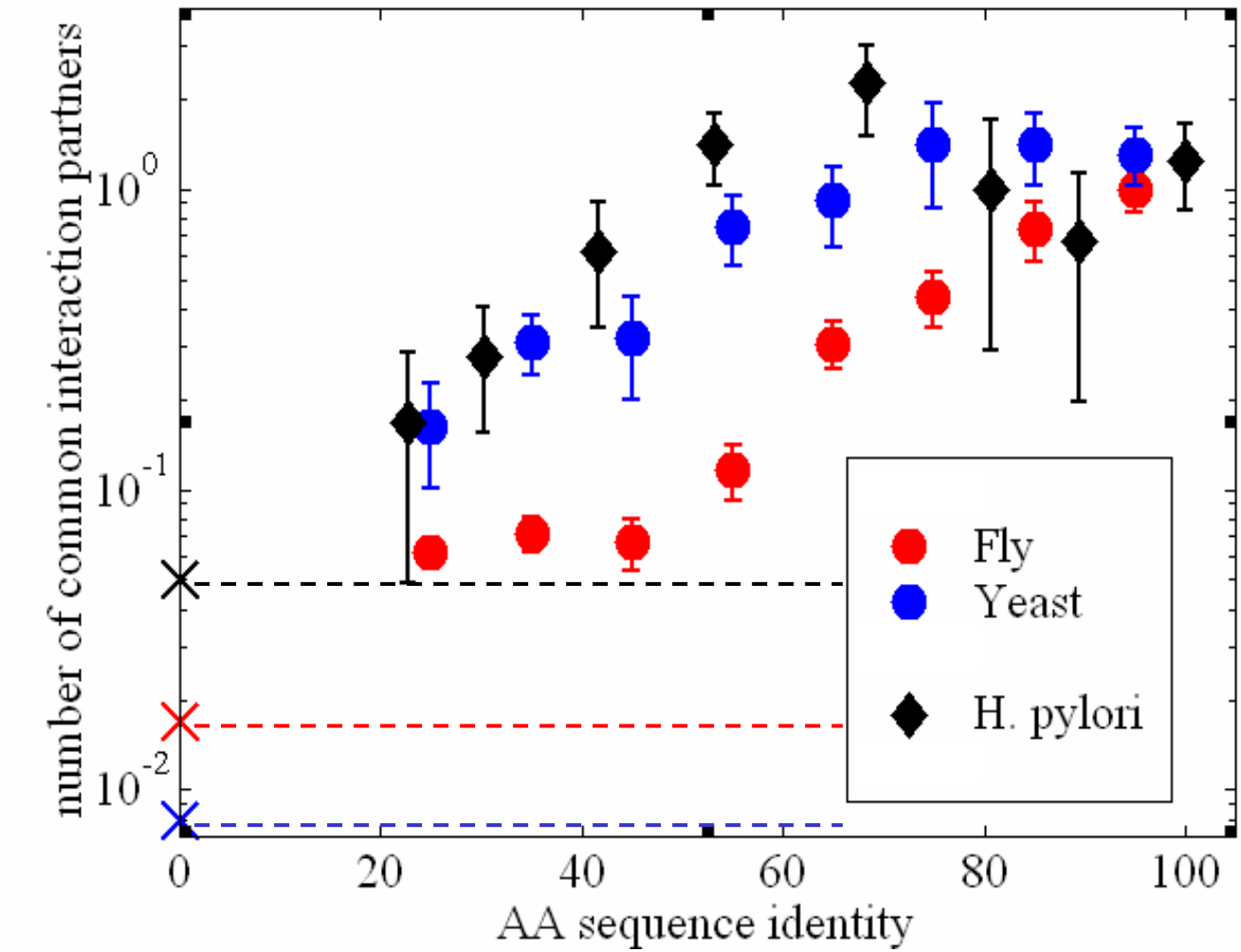
After some time

Pair of duplicated proteins



Shared interactions

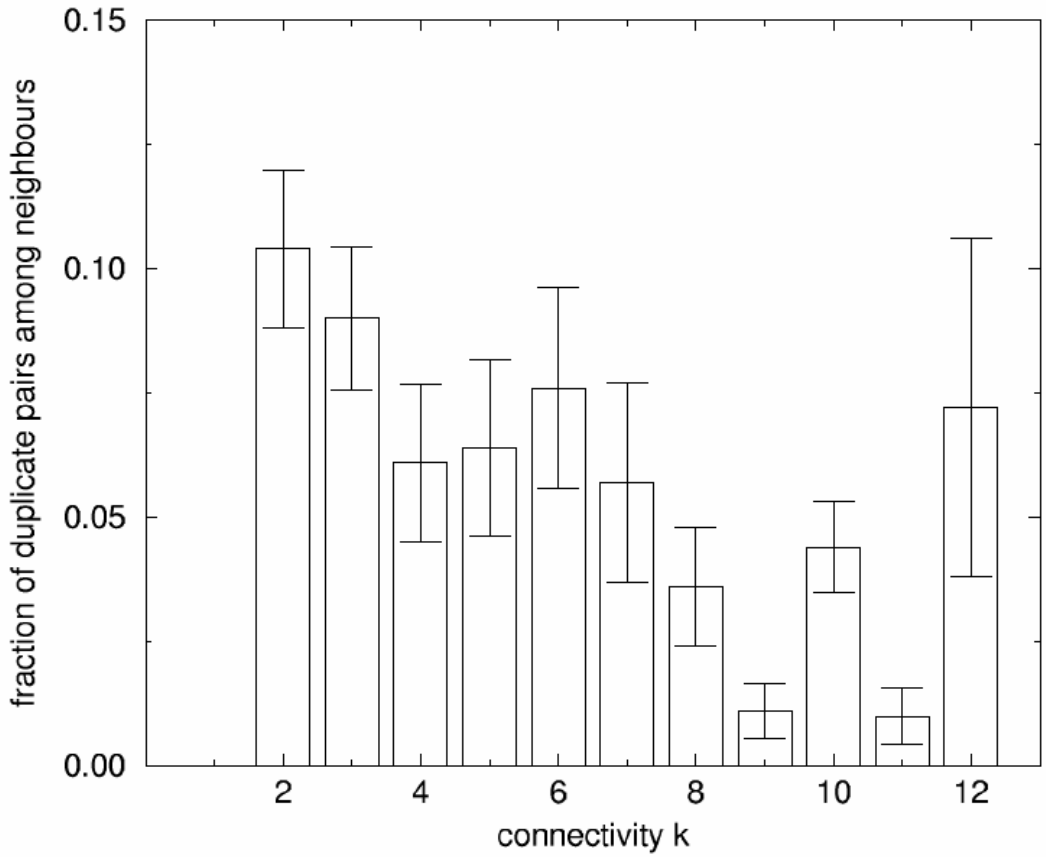
Traces of duplication in PPI networks



SM, K. Sneppen, K. Eriksen, and K-K. Yan, BMC Evol. Biol. **4**, 9 (2003)

(a similar but smaller scale-plot vs K_s in A. Wagner MBE 18, 1283 (2001)

But: how important are duplications for shaping hubs?



Duplication-divergence models could still be OK if sequences diverge relatively fast

J. Berg, M. Lässig, and A. Wagner, BMC Evol. Biol. (2004)

Biophysical explanation: "stickiness" models

- G. Caldarelli, A. Capocci, P. De Los Rios, M.A. Munoz, Scale-free Networks without Growth or Preferential Attachment: Good get Richer, cond-mat/0207366, (2002) published in PRL (2002)
- Followed by Deeds, E.J. and Ashenberg, O. and Shakhnovich, E.I., A simple physical model for scaling in protein-protein interaction networks, PNAS (2006)
- Then others including Yi Y. Shi, G.A. Miller, H. Qian, and K. Bomsztyk, Free-energy distribution of binary protein–protein binding suggests cross-species interactome differences, PNAS (2006).

- Nodes have intrinsic "stickiness" S_i.
- Stickiness could have exponential or Gaussian PDF.
- Binding edge i - j is drawn with probability $p_{ij}=F(S_i+S_j)$
- F is some (soft) threshold function, e.g.

$\exp(S_i+S_j-mu)/(1+\exp(S_i+S_j-mu))$

- Network does not have to grow

There are just TOO MANY homodimers

	N_{dimer}	$N^{(r)}_{\text{dimer}}$
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:

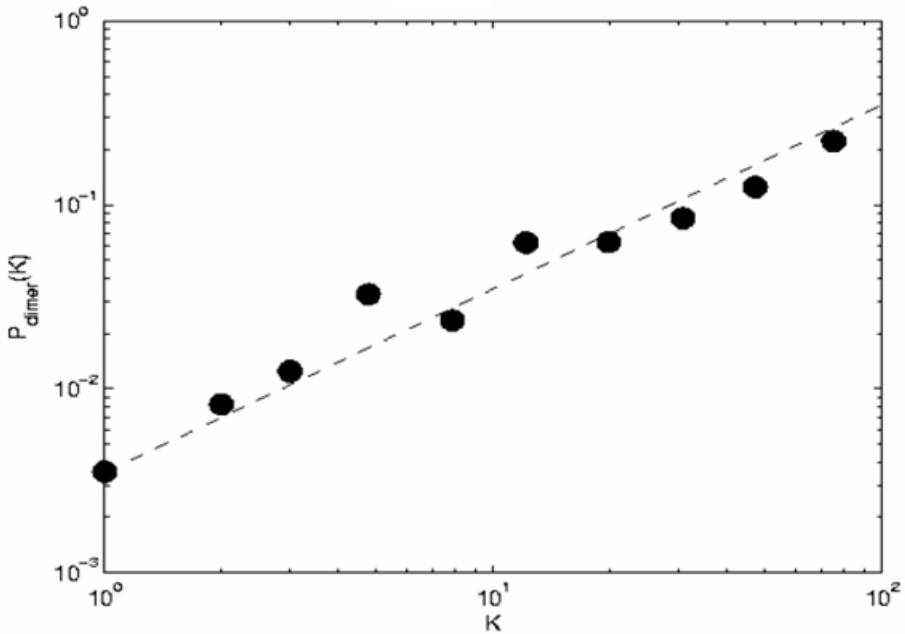
$P_{self} \sim \langle k \rangle / N$

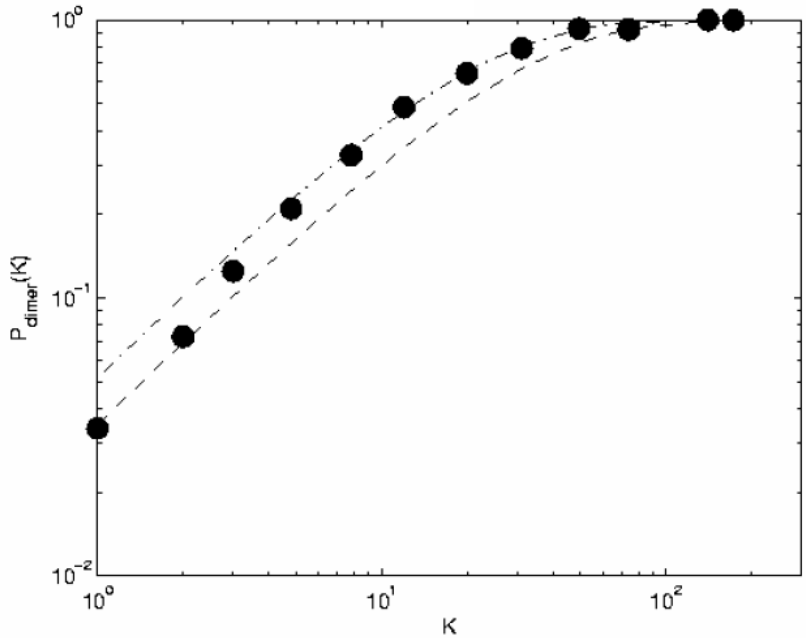
$N^{(r)}_{\text{dimer}} = N \bullet P_{self} = \langle k \rangle$

- Not surprising as homodimers have many functional roles

I. Ispolatov, A. Yuryev, I. Mazo, and SM, **33**, 3629 NAR (2005)

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





Fly: two-hybrid data

P_{self}~0.003, P_{others}~0.0002

Human: literature data

P_{self}~0.05, P_{others}~0.0002

I. Ispolatov, A. Yuryev, I. Mazo, and SM, **33**, 3629 NAR (2005)

Our interpretation

- Both the number of interaction partners K_i and the likelihood to self-interact are proportional to the same "stickiness" of the protein S_i which could depend on
  - the number of hydrophobic residues on the surface
  - protein abundance
  - its' popularity (in networks taken from many small-scale experiments)
  - etc.
- In random networks $p_{dimer}(K) \sim K^2$ not $\sim K$ like we observe empirically

I. Ispolatov, A. Yuryev, I. Mazo, and SM, **33**, 3629 NAR (2005)

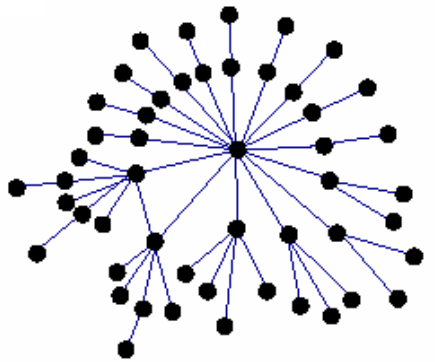
Functional explanation: there are as many binding partners as needed for function

- Not an explanation: why difficulty of functions is so heterogeneous?
- Difficult to check: the function of many binding interactions is poorly understood (quite clear in transcriptional regulatory networks e.g. in *E. coli*)
- The 3rd explanation does not exclude the previous two: Evolution by duplications combined with pure Biophysics (stickiness) provide raw materials from which functional interactions are selected

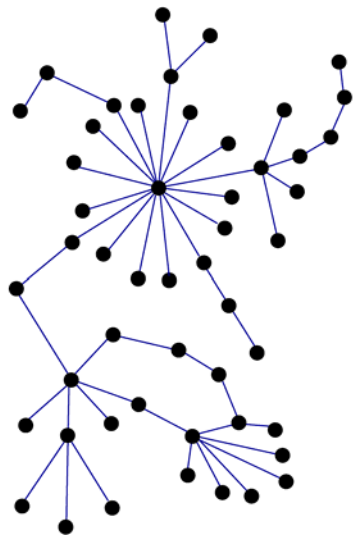
What are the common topological features?

- Broad distribution of the number of interaction partners (degree) of individual proteins
- Anti-correlation of degrees of interacting proteins

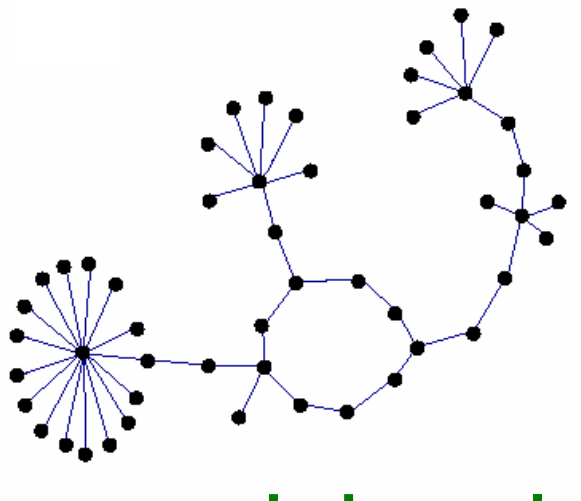
Central vs peripheral network architecture



central
(hierarchical)



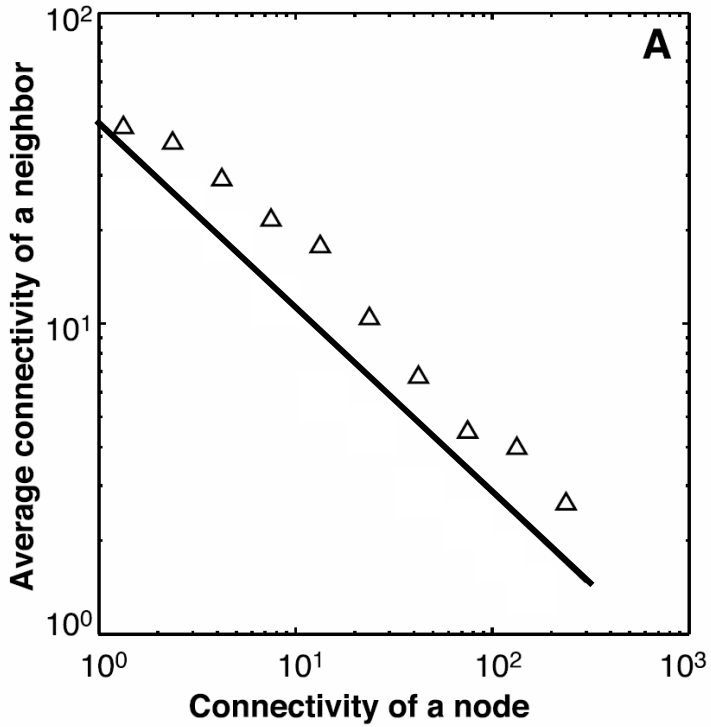
random

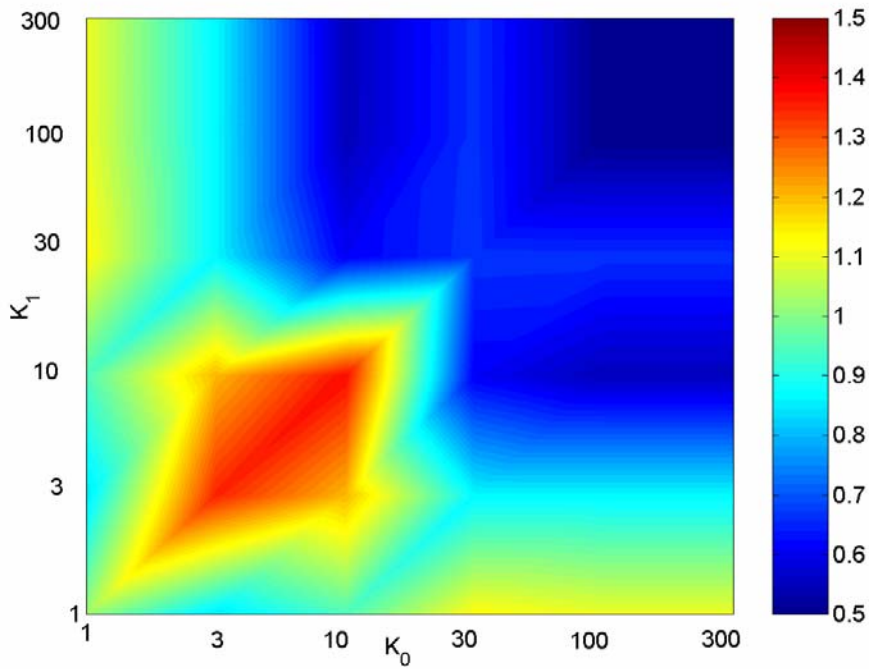


peripheral
(anti-hierarchical)

A. Trusina, P. Minnhagen, SM, K. Sneppen, *Phys. Rev. Lett.* ***92****, 17870, (2004*)

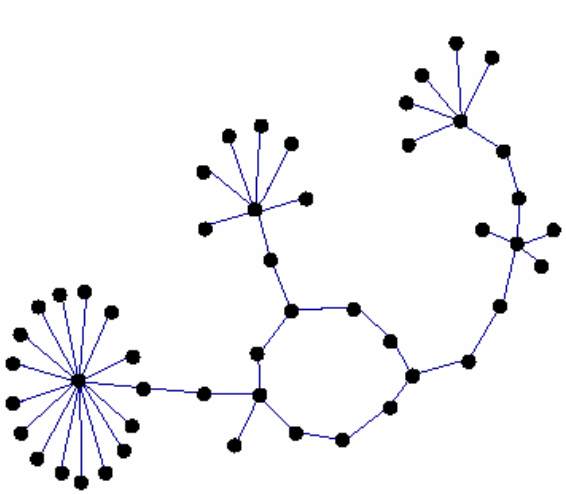
What is the case for protein interaction network



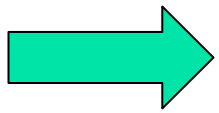


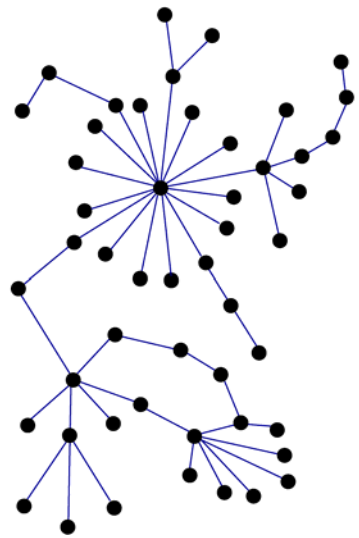
SM, K. Sneppen, *Science* ***296***, *910 (2002)*

Randomization



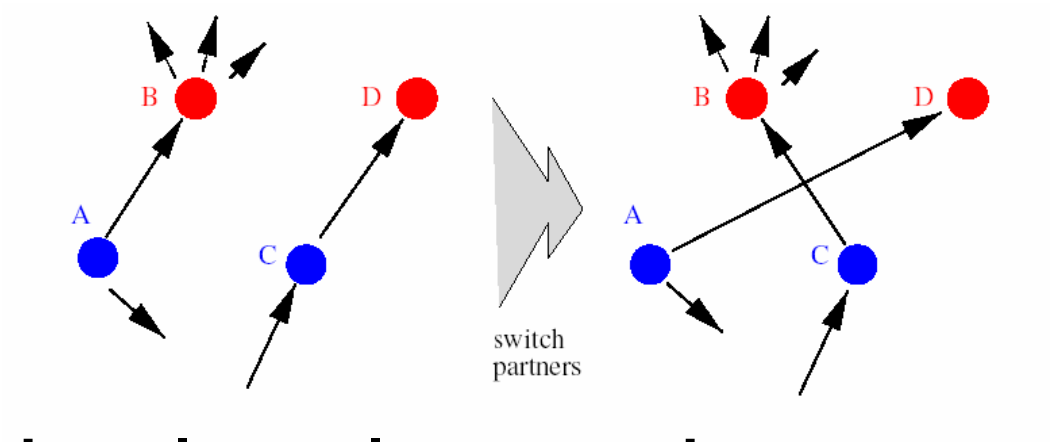
given complex network





random

Edge swapping (rewiring) algorithm

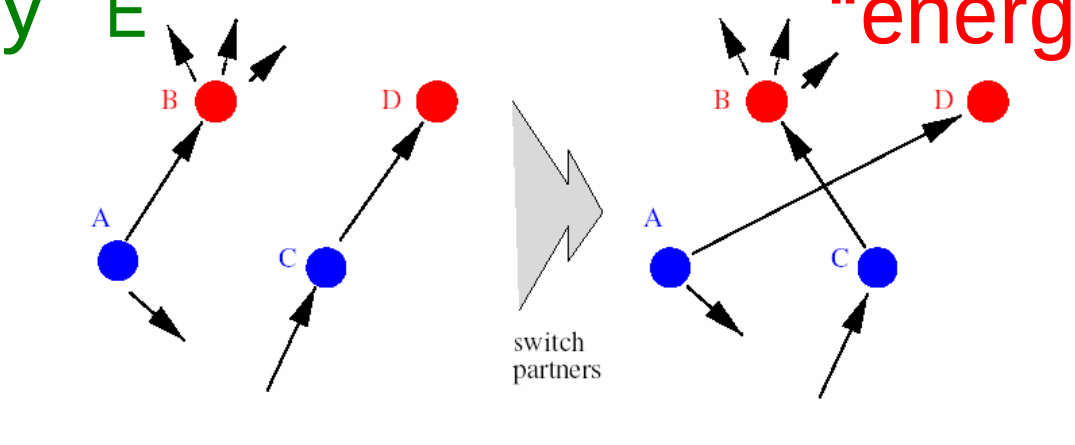


- Randomly select and rewire two edges
- Repeat many times

SM, K. Sneppen, *Science* ***296***, *910 (2002)*

Metropolis rewiring algorithm

"energy" E



"energy" E+ΔE

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

SM, K. Sneppen, *cond-mat/0205379 (2002)* Physica A 333, 529 (2004).

Anton Yuryev, AG
Kasper Eriksen,
U. of Lund

Iaroslav Ispolatov
Research scientist
Ariadne Genomics

Ilya Mazo
President
Ariadne Genomics

Kim Sneppen
NBI, Denmark

Koon-Kiu Yan,
PhD student @
Stony Brook U