

Bayesian Connectomics

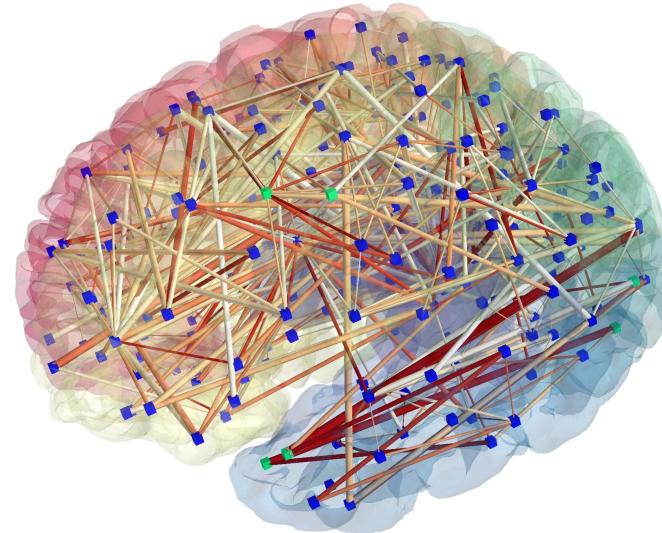
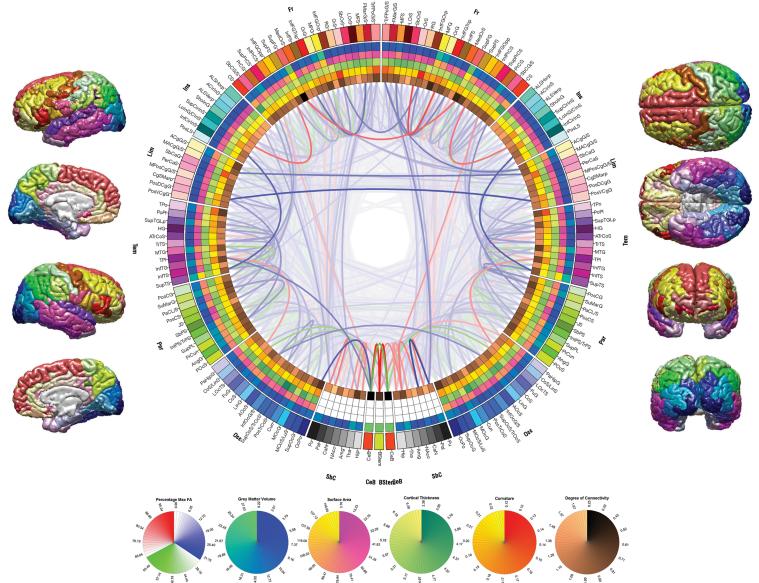
Max Hinne





Today

- The human brain shows both functional **segregation** and **integration**.
 - Function integration requires coordinated **interaction** between **distributed** neural populations.
 - Connectomics refers to the **mapping** and **analysis** of **brain connectivity** at different



www.connectomics.org/viewer



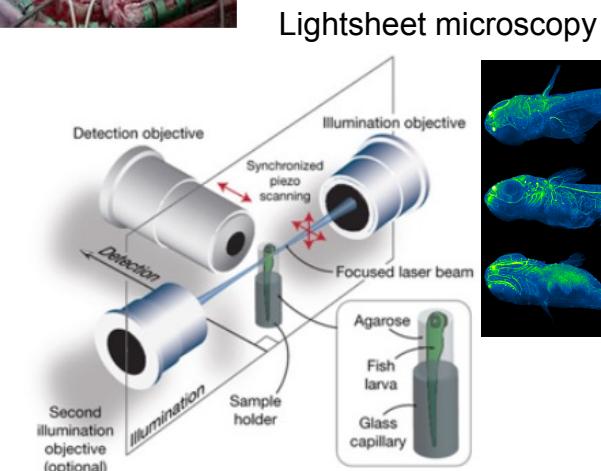
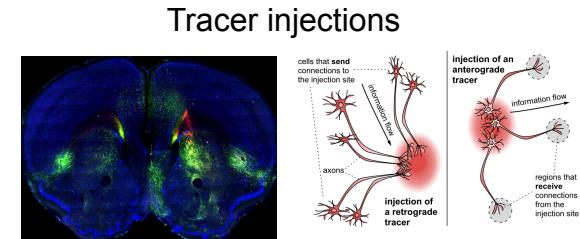
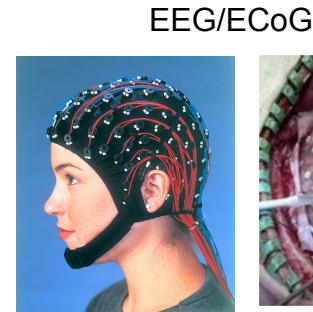
Flavours of connectomics

Connectivity has a structural, functional and effective component:

- **Structural connectivity:** the white-matter tracts that link neuronal populations
- **Functional connectivity:** the correlations between neuronal time series
- **Effective connectivity:** the causal interactions between neuronal populations

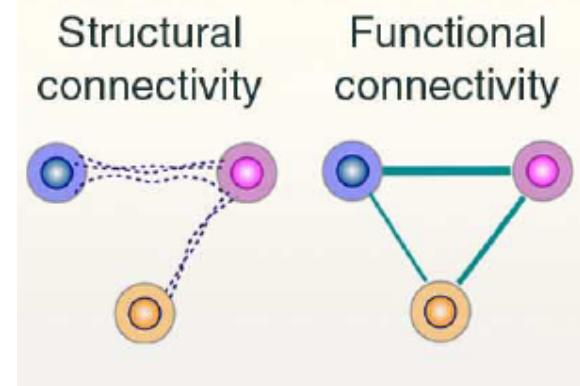
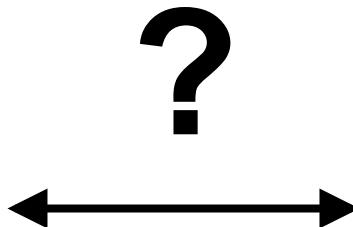
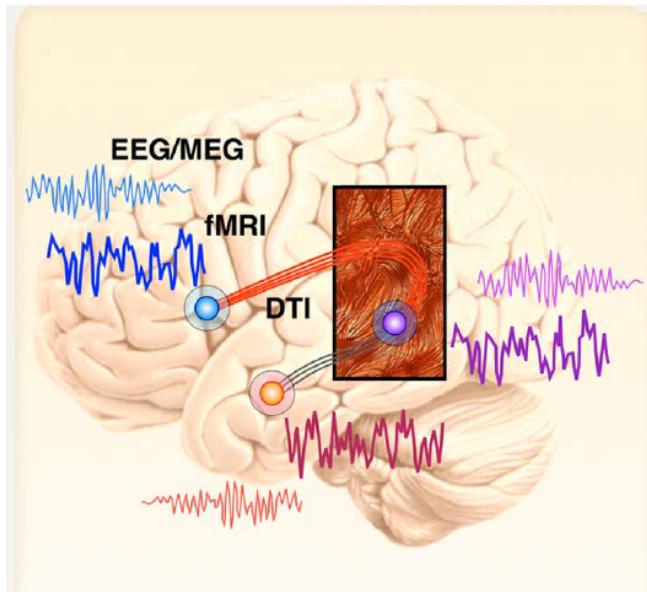
How can we measure connectivity?

Microscopic & macroscopic approaches





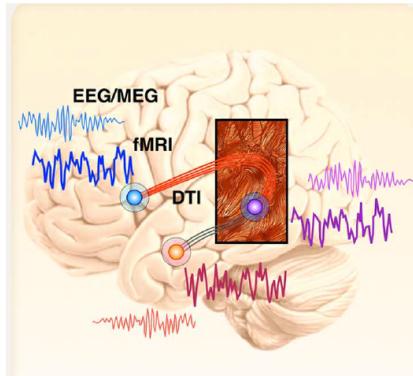
Connectivity analysis



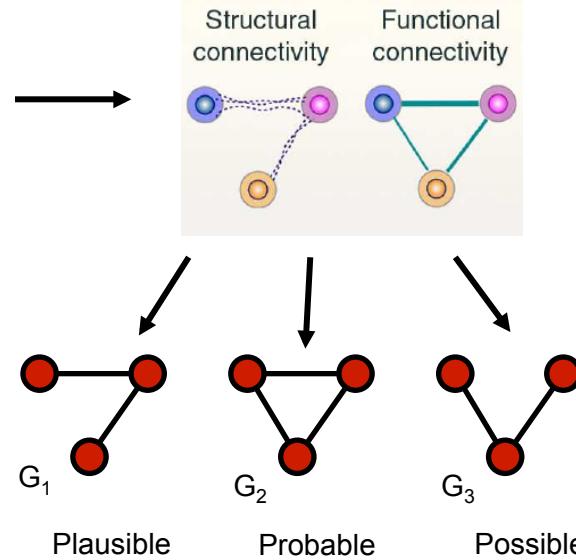
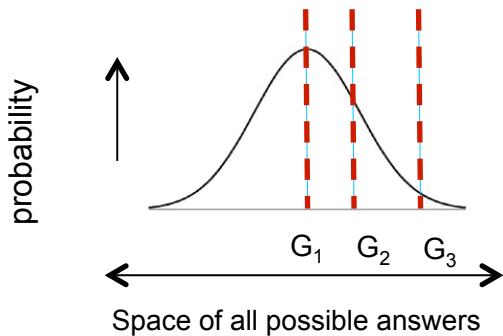
[Park & Friston, Science, 2013]



Connectivity analysis



[Park & Friston, Science, 2013]



How to express the certainty of our claims?



Connectivity analysis

- Connectivity analysis is inherently difficult due to the **noisy and incomplete** nature of the observations
- We advocate a Bayesian approach to connectomics: **Bayesian Connectomics** (BaCon for short)
 - Relies on defining **generative models** whose parameters are inferred from data, using various imaging modalities
 - Will be applied in the context of **macroscopic** connectomics:
 - Structural connectivity analysis using diffusion MRI or tract tracing
 - Functional connectivity analysis with functional MRI

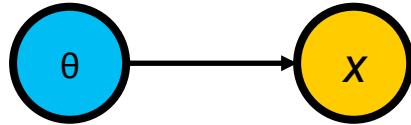
Generative models

Using Bayes to model data and assumptions



Generative modeling

- Which **random variables** depend on which other variables?
- Which **probability distribution** reflects this dependency?
- E.g.:

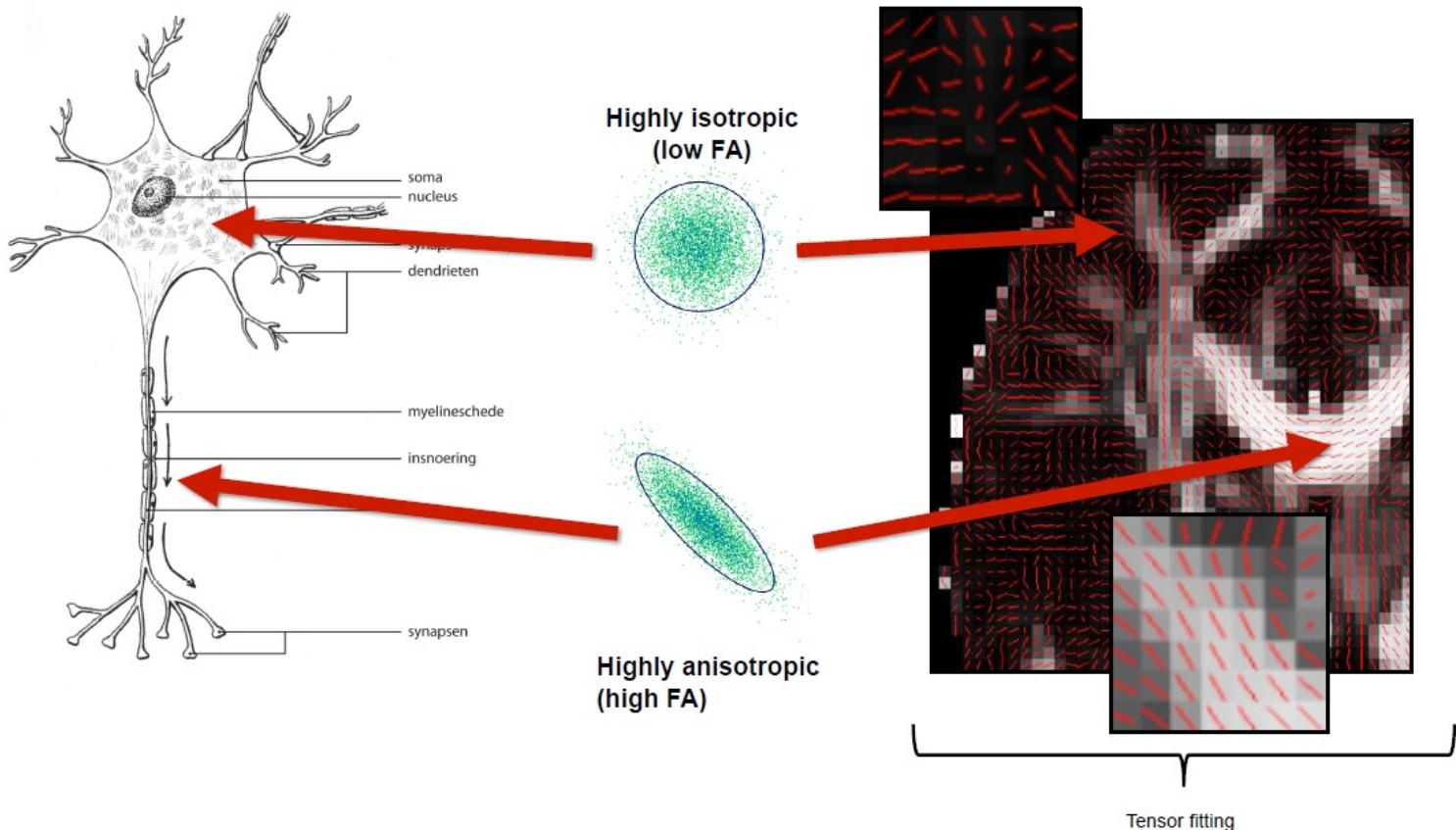


- $P(x|\theta)$ = Bernoulli, or $P(x|\theta)$ = Gaussian, etc., depending on what the variables represent.
- Use **Bayes' theorem** to compute $P(\theta|x) = P(x|\theta) P(\theta) / P(x)$ ← Not so easy in practice
- Modeling choices:
 - What is the **likelihood** $P(x|\theta)$, i.e. if we know the parameters, what process generates the data x ?
 - What is the **prior** $P(\theta)$, i.e. what parameters do we expect before we make observations?
- Challenge: compute posterior in reasonable time.

Bayesian structural connectivity analysis

What is the anatomical connectivity between brain
regions?

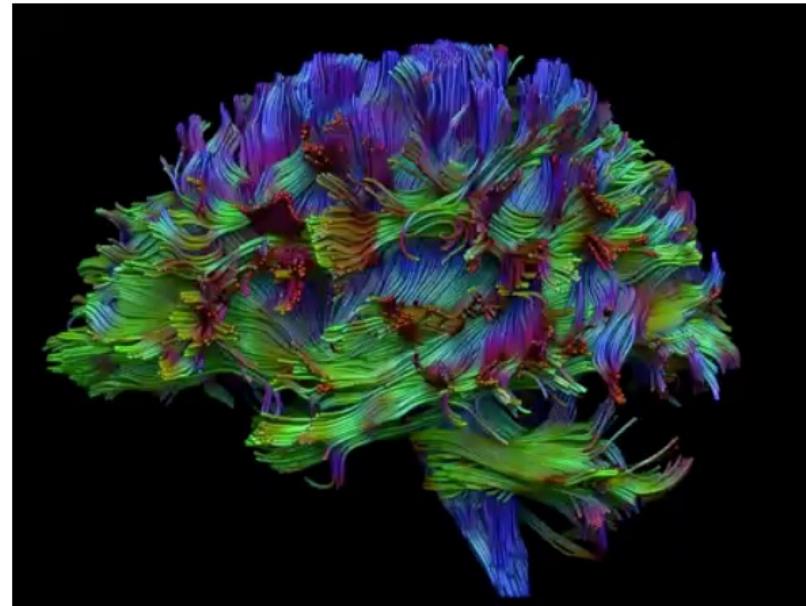
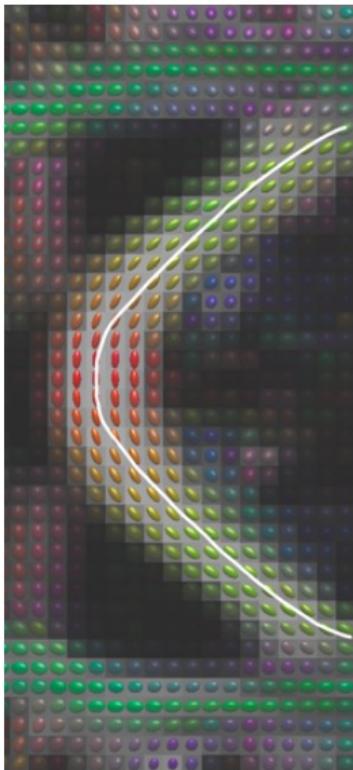
Diffusion imaging





Tractography

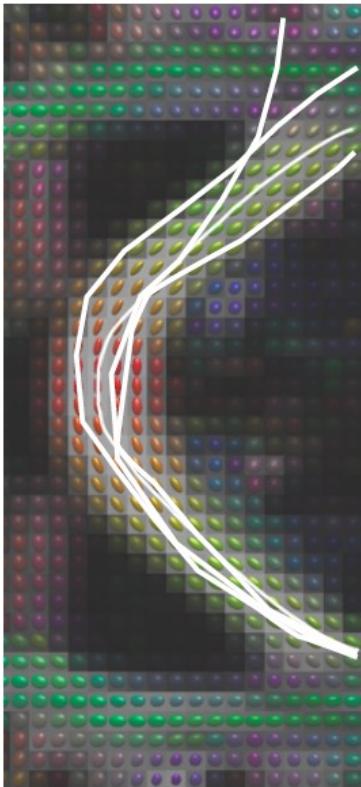
deterministic streamlining



- for each voxel, draw a streamline in the principal diffusion direction
- this gives a whole-brain structural connectivity estimate

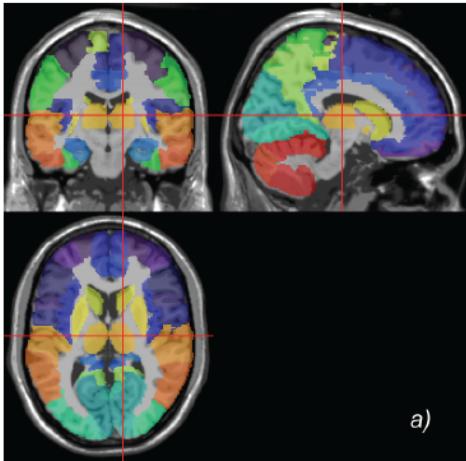
Tractography

probabilistic streamlining

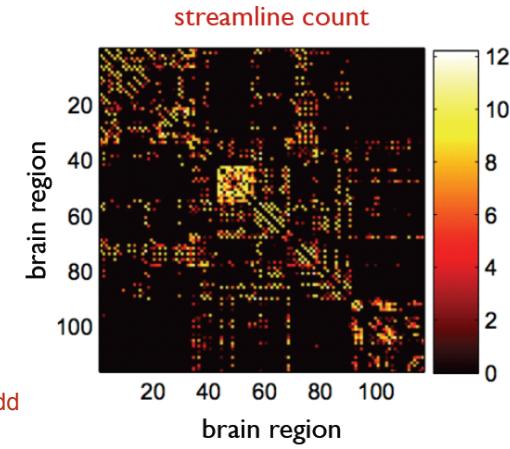


- for each voxel, draw N streamlines according to the local pdf on fiber direction
- this does not give a whole-brain structural connectivity estimate

Streamlines between brain regions

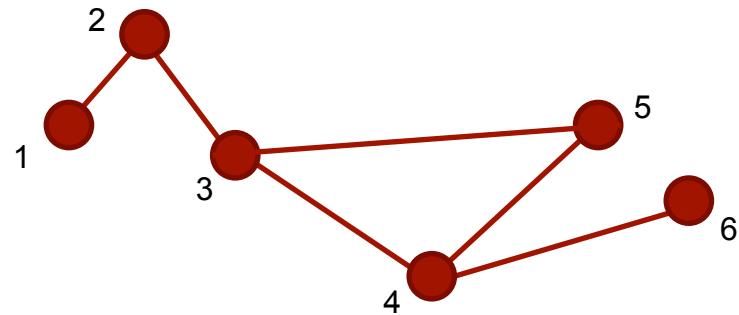


- Each node is **the center of mass** of one region.
- Streamlines connection region A to region B are **aggregated**.
- Intra-region streamlines are discarded.



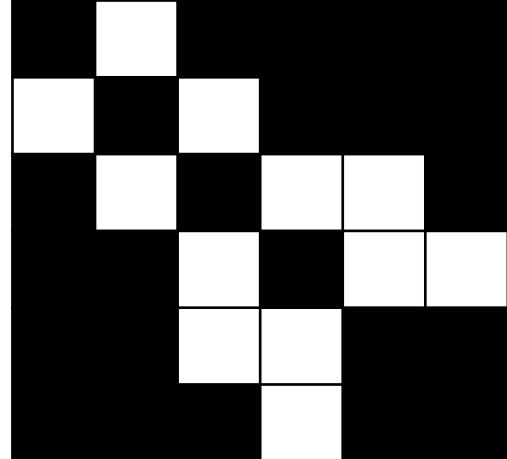
Representing a brain network

- A graph is a mathematical representation of sets of objects ('nodes' or 'vertices') and their relations ('edges', 'connections' or 'links').
- Formally, we write: $G = (V, E)$, with $E \subseteq V \times V$



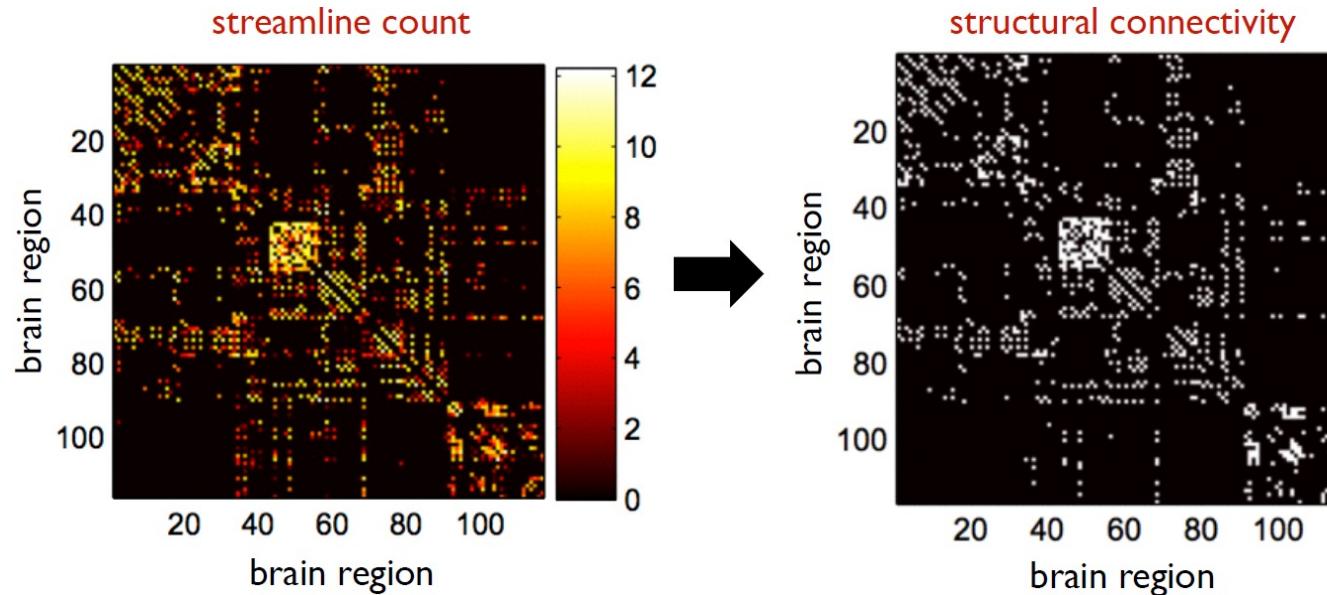
- Vertices = {1, 2, 3, 4, 5, 6}
- Edges = {{1,2}, {2,3}, {3,4}, {3,5}, {4,5}, {4,6}}

- Equivalent representation: adjacency matrix $G_{ij} = 1 \leftrightarrow (i,j) \in E$





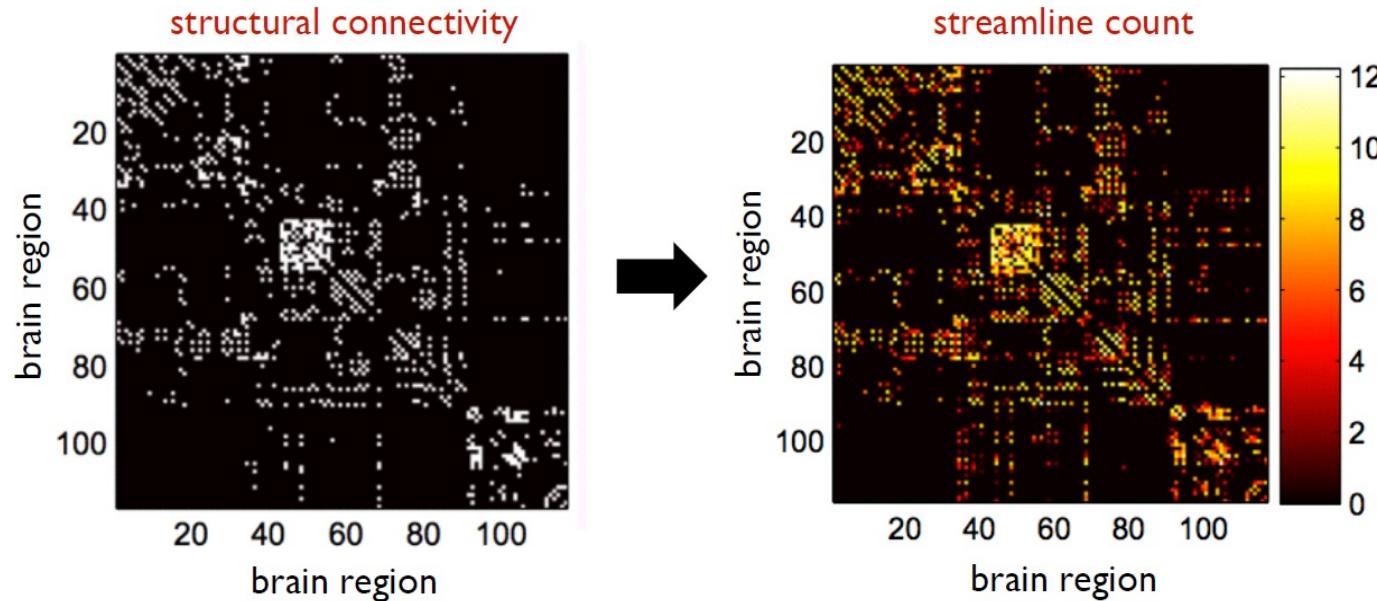
Goal: estimate (structural) connectivity from streamline counts



'Naive' approach: **threshold** count matrix → high counts become an edge.

- What threshold to choose?
- (Normalized) **streamline counts are not edge probabilities**.
 - They are conditioned on the (streamline) starting node.

Bayesian structural connectivity estimation



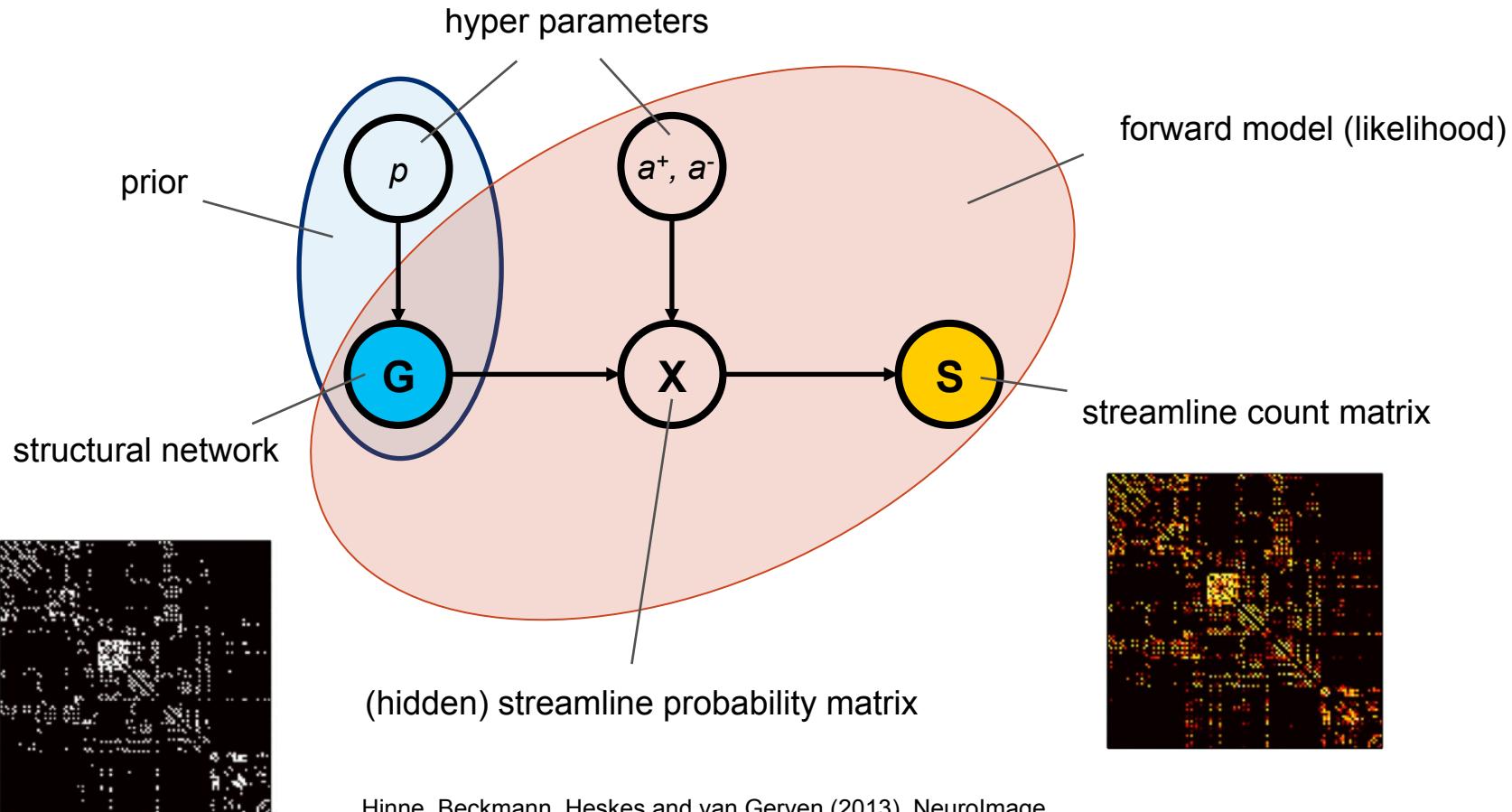
1. Model how a network **generates** the observed counts.
2. Use **Bayes' rule** to learn the network, given the observations.



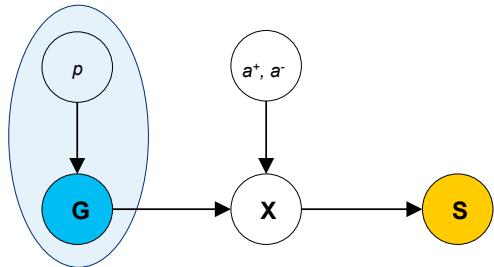
Bayesian structural connectivity estimation

- We want to take into account the **uncertainty** in probabilistic streamlining.
- We'd rather avoid arbitrary thresholds, which we don't know how to pick.
- We want to incorporate **prior knowledge**:
 - Graph-theoretical properties (e.g. degree distribution, small-worldness),
 - Anatomical constraints (i.e. minimize total edge length),
 - Known connections from other studies (i.e., edges found in tracer studies),
 - Data from other subjects/modalities.
- A **Bayesian approach** gives all of this (and more).

Generative model for structural connectivity



Prior on connectivity



$$\text{Bernoulli}(xp) = p \uparrow x (1-p) \uparrow 1-x$$

- A priori we assume **independent edges**.
- Each edge has a **Bernoulli distribution** as prior, i.e.:

$$P(\mathbf{G}|p) = \prod_{i < j} p(g_{ij} | p)$$

$$= \prod_{i < j} p(g_{ij}) (1-p)^{1-g_{ij}}$$

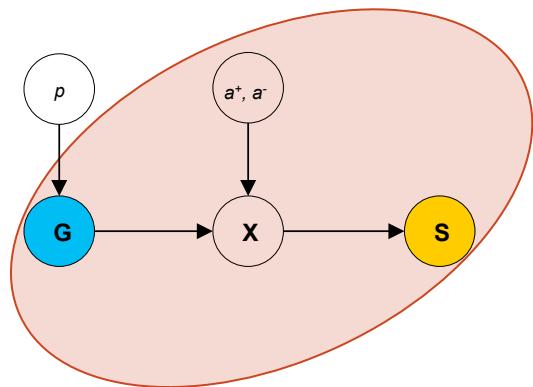
- $i < j$ because we assume the graph is **symmetric**.



(weighted) coin flip for each edge



Forward model: how does a graph ‘generate’ data



We focus on rows of **G**, **X** and **S** respectively:

- The connectivity of a region i to other regions is the vector:

$$\mathbf{g} \downarrow i = (g \downarrow i_1, \dots, g \downarrow i_K)$$

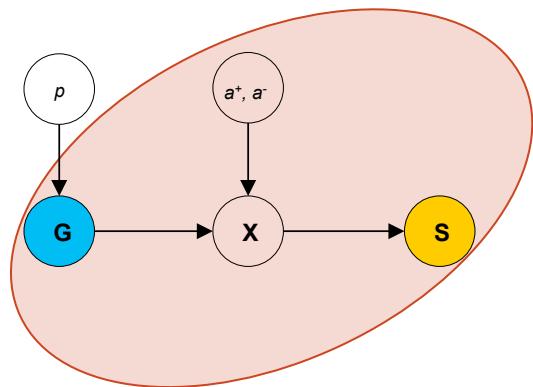
- This vector determines another vector containing the probabilities of streamlines:

$$\mathbf{x} \downarrow i = (x \downarrow i_1, \dots, x \downarrow i_K)$$

- Streamlines are drawn according to these probabilities:

$$\mathbf{s} \downarrow i = (s \downarrow i_1, \dots, s \downarrow i_K)$$

Forward model: how does a graph ‘generate’ data



- We assume again independence between regions, given \mathbf{G} , for the streamline count vectors.
- Hence, the likelihood term **factorizes over regions** as:

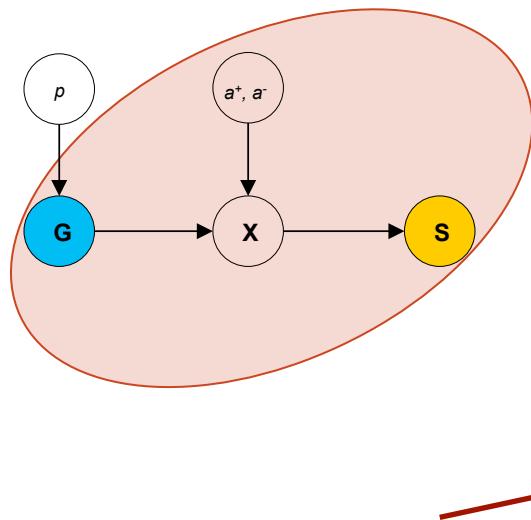
$$P_{\mathbf{S}, \mathbf{G}, a^{\uparrow+}, a^{\uparrow-}} = \prod_{i=1}^I p(\mathbf{s}_{\downarrow i} | \mathbf{g}_{\downarrow i}, a^{\uparrow+}, a^{\uparrow-})$$

where

$$p_{\mathbf{s} \downarrow i | \mathbf{g} \downarrow i, a^{\uparrow+}, a^{\uparrow-}} = \int_{\mathbf{x} \downarrow i} p_{\mathbf{s} \downarrow i | \mathbf{x} \downarrow i} p(\mathbf{x} \downarrow i | \mathbf{g} \downarrow i, a^{\uparrow+}, a^{\uparrow-})$$

- We **integrate out** the latent streamline probability vectors.
- Can be done analytically if we pick the right distributions for $p(\cdot)$.

Forward model: how does a graph ‘generate’ data



- For the streamline probability vectors $\mathbf{x} \downarrow i$ we need a distribution over probability vectors, so that $\sum_j x_{ij} = 1$ and $0 \leq x_{ij} \leq 1$. We assume:

$$p_{\mathbf{x} \downarrow i | \mathbf{g} \downarrow i, a \uparrow+, a \uparrow-} = \text{Dirichlet}(\mathbf{x} \downarrow i | \mathbf{b} \downarrow i)$$

with $\mathbf{b} \downarrow i = a \uparrow+ \mathbf{g} \downarrow i + a \uparrow- (1 - \mathbf{g} \downarrow i)$.

This is where the connections and hyperparameters are combined to inform the streamline probabilities.

- Next step: use the probability vector $\mathbf{x} \downarrow i$ to draw streamline counts:

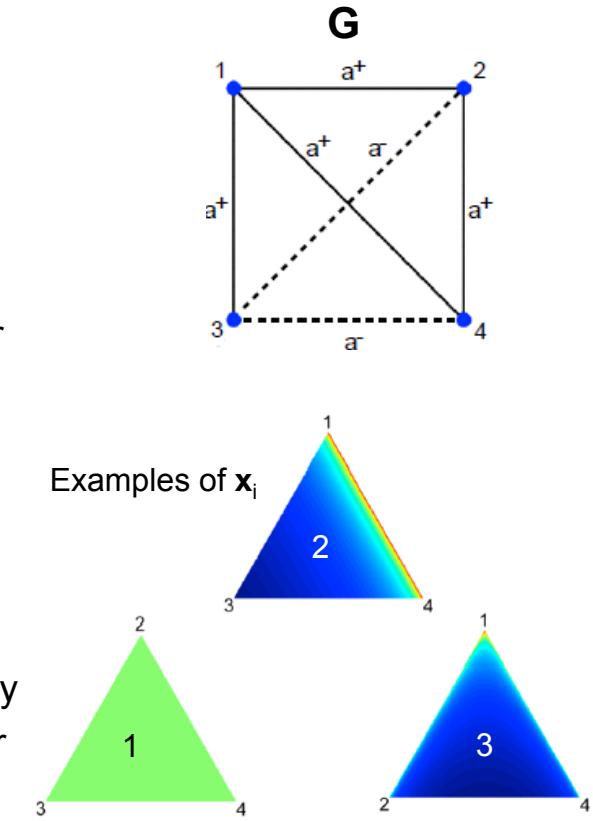
$$p_{\mathbf{s} \downarrow i | \mathbf{x} \downarrow i, S \downarrow i} = \text{Multinomial}(\mathbf{n} \downarrow i | \mathbf{x} \downarrow i, S \downarrow i)$$

with $S \downarrow i$ the total number of streamlines to draw for region i .



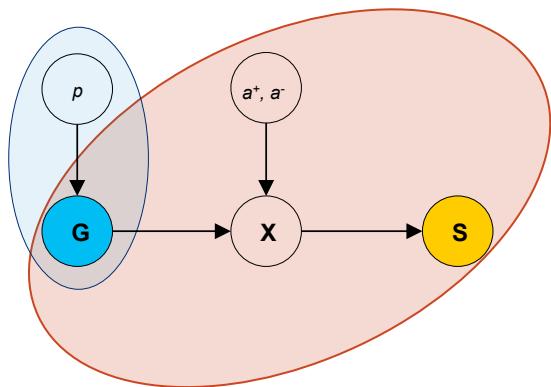
A note on Bernoulli's, Dirichlet's, Multinomials and more

- Distribution choices seem arbitrary, but they follow from the data type.
- Bernoulli** is appropriate for events with a binary outcome, e.g. success or no success: a weighted coinflip, with a weight parameter.
- Binomial** is the extension of the Bernoulli that is about the number of successes: how many successes out of N trials (same weight parameter).
- Categorical** is another extension of the Bernoulli, which has M outcomes: e.g. a 6-sided dice. Parameter is a loading vector with probabilities for each outcome.
- Multinomial** is to Categorical as Binomial is to Bernoulli: how many successes for each face of the dice. It has the same loading vector as the Categorical.
- Dirichlet** is a probability distribution over loading vectors: which





The generative model



- Putting it all together:
- For each vector i , we integrate out the latent variables $\mathbf{x} \downarrow i$:

$$p_{\mathbf{s} \downarrow i | \mathbf{g} \downarrow i, a \uparrow+, a \uparrow-} = \int_{\mathbf{x} \downarrow i} p_{\mathbf{s} \downarrow i | \mathbf{x} \downarrow i} p(\mathbf{x} \downarrow i | \mathbf{g} \downarrow i, a \uparrow+, a \uparrow-)$$

$$= S \downarrow i! / \prod_{j \uparrow i} s \downarrow ij \quad \Gamma(\sum_{j \uparrow i} b \downarrow ij) / \Gamma(\sum_{j \uparrow i})$$

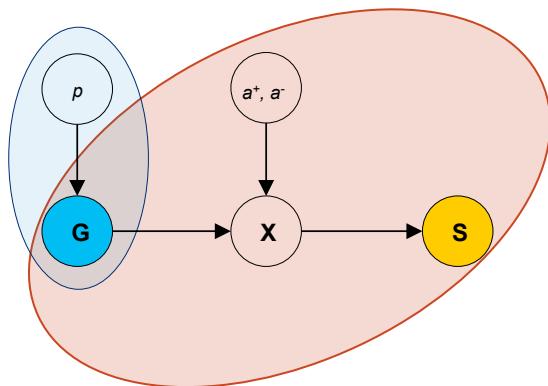
with $\mathbf{b} \downarrow i = a \uparrow+ \mathbf{g} \downarrow i + a \uparrow- (1 - \mathbf{g} \downarrow i)$ and $S \downarrow i = \sum_{j \uparrow i} s \downarrow ij$.

- And for the prior we had

$$P(\mathbf{G}|p) = \prod_{i < j} p^{g \downarrow ij} (1-p)^{1-g \downarrow ij}$$



The generative model



- Bayes rule time!

$$P(\mathbf{G}|\mathbf{S}, \theta) = P(\mathbf{S}|\mathbf{G}, \theta) P(\mathbf{G}|\theta) / P(\mathbf{S}|\theta) \quad 1 \downarrow \mathbf{G} \in \mathcal{M}$$

With $\theta = (a_1^+, a_1^-, p)$, to keep it readable and $1 \downarrow \mathbf{G} \in \mathcal{M}$ is 1 if \mathbf{G} is symmetric and without loops, and 0 otherwise.

- So far so good, but we **cannot actually compute** $P(\mathbf{G}|\mathbf{S}, \theta)$; for K regions there are $2 \uparrow K(K-1)/2$ possible networks!
 - 10 regions = 3.52×10^{13} networks
 - This problem is **intractable**, so we need **approximate inference** to explore the networks with reasonable probability.
 - Who cares about $P(\mathbf{G}|\mathbf{S}, \theta) \approx 0$ anyway!



Markov chain Monte Carlo

- We use MCMC to collect samples from all possible matrices, that are supported by the data and our prior beliefs.
- Starting from some matrix \mathbf{G}^t , we propose a new matrix \mathbf{G}^* by flipping one element. We compute:

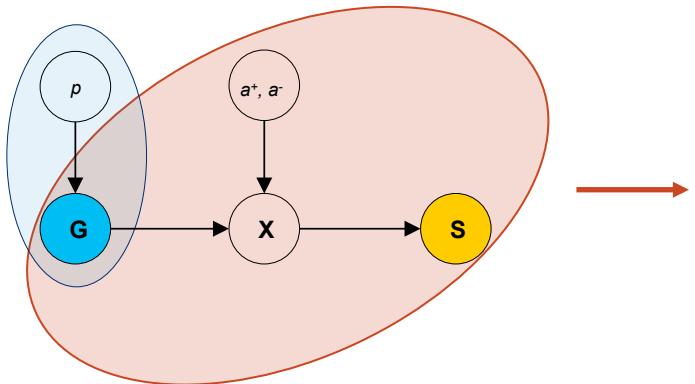
$$\alpha = \frac{P(\mathbf{S}|\mathbf{G}^*, \theta)}{P(\mathbf{S}|\mathbf{G}^t, \theta)} \cdot \frac{P(\mathbf{G}^t | \theta)}{P(\mathbf{G}^* | \theta)}$$

Note that we don't need $P(\mathbf{S}|\theta)$; it is the same for \mathbf{G}^t and \mathbf{G}^* .

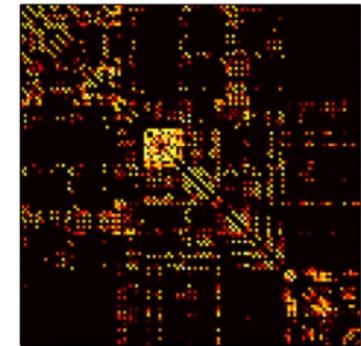
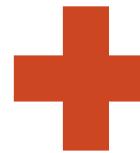
- With probability α we accept the proposed sample; $\mathbf{G}^{t+1} = \mathbf{G}^*$. Otherwise, we remain where we are; $\mathbf{G}^{t+1} = \mathbf{G}^t$.
- As $t \rightarrow \infty$, the collection of samples goes to $P(\mathbf{G}|\mathbf{S}, \theta)$.



Markov chain Monte Carlo in action

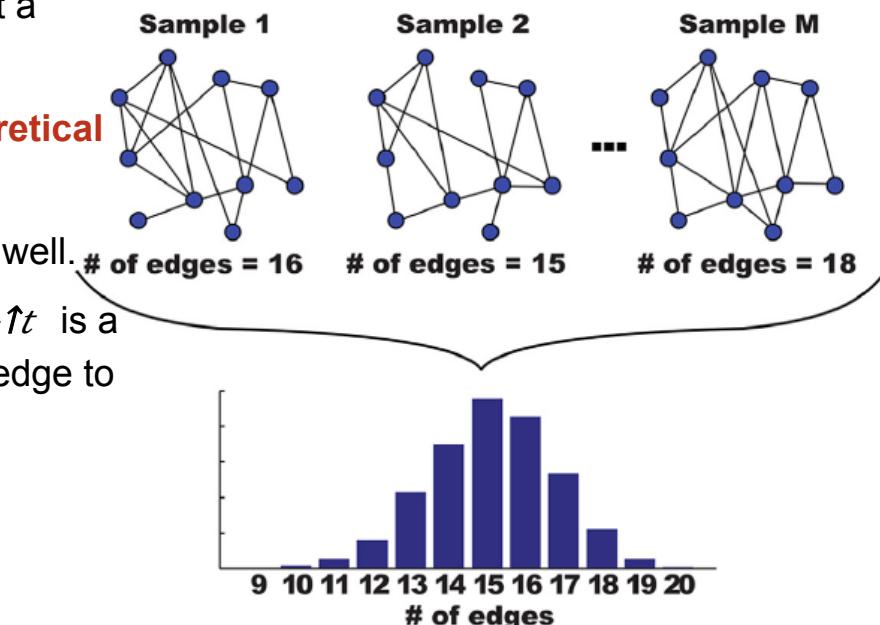
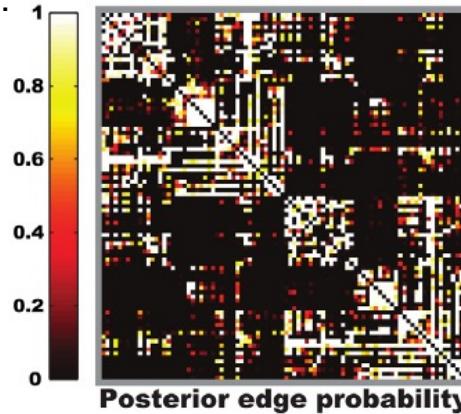


MCMC



Posterior distributions instead of point estimates

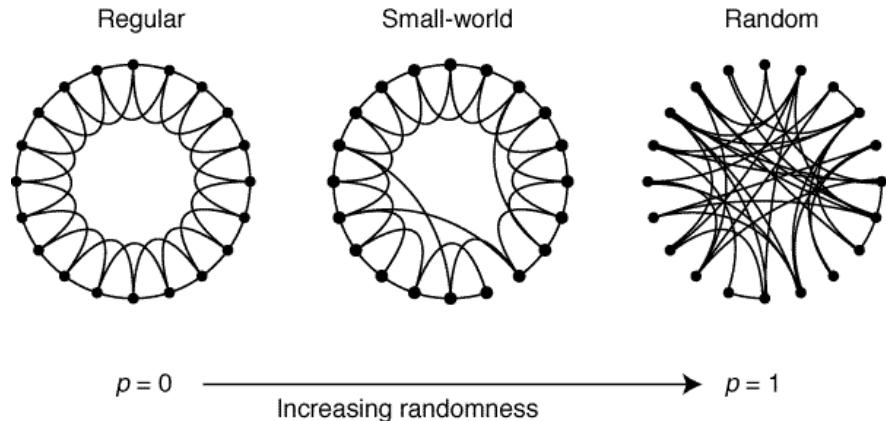
- We have a(n approximate) posterior distribution, not a point estimate.
- For each sample, we can compute any **graph-theoretical** quantity.
 - This implies a **distribution over this quantity** as well.
- The posterior expectation $\mathbb{E}(\mathbf{G}|\mathbf{S}, \theta) = 1/T \sum_{t=1}^T \mathbf{G}^\dagger_t$ is a summary statistic; it shows the probability for each edge to be present.



Janssen, Hinne, Heskes, van Gerven (2014), Front. Comp. Neur.

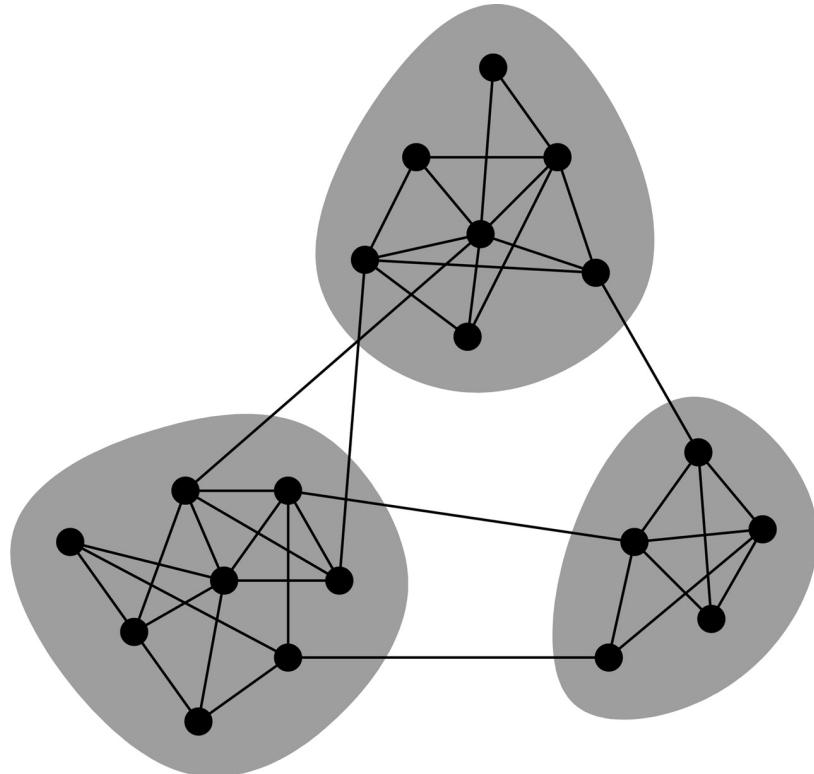
Network theory: Small-world phenomenon

- Seminal paper by Watts & Strogatz described the network of a small roundworm, *C. Elegans*, and showed that it had
 - **High clustering:**
 - Many connected nodes have the same neighbours.
 - **Short path length**
 - It takes only few ‘hops’ to reach any node in the network.
- These properties together are known as '**small-worldness**'.
- Brain networks are often small-world.



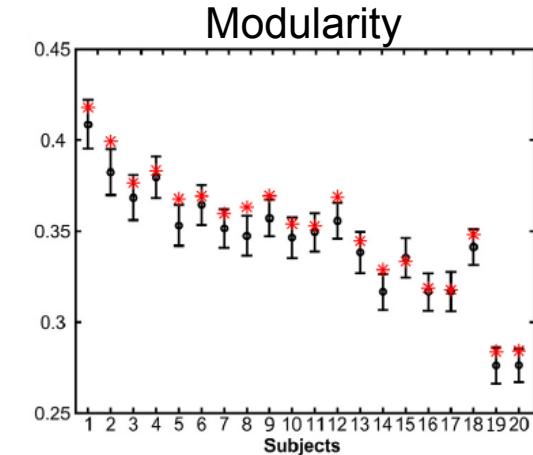
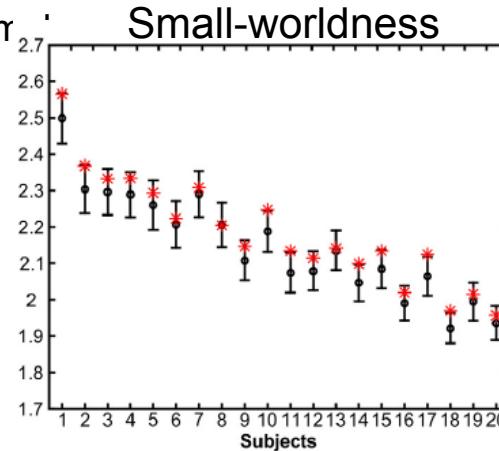
Network theory: Modularity

- Networks can be subdivided into **clusters** (a.k.a. modules, communities);
- A common way to define a cluster is to say that
 - A cluster has **many internal connections**, but
 - **Few connections to other clusters.**
- **Modularity** is a score that indicates how well a network can be clustered.
- Brain networks are often clustered by **optimizing modularity**.



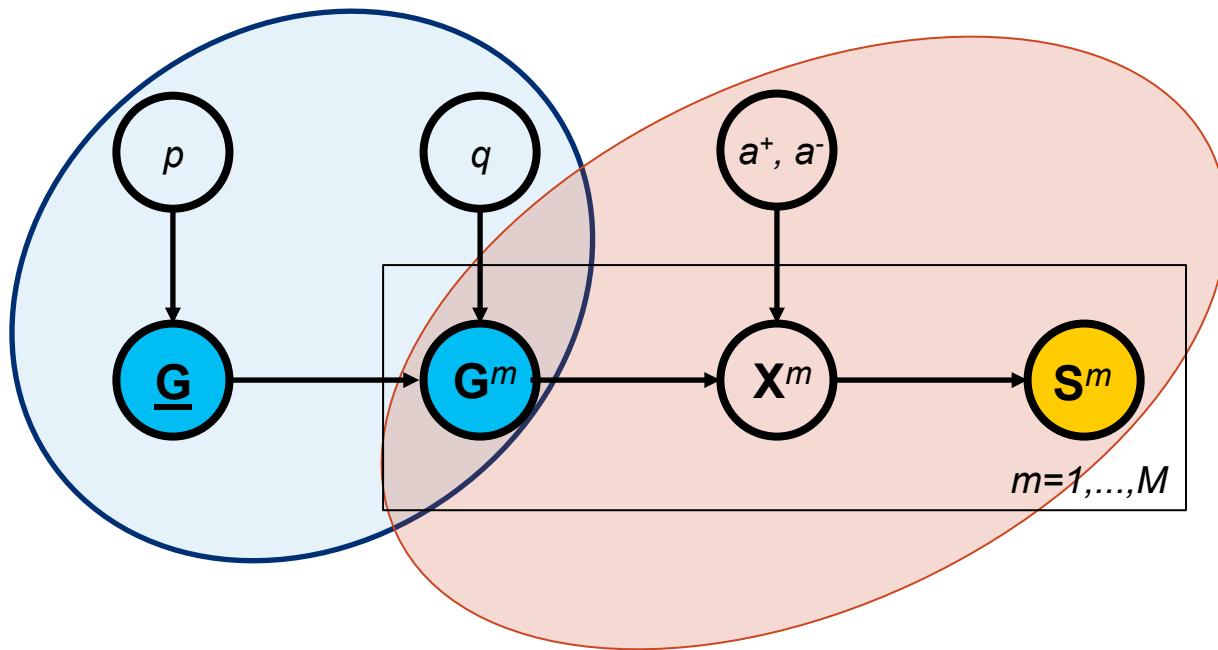
Posterior distributions instead of point estimates

- Posterior distribution of network measure f ;
 - Compute f for each sample.
 - Compare with thresholded network.
 - Threshold set to have network density match the mean density of the same



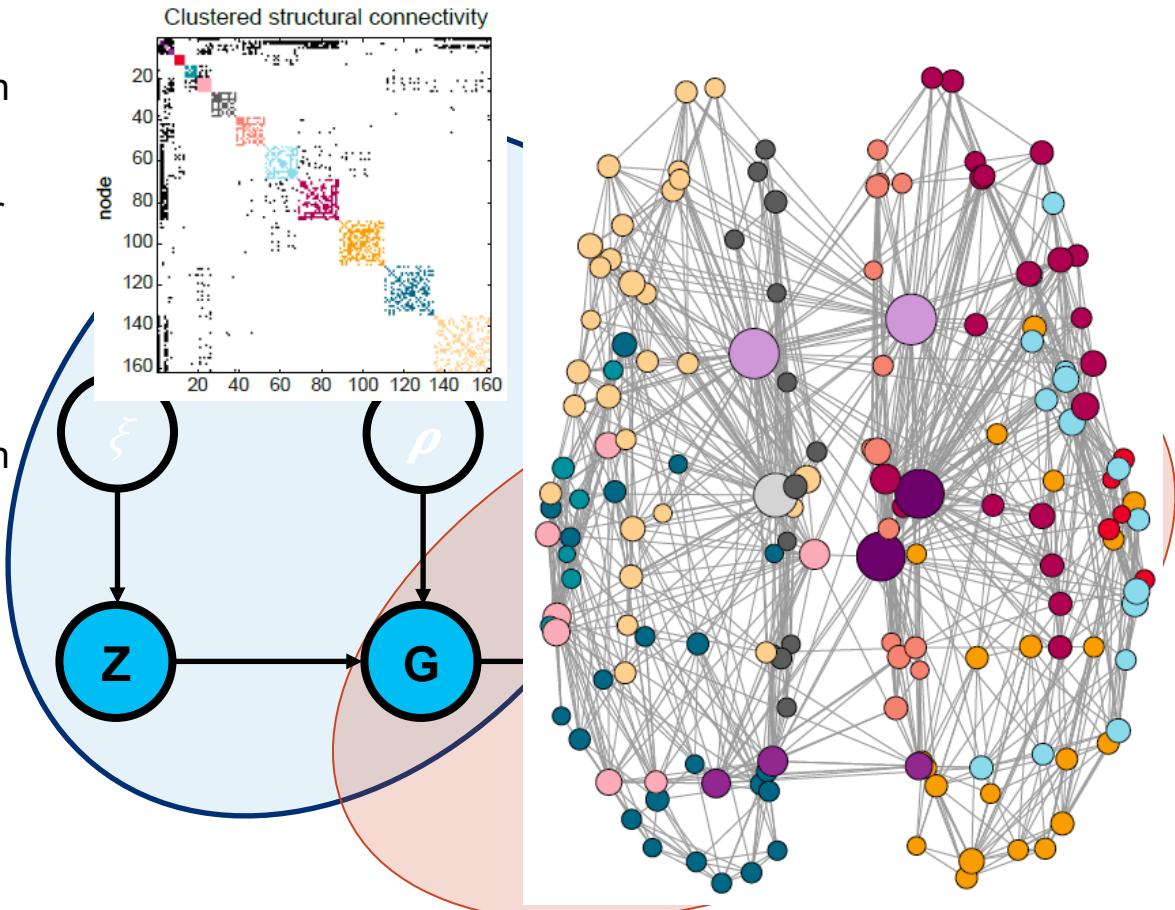
Model extensions

- Hierarchical model: Assume each subject's brain network \mathbf{G}^m is an instantiation of an ur-network $\underline{\mathbf{G}}$. Connections are inherited from $\underline{\mathbf{G}}$ with probability q .
- Consequence:
 - Subjects inform each other.



Model extensions

- Clustering: replace coin-flip prior with a clustering model.
- Connections from cluster a to cluster b all have the same probability.,.
- Nodes within a cluster are exchangable.
- Learn the latent $\rho_{G|Z,\xi}$ in addition to connectivity \mathbf{G} , i.e.:
- Models can be combined ad infinitum.

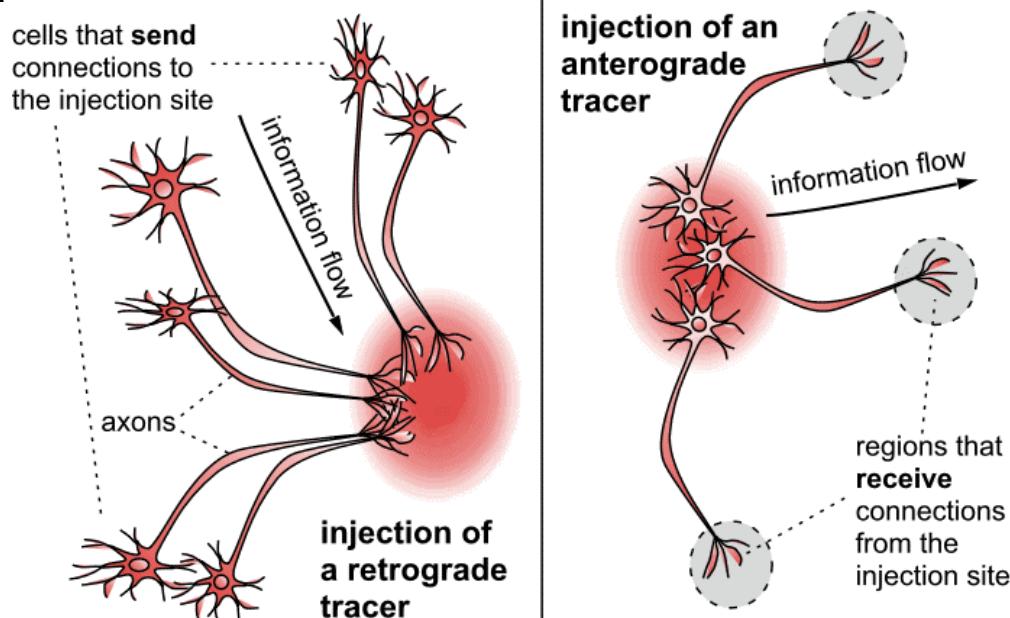


Predicting structural connectivity

Spatial embedding of network structures

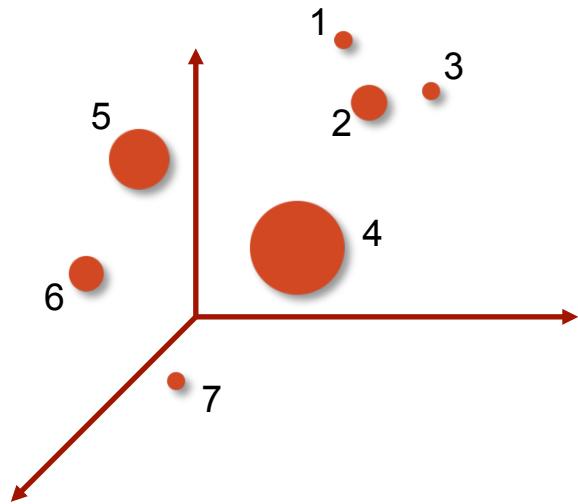
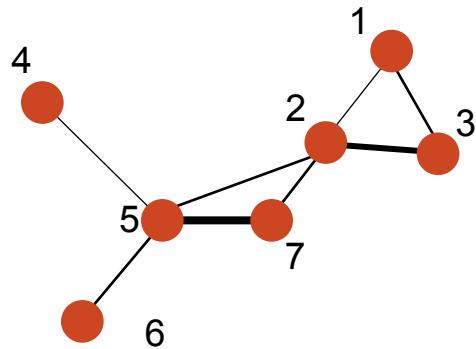
Different modalities for structural connectivity

- Downsides of diffusion MRI
 - **Indirect measurement.**
 - Lot of **noise/uncertainty**.
 - Difficulty with long-range connections.
- **Tracer studies** however
 - **Very accurate.**
 - Have **directionality**.
- But:
 - Requires **animal sacrifice**.
 - **Labour intensive**.

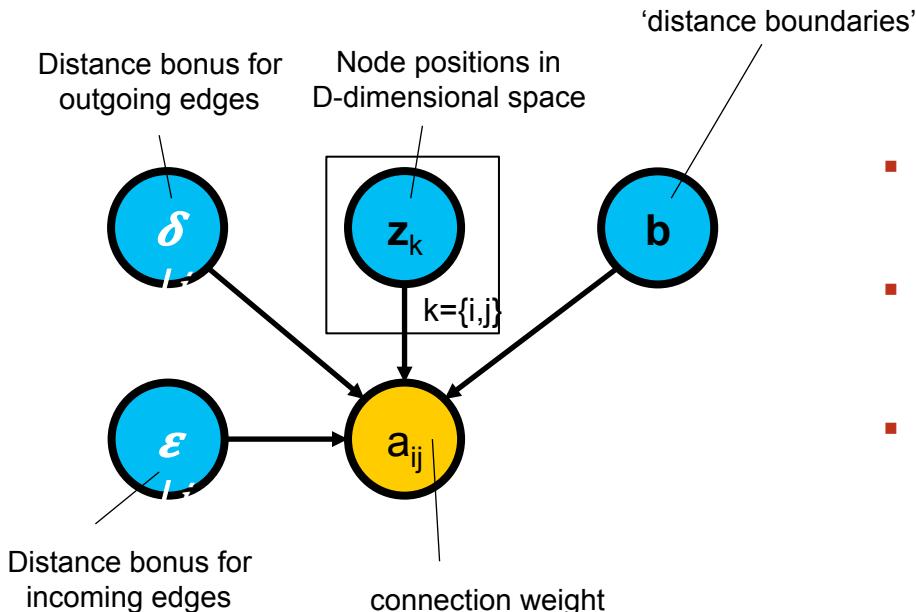


Link prediction

- Predicting connections avoids new data collection
- Idea:
 - Connections depend on latent factors
 - Learn factors from available data
 - Use factors to predict (absence/presence of) new connections
- Here: **latent space models**

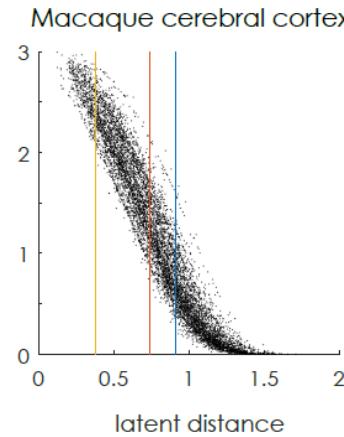
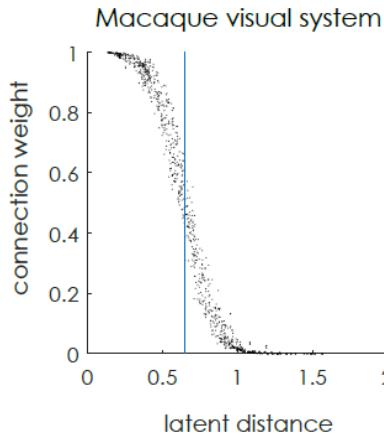


Ordinal latent space model

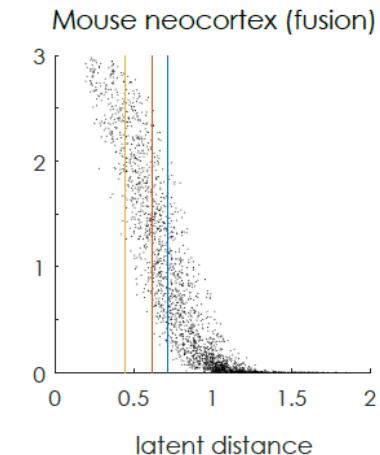
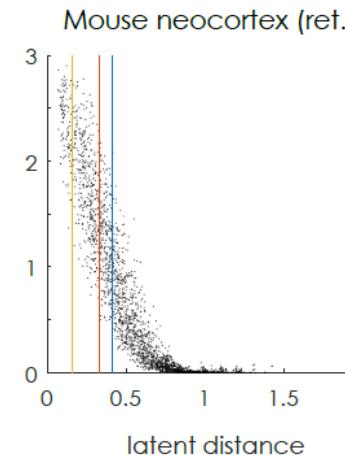
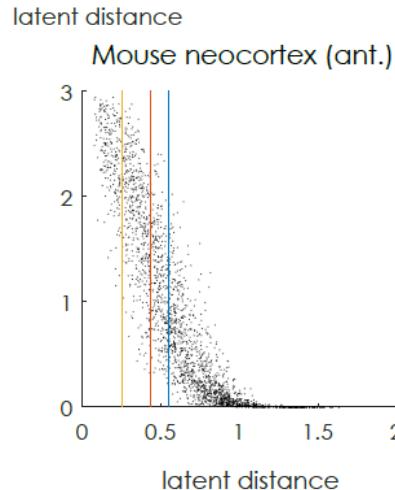


- Data: matrix A with elements *absent*, *weak*, *moderate* or *strong* (connection weights)
- Weak/uninformative priors not shown.
- Posterior $P(\mathbf{A}, \mathbf{Z}, \mathbf{b}, \boldsymbol{\delta}, \boldsymbol{\epsilon} | \boldsymbol{\theta})$, with $\boldsymbol{\theta}$ the hyper parameters not shown here, is approximated using the Stan software package (Hamiltonian Monte Carlo).
- Optimal latent dimensionality D is determined using cross-validation.

Learning latent embedding

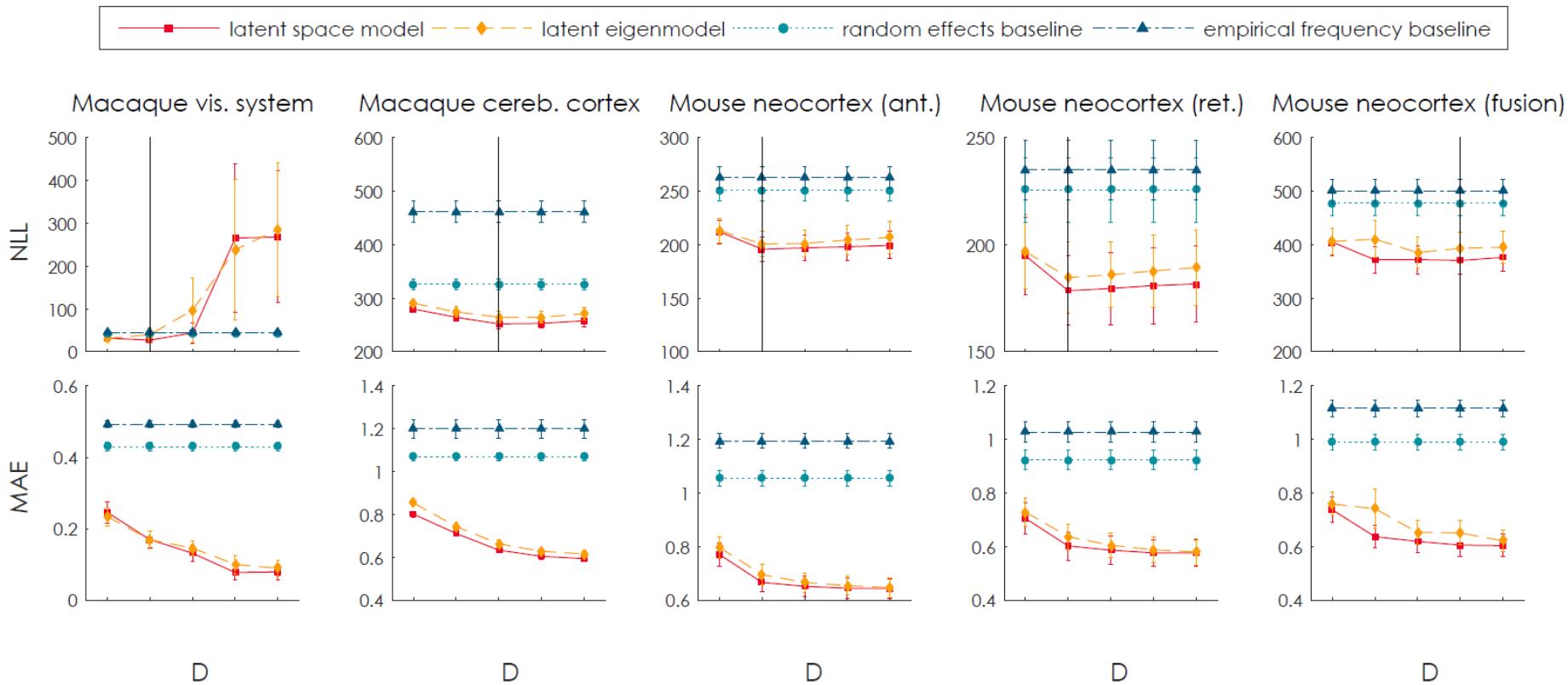


- Connection weight as function of (latent) distance, for optimal number of dimensions, $D = (2, 3, 2, 2, 4)$.
- Vertical variance due to δ and ϵ .
- Boundaries **b** indicated.



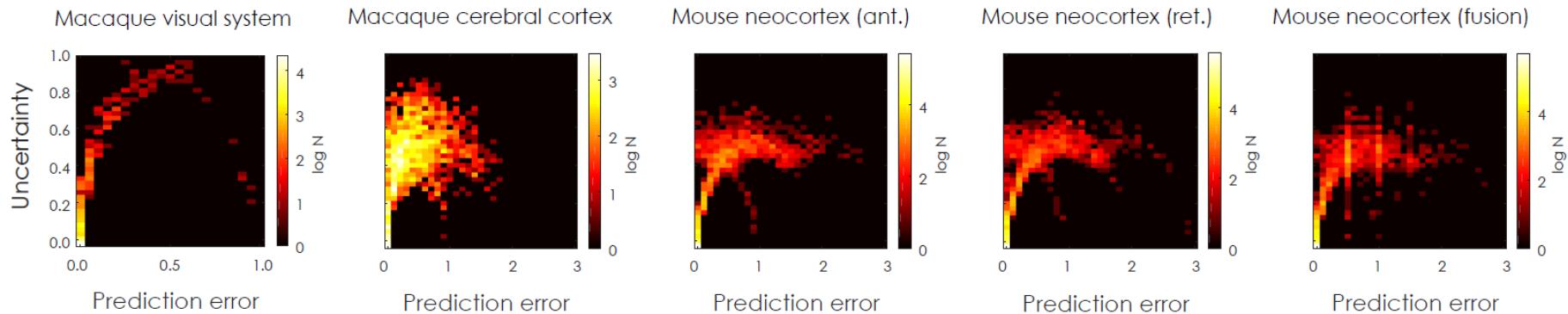


Experiments on macaque and mouse tracer data



Experiments on macaque and mouse tracer data

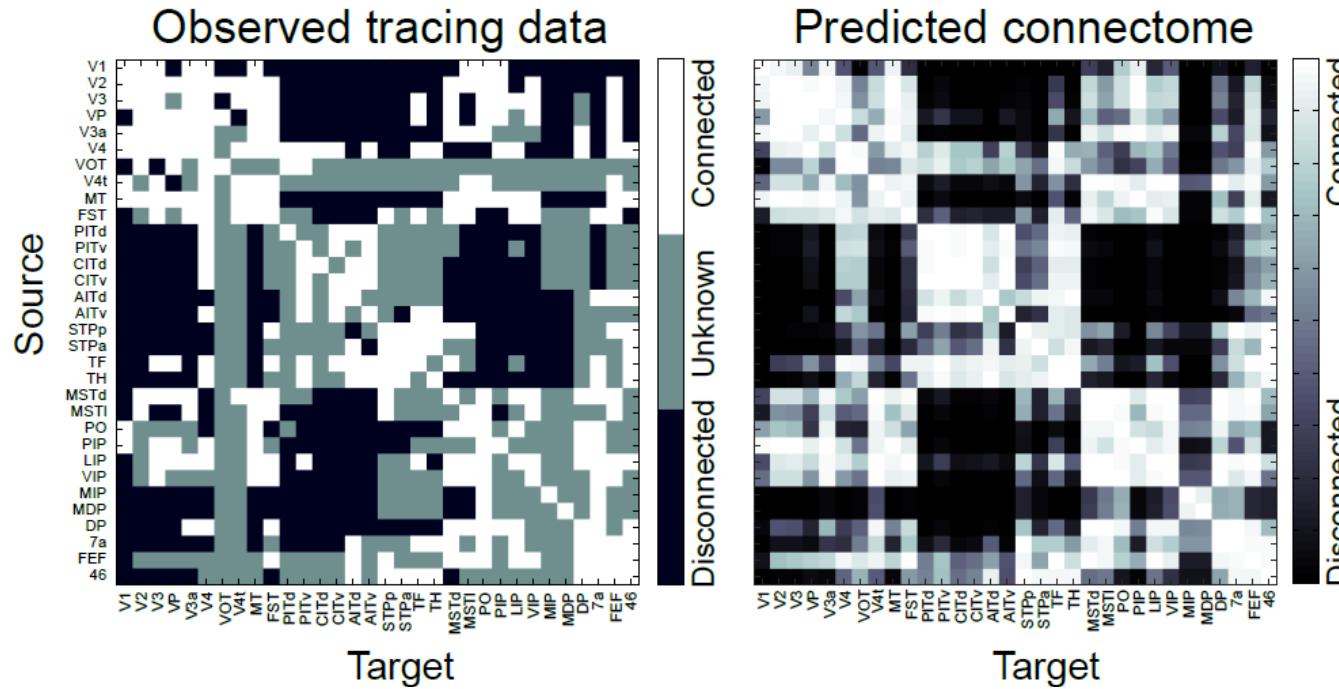
- Uncertainty in prediction reflects accuracy (in cross-validation).
- Low uncertainty goes hand-in-hand with low error.





Predicted connectome for the macaque visual system

- Using D=2 latent dimensions.

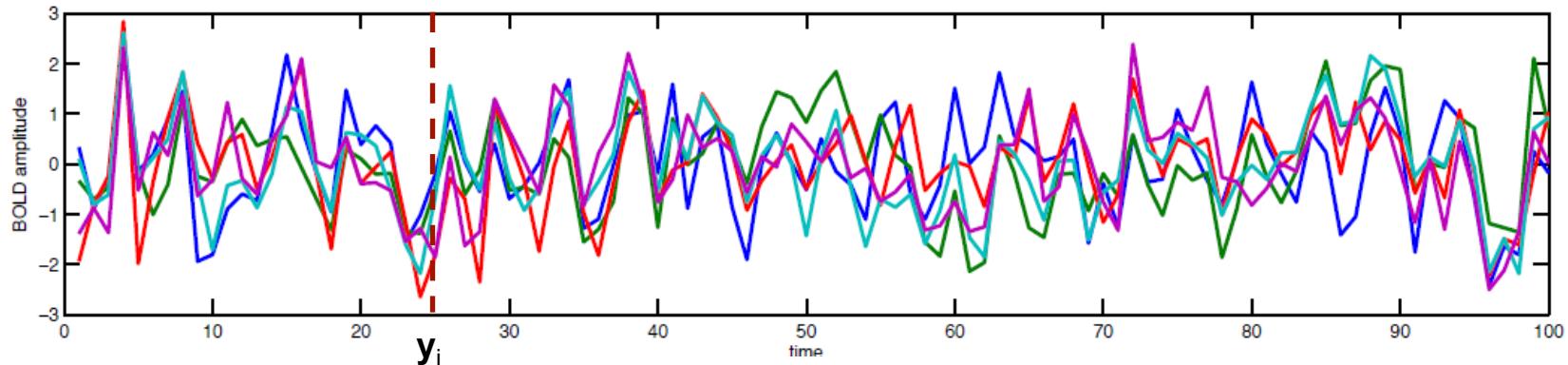


Bayesian functional connectivity analysis

How do different brain regions communicate?

Functional connectivity

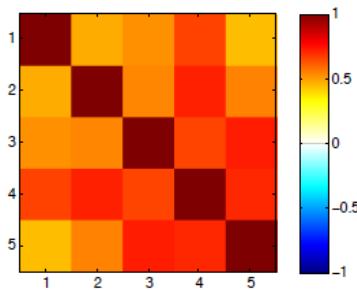
- Functional connectivity refers to **correlated activity** in spatially segregated brain regions.
- Consider the observed **BOLD response time series** in multiple brain regions:



- Let the p -dimensional vector \mathbf{y}_t denote observed BOLD response for p brain regions at time i .
- Let $\mathbf{Y} \in \mathbb{R}^{p \times n}$ be the full data set, with n observations (time points) for p regions.

Functional connectivity: covariance

- The easiest way to express functional connectivity is by computing the **covariance** between the time series of pairs of regions.
 - High covariance = strong connectivity.
- Covariances collected in a matrix give the covariance matrix Σ :

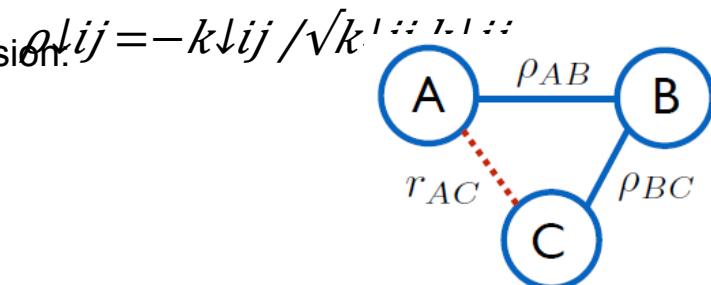


- If the data is standardized (0-mean and unit variance), the covariance matrix is equal to the **correlation** matrix.



Functional connectivity: precision & partial correlation

- Computing/ the (empirical) covariance is easy; $\Sigma = 1/n \mathbf{Y} \mathbf{Y}^T$.
- But covariance has an **inherent problem**; it captures **indirect** connectivity too!
 - If A-B and B-C, then also A-C.
 - We cannot distinguish direct from indirect connections by covariance only.
- Alternative: use the **inverse covariance** (precision) matrix $\mathbf{K} = \Sigma^{-1}$.
- Precision captures only (sort of) **direct connections**.
 - **Partial correlations** can be computed from the precision:



- The precision matrix has an interesting property:
 - **Zero elements** of the precision matrix reflect that the corresponding brain regions are **conditionally independent**, given the other regions.



Functional connectivity: sparse partial correlation matrix

- Zeros in the partial correlation matrix tell us which regions are **conditionally independent**.
- We obtain the precision matrix by computing the **maximum likelihood estimate** (MLE):

$$\mathbf{K} = \operatorname{argmax}_{\mathbf{K} \in \mathcal{P}^{\uparrow+}} [\log |\mathbf{K}| - \langle \boldsymbol{\Sigma}, \mathbf{K} \rangle] = \boldsymbol{\Sigma}^{-1}$$

- In practice, random fluctuations due to finite data will make par. corr. elements **non-zero**.
- With a graphical-lasso **prior** on the magnitude of the elements of \mathbf{K} we compute instead:

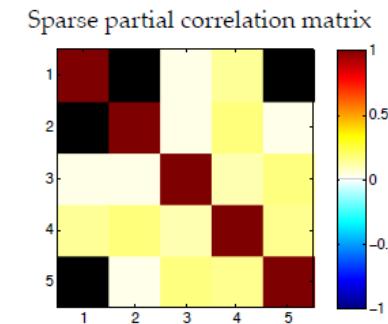
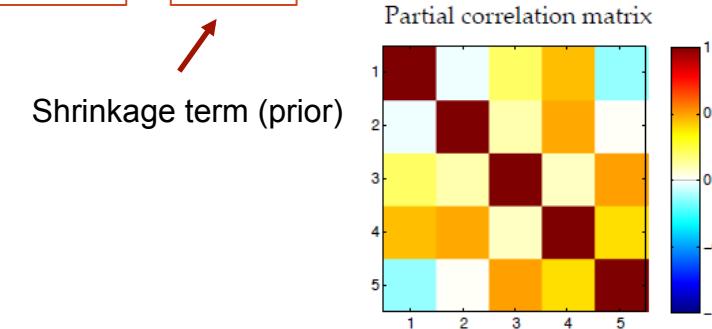
$$\mathbf{K} = \operatorname{argmax}_{\mathbf{K} \in \mathcal{P}^{\uparrow+}} [\log |\mathbf{K}| - \langle \boldsymbol{\Sigma}, \mathbf{K} \rangle - \lambda \|\mathbf{K}\|_1]$$

Likelihood term (as before)

Shrinkage term (prior)

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{A}^{\uparrow T} \mathbf{B})$$

$P^{\uparrow+}$ is the space of pos. def. matrices





Bayesian functional connectivity analysis

- The graphical lasso introduces a **bias** towards smaller partial correlations.
- Point estimates **ignore uncertainty**.

What could a **Bayesian approach to functional connectivity** analysis look like?

- We have a **multivariate Gaussian likelihood** term:

$$P(\mathbf{Y}|\mathbf{K}) = \prod_{i=1}^n P(\mathbf{y}_i | \mathbf{K}) = \prod_{i=1}^n \text{Normal}(\mathbf{y}_i | 0, K) \propto |\mathbf{K}|^{n/2} \exp[-\frac{1}{2} \langle \mathbf{K}, \mathbf{Y}^T \mathbf{Y} \rangle]$$

- We need a **prior** on the precision matrix \mathbf{K} .
- A default prior on precision matrices is known as the **Wishart distribution**.
 - It doesn't encourage zero partial correlations either.
 - We wish to explicitly model the conditional independence structure of the precision matrix, cf. the graphical lasso!



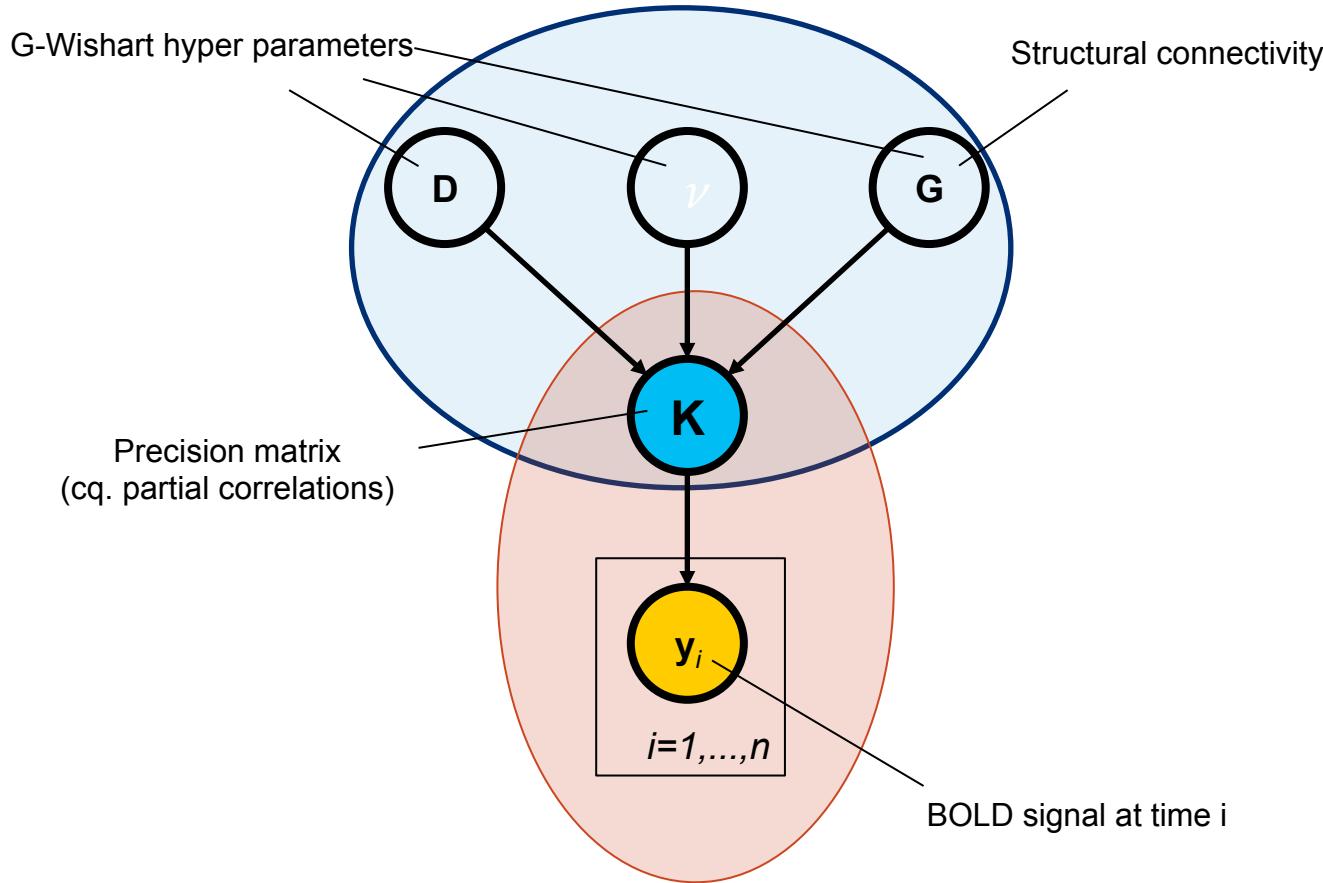
Bayesian functional connectivity analysis

- Conditional independence structure is represented as a graph $\mathbf{G} \rightarrow$ **independence = no edge**
- G-Wishart prior on precision matrix \mathbf{K}
$$P(\mathbf{K}|\mathbf{G}, \nu, \mathbf{D}) = |\mathbf{K}|^{\nu-2}/2^{\nu/2} \exp[-\frac{1}{2} \mathbf{K}^T \mathbf{D}^{-1} \mathbf{K}] \prod_{\mathbf{K} \in \mathcal{P}} \mathbf{1}^T \mathbf{K} + (\mathbf{G})$$

with hyper parameters ν and \mathbf{D} .

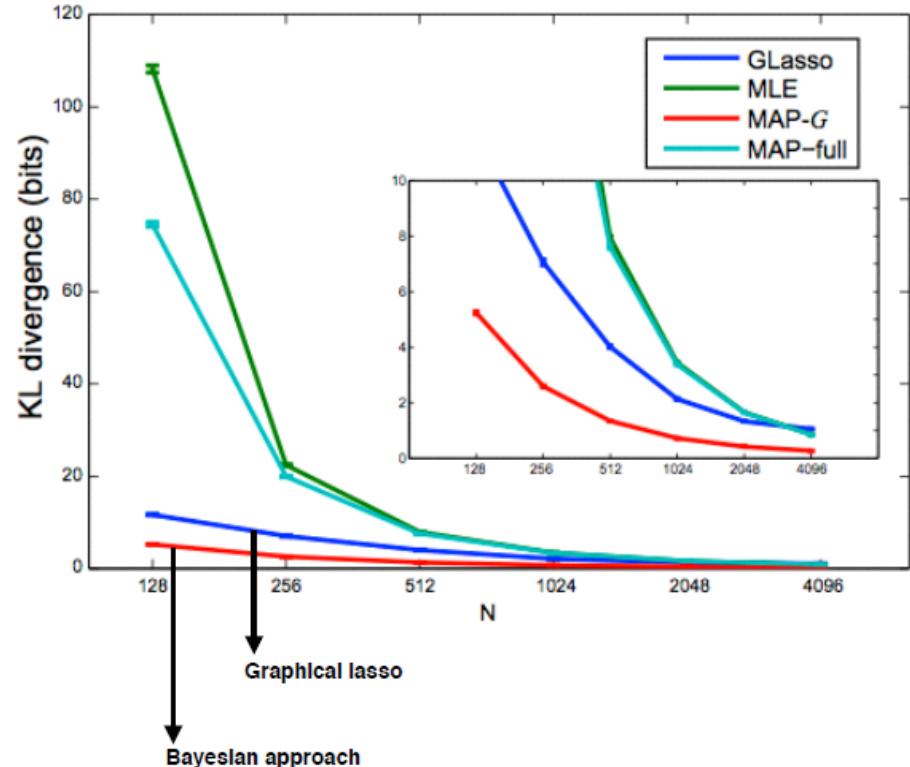
- Key idea: **structural connectivity** serves as an estimate for \mathbf{G} .
 - Rationale: two regions can only have functional communication if they're anatomically connected.
 - Partial correlations are **not shrunk**; they are either zero or they may have any weight.
 - Sampling from the posterior $P(\mathbf{K}|\mathbf{G}, \nu, \mathbf{D}, \mathbf{Y})$ is straightforward (Lenkoski, 2013).

Bayesian functional connectivity: generative model



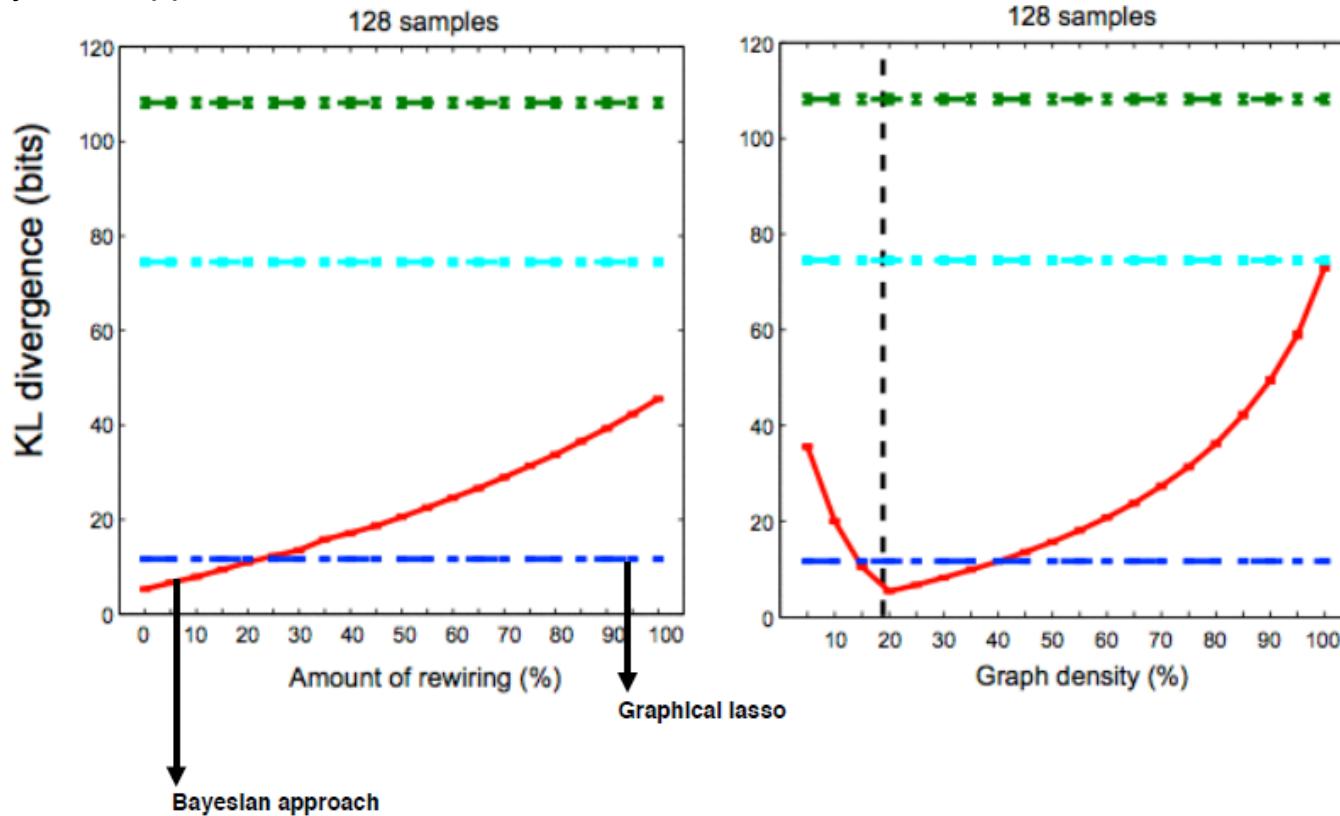
Simulation results

- Distance to ground-truth precision matrix (assuming known \mathbf{G})



Simulation results

- Bayesian approach is somewhat robust to errors in \mathbf{G} .



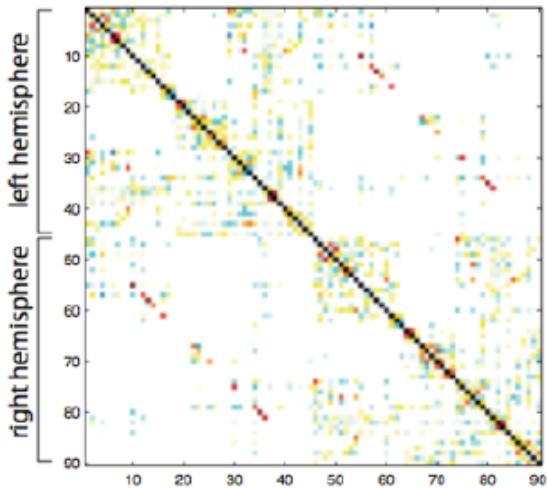
Empirical results: partial correlations for 90 brain regions

structural graph

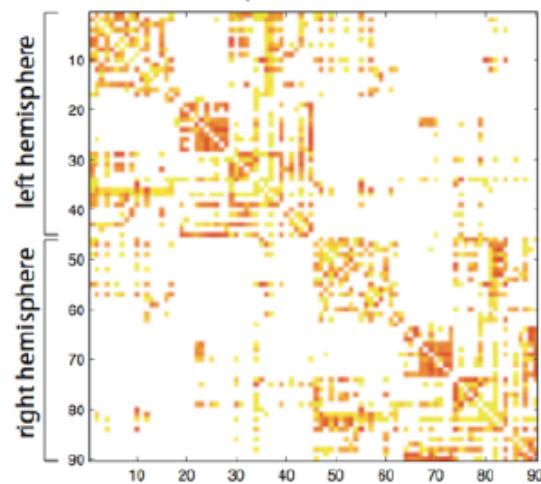


Hinne, Ambrogioni, Janssen, Heskes, van Gerven (2013), NeuroImage

S1 mean partial correlation

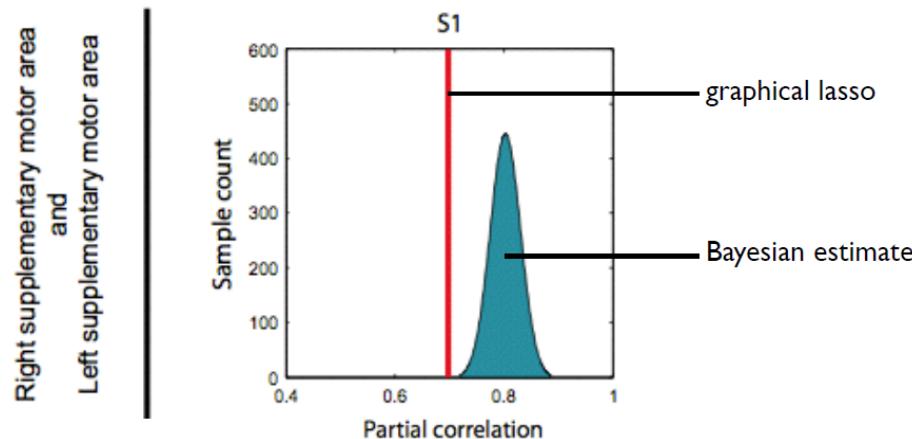


S1 SD partial correlation



Empirical results: partial correlations for 90 brain regions

- Highest partial correlations for:
 - **Functional homologues** in left and right hemisphere
 - Pallidum—putamen
 - Intra-hemispheric frontal areas
- Bayesian approach gives **credible intervals** on partial correlations

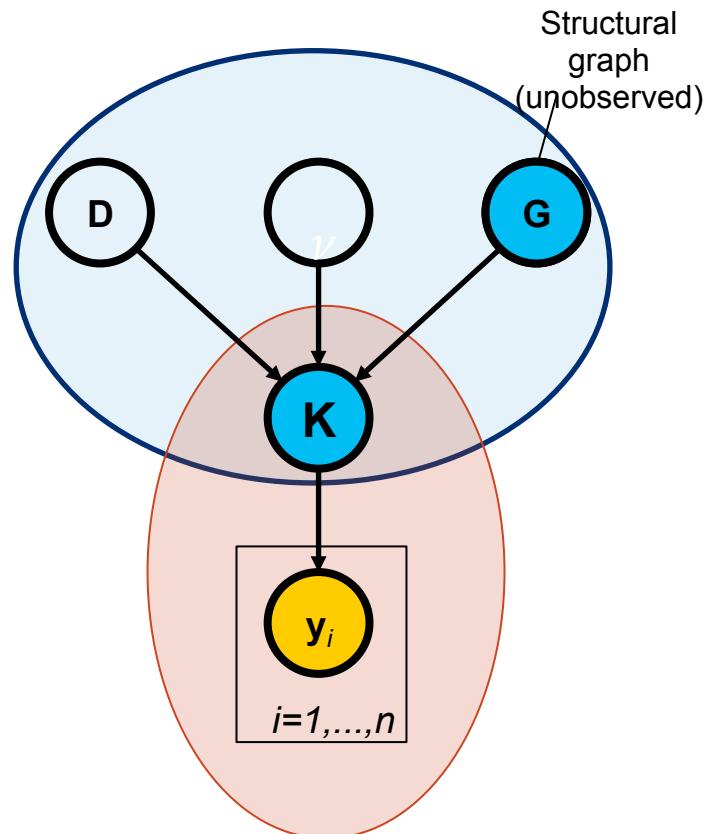


- Graphical lasso is biased towards small partial correlations.

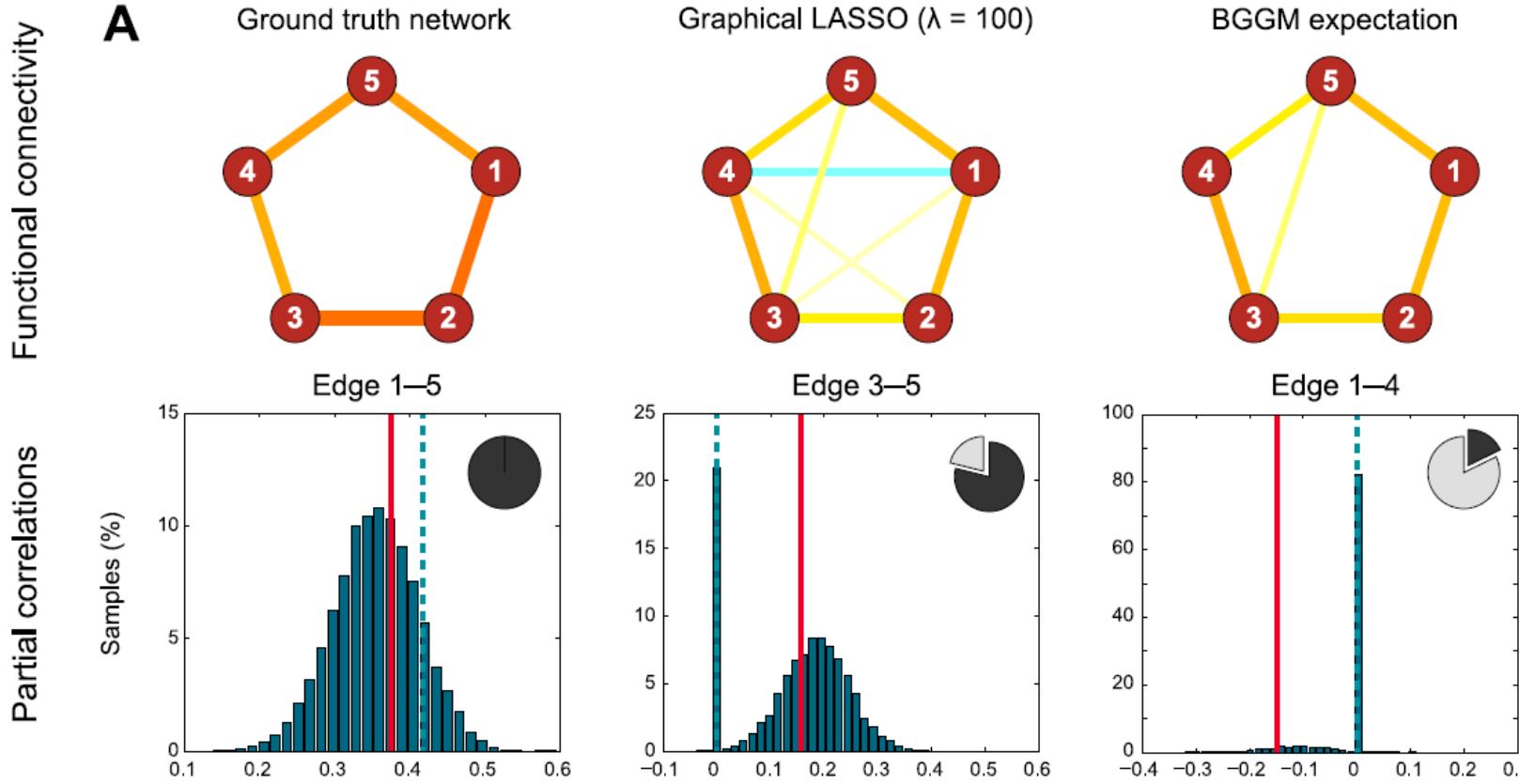


Joint inference of functional and structural connectivity

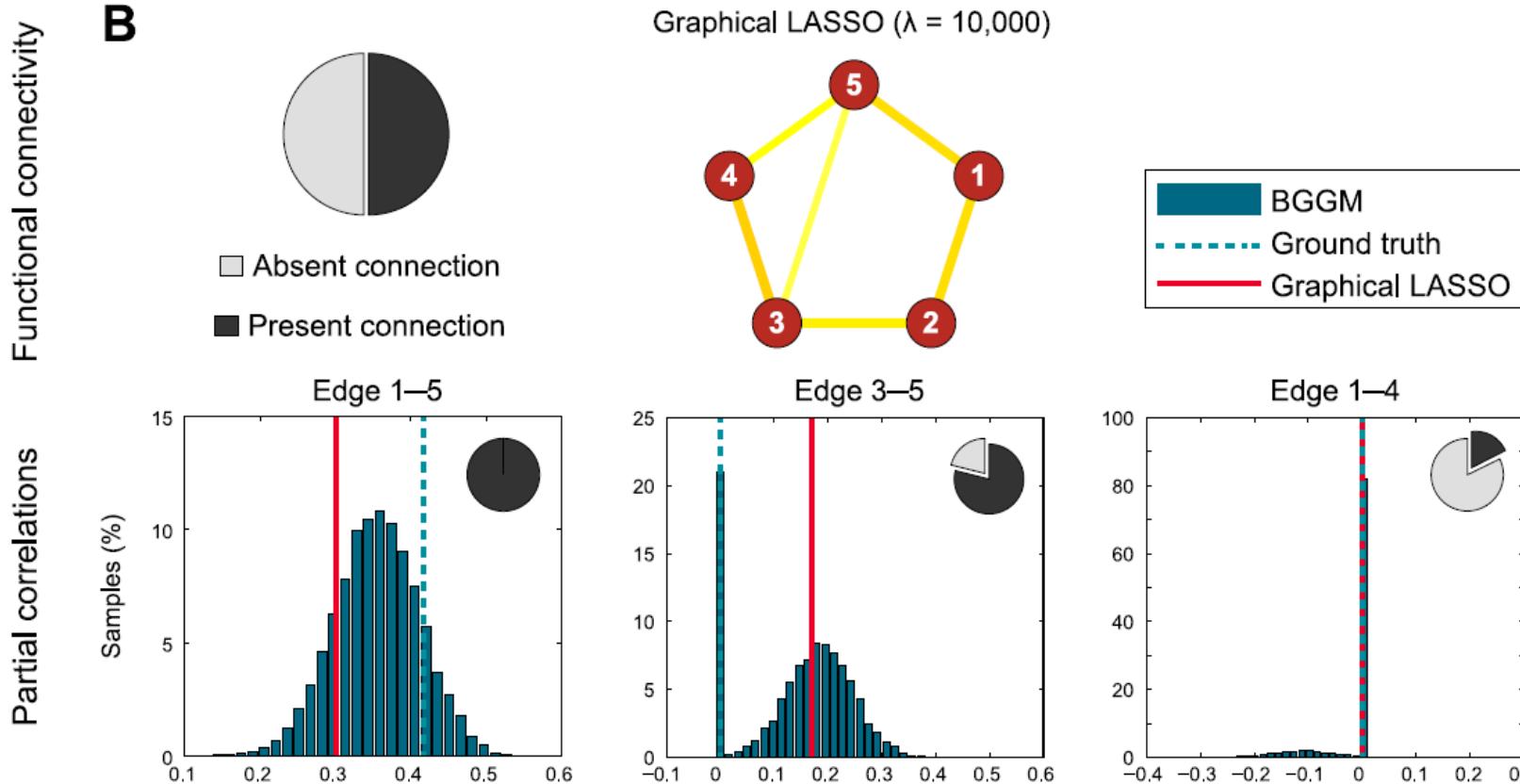
- We assumed \mathbf{G} was given.
 - Structural connectivity estimate was a useful starting point.
- Can we jointly estimate both **functional and structural** connectivity? E.g. can we obtain $P(\mathbf{G}, \mathbf{K} | \mathbf{Y}, \mathbf{D}, \nu)$?
- **Extremely hard** inference problem
 - Even computing acceptance proposals is intractable!
- Requires development of more efficient MCMC samplers.
- Recent work allows inference for small-scale problems.



Further comparison with graphical lasso



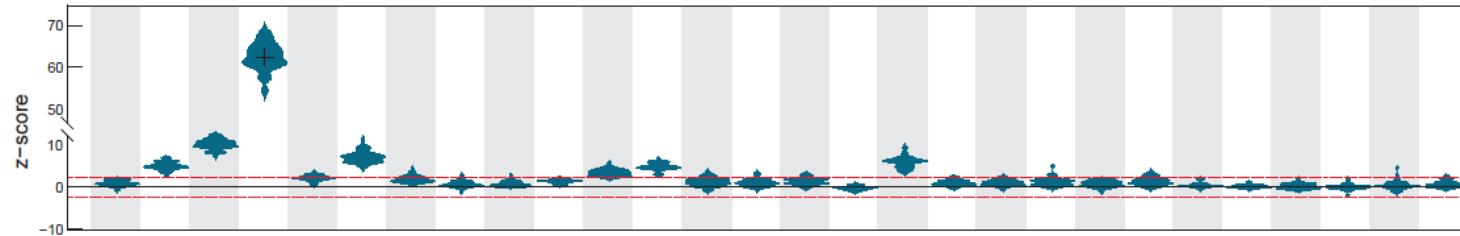
Further comparison with graphical lasso





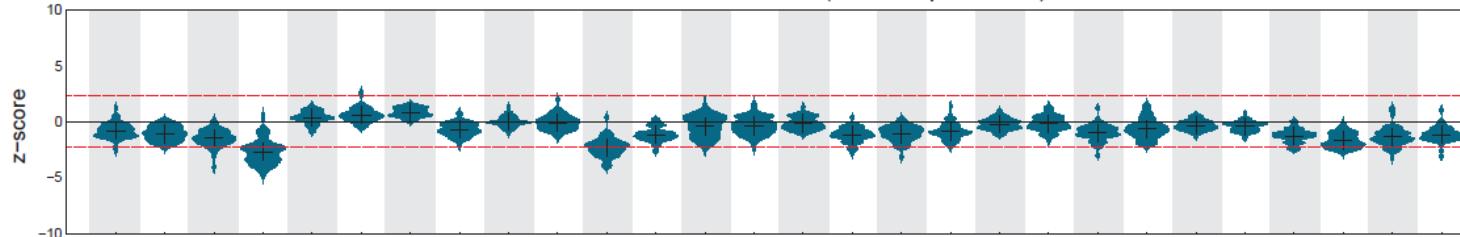
Further comparison with graphical lasso

Total distribution of error z-score



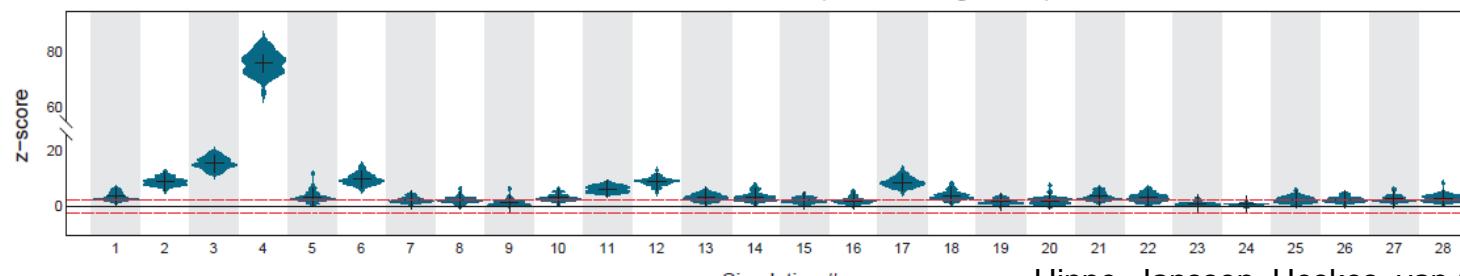
28 simulations with different properties (#nodes, noise levels, etc.)

Distribution of error z-score (for true positives)



Y-axis: z-score of G-lasso error vs errors of posterior distribution of our Bayesian approach.

Distribution of error z-score (for true negatives)



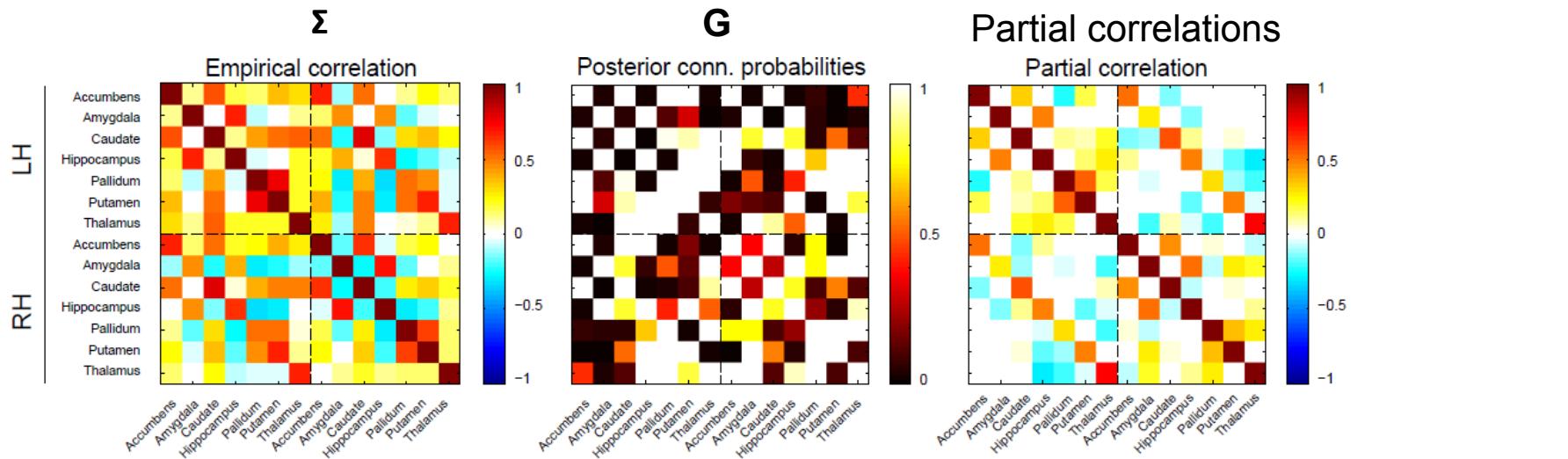
High z-score: G-lasso performs worse.

Simulation #

Hinne, Janssen, Heskes, van Gerven (2015), PLoS Comp

Joint inference of functional and structural connectivity

- Functional and structural connectivity using only functional data.
- 14 subcortical areas



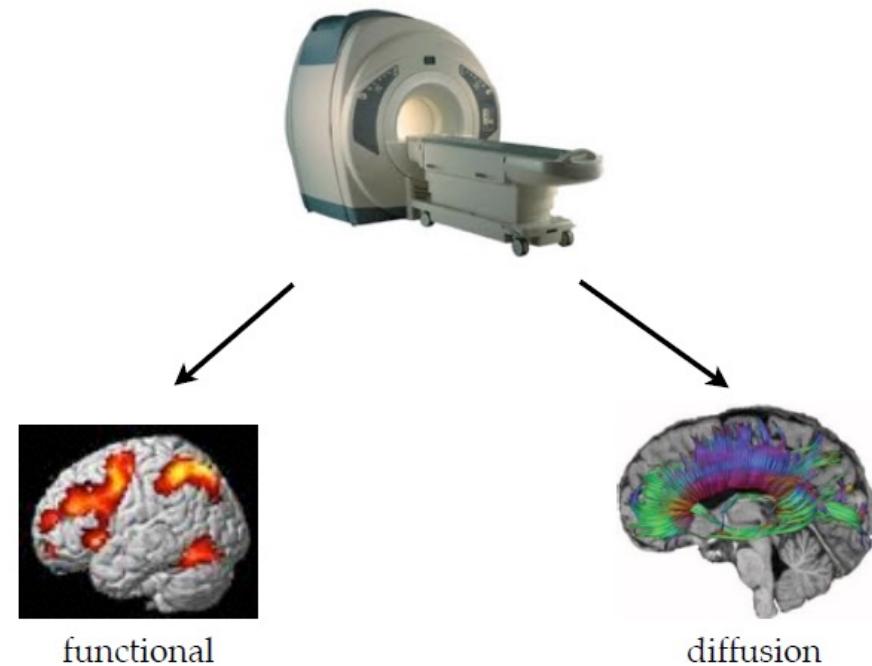
Bayesian data fusion

How do we combine different sources of information?



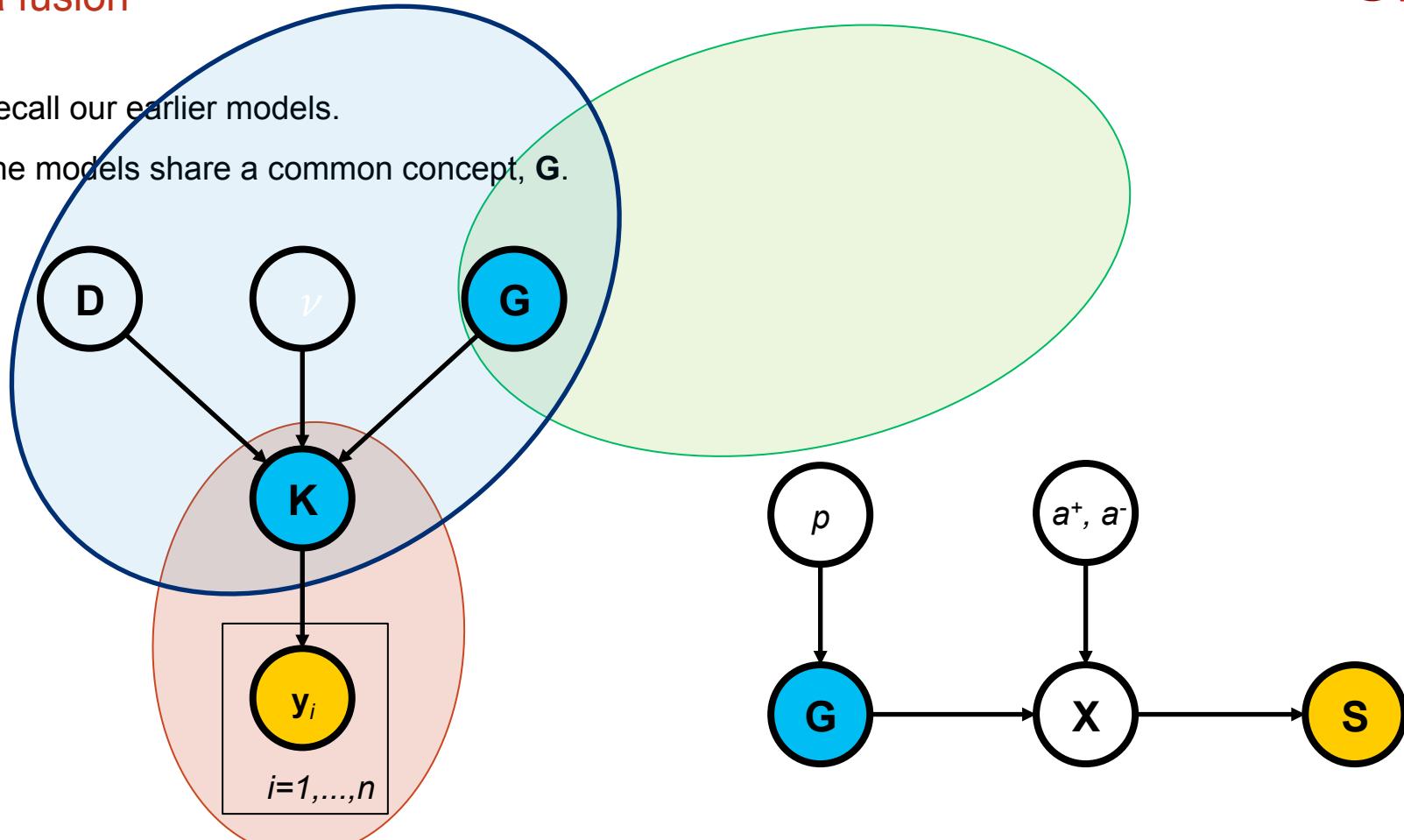
Data fusion

- Suppose we collect both resting-state fMRI and diffusion imaging data.
- How can we **integrate data** from multiple sources to get more reliable estimates of connectivity?
- This is a **data fusion** problem.



Data fusion

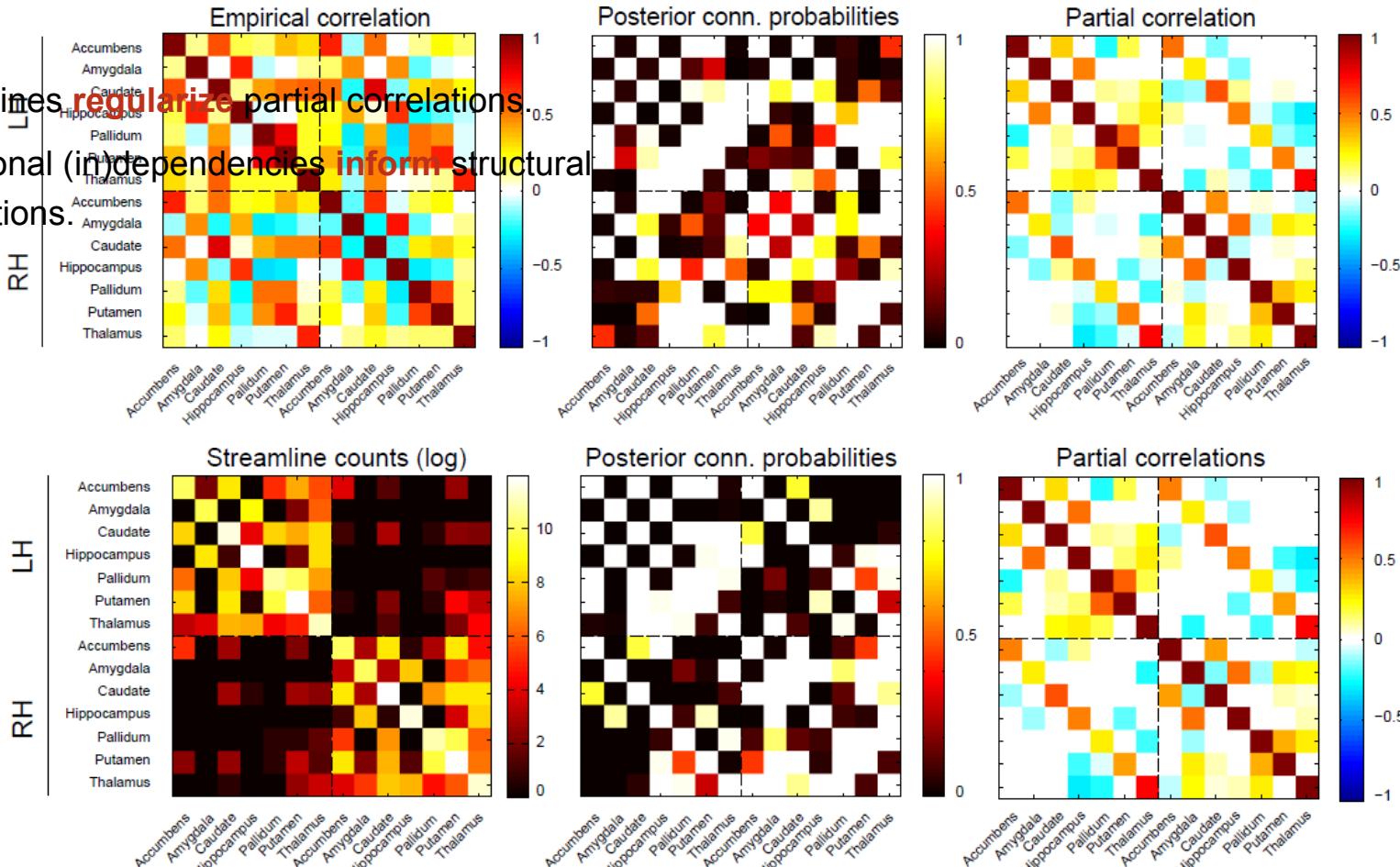
- Recall our earlier models.
- The models share a common concept, G .



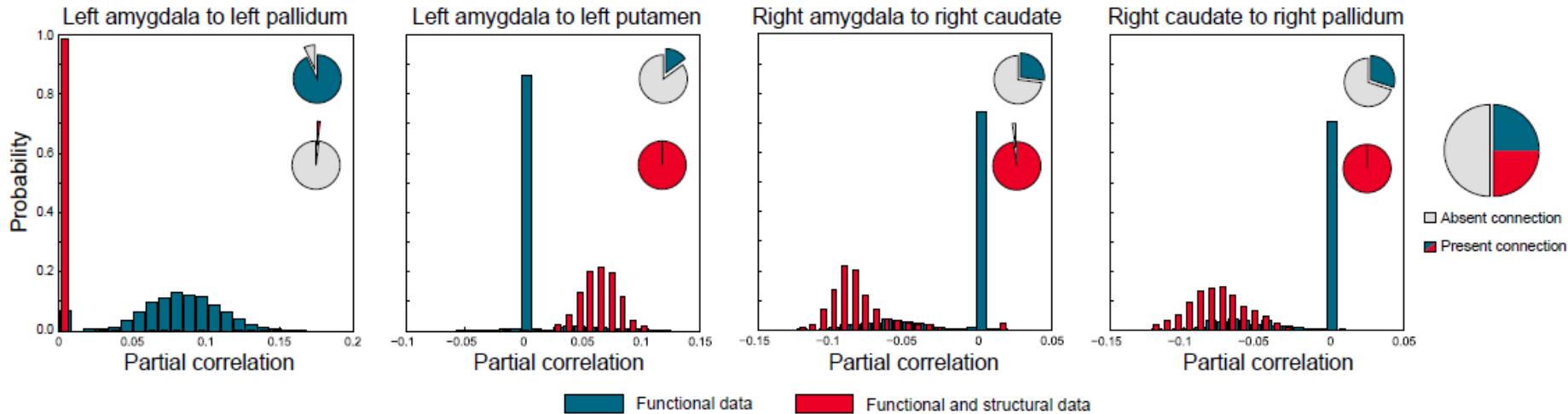


Data fusion

- Streamlines **regularize** partial correlations.
- Conditional (in)dependencies **inform** structural connections.



Data fusion



- Streamlining subcortical connectivity is prone to errors.
- Proof-of-concept shows value/applicability of data fusion.
- Bayesian approach allows **incorporating all available observations**.



Conclusions & future directions/challenges

Fundamental advantages

- Bayesian generative modeling gives an principled way of learning connectivity from neuroimaging data.
- Posterior distributions enable expression of uncertainty.
- Generative models can be combined in a plug-and-play fashion.

Practical challenges

- Large and complex models are difficult to sample.
- Application on large graphs requires development of more efficient approximate inference.

What lies ahead

- More realistic models.
- Dynamics of connectivity.
- Large-scale networks of microscopic connectivity.



Answers

Questions