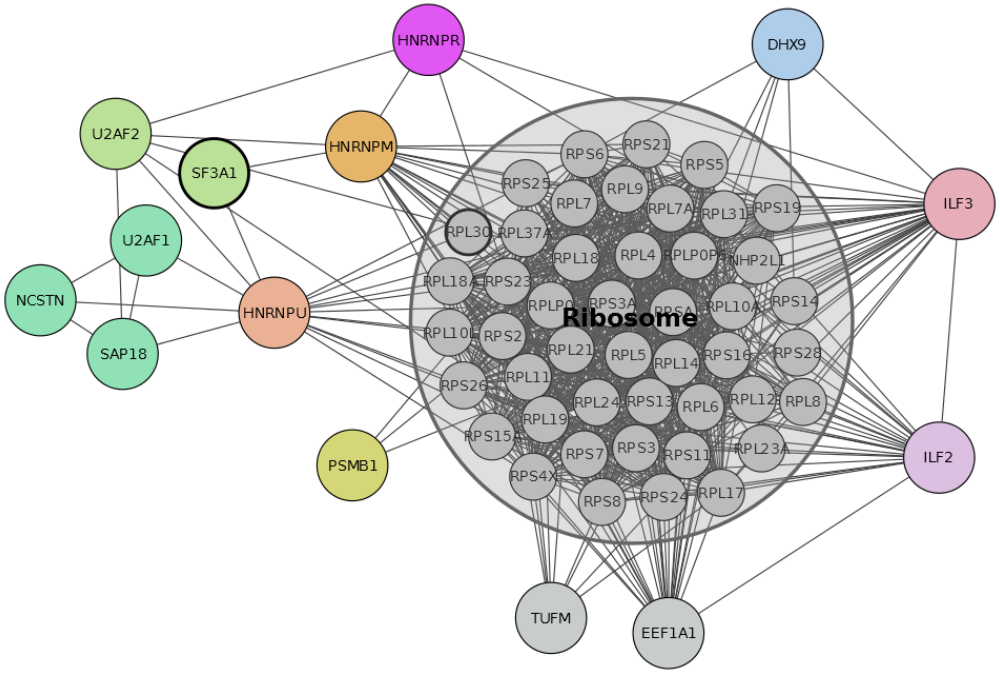
Quantitative tools for understanding and manipulating cellular signalling networks

ACC Coolen

King's College London / SaddlePoint Science



Statistical characterization and visualization

- Topology statistics beyond degree distributions
- Factor graphs
- Short loops and spectra

Quality of molecular interaction data

- Experimental bias in molecular interaction data
- Modelling the effect of experimental bias
- Experimental bias and loop statistics

Modelling cellular processes at non-local scales

- Statistical biology
- Signalling in the proteome

Hypothesis testing in signalling networks

- Random graphs as null models – the principles
- Common algorithms and their problems
- MCMC processes for hard-constrained networks

Identifying (in)vulnerabilities of signalling networks

- Path lengths and path multiplicities
- Disrupting or protecting signalling processes

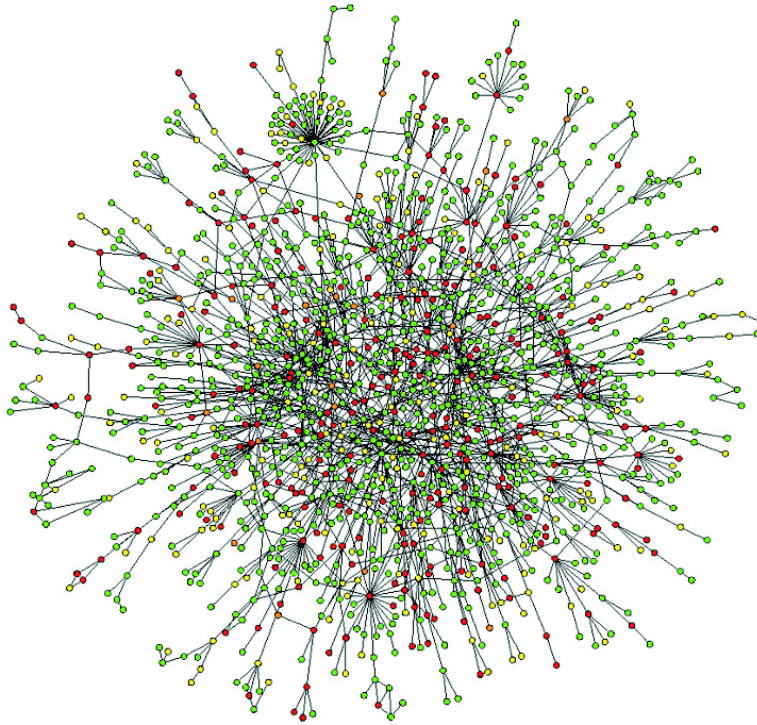
Cellular signalling networks

- *protein interactions:*

  nodes: proteins $i, j = 1 \ldots N$

  links: $A_{ij} = A_{ji} = 1$ if i can bind to j
  $A_{ij} = A_{ji} = 0$ otherwise

  nondirected,
  $N \sim 10^4$, links/node ~ 7

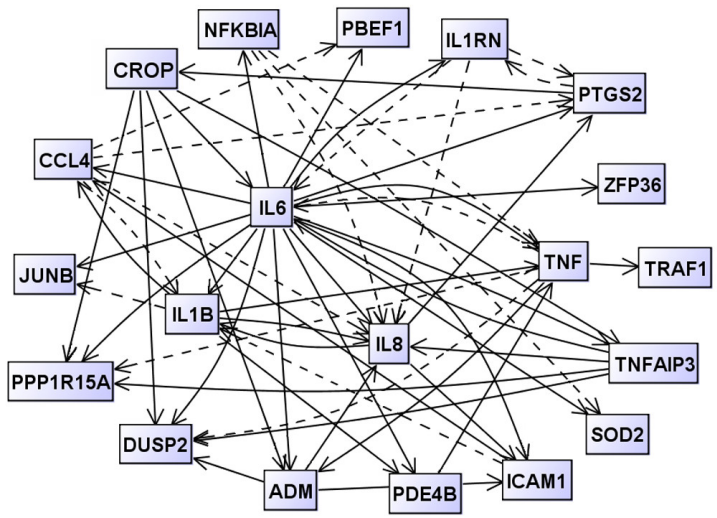


- *gene regulation*:

  nodes: genes $i, j = 1 \ldots N$

  links: $A_{ij} = 1$ if j codes for transcription factor of i
  $A_{ij} = 0$ otherwise

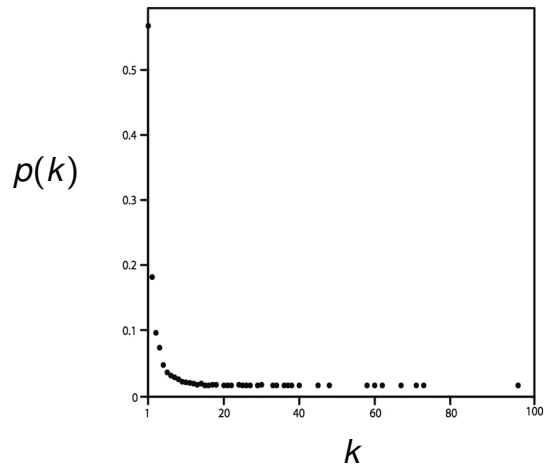
  directed,
  $N \sim 10^4$, links/node ~ 5



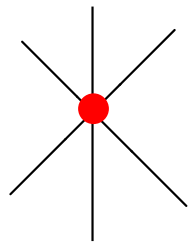
Topology statistics beyond degree distributions

- degrees: $k_i = \sum_j A_{ij}$

  distribution: $p(k) = \frac{1}{N} \sum_i \delta_{k,k_i(\mathbf{A})}$



pick node i at random:

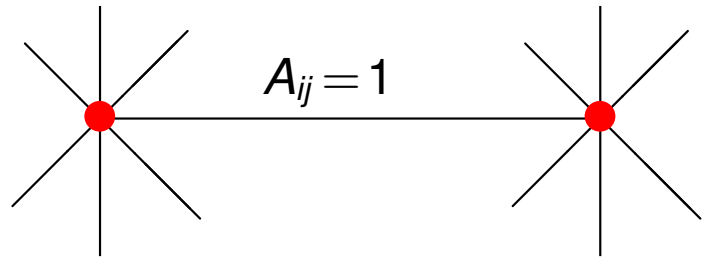


$k_i = k$?

- joint degree statistics of connected nodes

$$W(k,k') = \frac{1}{N\langle k\rangle} \sum_{ij} A_{ij} \delta_{k,k_i(\mathbf{A})} \delta_{k',k_j(\mathbf{A})}$$

pick link (i,j) at random



$k_i = k$? $k_j = k'$?

- relation between p and W:

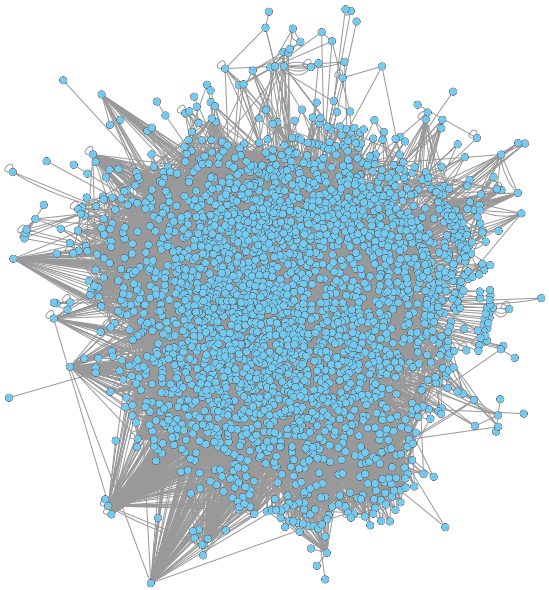
$$W(k) = \sum_{k'} W(k,k') = p(k)k/\langle k \rangle$$

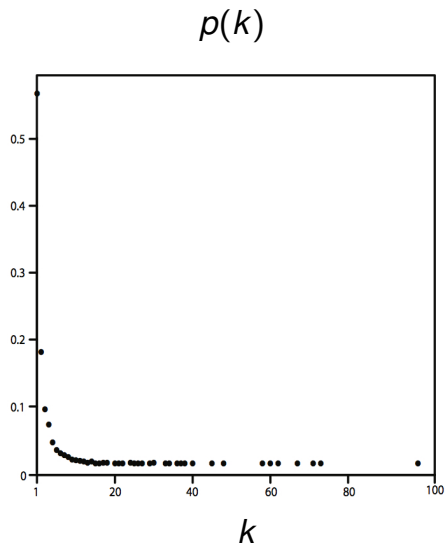
maginals of W carry no info beyond degree stats so focus on:

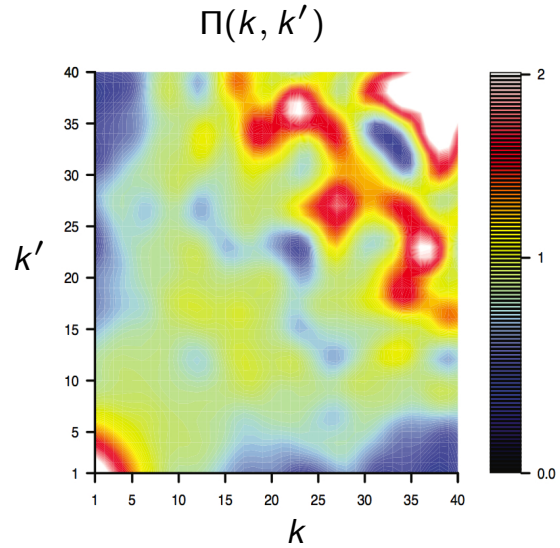
$$\Pi(k,k') = \frac{W(k,k')}{W(k)W(k')}$$

any (k,k') with $\Pi(k,k') \neq 1$:
structural information in degree correlations

Human PPIN:

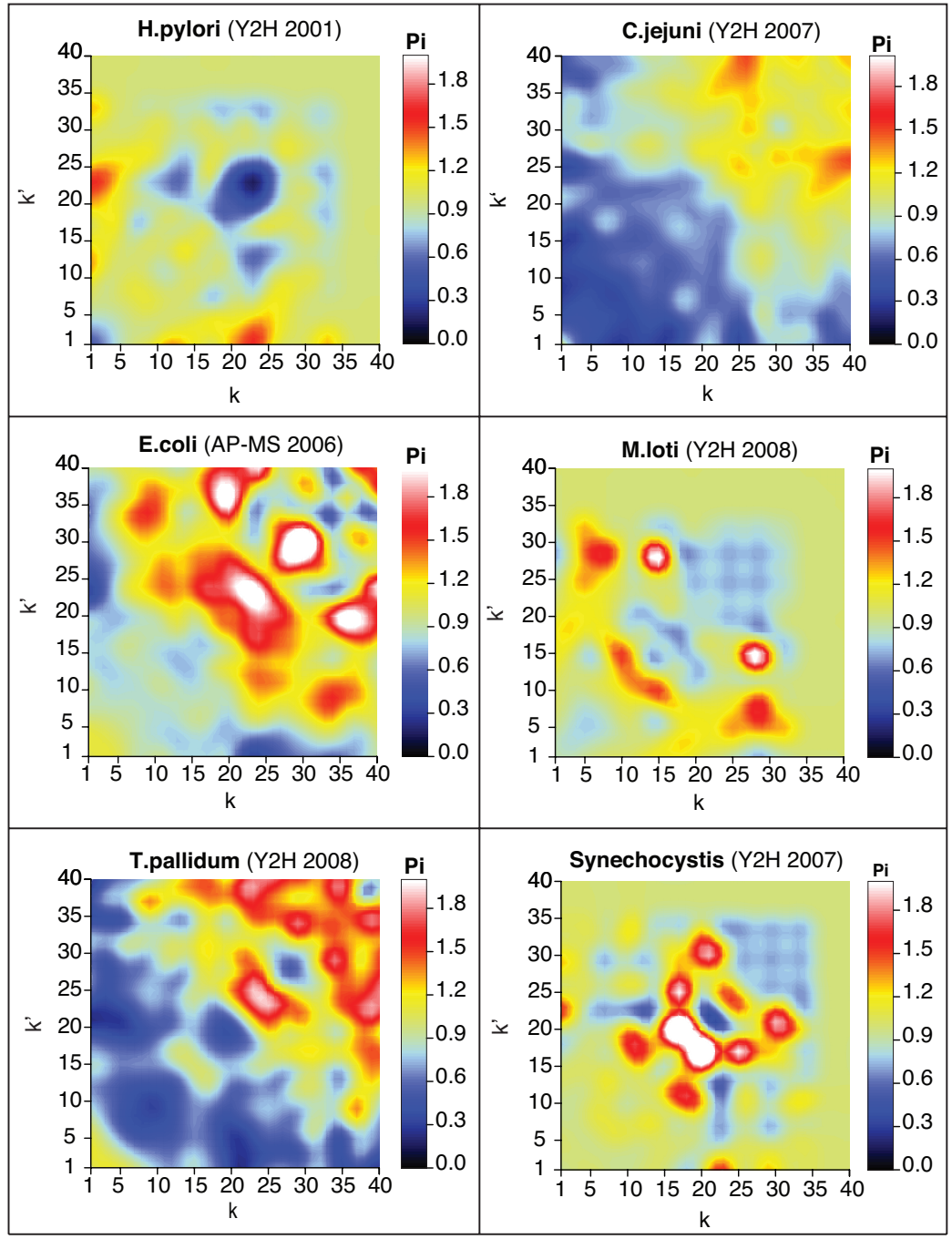






$$\Pi(k,k') = \frac{W(k,k')}{W(k)W(k')}$$

for protein
interaction networks



Directed graphs (e.g. gene regulation)

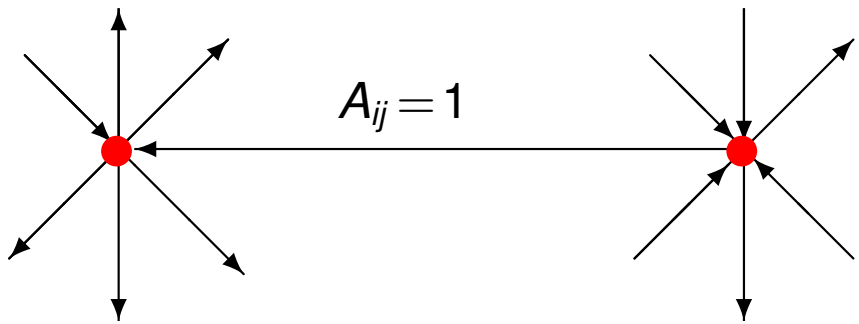
links become *arrows*

- degrees:

$$k_i^{\text{in}} = \sum_j A_{ij}, \qquad k_i^{\text{out}} = \sum_j A_{ji}, \qquad p(k_{\text{in}}, k_{\text{out}}) = \frac{1}{N} \sum_i \delta_{k_{\text{in}}, k_i^{\text{in}}} \delta_{k_{\text{out}}, k_i^{\text{out}}}$$

- joint in-out degree statistics of connected nodes

$W(k_{\text{in}}, k_{\text{out}}; k'_{\text{in}}, k'_{\text{out}})$



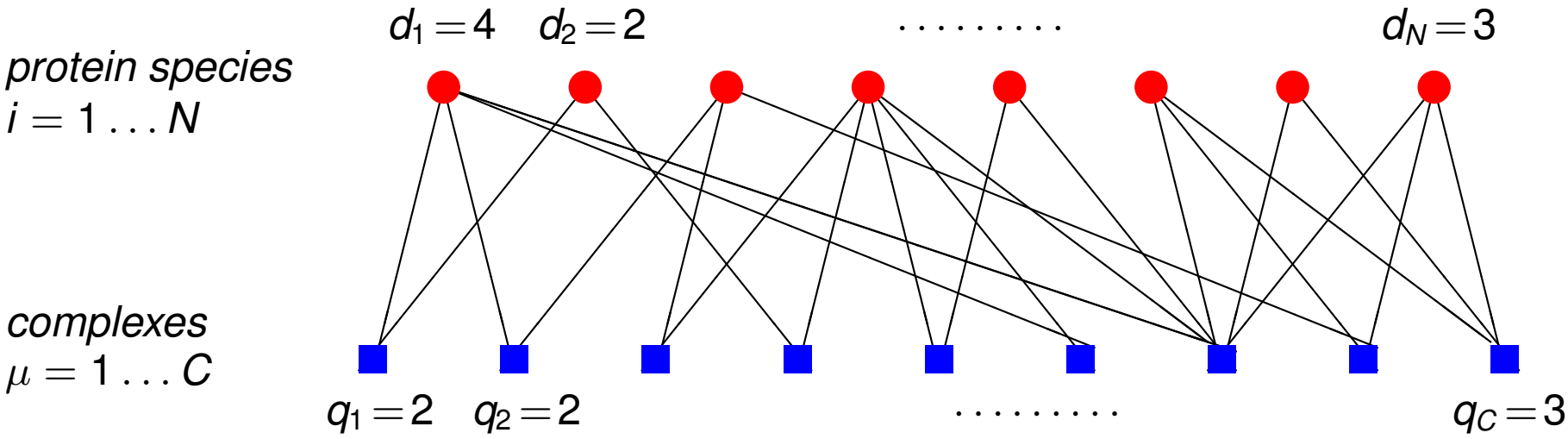
$(k_{\text{in}}, k_{\text{out}})_i = (k_{\text{in}}, k_{\text{out}})$? $(k_{\text{in}}, k_{\text{out}})_j = (k'_{\text{in}}, k'_{\text{out}})$?

note:

$W(k_{\text{in}}, k_{\text{out}}; k'_{\text{in}}, k'_{\text{out}}) \neq W(k'_{\text{in}}, k'_{\text{out}}; k_{\text{in}}, k_{\text{out}})$

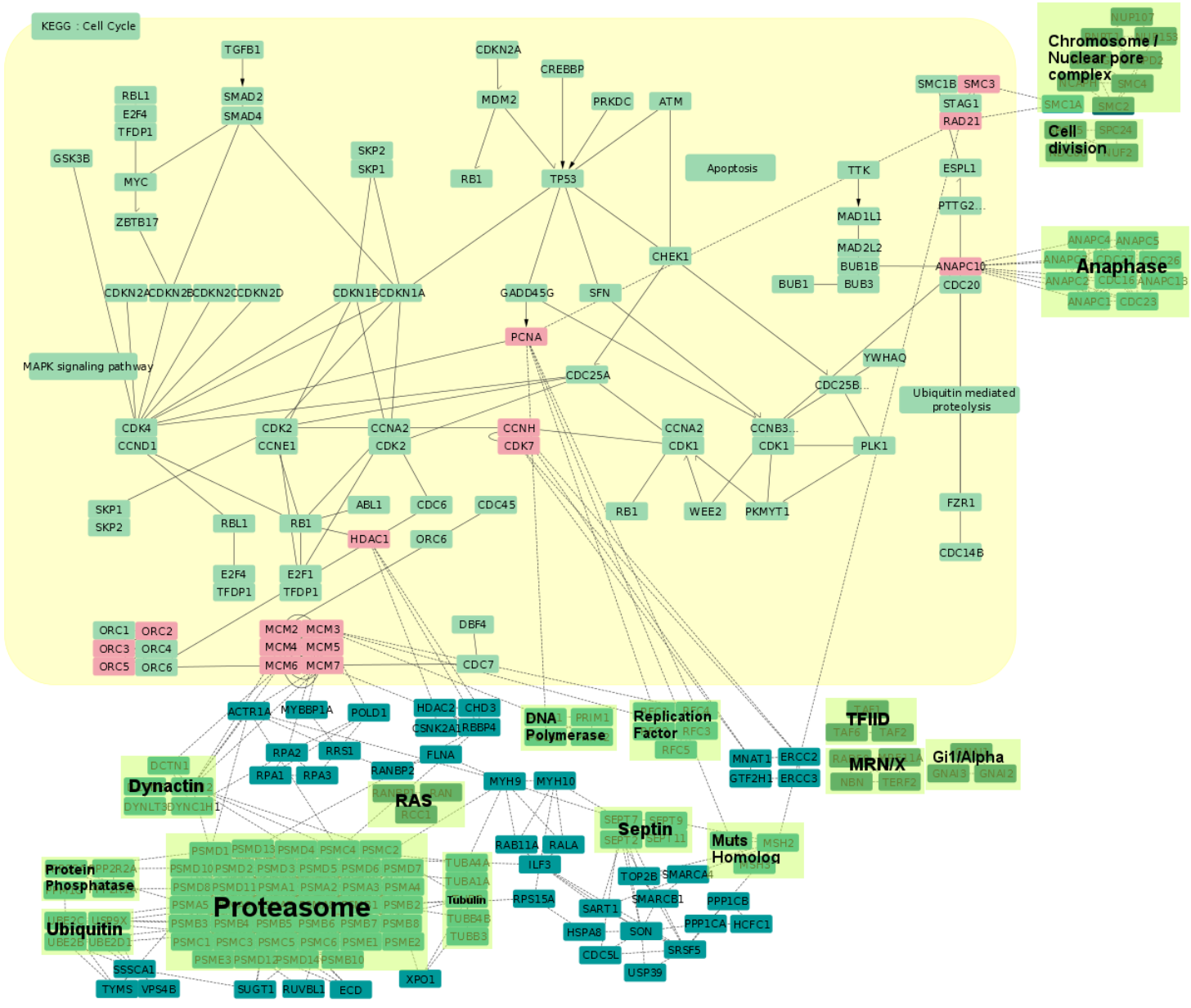
represent PPINs more effectively

Factor graphs

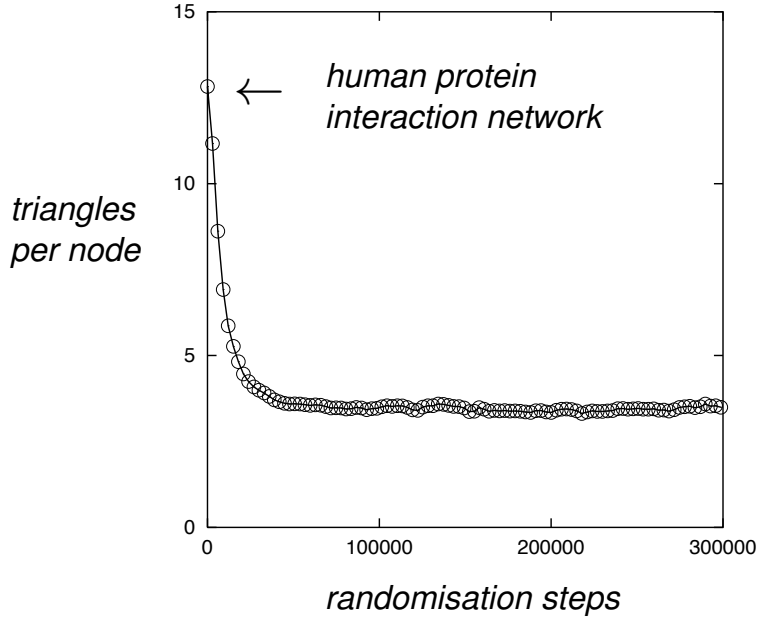


- more informative than standard PPIN graph (e.g. ‘party hubs’ vs ‘date hubs’)
- links directly to reaction equations
- relations between stats $p(q)$ of complex sizes and stats $p(d)$ of protein promiscuities
- formulae linking loop statistics in standard PPIN to stats of complex sizes and protein promiscuity

Beyond degree distributions and degree correlations



protein interaction networks
have many short loops ...



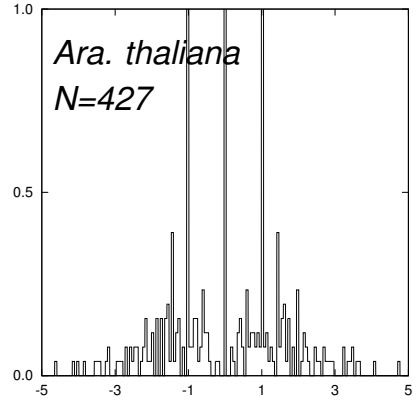
nr of closed paths of length ℓ:

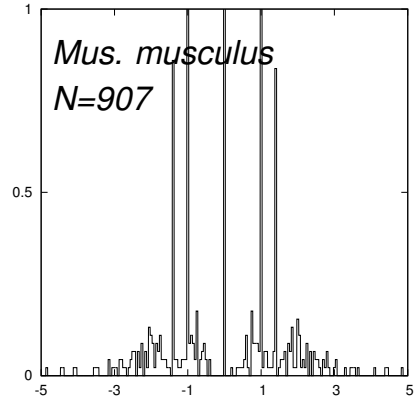
$$
\begin{aligned}
n_\ell &= \frac{1}{2\ell} \sum_{i_1, i_2, \ldots, i_{\ell-1}} A_{i_1 i_2} A_{i_2 i_3} \ldots A_{i_\ell i_1} \\
&= \frac{1}{2\ell} \sum_i (\mathbf{A}^\ell)_{ii} \\
&= \frac{N}{2\ell} \int \mathrm{d}\mu \; \varrho(\mu) \mu^\ell \qquad \varrho(\mu): \quad \textit{spectrum of eigenvalues of } \mathbf{A}
\end{aligned}
$$

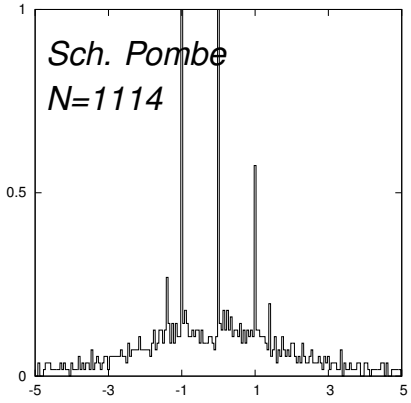
- information on closed paths of *all* lengths
  is encoded in spectrum $\varrho(\mu)$
- statistical analysis of 'loopy' graphs *very tricky*
- requires new mathematical tools (now being developed)

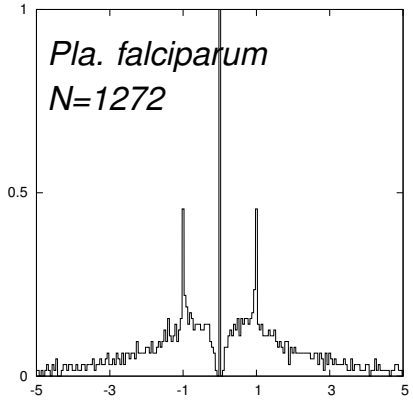
Spectra $\varrho(\mu)$ of protein interaction networks

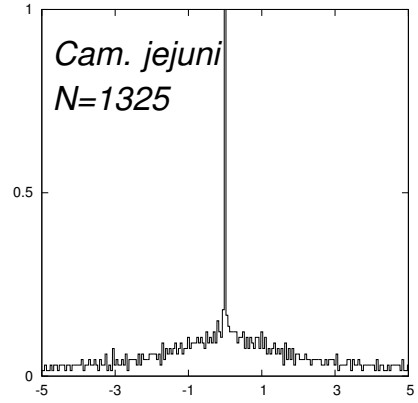
access to statistics of short loops ...

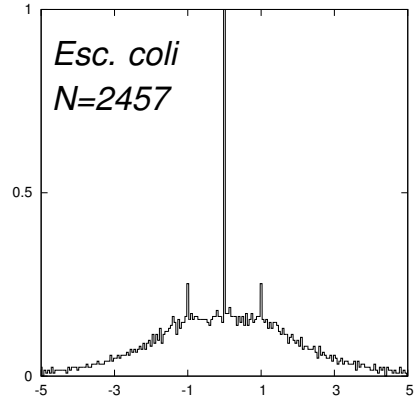


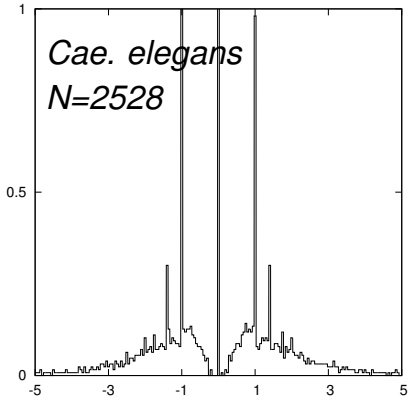


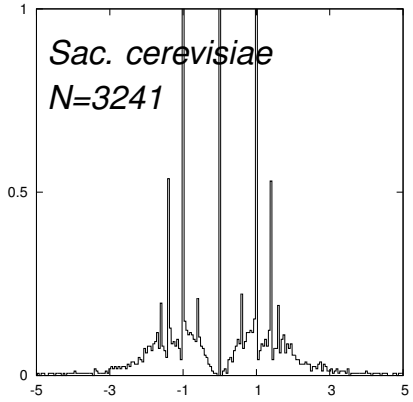


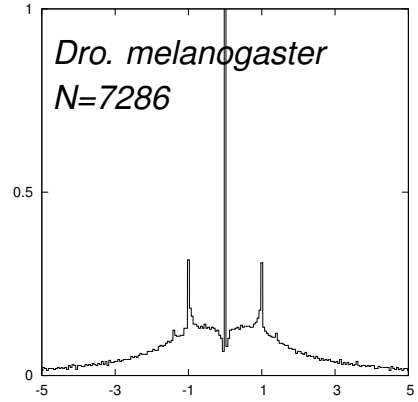


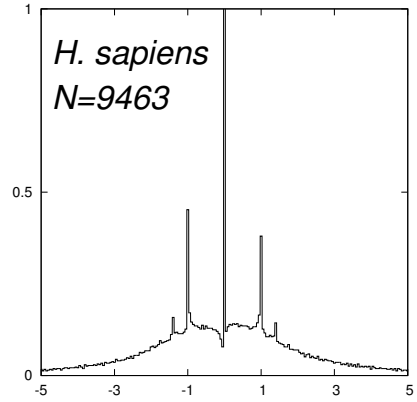




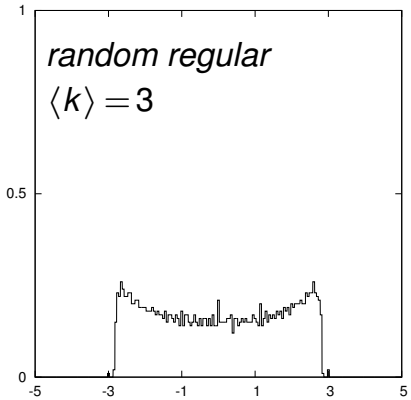


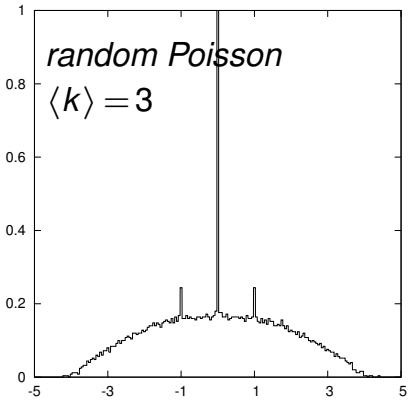


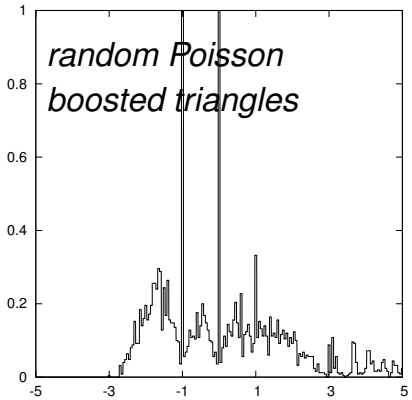


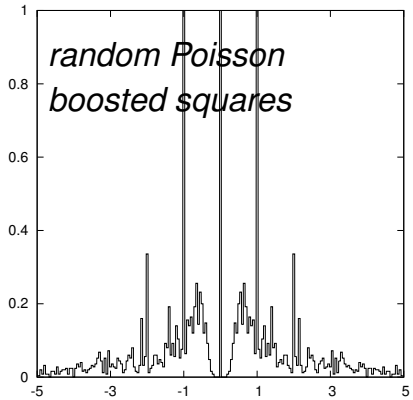


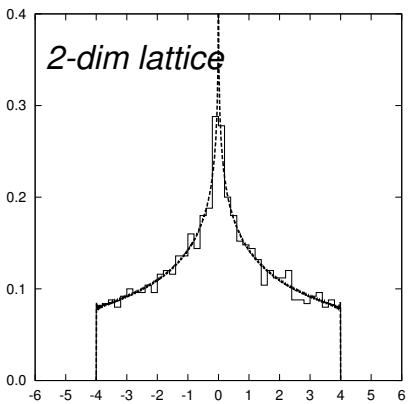
generally quite different from
spectra of non-biological networks!

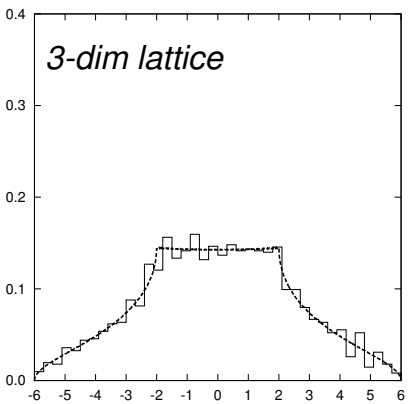


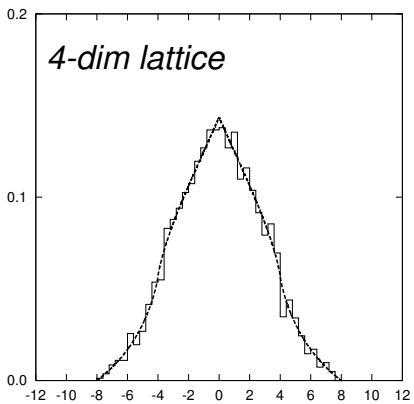


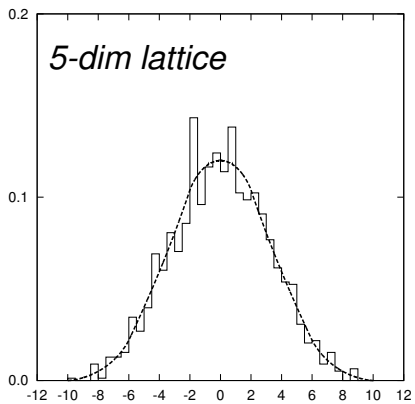












nature biotechnology

Commentary

Nature Biotechnology **26**, 69 - 72 (2008)
doi:10.1038/nbt0108-69

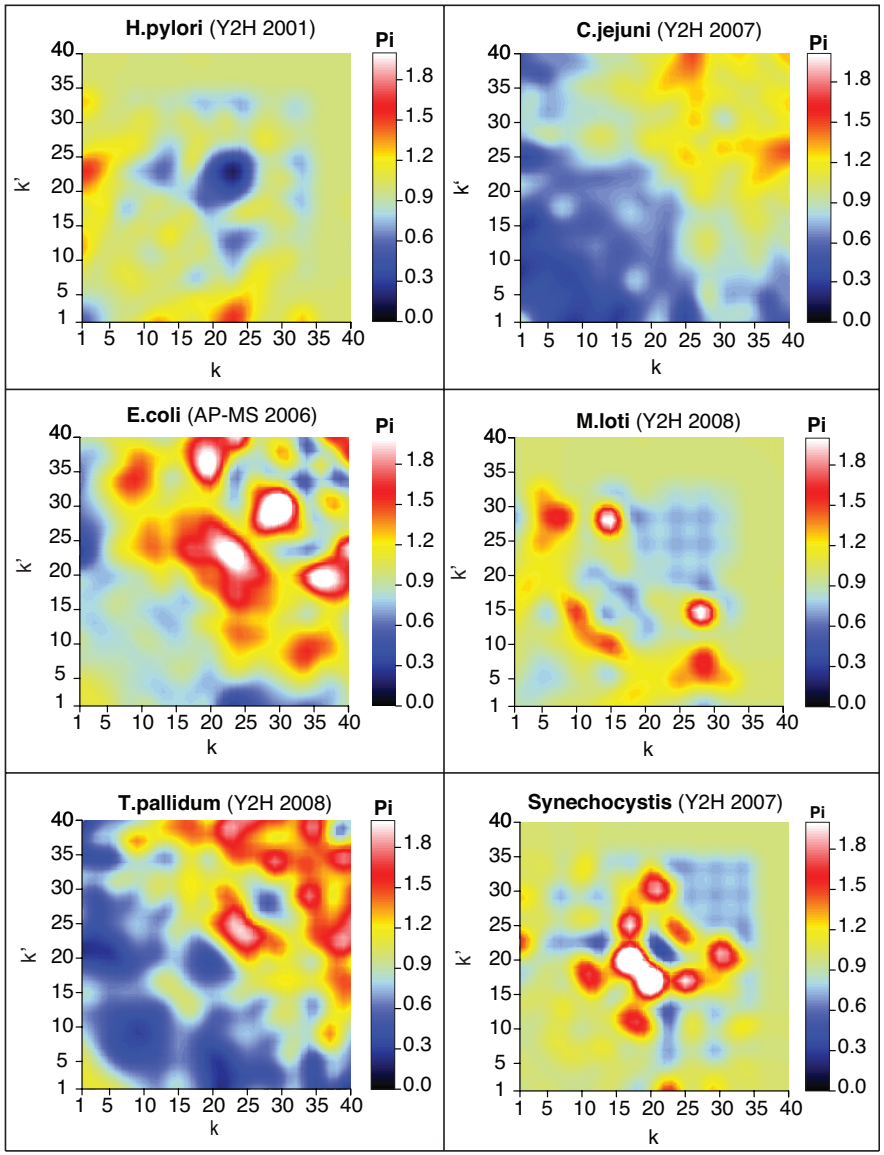
Protein-protein interaction networks and biology—what's the connection?

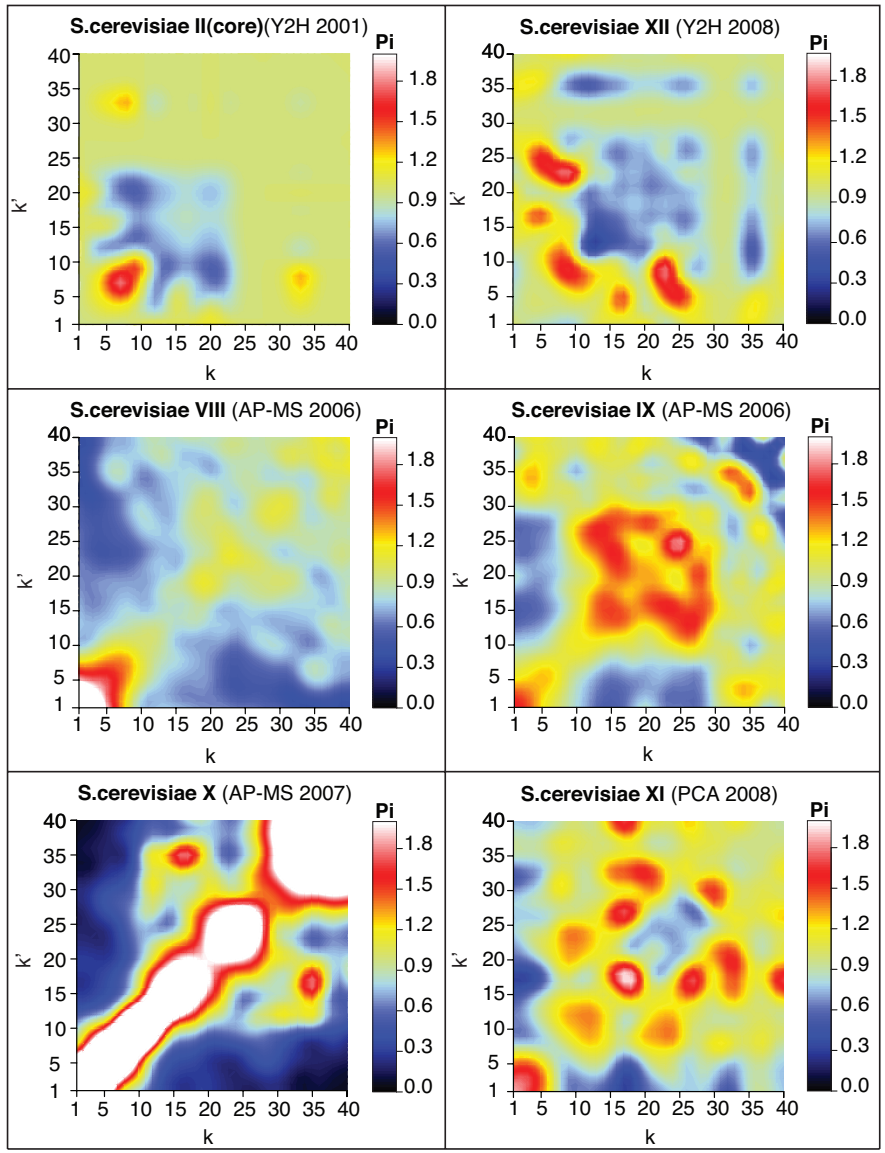
Luke Hakes[1], John W Pinney[1], David L Robertson[1] & Simon C Lovell[1]

Analysis of protein-protein interaction networks is an increasingly popular means to infer biological insight, but is close enough attention being paid to data handling protocols and the degree of bias in the data?

The availability of large-scale protein-protein interaction data has led to the recent popularity of the study of protein interaction networks. Just as the enormous amount of available sequence data has made it

$\Pi(k, k')$ for PPIN: an accurate problem sensor ...





how to measure dissimilarity between networks?

information theory: regard network as noisy realization
of a graph with characteristics $\{p, \Pi\}$

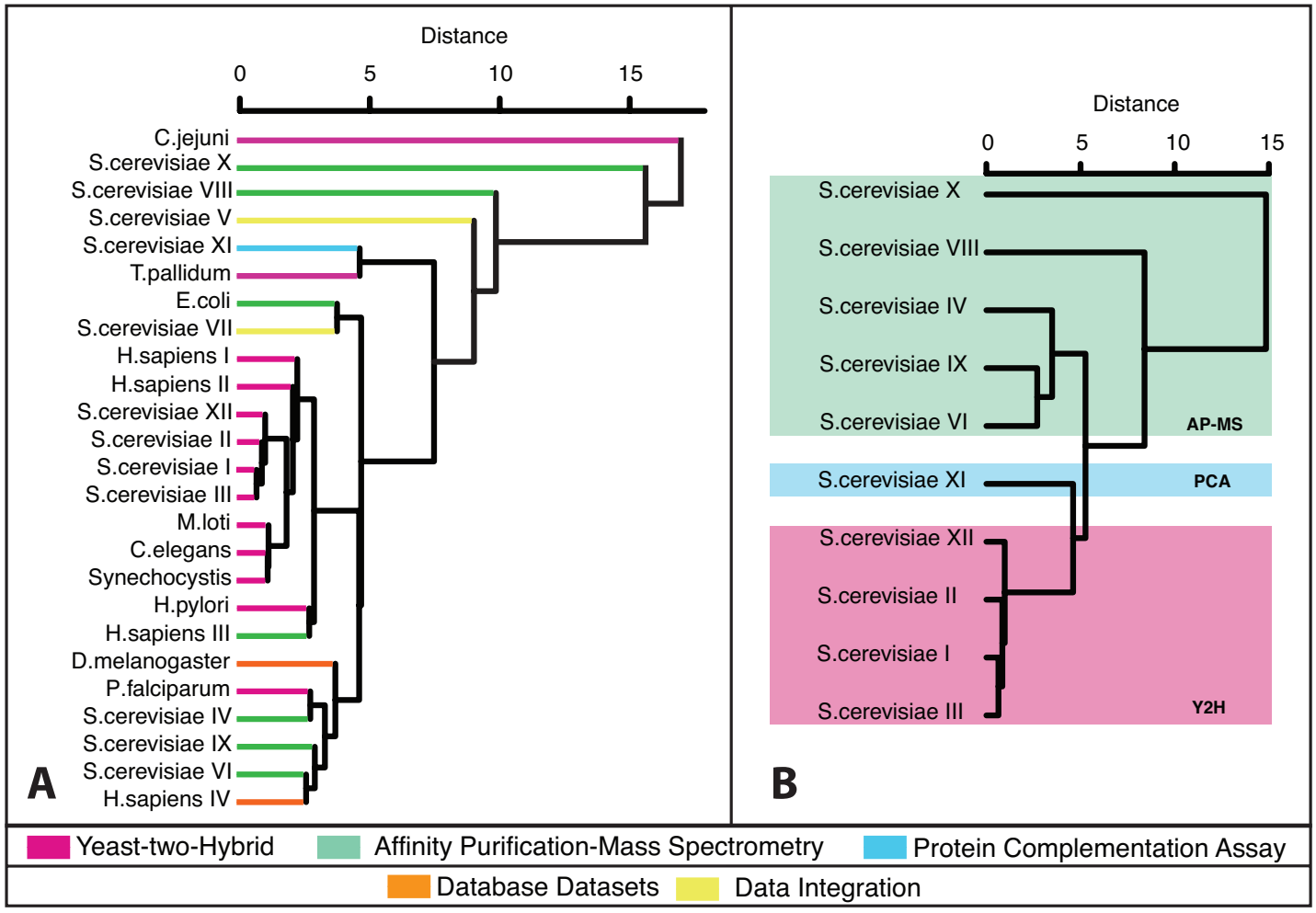
$$
\begin{aligned}
D_{AB} \quad = \quad & \frac{1}{2N} \sum_{\mathbf{A}} p(\mathbf{A}|p_A, \Pi_A) \log \left[\frac{p(\mathbf{A}|p_A, \Pi_A)}{p(\mathbf{A}|p_B, \Pi_B)} \right] \\
& + \frac{1}{2N} \sum_{\mathbf{A}} p(\mathbf{A}|p_B, \Pi_B) \log \left[\frac{p(\mathbf{A}|p_B, \Pi_B)}{p(\mathbf{A}|p_A, \Pi_A)} \right]
\end{aligned}
$$

max entropy distr $p(\mathbf{A}|p, \Pi)$,
result of calculation, large N:

$$
\begin{aligned}
D_{AB} = & \frac{1}{2} \sum_k p_A(k) \log \left[\frac{p_A(k)}{p_B(k)} \right] + \sum_{kk'} \frac{p_A(k) p_A(k') k k'}{4 \langle k \rangle_A} \Pi_A(k, k') \log \left[\frac{\Pi_A(k, k')}{\Pi_B(k, k')} \right] \\
& + \frac{1}{2} \sum_k p_B(k) \log \left[\frac{p_B(k)}{p_A(k)} \right] + \sum_{kk'} \frac{p_B(k) p_B(k') k k'}{4 \langle k \rangle_B} \Pi_B(k, k') \log \left[\frac{\Pi_B(k, k')}{\Pi_A(k, k')} \right] \\
& + \frac{1}{2} \sum_k k \Big[p_A(k) \log \varrho_{BA}(k) + p_B(k) \log \varrho_{AB}(k) \Big]
\end{aligned}
$$

ϱ_{AB} : *solution of* $\varrho(k) = \sum_{k'} \Pi_A(k, k') k' p_B(k') / \langle k \rangle_B \varrho(k')$

clustering of PPIN data with information-theoretic distance measure



- PPINs of same species, measured via *same experimental method*: similar (in spite of limited overlap)
- PPINs measured via same method cluster together: strong experimental bias, explains many reproducibility problems ...

Contamination of molecular interaction data

- node undersampling:

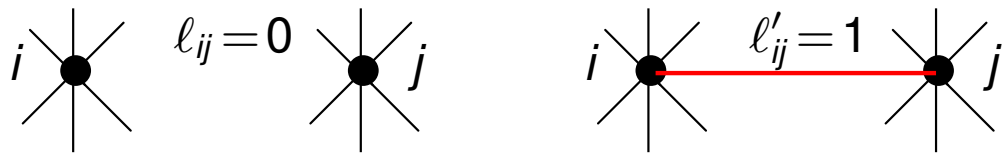
  $x(k_i)$: prob to detect protein i

- link undersampling:

  $y(k_i, k_j)$: prob to detect interaction (i, j)

- link oversampling:

  $z(k_i, k_j)/N$: prob to report false positive interaction



calculate relation between:

measured $p(k)$ *and* $W(k, k')$
and true $p(k)$ *and* $W(k, k')$

in terms of $x(k), y(k), z(k, k')$

core result

can be done *analytically*, for large N,
for *all* sampling protocols,
(via path integrals, steepest descent, ...):

$$p(k|x,y,z) = \frac{\sum_q x(q)p(q)\Big\{a(q)\mathcal{J}(k|q) + qb(q)\mathcal{L}(k|q)\Big\}}{k\sum_q p(q)x(q)}$$

$$W(k,k'|x,y,z) = \frac{\sum_{q,q'>0} x(q)x(q')\Big\{p(q)p(q')z(q,q')\mathcal{J}(k|q)\mathcal{J}(k'|q') + \langle k\rangle W(q,q')y(q,q')\mathcal{L}(k|q)\mathcal{L}(k'|q')\Big\}}{\bar{k}(x,y,z)\sum_q p(q)x(q)}$$

with

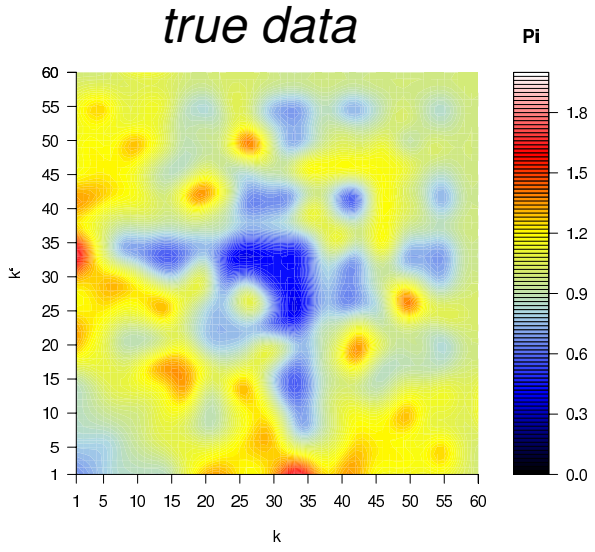
$$\mathcal{J}(k|q) = \mathrm{e}^{-a(q)}\sum_{n=0}^{\min\{k-1,q\}}\left(\begin{array}{c} q \\ n \end{array}\right)\frac{a^{k-1-n}(q)}{(k-1-n)!}b^n(q)(1-b(q))^{q-n}$$

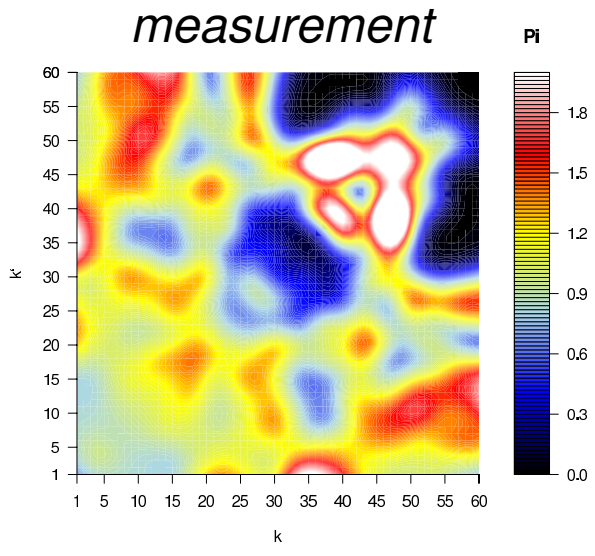
$$\mathcal{L}(k|q) = \mathrm{e}^{-a(q)}\sum_{n=0}^{\min\{k-1,q-1\}}\left(\begin{array}{c} q-1 \\ n \end{array}\right)\frac{a^{k-1-n}(q)}{(k-1-n)!}b^n(q)(1-b(q))^{q-1-n}$$

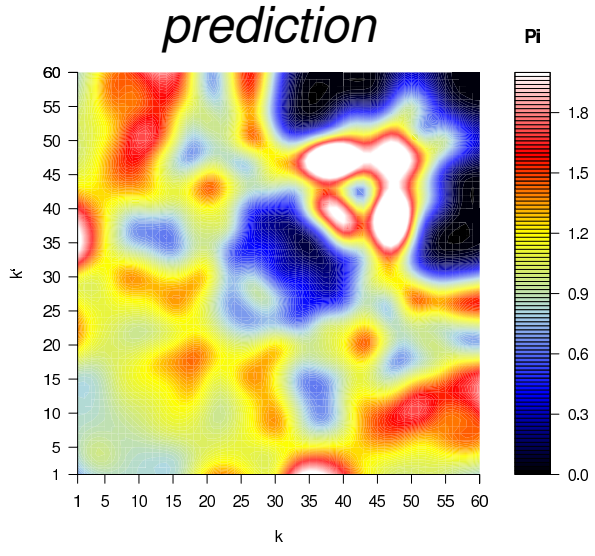
$$a(q) = \sum_{q'\geq 0} p(q')x(q')z(q,q'), \qquad b(q) = \frac{\langle k\rangle}{qp(q)}\sum_{q'\geq 0} x(q')y(q,q')W(q,q')$$

$$\bar{k}(x,y,z) = \frac{\sum_q x(q)p(q)[a(q)+q\,b(q)]}{\sum_q p(q)x(q)}$$

heat maps of
$W(k, k')/W(k)W(k')$:







- *predict* what will be measured, given contamination parameters
- hence we can infer contamination parameters and true data ...

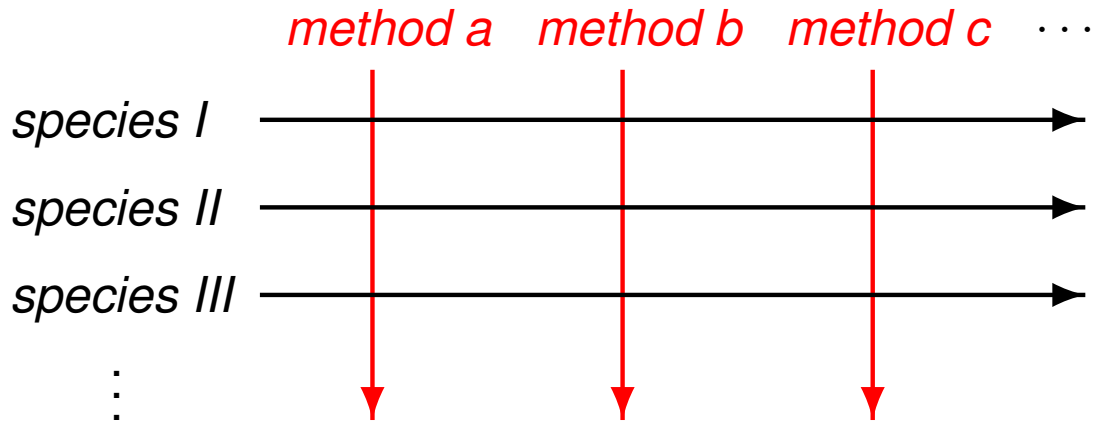
Ongoing work: decontamination of PPIN data

available PPIN data:

– for L different species $\ell = 1 \ldots L$
  each with unknown network $\mathbf{A}^{\ell}$

– measured via M different protocols $\alpha = 1 \ldots M$ (e.g. Y2H, PCA, MS)
  each with unknown sampling parameters $\theta^{\alpha} = \{x^{\alpha}, y^{\alpha}, z^{\alpha}\}$

matrix of $M \times L$
observed networks $\mathbf{A}^{\ell,\alpha}$: $A_{ij}^{\ell,\alpha} = \sigma_i^{\ell,\alpha}\sigma_j^{\ell,\alpha}[\tau_{ij}^{\ell,\alpha}A_{ij}^{\ell} + (1 - A_{ij}^{\ell})\lambda_{ij}^{\ell,\alpha}]$



objective:

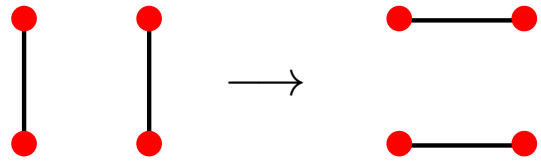
find true PINs $\{\mathbf{A}^1, \ldots, \mathbf{A}^L\}$ and
sampling pars $\{\theta^1, \ldots, \theta^M\}$ (via Bayesian methods)

Experimental bias and loop statistics

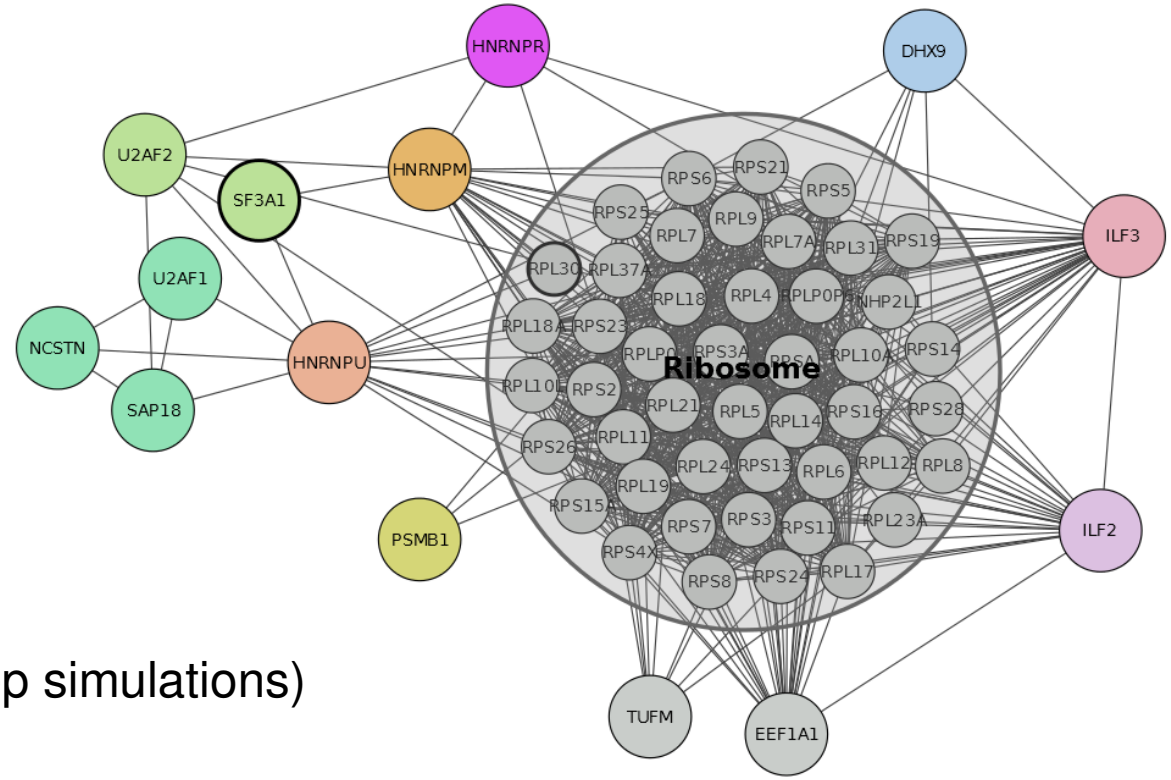
- *resilient short loops*

  those preserved after
  edge-swap randomization

  type 1: leave all degrees invariant
  type 2: leave also $W(k, k')$ invariant



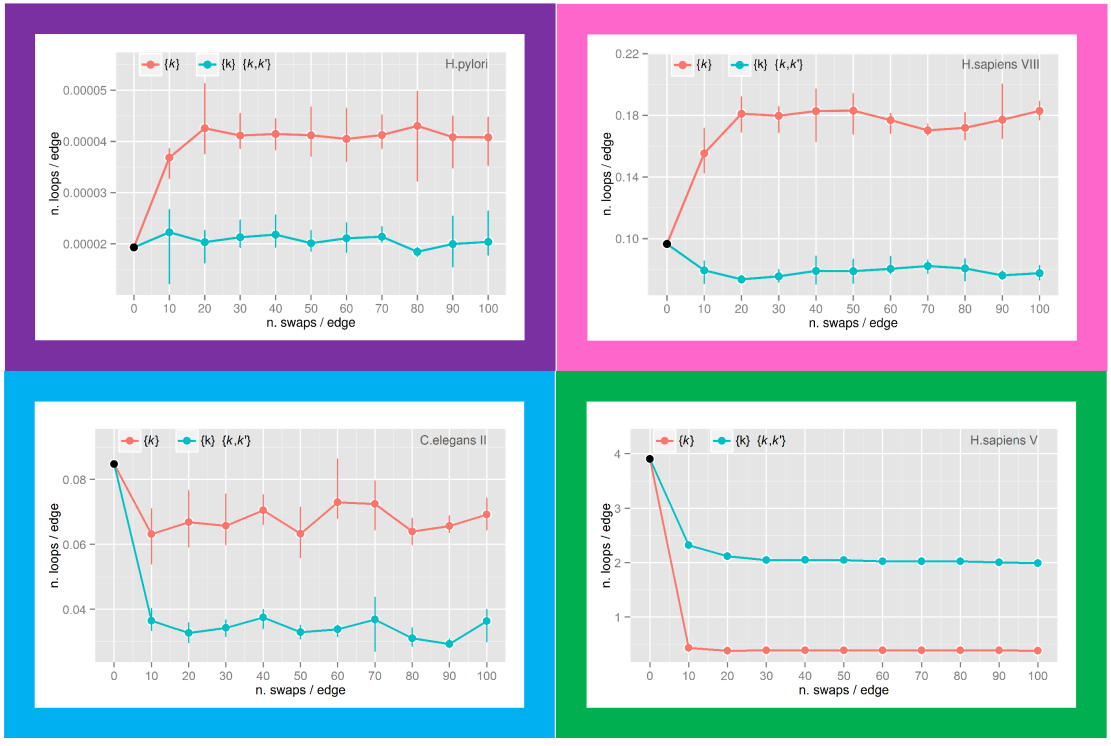
- resilient loops of
  lengths 3 and 4
  (type 2) in
  human PPIN:



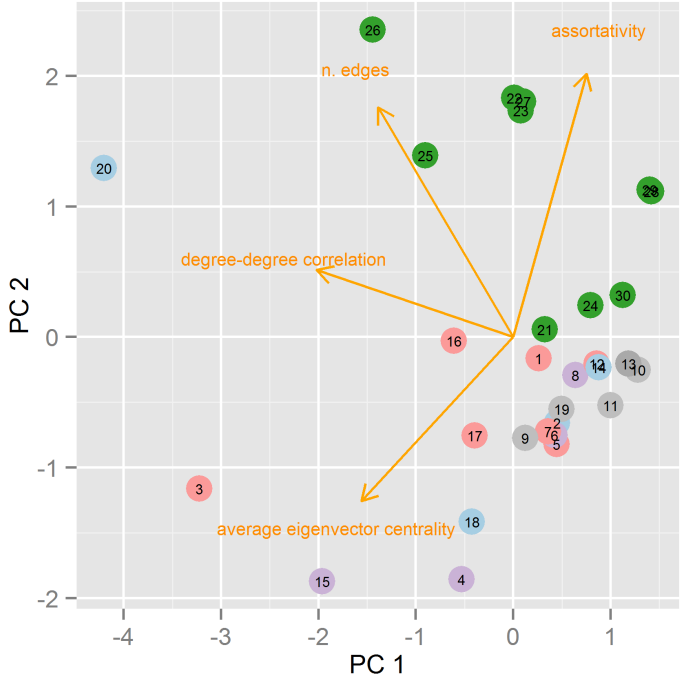
(preserved in 5 indep simulations)

- distinct effects of randomization on nrs of loops
- strong correlation with experimental protocol
- green: higher quality PPIN datasets

a)



b)



Analysis of molecular signalling processes

proteome:

usual description
reaction equations

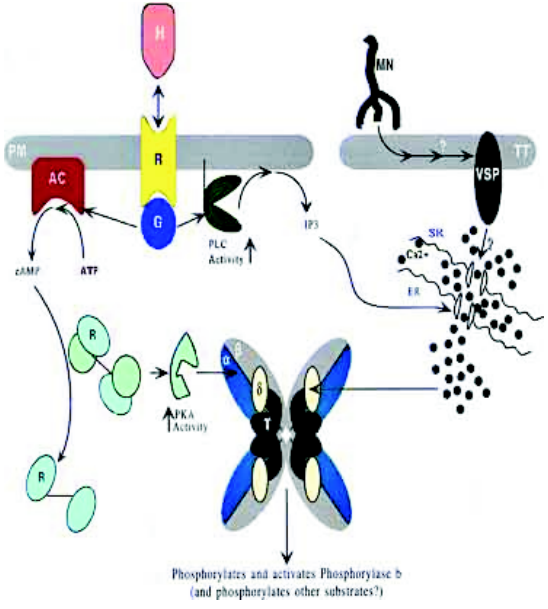


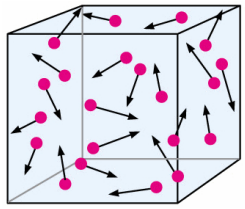
Table 2. Model Equations

$$
\begin{aligned}
d(RD)/dt &= k_{81}RDA - k_{18}RD \cdot A + k_{31}RDE - k_{13}RD \cdot E - k_{19}RD + k_{91}R \cdot D + k_{21}RT - k_{12}RD \cdot M \\
d(RT)/dt &= k_{52}RTE - k_{25}RT \cdot E + k_{92}R \cdot T - k_{29}RT - k_{21}RT + k_{62}RTA - k_{26}RT \cdot A - k_{2M}RT \cdot E + k_{M2}M + k_{12}R \\
d(RDE)/dt &= k_{13}RD \cdot E - k_{31}RDE + k_{43}RE \cdot D - k_{34}RDE + k_{53}RTE \\
d(RE)/dt &= k_{34}RDE - k_{43}RE \cdot D + k_{54}RTE - k_{45}RE \cdot T + k_{94}R \cdot E - k_{49}RE \\
d(RTE)/dt &= k_{45}RE \cdot T - k_{54}RTE + k_{25}RT \cdot E - k_{52}RTE - k_{53}RTE \\
d(RTA)/dt &= k_{26}RT \cdot A - k_{62}RTA - k_{68}RTA + k_{76}RA \cdot T - k_{67}RTA \\
d(RA)/dt &= k_{67}RTA - k_{76}RA \cdot T + k_{97}R \cdot A - k_{79}RA + k_{87}RDA - k_{78}RA \cdot D \\
d(RDA)/dt &= k_{68}RTA + k_{78}RA \cdot D - k_{87}RDA + k_{18}RD \cdot A - k_{81}RDA \\
d(R)/dt &= k_{29}RT - k_{92}R \cdot T + k_{49}RE - k_{94}R \cdot E + k_{19}RD - k_{91}R \cdot D + k_{79}RA - k_{97}R \cdot A \\
d(E)/dt &= k_{31}RDE - k_{13}RD \cdot E + k_{52}RTE - k_{25}RT \cdot E + k_{49}RE - k_{94}R \cdot E - k_{2M}RT \cdot E + k_{M2}M \\
d(A)/dt &= k_{81}RDA - k_{18}RD \cdot A + k_{62}RTA - k_{26}RT \cdot A + k_{79}RA - k_{97}R \cdot A \\
d(M)/dt &= k_{2M}RT \cdot E - k_{M2}M
\end{aligned}
$$

Model equations correspond to the reaction scheme shown in Figure 1. Numbering of the reaction rate cor
follows the conventions introduced in Table 3.

- cannot solve eqns analytically ...
- uncertain pathways and parameters ...
- too many components for numerical exploration ...

statistical physics



$\sim 10^{24}$ positions, velocities

$(\vec{x}_1, \vec{v}_1),\ (\vec{x}_2, \vec{v}_2),\ \ldots$

Newton's equations

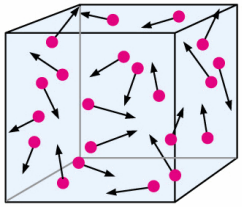
$\frac{d}{dt}(\vec{x}_1, \vec{v}_1) = \ldots,\ \frac{d}{dt}(\vec{x}_2, \vec{v}_2) = \ldots$ $\leftarrow$ don't try to solve these!

macroscopic description:

densities, correlation functions,
perturbation response functions,
phase transitions ...

'self-averaging': macroscopic theory only
dependent on *statistics* of model parameters ...

statistical physics



$\sim 10^{24}$ positions, velocities

$(\vec{x}_1, \vec{v}_1),\ (\vec{x}_2, \vec{v}_2),\ \ldots$

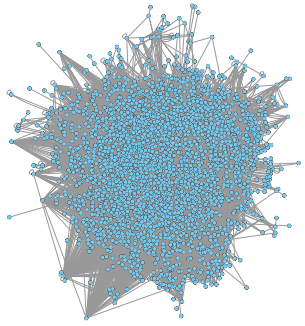
Newton's equations

$\frac{d}{dt}(\vec{x}_1, \vec{v}_1) = \ldots,\ \frac{d}{dt}(\vec{x}_2, \vec{v}_2) = \ldots$

macroscopic theory:

densities, correlation functions,
response functions (to perturbations),
phase transitions ...

statistical biology



$\sim 10^4$ concentr of proteins & complexes

$\vec{x}_1,\ \vec{x}_2,\ \vec{x}_3,\ \ldots$

reaction equations

$\frac{d}{dt}\vec{x}_1 = \ldots,\ \frac{d}{dt}\vec{x}_2 = \ldots, \frac{d}{dt}\vec{x}_3 = \ldots$

macroscopic theory:

???

'self-averaging': macroscopic theory only
dependent on *statistics* of model parameters ...

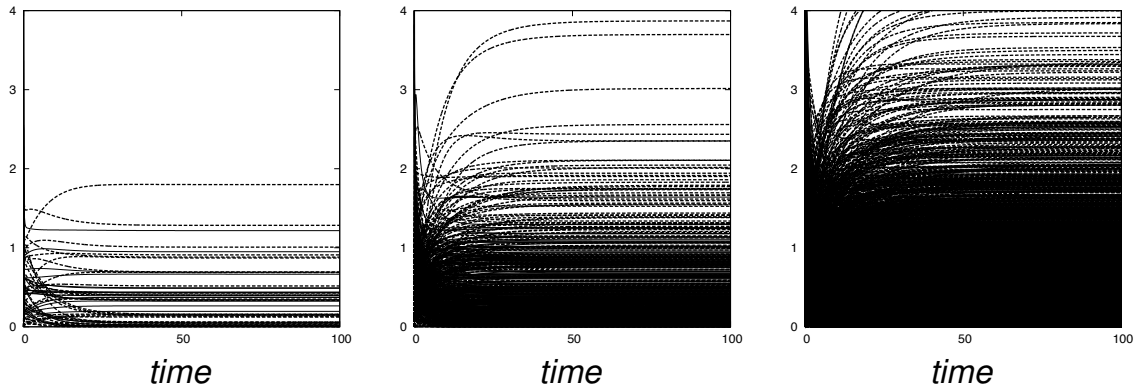
reaction eqn systems are also ‘self-averaging’!
numerical illustration

dashed: complexes
solid: unbound proteins

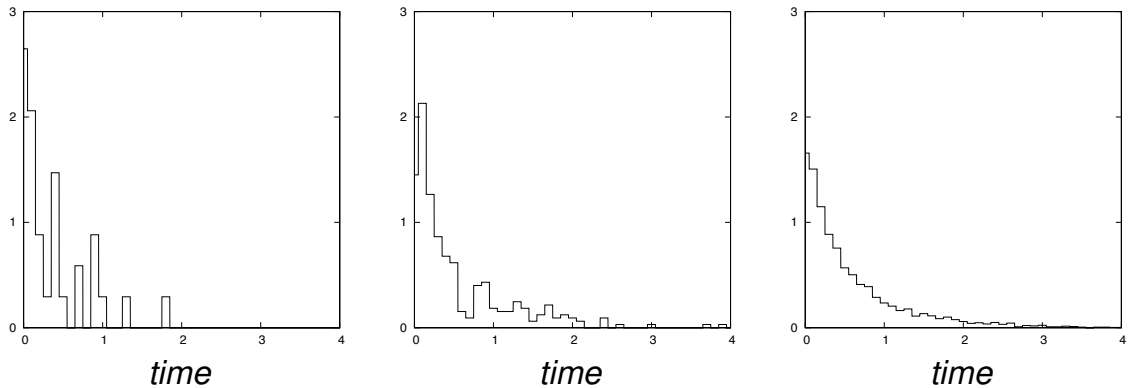
2 post-transl states/protein,
binary complexes,
random topologies & rates,
7 partners on average

10 species *100 species* *1000 species*

individual concentrations



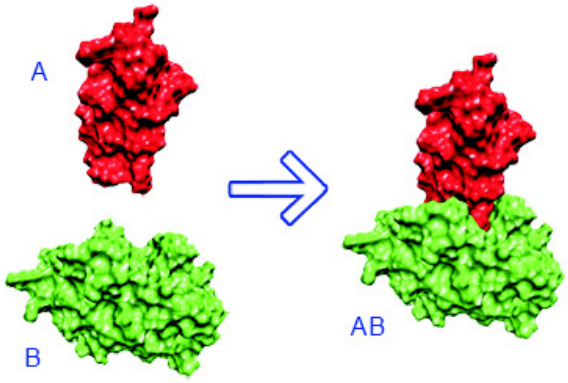
stationary state distribution of concentrations



depends only on param & network *statistics*!

Signalling dynamics in the proteome

from many-particle physics
to *many-particle biology*



- notation:

  $i = 1 \ldots N$ labels proteins
  x_i^α: concentr of protein i in state α
  x_{ij}: concentration of dimer $i \asymp j$

- events:

		rate:
complex formation:	$(i,\alpha) + (j,\beta) \to (i \asymp j)$	$k_{ij}^{\alpha\beta+} x_i^\alpha x_j^\beta$
complex dissociation:	$(i \asymp j) \to (i,\alpha) + (j,\beta)$	$k_{ij}^{\alpha\beta-} x_{ij}$
conformation change:	$(i,\alpha) \to (i,\beta)$	$\lambda_i^{\alpha\beta} x_i^\alpha$
protein degradation:	$(i,\alpha) \to \emptyset$	$\gamma_i^\alpha x_i^\alpha$
protein synthesis:	$\emptyset \to (i,\alpha)$	θ_i^α

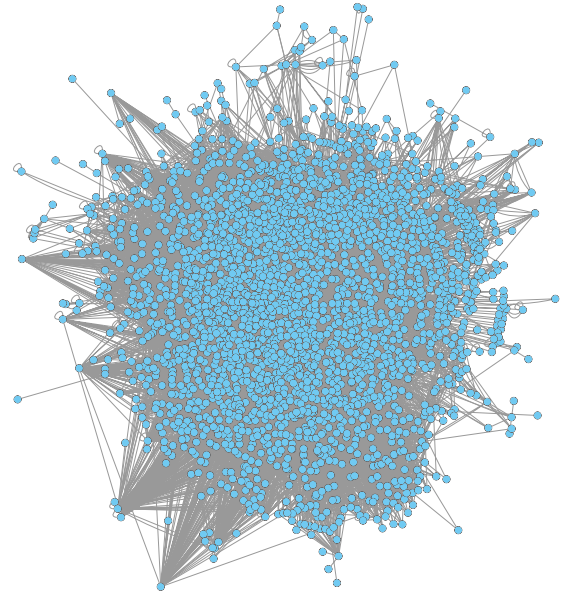
- reaction eqns

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i^\alpha = \sum_j A_{ij} \overbrace{\sum_\beta [k_{ij}^{\alpha\beta-} x_{ij} - k_{ij}^{\alpha\beta+} x_i^\alpha x_j^\beta]}^{\textit{complex formation \& dissociation}} + \overbrace{\sum_\beta [\lambda_i^{\beta\alpha} x_i^\beta - \lambda_i^{\alpha\beta} x_i^\alpha]}^{\textit{post-transl modification}} - \overbrace{\gamma_i^\alpha x_i^\alpha}^{\textit{decay}}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x_{ij} = A_{ij} \sum_{\alpha\beta} [k_{ij}^{\alpha\beta+} x_i^\alpha x_j^\beta - k_{ij}^{\alpha\beta-} x_{ij}]$$

- tailored random PPIN (prescribed degrees)
  $A_{ij} = 0, 1$

$$p(\mathbf{A}) = \frac{\prod_i \delta_{k_i, \sum_{j\neq i} A_{ij}}}{Z} \prod_i \left[c_0 \delta_{c_{ii},1} + (1-c_0)\delta_{c_{ii},0} \right]$$



- draw reaction rates randomly
  from realistic distributions $P(k^+, k^-)$, $P(\lambda, \gamma)$
  (experimental data!)

Generating functional analysis

how to calculate *properties* of solutions
$\mathbf{x}^\star(t)$ *of dynamical equations*

without solving equations ...

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i(t) = F_i[\mathbf{x}(t), \boldsymbol{\theta}], \qquad \mathbf{x} = (x_1, \ldots, x_N)$$

$\boldsymbol{\theta}$: *network topology*, *reaction rates*, ...

- *generating functional*

$$Z[\boldsymbol{\psi}] = \int \Big[\prod_t \mathrm{d}\mathbf{x}(t)\Big] \mathrm{e}^{\sum_{i,t}\psi_i(t)x_i(t)} \overbrace{\prod_{i,t}\delta\Big[x_i(t+\mathrm{d}t) - x_i(t) - F_i[\mathbf{x}(t),\boldsymbol{\theta}]\mathrm{d}t\Big]}^{\text{picks up solution of eqns}}$$

delta function:
$\delta[z]=0$ for $z\neq 0$, $\int \mathrm{d}z\ \delta[z] = 1$

- *note:*

$$Z[\mathbf{0}] = 1, \qquad \lim_{\boldsymbol{\psi}\to\mathbf{0}} \frac{\partial Z[\boldsymbol{\psi}]}{\partial\psi_i(t)} = x_i^\star(t), \qquad \lim_{\boldsymbol{\psi}\to\mathbf{0}} \frac{\partial^2 Z[\boldsymbol{\psi}]}{\partial x_i(t)\partial; x_j(t')} = x_i^\star(t)x_j^\star(t'), \quad \ldots$$

if macroscopic quantities *self-averaging* for large N:

- *average* $Z[\psi]$ *over pars* θ

$$X(t) = \frac{1}{N}\sum_i \langle x_i^\star(t)\rangle_{\boldsymbol{\theta}} = \frac{1}{N}\sum_i \lim_{\boldsymbol{\psi}\to\mathbf{0}} \frac{\partial}{\partial\psi_i(t)}\langle Z[\boldsymbol{\psi}]\rangle_{\boldsymbol{\theta}}$$

$$C(t,t') = \frac{1}{N}\sum_i \langle x_i^\star(t)x_i^\star(t')\rangle_{\boldsymbol{\theta}} = \frac{1}{N}\sum_i \lim_{\boldsymbol{\psi}\to\mathbf{0}} \frac{\partial}{\partial\psi_i(t)\partial\psi_i(t')}\langle Z[\boldsymbol{\psi}]\rangle_{\boldsymbol{\theta}}$$

etc

- *proteome models: after calculations ...*
  (path integral techniques, saddle-point integration, etc)

  large N:

$$W = \mathcal{G}_1[W], \quad D = \mathcal{G}_2[W], \qquad \mathcal{G}_{1,2}: \text{ tricky but } \underline{\text{exact}} \text{ formulas}$$

macroscopic quantities:

$D[\{x\}|\{y\}]$, $W[\{x\}|\{y\}]$

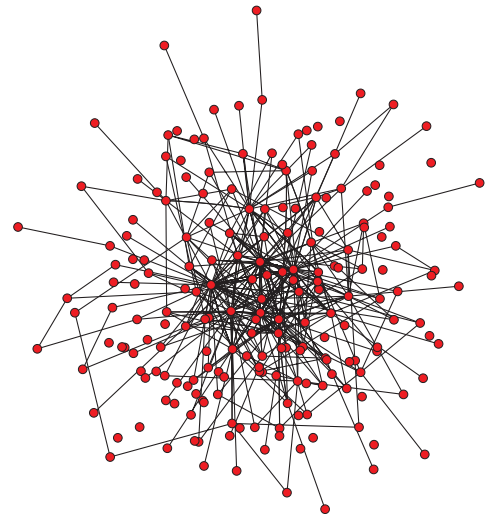
$\{x\}$: *evolving protein concentr* $x_\alpha(t)$

$\{y\}$: *time dep production rates* $y_\alpha(t)$

$D[\{x\}|\{y\}]$: response of (free) protein concentrations to time-dep gene expression perturbations

Random graphs as null models – principles

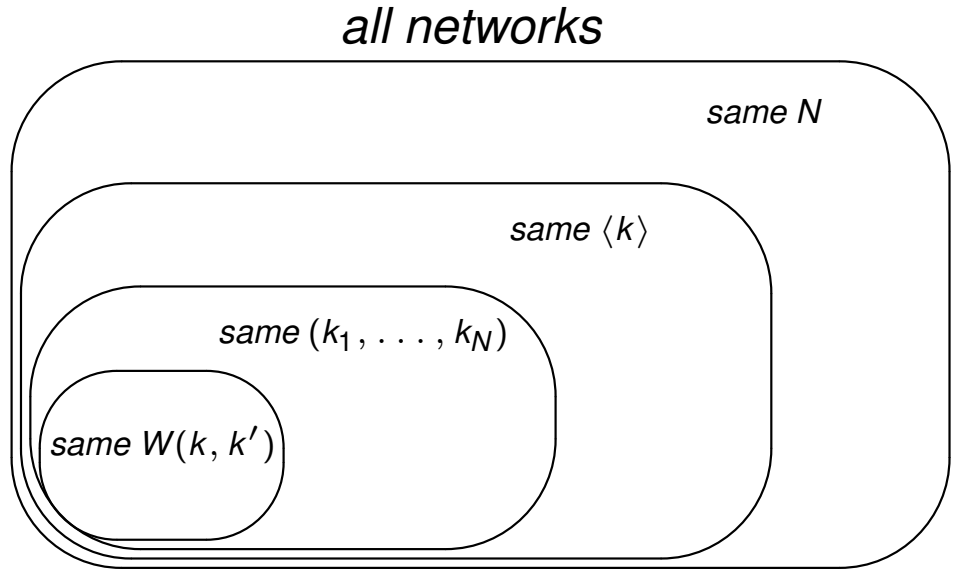
- make an observation to test a hypothesis
  (density of motif, average path length, assortativity, ...)
- is observation statistically significant?
  – how likely is observation in a *null model*?
  – null model: random signalling network



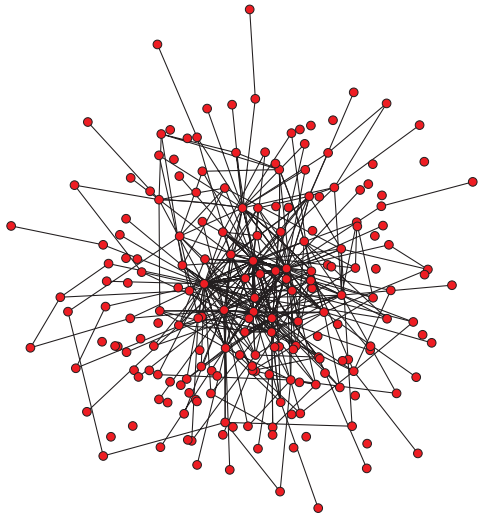
devil is in the detail:
what do we mean by 'random'?

- 1. if null model trivial: test is pointless ...
  2. if null model biased: test is flawed ...

1. which features must the random network inherit from the real one?
2. how to generate random network, with right features but otherwise unbiased?



Common algorithms and their problems



soft constraints only:
standard MCMC dynamics

objective: generate random nondirected graph
with specified probabilities $p(\mathbf{A})$

strategy: start from *any* graph $\mathbf{A}$
propose moves $\mathbf{A} \to F\mathbf{A}$,

use acceptance probabilities $\mathcal{A}(F\mathbf{A}|\mathbf{A})$
obtained from detailed balance condition

$$\mathcal{A}(\mathbf{c}'|\mathbf{A}) = \Big[1 + p(\mathbf{A})/p(\mathbf{A}')\Big]^{-1}$$

stochastic process is ergodic, and converges to $p(\mathbf{A})$

problems:

not all average values of features accessible in practice ...
equilibration can take a *very long* time ...

The problem of phase transitions

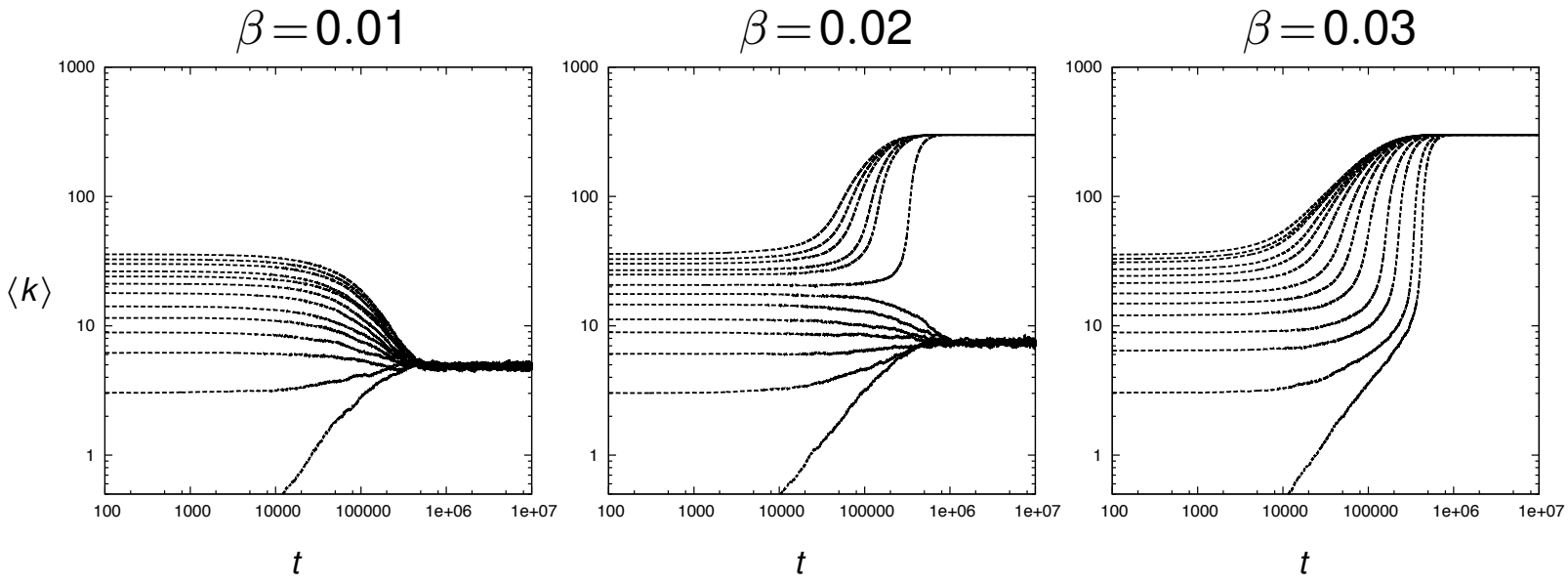
example null model:

random N-node graph with prescribed average and width of $p(k)$

$$p(\mathbf{A}) = \frac{1}{Z} e^{\alpha \sum_i k_i(\mathbf{A}) + \beta \sum_i k_i^2(\mathbf{A})}$$

$N = 300$

$\alpha = 4$



- phase transitions can prevent us from controlling features in soft-constrained ensembles
- need hard-constrained ensembles ...

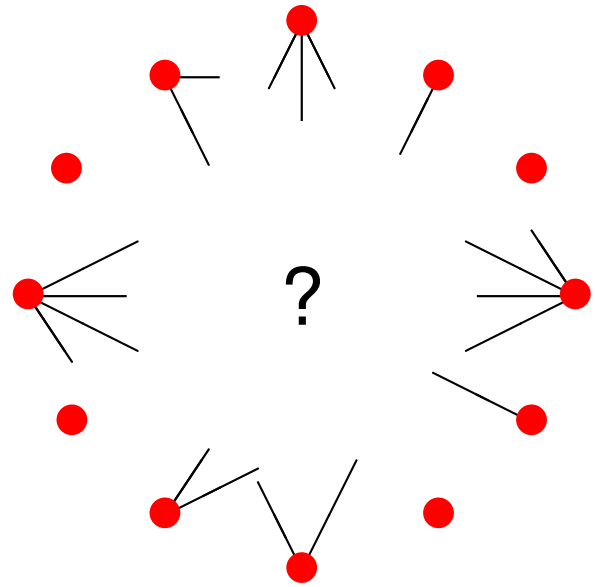
why is generation of graphs with hard constraints nontrivial?

- many users misjudge what the real problem is:

  sampling all networks with given features: usually easy ...
  sampling them with specified probabilities: *nontrivial!*

- many ad-hoc graph generation algorithms *appear* sensible, but lack analysis of which probability $p(\mathbf{A})$ they converge to

Matching algorithm
(Bender and Canfield, 1978)



aim: generate random nondirected graph
with specified degrees $(k_1, \ldots, k_N)$

strategy: stochastic growth,
starting from empty graph

repeat:

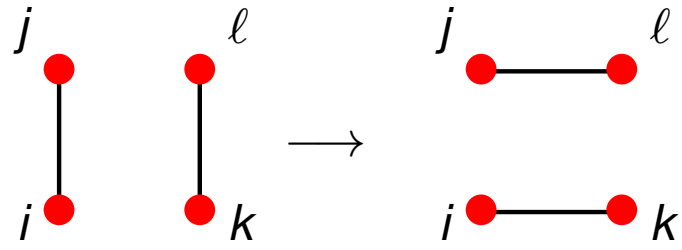
1. pick at random two nodes (i, j)
2. if $\sum_\ell A_{i\ell} < k_i$ and $\sum_\ell A_{j\ell} < k_j$: connect i and j

terminate if $\sum_j A_{ij} = k_i$ for all i

problems:

- impossible to control $p(\mathbf{A})$
- convergence not guaranteed,
process 'hangs' if remaining 'stubs' require self-loops ...
- if process 'hangs', users often don't reject the graph,
this creates correlations between graph realisations $\rightarrow$ *bias*

Edge switching algorithm
(Seidel, 1976)



aim: generate random nondirected graph
with specified degrees $(k_1, \ldots, k_N)$

strategy: degree-preserving 'shuffling' ,
starting from any graph with $(k_1, \ldots, k_N)$

repeat:

1. pick at random four nodes (i, j, k, ℓ)
that are *pairwise connected*

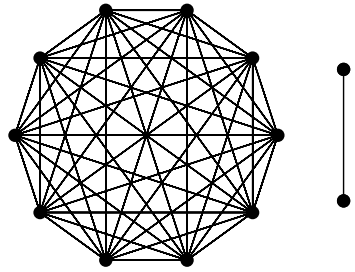
2. carry out an 'edge swap' (preserves degrees!)

terminate if process has equilibrated

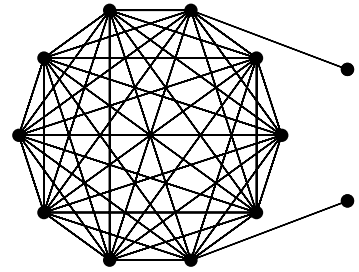
problems:

- cannot control $p(\mathbf{A})$
- sampling is *biased*,
favours graphs on which
many moves are possible

many possible moves

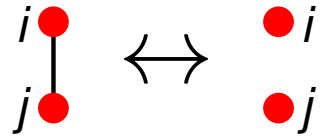
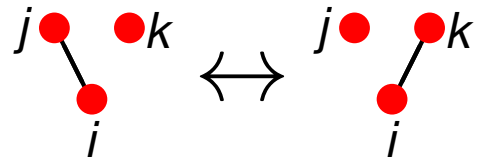
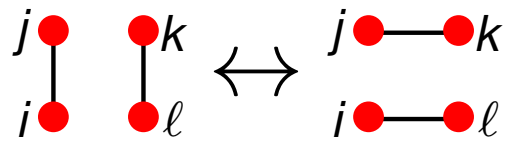


few moves ...



MCMC processes for hard-constrained networks

need to think more carefully about
elementary moves in space of networks

MOVE SET	INVARIANTS	ACTION
Link flips $\{F_{ij}\}$	none	
Hinge flips $\{F_{ijk}\}$	average degree $\bar{k} = \frac{1}{N}\sum_{rs} A_{rs}$	
Edge swaps $\{F_{ijk\ell}\}$	all individual degrees $k_r = \sum_s A_{rs}, \quad r = 1 \ldots N$	

canonical Markov chain

ergodic auto-invertible moves F,
G: all N-node networks that satisfy hard constraints

convergence to $p(\mathbf{A}) = Z^{-1}\mathrm{e}^{-H(\mathbf{A})}$ on G when:

1. pick a candidate move F uniformly at random
2. accept (and execute) with acceptance probabilities:

$$\mathcal{A}(\mathbf{A}|\mathbf{A}') = \frac{n(\mathbf{A}')\mathrm{e}^{-\frac{1}{2}[H(\mathbf{A})-H(\mathbf{A}')]}}{n(\mathbf{A}')e^{-\frac{1}{2}[H(\mathbf{A})-H(\mathbf{A}')]} + n(\mathbf{A})\mathrm{e}^{\frac{1}{2}[H(\mathbf{A})-H(\mathbf{A}')]}}$$

$n(\mathbf{A})$: *nr of moves F that can act on* $\mathbf{A}$

naive edge-swapping?
$\mathcal{A}(\mathbf{A}|\mathbf{A}') = 1$

would give: *sampling bias* : $p(\mathbf{A}) = \dfrac{n(\mathbf{A})}{\sum_{\mathbf{A}' \in G} n(\mathbf{A}')}$

picking candidate moves randomly ...

even this is tricky!

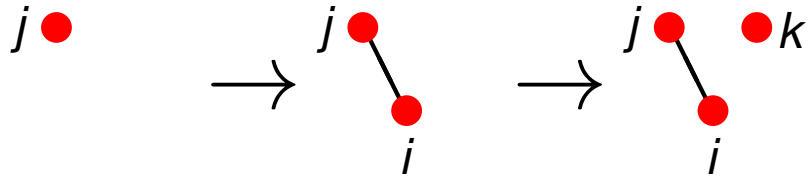
required: $p(F|\mathbf{A}) = 1/n(\mathbf{A})$ ...

PROTOCOL 1:

(i) pick a site j with $k_j(\mathbf{A}) > 0$

(ii) pick a site $i \in \partial_j(\mathbf{A})$

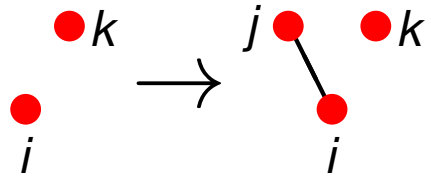
(iii) pick a site $k \notin \partial_i(\mathbf{A}) \cup \{i\}$



PROTOCOL 2:

(i) pick two disconnected sites (i, k) with $k_i(\mathbf{A}) > 0$

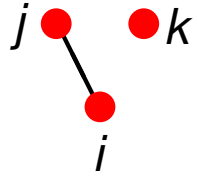
(ii) pick a site $j \in \partial_i(\mathbf{A})$



PROTOCOL 3:

(i) pick two connected sites (i, j) and a third site k

(ii) while $A_{ik} = 1$ return to (i)

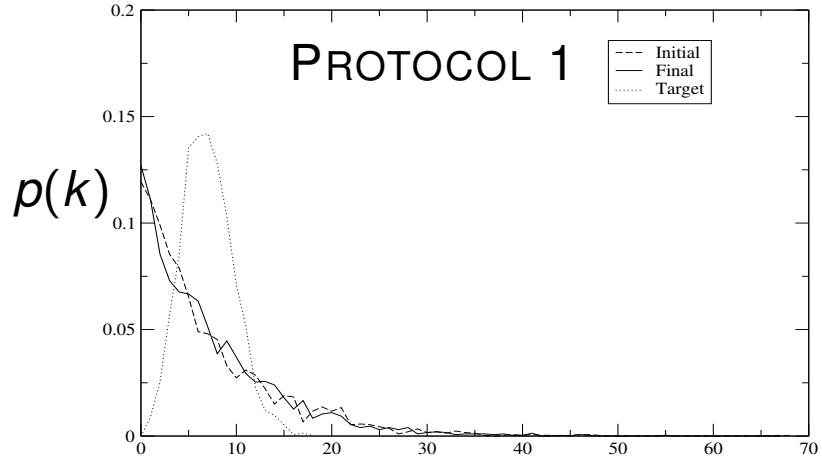


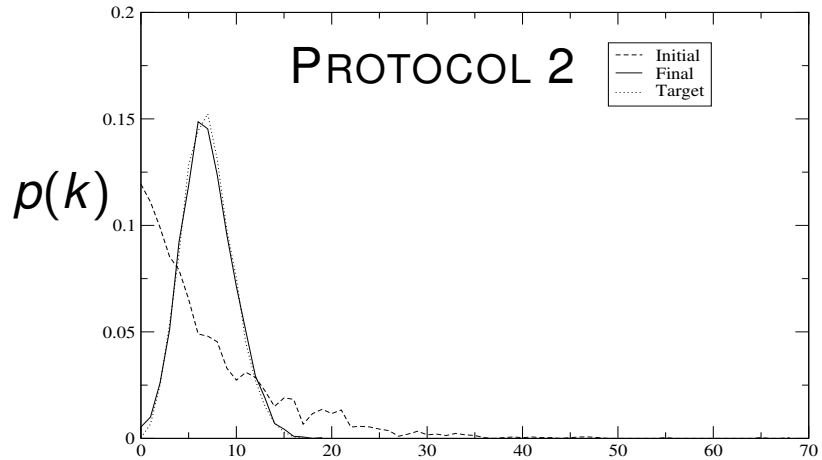
$N = 3000$, $\langle k \rangle = 7$

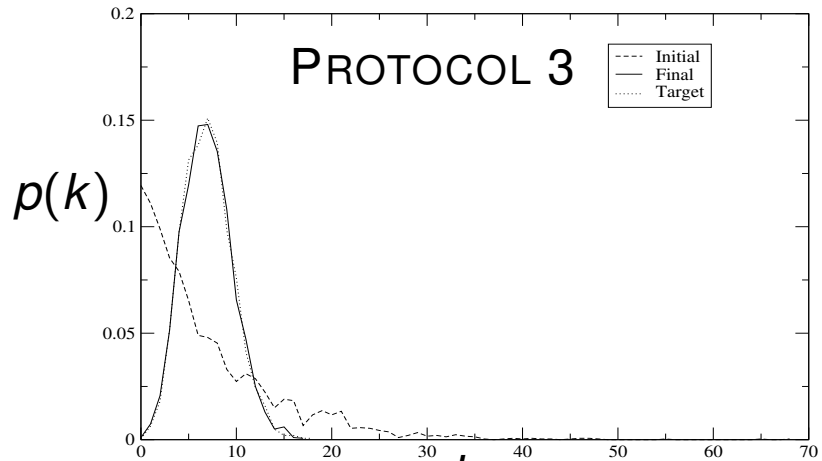
dashed: start graph

dotted: $p(k)$ of target $p(\mathbf{A})$

solid: MCMC result







Mobility of nondirected networks

to implement the Markov chain,
need *analytical formula* for graph mobility $n(\mathbf{A})$

work out combinatorics:

$$n(\mathbf{A}) = \underbrace{\frac{1}{4}N^2\langle k\rangle^2 + \frac{1}{4}N\langle k\rangle - \frac{1}{2}N\langle k^2\rangle}_{\textit{invariant}} + \underbrace{\frac{1}{4}\mathrm{Tr}(\mathbf{A}^4) + \frac{1}{2}\mathrm{Tr}(\mathbf{A}^3) - \frac{1}{2}\sum_{ij} k_i A_{ij} k_j}_{\textit{state dependent}}$$

- state-dep part can be ignored if $\langle k^2\rangle k_{\max}/\langle k\rangle^2 \ll N$
- avoid calculating $n(\mathbf{A})$
  at each iteration step:

  (i) calculate $n(\mathbf{A})$ at time $t = 0$
  (ii) update dynamically, compute $\Delta_F n(\mathbf{A})$ for executed move F

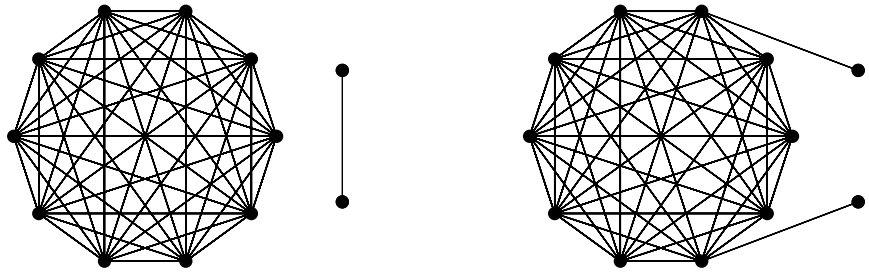
Example:

target: $p(\mathbf{A})$ constant

$N = 100$

naive versus correct
acceptance probabilities

many possible moves *few moves*



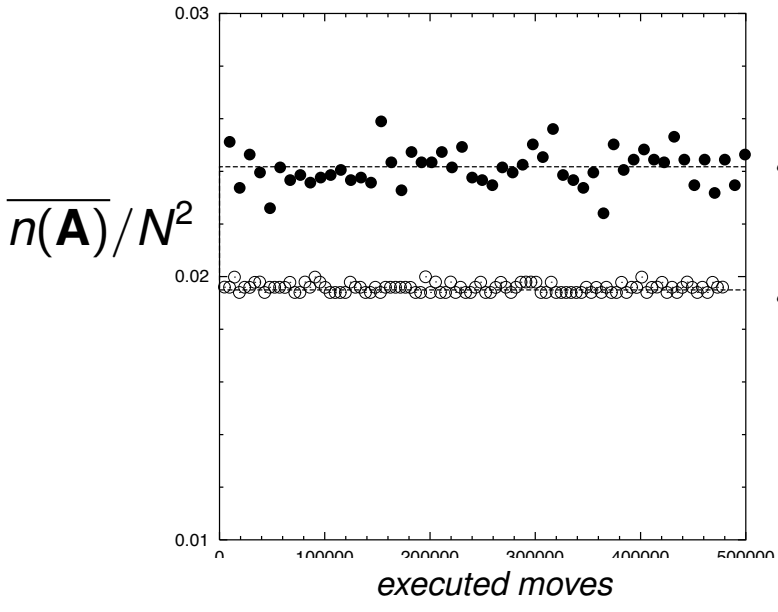
predictions:

$p(\mathbf{A}) = \textit{constant}$:

$\overline{n(\mathbf{A})}/N^2 \approx 0.0195$

$p(\mathbf{A}) = n(\mathbf{A})/Z$:

$\overline{n(\mathbf{A})}/N^2 \approx 0.0242$

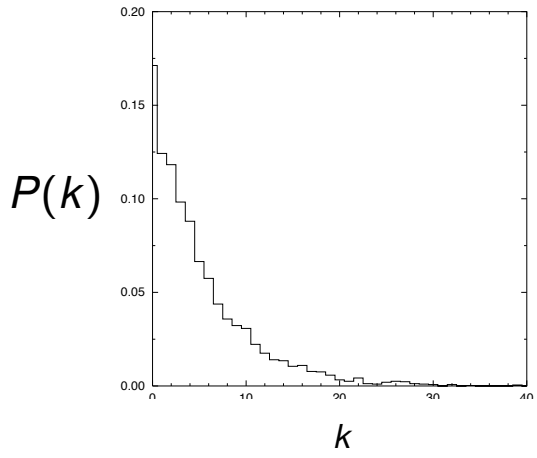


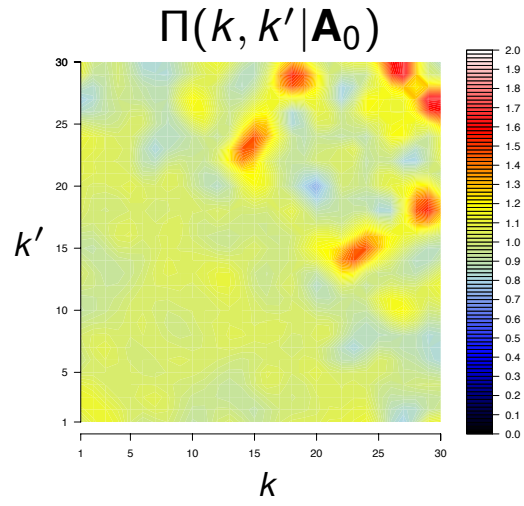
$\mathcal{A}(\mathbf{A}|\mathbf{A}') = 1$

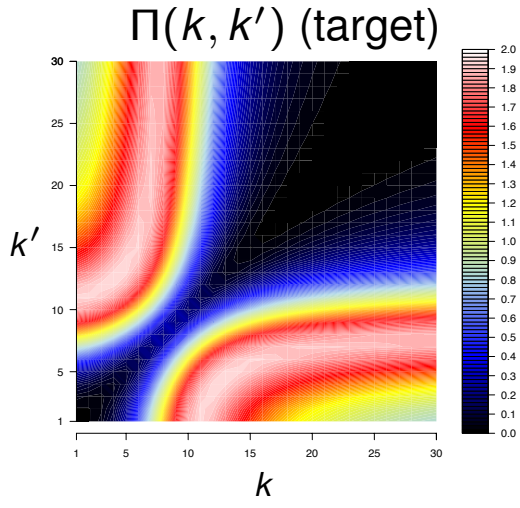
$\mathcal{A}(\mathbf{A}|\mathbf{A}') = [1 + \frac{n(\mathbf{A})}{n(\mathbf{A}')}]^{-1}$

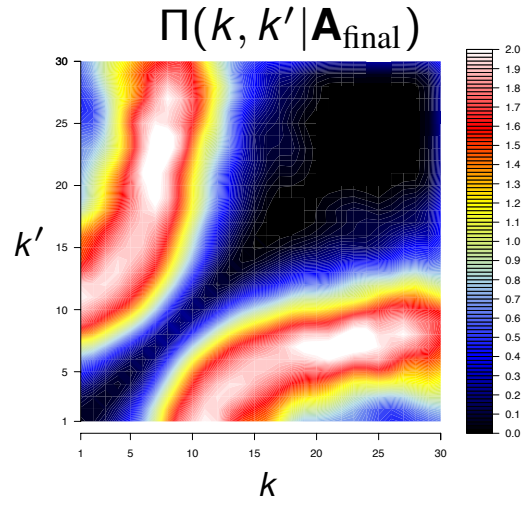
Example

target =
degree-correlated
$p(\mathbf{A})$ on G







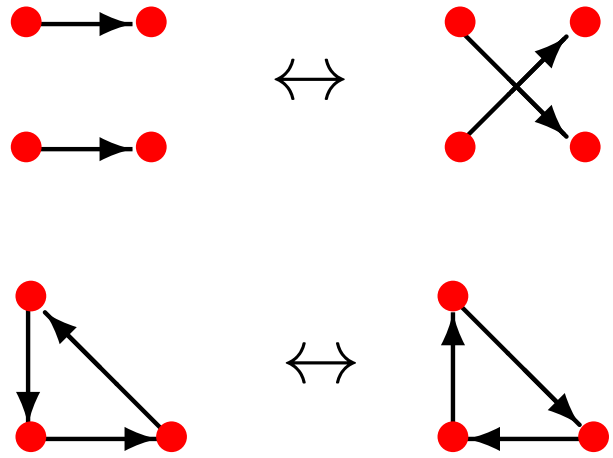


$N = 4000$,
$\langle k \rangle = 5$

$$\Pi(k,k') = \frac{(k-k')^2}{[\beta_1 - \beta_2 k + \beta_3 k^2][\beta_1 - \beta_2 k' + \beta_3 k'^2]}$$

Directed networks

- constraints: in- and out-degrees, $(k_1^{\text{in}}, \ldots, k_N^{\text{in}})$, $(k_1^{\text{out}}, \ldots, k_N^{\text{out}})$
- moves: *directed edge swaps*
- further move type required to restore ergodicity: 3-loop reversal



mobilities $n(\mathbf{A})$:

$$n_{\square}(\mathbf{A}) = \underbrace{\frac{1}{2}N^2\langle k\rangle^2 - \sum_j k_j^{\text{in}} k_j^{\text{out}}}_{\textit{invariant}} + \underbrace{\frac{1}{2}\text{Tr}(\mathbf{c}^2) + \frac{1}{2}\text{Tr}(\mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger\mathbf{A}) + \text{Tr}(\mathbf{A}^2\mathbf{A}^\dagger) - \sum_{ij} k_i^{\text{in}} A_{ij} k_j^{\text{out}}}_{\textit{state dependent}}$$

$$n_{\triangle}(\mathbf{A}) = \underbrace{\frac{1}{3}\text{Tr}(\mathbf{A}^3) - \text{Tr}(\hat{\mathbf{A}}\mathbf{c}^2) + \text{Tr}(\hat{\mathbf{A}}^2\mathbf{A}) - \frac{1}{3}\text{Tr}(\hat{\mathbf{A}}^3)}_{\textit{state dependent}}$$

with: $(\mathbf{A}^\dagger)_{ij} = A_{ji}$, $\hat{\mathbf{A}}_{ij} = A_{ij}A_{ji}$

Statistical characterization and visualization
- Topology statistics beyond degree distributions
- Factor graphs
- Short loops and spectra

Quality of molecular interaction data
- Experimental bias in molecular interaction data
- Modelling the effect of experimental bias
- Experimental bias and loop statistics

Modelling cellular processes at non-local scales
- Statistical biology
- Signalling in the proteome

Hypothesis testing in signalling networks
- Random graphs as null models – the principles
- Common algorithms and their problems
- MCMC processes for hard-constrained networks

Identifying (in)vulnerabilities of signalling networks
- Path lengths and path multiplicities
- Disrupting or protecting signalling processes

Path lengths and path multiplicities

somatic mutations triggering cancer ...
interventions in signalling pathways ...

what are 'crucial' nodes in networks?

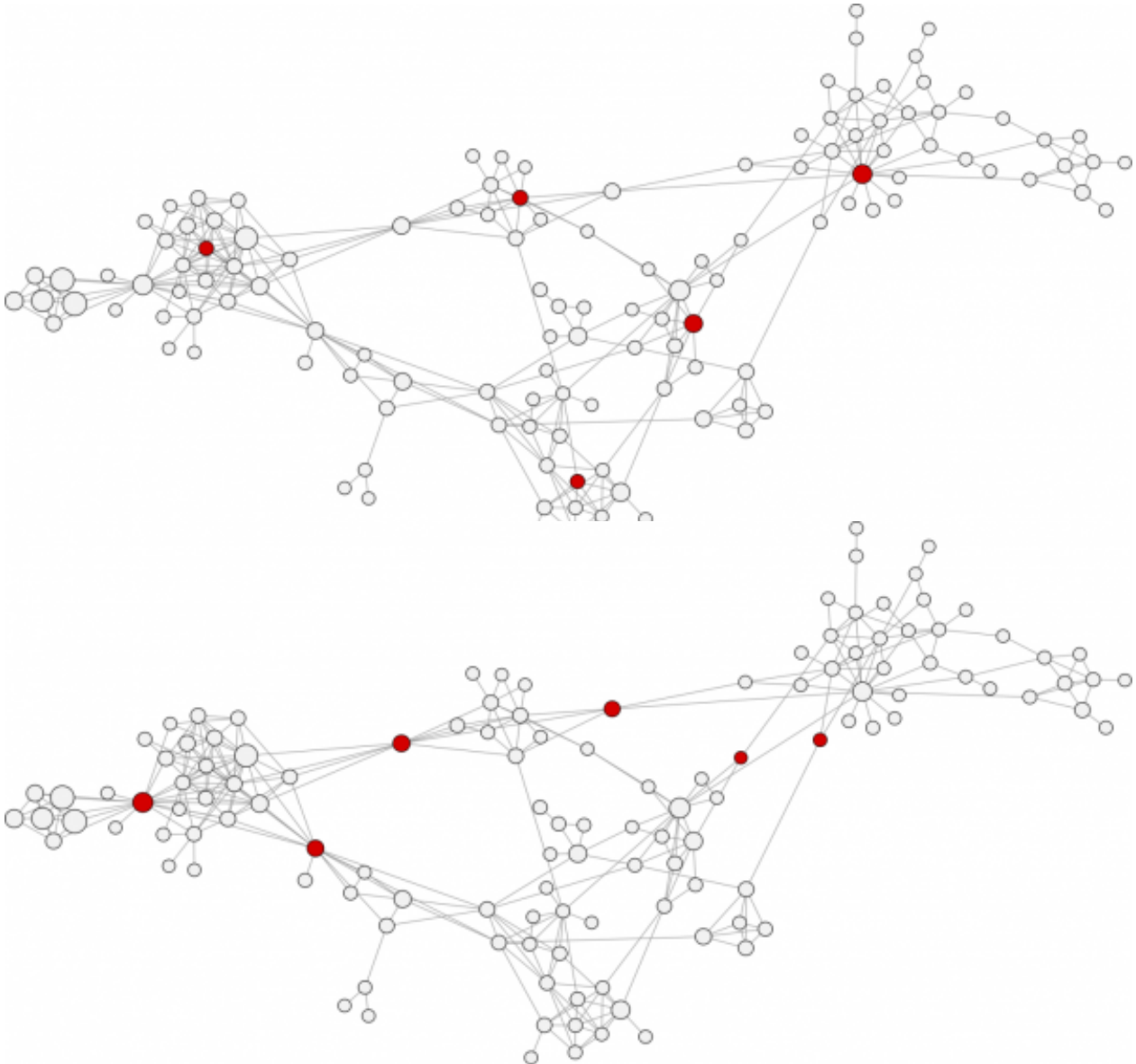
e.g.

closeness centrality of node i:

average length of shortest path between *i* and all other nodes

betweenness centrality of node i:

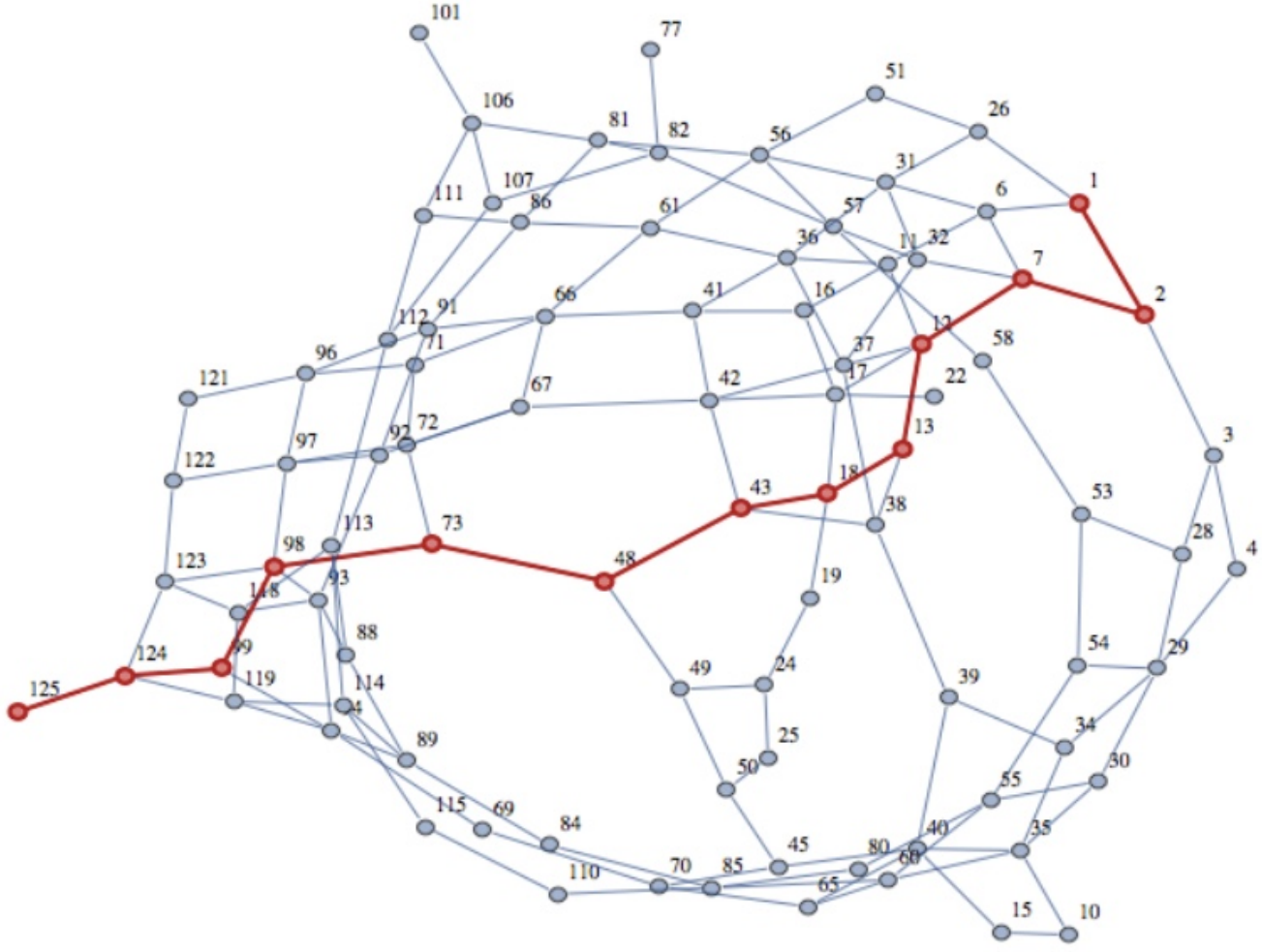
fraction of all shortest paths between node pairs that pass through *i*



all node centrality measures based on *distance*:

d_{ij} :
length of shortest path between nodes (i, j)

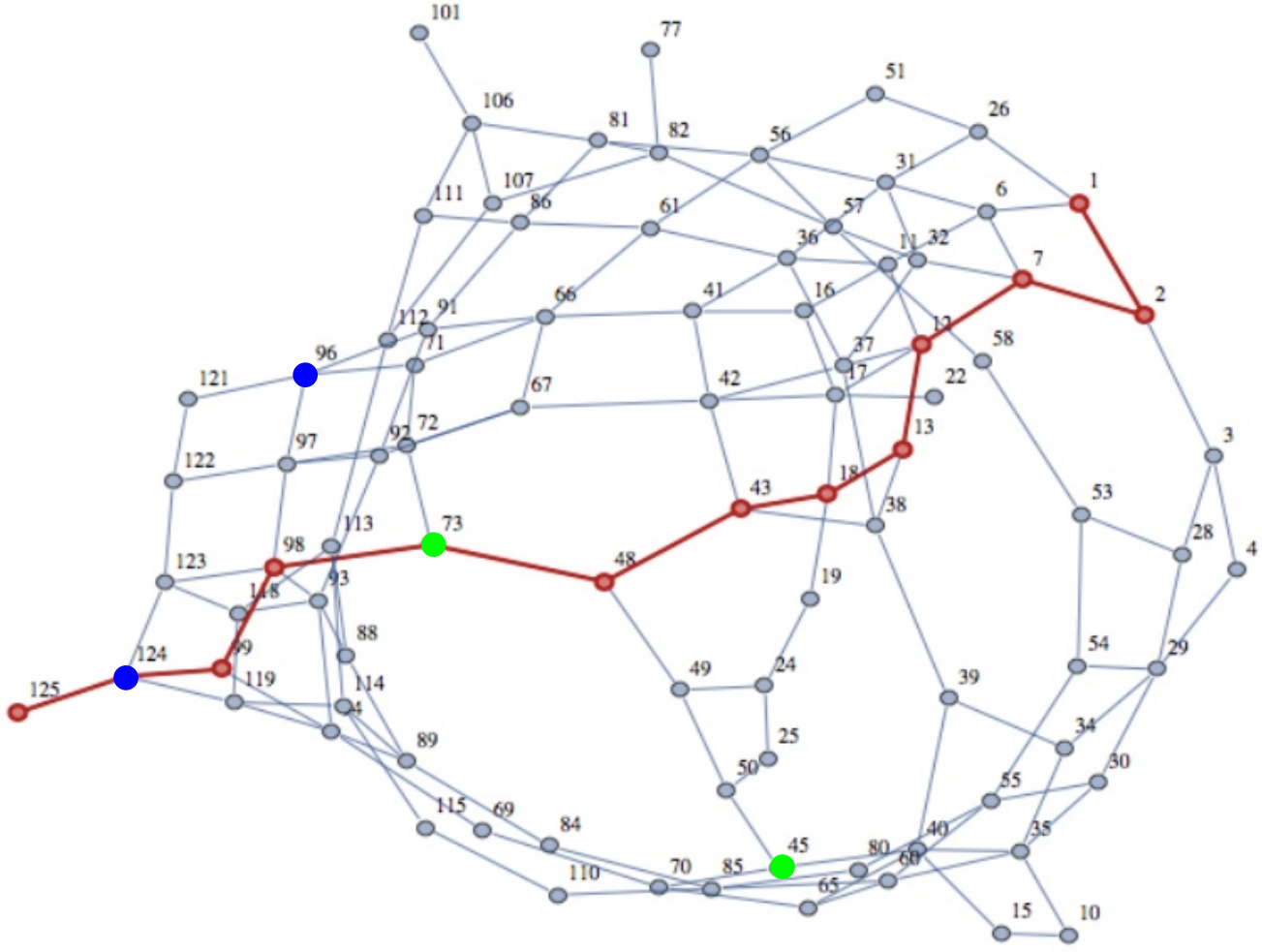
$d_{1,125} = 12$



all node centrality measures based on *distance*:

d_{ij} :
length of shortest path between nodes (i, j)

$d_{1,125} = 12$



completely disregarding multiplicities of paths!
relevant for understanding acquired resistance to therapy?

we would like:

$d_{124,96} < d_{45,73} < 4$ ...

Alternative measures of distance between nodes

intuition:

more paths connecting i and j $\leftrightarrow$ *shorter* distance D_{ij}

- effective connectivity between sites i and j: $$S_{ij}(\Delta) = \sum_{\ell \geq 0} n_{ij}(\ell) \mathrm{e}^{-\ell/\Delta}$$

  $n_{ij}(\ell)$: nr of length-ℓ paths between i and j

  Δ: range over which paths contribute

- effective distance: defined via $S_{ij}(\Delta) = \mathrm{e}^{-D_{ij}(\Delta)/\Delta}$ $$D_{ij}(\Delta) = -\Delta \log S_{ij}(\Delta)$$

work out formulae:

$$D_{ij}(\Delta) = -\Delta \log[(\mathbf{I} - \mathrm{e}^{-1/\Delta}\mathbf{A})^{\mathrm{inv}}]_{ij} \qquad \lim_{\Delta \to 0} D_{ij}(\Delta) = d_{ij}$$

Statistical characterization and visualization
- Topology statistics beyond degree distributions
- Factor graphs
- Short loops and spectra

Quality of molecular interaction data
- Experimental bias in molecular interaction data
- Modelling the effect of experimental bias
- Experimental bias and loop statistics

Modelling cellular processes at non-local scales
- Statistical biology
- Signalling in the proteome

Hypothesis testing in signalling networks
- Random graphs as null models – the principles
- Common algorithms and their problems
- MCMC processes for hard-constrained networks

Identifying (in)vulnerabilities of signalling networks
- Path lengths and path multiplicities
- Disrupting or protecting signalling processes

oncogenesis: *dangerous somatic perturbations*
therapy: *targeted disruptive perturbations*
resistance: *defense against therapeutic perturbations*

Disrupting or protecting networks intelligently

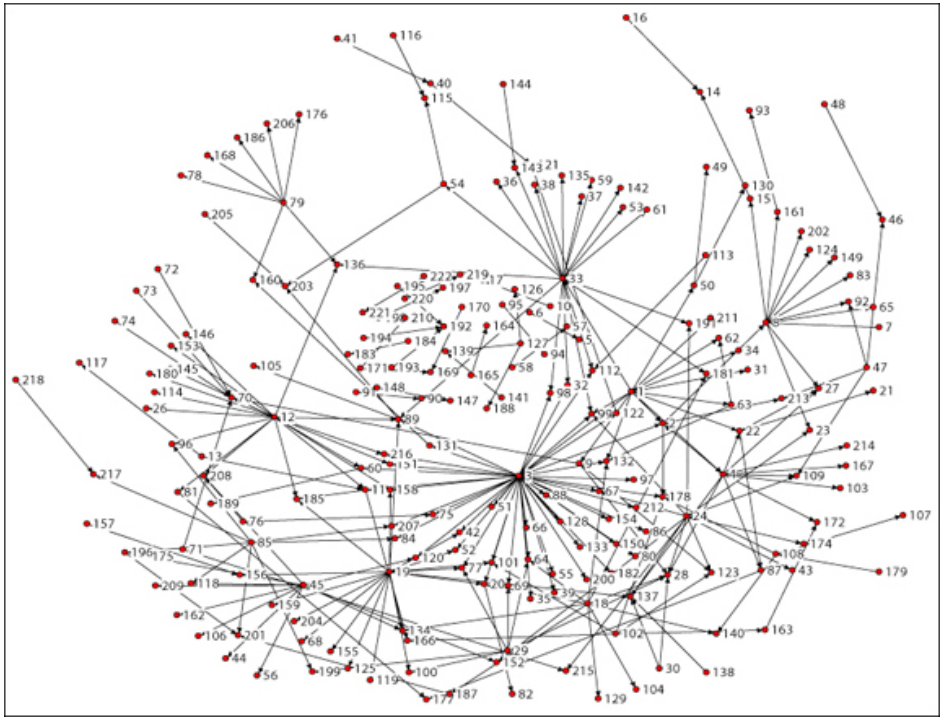
conventional wisdom:

- scale-free networks:

  robust against random attacks
  ('hubs' unlikely to be hit)

- random networks:

  robust against clever attacks
  (no 'hubs' to focus on)



realistic?

remove fixed fraction of nodes ...
unconstrained attack resources ...
success/failure defined in terms of path lengths ...

simplest setup

attacker:

seeks to derail process running on network

defender:

seeks to protect it

- *attack strategy*

  prob $q(\xi|k)$ that a node with degree k is either removed ($\xi = 1$) or left alone ($\xi = 0$)

  constraint: limited attack resources, $\sum_k p(k)q(1|k)\phi(k) \leq 1$,
  $\phi(k)$: 'cost' of removing node with degree k

- *defence strategy*

  degree distribution $p(k)$,
  constraint: limited nr of links, $\sum_k p(k)k \leq c$

what if we forget about path lengths ...

Effect of node attacks on signalling processes

- integrity of process:
  measured by critical noise level $T_c[p,q]$ of ordered state,

  optimal attack $q^\star[p]$: *minimize* $T_c[p,q]$ *subject to* $\sum_k p(k)q(1|k)\phi(k) \leq 1$

  optimal defence $p^\star$: *maximize* $T_c[p,q^\star[p]]$ *subject to* $\sum_k p(k)k \leq c$

- *formulae for* $T_c[p,q]$

  for several process types
  (interacting binary vars, coupled oscillators, ...)

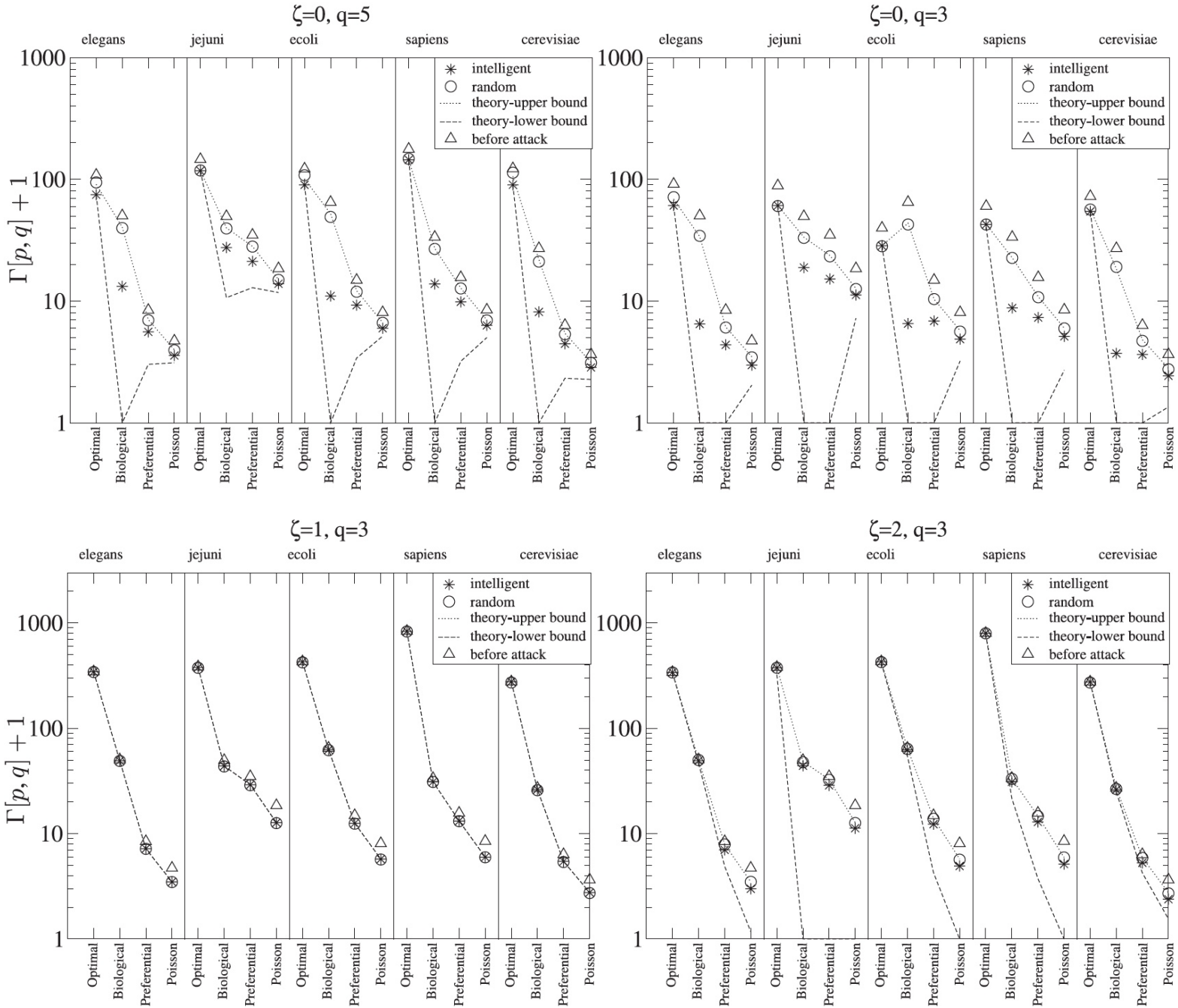
  $T_c[p,q]$ is monotonic function of

$$\Gamma[p,q] = \frac{1}{\langle k \rangle} \sum_k p(k)q(1|k)k(k-1)$$

- intelligent attack cannot improve on random attack when:

  (i) $\phi(k) \propto k(k-1)$ (benefit of knowing degrees balances cost of using it)
  (ii) $p(k) = \delta_{k,\langle k \rangle}$ (regular graphs, no degree info)

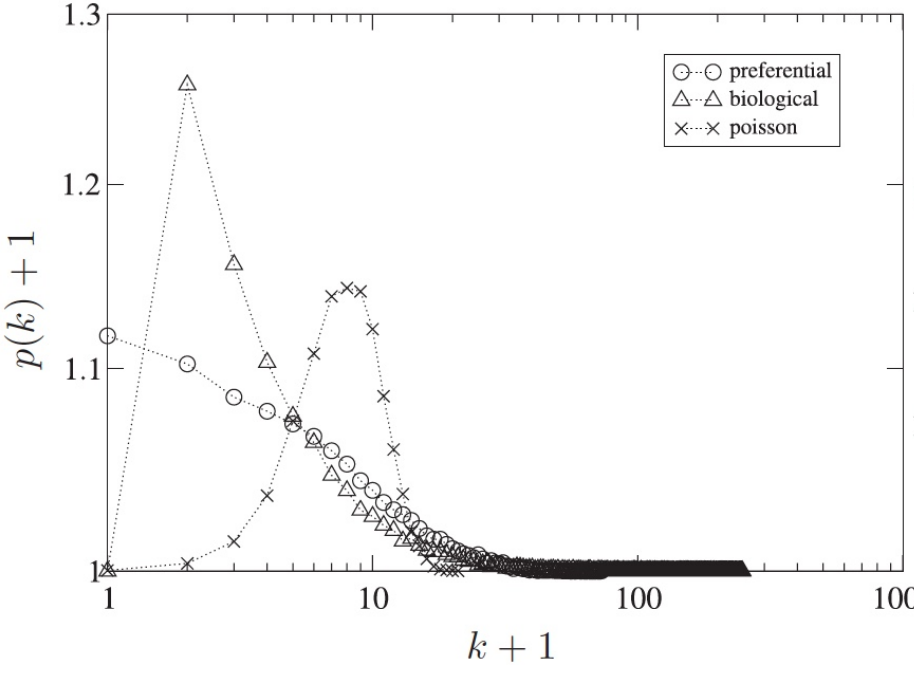
$$\phi(k) \propto qk^{\zeta}(k-1)$$



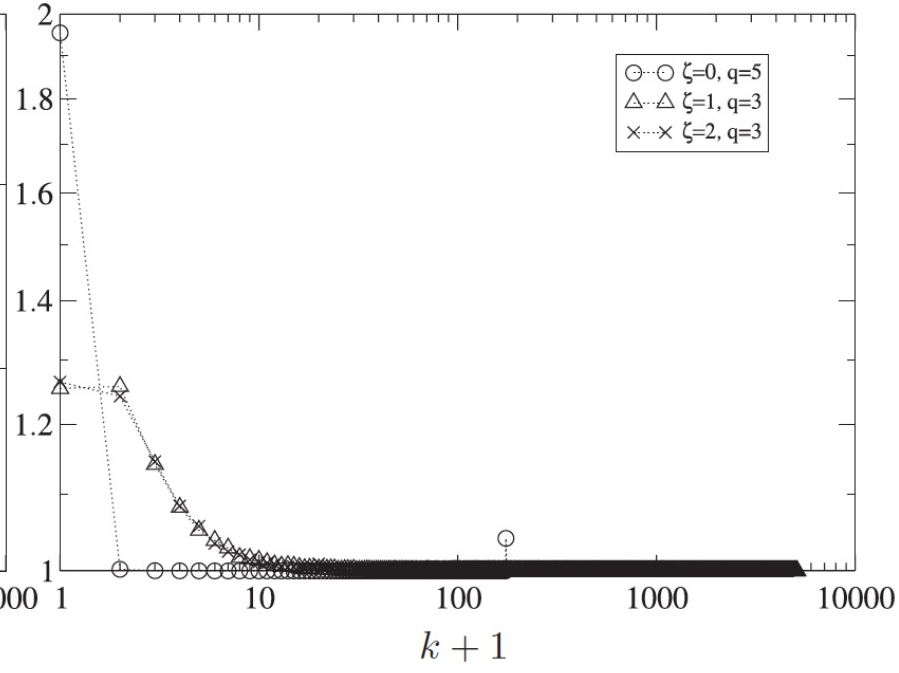
all networks:
same N and $\langle k \rangle$ as human PPIN

$\phi(k) \propto qk^{\zeta}(k-1)$

example degree distributions



optimally resistant degree distribution



Collaborators

King's College London

Alessia Annibale, Luis Fernandes, Franca Fraternali,
Nuria Planell-Morell, Tony Ng, Kate Roberts, Thomas Schlitt

Queen Mary Univ of London

Ginestra Bianconi

Francis Crick Institute

Jens Kleinjung

Univ di Roma La Sapienza

Andrea De Martino

Funding

EPSRC, BBSRC, EU

Papers, seminars, lecture notes

https://nms.kcl.ac.uk/ton.coolen