V6 – Biological PPI Networks

- are they really scale-free?
- network growth
- functional annotation in the network

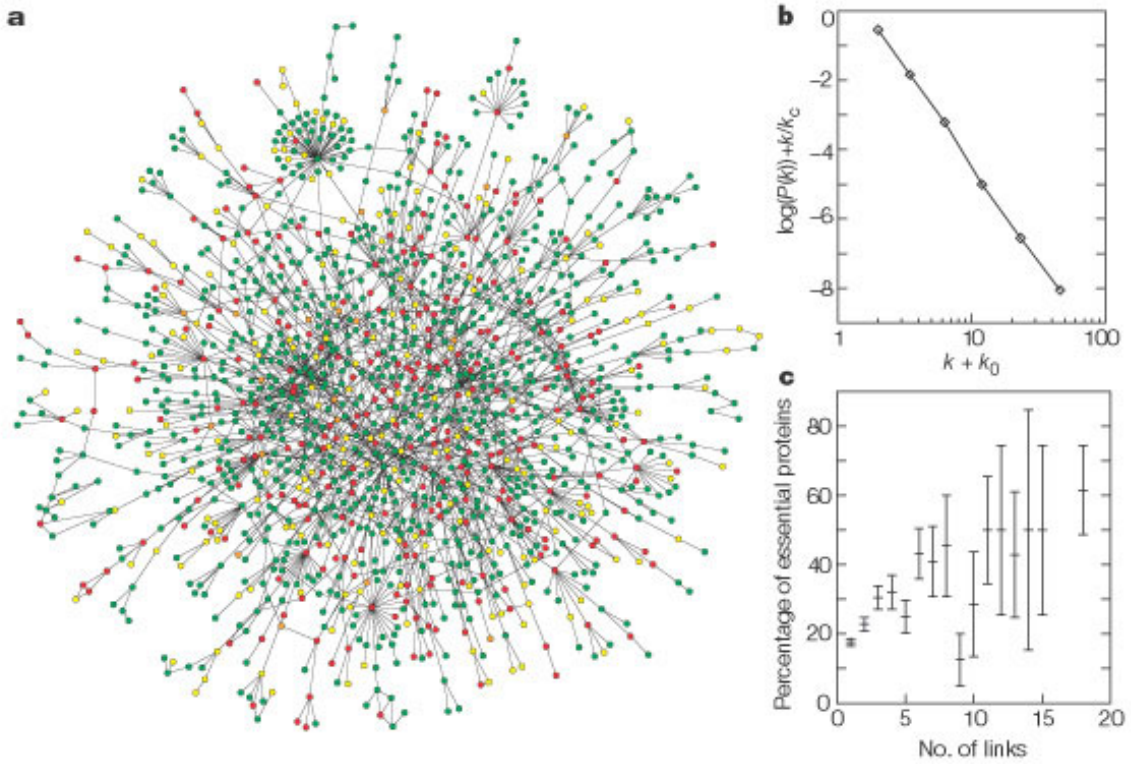
Mon, Nov 14, 2016

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…"

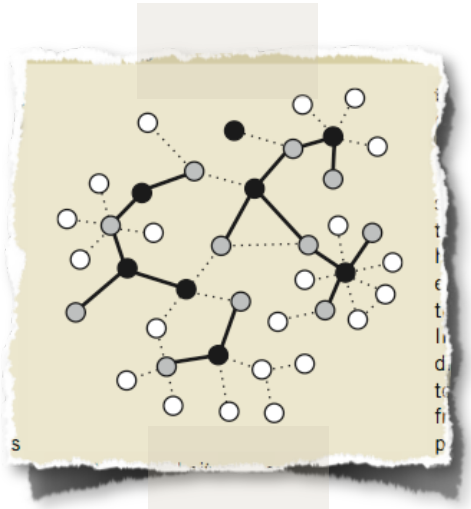
"Are" they scale-free
or
"Do they look like" scale-free???

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
→ determine degree distribution



- Random (ER / Erdös-Renyi) → $P(k)$ = Poisson
- Exponential (EX) → $P(k) \sim \exp[-k]$
- scale-free / power-law (PL) → $P(k) \sim k^{-\gamma}$
- $P(k)$ = truncated normal distribution (TN)

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 experiments **detected** only **3...9%** of the complete interactome!

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

R square

Given: a data set with n values marked $y_1,...,y_n$ and
a set of fitted / predicted / modeled) values $f_1,...,f_n$ e.g. from linear regression.

We call their difference **residuals** $e_i = y_i - f_i$

and the mean value $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The total sum of squares (proportional to the variance of the data) is:

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

The sum of squares of residuals is:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The **coefficient of determination**, R^2 or r^2 is often defined as:

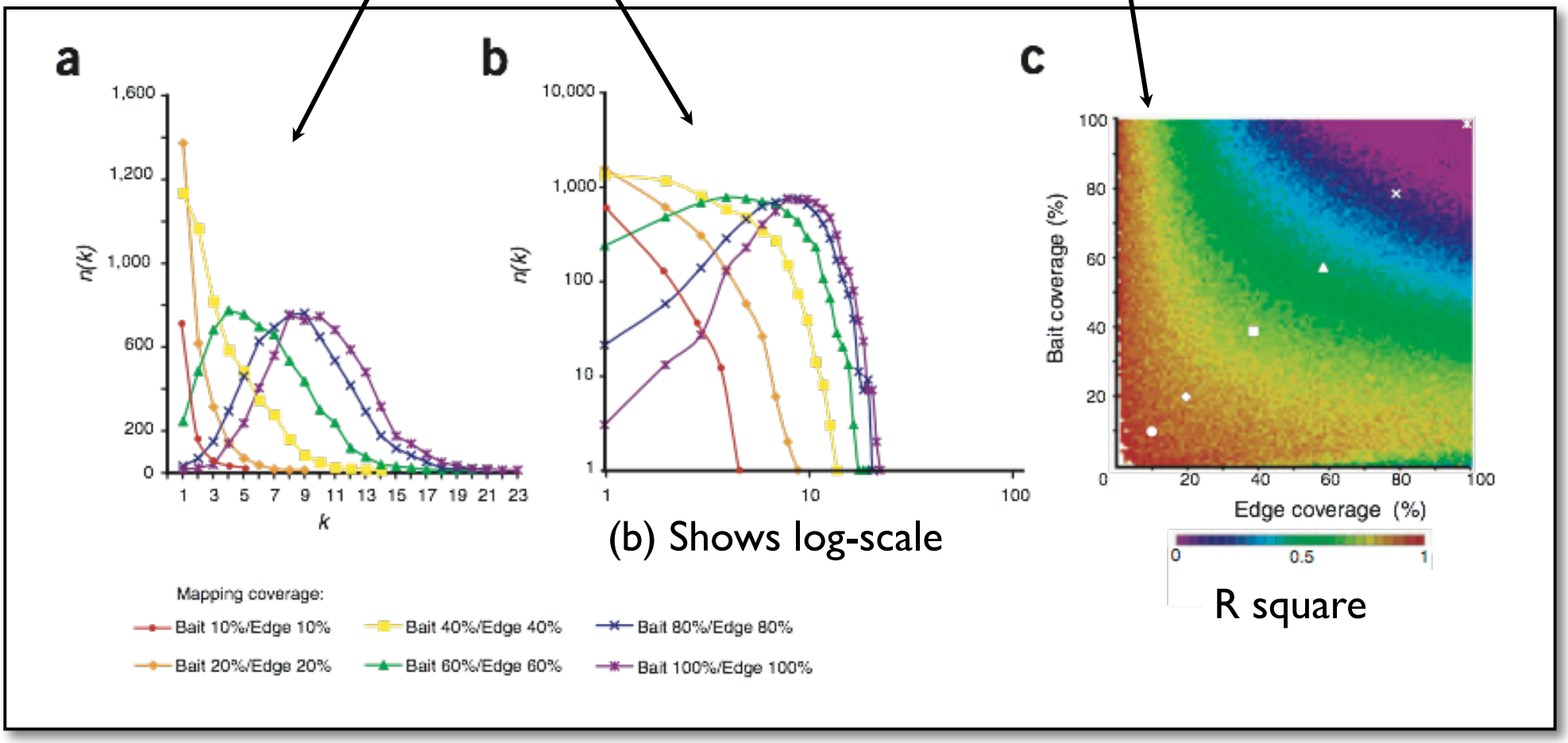
$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

www.wikipedia.org

Sparsely Sampled random (ER) Network

resulting *P(k)* for different coverages

(c) Shows linearity (R square) between detected *P(k)* and ideal power law; good agreement (red; R ≈1 for low edge coverage)

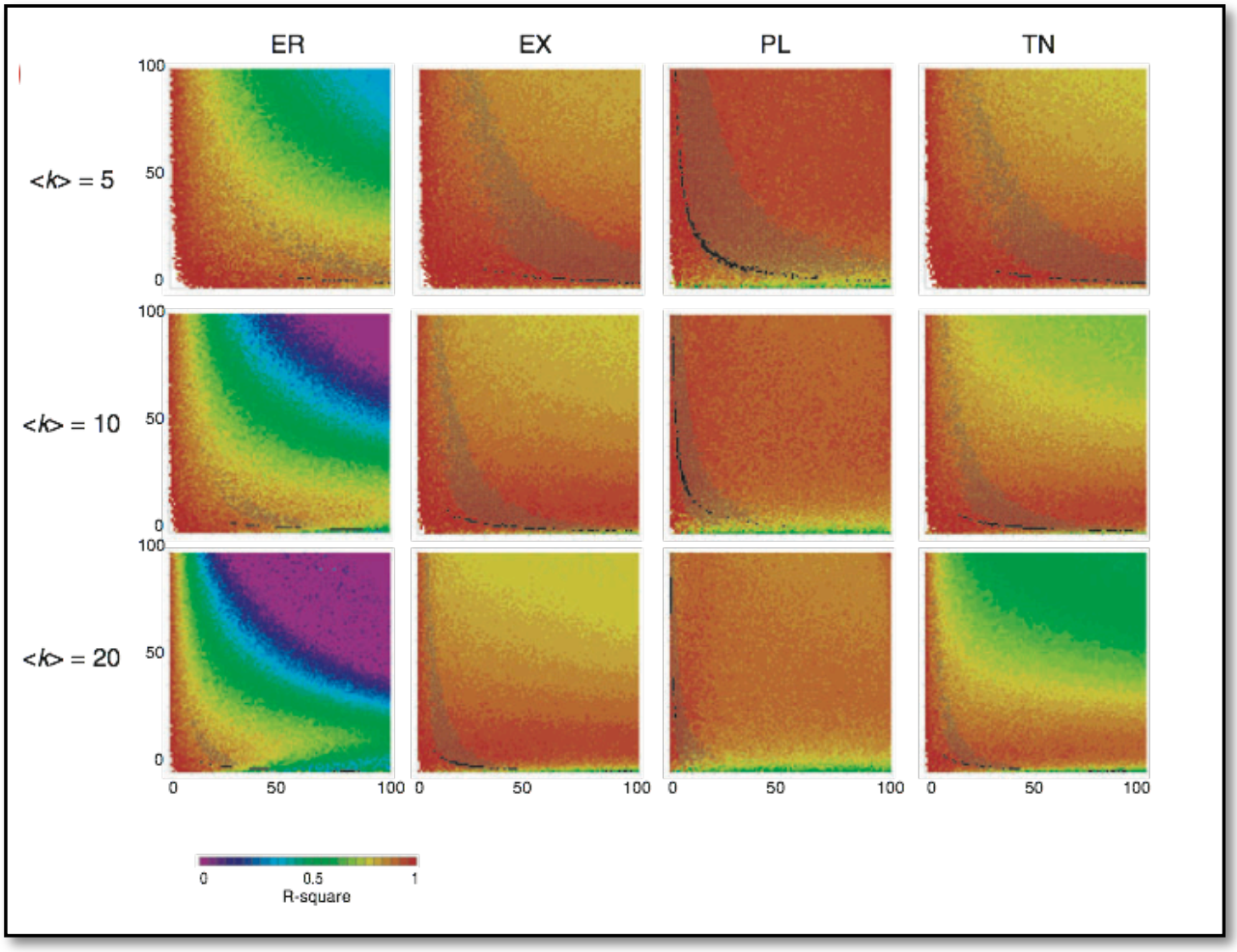


→ for **sparse** sampling (10-20%), even an ER network

"looks" scale-free (when only *P(k)* is considered)

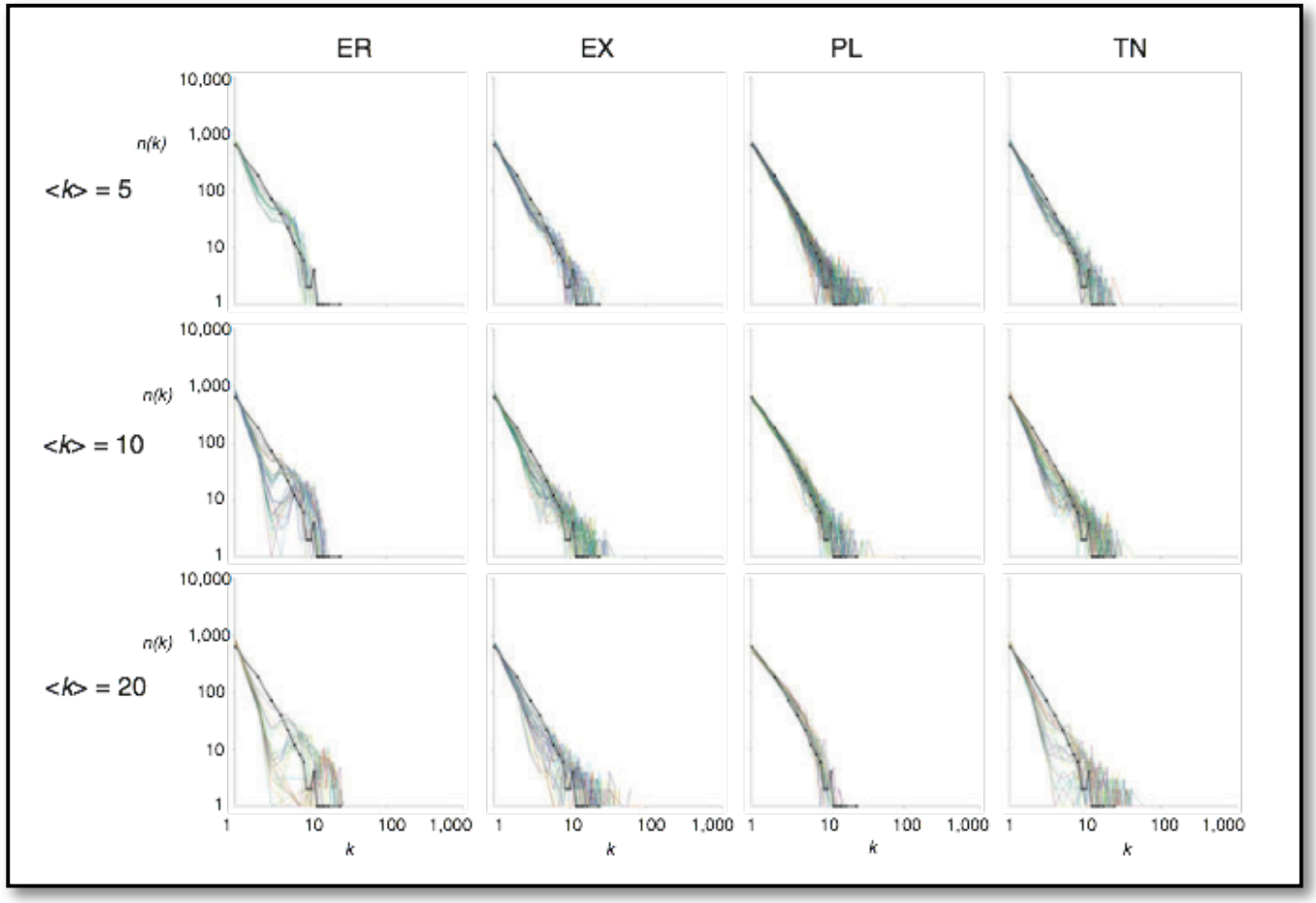
Han et al, *Nature Biotech* **23** (2005) 839

Anything Goes – different topologies



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

Compare to Uetz et al. data



Uetz et al. data (solid line) is compared to sampled networks of similar size.

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

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

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 comparable

The Fly Network

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

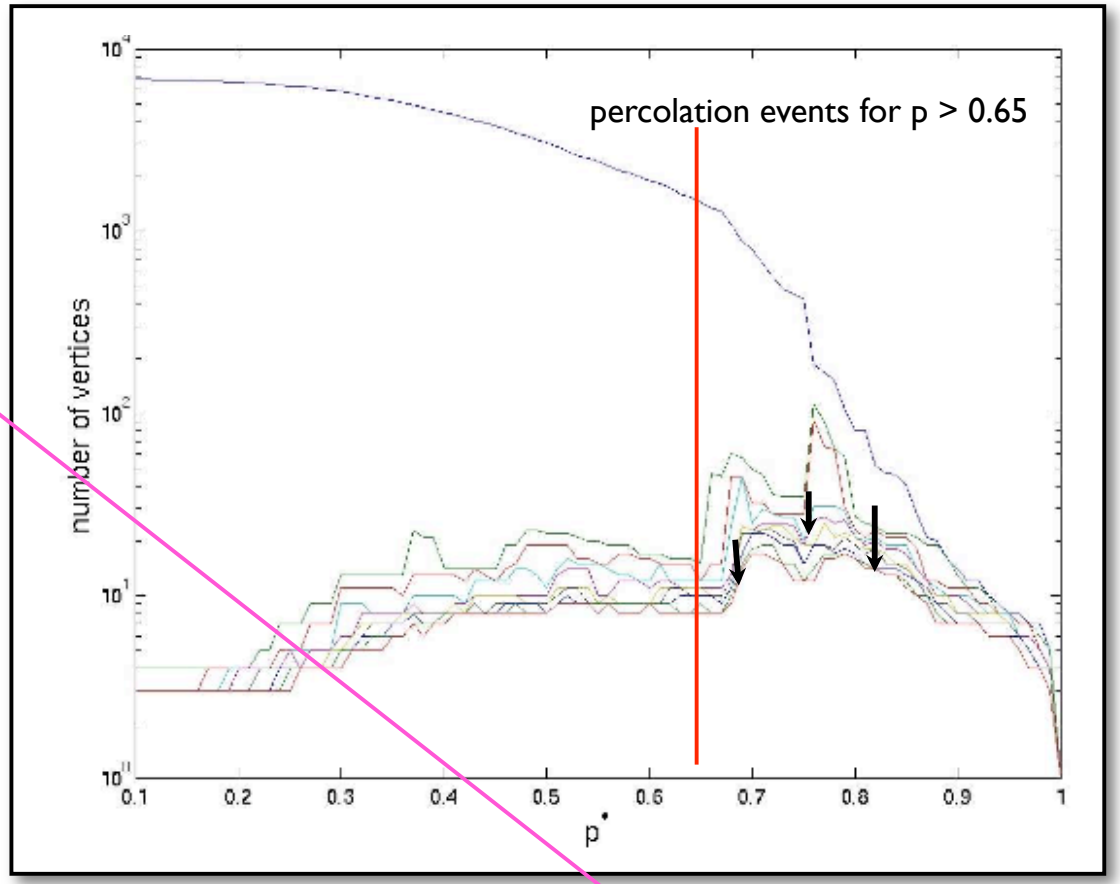
They assigned a confidence score [0, 1] for every observed interaction

→ 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



Size of largest components. At p = 0.65, there is one large component with 1433 nodes and the other 703 components contain at most 15 nodes.

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

Network subgraphs -> 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 7 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 (Albert-Barabasi)
AGV	Aging vertices network
SMW	Small world network

Growth Type 1: DMC

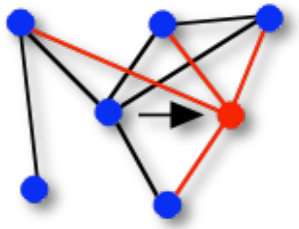
"Duplication – mutation with preserved complementarity"

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

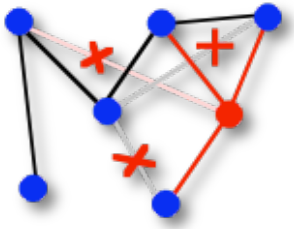
Algorithm:

Start from two connected nodes

- duplicate existing node with all interactions



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



Repeat these steps many (e.g. $N - 2$) times

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 ($k_i + \alpha$), $\alpha = 0 \ldots 5$
(BA for $\alpha = 0$)

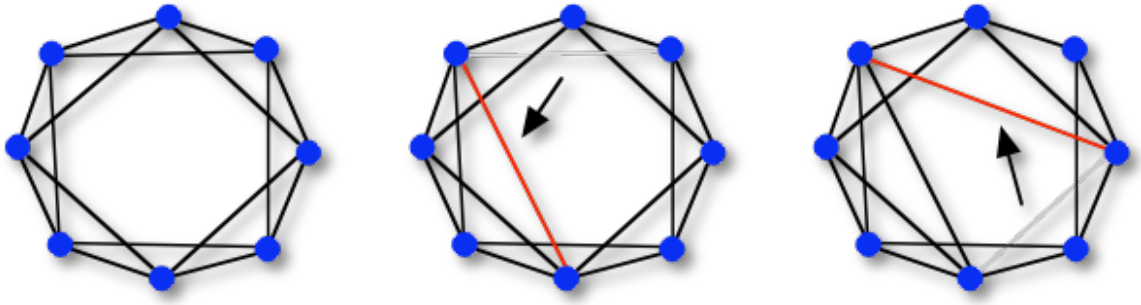
Growth Types 6-7: AGV and SMW

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)

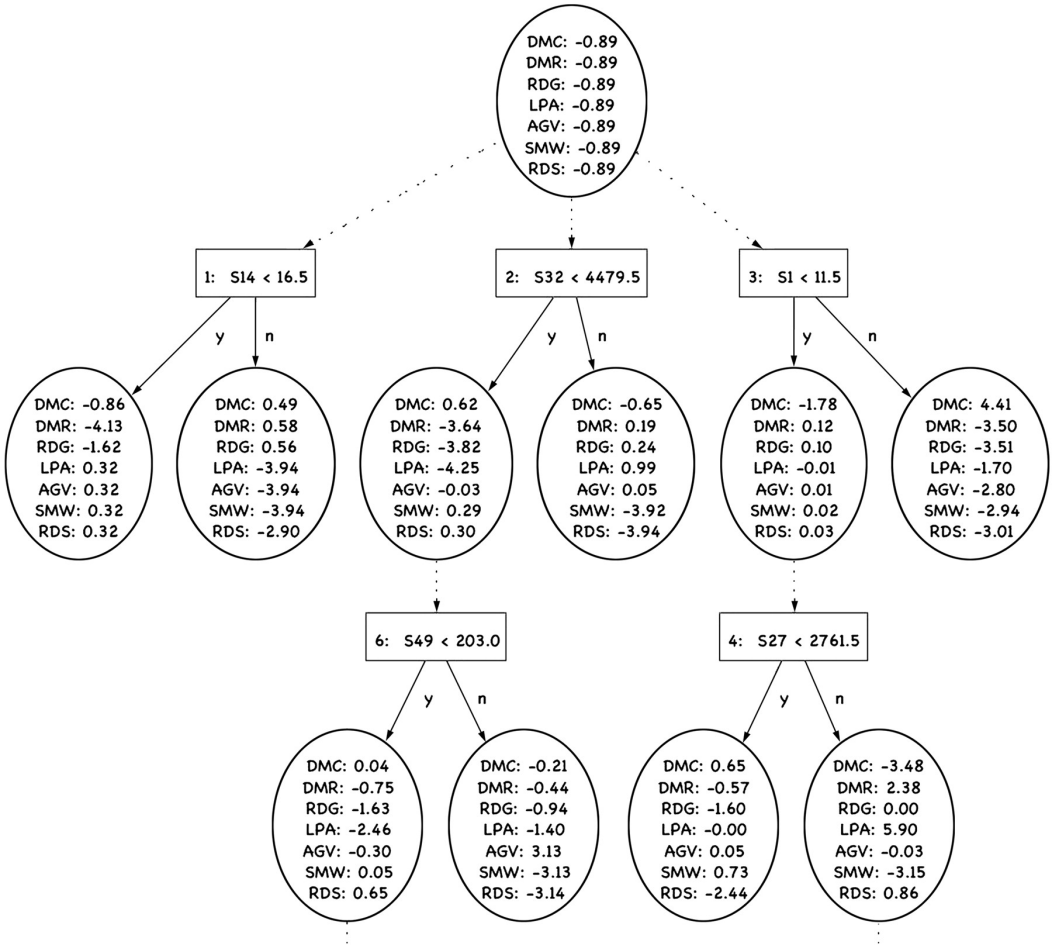
Randomly rewire regular ring lattice



Alternating Decision Tree Classifier

Trained with the motif counts from 1000 networks of each of the 7 types

→ prototypes are well separated and reliably classified



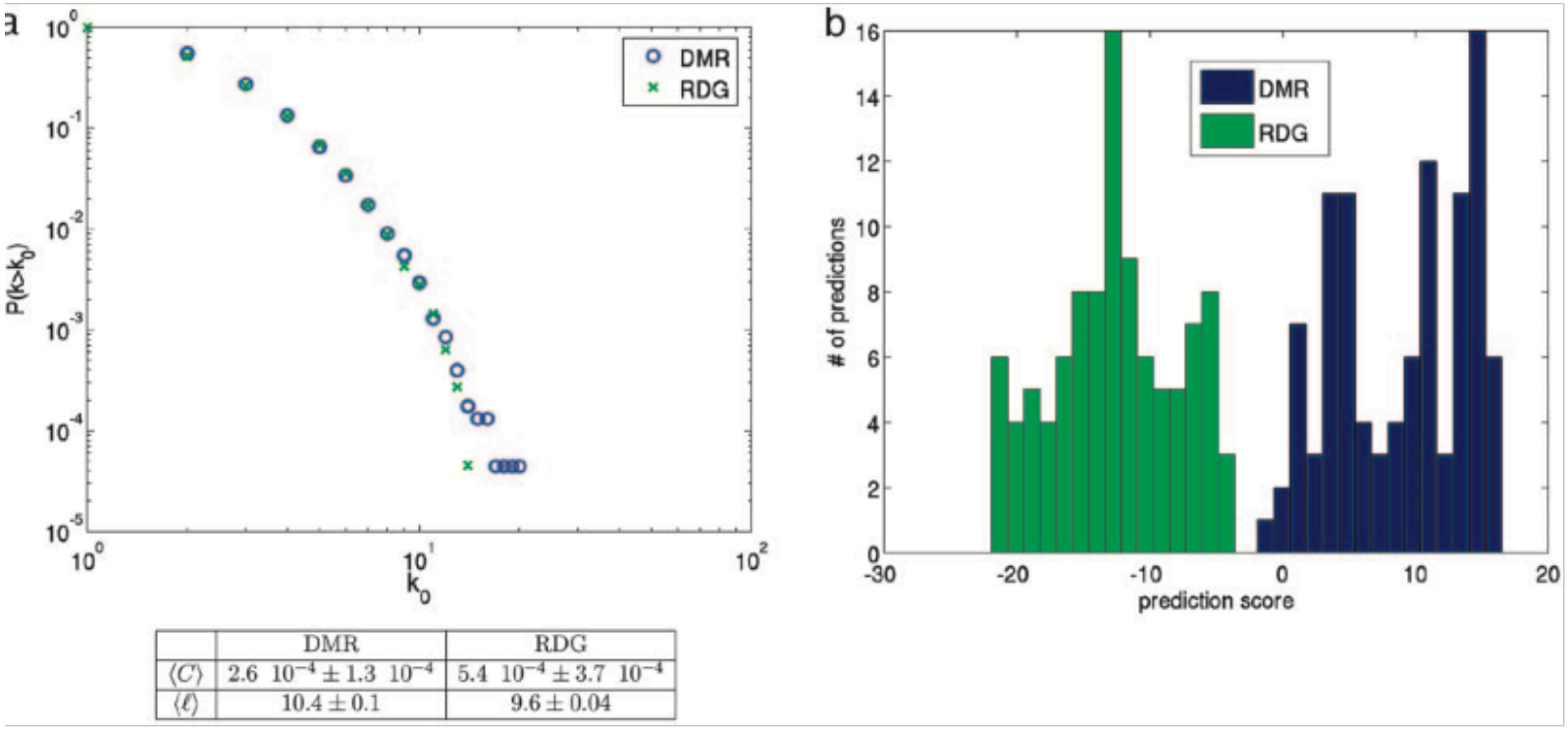
Decision nodes count occurrence of subgraphs

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 the generated networks 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 <C> and <l> (left),
but different counts of the network motifs (right)

-> networks can (only) be perfectly separated by motif-based classifier

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

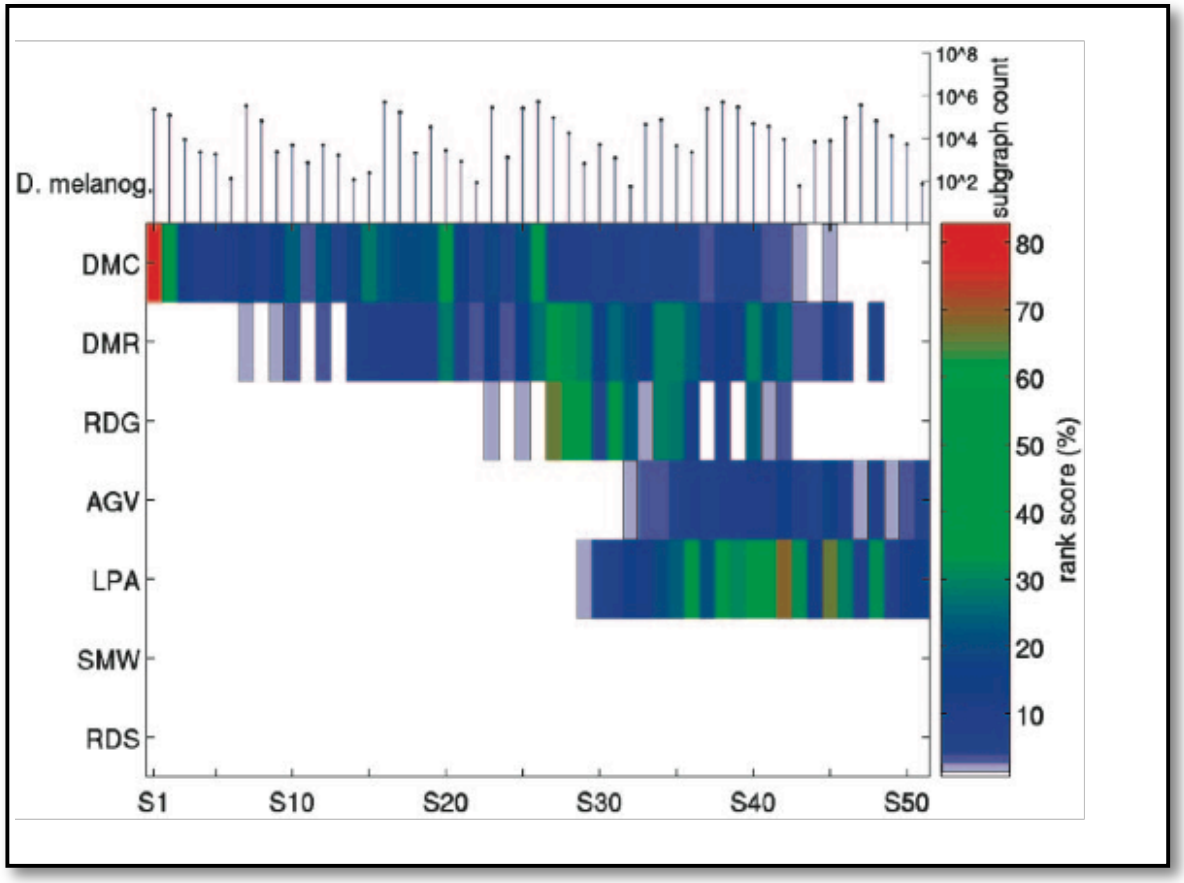
Drosophila is consistently (independently of the cut-off in subgraph size) classified as a DMC network, with an especially strong prediction for a confidence threshold of $p^* = 0.65$.

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

→ Scale-free or random networks are very unlikely

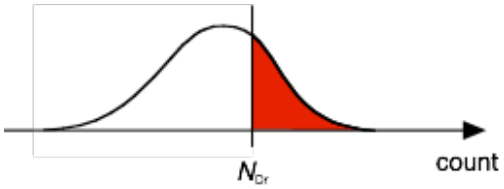
Middendorf et al, *PNAS* **102** (2005) 3192

Motif Count Frequencies



-> DMC and DMR networks contain most subgraphs in similar amount as fly network (top).

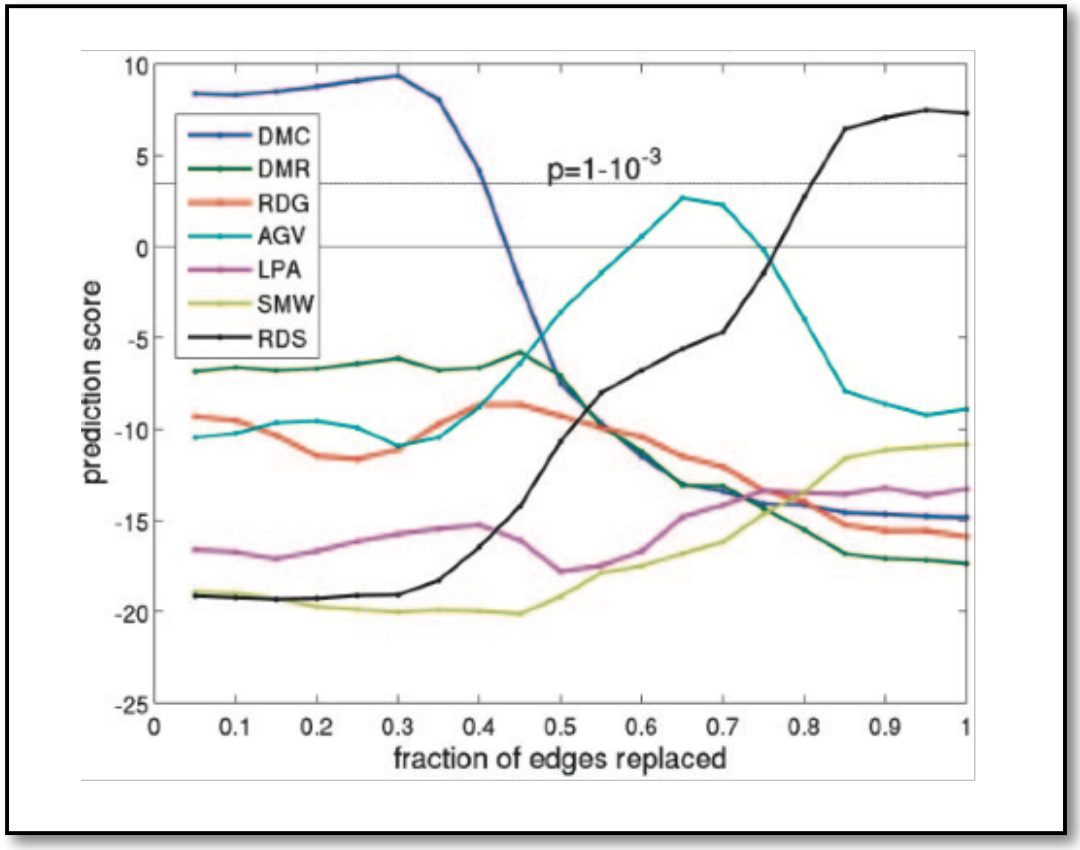
rank score: fraction of test networks with a higher count than Drosophila
(50% = same count as fly on avg.)



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

Experimental Errors?

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



→ Classification **unchanged** for ≤ **30%** incorrect edges,
at higher values RDS takes over (as to be expected)

Summary (I)

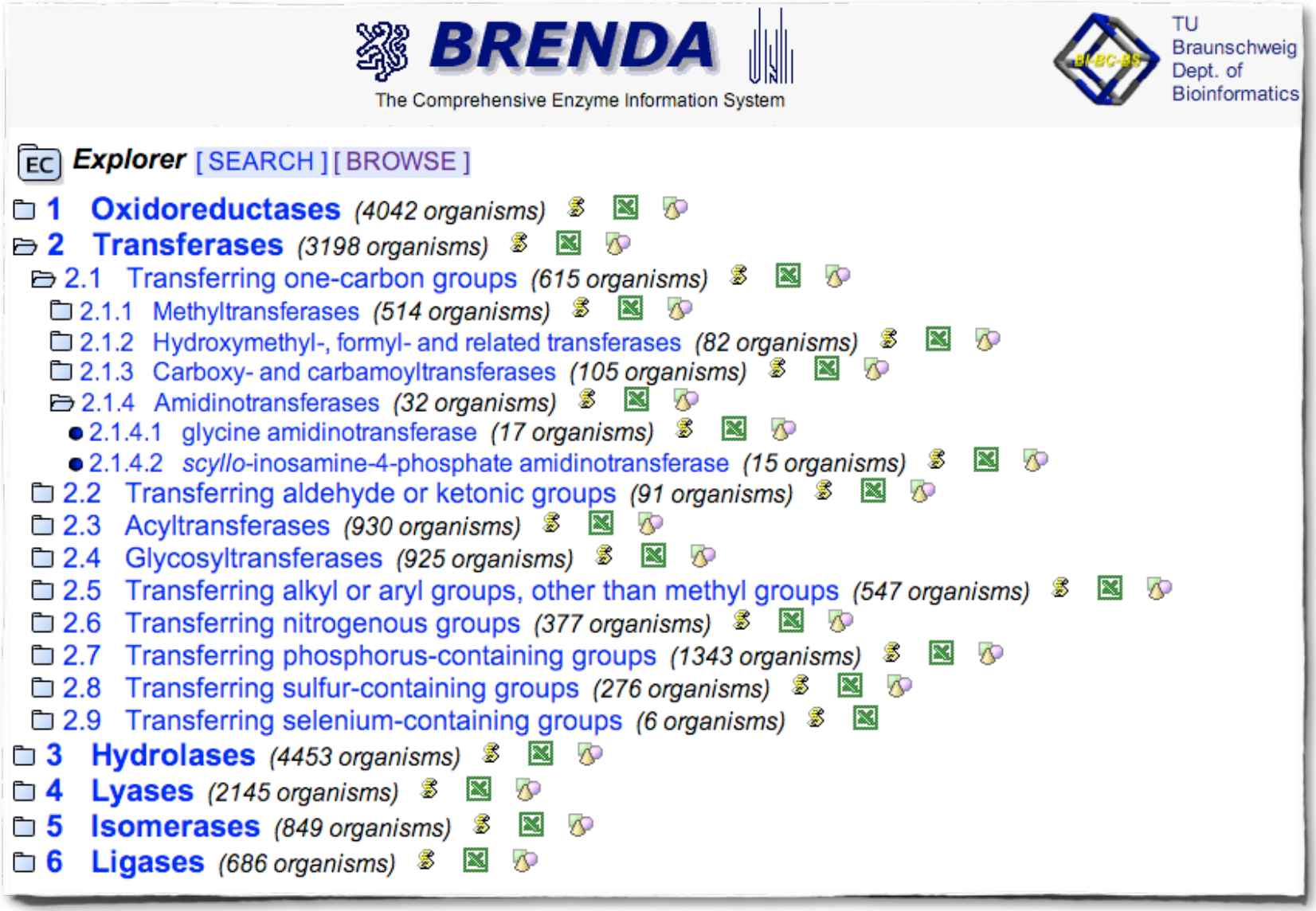
Sampling matters!

→ "Scale-free" *P(k)* obtained by sparse sampling
from many network types

Test different **hypotheses** for

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

What Does a Protein Do?



Enzyme **C**lassification scheme
(from http://www.brenda-enzymes.org/)

What about Un-Classified Proteins?

BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i302–i310
doi:10.1093/bioinformatics/bti1054

Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps

Elena Nabieva[1,2], Kam Jim[2], Amit Agarwal[1], Bernard Chazelle[1] and Mona Singh[1,2,]*

[1]Computer Science Department and [2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

Received on January 15, 2005; accepted on March 27, 2005

Many **unclassified proteins**:

→ estimate: ~1/3 of the yeast proteome not annotated functionally

→ BioGRID: 4495 proteins in the largest cluster of the yeast physical interaction map.

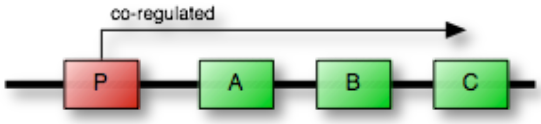
only 2946 have a MIPS functional annotation

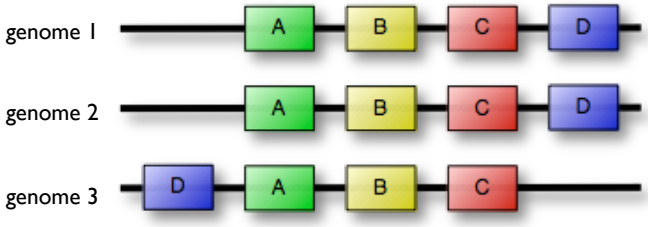
Partition the Graph

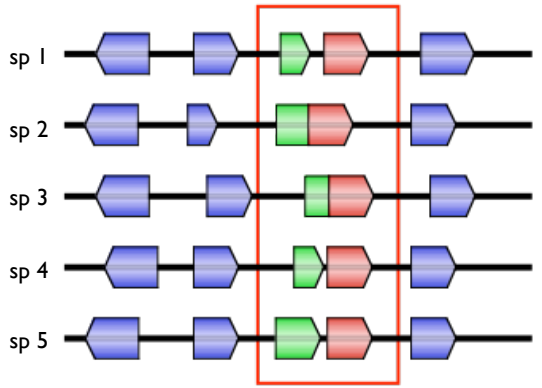
Large **PPI networks** can be built from:

- HT experiments (Y2H, TAP, synthetic lethality, coexpression, coregulation, …)
- predictions (gene profiling, gene neighborhood, phylogenetic profiles, …)

→ proteins that are functionally linked





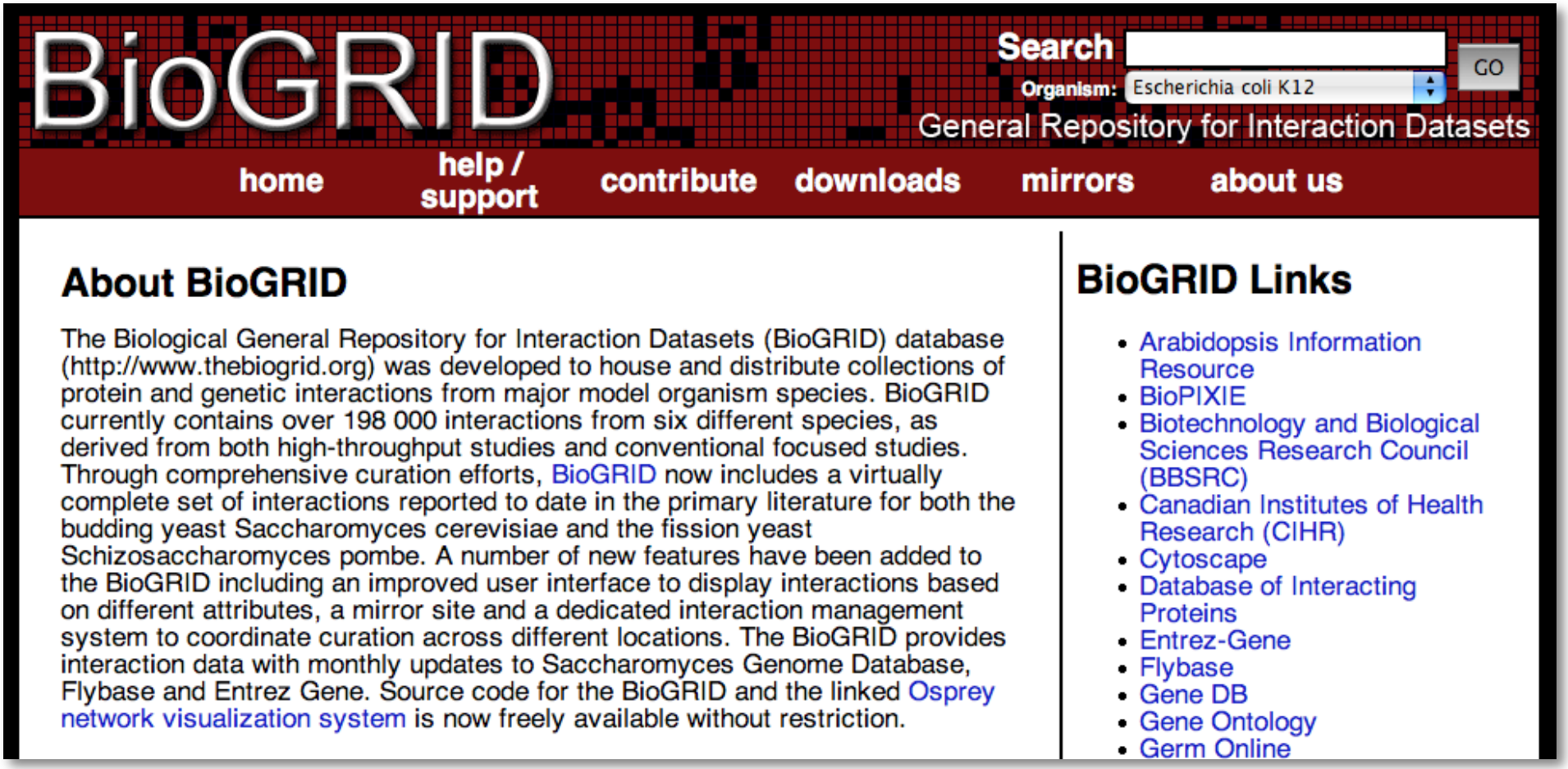


Identify **unknown functions** from **clustering** of these networks by, e.g.:

- shared interactions (similar neighborhood)
- membership in a community
- similarity of shortest path vectors to all other proteins (= similar path into the rest of the network)

Protein Interactions

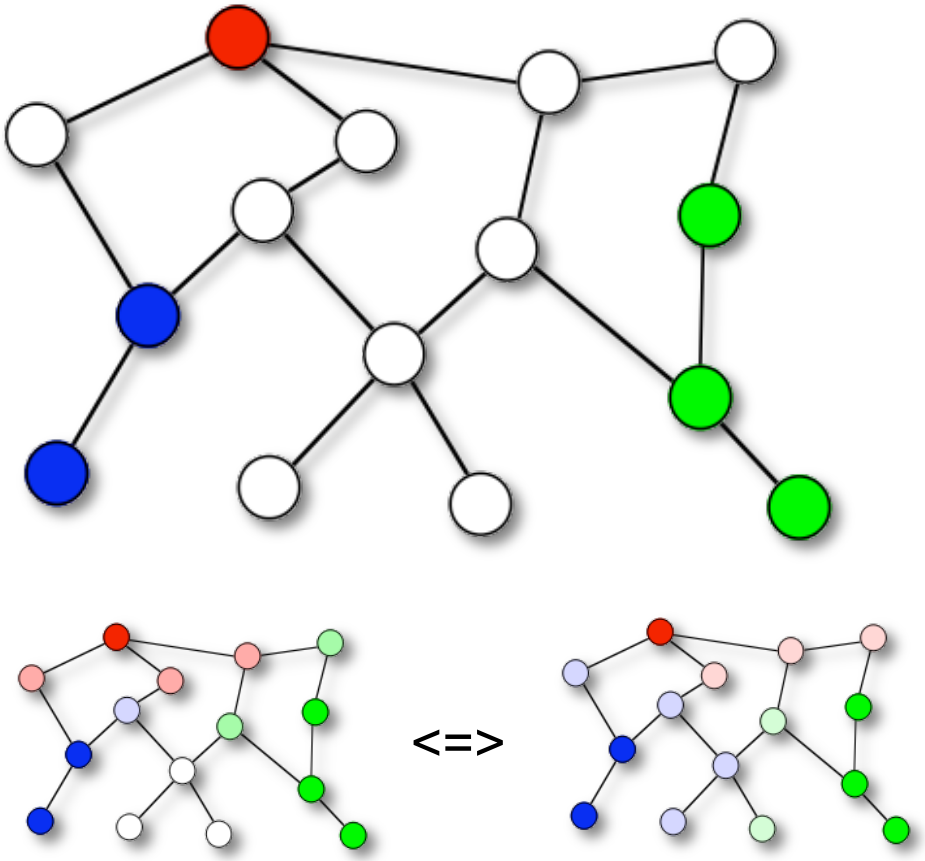
Nabieva et al used the *S. cerevisiae* dataset from GRID of 2005 (now BioGRID)
→ 4495 proteins and 12 531 physical interactions in the largest cluster



http://www.thebiogrid.org/about.php

Function Annotation

Task: **predict** function (= functional annotation) for an unlabeled protein from the **available** annotations of other proteins in the network



Similar task:
How to **assign colors** to the white nodes?

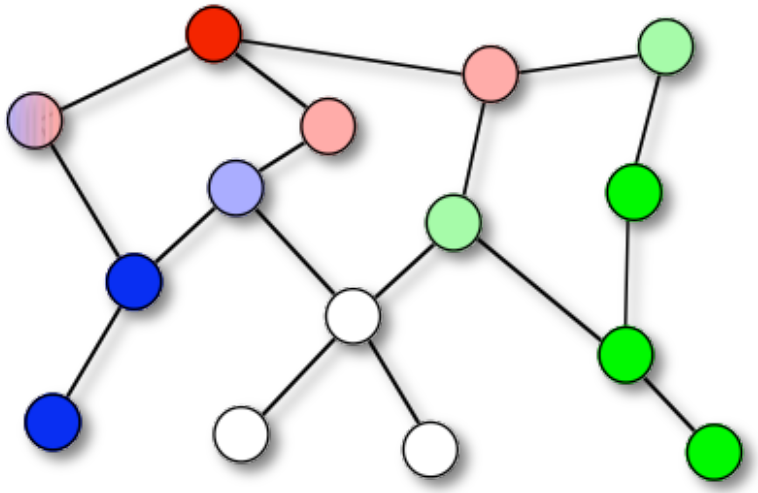
Use information on:

- distance to colored nodes
- local connectivity
- reliability of the links
- …

Algorithm I: Majority

Schwikowski, Uetz, and Fields, " A network of protein–protein interactions in yeast"
Nat. Biotechnol. **18** (2000) 1257

Consider all neighbors and **sum up** how often a certain **annotation occurs**
→ score for an annotation = count among the direct neighbors
→ take the 3 most frequent functions



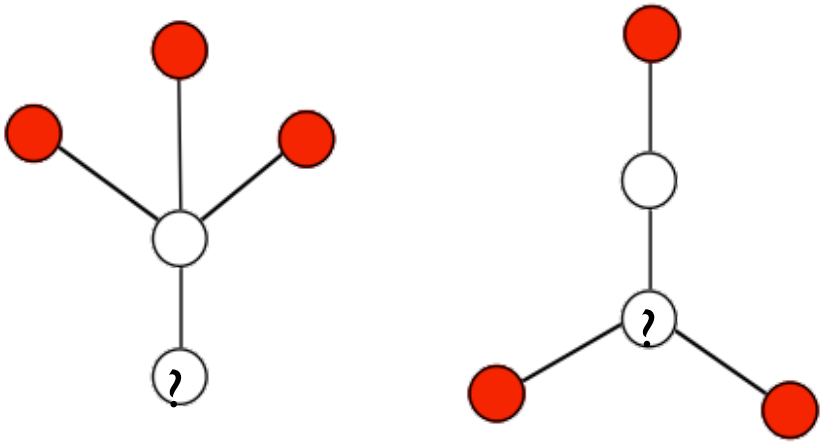
Majority makes only limited use of the local connectivity
→ cannot assign function to next-neighbors

For weighted graphs:
→ weighted sum

Extended Majority: Neighborhood

Hishigaki, Nakai, Ono, Tanigami, and Takagi, "Assessment of prediction accuracy of protein function from protein–protein interaction data",
Yeast **18** (2001) 523

Look for **overrepresented** functions within a given **radius** of 1, 2, or 3 links
→ use as function score the value of a χ^2–test



Neighborhood algorithm does not consider local network topology

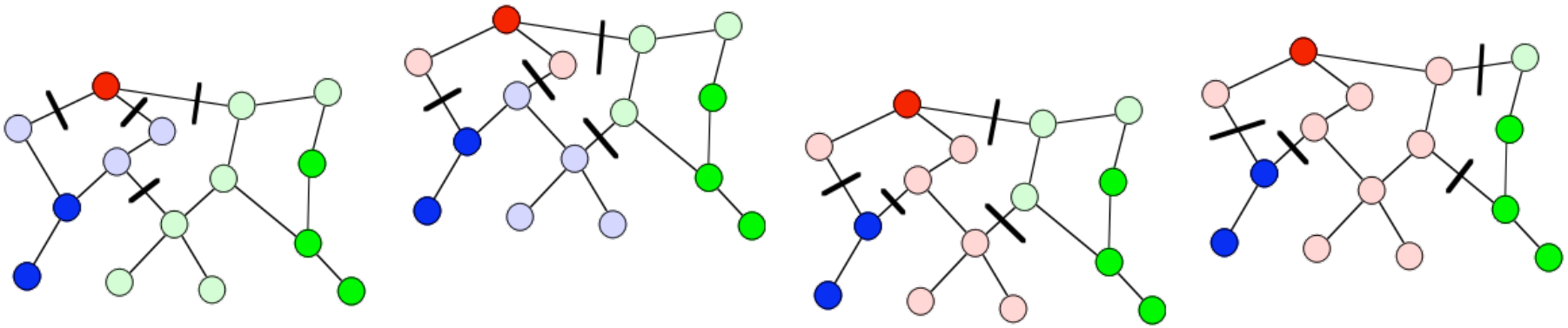
Both examples (left) are treated **identically** with $r = 2$

Minimize Changes: GenMultiCut

Karaoz, Murali, Letovsky, Zheng, Ding, Cantor, and Kasif, "Whole-genome annotation by using evidence integration in functional-linkage networks"
PNAS **101** (2004) 2888

"Annotate proteins so as to **minimize** the number of times that **different** functions are associated to **neighboring** (i.e. interacting) proteins"

→ generalization of the multiway *k*-cut problem for weighted edges,
can be stated as an integer linear program (ILP)



Multiple possible solutions → scores from **frequency** of annotations

Nabieva *et al*: FunctionalFlow

Extend the idea of **"guilty by association"**

→ each annotated protein is considered as a source of "function"-flow

→ propagate/simulate for a few time steps

→ choose the annotation *a* with the highest accumulated flow

Each node *u* has a reservoir $R_t(u)$, each edge a capacity constraint (weight) $w_{u,v}$

Initially: $R_0^a(u) = \begin{cases} \infty, & \text{if } u \text{ is annotated with } a, \\ 0, & \text{otherwise.} \end{cases}$ and $g_0^a(u,v) = 0$

Then: **downhill flow** from node *u* to node *v* with capacity constraints

$$g_t^a(u,v) = \begin{cases} 0, & \text{if } R_{t-1}^a(u) < R_{t-1}^a(v) \\ \min\left(w_{u,v}, \frac{w_{u,v}}{\sum_{(u,y)\in E} w_{u,y}}\right), & \text{otherwise.} \end{cases}$$

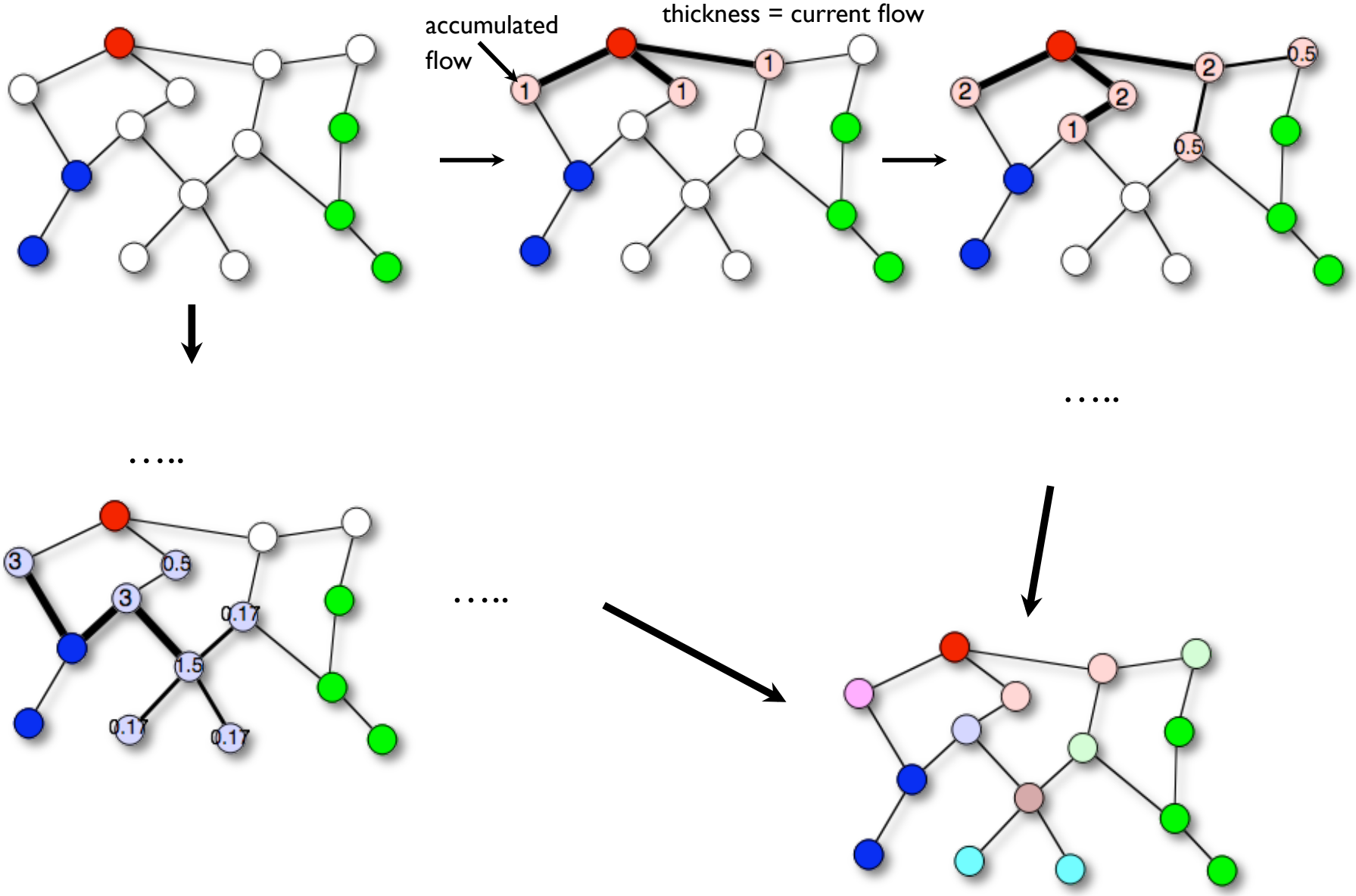
Idea: Node v has already „more function" than node u → no flow uphill

Score from accumulated in-flow:

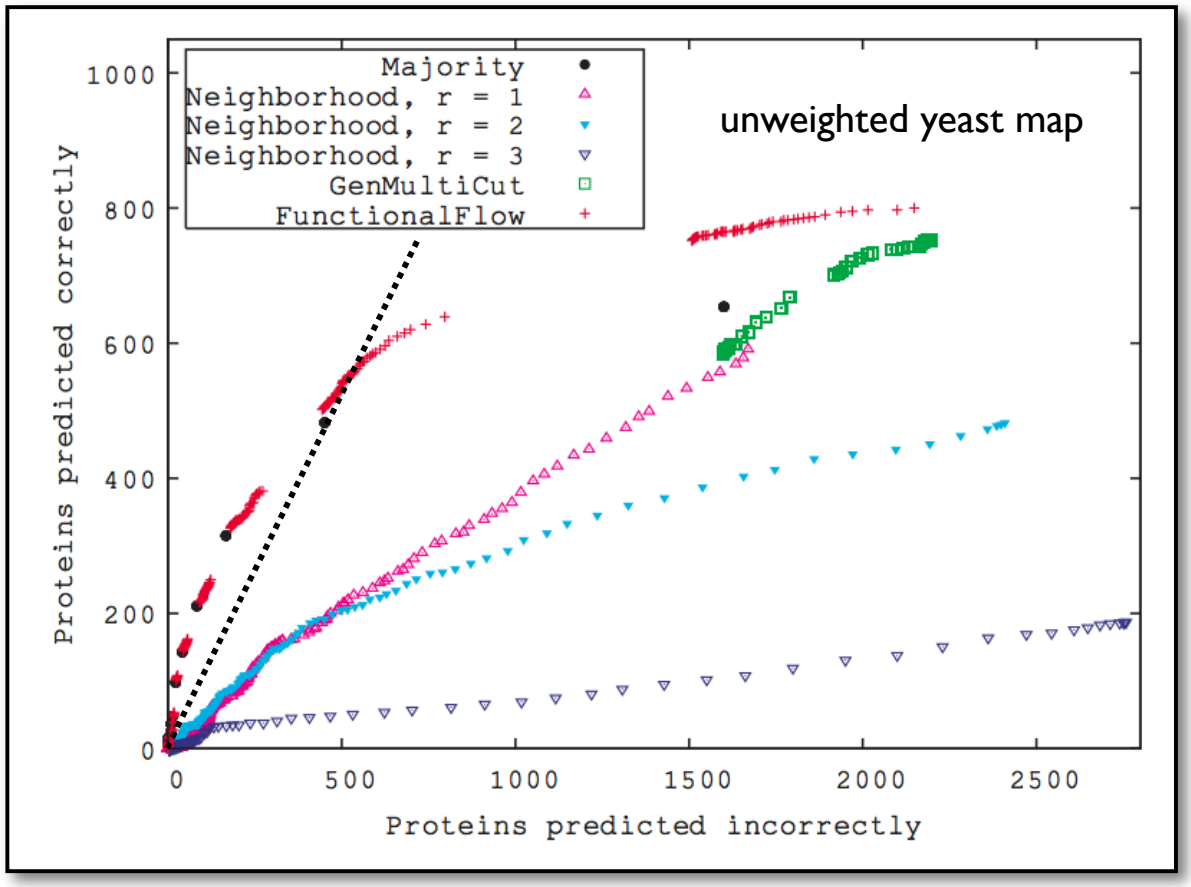
$$f_a(u) = \sum_{t=1}^{d} \sum_{v:(u,v)\in E} g_t^a(v,u)$$

Nabieva et al, Bioinformatics 21 (2005) i302

An Example


accumulated
flow
thickness = current flow
1
1
1
2
2
1
2
0.5
0.5
.....
.....
3
0.5
3
0.17
1.5
0.17
0.17
.....

Comparison



For FunctionalFlow:
six propagation steps were simulated; this is comparable to the diameter of the yeast network ≈ 12

Majority results are initially very good, but reduced coverage

Results with neighborhood get more unprecise for larger radii r

Change **score threshold** for accepting annotations → ratio **TP/FP**
→ **FunctionalFlow** performs **best** in the high-confidence region
→ but many false predictions!!!

Nabieva et al, Bioinformatics 21 (2005) i302

Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks

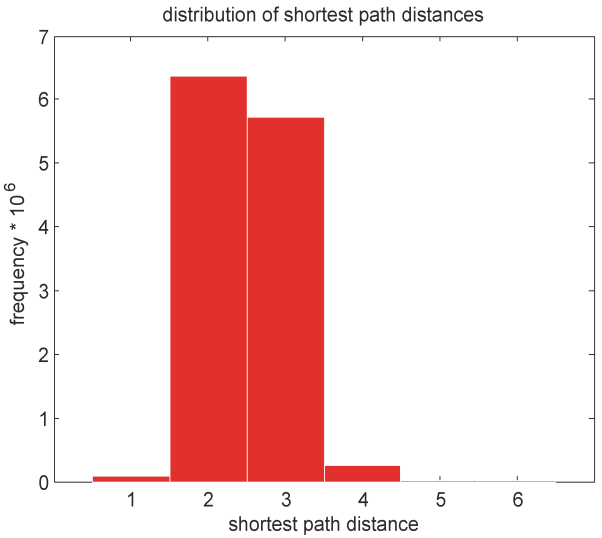
Citation: Cao M, Zhang H, Park J, Daniels NM, Crovella ME, et al. (2013) Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks. PLoS ONE 8(10): e76339. doi:10.1371/journal.pone.0076339

Relying on the ordinary shortest-path distance metric in PPI networks is problematic because PPI networks are "small world" networks.
Most nodes are close to all other nodes.

→ any method that infers similarity based on proximity will find that a large fraction of the network is proximate to any typical node.

Largest connected component of *S. cerevisiae* PPI network (BioGRID) has 4990 nodes and 74,310 edges (physical interactions).

Right Fig. shows the histogram of shortest-path lengths from this network. Over 95% of all pairs of nodes are either 2 hops or 3 hops apart



(a)

What nodes mediate short contacts?

The 2-hop neighborhood of a typical node
probably includes around half of **all nodes** in the graph.

One of the **reasons** that paths are typically short in biological networks
like the PPI network is due to the **presence of hubs**.

But hub proteins often represent proteins with
different functional roles than their neighbors.

Hub proteins also likely have multiple, distinct functions.

→not all short paths provide equally strong evidence
of similar function in PPI networks.

DSD Distance Metric

Given some fixed $k>0$, we define $He^{\{k\}}(A,B)$ to be the expected number of times that a random walk starting at A and proceeding for k steps, will visit B. If there is no ambiguity about *k*, we can drop *k*.

$$He(v_i)=(He(v_i,v_1),He(v_i,v_2),\ldots,He(v_i,v_n))$$

He(v_i) is a „random walk distance vector" of node v_i from all other nodes.

$$DSD(u,v)=||He(u)-He(v)||_1 \quad \text{where}$$

$||He(u)-He(v)||_1$ denotes the L_1 norm of the He vectors

Two nodes *u* and *v* have small DSD if they have similar distance from all other nodes.

Explanation:

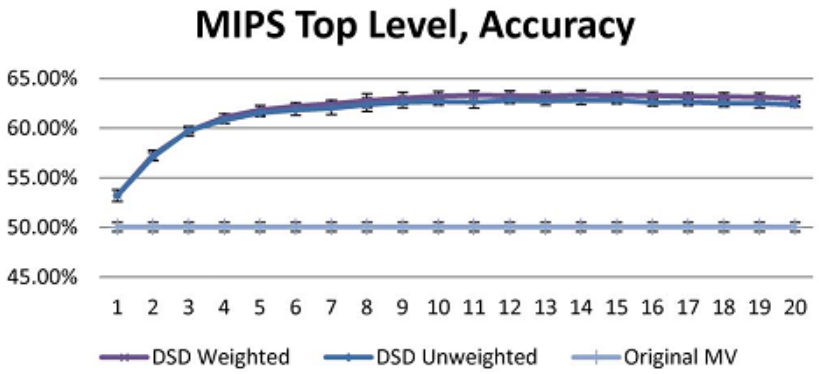
The one-norm (also known as the L_1-norm, ℓ_1 norm, or mean norm) of a vector $\vec{v}$ is denoted $\|\vec{v}\|_1$ and is defined as the sum of the absolute values of its components:

$$\|\vec{v}\|_1 = \sum_{i=1}^{n} |v_i| \tag{1}$$

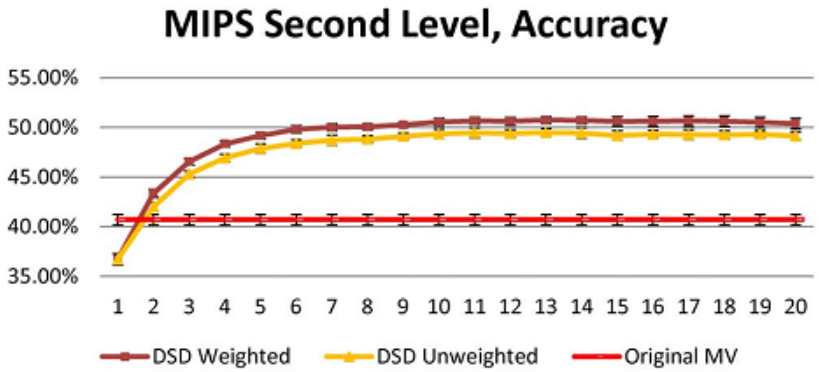
for example, given the vector $\vec{v} = (1, -4, 5)$, we calculate the one-norm:

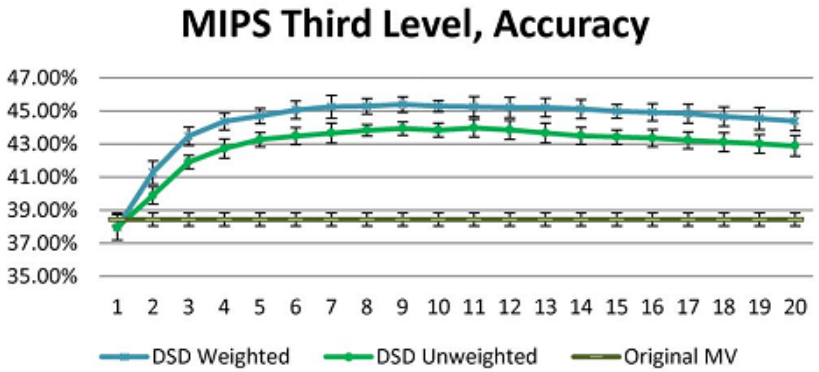
$$\|(1,-4,5)\|_1 = |1| + |-4| + |5| = 10$$

DSD clearly improves functional predictions



MV: majority voting





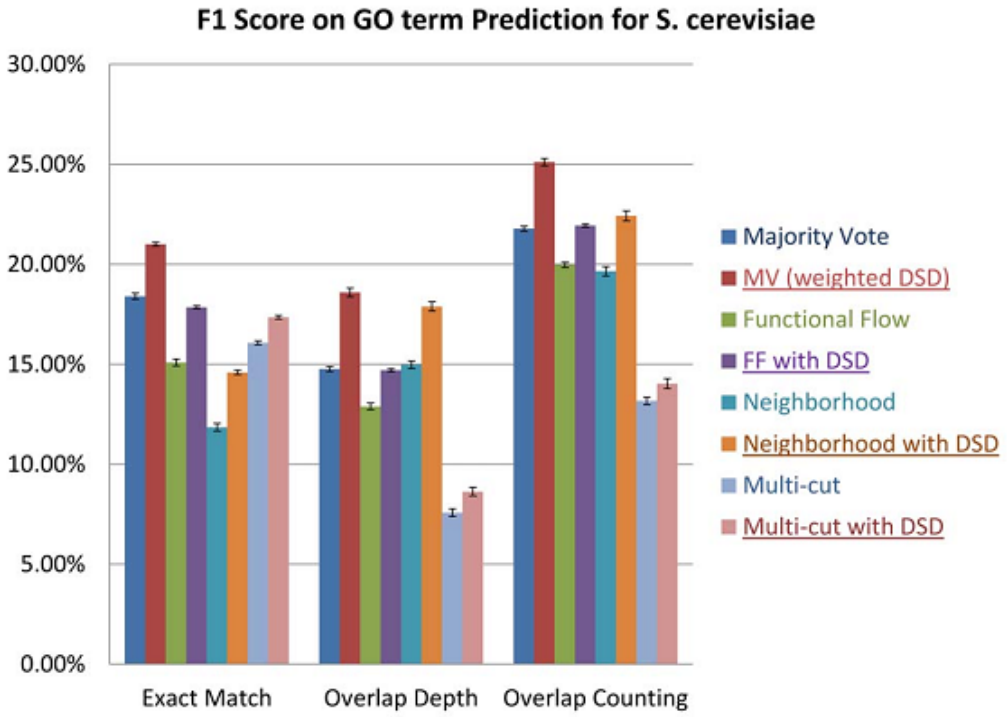


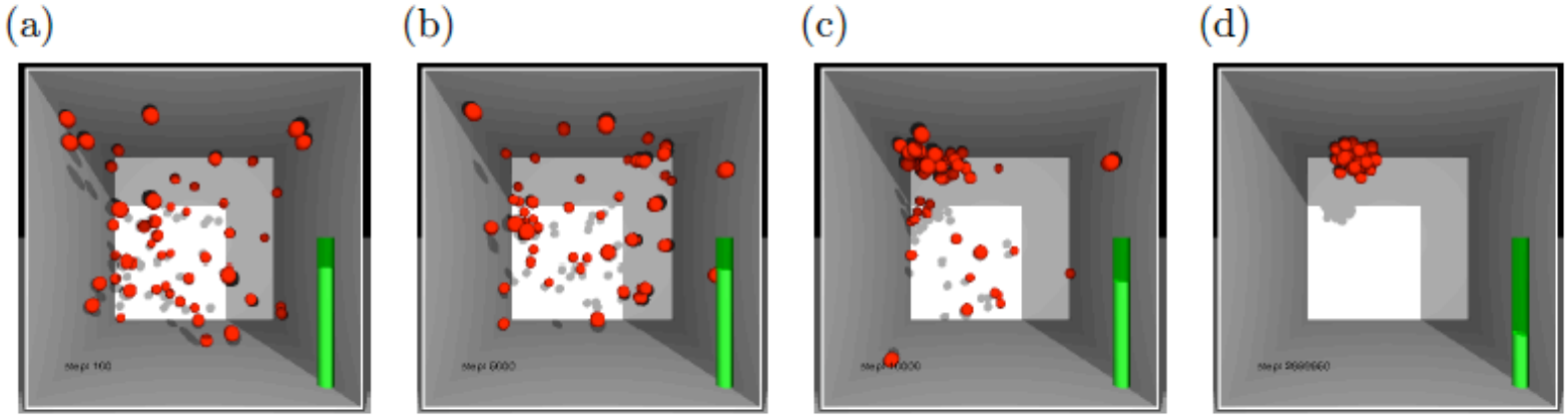
Figure 6. Improvement on F1 Score for DSD using three evaluation methods: exact match, overlap depth and overlap counting, on informative GO terms for the four algorithms for *S. cerevisiae* in 10 runs of 2-fold cross validation.

What you can else do with Interaction graphs?

E.g. efficiently track interactions
between many particles
in dynamic simulations

Strongly attracting particles form large “blob”

(a) to (d) are 4 snapshots of a simulation with ca. N = 50 interacting particles in a box.



How can one analyze the particle connectivity efficiently?

M.Sc. thesis

Florian Lauck (2006)

For i = 1 to N - 1

  For j = i + 1 to N

    For k = j + 1 to N

      If (i .is bound to. j) then

        If (j .is bound to. k) then ….

this is impractical!

Map simulation to interaction graph

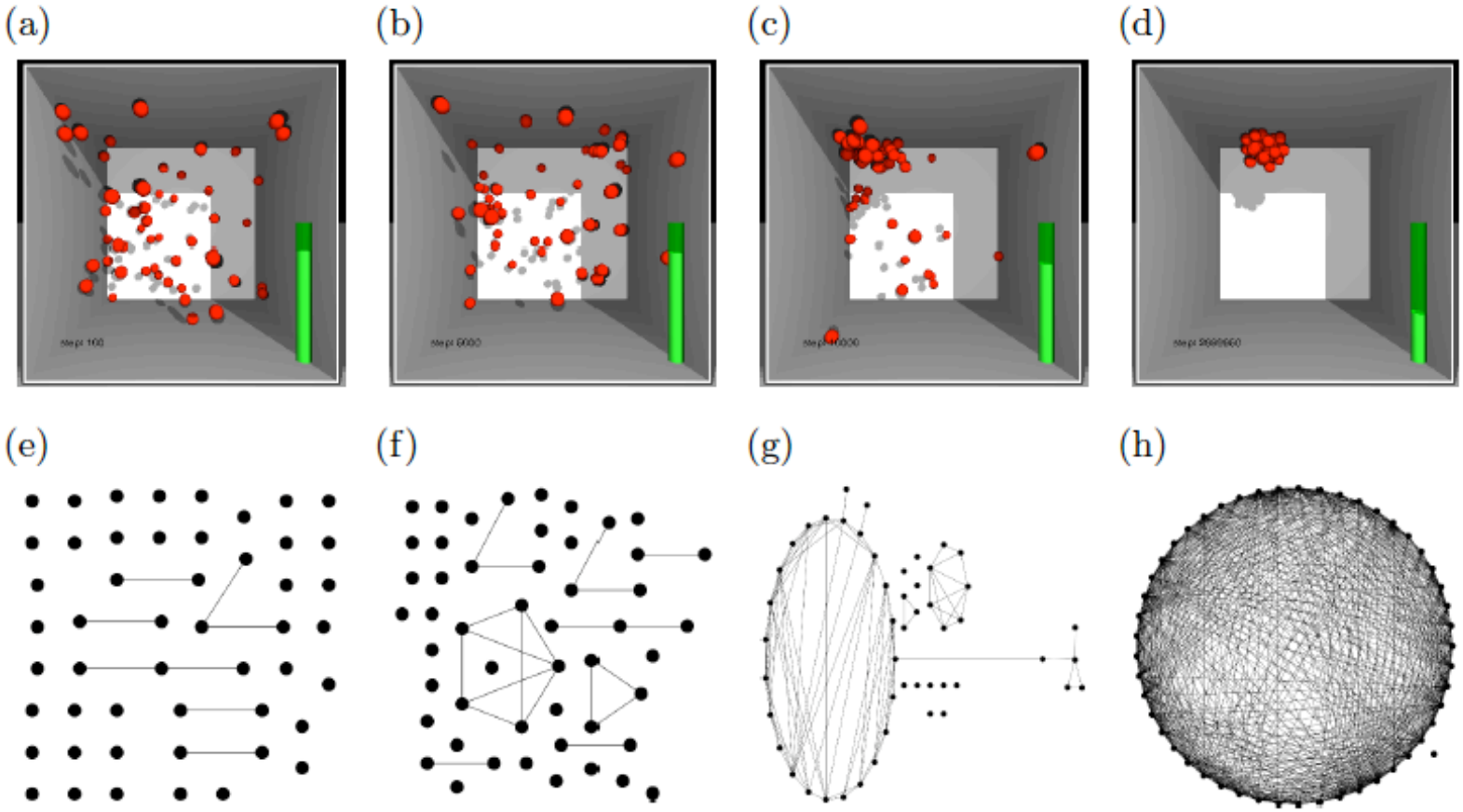
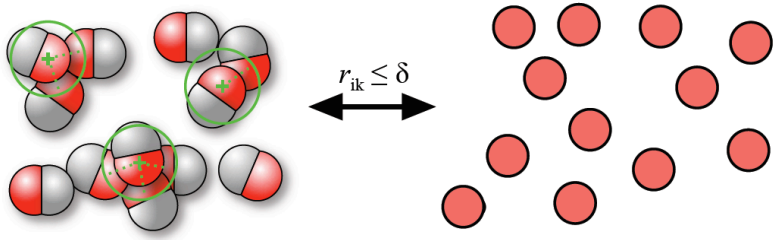


Figure 2.7: Graph and spatial view of a simulation with 50 particles at four different points in time. The green bar denotes the energy of the system.

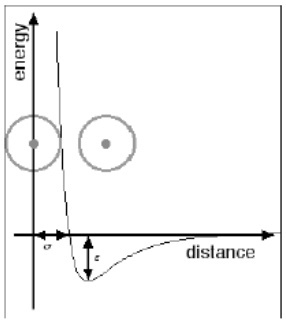
M.Sc. thesis Florian Lauck (2006)

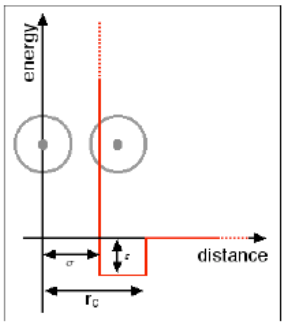
Large number of simultaneous assocications: map simulations to interaction graphs

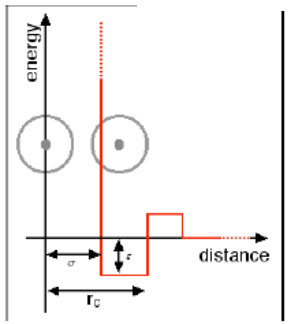


Simple MC scheme
for diffusion + association/
dissociation

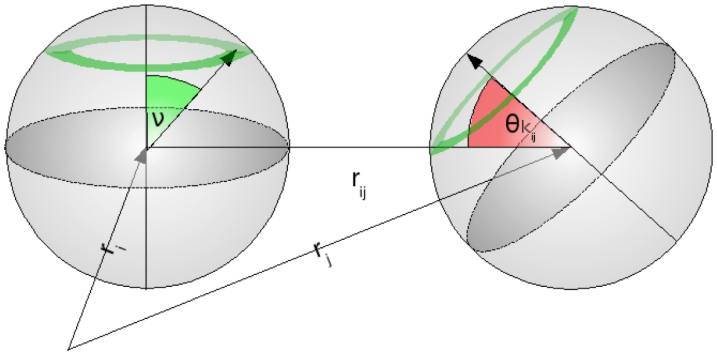
```
function INITIALIZE(N)
    for P ∈ List of Particles do
        CREATE RANDOM COORDINATES(P)
    CREATE GRAPH(N)
for all Iterations do
    for P ∈ List of Particles do
        MOVE AND ROTATE(P)
        for all P_i ∈ (List of Particles - P) do
            d = DISTANCE(P, P_i)
            e_i = POTENTIAL(d)
            if d ≤ r_C then APPEND(List of Interactions, (P, P_i))
            E_new += e_i
        a = TRANSITION PROBABILITY(E_new, E_old)
        x = RANDOMNUMBER

        if x ≤ p then                                  ▷ accept new state
            APPEND(List of ALL interactions, List of Interactions)
            E_old = E_new
        else                                           ▷ discard new state
            RESET(P) CLEAR(List of Interactions)
    UPDATE(Graph, List of ALL Interactions)
    ANALYSIS(Graph)
```







Interaction patches define complex geometry



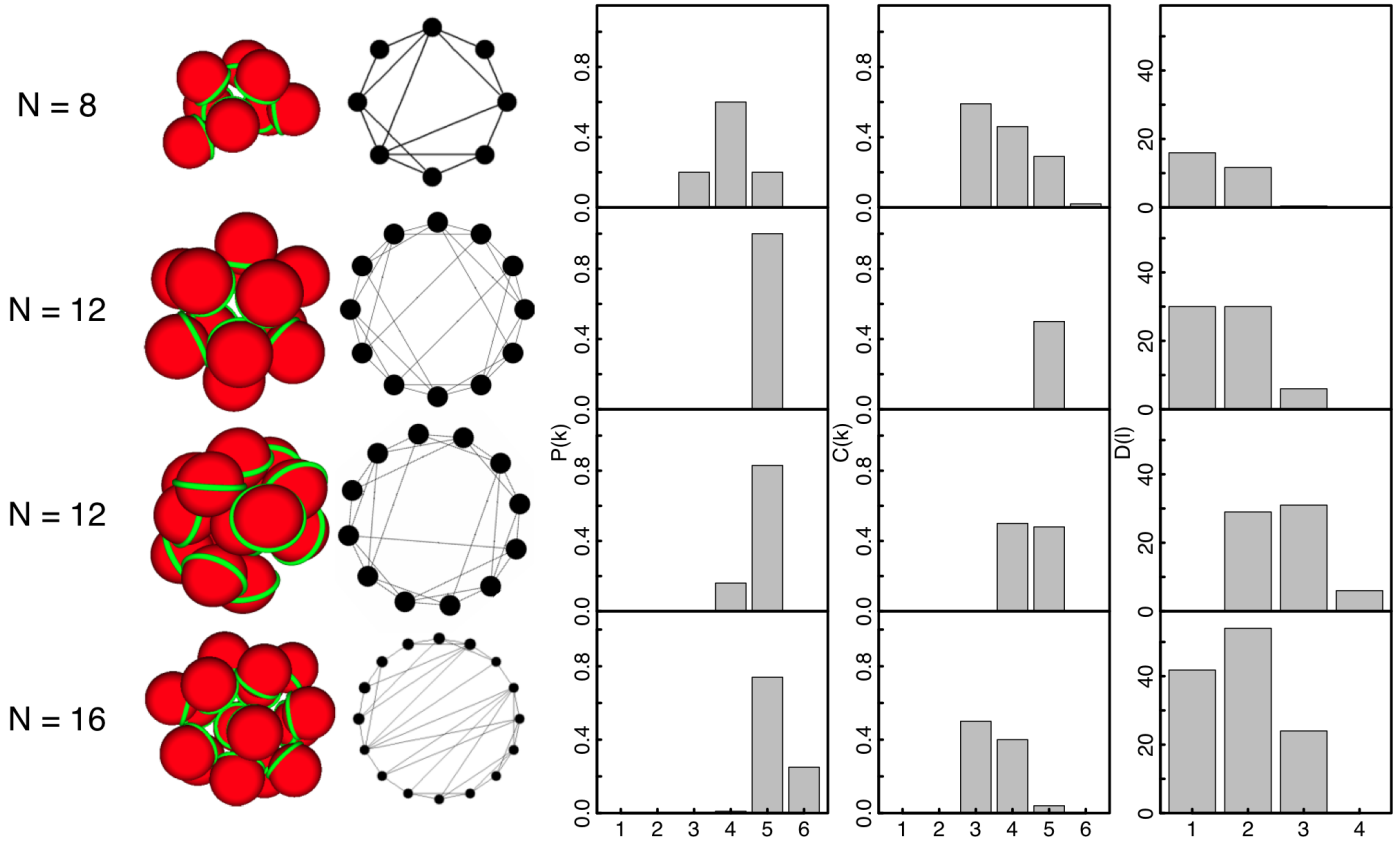
$$G_{ij}(r_{ij}, \theta_{ij}) = \exp\left[\frac{(\theta_{ij} - \nu)}{2\sigma^2_{PW}}\right]$$

$$V_{total} = V(r_{ij}) \times G_{ij}(r_{ij}, \theta_{ij}) \times G_{ji}(r_{ij}, \theta_{ji})$$

Interaction potential = distance dependent term ×orientation dep. terms

Lauck et al. , *JCTC 5, 641* (2009)

Assembly of icosahedral complexes



$$P(k) = \frac{N(k)}{N} \quad C(k) = \frac{\sum C_k}{n_k}$$

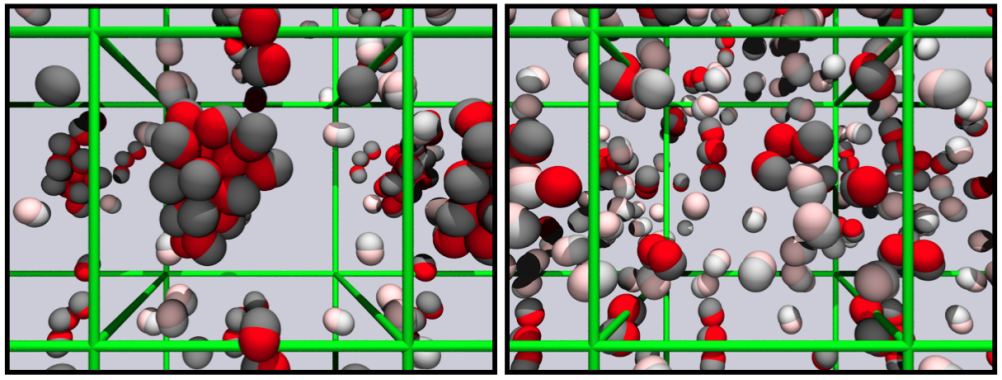
Degree distribution

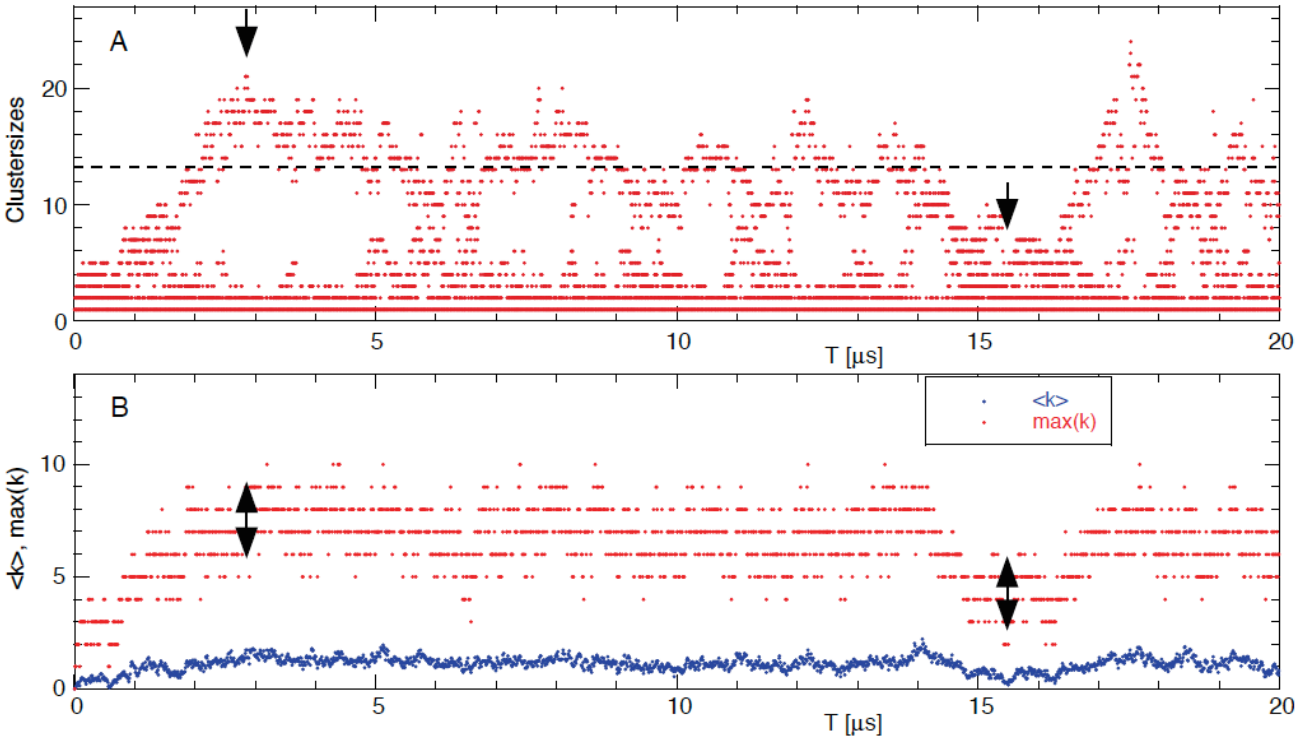
Average Cluster coefficient

shortest pathways between nodes

Lauck et al. , *JCTC 5, 641* (2009)

Dynamic view at particle agglomeration





Two snapshots

T = 2.85 µs
most of the
particles are part
of a large cluster,

T = 15.44 µs
largest cluster
has 3 particles.

Geyer,

BMC Biophysics (2011)

Summary: Static PPI-Networks

"Proteins are **modular machines**" <=> How are they related to each other?

1) **Understand** "Networks"
   prototypes (ER, SF, …) and their properties (*P(k)*, *C(k)*, clustering, …)

2) Get the **data**
   experimental and computational approaches (Y2H, TAP, co-regulation, …),
   quality control and data integration (Bayes)

3) **Analyze** the data
   compare *P(k)*, *C(k)*, clusters, … → highly modular, clustered
   obscured by sparse sampling → PPI networks are not strictly scale-free

4) **Predict** missing information
   network structure combined from multiple sources → functional annotation

Next part of lecture: gene-regulatory networks