Network Inference

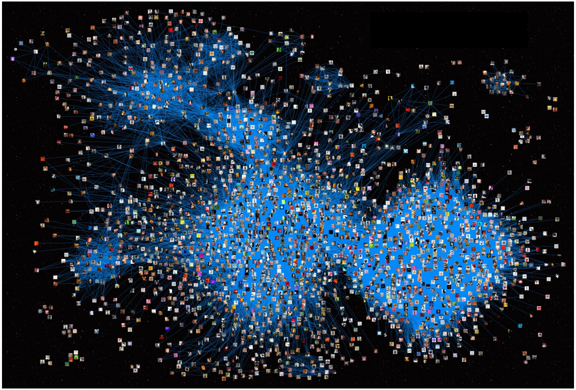
Part 2

Tiago P. Peixoto

University of Bath

Tehran, August 2018

Network measurements are noisy



(A social network)

- As with any empirical measurement, network data are unreliable.
- However, very few datasets contain any kind of error estimate!
- We know there must be errors, but we do not know how many, or where they are located.

Network reconstruction task

$$\underbrace{\boldsymbol{A}^*}_{\text{True network}} \longrightarrow \boxed{\text{Noisy measurement}} \longrightarrow \underbrace{\boldsymbol{A}}_{\text{Observed network}}$$

$$\underbrace{\boldsymbol{A}}_{\text{Observed network}} \longrightarrow \boxed{\text{Reconstruction}} \longrightarrow \underbrace{\hat{\boldsymbol{A}}}_{\text{Reconstructed network}}$$

So that $\hat{\boldsymbol{A}}$ is as close as possible to $\boldsymbol{A}^*$.

Caveats:

- With a single copy of $\boldsymbol{A}$.
- Without knowing how strong the noise is (i.e. the number of missing or spurious edges).

How is reconstruction possible?

(a) (b)

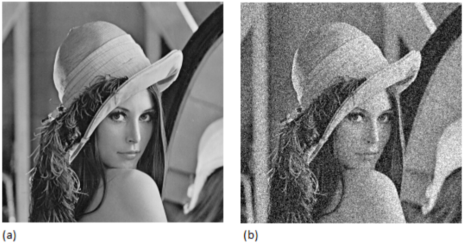
How is reconstruction possible?

(a) (b)

We need:

- A model for structure.
- A model for noise.

How is reconstruction possible?



We need:

- A model for structure.
- A model for noise.

(but for networks)

Nonparametric Bayesian inference

- A model for structure, $P(\boldsymbol{A}|\theta)$
- A model for noise, $P(\boldsymbol{\mathcal{D}}|\boldsymbol{A},\phi)$

$\boldsymbol{A} \to$ Network, $\boldsymbol{\mathcal{D}} \to$ Measured data, $(\theta,\phi) \to$ Parameters

Posterior distribution:

$$P(\boldsymbol{A}|\boldsymbol{\mathcal{D}}) = \frac{P(\boldsymbol{\mathcal{D}}|\boldsymbol{A})P(\boldsymbol{A})}{P(\boldsymbol{\mathcal{D}})}$$

Marginal probabilities:

$$P(\boldsymbol{\mathcal{D}}|\boldsymbol{A}) = \int P(\boldsymbol{\mathcal{D}}|\boldsymbol{A},\phi)P(\phi)\mathrm{d}\phi$$

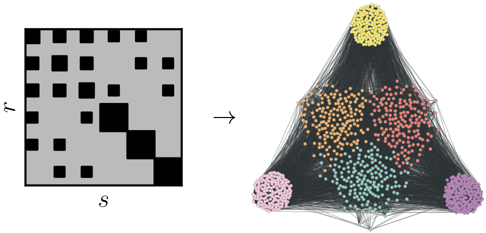
$$P(\boldsymbol{A}) = \int P(\boldsymbol{A}|\theta)P(\theta)\mathrm{d}\theta$$

Structure: The stochastic block model (SBM)

Planted partition: N nodes divided into B groups.

Parameters: $b_i \to$ group membership of node i

$\lambda_{rs} \to$ edge probability from group r to s.



Degree-corrected: Arbitrary degree sequence: $\{\kappa_i\}$

- Not restricted to assortative structures ("communities").
- Easily generalizable (edge direction, overlapping groups, etc.)

Bayesian SBM

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{b}) = \prod_{i<j}\frac{(\kappa_i\kappa_j\lambda_{b_ib_j})^{A_{ij}}e^{-\kappa_i\kappa_j\lambda_{b_ib_j}}}{A_{ij}!} \times \prod_i \frac{(\kappa_i^2\lambda_{b_ib_i}/2)^{A_{ii}/2}e^{-\kappa_i^2\lambda_{b_ib_i}/2}}{(A_{ii}/2)!}$$

Noninformative priors:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r\le s} e^{-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}}/(1+\delta_{rs})\bar{\lambda}$$

$$P(\boldsymbol{\kappa}|\boldsymbol{b}) = \prod_r (n_r-1)!\,\delta(\textstyle\sum_i \kappa_i\delta_{b_i,r}-1)$$

Bayesian SBM

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{b}) = \prod_{i<j}\frac{(\kappa_i\kappa_j\lambda_{b_ib_j})^{A_{ij}}e^{-\kappa_i\kappa_j\lambda_{b_ib_j}}}{A_{ij}!}\times\prod_i\frac{(\kappa_i^2\lambda_{b_ib_i}/2)^{A_{ii}/2}e^{-\kappa_i^2\lambda_{b_ib_i}/2}}{(A_{ii}/2)!}$$

Noninformative priors:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r\le s} e^{-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}}/(1+\delta_{rs})\bar{\lambda}$$

$$P(\boldsymbol{\kappa}|\boldsymbol{b}) = \prod_r (n_r-1)!\delta(\textstyle\sum_i \kappa_i\delta_{b_i,r}-1)$$

Marginal likelihood:

$$\begin{aligned}P(\boldsymbol{A}|\boldsymbol{b}) &= \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})P(\boldsymbol{\kappa}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\lambda}\mathrm{d}\boldsymbol{\kappa}\\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}}\times\frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!}\times\prod_r\frac{(n_r-1)!}{(e_r+n_r-1)!}\times\prod_i k_i!\\ &= P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e})\end{aligned}$$

Bayesian SBM

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{b}) = \prod_{i<j}\frac{(\kappa_i\kappa_j\lambda_{b_ib_j})^{A_{ij}}e^{-\kappa_i\kappa_j\lambda_{b_ib_j}}}{A_{ij}!}\times\prod_i\frac{(\kappa_i^2\lambda_{b_ib_i}/2)^{A_{ii}/2}e^{-\kappa_i^2\lambda_{b_ib_i}/2}}{(A_{ii}/2)!}$$

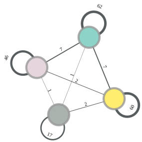
Noninformative priors:

$$P(\boldsymbol{\lambda}|\boldsymbol{b}) = \prod_{r\le s} e^{-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}}/(1+\delta_{rs})\bar{\lambda}$$

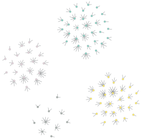
$$P(\boldsymbol{\kappa}|\boldsymbol{b}) = \prod_r (n_r-1)!\,\delta(\textstyle\sum_i\kappa_i\delta_{b_i,r}-1)$$

Marginal likelihood:

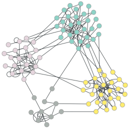
$$\begin{aligned} P(\boldsymbol{A}|\boldsymbol{b}) &= \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})P(\boldsymbol{\kappa}|\boldsymbol{b})\,\mathrm{d}\boldsymbol{\lambda}\mathrm{d}\boldsymbol{\kappa} \\ &= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}}\times\frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_{i<j}A_{ij}!\prod_i A_{ii}!!}\times\prod_r\frac{(n_r-1)!}{(e_r+n_r-1)!}\times\prod_i k_i! \\ &= P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}) \end{aligned}$$



Edge counts $\boldsymbol{e}$.

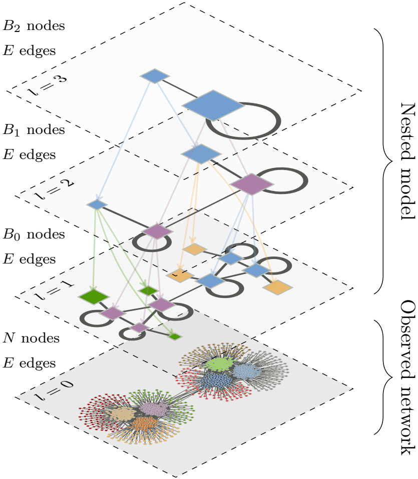


Degrees, $\boldsymbol{k}$.



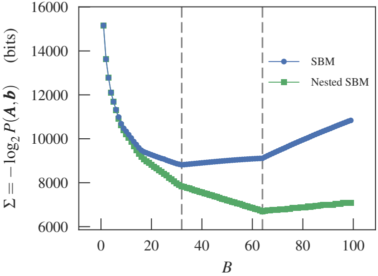
Network, $\boldsymbol{A}$.

Nested SBM: Group hierarchies



Deeper Bayesian hierarchy:

- Prevents underfitting.
- Multiple scales of description.



Measurement model

Edge-o-meter

$p \to$ probability of a missing edge $(1 \to 0)$
$q \to$ probability of a spurious edge $(0 \to 1)$

$n_{ij} \to$ number of measurements of pair (i, j)
$x_{ij} \to$ number of edges recorded

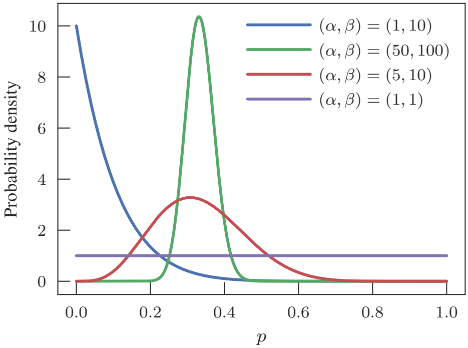
$$P(x_{ij}|n_{ij}, A_{ij}, p, q) = \binom{n_{ij}}{x_{ij}} \left[(1-p)^{x_{ij}} p^{n_{ij}-x_{ij}}\right]^{A_{ij}} \left[q^{x_{ij}} (1-q)^{n_{ij}-x_{ij}}\right]^{1-A_{ij}}$$

$$P(\boldsymbol{x}|\boldsymbol{n}, \boldsymbol{A}, \alpha, \beta, \mu, \nu) = \int P(\boldsymbol{x}|\boldsymbol{n}, \boldsymbol{A}, p, q) P(p|\alpha, \beta) P(q|\mu, \nu) \, \mathrm{d}p \, \mathrm{d}q$$

$P(p|\alpha, \beta), P(q|\mu, \nu) \to$ Beta priors

The Edge-o-meter

$$P(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\alpha-1}}{\mathcal{B}(\alpha,\beta)} \qquad P(q|\mu,\nu) = \frac{q^{\mu-1}(1-q)^{\mu-1}}{\mathcal{B}(\mu,\nu)}$$



- $(\alpha,\beta) = (1,10) \rightarrow$ accurate measurement (low noise)
- $(\alpha,\beta) = (50,100) \rightarrow$ high noise, good calibration
- $(\alpha,\beta) = (5,10) \rightarrow$ high noise, bad calibration
- $(\alpha,\beta) = (1,1) \rightarrow$ <u>noninformative</u> (i.e. uniform distribution)

The full reconstruction method

Posterior distribution:

$$P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{n},\boldsymbol{x},\alpha,\beta,\mu,\nu) = \frac{P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},\alpha,\beta,\mu,\nu)P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{x}|\alpha,\beta,\mu,\nu)}.$$

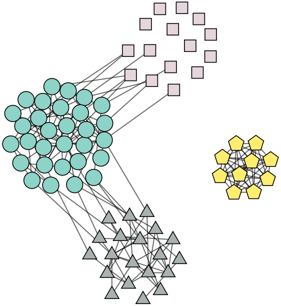
We infer both the network $\boldsymbol{A}$ as well as the SBM latent variables $\boldsymbol{b}$ via MCMC:

Move proposals $P(\boldsymbol{b}'|\boldsymbol{A},\boldsymbol{b})$ and $P(\boldsymbol{A}'|\boldsymbol{A},\boldsymbol{b})$, accept with probability

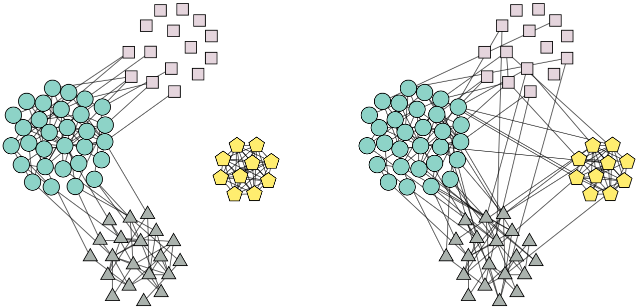
$$\min\left(1, \frac{P(\boldsymbol{A}',\boldsymbol{b}'|\boldsymbol{\mathcal{D}})P(\boldsymbol{A}|\boldsymbol{A}',\boldsymbol{b}')P(\boldsymbol{b}|\boldsymbol{A}',\boldsymbol{b}')}{P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{\mathcal{D}})P(\boldsymbol{A}'|\boldsymbol{A},\boldsymbol{b})P(\boldsymbol{b}'|\boldsymbol{A},\boldsymbol{b})}\right).$$

(Efficient, scales to very large networks.)

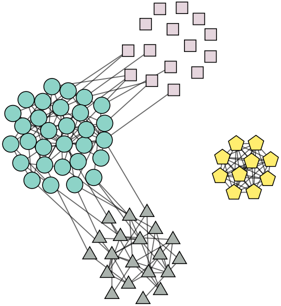
How does it work?

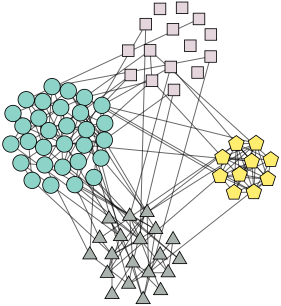


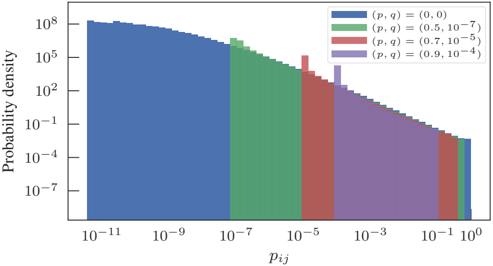
How does it work?



How does it work?

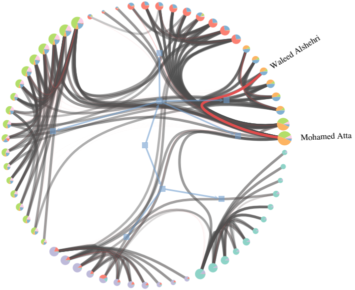


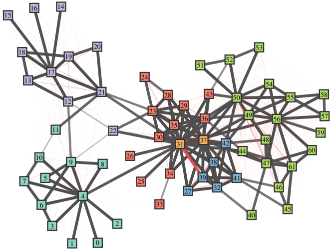


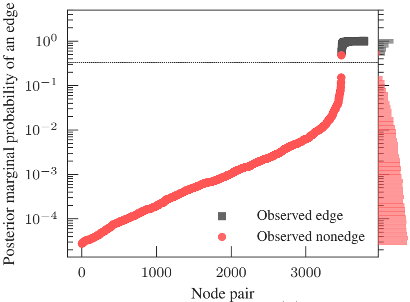


$$p'_{ij} = (1 - p - q)p_{ij} + q$$

Example: Terrorist associations







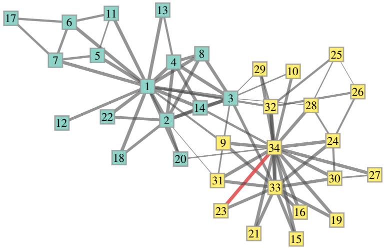
V. Krebs, "Mapping networks of terrorist cells", Connections 24 (3), 43-52

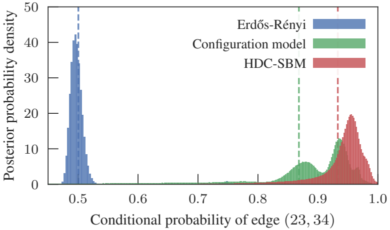
Example: Zachary's karate club

Individual Number

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
3	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0
4	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
17	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
20	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
22	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1
31	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
32	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1
33	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	1	1	1	0	1
34	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	0

(credit: Aaron Clauset)





W. W. Zachary, J. Anthro. Research 33(4), 452-473 (1977)

Wait! Is this just edge prediction?

It *is* edge prediction, but it yields a full posterior distribution $P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x})$ that is **nonparametric**.

We can:

- Perform maximum marginal posterior estimation,

$$\hat{A}_{ij} = \begin{cases} 1 & \text{if } \pi_{ij} > 1/2 \\ 0 & \text{if } \pi_{ij} < 1/2, \end{cases}$$

  where $\pi_{ij} = \sum_{\boldsymbol{A}} A_{ij} P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x})$ is the marginal posterior edge probability.

- Estimate network properties $y(\boldsymbol{A})$ and their error estimates:

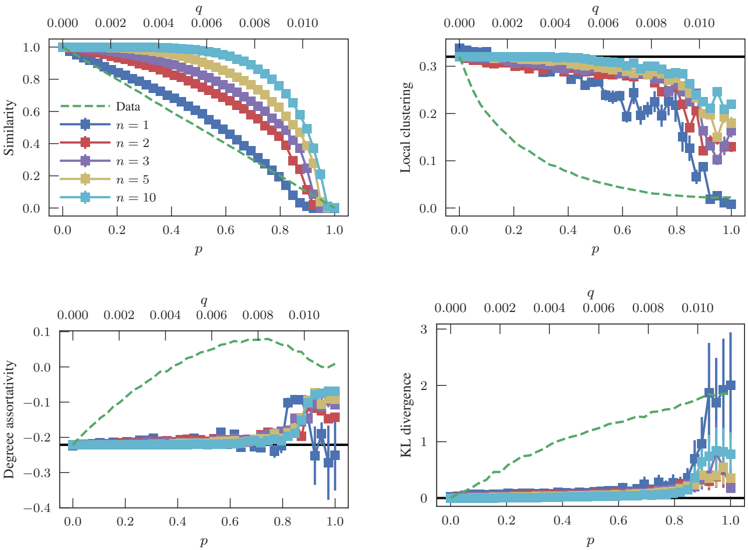
$$\hat{y} = \sum_{\boldsymbol{A}} y(\boldsymbol{A}) P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x})$$

$$\sigma_y^2 = \sum_{\boldsymbol{A}} (\hat{y} - y(\boldsymbol{A}))^2 P(\boldsymbol{A}|\boldsymbol{n},\boldsymbol{x}).$$

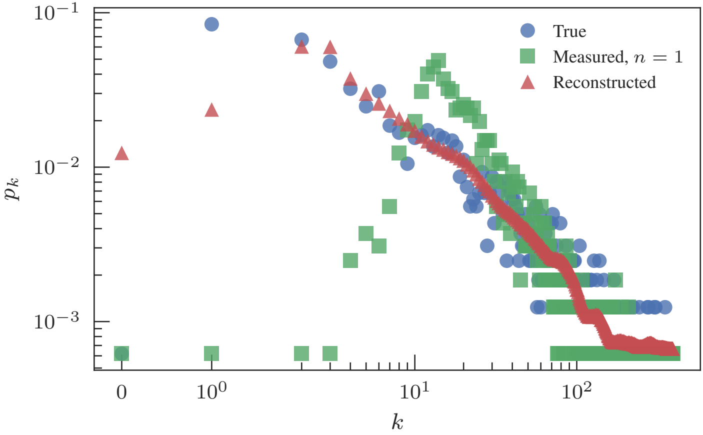
Reconstruction performance

Real network (political blogs) + simulated noise:

$$p \in [0,1],\ q = pE/[\binom{N}{2} - E]$$

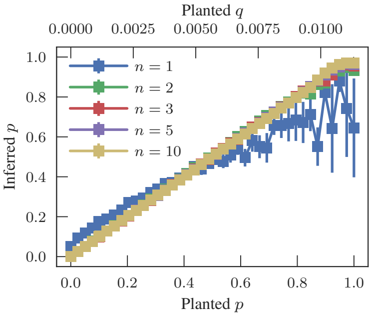


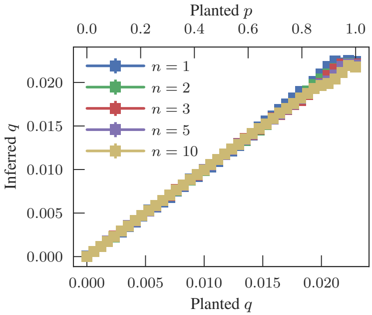
Reconstruction performance: Degrees



$$(p, q) = (0.41, 0.0094)$$

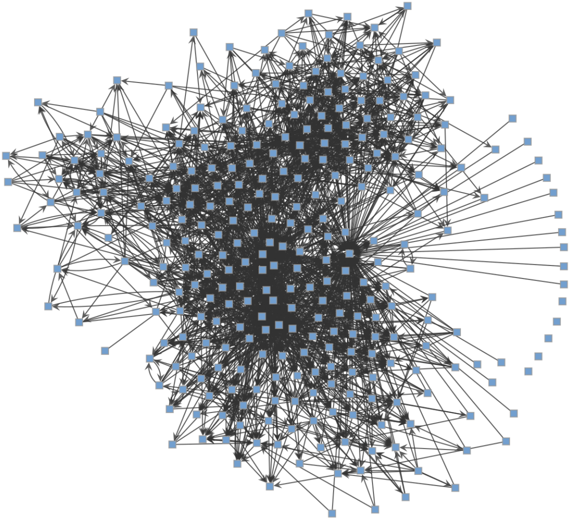
Inferring the noise





Empirical reconstruction redux

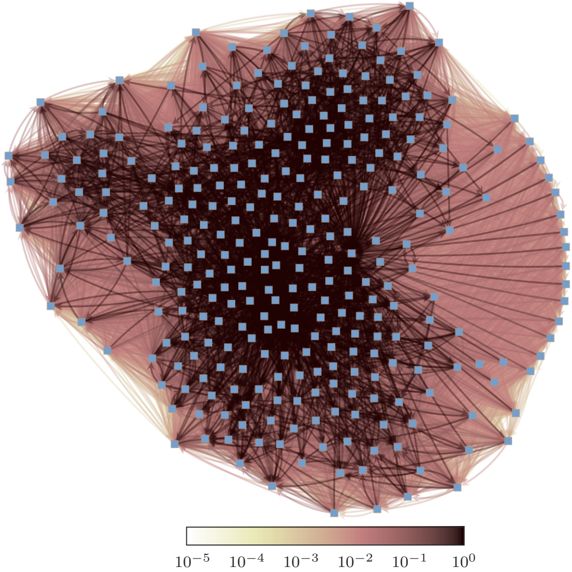
C. elegans neural network



White et al., Phil. Trans. R. Soc. London 314, 1 (1986).

Empirical reconstruction redux

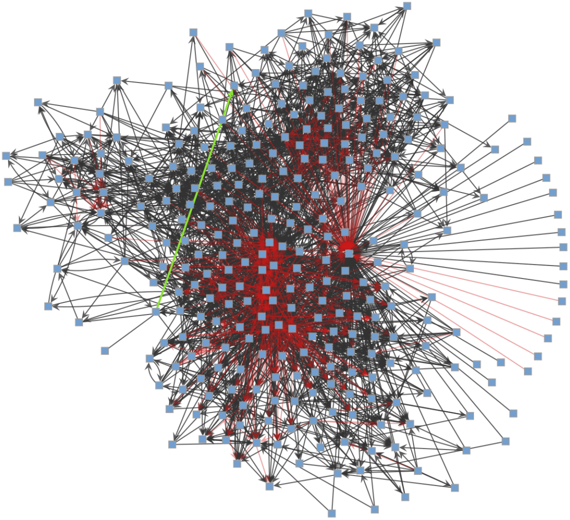
C. elegans neural network



White et al., Phil. Trans. R. Soc. London 314, 1 (1986).

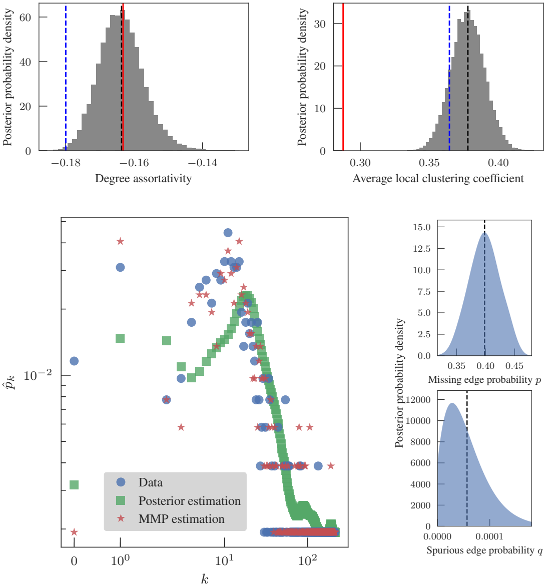
Empirical reconstruction redux

C. elegans neural network

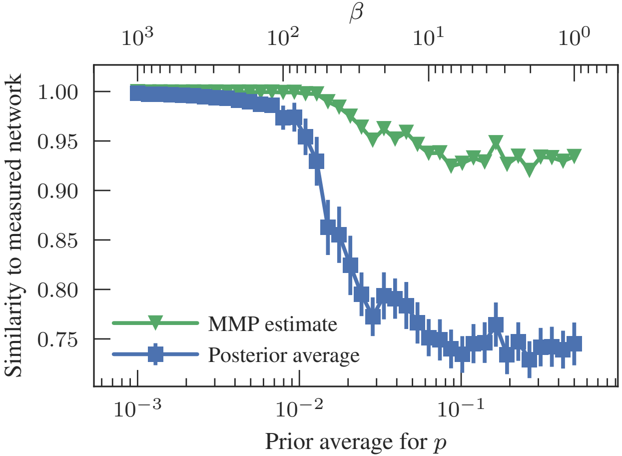


White et al., Phil. Trans. R. Soc. London 314, 1 (1986).

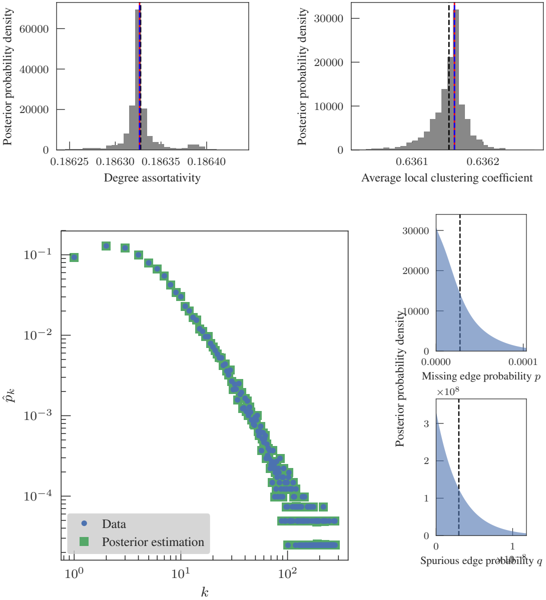
C. elegans NEURAL NETWORK



C. elegans NEURAL NETWORK



ARXIV.ORG CO-AUTHORSHIP NETWORK



Uncertainty assessment of empirical data

Dataset	Similarity	Nodes	Edges		Degree assortativity		Local clustering		B_e	$\hat{p}$	$\hat{q}$
			Direct	Estimated	Direct	Estimated	Direct	Estimated			
karate	0.94(4)	34	78	77(7)	−0.475 61	−0.49(5)	0.570 64	0.58(5)	2.7(6)	0.06(5)	0.012(10)
terrorists	0.96(2)	62	152	154(8)	−0.080 48	−0.096(20)	0.486 37	0.50(2)	5.4(5)	0.05(4)	0.003(2)
football	0.857(16)	115	613	500(18)	0.162 44	0.18(7)	0.403 22	0.68(4)	12.7(3)	0.05(3)	0.0226(19)
netscience	0.9981(17)	379	914	915(3)	−0.081 68	−0.0823(18)	0.741 23	0.741(3)	29.6(14)	0.004(3)	$3.1(19) \times 10^{-5}$
celegans	0.754(20)	302	2345	3850(150)	−0.163 20	−0.165(7)	0.287 52	0.374(12)	17.25(19)	0.39(3)	$6(3) \times 10^{-5}$
malaria	0.9981(15)	1103	2965	2973(9)	−0.300 13	−0.2997(20)	0	0(0)	30.8(3)	0.004(3)	$4(3) \times 10^{-6}$
power	0.80(7)	4941	6594	9900(1300)	0.003 46	0.043(17)	0.080 10	0.058(7)	15.6(7)	0.33(10)	$2.5(19) \times 10^{-7}$
polblogs	0.965(5)	1222	16 714	17 860(190)	−0.221 33	−0.2226(16)	0.320 25	0.343(5)	16.6(3)	0.066(10)	$4.4(17) \times 10^{-5}$
dlbp	0.64(1)	12 590	49 744	106 000(2000)	−0.045 72	−0.0559(19)	0.117 18	0.164(7)	86.4(20)	0.529(11)	$9(5) \times 10^{-9}$
openflights	0.9916(9)	3286	39 430	40 100(70)	−0.005 31	−0.0071(11)	0.496 47	0.507(2)	117.1(5)	0.0167(18)	$1.0(3) \times 10^{-7}$
reactome	0.999 977(10)	6327	146 160	146 164(3)	0.244 87	0.244 87(4)	0.588 38	0.5887(3)	318.7(10)	$4.1(18) \times 10^{-5}$	$1.3(8) \times 10^{-7}$
cond-mat	0.999 986(13)	40 421	175 693	175 695(4)	0.186 33	0.186 33(2)	0.636 16	0.636 15(3)	1014(6)	$3(2) \times 10^{-5}$	$3(2) \times 10^{-9}$
Enron	0.999 86(5)	36 692	183 831	183 885(18)	−0.110 76	−0.110 75(2)	0.496 98	0.496 92(8)	188.9(11)	0.000 28(10)	$2.9(19) \times 10^{-9}$
linux	0.9973(3)	30 837	213 424	214 600(120)	−0.174 68	−0.174 67(7)	0.128 49	0.1322(10)	351.2(7)	0.0055(5)	$1.7(10) \times 10^{-9}$
brightkite	0.9985(3)	58 228	214 078	214 740(80)	0.010 82	0.011 00(11)	0.172 33	0.172 34(10)	151(3)	0.0029(5)	$1.7(12) \times 10^{-8}$
pgp	0.997 99(9)	39 796	301 498	301 660(60)	0.000 76	0.000 49(8)	0.461 09	0.4617(2)	929(2)	0.002 27(16)	$3.35(18) \times 10^{-7}$
caida	0.999 67(13)	53 387	496 731	497 070(130)	−0.186 97	−0.186 959(17)	0.680 97	0.681 26(14)	218.0(16)	0.0007(3)	$1.0(8) \times 10^{-9}$
web-Stanford	0.999 998 7(8)	281 903	2 312 497	2 312 494(4)	−0.112 44	−0.112 444 7(2)	0.597 63	0.597 634(3)	4168(2)	$1.0(2) \times 10^{-6}$	$7(5) \times 10^{-11}$
flickr	0.999 976(13)	105 938	2 316 948	2 316 830(60)	0.246 85	0.246 823(16)	0.089 13	0.089 138(7)	617(2)	$6(3) \times 10^{-7}$	$2.0(11) \times 10^{-8}$

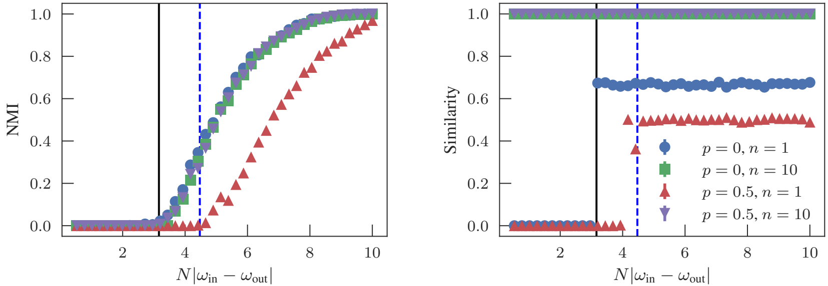
Noise and detectability of communities

Planted partition model:

$$\omega_{rs} = \omega_{\text{in}}\delta_{rs} + \omega_{\text{out}}(1-\delta_{rs})$$

Single observation, $n = 1$, effectively:

$$\omega'_{rs} = (1-p-q)\omega_{rs} + q$$



$$N|\omega_{\text{in}} - \omega_{\text{out}}| < B\sqrt{\langle k\rangle}, \qquad N|\omega_{\text{in}} - \omega_{\text{out}}| < \frac{B\sqrt{(1-p-q)\langle k\rangle + qN}}{(1-p-q)}.$$

Multiple measurements and heterogeneous errors

Observational error does not need to be uniform for every pair (i, j).
Non-uniform model, w/ pair-specific error rates: p_{ij} and q_{ij}

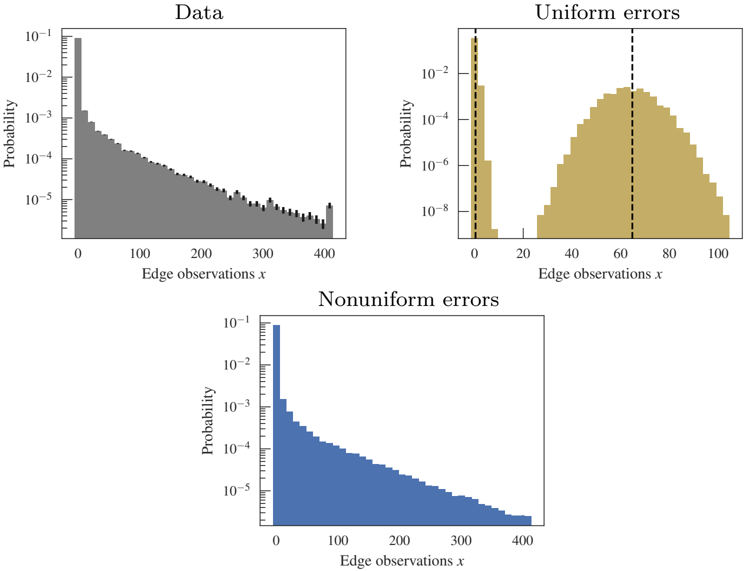
$$P(x_{ij}|n_{ij}, A_{ij}, p_{ij}, q_{ij}) = \binom{n_{ij}}{x_{ij}} \left[(1-p_{ij})^{x_{ij}} p_{ij}^{n_{ij}-x_{ij}}\right]^{A_{ij}} \left[q_{ij}^{x_{ij}}(1-q_{ij})^{n_{ij}-x_{ij}}\right]^{1-A_{ij}}$$

Marginal probability,

$$\begin{aligned} &P(x_{ij}|n_{ij}, A_{ij}, \alpha, \beta, \mu, \nu) \\ &\quad = \int P(x_{ij}|n_{ij}, A_{ij}, p_{ij}, q_{ij}) P(p_{ij}|\alpha, \beta) P(q_{ij}|\mu, \nu)\, \mathrm{d}p_{ij} \mathrm{d}q_{ij} \\ &\quad = \binom{n_{ij}}{x_{ij}} \left[\frac{\mathcal{B}(n_{ij}-x_{ij}+\alpha, x_{ij}+\beta)}{\mathcal{B}(\alpha, \beta)}\right]^{A_{ij}} \times \\ &\qquad \left[\frac{\mathcal{B}(x_{ij}+\mu, n_{ij}-x_{ij}+\nu)}{\mathcal{B}(\mu, \nu)}\right]^{1-A_{ij}}. \end{aligned}$$

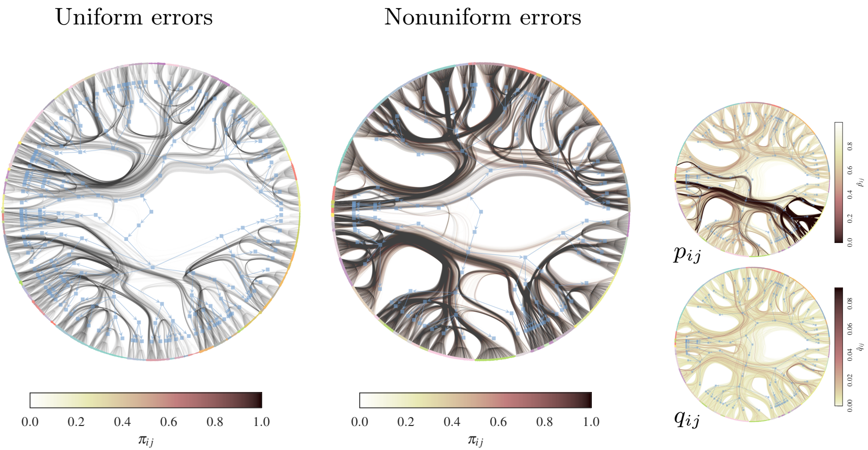
Human connectome

418 individuals



Human connectome

418 individuals



Extrinsic error estimates

$Q_{ij} \in [0, 1] \rightarrow$ experimentally determined uncertainties

$$P_Q(\boldsymbol{A}|\boldsymbol{Q}) = \prod_{i<j} Q_{ij}^{A_{ij}} (1 - Q_{ij})^{1-A_{ij}}.$$

Example:

STRING Protein-Protein interaction network database, Szklarczyk et al, Nucleic Acids Research 45, D362–D368 (2017).

Errors are estimated via a combination of: (i) direct experiments, (ii) database curation, (iii) publication text-mining, (iv) co-expression data, (v) genome proximity, (vi) ortholog fusion, (vii) phylogenetic co-ocurrence.

Extrinsic error estimates

The distribution $P_Q(\boldsymbol{A}|\boldsymbol{Q})$ implies the following noisy measurement process,

$$P(\boldsymbol{Q}|\boldsymbol{A}) = \frac{P_Q(\boldsymbol{A}|\boldsymbol{Q})P_Q(\boldsymbol{Q})}{P_Q(\boldsymbol{A})},$$

with prior

$$P_Q(\boldsymbol{Q}) = \prod_{i<j} P(Q_{ij}),$$

and normalization constant

$$P_Q(\boldsymbol{A}) = \int P_Q(\boldsymbol{A}|\boldsymbol{Q})P_Q(\boldsymbol{Q})\,\mathrm{d}\boldsymbol{Q} = \prod_{i<j} \bar{Q}^{A_{ij}}(1-\bar{Q})^{1-A_{ij}},$$

with $\bar{Q} = \int_0^1 QP(Q)\mathrm{d}Q$. Combining these together we have

$$P(\boldsymbol{Q}|\boldsymbol{A}) = P_Q(\boldsymbol{Q}) \prod_{i<j} \left(\frac{Q_{ij}}{\bar{Q}}\right)^{A_{ij}} \left(\frac{1-Q_{ij}}{1-\bar{Q}}\right)^{1-A_{ij}},$$

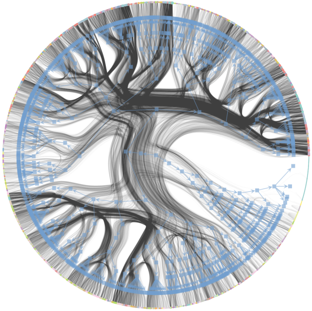
$$P(\boldsymbol{A}|\boldsymbol{Q}) = \frac{P(\boldsymbol{Q}|\boldsymbol{A})P(\boldsymbol{A})}{P(\boldsymbol{Q})}, \qquad \bar{Q} = \frac{\sum_{i<j} Q_{ij}}{\binom{N}{2}}.$$

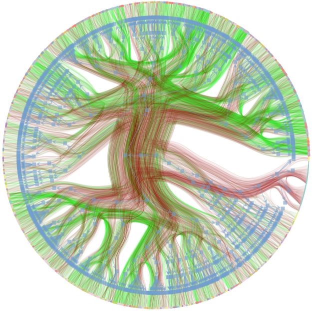
Extrinsic error estimates

$$P(\boldsymbol{A}|\boldsymbol{Q}) = \frac{P(\boldsymbol{Q}|\boldsymbol{A})P(\boldsymbol{A})}{P(\boldsymbol{Q})}, \qquad P(\boldsymbol{A}|\boldsymbol{Q}) \neq P_Q(\boldsymbol{A}|\boldsymbol{Q})!$$

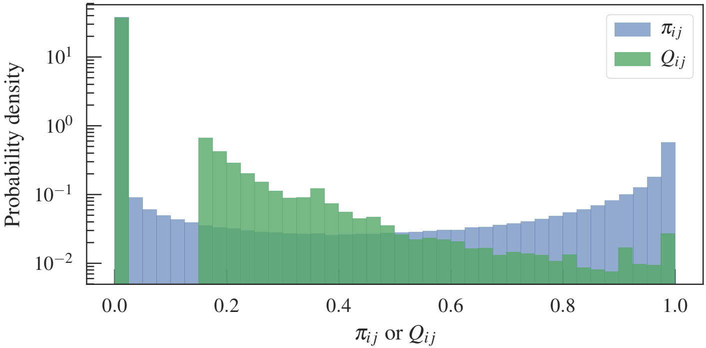
We are keeping the same noise generating process, but changing our prior assumption about the data.

E. coli proteins:





Extrinsic error estimates



For code, see:

`https://graph-tool.skewed.de`

(See also HOWTO at: `https://graph-tool.skewed.de/static/doc/demos/inference/inference.html`)