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
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 \( \sim 7 \)

▶ 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 \( \sim 5 \)
Topology statistics beyond degree distributions

- **degrees:** \[ k_i = \sum_j A_{ij} \]
  
- **distribution:** \[ p(k) = \frac{1}{N} \sum_i \delta_{k,k_i(A)} \]

  pick node \( i \) at random: \( k_i = k \) ?

- **joint degree statistics of connected nodes**

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

  pick link \( (i, j) \) at random

  \[ A_{ij} = 1 \]

  \( k_i = k \) ?

  \( k_j = k' \) ?
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:

\[ k_{i}^{\text{in}} = \sum_{j} A_{ij}, \quad k_{i}^{\text{out}} = \sum_{j} A_{ji}, \quad p(k_{\text{in}}, k_{\text{out}}) = \frac{1}{N} \sum_{i} \delta_{k_{\text{in}}, k_{\text{in}}}^{i} \delta_{k_{\text{out}}, k_{\text{out}}}^{i} \]

▶ joint in-out degree statistics of connected nodes

\[ W(k_{\text{in}}, k_{\text{out}}; k_{\text{in}}', k_{\text{out}}') \]

\[ (k_{\text{in}}, k_{\text{out}})_i = (k_{\text{in}}, k_{\text{out}})? \quad (k_{\text{in}}, k_{\text{out}})_j = (k_{\text{in}}', k_{\text{out}}')? \]

note:

\[ W(k_{\text{in}}, k_{\text{out}}; k_{\text{in}}', k_{\text{out}}') \neq W(k_{\text{in}}', k_{\text{out}}'; k_{\text{in}}, k_{\text{out}}) \]
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
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
**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
Beyond degree distributions and degree correlations
protein interaction networks have many short loops ...

nr of closed paths of length $\ell$:

$$n_\ell = \frac{1}{2\ell} \sum_{i_1,i_2,\ldots,i_{\ell-1}} A_{i_1i_2} A_{i_2i_3} \cdots A_{i_\ell i_1}$$

$$= \frac{1}{2\ell} \sum_i (A^\ell)_{ii}$$

$$= \frac{N}{2\ell} \int d\mu \varrho(\mu) \mu^\ell \quad \varrho(\mu) : \text{spectrum of eigenvalues of } A$$

- 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 $\rho(\mu)$ of protein interaction networks

access to statistics of short loops ...
generally quite different from spectra of non-biological networks!
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
Protein-protein interaction networks and biology—what's the connection?

Luke Hakes¹, John W Pinney¹, David L Robertson¹ & Simon C Lovell¹

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 possible to identify protein homologues with useful accuracy, the growing amount of protein-protein interaction data is beginning to allow a growing body of researchers to ask questions of the proteome in new and powerful ways.
Π(\(k, k'\)) for PPIN:
an accurate problem sensor ...
how to measure dissimilarity between networks?

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

\[
D_{AB} = \frac{1}{2N} \sum_{A} p(A|p_A, \Pi_A) \log \left[ \frac{p(A|p_A, \Pi_A)}{p(A|p_B, \Pi_B)} \right] \\
+ \frac{1}{2N} \sum_{A} p(A|p_B, \Pi_B) \log \left[ \frac{p(A|p_B, \Pi_B)}{p(A|p_A, \Pi_A)} \right]
\]

max entropy distr \( p(A|p, \Pi) \),

result of calculation, large \( N \):

\[
D_{AB} = \frac{1}{2} \sum_{k} p_A(k) \log \left[ \frac{p_A(k)}{p_B(k)} \right] + \sum_{kk'} \frac{p_A(k)p_A(k')kk'}{4\langle k \rangle_A} \Pi_A(k, k') \log \left[ \frac{\Pi_A(k, k')}{\Pi_B(k, k')} \right] \\
+ \frac{1}{2} \sum_{k} p_B(k) \log \left[ \frac{p_B(k)}{p_A(k)} \right] + \sum_{kk'} \frac{p_B(k)p_B(k')kk'}{4\langle k \rangle_B} \Pi_B(k, k') \log \left[ \frac{\Pi_B(k, k')}{\Pi_A(k, k')} \right]
\]

\[
\rho_{AB} : \quad \text{solution of } \quad \rho(k) = \sum_{k'} \Pi_A(k, k')k' p_B(k')/\langle k \rangle_B \rho(k')
\]
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 ...
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
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 \( W(k, k') \)
- true \( p(k) \) and \( W(k, k') \)

In terms of \( x(k), y(k), z(k, k') \)
core result

can be done \textit{analytically}, for large $N$, for all sampling protocols, (via path integrals, steepest descent, ...):

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

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

with

$$\mathcal{J}(k|q) = e^{-a(q)} \sum_{n=0}^{\min\{k-1, q\}} \binom{q}{n} \frac{a^{k-1-n}(q)}{(k-1-n)!} b^n(q)(1 - b(q))^{q-n}$$

$$\mathcal{L}(k|q) = e^{-a(q)} \sum_{n=0}^{\min\{k-1, q-1\}} \binom{q-1}{n} \frac{a^{k-1-n}(q)}{(k-1-n)!} b^n(q)(1 - b(q))^{q-1-n}$$

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

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

– 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 \( A^{\ell,\alpha} \):

\[
A^{\ell,\alpha}_{ij} = \sigma^{\ell,\alpha}_i \sigma^{\ell,\alpha}_j [\tau^{\ell,\alpha}_{ij} A^{\ell}_{ij} + (1 - A^{\ell}_{ij}) \lambda^{\ell,\alpha}_{ij}]
\]

method a method b method c \ldots

<table>
<thead>
<tr>
<th>species I</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>species II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>species III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\vdots</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

objective:

find true PINs \( \{A^1, \ldots, A^L\} \) and
sampling pars \( \{\theta^1, \ldots, \theta^M\} \) (via Bayesian methods)
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
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
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
Analysis of molecular signalling processes

proteome:
usual description
reaction equations

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

<table>
<thead>
<tr>
<th>Table 2. Model Equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{d(RD)}{dt} = k_{81}RDA - k_{18}RD \cdot A + k_{31}RDE - k_{13}RD \cdot E - k_{19}RD + k_{91}R \cdot D + k_{21}RT - k_{12}RD \cdot M$</td>
</tr>
<tr>
<td>$\frac{d(RT)}{dt} = k_{53}RTE - k_{25}RT \cdot E + k_{93}R \cdot T - k_{29}RT - k_{21}RT + k_{62}RTA - k_{26}RT \cdot A - k_{2mRT} \cdot E + k_{M2M} + k_{12}RT$</td>
</tr>
<tr>
<td>$\frac{d(RDE)}{dt} = k_{13}RD \cdot E - k_{31}RDE + k_{43}RE \cdot D - k_{34}RDE + k_{53}RTE$</td>
</tr>
<tr>
<td>$\frac{d(RE)}{dt} = k_{34}RDE - k_{43}RE \cdot D + k_{54}RTE - k_{45}RE \cdot T + k_{94}R \cdot E - k_{49}RE$</td>
</tr>
<tr>
<td>$\frac{d(RTE)}{dt} = k_{45}RE \cdot T - k_{54}RTE + k_{25}RT \cdot E - k_{52}RTE - k_{53}RTE$</td>
</tr>
<tr>
<td>$\frac{d(RTA)}{dt} = k_{26}RT \cdot A - k_{62}RTA - k_{68}RTA + k_{76}RA \cdot T - k_{67}RTA$</td>
</tr>
<tr>
<td>$\frac{d(RA)}{dt} = k_{69}RTA - k_{76}RA \cdot T + k_{97}R \cdot A - k_{79}RA + k_{87}RDA - k_{78}RA \cdot D$</td>
</tr>
<tr>
<td>$\frac{d(RDA)}{dt} = k_{68}RTA + k_{78}RA \cdot D - k_{87}RDA + k_{18}RD \cdot A - k_{81}RDA$</td>
</tr>
<tr>
<td>$\frac{d(R)}{dt} = k_{29}RT - k_{93}R \cdot T + k_{49}RE - k_{94}R \cdot E + k_{19}RD - k_{91}R \cdot D + k_{79}RA - k_{97}R \cdot A$</td>
</tr>
<tr>
<td>$\frac{d(E)}{dt} = k_{31}RDE - k_{13}RD \cdot E + k_{52}RTE - k_{25}RT \cdot E + k_{49}RE - k_{94}R \cdot E - k_{2mRT} \cdot E + k_{M2M}$</td>
</tr>
<tr>
<td>$\frac{d(A)}{dt} = k_{81}RDA - k_{18}RD \cdot A + k_{62}RTA - k_{26}RT \cdot A + k_{79}RA - k_{97}R \cdot A$</td>
</tr>
<tr>
<td>$\frac{d(M)}{dt} = k_{2mRT} \cdot E - k_{M2M}$</td>
</tr>
</tbody>
</table>

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

\[ \sim 10^{24} \text{ positions, velocities} \]

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

Newton’s equations

\[ \frac{d}{dt} (\vec{x}_1, \vec{v}_1) = \ldots, \frac{d}{dt} (\vec{x}_2, \vec{v}_2) = \ldots \quad \leftarrow \text{don’t try to solve these!} \]

macroscopic description:
densities, correlation functions,
perturbation response functions,
phase transitions ...

‘self-averaging’: macroscopic theory only
dependent on statistics of model parameters ...
statistical physics

\[ \sim 10^{24} \text{ positions, velocities} \]
\[(\vec{x}_1, \vec{v}_1), (\vec{x}_2, \vec{v}_2), \ldots \]

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 ... 

‘self-averaging’: macroscopic theory only dependent on statistics of model parameters ...

statistical biology

\[ \sim 10^4 \text{ concentr of proteins \& complexes} \]
\[ \vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots \]

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:

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

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

\[ \text{dashed: complexes} \]
\[ \text{solid: unbound proteins} \]

\[ \text{individual concentrations} \]

\[ \text{stationary state distribution of concentrations} \]

depends only on param & network statistics!
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
Signalling dynamics in the proteome

from many-particle physics
to \textit{many-particle biology}

\begin{itemize}
\item notation:
\begin{align*}
i &= 1 \ldots N \text{ labels proteins} \\
x_i^\alpha &\text{: concentr of protein } i \text{ in state } \alpha \\
x_{ij} &\text{: concentration of dimer } i \rightleftharpoons j
\end{align*}
\end{itemize}

\begin{itemize}
\item events:
\end{itemize}

\begin{align*}
\text{complex formation:} & \quad (i, \alpha) + (j, \beta) \rightarrow (i \rightleftharpoons j) \quad k_{ij}^{\alpha\beta} x_i^\alpha x_j^\beta \\
\text{complex dissociation:} & \quad (i \rightleftharpoons j) \rightarrow (i, \alpha) + (j, \beta) \quad k_{ij}^{\alpha\beta} x_{ij} \\
\text{conformation change:} & \quad (i, \alpha) \rightarrow (i, \beta) \quad \lambda_i^{\alpha\beta} x_i^\alpha \\
\text{protein degradation:} & \quad (i, \alpha) \rightarrow \emptyset \quad \gamma_i^\alpha x_i^\alpha \\
\text{protein synthesis:} & \quad \emptyset \rightarrow (i, \alpha) \quad \theta_i^\alpha
\end{align*}
reaction eqns

\[
\frac{d}{dt} x_i^\alpha = \sum_j A_{ij} \left( \sum_{\beta} \left[ k_{ij}^{\alpha \beta} x_{ij} - k_{ij}^{\beta \alpha} x_i^\alpha x_j^\beta \right] + \sum_{\beta} \left[ \lambda_i^{\beta \alpha} x_i^\beta - \lambda_i^{\alpha \beta} x_i^\alpha \right] - \gamma_i^\alpha x_i^\alpha \right)
\]

\[
\frac{d}{dt} x_{ij} = A_{ij} \left( \sum_{\alpha \beta} \left[ k_{ij}^{\alpha \beta} x_i^\alpha x_j^\beta - k_{ij}^{\beta \alpha} x_{ij} \right] \right)
\]

tailored random PPIN (prescribed degrees)

\[ A_{ij} = 0, 1 \]

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

draw reaction rates randomly from realistic distributions \[ P(k^+, k^-), P(\lambda, \gamma) \]
(experimental data!)
Generating functional analysis

how to calculate properties of solutions \( \mathbf{x}^*(t) \) of dynamical equations without solving equations ...

\[
\frac{d}{dt} x_i(t) = F_i[\mathbf{x}(t), \theta], \quad \mathbf{x} = (x_1, \ldots, x_N)
\]
\[\theta: \text{network topology, reaction rates, ...}\]

▶ generating functional

\[
Z[\psi] = \int \left[ \prod_t d\mathbf{x}(t) \right] e^{\sum_{i,t} \psi_i(t)x_i(t)} \prod_{i,t} \delta \left[ x_i(t+dt) - x_i(t) - F_i[\mathbf{x}(t), \theta]dt \right]
\]

delta function:
\[\delta[z] = 0 \text{ for } z \neq 0, \quad \int dz \delta[z] = 1\]

▶ note:

\[
Z[0] = 1, \quad \lim_{\psi \to 0} \frac{\partial Z[\psi]}{\partial \psi_i(t)} = x_i^*(t), \quad \lim_{\psi \to 0} \frac{\partial^2 Z[\psi]}{\partial x_i(t) \partial x_j(t')} = x_i^*(t)x_j^*(t'), \quad \ldots
\]
if macroscopic quantities *self-averaging* for large $N$:

- **average $Z[\psi]$ over pars $\theta$**

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

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

- **proteome models: after calculations ...**

(path integral techniques, saddle-point integration, etc)

large $N$:

\[
W = G_1[W], \quad D = G_2[W], \quad G_{1,2} : \text{tricky but exact formulas}
\]

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
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
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?

\[
\text{all networks} \quad \begin{cases}
    \text{same } N \\
    \text{same } \langle k \rangle \\
    \text{same } (k_1, \ldots, k_N) \\
    \text{same } W(k, k')
\end{cases}
\]
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
Common algorithms and their problems

soft constraints only:
standard MCMC dynamics

objective: generate random nondirected graph with specified probabilities \( p(A) \)

strategy: start from any graph \( A \)
propose moves \( A \rightarrow FA \),
use acceptance probabilities \( \mathcal{A}(FA|A) \)
obtained from detailed balance condition

\[
\mathcal{A}(c'|A) = \left[ 1 + \frac{p(A)}{p(A')} \right]^{-1}
\]

stochastic process is ergodic, and converges to \( p(A) \)

problems:
not all average values of features accessible in practice ...
equilibration can take a very long time ...
The problem of phase transitions

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

$$ p(A) = \frac{1}{Z} e^{\alpha \sum_i k_i(A) + \beta \sum_i k_i^2(A)} \quad N = 300 $$
$$ \alpha = 4 $$

- 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 \( p(A) \) 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 $p(A)$

▶ 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 \((k_1, \ldots, k_N)\)

\[\text{repeat:}\]
1. pick at random four nodes \((i, j, k, \ell)\) that are pairwise connected
2. carry out an ‘edge swap’ (preserves degrees!)

terminate if process has equilibrated

problems:
- cannot control \(p(A)\)
- sampling is biased, favours graphs on which many moves are possible
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
MCMC processes for hard-constrained networks

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

<table>
<thead>
<tr>
<th>MOVE SET</th>
<th>INVARIANTS</th>
<th>ACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link flips</td>
<td>none</td>
<td><img src="https://example.com/link-flip-diagram" alt="Link flip" /></td>
</tr>
<tr>
<td>${F_{ij}}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hinge flips</td>
<td>average degree $\bar{k} = \frac{1}{N} \sum_{rs} A_{rs}$</td>
<td><img src="https://example.com/hinge-flip-diagram" alt="Hinge flip" /></td>
</tr>
<tr>
<td>${F_{ijk}}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge swaps</td>
<td>all individual degrees $k_r = \sum_s A_{rs}, \ r = 1 \ldots N$</td>
<td><img src="https://example.com/edge-swap-diagram" alt="Edge swap" /></td>
</tr>
<tr>
<td>${F_{ijk\ell}}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
canonical Markov chain

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

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

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

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

$n(A)$: nr of moves $F$ that can act on $A$

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

would give: sampling bias:
$$
p(A) = \frac{n(A)}{\sum_{A' \in G} n(A')}
$$
picking candidate moves randomly ... even this is tricky!

required: \( p(F|A) = 1/n(A) \) ...

---

**PROTOCOL 1:**
(i) pick a site \( j \) with \( k_j(A) > 0 \)
(ii) pick a site \( i \in \partial_j(A) \)
(iii) pick a site \( k \notin \partial_i(A) \cup \{i\} \)

![Diagram for Protocol 1](image)

---

**PROTOCOL 2:**
(i) pick two disconnected sites \( (i, k) \) with \( k_i(A) > 0 \)
(ii) pick a site \( j \in \partial_i(A) \)

![Diagram for Protocol 2](image)

---

**PROTOCOL 3:**
(i) pick two connected sites \( (i, j) \) and a third site \( k \)
(ii) while \( A_{ik} = 1 \) return to (i)

![Diagram for Protocol 3](image)
$N = 3000, \langle k \rangle = 7$

dashed: start graph
dotted: $p(k)$ of target $p(A)$
solid: MCMC result
Mobility of nondirected networks

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

work out combinatorics:

$$n(A) = \frac{1}{4} N^2 \langle k \rangle^2 + \frac{1}{4} N \langle k \rangle - \frac{1}{2} N \langle k^2 \rangle + \left( \frac{1}{4} \text{Tr}(A^4) + \frac{1}{2} \text{Tr}(A^3) - \frac{1}{2} \sum_{ij} k_i A_{ij} k_j \right)$$

- state-dep part can be ignored if $\langle k^2 \rangle k_{\text{max}}^2 / \langle k \rangle^2 \ll N$
- avoid calculating $n(A)$ at each iteration step:
  (i) calculate $n(A)$ at time $t = 0$
  (ii) update dynamically, compute $\Delta_F n(A)$ for executed move $F$
Example:

target: $p(A)$ constant

$N = 100$

naive versus correct acceptance probabilities

predictions:

$p(A) = \text{constant}$:

$\frac{n(A)}{N^2} \approx 0.0195$

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

$\frac{n(A)}{N^2} \approx 0.0242$
Example

\[ \text{target} = \text{degree-correlated} \]

\( p(A) \) on \( G \)

\[ N = 4000, \quad \langle k \rangle = 5 \]

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

\[
\begin{align*}
\mathcal{n}_{\square}(A) &= 1/2N^2 \langle k \rangle^2 - \sum_j k_j^{\text{in}} k_j^{\text{out}} + \frac{1}{2} \text{Tr}(c^2) + \frac{1}{2} \text{Tr}(A^\dagger AA^\dagger A) + \text{Tr}(A^2 A^\dagger) - \sum_{ij} k_j^{\text{in}} A_{ij} k_j^{\text{out}} \\
\mathcal{n}_{\triangle}(A) &= \frac{1}{3} \text{Tr}(A^3) - \text{Tr}(\hat{A}c^2) + \text{Tr}(\hat{A}^2 A) - \frac{1}{3} \text{Tr}(\hat{A}^3)
\end{align*}
\]

\(n_{\square}(A)\) is **invariant**

\(n_{\triangle}(A)\) is **state dependent**

with: \((A^\dagger)_{ij} = A_{ji}, \ \hat{A}_{ij} = A_{ij}A_{ji}\)
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 ... what are ‘crucial’ interventions in signalling pathways ... 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:

\[ d_{124,96} < d_{45,73} < 4 \ldots \]
Alternative measures of distance between nodes

intuition: more paths connecting \( i \) and \( j \) \( \leftrightarrow \) shorter distance \( D_{ij} \)

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

  \( n_{ij}(\ell) \): nr of length-\( \ell \) paths between \( i \) and \( j \)
  \( \Delta \): range over which paths contribute

- effective distance: defined via
  \[
  D_{ij}(\Delta) = -\Delta \log S_{ij}(\Delta)
  \]

work out formulae:

\[
D_{ij}(\Delta) = -\Delta \log[(I - e^{-1/\Delta}A)^{\text{inv}}]_{ij}
\]

\[
\lim_{\Delta \to 0} D_{ij}(\Delta) = d_{ij}
\]
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,

\[
\text{optimal attack } q^*[p]: \quad \text{minimize } T_c[p, q] \text{ subject to } \sum_k p(k)q(1|k)\phi(k) \leq 1
\]

\[
\text{optimal defence } p^*: \quad \text{maximize } T_c[p, q^*[p]] \text{ subject to } \sum_k p(k)k \leq c
\]

- 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

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