Bayesian Graphical Models for Multiple Networks

Elin Shaddox

4 November 2019

Introduction

Overall Goal

- Infer multiple biological networks across multiple sample groups and data types
- Improve understanding of disease progression

Statistical Challenges

- Networks are complex, can be difficult to infer from small samples
- Computational burden is a major challenge for Bayesian graphical models

Proposed Methodology

Designed to improve scalability and reliability of inference

Outline

- Motivating Disease
- Background on Graphical Models
- Inference of Multiple Networks Across Multiple Sample Groups
- Inference of Multiple Networks For Sample Groups Across Multiple Data Platforms
- Inference of Joint Networks Using a Spiked Dirichlet Process Prior
- Concluding Remarks

Disease of Interest

Chronic Obstructive Pulmonary Disease (COPD)

- 3rd leading cause of death in the US [1]
- Acute Exacerbations of COPD are 2nd leading cause of hospital stays[2]
- 90% of COPD patients are smokers, but about 75% of smokers do not develop COPD
- Poor understanding of risk factors for disease susceptibility and disease progression

---

[1] National Center for Health Statistics, Health, United States, 2015: with special feature on racial and ethnic health disparities, 2016.

[2] Perera et al. (2012). Acute exacerbations of COPD in the United States: inpatient burden and predictors of costs and mortality. COPD: *J Chron Obstruct Pulmon Dis* **9**, 131-141.

COPD Data

Two Studies on Genetic Epidemiology of COPD

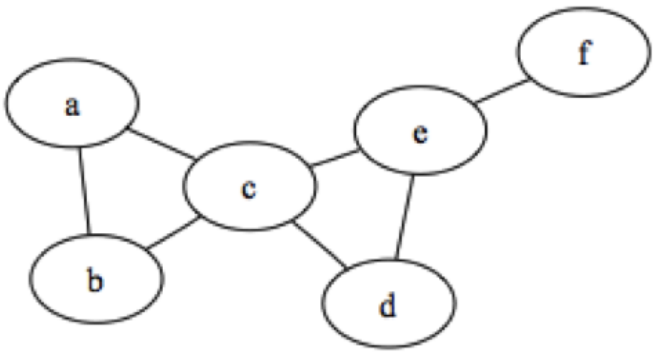
Analysis focuses on a single pathway of interest at a time

- There are p variables per pathway
- For each of the n subjects we have p expression levels or abundances
- Variables are corrected for age, sex, body mass index, and current smoking status
- The n subjects are allocated into K subgroups based on disease stage resulting in n_k subjects per group
- For each group we infer the network for the p variables

Graphical Models Framework

Undirected Graphical Models

- Use a graph structure to summarize conditional independence between variables
- Each node represents a variable
- No edge exists between two nodes if and only if variables are independent given all others



Fundamental assumption

- True conditional dependence structure is reasonably sparse

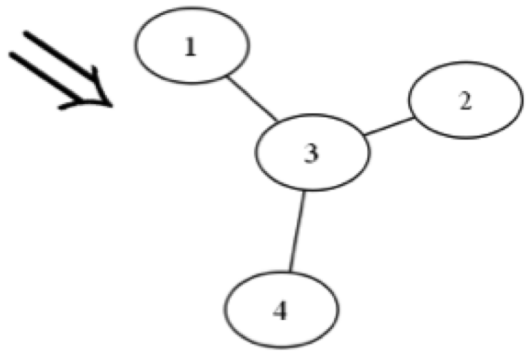
Gaussian Graphical Model (GGM)

When the variables of the undirected graphical model follow a joint multivariate normal distribution

$$x_l \sim \mathcal{N}(\mu, \Omega^{-1}) \quad l = 1 \ldots n$$

- $\omega_{i,j} = 0$ if no edge exists between genes i and j in graph G
- For simplicity, we column-center the data to assume $\mu = 0$

$$\begin{pmatrix} \omega_{11} & 0 & \omega_{13} & 0 \\ & \omega_{22} & \omega_{23} & 0 \\ & & \omega_{33} & \omega_{34} \\ & & & \omega_{44} \end{pmatrix}$$



Gaussian Graphical Model for Multiple Groups

$$x_{k,l} \sim \mathcal{N}(\mu_k, \Omega_k^{-1}) \quad l = 1, \ldots, n_k \quad k = 1, \ldots, K$$

- Each subgroup has a unique network
- $x_{k,l}$ gene expression levels for subject l in group k
- $\Omega_k = \Sigma_k^{-1} = (\omega_{i,j,k})$ is the precision matrix for group k

Common Approaches to Multiple Network Inference Across Sample Groups

Frequentist Penalized Methods

- Extensions to the graphical lasso [3]
- Fused graphical lasso encourages shared structure and edge values
- Group graphical lasso encourages shared structure
- Penalty parameter selection according to AIC

Bayesian Approaches

Commonly assign a G-Wishart prior on Ω_k [4]

- Provides flexible formulation for modeling
- Method is limited in scalability and computation

---

[3] Danaher, P., Wang, P., and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. B* **76**, 373-397.

[4] Peterson, C., Stingo, F., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association* **110**, 159 – 174.

Bayesian Inference of Joint Networks

Objective

Improve scalability

Statistical Approach

Defines a continuous shrinkage prior on Ω_k to address difficulties of G-Wishart approach [5]

Motivation

- Diseases are multi-level illnesses defined by changes at the cellular level
- Network-based inference can elucidate underlying pathogenic mechanisms influencing disease progression

Translational Impact

- Pinpoint molecular targets for further study and therapies

[5] Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis* **10**, 351-377.

Proposed Model for Multiple Network Inference

Builds on Features of the Bayesian Approach

- Markov Random Field (MRF) prior to encourage common network structure when supported by the data
- Spike and slab prior on network relatedness parameters to share information when appropriate
- **Continuous Shrinkage Prior**

$$p(\Omega_k|\mathbf{G}, v_0, v_1, \lambda) \propto \prod_{i<j} \mathcal{N}(\omega_{i,j}|0, v^2_{g_{ij}}) \prod_i \text{Exp}(\omega_{i,i}|\frac{\lambda}{2})$$

defined by introducing binary latent variables of edge-inclusion $\mathbf{G} \equiv (g_{i,j})_{i<j} \in \mathcal{G} \equiv \{0,1\}^{p(p-1)/2}$

Advantages

- Avoid use of priors over the space of positive definite matrices with fixed zeroes
- Improve computational efficiency

Linking Networks with a MRF Prior

Captures and models the dependence structure between binary latent variables of edge inclusion $\mathbf{g_{ij}} = \{g_{1ij}, \ldots, g_{kij}\}$

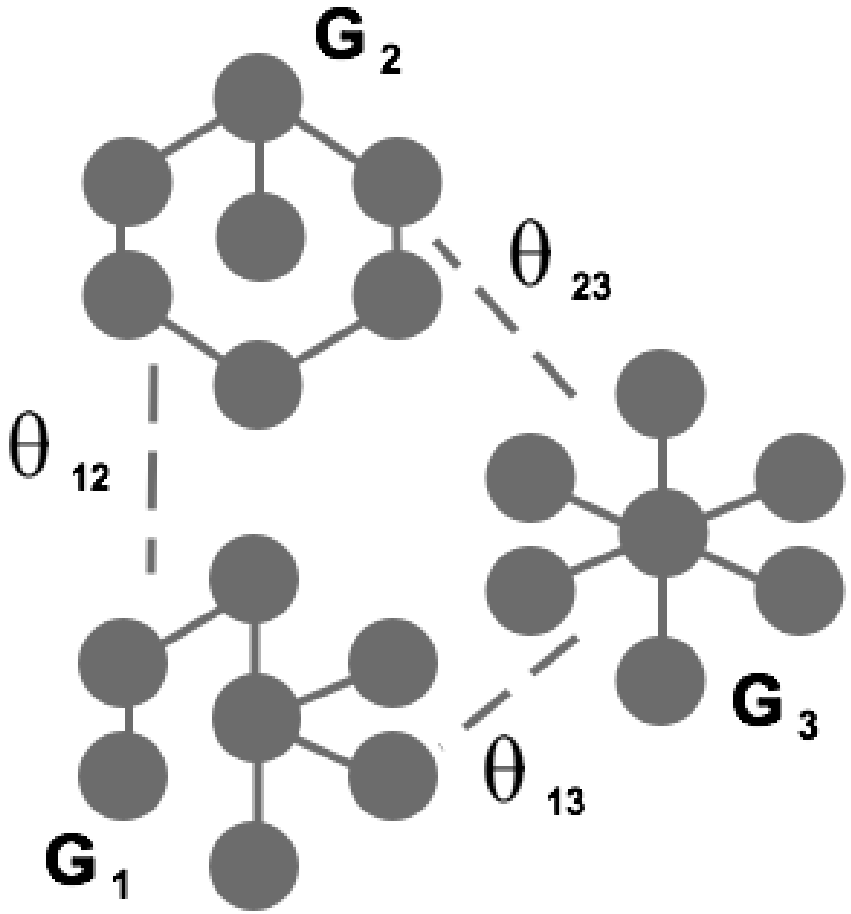
$$p(\mathbf{g_{ij}}|\nu_{ij}, \Theta) = \frac{\exp(\nu_{ij}\mathbf{1}^T\mathbf{g_{ij}} + \mathbf{g_{ij}}^T\Theta\mathbf{g_{ij}})}{C(\nu_{ij}, \Theta)}$$

- Encourages selection of common edges if subgroups are similar
- Conditional probability of edge inclusion

$$p(g_{kij}|\{g_{mij}\}_{m\neq k}, \nu_{ij}, \Theta) = \frac{\exp(g_{kij}(\nu_{ij} + 2\sum_{m\neq k}\theta_{km}g_{mij}))}{1 + \exp(\nu_{ij} + 2\sum_{m\neq k}\theta_{km}g_{mij})}$$

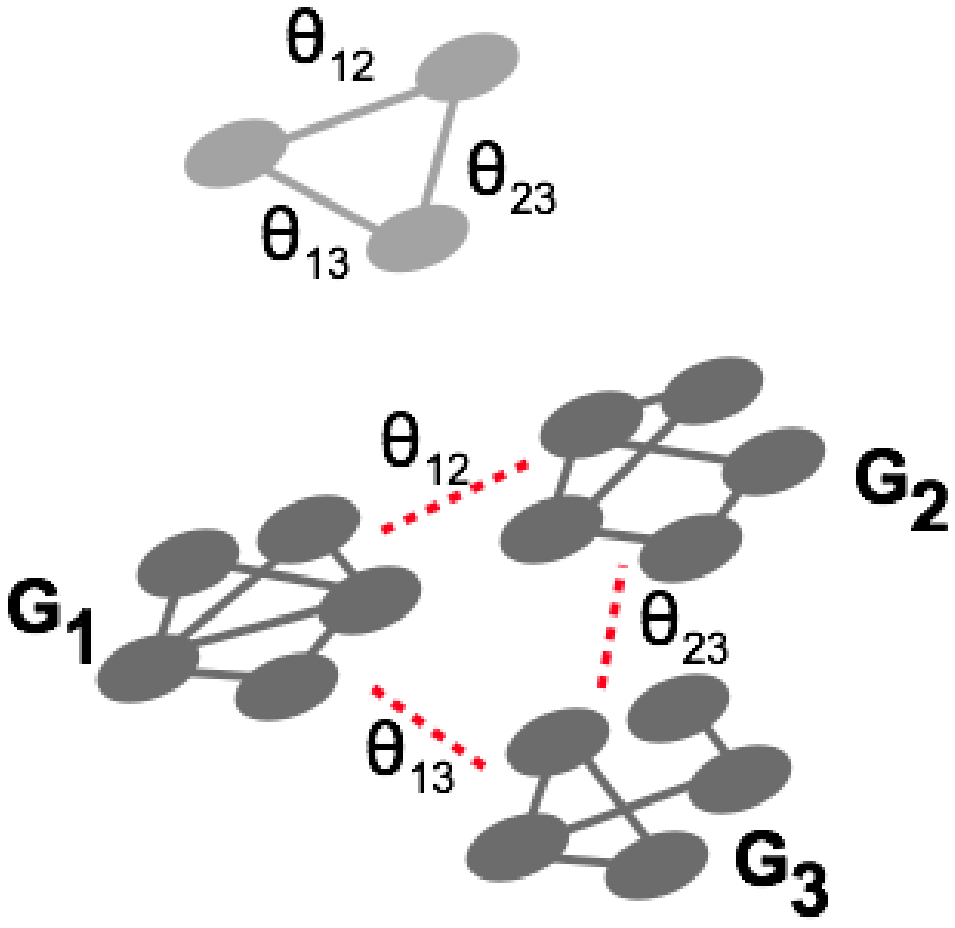
  - ν_{ij} is a specific parameter for each set of edges $\mathbf{g_{ij}}$
  - Θ is a $K \times K$ symmetric matrix denoting pairwise relatedness of graphs
  - Higher θ_{km} values promote common edges

Proposed Model for Multiple Network Inference



Gaussian graphical model for each sample group

Proposed Model for Multiple Network Inference



Imposed Network of Relative Cross-Group Similarity Measures

Selection Prior for Network Similarity

Spike and slab prior on off-diagonal entries of Θ

$$p(\theta_{km}|\gamma_{km}) = (1-\gamma_{km})\delta_0 + \gamma_{km}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta_{km}^{\alpha-1}\exp(-\beta\,\theta_{km})$$

- Fixed hyperparameters α and β
- Latent indicator variable γ_{km} of graph relatedness
- Independent Bernoulli priors defined on latent indicators with probability $w \in [0,1]$

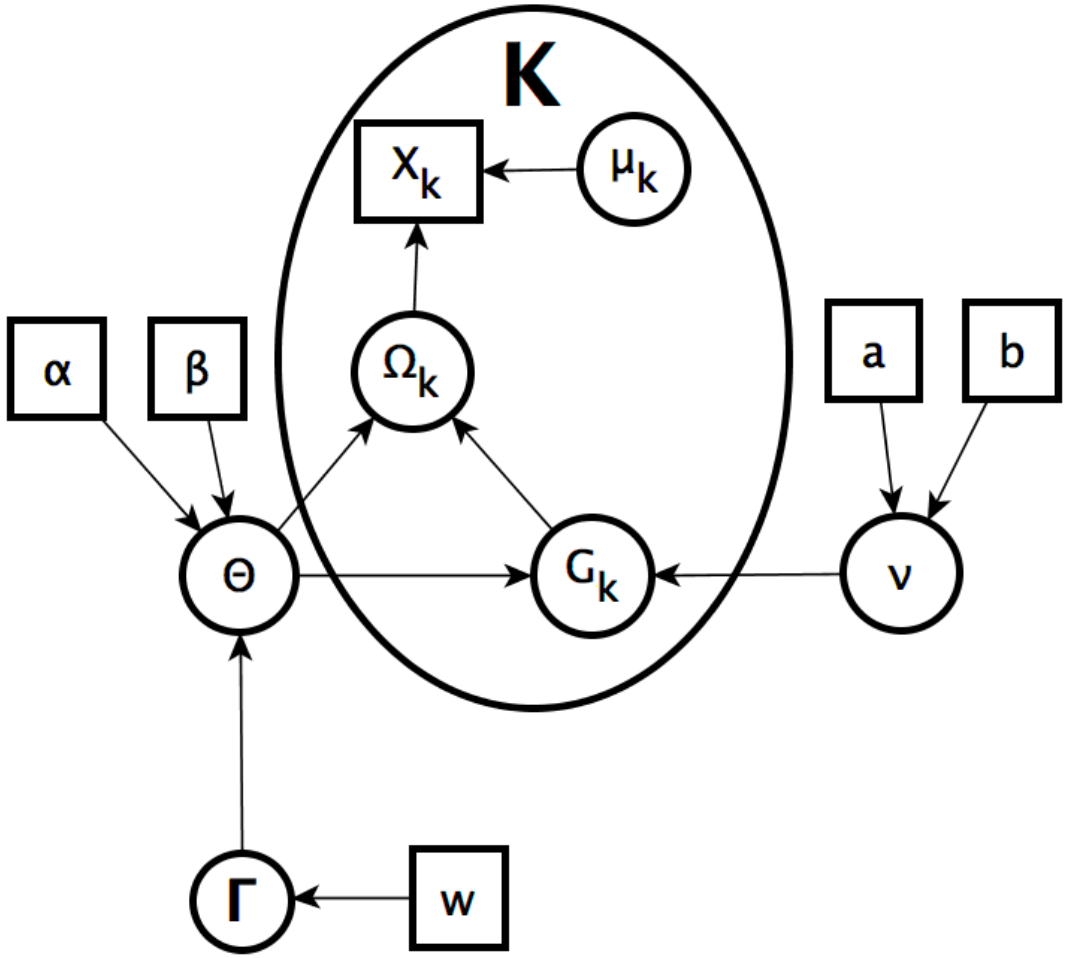
Completing the Model

Edge specific prior for edge inclusion probability

$$P(\nu_{ij}) = \frac{1}{\beta(a,b)} \frac{\exp(a\nu_{ij})}{(1+\exp(\nu_{ij}))^{a+b}}$$

- Can encourage sparsity of the graphs $G_1, \ldots, G_K$
- Can be used to incorporate prior knowledge of connections between genes

Diagram of the Joint Graph Model



Posterior Inference

MCMC Sampling Scheme

At each iteration t of our algorithm, sample from posterior full conditionals

(i) Update graph $G_k^{(t)}$ and precision matrix $\Omega_k^{(t)}$ for $k = 1, \ldots, K$

- Block Gibbs sampler
- p-coupled stochastic search variable selection algorithm

(ii) Update network relatedness parameters θ_{km} and γ_{km} for $1 \leq k < m \leq K$ using Metropolis-Hastings steps

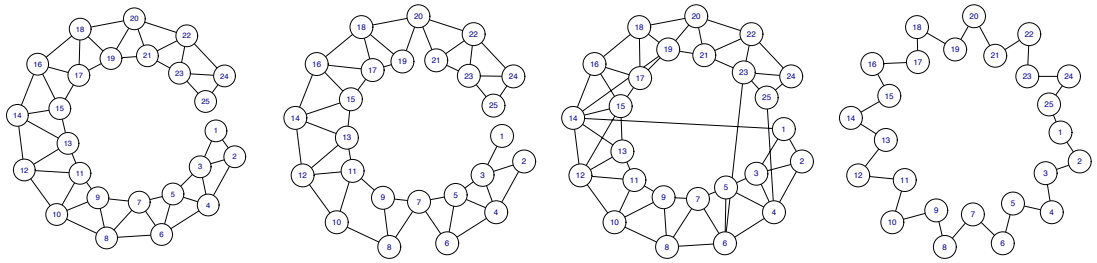
- Incorporates both between-model and within-model moves

(iii) Update v_{ij} for $1 \leq i < j \leq p$ using a standard Metropolis-Hastings step

Model Selection

- Posterior marginal probabilities of edge inclusion (MPP)

Simulation Studies to Assess Performance



Simulation Set Up

- Three scenarios considered: $p = 25$, $p = 50$, $p = 100$
- Network for Ω_1 is an AR(2) model, edges added or removed to get subsequent graphs
- Data generated from multivariate normal for each group

$$MPP(\Theta) = \begin{pmatrix} \cdot & .998 & .992 & .999 \\ & \cdot & .650 & .583 \\ & & \cdot & .516 \\ & & & \cdot \end{pmatrix}$$

Simulation Results

Performance Compared with Fused and Group Lasso

25-node Setting	**TPR**	**FPR**	**MCC**	**AUC**
Fused Lasso	0.96	0.47	0.34	0.90
Group Lasso	0.96	0.47	0.34	0.85
Proposed Method	0.58	0.001	0.68	0.95
50-node Setting	**TPR**	**FPR**	**MCC**	**AUC**
Fused Lasso	0.93	0.33	0.32	0.95
Group Lasso	0.93	0.33	0.32	0.88
Proposed Method	0.73	0.01	0.78	0.96
100-node setting	**TPR**	**FPR**	**MCC**	**AUC**
Fused Lasso	0.87	0.26	0.28	0.97
Group Lasso	0.85	0.17	0.31	0.90
Proposed Method	0.71	0.01	0.75	0.96

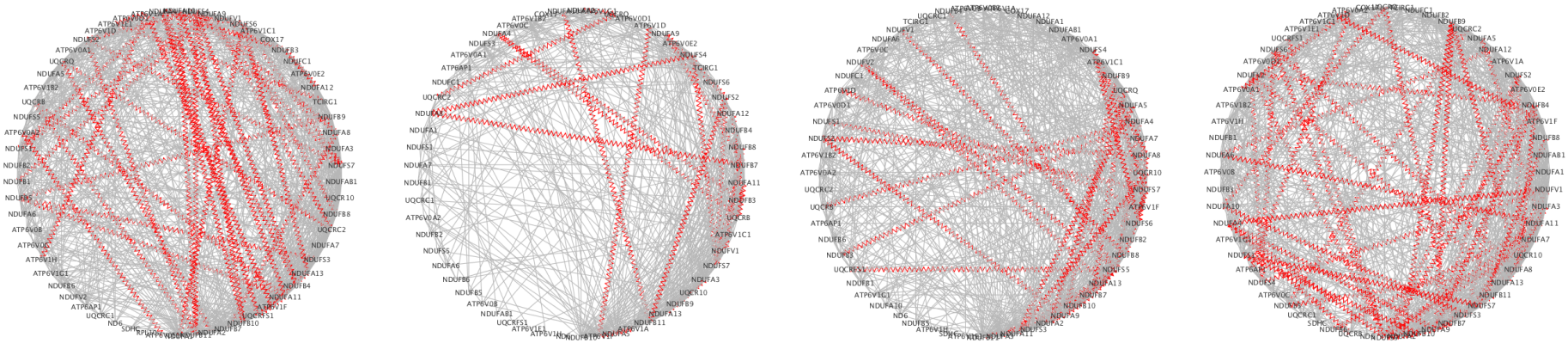
ECLIPSE COPD Cohort Study

- Whole-blood gene expression data was generated for 226 subjects using the Affymetrix Human Genome U133 Plus 2.0 Array available at NCBI GEO GSE22148
- Four candidate gene pathways were selected
  - Glycerophospholipid Metabolism (GPL)
  - Oxidative Phosphorylation (OxPhos)
  - Regulation of Autophagy (RegAuto)
  - FcγR Mediated Phagocytosis (FcyR)
- Subjects were classified into four groups by severity of radiologic emphysema with sample sizes 43 - 61

Goal: For each separate pathway, infer unique gene networks for all sample groups

Application Results

Estimated Networks for OxPhos pathway



Control Mild Moderate Severe

Red zig-zag edges denote known protein-protein interactions

Numbers of disease disrupted pairs

	Total Pairs	1000	1100	1110	0111	0011	0001	Total disrupted
GPL	539 (1)	58	26	28	21	40 (1)	59	232 (1)
FcγR	892 (50)	102 (8)	34 (3)	30 (1)	31	40 (1)	125 (9)	362 (22)
OxPhos	1072 (275)	127 (27)	37 (7)	25 (6)	23 (6)	62 (17)	120 (25)	394 (88)
RegAuto	153 (9)	13	2	11 (1)	8	4	11 (1)	49 (2)

Inference of Multiple Networks Across Multiple Data Platforms

Objective

- Improve inference of networks across multiple data platforms and samples

Statistical Approach

- Extension of Multiple Network Approach
- Bayesian prior construction to achieve joint estimation in a flexible manner
  - No assumptions on similarity of networks
  - No assumptions on directionality of influence

Motivation

- Medical study settings where heterogeneous subjects are profiled across multiple platforms

Translational Impact

- Pinpoint molecular targets for further study and therapies

Current Approaches for Multiple Platform Joint Network Inference

Separate inference for each platform

- Reduces statistical power
- Ignores potential commonalities

Existing methods for integrative analysis

- Model the association between different layers / types of variables
- Assume a direction of influence

Proposed model for multiple platform approach

Features of the Model

- MRF priors encourage
  - Common edges across related sample groups
  - Common cross-group relation across platforms
- Spike and slab prior on
  - Parameters measuring network relatedness
  - Parameters measuring platform relatedness

Flexibility of approach

- Number of subjects and variables may differ across data types
- Does not enforce similarity unless supported by the data
- Shares information when appropriate

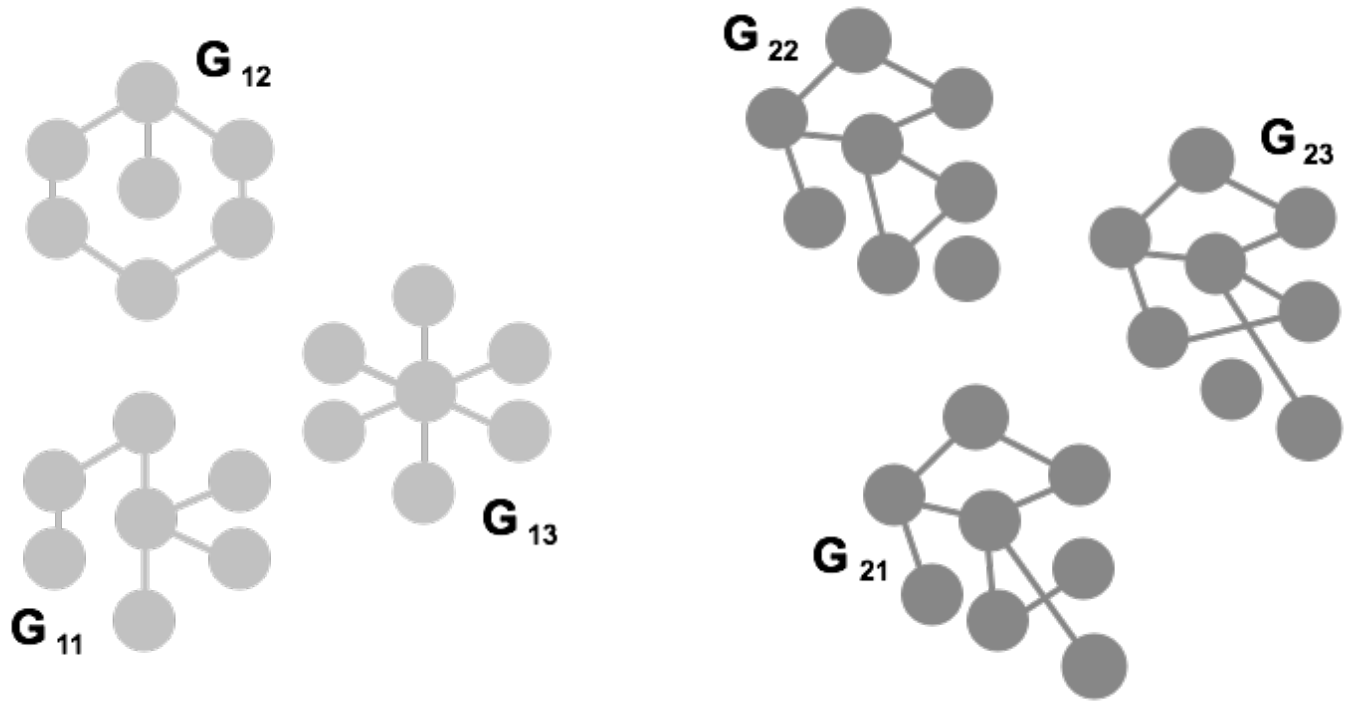
Gaussian Graphical Model for Multiple Groups

$$x_{s,k,l} \sim \mathcal{N}(\mu_{sk}, \Omega_{sk}^{-1})$$

$$l = 1, \ldots, n_{sk} \qquad k = 1, \ldots, K \qquad s = 1, \ldots, S$$

- For simplicity, we column-center the data for each group and assume $\mu_{sk} = 0$
- Each subgroup of each platform has a unique network
- $x_{s,k,l}$ platform levels for subject l in group k
- $\Omega_{sk} = \Sigma_{sk}^{-1} = (\omega_{ijsk})$ is the precision matrix for group k of platform s
- $\omega_{ijsk} = 0$ if no edge exists between genes i and j in graph G_{sk}

Graphical representation of model



Gaussian graphical model for each sample group and platform

$$x_{skl} \sim \mathcal{N}(\mu_{sk}, \Omega_{sk}^{-1}), \quad l = 1, \ldots, n_{sk}, \quad k = 1, \ldots, K, \quad s = 1, \ldots, S$$

Prior linking Networks across sample groups

where $\mathbf{g}_{sij} = \{g_{s1ij}, \ldots, g_{sKij}\}^T$ is a vector of binary edge inclusion indicators across K graphs for platform s, we construct MRF and spike-and-slab priors on $\mathbf{g}_{sij}$ and Θ_s as before

$$p(\mathbf{g}_{sij}|\nu_{sij}, \Theta_s) = \frac{\exp(\nu_{sij}\mathbf{1}^T\mathbf{g}_{sij} + \mathbf{g}_{sij}{}^T\Theta_s\mathbf{g}_{sij})}{C(\nu_{sij}, \Theta_s)},$$

- ν_{sij} is a sparsity parameter
- Θ_s is a $K \times K$ symmetric matrix of pairwise relatedness

$$P(\theta_{skm}|\gamma_{skm}) = (1-\gamma_{skm})\delta_0 + \gamma_{skm}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta_{skm}^{\alpha-1}e^{-\beta\theta_{skm}},$$

- $\Gamma(\cdot)$ represents the gamma function
- α and β are fixed hyper parameters
- γ_{skm} indicates related network event

Prior linking Networks across Platforms

Rather than independent Bernoulli priors on γ_{skm} variables, we construct a second MRF prior linking platforms; where $\gamma_{km} = \{\gamma_{1km}, \ldots, \gamma_{Skm}\}^T$ is a vector of binary indicators for network relatedness between subgroups across platforms

$$p(\gamma_{km}|w_{km}, \Phi) = C(w_{km}, \Phi)^{-1} \exp(w_{km}\mathbf{1}^T\gamma_{km} + \gamma_{km}^T\Phi\gamma_{km}),$$

- w_{km} is a sparsity parameter
- Φ_s is a $S \times S$ symmetric matrix of pairwise relatedness

$$p(\phi_{st}|\zeta_{st}) = (1-\zeta_{st})\delta_1 + \zeta_{st}\frac{\kappa^\eta}{\Gamma(\eta)}\phi_{st}^{\eta-1}e^{-\kappa\phi_{st}},$$

- κ and η are fixed hyper parameters
- ζ_{st} latent binary variable indicating related cross-platform dependency

Completing the Model

Continuous Shrinkage Prior on precision matrices Ω_{sk}

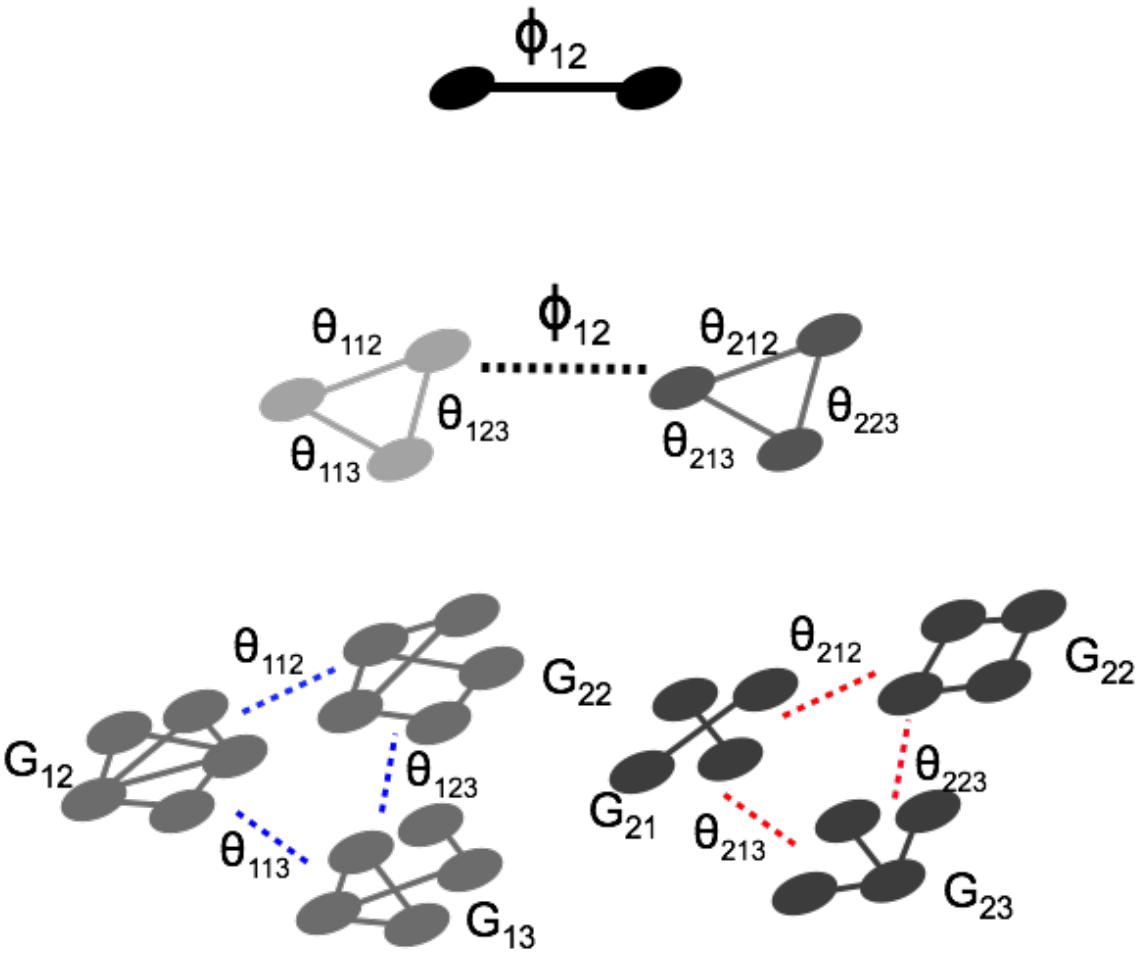
- Allows direct modeling of the latent graph $\mathbf{G}_{sk}$
- Computationally scalable

Beta priors on sparsity parameters v_{sij} and w_{km}

- Platform specific hyper parameters may be chosen in cases where sparsity is known to vary
- Can incorporate reference information on particular connections
- Small values encourage overall sparsity

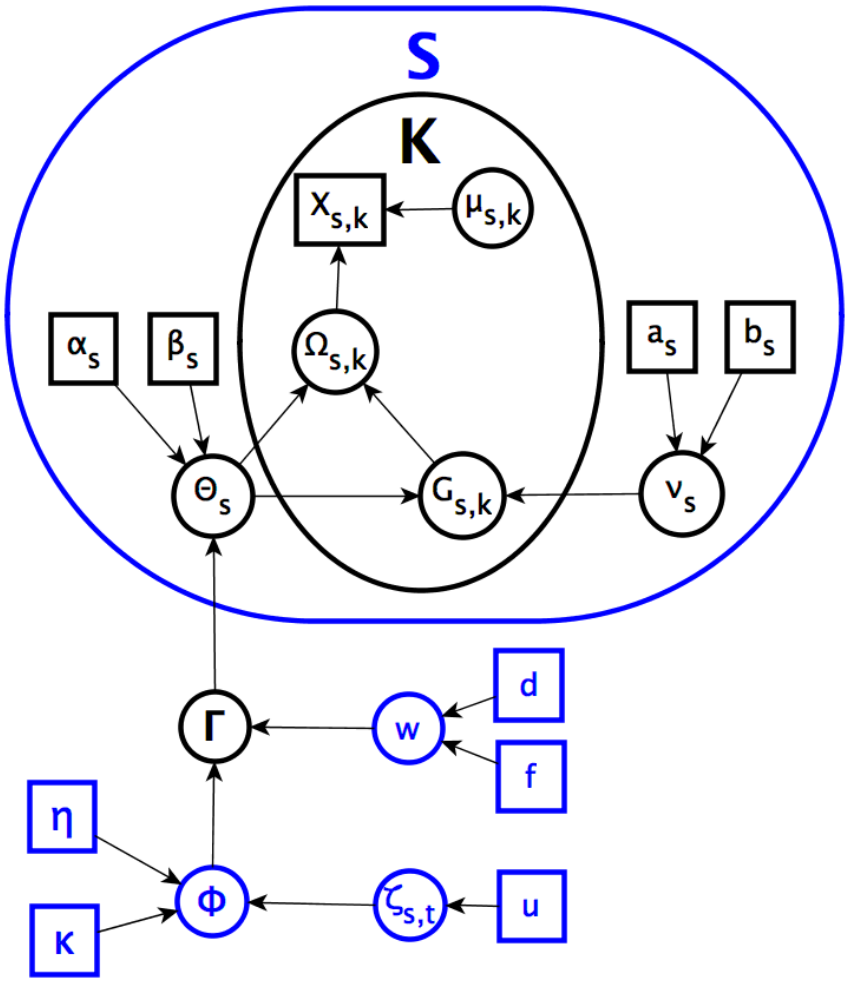
Independent Bernoulli(u) priors on latent indicators ζ_{st}

Graphical representation of model



Imposed networks of cross-group and cross-platform similarity measures for a 2-platform setting

Diagram of the Joint Platform Model



COPDGene Study

NIH funded multi center observational study

- Subjects 45-80 with at least a 10 pack year smoking history were recruited and biomarker measurements were attained from blood
- To understand genetic, clinical, and molecular factors that determine COPD development

Apportion subjects according to Global Initiative for Chronic Obstructive Lung Disease (GOLD)

- Control group (GOLD stage = 0), $n = 42$
- Mild or moderate group (GOLD stage = 1 or 2), $n_2 = 42$
- Severe group (GOLD stage = 3 or 4), $n_3 = 42$

COPDGene Study

Gene expression

- Measured from peripheral blood mononuclear cells
- Expression levels for 28 (RegAuto) and 104 (FcyR) probesets
- Collapsed to 20 and 58 unique genes

Metabolite Abundances

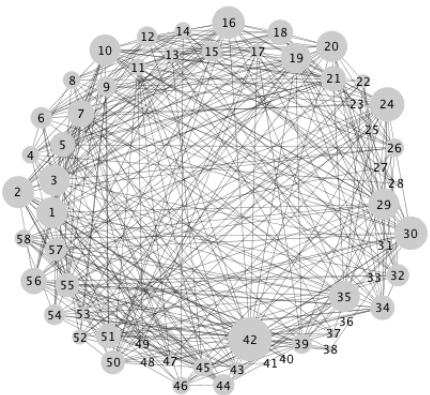
- Plasma metabolite abundances generated from liquid chromatography/mass spectrometry
- Matched to lipid and aqueous annotation to extract KEGG IDs
- Metabolite measurements for 117 (RegAuto) and 60 (FcyR) collapsed to 21 (RegAuto) and 23 (FcyR)

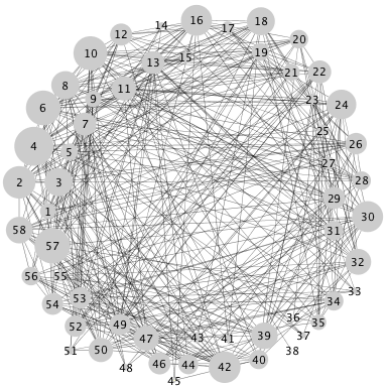
Goal: For each separate pathway, infer unique gene AND metabolite networks for all sample groups

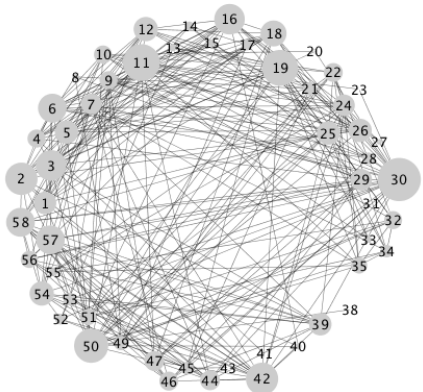
Case Study Conclusions

Results indicated a general preference for shared structure

$$\Phi_{FcyR} = .9771$$

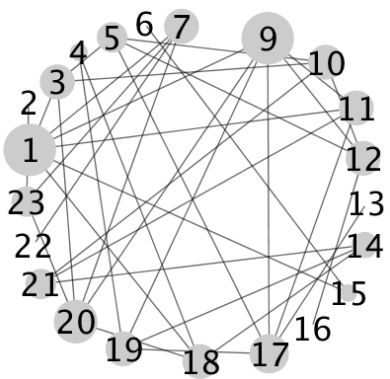




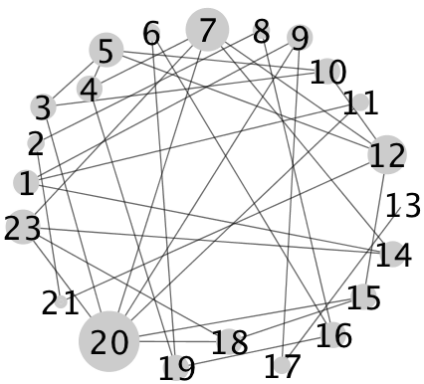


$$\Theta_G = \begin{pmatrix} \cdot & 1.0 & 1.0 \\ & \cdot & 1.0 \\ & & \cdot \end{pmatrix}$$

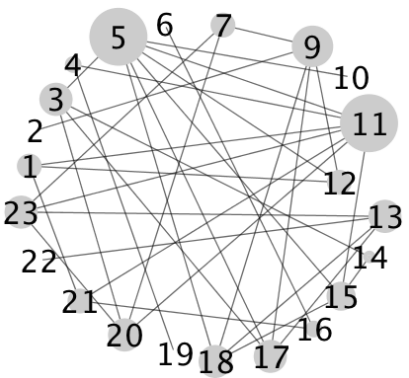
Control



Moderate



Severe



$$\Theta_M = \begin{pmatrix} \cdot & .95 & .97 \\ & \cdot & .95 \\ & & \cdot \end{pmatrix}$$

Inferred Network Structure Results

Numbers of disease disrupted pairs

Pathway	Platform	Total Pairs	100	110	011	001	Total Disrupted
FcγR	**Metabolites**	73	17	5	3	18	43
FcγR	**Genes**	656 (49)	151 (8)	63 (7)	74 (3)	63 (4)	351 (22)
Reg Auto	**Metabolites**	66	14	4	5	17	40
Reg Auto	**Genes**	101 (6)	23 (2)	13	7	8	51 (2)

Network Characteristics

	METABOLITES			GENES		
FcγR Pathway	**Control**	**Moderate**	**Severe**	**Control**	**Moderate**	**Severe**
Number of edges	59	57	58	405	444	332
Global clustering	0.1665	0.2430	0.1683	0.4268	0.4495	0.4442
Betweenness centrality	0.2122	0.2783	0.3348	0.0771	0.0483	0.0995
Count of hub nodes	12	5	10	50	53	46
Reg Auto Pathway	**Control**	**Moderate**	**Severe**	**Control**	**Moderate**	**Severe**
Number of edges	49	51	54	71	76	49
Global clustering	0.0881	0.2143	0.1003	0.4649	0.5175	0.4123
Betweenness centrality	0.1524	0.21117	0.1862	0.1800	0.1205	0.1435
Count of hub nodes	9	8	6	14	15	8

A Spiked Dirichlet Process Prior for Joint Network Inference

Objective

- Improve inference of networks by incorporating edge dependent cross-group similarity measures

Statistical Approach

- Extension of Multiple Network Approach
  - No assumptions on similarity of networks
- Dirichlet Process (DP) prior construction to
  - Reduce over-parameterization of the model
  - Cluster edge-specific similarity measures

Proposed model for DP prior approach

Features of the Model

- MRF priors encourage
  - Common edges across related sample groups
- Spike and slab prior on parameters measuring network relatedness
  - Dirichlet process defined as the slab portion

Advantages of approach

- Does not enforce similarity unless supported by the data
- Shares information when appropriate at the edge-level rather than group-level
- Parameters of the model space are reduced by clustering selected similarity measures

Proposed model for DP prior approach

Build on Features of the Multiple Network Approach

- Markov Random Field (MRF) prior defined on the vector of binary edge inclusion indicators for each edge (i,j)

$$p(\mathbf{g_{ij}}|\nu_{ij},\Theta_{ij}) = \frac{\exp(\nu_{ij}\mathbf{1}^T\mathbf{g_{ij}} + \mathbf{g_{ij}}^T\Theta_{ij}\mathbf{g_{ij}})}{C(\nu_{ij},\Theta_{ij})}$$

where Θ_{ij} corresponds to the $\frac{p(p-1)}{2}$ $K \times K$ symmetric matrices for edge specific pairwise relatedness across graphs

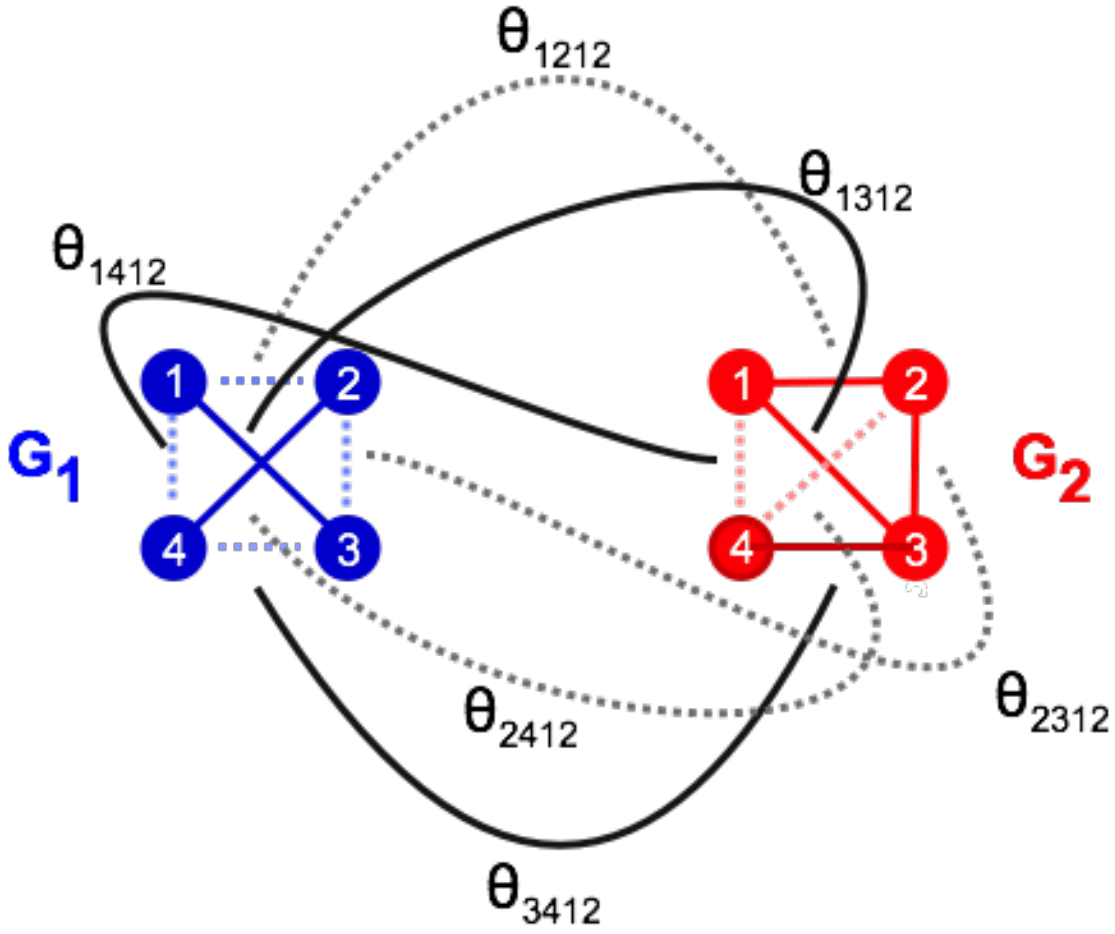
Selection Prior for Edge-Specific Similarity

Spike and slab prior on off-diagonal entries of Θ_{ij}

$$P(\theta_{ijkm}|\gamma_{ijkm}, H) = (1-\gamma_{ijkm})\delta_0 + \gamma_{ijkm}H$$
$$H \sim DP(\zeta, H_0)$$
$$H_0 \sim Gamma(\alpha, \beta)$$

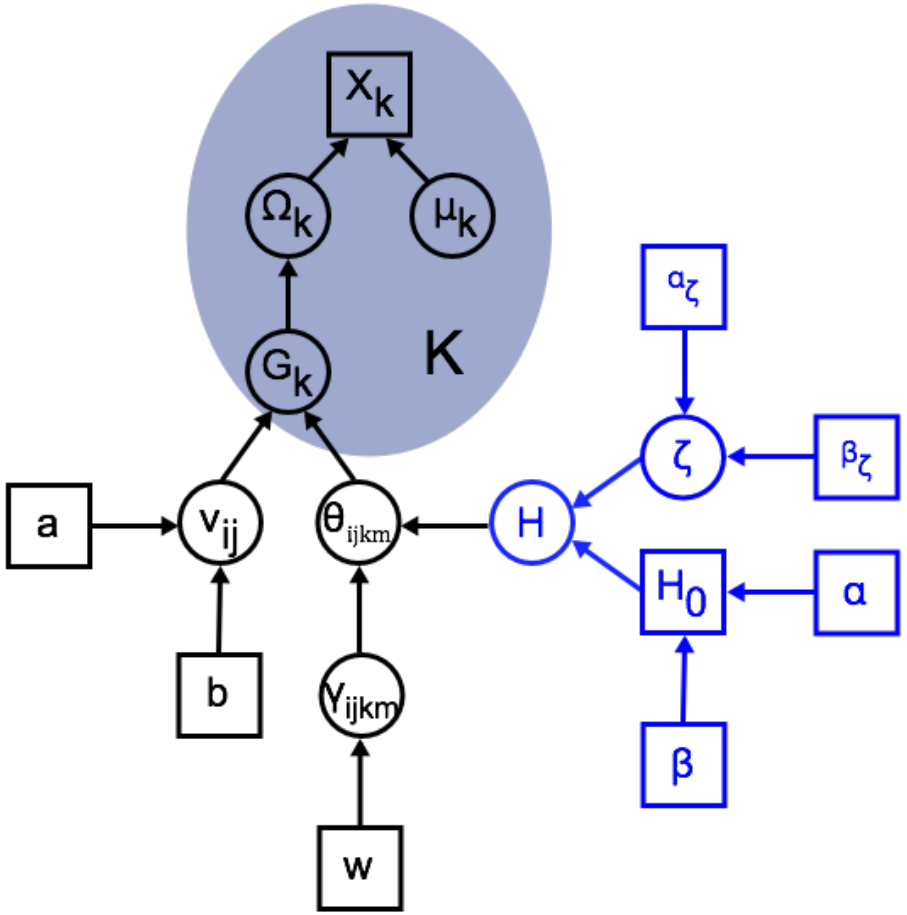
- Latent indicator γ_{ijkm} for edge-specific graph relatedness
  - Independent Bernoulli priors defined on latent indicators with fixed probability $w \in [0,1]$
- DP H defined by Gamma base measure with fixed hyperparameters α and β, and concentration parameter ζ
  - Concentration parameter defined by Gamma prior with fixed hyperparameters α_ζ and β_ζ

Graphical representation of model



Edge-specific similarity measures for a 2-subgroup setting

Diagram of the DP Prior Joint Model



Current Explorations of the spiked DP approach

- Investigating simulation set-up and how to incorporate clustering across networks
- Extracting the optimal clustering structure from the posterior distribution
- Examining patterns in network structure that are learned via DP clustering
- Investigating sensitivity analysis of hyper parameter settings

Discussion

Statistical Impact of Method Development

- Improved reliability of network inference

Advantages of Approaches

- Improved scalability of Bayesian joint network inference
- Integrated platform approach takes advantage of commonalities, yet can also highlight platform specific differences
- Explore edge-specific similarity clustering patterns

Translational Impact

- Improving knowledge of the relationship between metabolites and genes
- Pinpointing notable interactions from control and disease stages of COPD

Acknowledgements

This work was supported by NLM grant T15 LM007093 and NIH grant 5T32-CA096520-07.

Thanks to my advisors, mentors, and collaborators

- Dr. Marina Vannucci
- Dr. Christine B. Peterson
- Dr. Francesco Stingo
- Dr. Nicola Hanania
- Dr. Katerina Kechris
- Dr. Russell Bowler
- Dr. Charmion Cruickshank-Quinn

References

Shaddox, E., Peterson, C., Stingo, F., Hanania, N., Cruickshank-Quinn, C., Kechris, K., Bowler, R., and Vannucci, M. (2018). Bayesian Inference of Networks Across Multiple Sample Groups and Data Types. *Biostatistics* Accepted.

Shaddox, E., Stingo, F., Peterson, C., Jacobson, S., Cruickshank-Quinn, C., Kechris, K., Bowler, R., and Vannucci, M. (2018). A Bayesian approach for learning gene networks underlying disease severity in COPD. *Statistics in Biosciences* 10(1):59-85.