Bioinformatics 3

V6 – Biological Networks are Scale-free, aren't they?

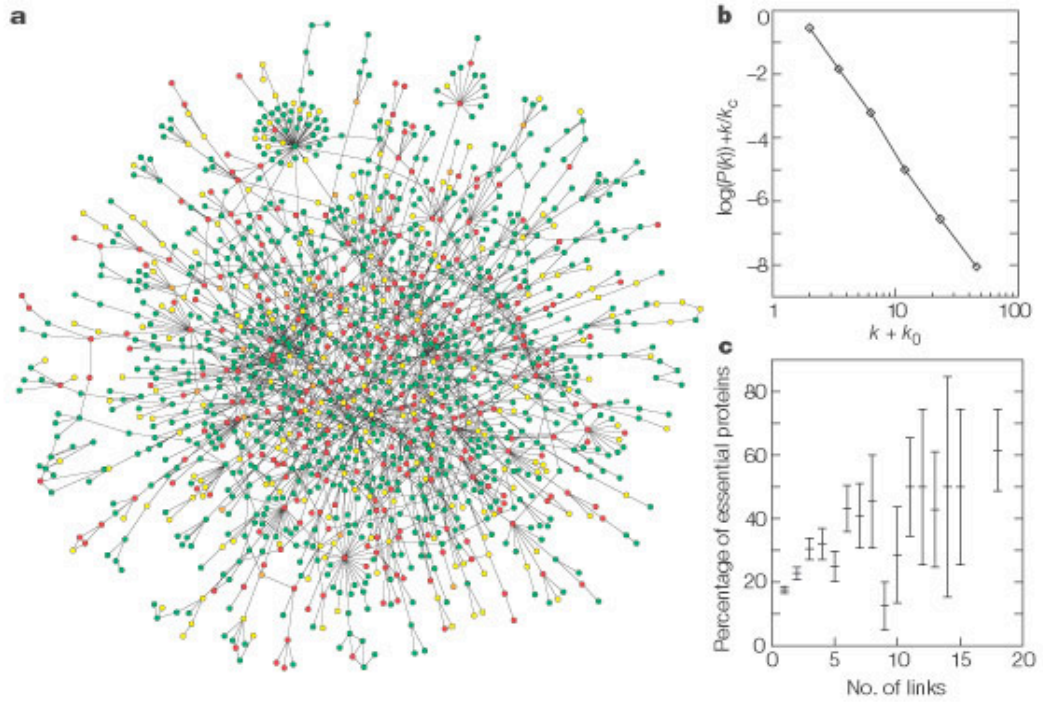
Tue, Nov 8, 2011

brief communications

Lethality and centrality in protein networks

The most highly connected proteins in the cell are the most important for its survival.

Jeong, Mason, Barabási, Oltvai, *Nature* **411** (2001) 41



largest cluster of the yeast proteome (at 2001)

=> "PPI networks apparently are scale-free…"

"Are" they scale-free
or
"Do they look like" scale-free???

Partial Sampling

Estimated for yeast: 6000 proteins, 30000 interactions

Table 1 Topological properties of interactome maps

Data set	Ito *et al.* (yeast)	Uetz *et al.* (yeast)	Ito-Uetz combined	Li *et al.* (worm)	Giot *et al.* (fly)	Minimum value	Maximum value
Total number of nodes	797	1,005	1,417	1,415	4,651	797	4,651
Nodes in main component	417 (52%)	473 (47%)	970 (68%)	1,260 (89%)	3,039 (65%)	47%	89%
Total number of interactions	806	948	1,520	2,135	4,787	806	4,787
Interactions in main component	544	558	1,229	2,038	3,715	544	3,715
R-square	0.843	0.954	0.899	0.885	0.91	0.843	0.954
γ	–1.82	–2.42	–1.91	–1.59	–2.75	–2.75	–1.59
$\langle k \rangle$	1.96	1.84	2.15	2.98	2.04	1.84	2.98
Average clustering coefficient	0.2	0.11	0.09	0.09	0.06	0.06	0.2
Number of network components	143	177	160	70	591	70	591
Average component size	5.6	5.7	8.9	20.2	7.9	5.6	20.2
Characteristic path length	6.14	7.48	6.55	4.91	9.43	4.91	9.43
Number of baits	455	512	827	502	2,820	455	2,820

The linear regression R-square measures the linearity between log(n(k)) and log(k) i.e. the fit to a power-law distribution. γ is the exponent of the power law distribution formula that best fits the observed distribution. $\langle k \rangle$ is the average number of interactions per protein observed in the network. For the Ito, Li and Giot data sets only the high confidence interactions were considered (core).

Y2H **covers** only **3...9%** of the complete interactome!

Han et al, *Nature Biotech* **23** (2005) 839

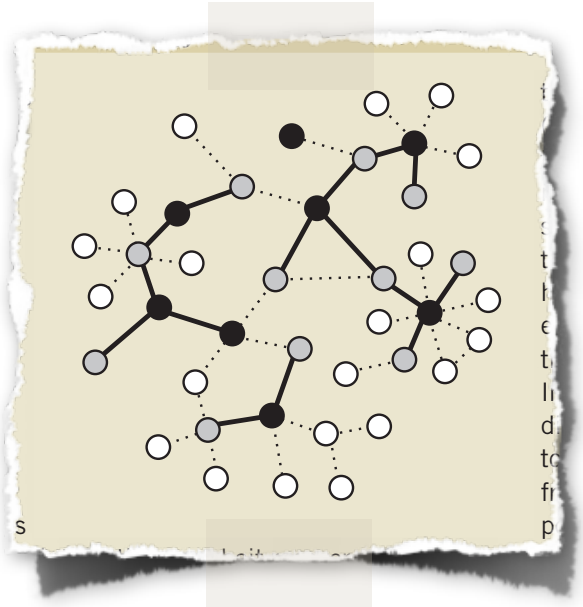
Effect of sampling on topology predictions of protein-protein interaction networks

Jing-Dong J Han[1–3], Denis Dupuy[1,3], Nicolas Bertin[1], Michael E Cusick[1] & Marc Vidal[1]

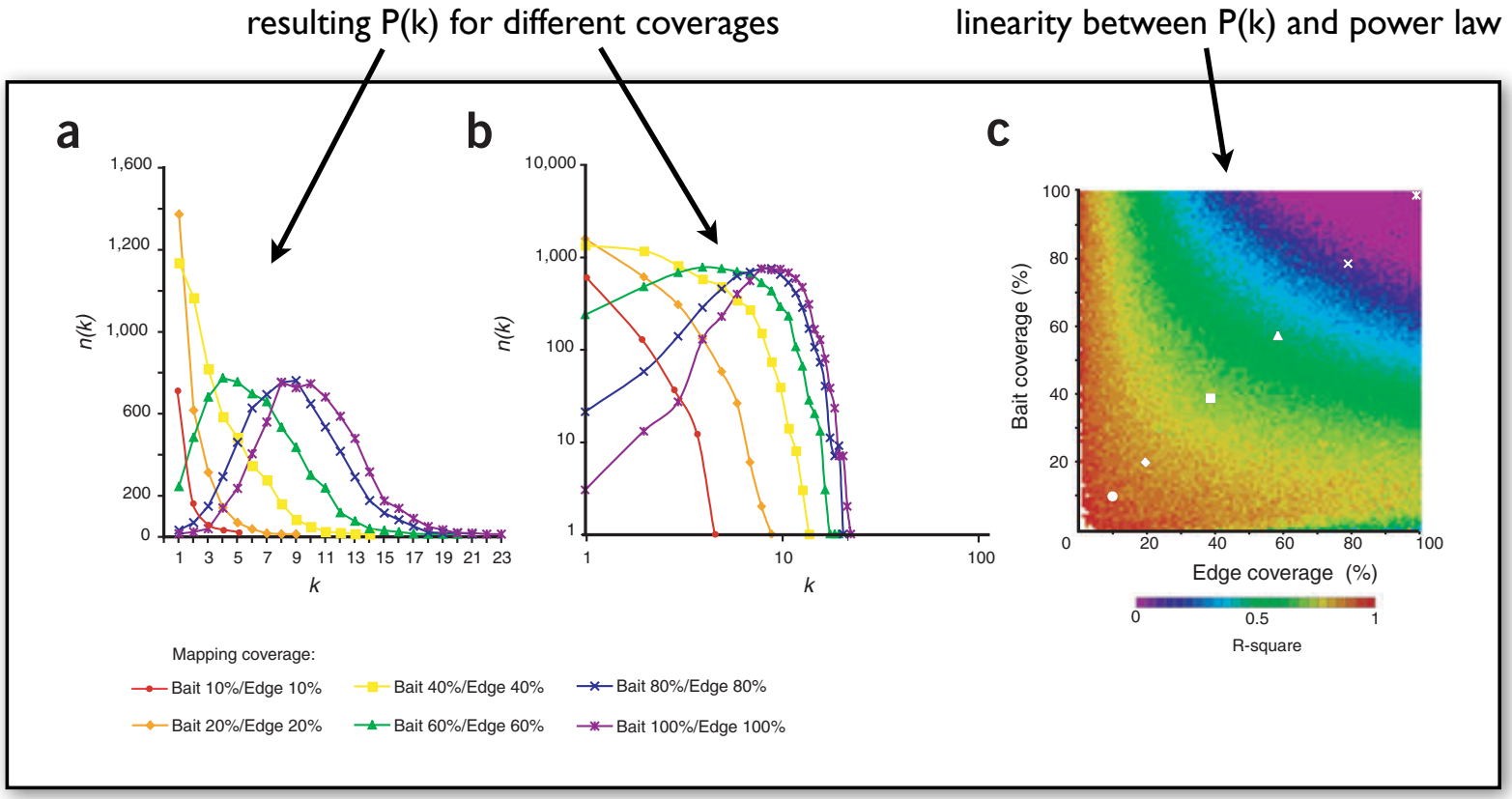
Nature Biotech **23** (2005) 839

Generate networks of various types,
sample sparsely from them
=> degree distribution?

- Random (ER) => P(k) = Poisson
- Exponential => P(k) ~ exp[-k]
- scale-free => P(k) ~ $k^{-\gamma}$
- P(k) = truncated normal distribution



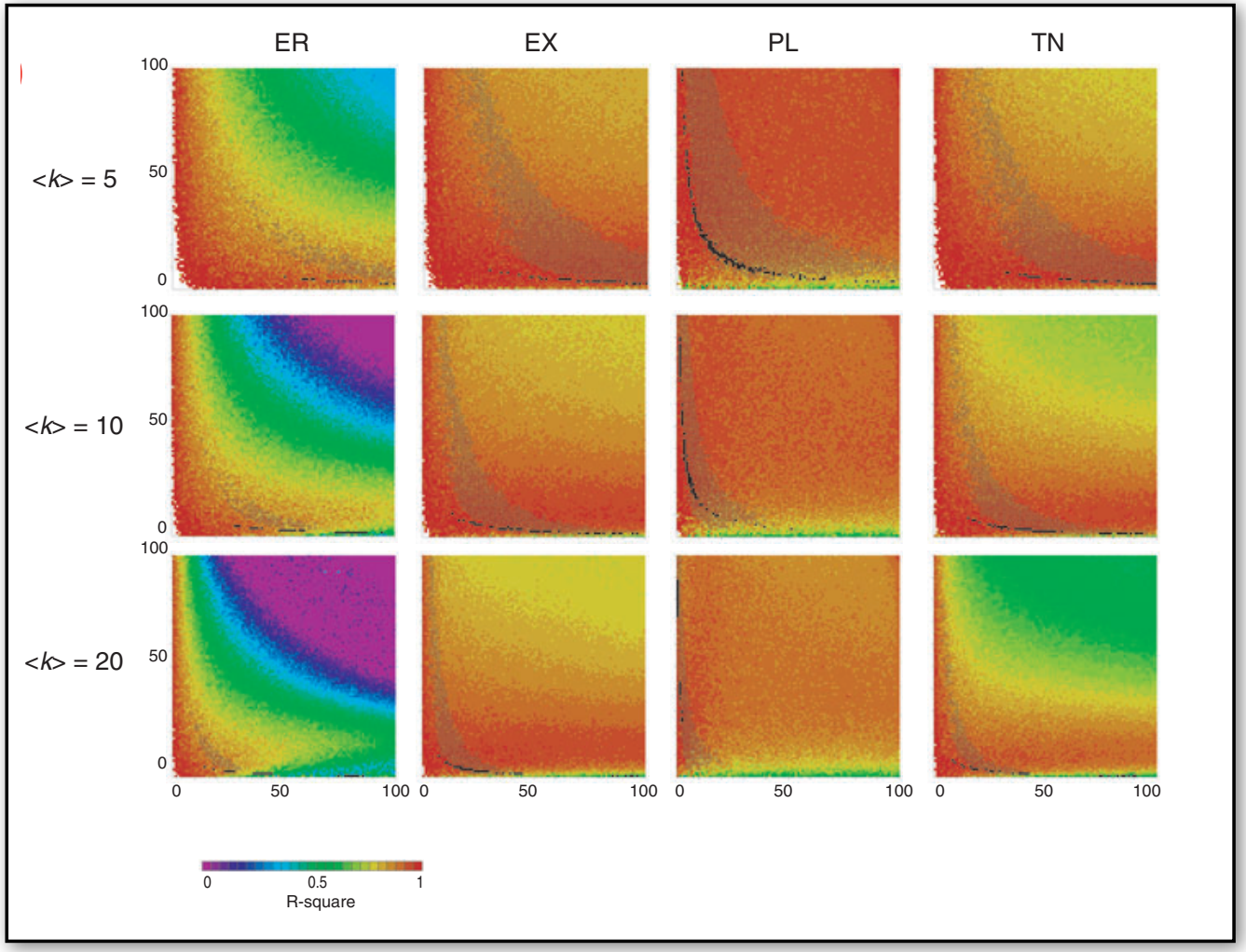
Sparsely Sampled ER Network



=> for **sparse** sampling, even an ER networks "**looks**" **scale-free**
(when only P(k) is considered)

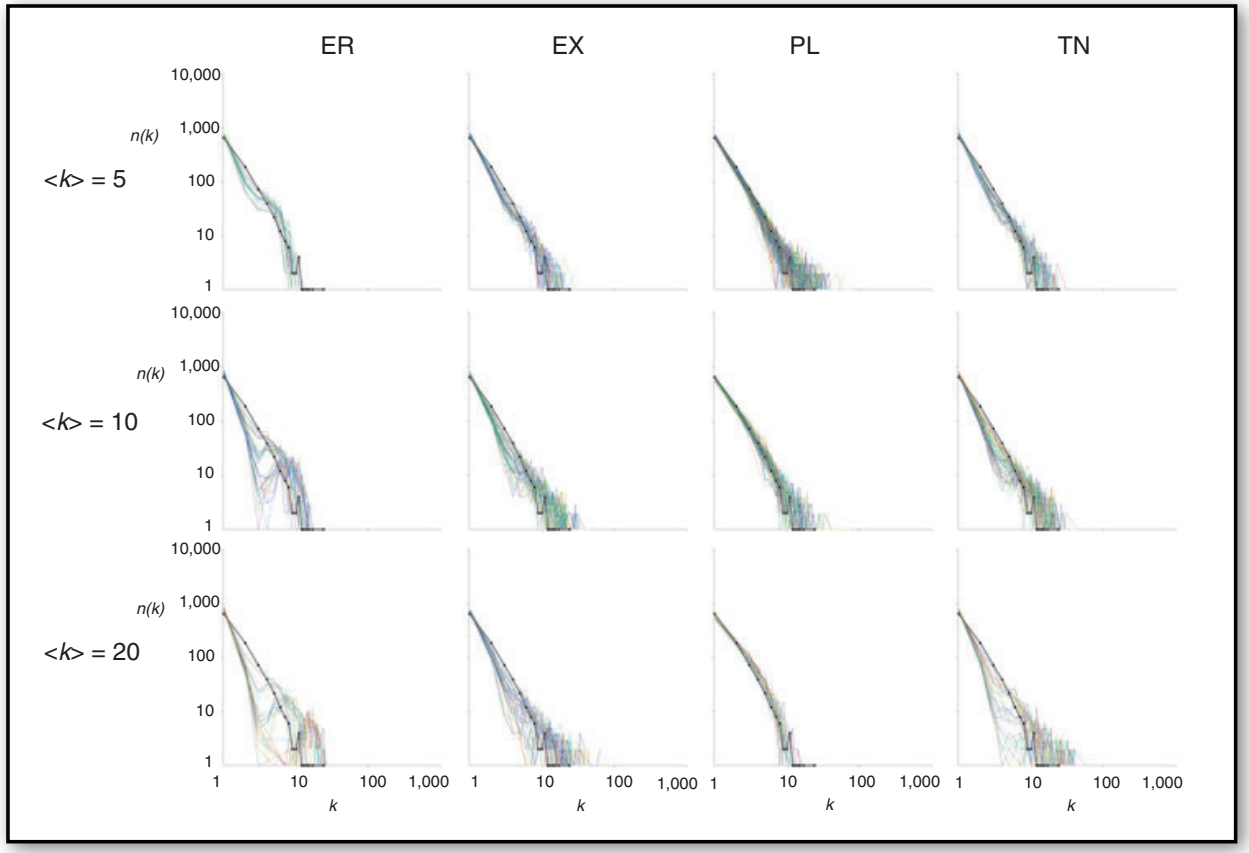
Han et al, *Nature Biotech* **23** (2005) 839

Anything Goes



Han et al, *Nature Biotech* **23** (2005) 839

Compare to Uetz et al. Data



Sampling density affects observed degree distribution
=> true underlying network cannot be identified from available data

Han et al, *Nature Biotech* **23** (2005) 839

Which Network Type?

On the structure of protein–protein interaction networks

A. Thomas*, R. Cannings†, N.A.M. Monk‡[1] and C. Cannings‡

*Genetic Epidemiology, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, U.S.A., †10 Peterborough Drive, Sheffield S10 4JB, U.K., and ‡Centre for Bioinformatics and Computational Biology, and Division of Genomic Medicine, University of Sheffield, Royal Hallamshire Hospital, Sheffield S10 2JF, U.K.

Abstract

We present a simple model for the underlying structure of protein–protein pairwise interaction graphs that is based on the way in which proteins attach to each other in experiments such as yeast two-hybrid assays. We show that data on the interactions of human proteins lend support to this model. The frequency of the number of connections per protein under this model does not follow a power law, in contrast to the reported behaviour of data from large-scale yeast two-hybrid screens of yeast protein–protein interactions. Sampling sub-graphs from the underlying graphs generated with our model, in a way analogous to the sampling performed in large-scale yeast two-hybrid searches, gives degree distributions that differ subtly from the power law and that fit the observed data better than the power law itself. Our results show that the observation of approximate power law behaviour in a sampled sub-graph does not imply that the underlying graph follows a power law.

Biochem. Soc. Trans. **31** (2001) 1491

Protein Association Network

Proteins interact (bind) via **complementary domains**

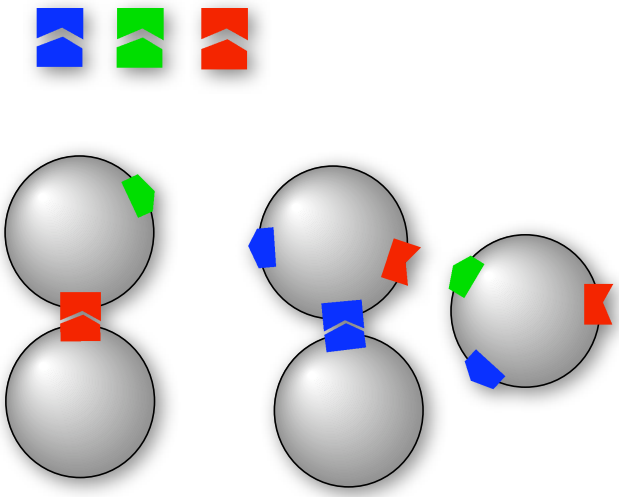
=> randomly distribute $2m$ domains onto n proteins with prob. p

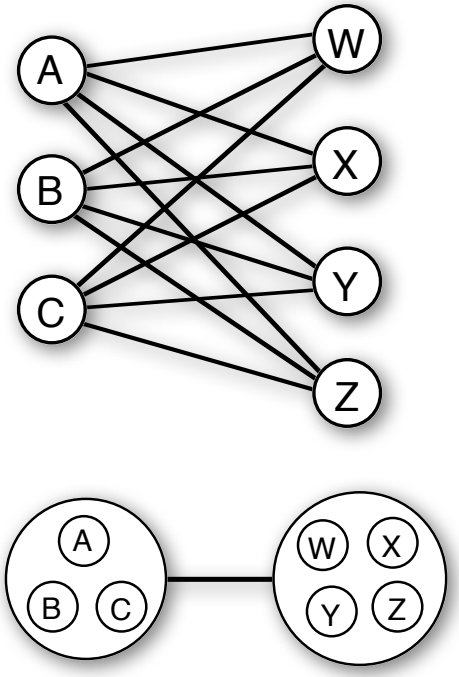
=> on avg. $\lambda = 2mp$ domains per protein

Typical **numbers** (yeast): $n = 6000$, $m = 1000$, $\lambda = 1 \ldots 2$

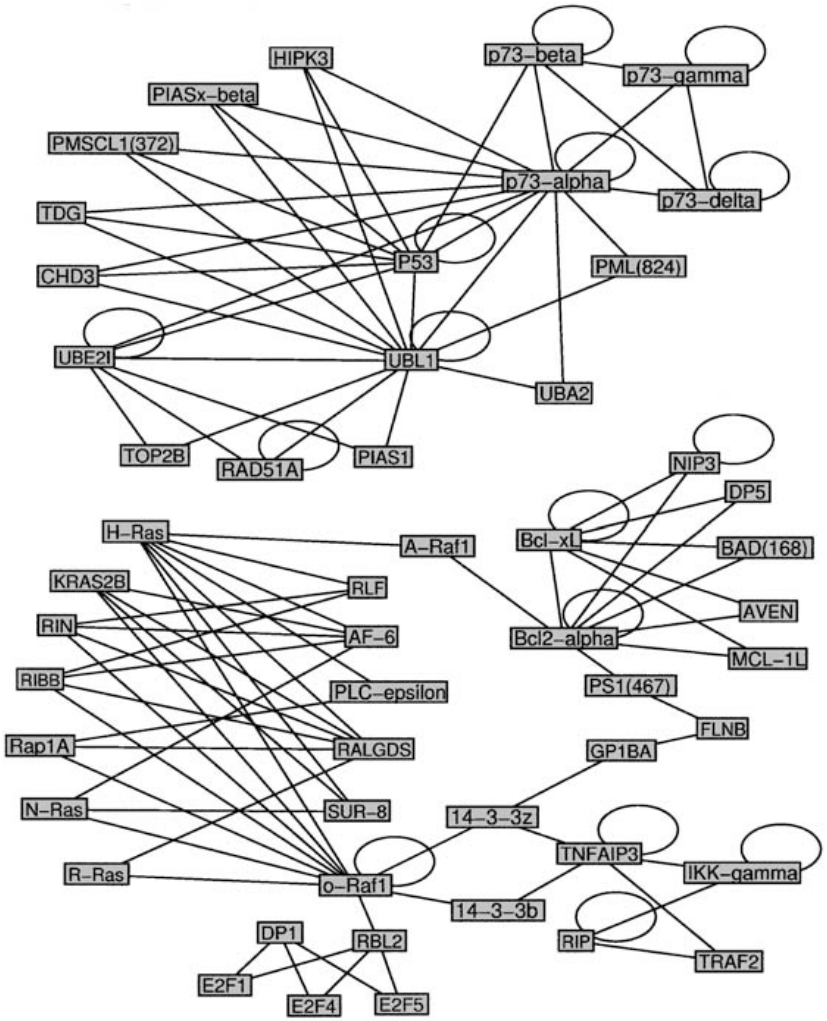
Central network sub-structure:

complete **bi-partite graphs**





Human Bipartite Graphs



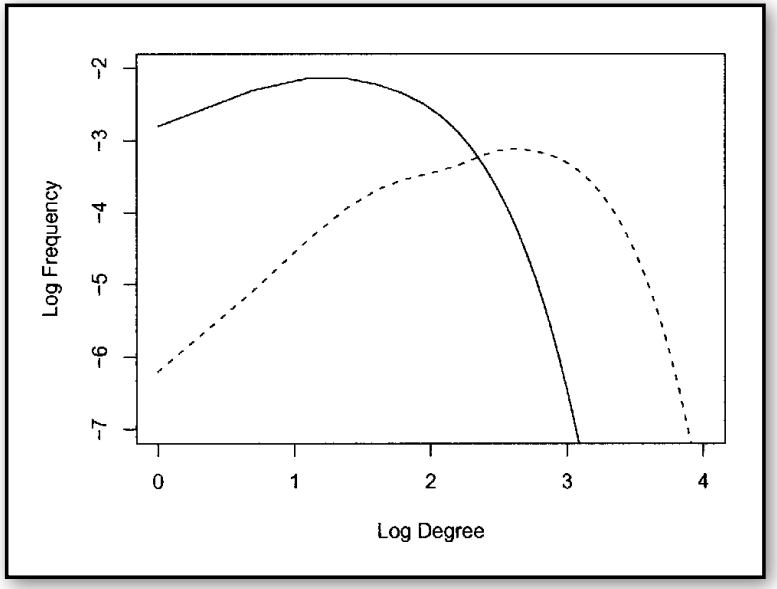
Parts of the human interactome from the Pronet database (www.myriad-pronet.com)

Thomas et al., *Biochem. Soc. Trans.* **31** (2001) 1491

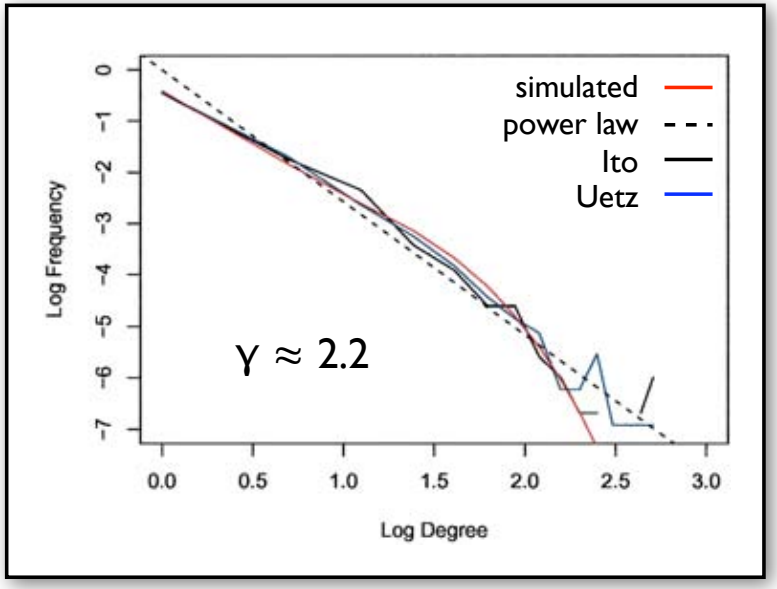
Partial Sampling

P(k) of the modelled interactome: n = 6000, m = 1000, λ = 1, 2

all nodes and vertices



450 proteins with avg 5 neighbors



Sparsely sampled protein-domain-interaction network fits very well
=> is this the correct mechanism?

Thomas et al., *Biochem. Soc. Trans.* **31** (2001) 1491

Network Growth Mechanisms

Given: an observed PPI network => how did it grow (evolve)?

Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network

Manuel Middendorf[†], Etay Ziv[‡], and Chris H. Wiggins[§¶‖]

[†]Department of Physics, [‡]College of Physicians and Surgeons, [§]Department of Applied Physics and Applied Mathematics, and [¶]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10027

Communicated by Barry H. Honig, Columbia University, New York, NY, December 20, 2004 (received for review September 7, 2004)

PNAS **102** (2005) 3192

Look at **network motifs** (local connectivity):
compare motif distributions from various network prototypes to fly network

Idea: each growth **mechanism** leads to a typical motif **distribution**,
even if global measures are equal

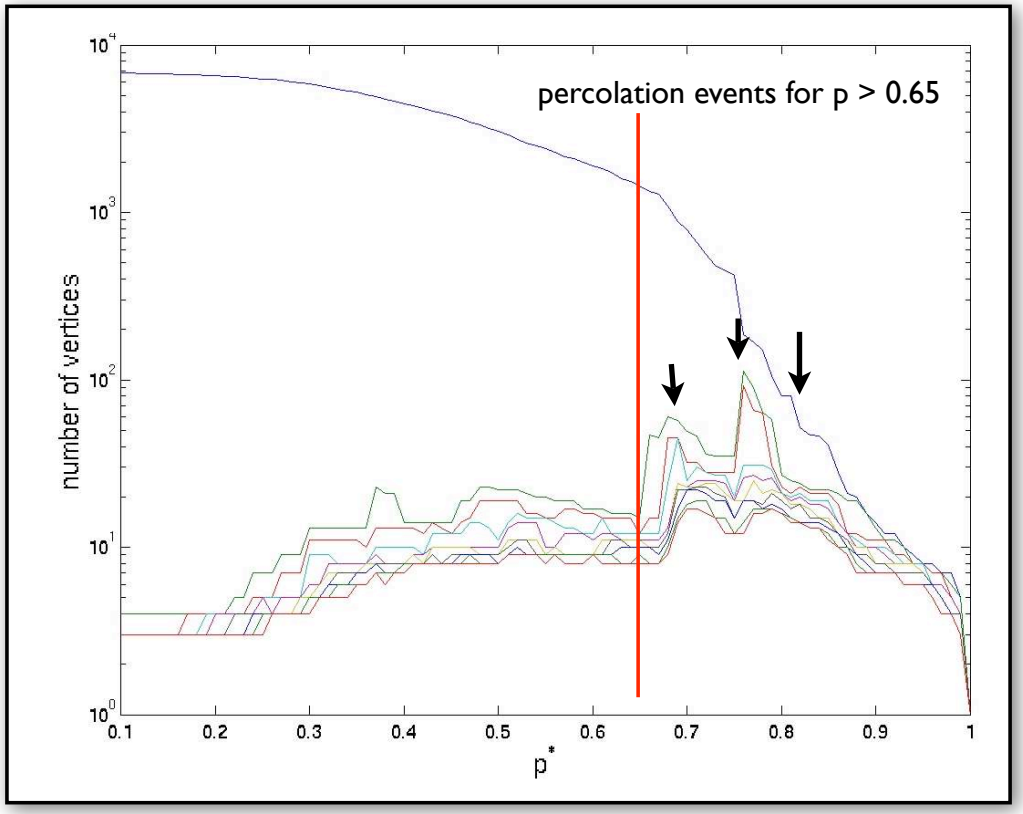
The Fly Network

Y2H PPI network for *D. melanogaster* from Giot et al. [*Science* **302** (2003) 1727]

Confidence score [0, 1] for every observed interaction
=> use only data with
p > 0.65 (0.5)
=> remove self-interactions and isolated nodes

High confidence network with 3359 (4625) nodes and 2795 (4683) edges

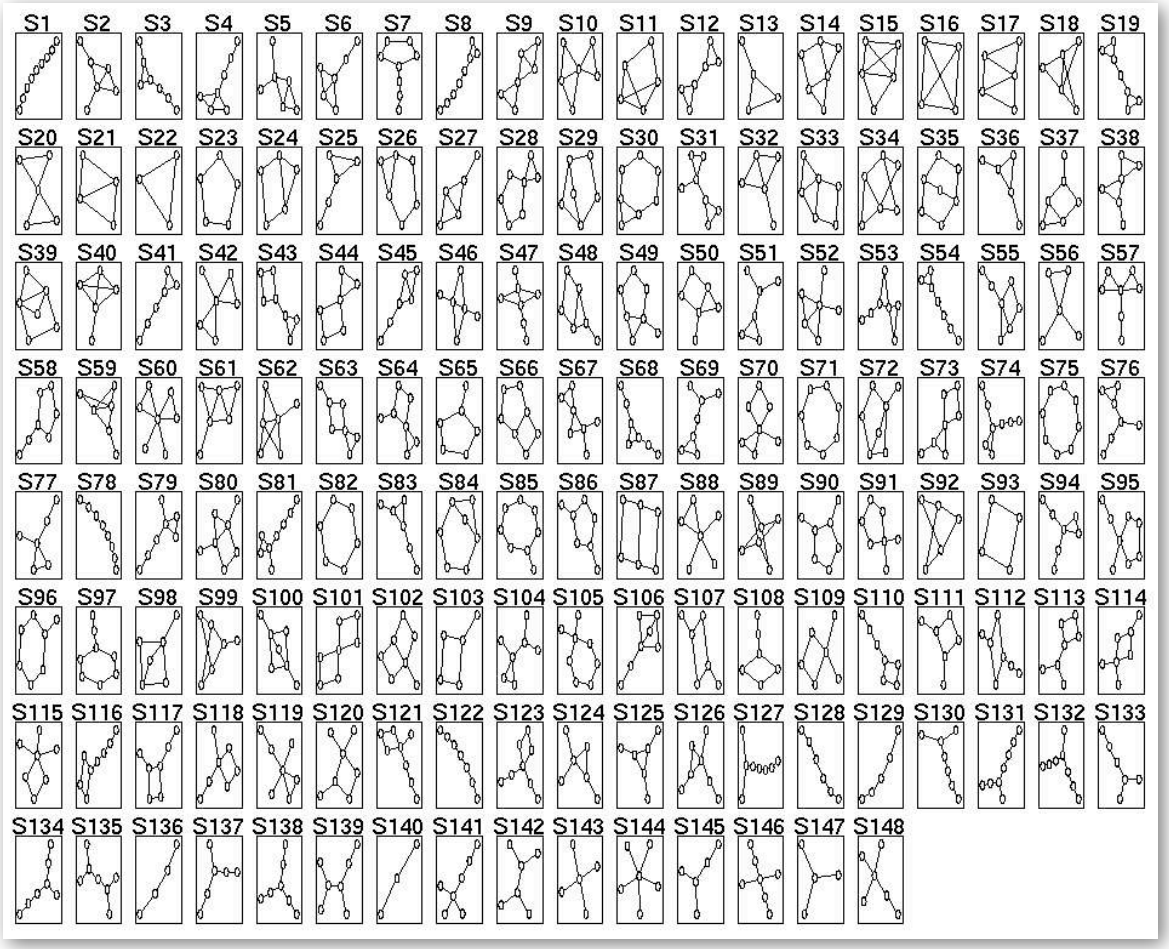
Use prototype networks of same size for training



Middendorf et al, *PNAS* **102** (2005) 3192

Network Motives

All non-isomorphic subgraphs that can be generated with a walk of length 8



Middendorf et al, *PNAS* **102** (2005) 3192

Growth Mechanisms

Generate 1000 networks, each, of the following seven types
(Same size as fly network, undefined parameters were scanned)

DMC	Duplication-mutation, preserving complementarity
DMR	Duplication with random mutations
RDS	Random static networks
RDG	Random growing network
LPA	Linear preferential attachment network
AGV	Aging vertices network
SMW	Small world network

Growth Type 1: DMC

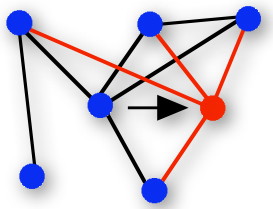
"Duplication – mutation with preserved complementarity"

Evolutionary idea: gene **duplication**, followed by a partial **loss** of function of one of the copies, making the other copy essential

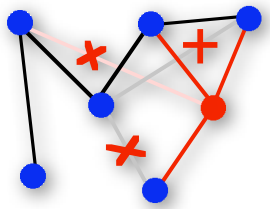
Algorithm:

Start from two connected nodes,
repeat N-2 times:

- duplicate existing node with all interactions



- for all neighbors: delete with probability q_{del} either link from original node **or** from copy



Growth Type 2: DMR

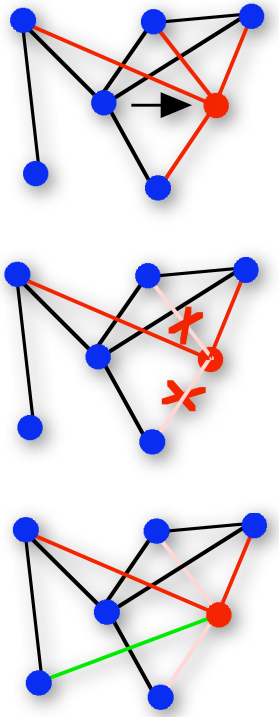
"Duplication with random mutations"

Gene duplication, but no correlation between original and copy
(original unaffected by copy)

Algorithm:

Start from five-vertex cycle,
repeat N-5 times:

- duplicate existing node with all interactions
- for all neighbors: delete with probability q_{del} link from copy
- add new links to non-neighbors with probability q_{new}/n



Growth Types 3–5: RDS, RDG, and LPA

RDS = static random network

Start from N nodes, add L links randomly

RDG = growing random network

Start from small random network, add nodes,
then edges between all existing nodes

LPA = linear preferential attachment

Add new nodes similar to Barabási-Albert algorithm,
but with preference according to (k_i + α), α = 0…5
(BA for α = 0)

For larger α: preference only for larger hubs, no difference for lower k_i

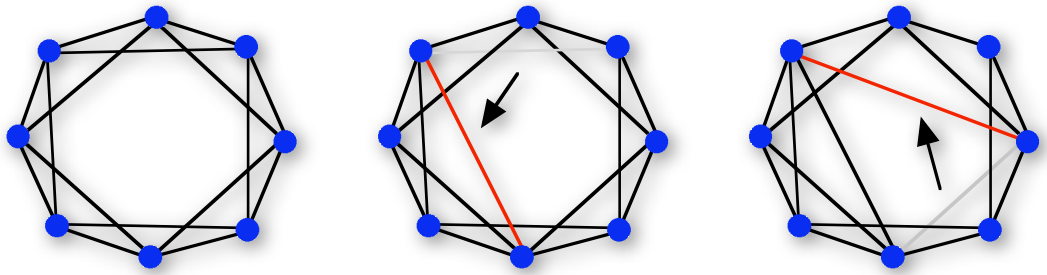
Growth Types 6-7: AGV and SMW

AGV = aging vertices network

Like growing random network,
but preference decreases with age of the node
=> citation network: more recent publications are cited more likely

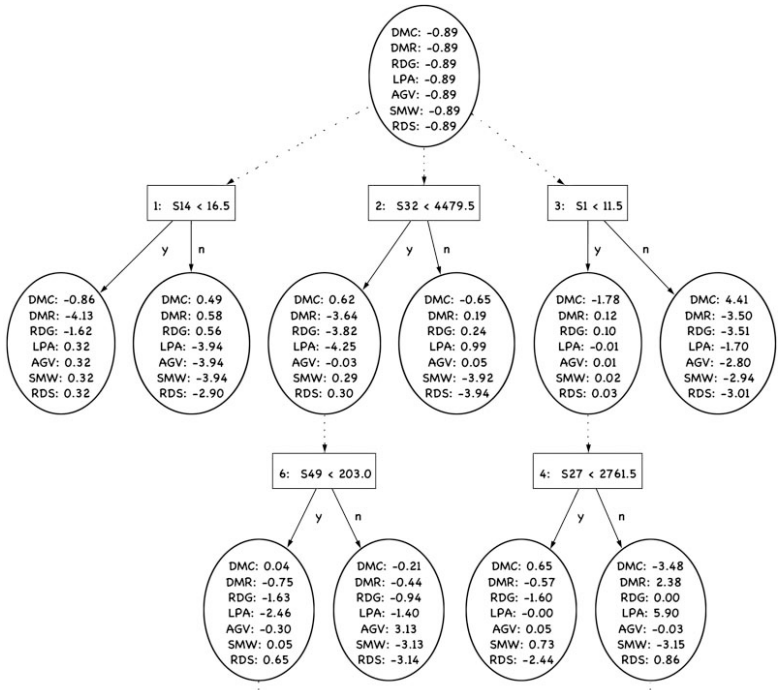
SMW = small world networks (Watts, Strogatz, *Nature* **363** (1998) 202)

Randomly rewire regular ring lattice



Alternating Decision Tree Classifier

Trained with the motif counts from 1000 networks of each of the seven types
=> prototypes are well separated and reliably classified



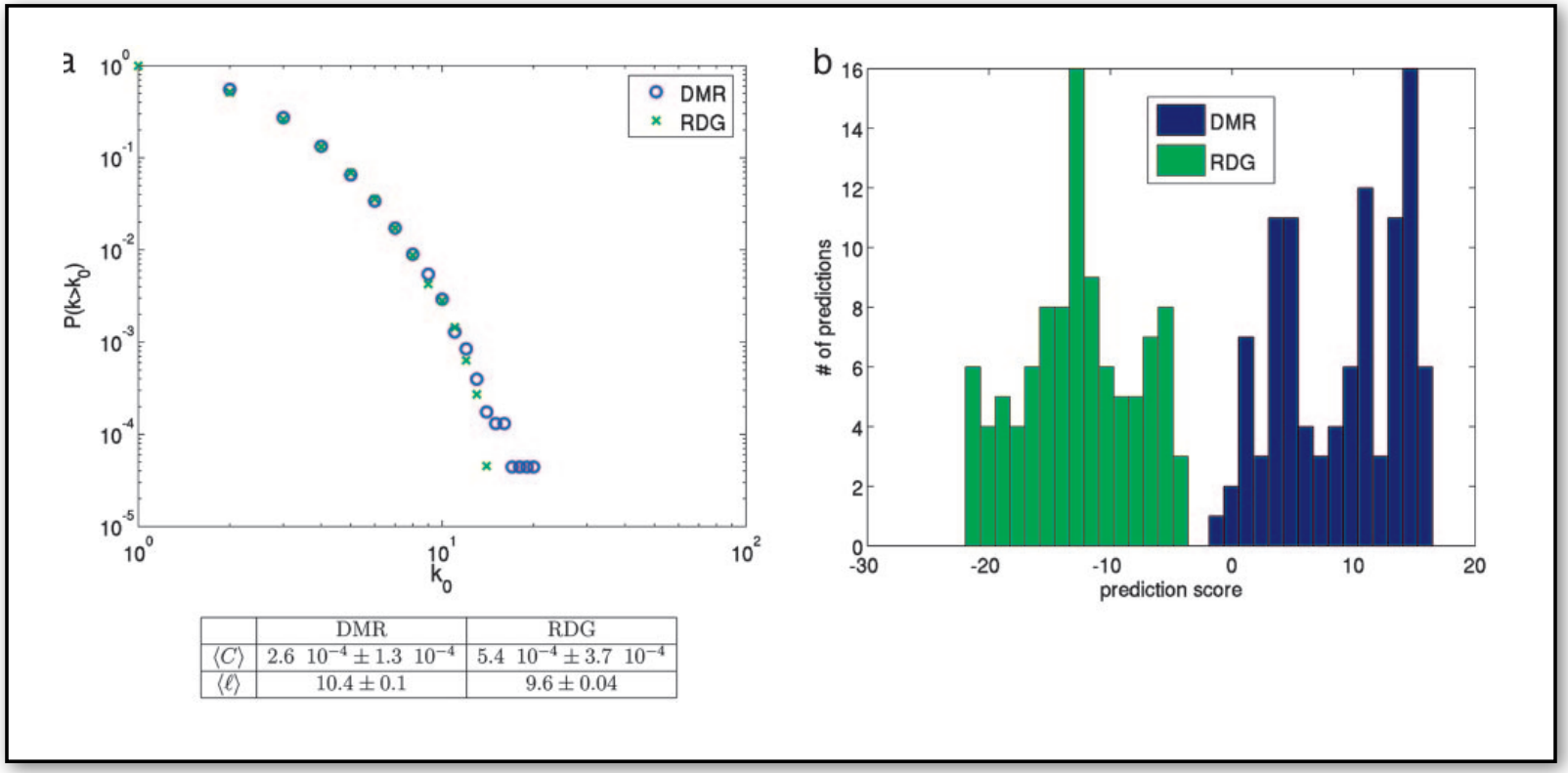
Part of a trained ADT

Prediction accuracy for networks similar to fly network with p = 0.5:

	Prediction						
Truth	DMR	DMC	AGV	LPA	SMW	RDS	RDG
DMR	99.3	0.0	0.0	0.0	0.0	0.1	0.6
DMC	0.0	99.7	0.0	0.0	0.3	0.0	0.0
AGV	0.0	0.1	84.7	13.5	1.2	0.5	0.0
LPA	0.0	0.0	10.3	89.6	0.0	0.0	0.1
SMW	0.0	0.0	0.6	0.0	99.0	0.4	0.0
RDS	0.0	0.0	0.2	0.0	0.8	99.0	0.0
RDG	0.9	0.0	0.0	0.1	0.0	0.0	99.0

Middendorf et al, *PNAS* **102** (2005) 3192

Are They Different?



	DMR	RDG
$\langle C \rangle$	$2.6\ 10^{-4} \pm 1.3\ 10^{-4}$	$5.4\ 10^{-4} \pm 3.7\ 10^{-4}$
$\langle \ell \rangle$	10.4 ± 0.1	9.6 ± 0.04

Example DMR vs. RDG: Similar global parameters,
but different counts of the network motifs

Middendorf et al, *PNAS* **102** (2005) 3192

How Did the Fly Evolve?

	Eight-step subgraphs ($p^* = 0.65$)		Subgraphs with up to seven edges ($p^* = 0.65$)		Eight-step subgraphs ($p^* = 0.5$)	
Rank	Class	Score	Class	Score	Class	Score
1	DMC	8.2 ± 1.0	DMC	8.6 ± 1.1	DMC	0.8 ± 2.9
2	DMR	−6.8 ± 0.9	DMR	−6.1 ± 1.7	DMR	−2.1 ± 2.0
3	RDG	−9.5 ± 2.3	RDG	−9.3 ± 1.6	AGV	−3.1 ± 2.2
4	AGV	−10.6 ± 4.2	AGV	−11.5 ± 4.1	LPA	−10.1 ± 3.1
5	LPA	−16.5 ± 3.4	LPA	−14.3 ± 3.2	SMW	−20.6 ± 1.9
6	SMW	−18.9 ± 0.7	SMW	−18.3 ± 1.9	RDS	−22.3 ± 1.7
7	RDS	−19.1 ± 2.3	RDS	−19.9 ± 1.5	RDG	−22.5 ± 4.7

Drosophila is consistently (independently of the cut-off in subgraph size) classified as a DMC network, with an especially strong prediction for a confidence threshold of $p^* = 0.65$.

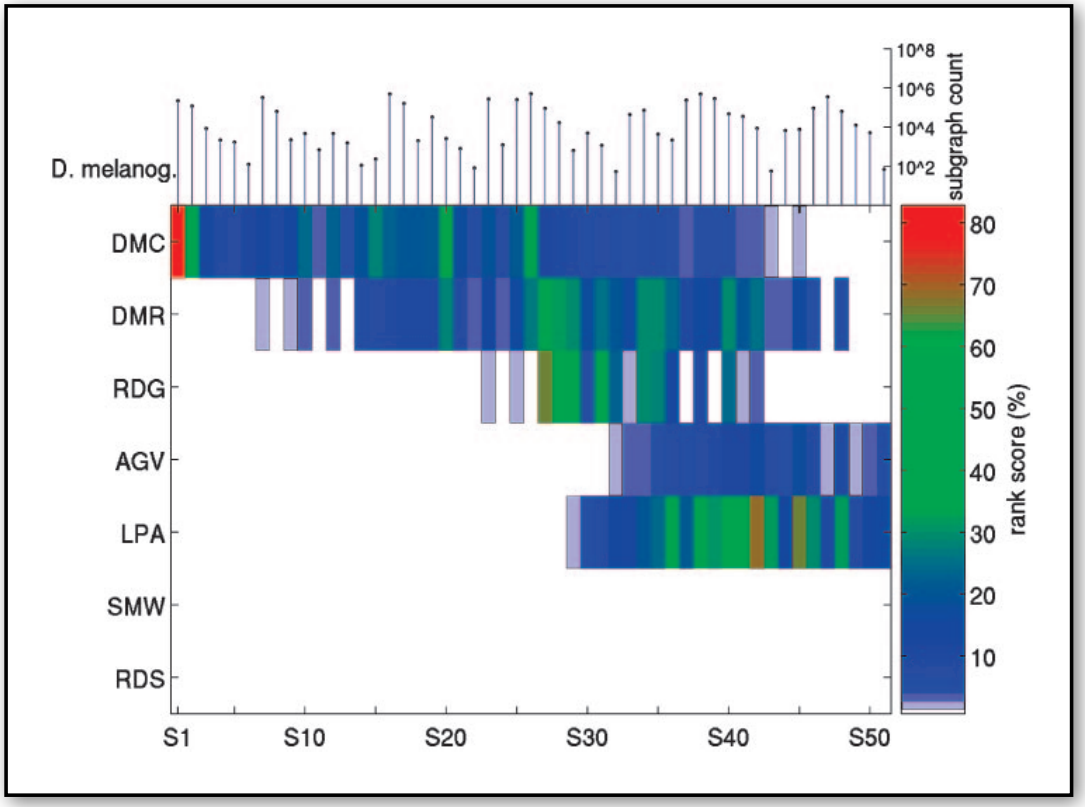
=> Best overlap with DMC (Duplication-mutation, preserved complementarity)

=> Scale-free or random networks are very unlikely

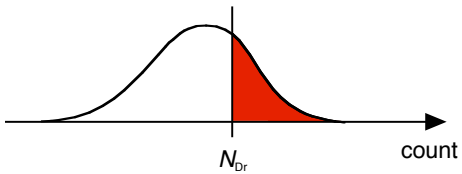
=> what about protein-domain-interaction network of Thomas etal?

Middendorf et al, *PNAS* **102** (2005) 3192

Motif Count Frequencies



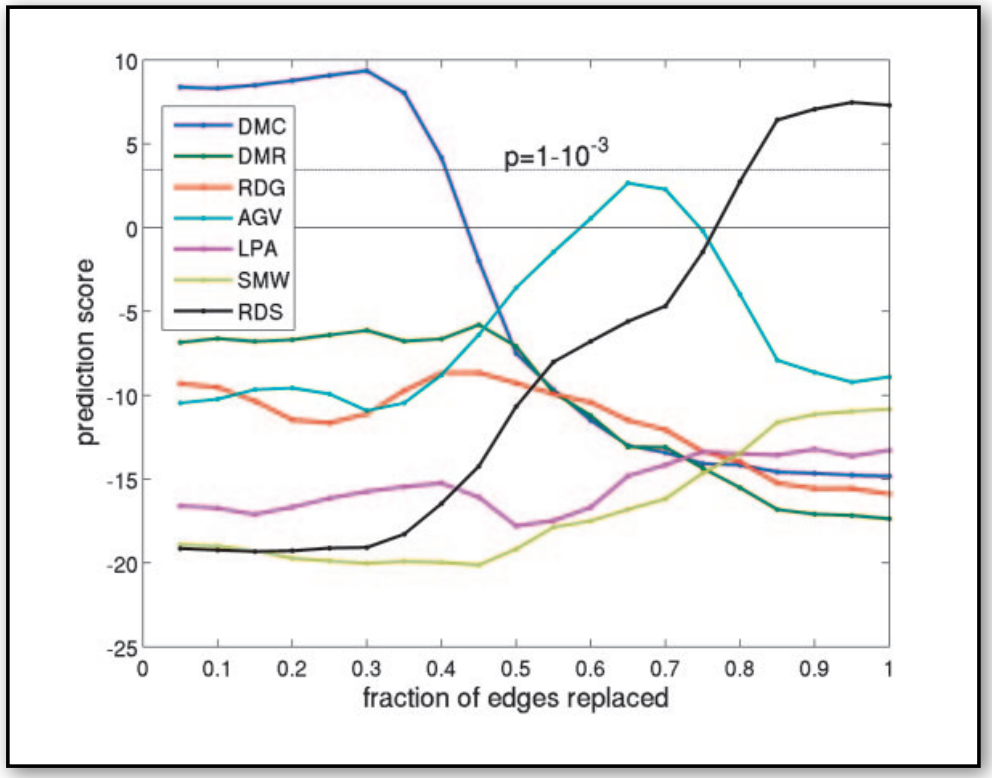
rank score: fraction of test networks with a higher count than Drosophila
(50% = same count as fly on avg.)



Middendorf et al, *PNAS* **102** (2005) 3192

Experimental Errors?

Randomly replace edges in **fly** network and **classify** again:



=> Classification **unchanged** for ≤ **30%** incorrect edges

Elvis is Alive…

… and the Penrose triangle is real.

wikipedia: Penrose-Dreieck

Suggested Reading

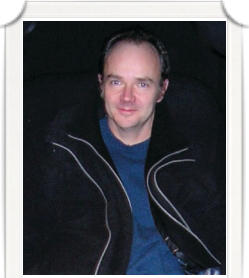
The powerful law of the power law and other myths in network biology†

Gipsi Lima-Mendez* and Jacques van Helden*

Received 5th May 2009, Accepted 12th August 2009
First published as an Advance Article on the web 2nd October 2009
DOI: 10.1039/b908681a

For almost 10 years, topological analysis of different large-scale biological networks (metabolic reactions, protein interactions, transcriptional regulation) has been highlighting some recurrent properties: power law distribution of degree, scale-freeness, small world, which have been proposed to confer functional advantages such as robustness to environmental changes and tolerance to random mutations. Stochastic generative models inspired different scenarios to explain the growth of interaction networks during evolution. The power law and the associated properties appeared so ubiquitous in complex networks that they were qualified as "universal laws". However, these properties are no longer observed when the data are subjected to statistical tests: in most cases, the data do not fit the expected theoretical models, and the cases of good fitting merely result from sampling artefacts or improper data representation. The field of network biology seems to be founded on a series of myths, *i.e.* widely believed but false ideas. The weaknesses of these foundations should however not be considered as a failure for the entire domain. Network analysis provides a powerful frame for understanding the function and evolution of biological processes, provided it is brought to an appropriate level of description, by focussing on smaller functional modules and establishing the link between their topological properties and their dynamical behaviour.

Gipsi Lima-Mendez



Jacques van Helden

Molecular BioSystems **5** (2009)1482

Summary

What you learned **today**: **Sampling matters!**

=> "Scale-free" P(k) by sparse sampling from many network types

Test different **hypotheses** for

- **global** features
  => depends on unkown parameters and sampling
  => no clear statement possible
- **local** features (motifs)
  => better preserved
  => DMC best among tested prototypes

Next lecture:

- Functional annotation of proteins
- Gene regulation networks: how causality spreads