# What you see is *not* what you get: how sampling affects macroscopic features of biological networks

ACC Coolen and A Annibale

Dept of Mathematics and Randall Division
King's College London

*published in:*

*Royal Soc Interface Focus 1 (2011) 836-856*

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

# Outline

# Background

**protein interaction networks** (PIN):

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

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

data in public databases,

*used for:*

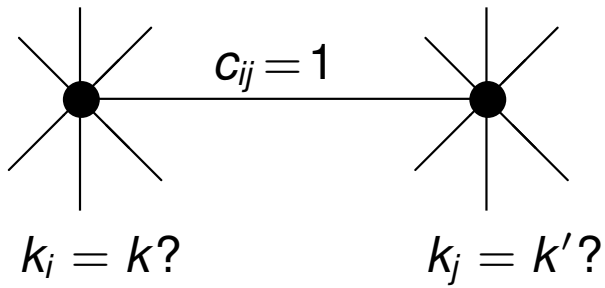
*understanding biology, filtering expression data in medical prediction*

**Quantify topology**:

- degree of a node (nr of links): $k_i(\mathbf{c}) = \sum_j c_{ij}$
  distribution: $p(k) = N^{-1} \sum_i \delta_{k,k_i(\mathbf{c})}$
- degree stats of *connected* nodes

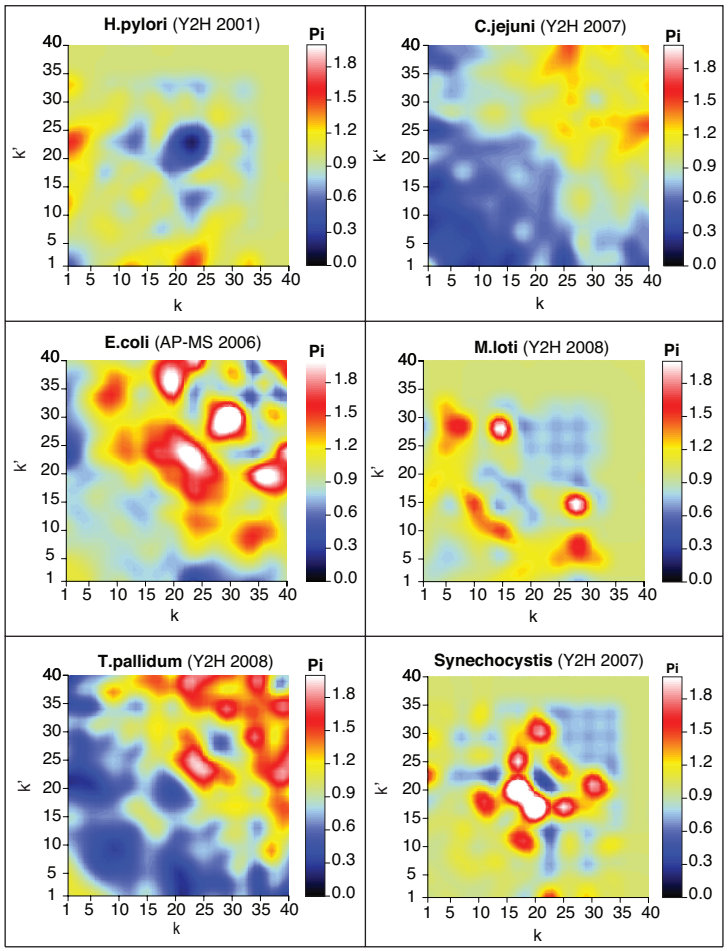
$$W(k,k') = \frac{1}{N\langle k \rangle} \sum_{ij} c_{ij}\ \delta_{k,k_i(\mathbf{c})} \delta_{k',k_j(\mathbf{c})}$$

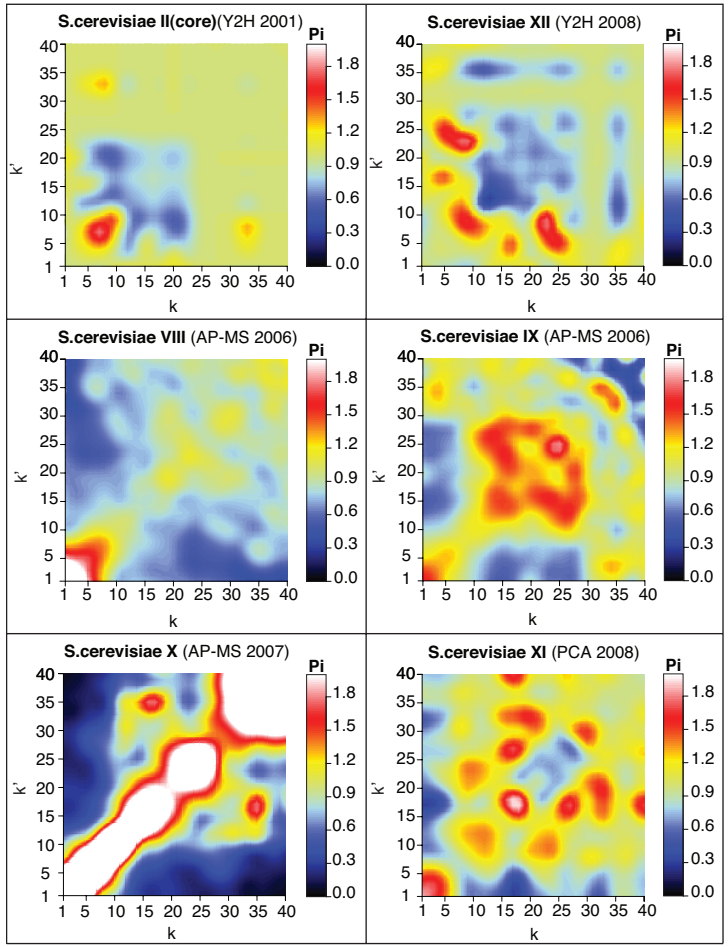
$W(k,k') \neq W(k)W(k')$:
*structural info in degree correlations*



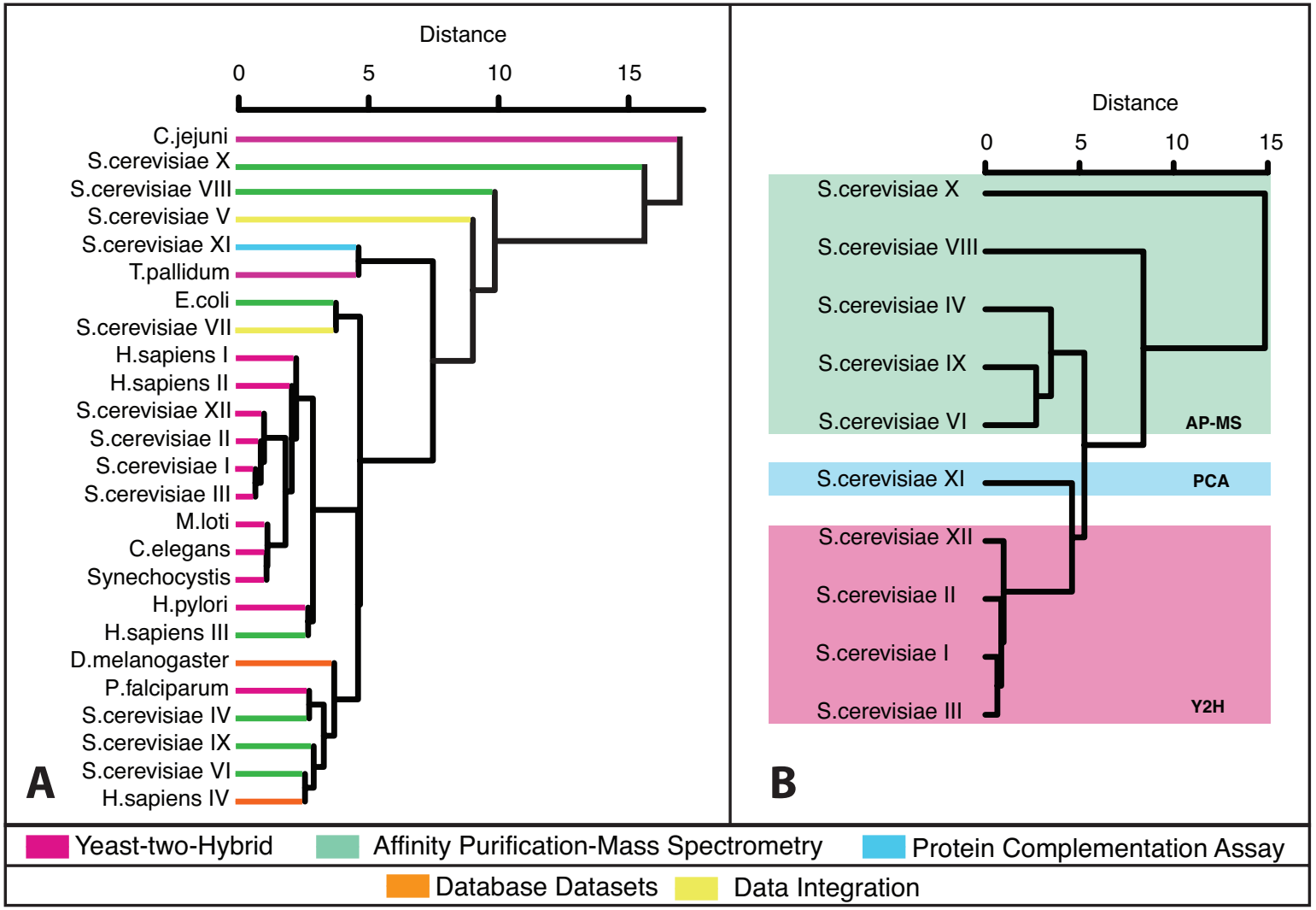
# Information in degree correlations?

plot $\Pi(k, k') = W(k, k')/W(k)W(k')$
the PIN reproducibility problem ...



clustering of PINs using information-theoretic distance measure



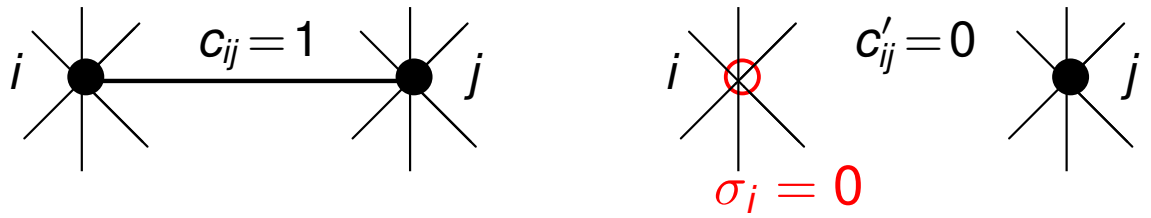

Strong
**bias introduced by experimental method**
that overrules species information ...

# Analysis of the network sampling process

## types of sampling errors

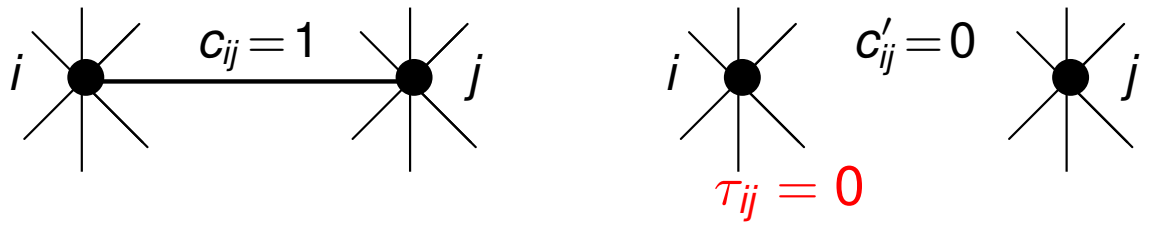
- node undersampling:

$c'_{ij} = \sigma_i \sigma_j c_{ij}$
$\sigma_i = 0, 1$



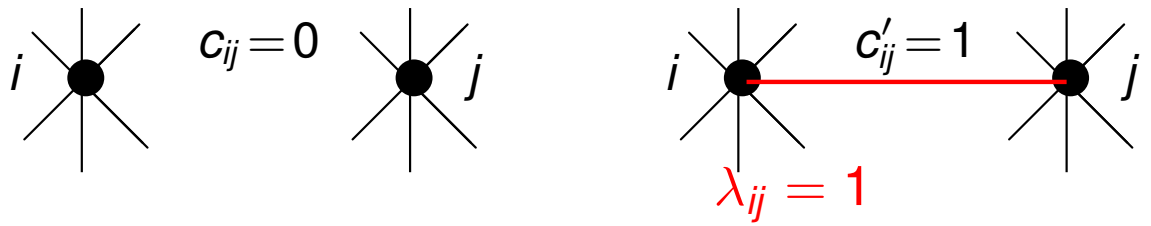

- link undersampling:

$c'_{ij} = \tau_{ij} c_{ij}$
$\tau_{ij} = 0, 1$




- link oversampling:

$c'_{ij} = c_{ij} + \lambda_{ij}(1 - c_{ij})$
$\lambda_{ij} = 0, 1$




combined:

$$c'_{ij} = \sigma_i \sigma_j [\tau_{ij} c_{ij} + (1 - c_{ij}) \lambda_{ij}], \qquad N' = \sum_{i=1}^{N} \sigma_i$$

sampling variables: $\boldsymbol{\sigma} = \{\sigma_i\},\ \boldsymbol{\tau} = \{\tau_{ij}\},\ \boldsymbol{\lambda} = \{\lambda_{ij}\}$

drawn randomly and independently,

| | |
|---|---|
| $x(k)$: | prob to *detect node* with degree $k$ |
| $y(k,k')$: | prob to *detect link* between nodes with degrees $(k,k')$ |
| $z(k,k)/N$: | prob to *report nonexisting link* between nodes with degrees $(k,k')$ |

unbiased sampling: $x(k)=x,\ y(k,k')=y,\ z(k,k')=z$

**aims:**

– express features of $\mathbf{c}'$ in term of those of $\mathbf{c}$, and *vice versa*
– analytically if possible!
– decontaminate public data bases for protocol-specific bias

*studies so far (e.g. Stumpf et al 2005, 2006):*
*limited analytical results, for random sampling only,*
*no results on oversampling*

**strategy**:

- family of networks $\mathbf{c}$ with $N$ nodes, degrees $(k_1, \ldots, k_N)$, and characteristics $p(k)$ and $W(k,k')$:

$$p(\mathbf{c}) = \frac{1}{Z_N}\Big[\prod_i \delta_{k_i,k_i(\mathbf{c})}\Big]\prod_{i<j}\Big[\frac{\langle k\rangle}{N}\frac{W(k_i,k_j)}{p(k_i)p(k_j)}\delta_{c_{ij},1} + \Big(1-\frac{\langle k\rangle}{N}\frac{W(k_i,k_j)}{p(k_i)p(k_j)}\Big)\delta_{c_{ij},0}\Big]$$

- sample networks $\mathbf{c}'$:

$$c'_{ij} = \sigma_i\sigma_j[\tau_{ij}c_{ij} + (1-c_{ij})\lambda_{ij}]$$

sampling protocol:

$x(k)$ *node undersampling* $\quad$ $y(k,k')$ *link undersampling* $\quad$ $z(k,k')$ *link oversampling*

- calculate average characteristics of *sampled* networks

$$p(k|x,y,z) = \Big\langle\!\Big\langle \frac{\sum_i \sigma_i\delta_{k,k_i(\mathbf{c}')}}{\sum_i \sigma_i}\Big\rangle\!\Big\rangle_{\mathbf{c},\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}}$$

$$W(k,k'|x,y,z) = \Big\langle\!\Big\langle \frac{\sum_{ij} c'_{ij}\delta_{k,k_i(\mathbf{c}')}\delta_{k',k_j(\mathbf{c}')}}{\sum_{ij} c'_{ij}}\Big\rangle\!\Big\rangle_{\mathbf{c},\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}}$$

in terms of true $p(k)$ and $W(k,k')$ ...

## **core result**:

can be done *analytically*, as $N \to \infty$,
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

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

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

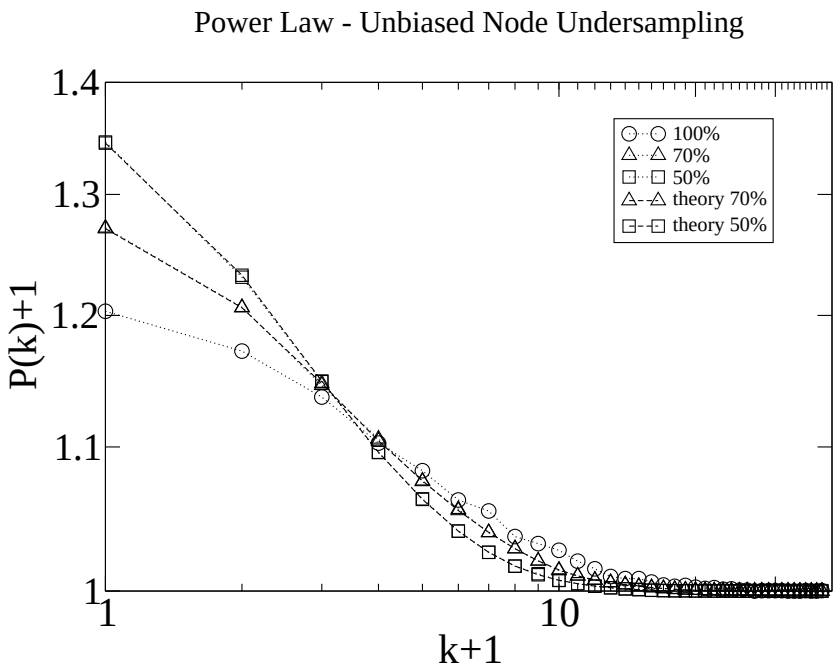
$$a(q) = \sum_{q'\geq 0} p(q')x(q')z(q,q'), \qquad b(q) = \frac{\langle k\rangle}{qp(q)} \sum_{q'\geq 0} x(q')y(q,q')W(q,q')$$

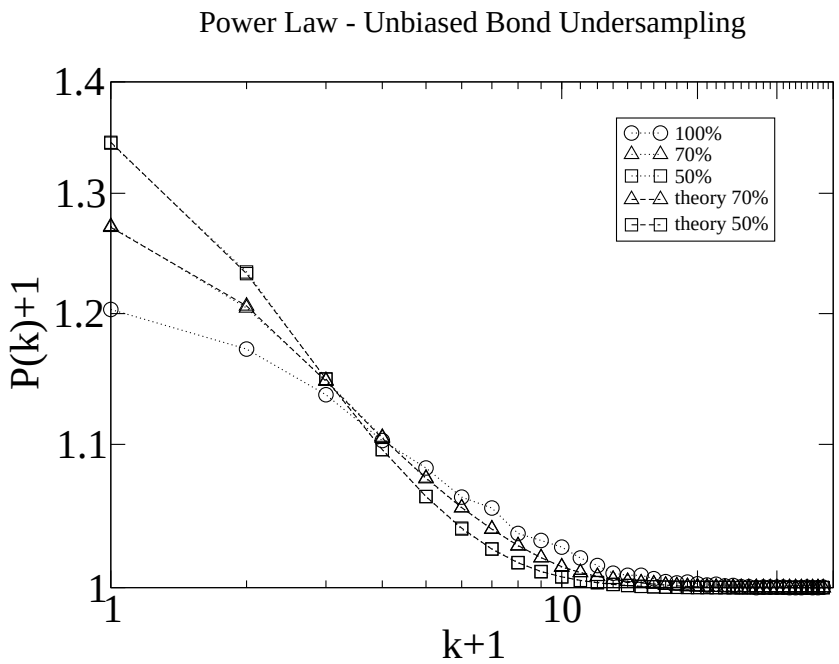
$$\bar{k}(x,y,z) = \frac{\sum_q x(q)p(q)[a(q)+q\,b(q)]}{\sum_q p(q)x(q)}$$

# Applications of the theory

## impact on network degree distributions

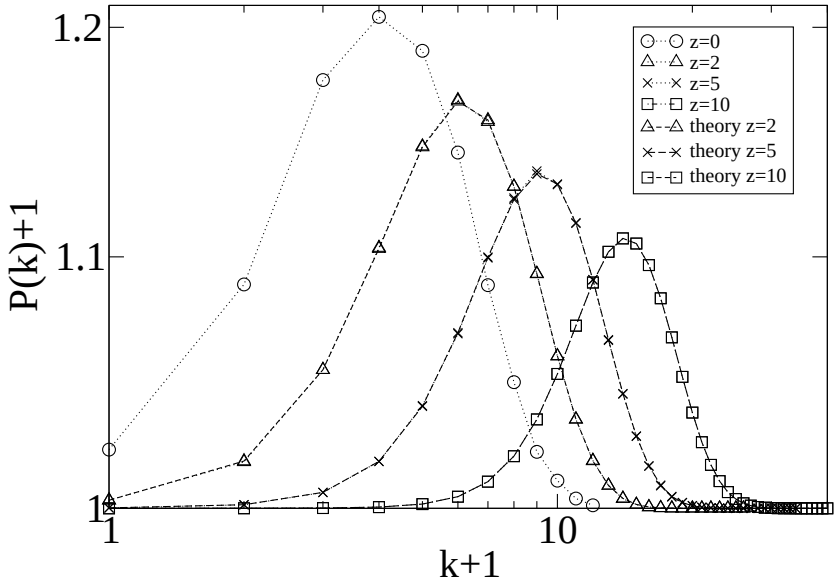
- only Poisson distribution retains shape
- unbiased sampling: recover formulae of Stumpf et al 2005
  node & link undersampling equivalent
- biased sampling: node & link undersampling not equivalent



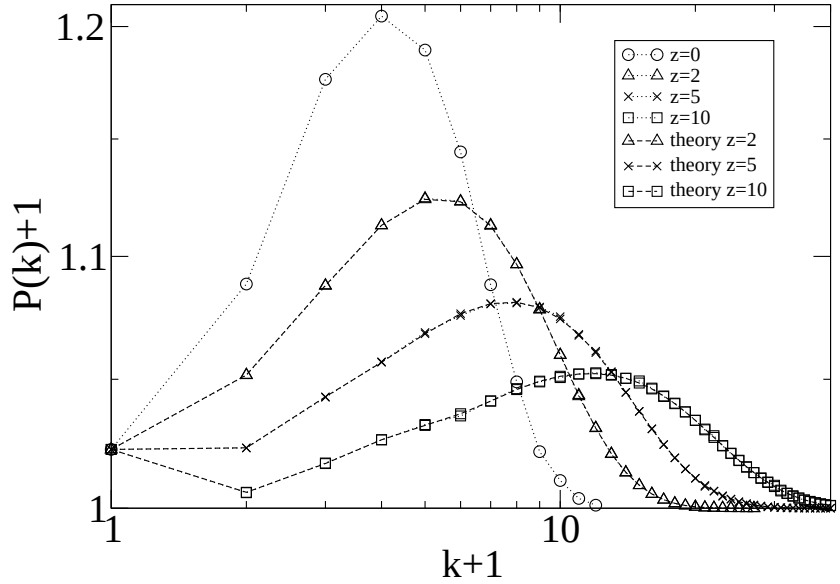




$N = 3512$

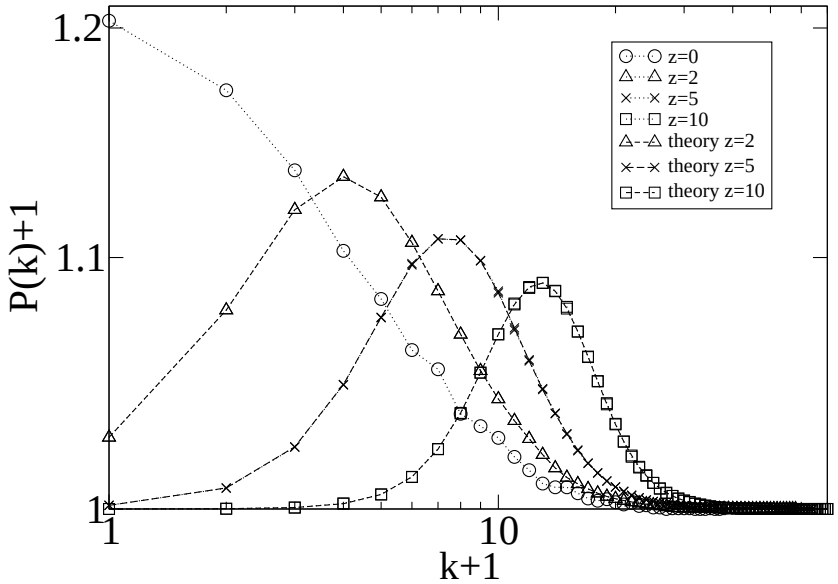
Poisson - Unbiased Bond Oversampling



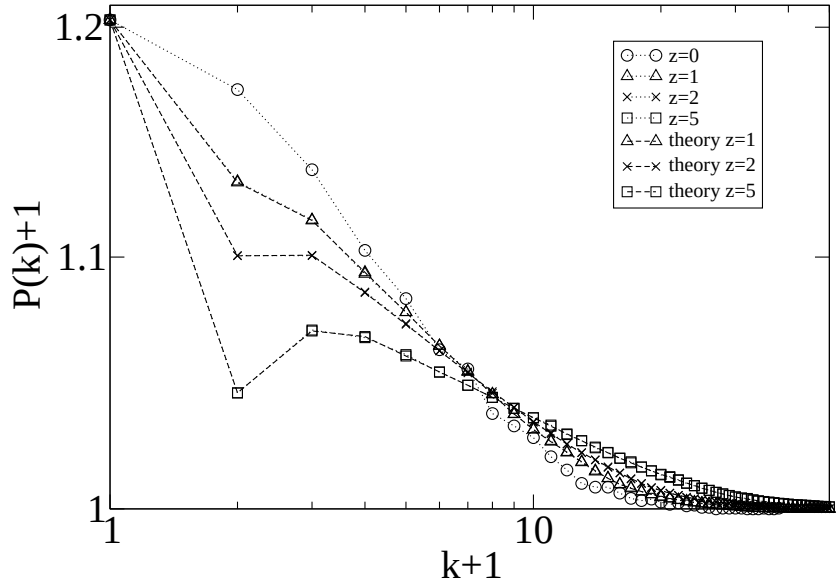

Poisson - Biased Bond oversampling




Power Law - Unbiased Bond Oversampling




Power Law - Biased Bond Oversampling

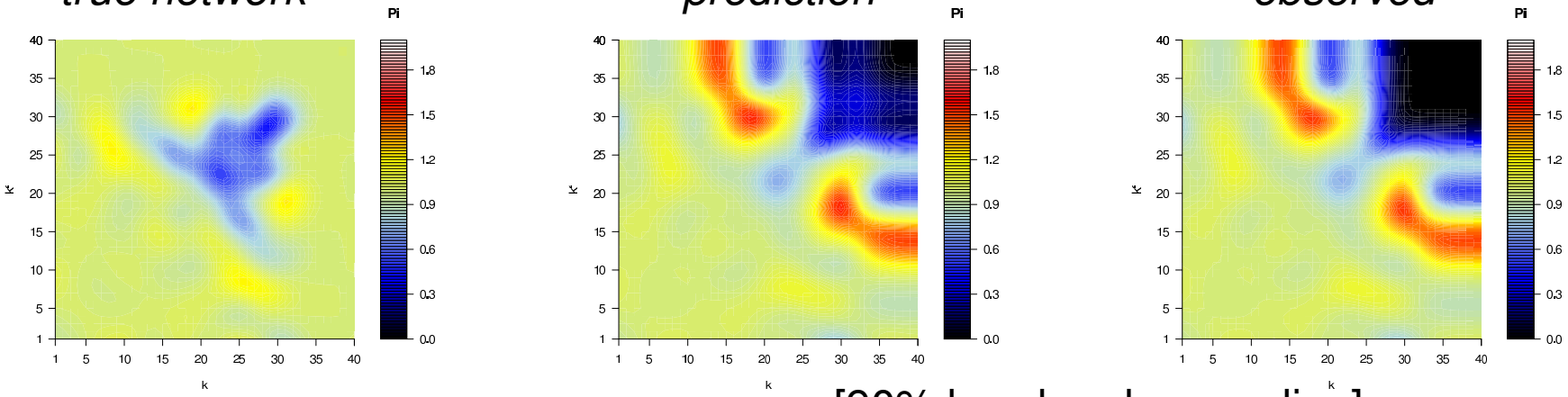


# unbiased undersampling

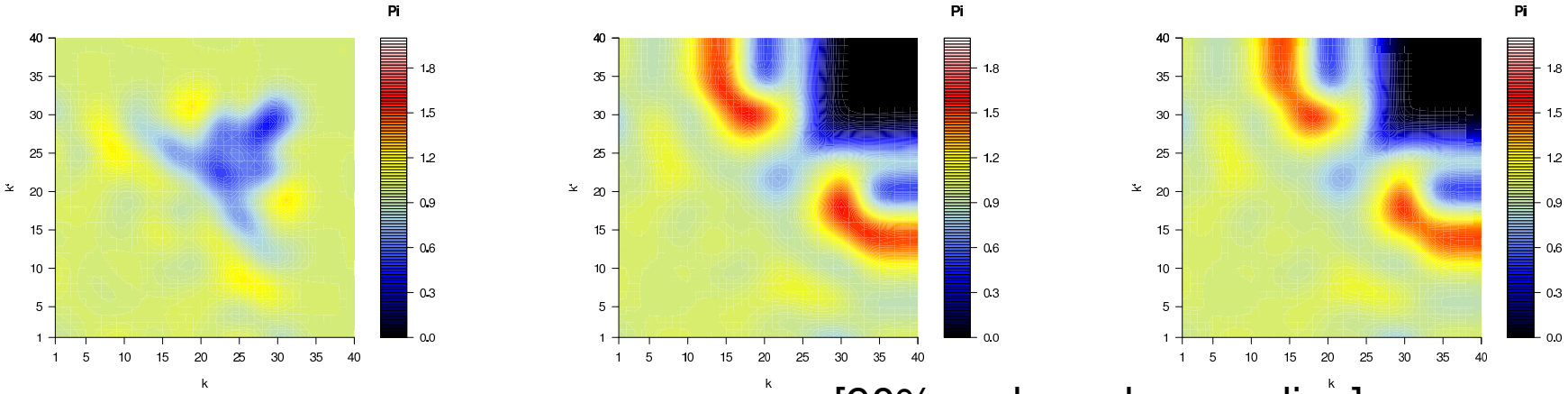
$N = 3512$, $\langle k \rangle = 3.7$

$\Pi(k, k') = W(k, k')/W(k)W(k')$
(power law network, averaged over $10^4$ samples)

*true network* *prediction* *observed*




[90% bond undersampling]



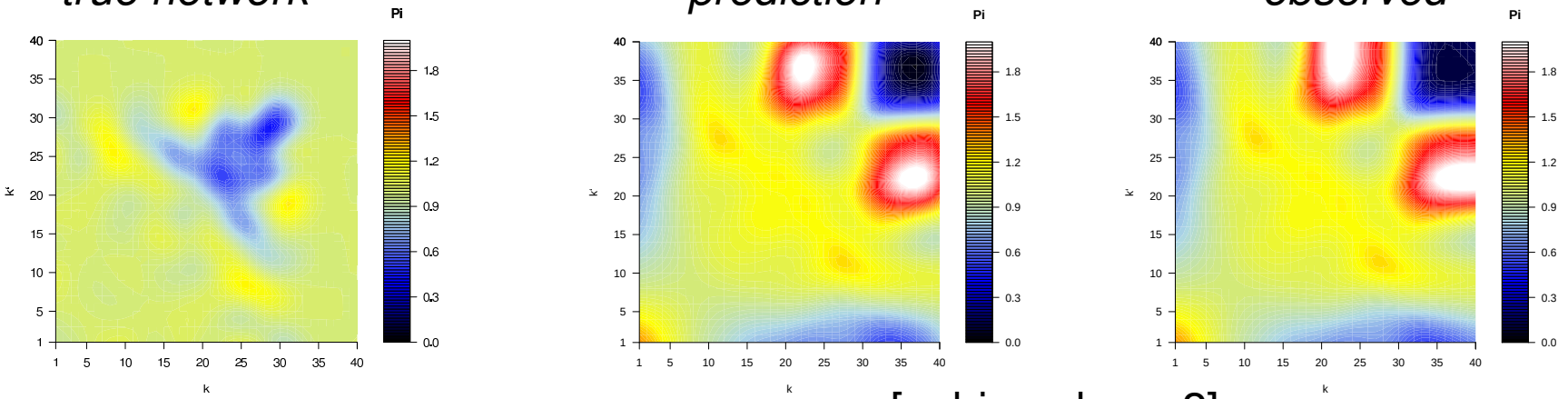

[90% node undersampling]

# oversampling

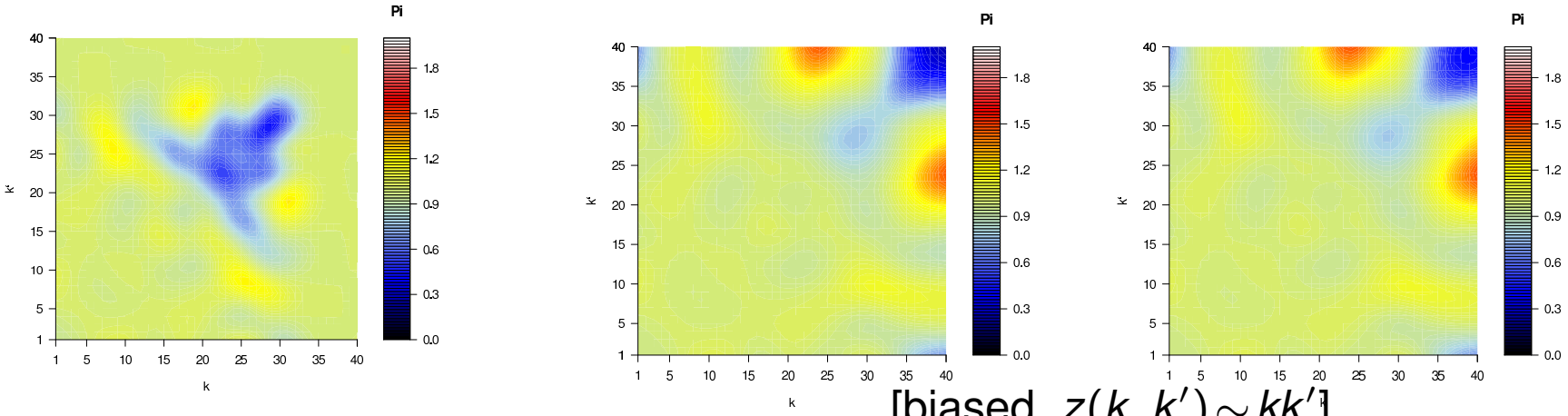
$N = 3512$, $\langle k \rangle = 3.7$

$\Pi(k, k') = W(k, k')/W(k)W(k')$
(power law network, averaged over $10^4$ samples)

*true network* *prediction* *observed*




[unbiased, $z = 2$]



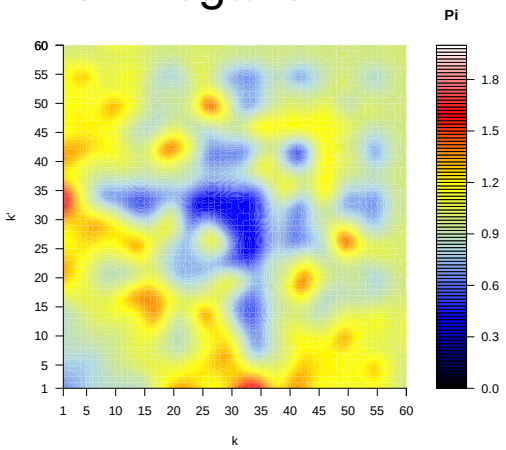

[biased, $z(k, k') \sim kk'$]

# protein interaction network *(C. Elegans)*

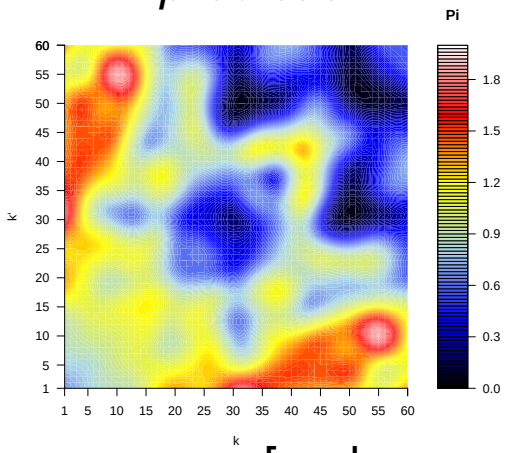
unbiased link sampling, $N = 3512$, $\langle k \rangle = 3.7$

$\Pi(k, k') = W(k, k')/W(k)W(k')$
(averaged over $10^4$ samples)

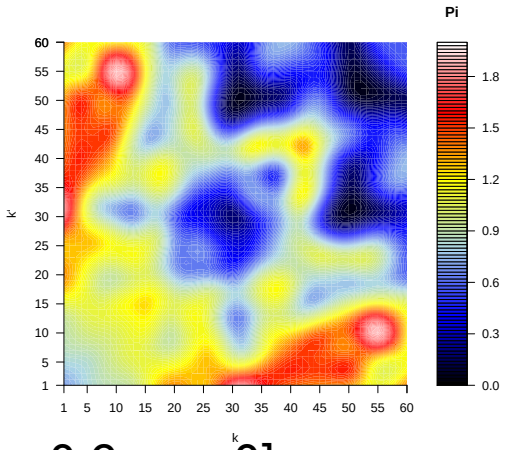
C. Elegans *PIN*



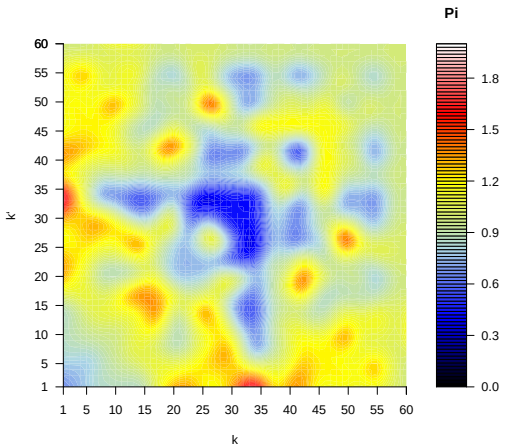

*prediction*

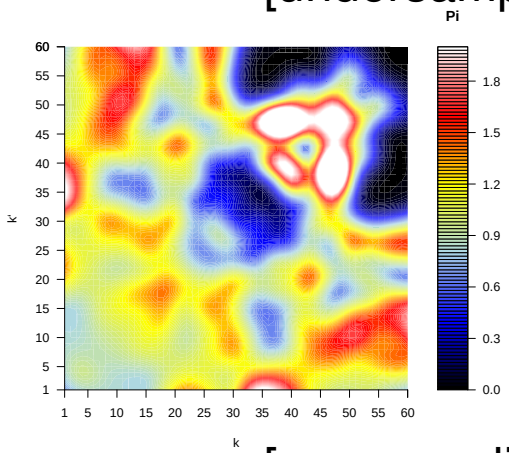


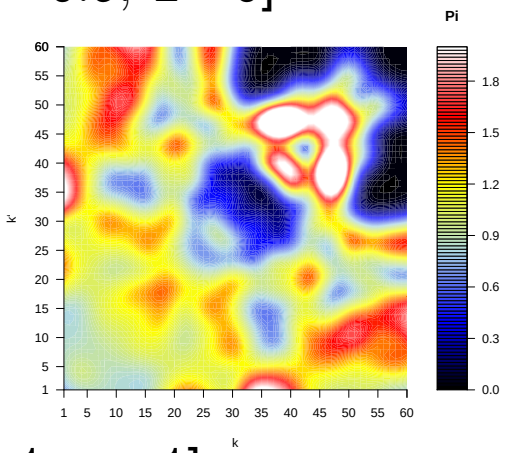

*observed*




[undersampling: $y = 0.9,\ z = 0$]




[oversampling: $y = 1,\ z = 1$]

# Summary

explicit formulae for $p'(k)$ and $W'(k,k')$ in terms of $p(k)$ $W(k,k')$,
for biased/unbiased node/link under/oversampling
(exact for $N \to \infty$, precise already for $N \sim 10^3$)

sampling generally affects shapes of $p(k)$ and $W(k,k')$,
except for trivial (i.e. uncorrelated Poissonian) networks

- unbiased undersampling:
  - node/link undersampling are equivalent
  - uncorrelated networks remain uncorrelated
- biased undersampling:
  - node/link undersampling not equivalent
  - uncorrelated networks *appear correlated*
- link oversampling:
  - always generates degree correlations

# Ongoing work: decontamination of PIN data

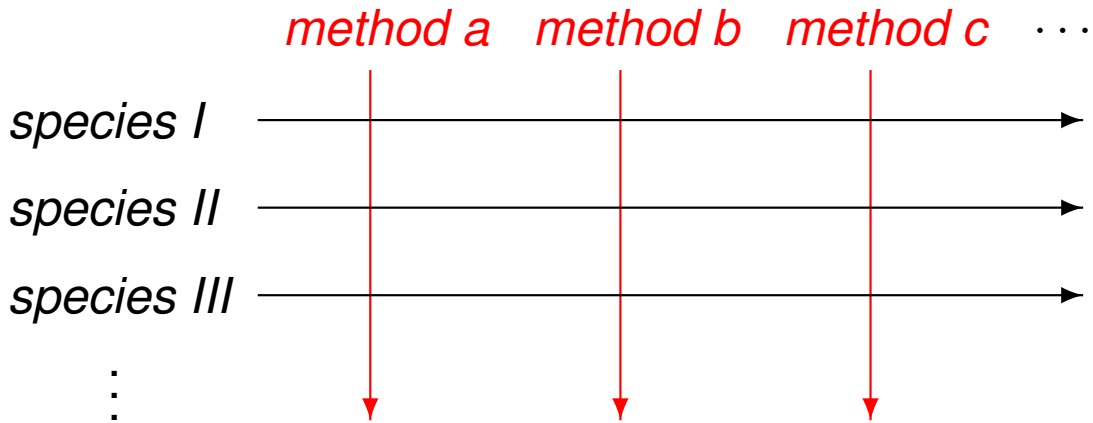
**available PIN data**:

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

– 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 $\mathbf{c}^{\ell,\alpha}$:

$$c_{ij}^{\ell,\alpha} = \sigma_i^{\ell,\alpha} \sigma_j^{\ell,\alpha} [\tau_{ij}^{\ell,\alpha} c_{ij}^{\ell} + (1 - c_{ij}^{\ell}) \lambda_{ij}^{\ell,\alpha}]$$




**objective**:

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