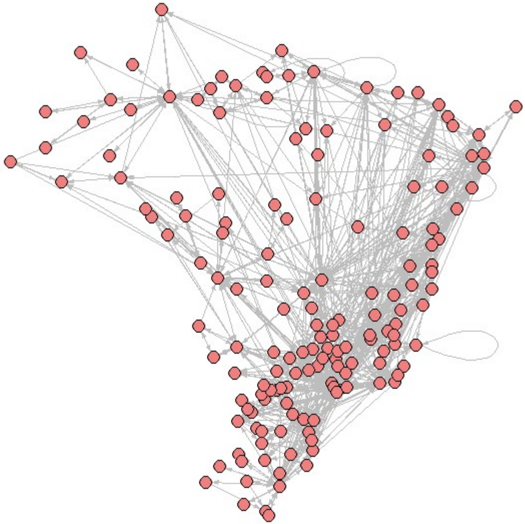
Community detection in weighted networks

Andressa Cerqueira

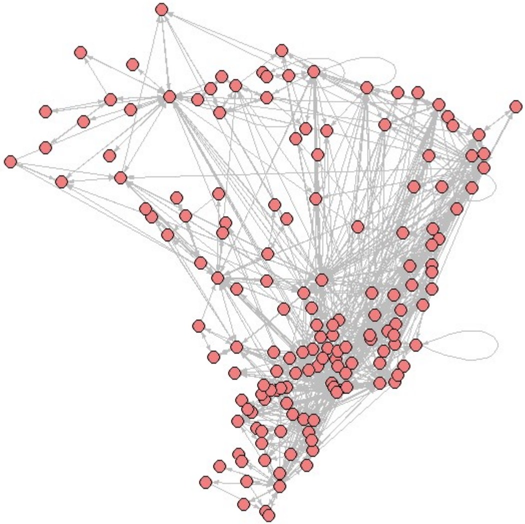
Department of Statistics
Universidade Federal de São Carlos, UFSCar

May 17, 2023

Motivation: Graphs/Networks



Motivation: Graphs/Networks

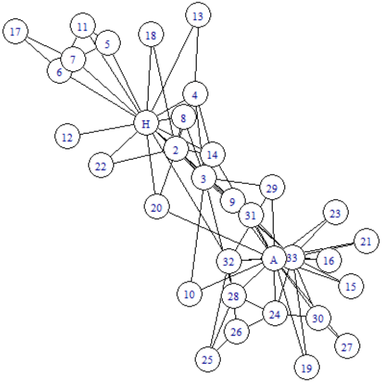


Statistics/ Probability

- statistical model
- infer the parameters
- hypothesis testing
- clustering nodes
- probabilistic model
- study asymptotic properties of the model
- study dynamics on networks

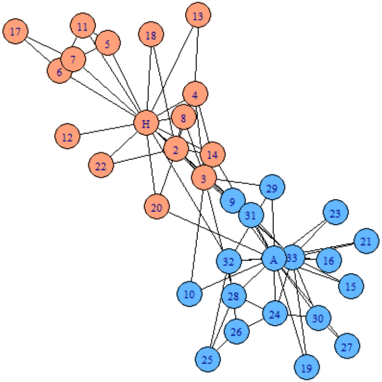
Motivation: Community detection

- Zachary's karate club: social relationship between 34 members of a karate club

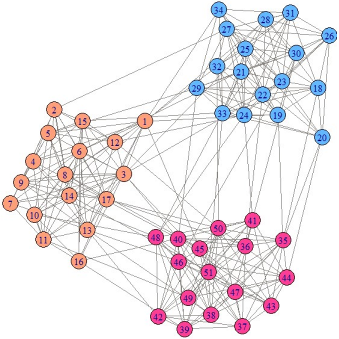


Motivation: Community detection

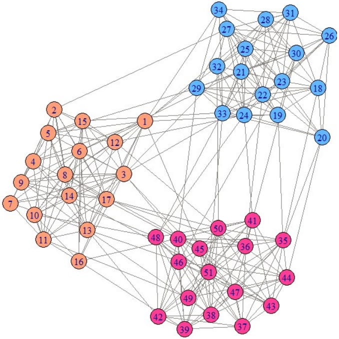
- Zachary's karate club: social relationship between 34 members of a karate club



Motivation: Community detection

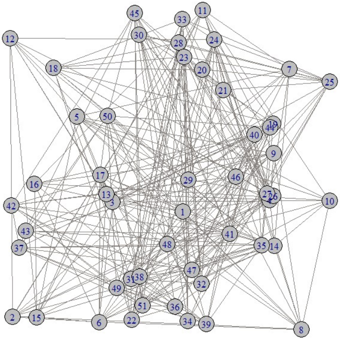


Motivation: Community detection

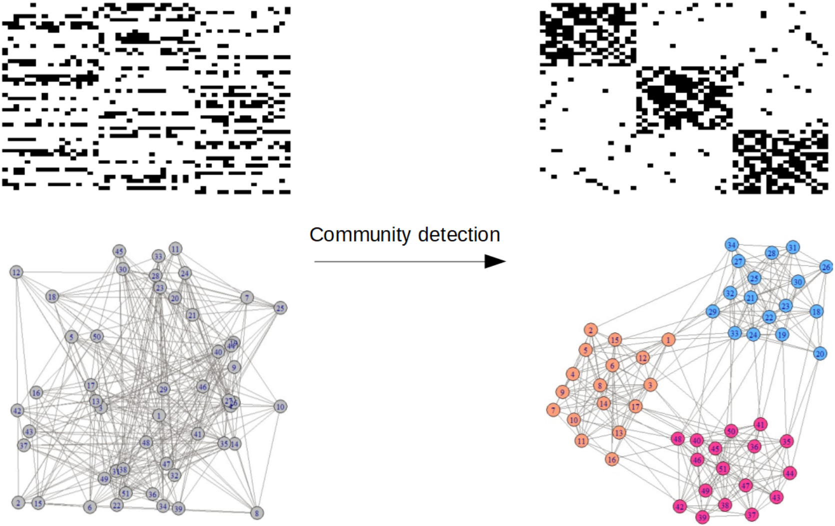


Motivation: Community detection

Motivation: Community detection



Motivation: Community detection



Stochastic Block Models

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. Social networks, 5(2), 109-137.

Latent variables: $C_1, C_2, \ldots, C_n$ i.i.d. with $\mathbb{P}(C_i = a) = \pi_a$, $a = 1, \ldots, K$ and $\pi = (\pi_1, \ldots, \pi_K)$.
K is known.

Stochastic Block Models

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. Social networks, 5(2), 109-137.

Latent variables: $C_1, C_2, \ldots, C_n$ i.i.d. with $\mathbb{P}(C_i = a) = \pi_a$, $a = 1, \ldots, K$ and $\pi = (\pi_1, \ldots, \pi_K)$.
K is known.
Observed variables: $\mathbf{A}$ adjacent matrix of the graph with n vertices.
$A_{ij} \in \{0, 1\}$ and

$$A_{ij}|(C_i = a, C_j = b) \sim Bernoulli(P_{a,b})$$

P is a symmetric matrix $K \times K$.

Let $(\mathbf{A}, \mathbf{c})$ be a sample of the SBM with K blocks and parameter $\theta = (\pi, P)$.

Let $(\mathbf{A}, \mathbf{c})$ be a sample of the SBM with K blocks and parameter $\theta = (\pi, P)$.

Problems:

1. Estimation of nodes' labels $\mathbf{c} = (c_1, \cdots, c_n)$ (Community detection);

Let $(\mathbf{A}, \mathbf{c})$ be a sample of the SBM with K blocks and parameter $\theta = (\pi, P)$.

Problems:

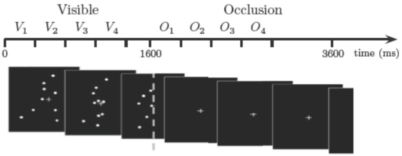
1. Estimation of nodes' labels $\mathbf{c} = (c_1, \cdots, c_n)$ (Community detection);
2. Estimation of the parameter $\theta = (\pi, P)$;

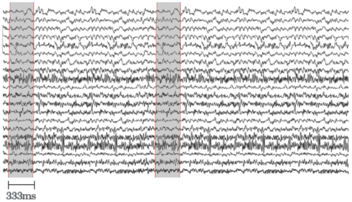
Let $(\mathbf{A}, \mathbf{c})$ be a sample of the SBM with K blocks and parameter $\theta = (\pi, P)$.

Problems:

1. Estimation of nodes' labels $\mathbf{c} = (c_1, \cdots, c_n)$ (Community detection);
2. Estimation of the parameter $\theta = (\pi, P)$;
3. Estimation of the number of blocks K (Model selection problem).

Motivation: weighted networks





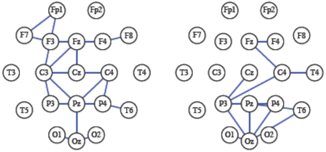
stimulus → EEG Signal

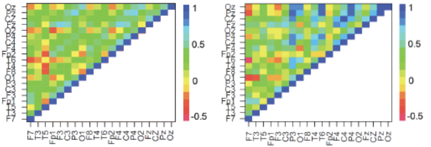
interaction criterion

correlation matrix

network criterion

graph





Motivation: weighted networks

- Number of flights between two airports in an air transportation network
- Brain connectivity networks with edge weights measured as Fisher-transformed Pearson correlations between brain regions

Weighted Stochastic Block Models

Network with n vertices
Latent variables: $C_1, C_2, \ldots, C_n$ i.i.d. with $\mathbb{P}(C_i = a) = \pi_a$, $a = 1, \ldots, K$ and $\pi = (\pi_1, \ldots, \pi_K)$.
K is known.

Weighted Stochastic Block Models

Network with n vertices
Latent variables: $C_1, C_2, \ldots, C_n$ i.i.d. with $\mathbb{P}(C_i = a) = \pi_a$, $a = 1, \ldots, K$ and $\pi = (\pi_1, \ldots, \pi_K)$.
K is known.

Observed variables: Edge-weighted network $\mathbf{W}$ such that

$$W_{ij} \,|\, \mathbf{C} = \mathbf{c} \sim N(B_{c_i c_j}, \Sigma_{c_i c_j}), \qquad 1 \le i < j \le n$$
$$W_{ii} = 0, \qquad i = 1, \ldots, n.$$

where $B \in \mathbb{R}^{K \times K}$ and $\Sigma \in \mathbb{R}_+^{K \times K}$.
$\mathbf{W}$ is a symmetric $n \times n$ matrix.

Weighted Stochastic Block Models

Network with n vertices
Latent variables: $C_1, C_2, \ldots, C_n$ i.i.d. with $\mathbb{P}(C_i = a) = \pi_a$, $a = 1, \ldots, K$ and $\pi = (\pi_1, \ldots, \pi_K)$.
K is known.

Observed variables: Edge-weighted network $\mathbf{W}$ such that

$$W_{ij} \,|\, \mathbf{C} = \mathbf{c} \sim N(B_{c_i c_j}, \Sigma_{c_i c_j}), \qquad 1 \leq i < j \leq n$$
$$W_{ii} = 0, \qquad i = 1, \ldots, n.$$

where $B \in \mathbb{R}^{K \times K}$ and $\Sigma \in \mathbb{R}_+^{K \times K}$.
$\mathbf{W}$ is a symmetric $n \times n$ matrix.

Goal: Recover the labels of the nodes using the observed weighted network $\mathbf{W}$.

Weighted Stochastic Block Models: Likelihood

The likelihood is given by

$$L(\pi, B, \Sigma; \mathbf{w}) = \sum_{\mathbf{c} \in \{1, \cdots, K\}^n} p(\mathbf{c}|\pi) p(\mathbf{w}|\mathbf{c}, B, \Sigma)$$

and it is not tractable.

Aicher, C., Jacobs, A. Z., & Clauset, A. (2014). Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2), 221-248.

- Instead of using the variables W_{ij}, $1 \leq i < j \leq n$ we use the variables s_{ik}, $1 \leq i \leq n$ and $1 \leq k \leq K$.
- For any label's vector $e = (e_1, \ldots, e_n)$, define

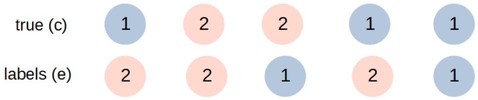
$$s_{ik}(e) = \sum_{j=1}^{n} W_{ij} \mathbb{1}\{e_j = k\}.$$

true (c)	1	2	2	1	1
labels (e)	2	2	1	2	1

i=2	$N(B_{21}, \Sigma_{21})$	$N(B_{22}, \Sigma_{22})$	$N(B_{22}, \Sigma_{22})$	$N(B_{21}, \Sigma_{21})$	$N(B_{21}, \Sigma_{21})$

- Instead of using the variables W_{ij}, $1 \leq i < j \leq n$ we use the variables s_{ik}, $1 \leq i \leq n$ and $1 \leq k \leq K$.
- For any label's vector $e = (e_1, \ldots, e_n)$, define

$$s_{ik}(e) = \sum_{j=1}^{n} W_{ij} \mathbb{1}\{e_j = k\}.$$



i=2	$N(B_{21}, \Sigma_{21})$	$N(B_{22}, \Sigma_{22})$	$N(B_{22}, \Sigma_{22})$	$N(B_{21}, \Sigma_{21})$	$N(B_{21}, \Sigma_{21})$

- Instead of using the variables W_{ij}, $1 \leq i < j \leq n$ we use the variables s_{ik}, $1 \leq i \leq n$ and $1 \leq k \leq K$.

- For any label's vector $e = (e_1, \ldots, e_n)$, define

$$s_{ik}(e) = \sum_{j=1}^{n} W_{ij} \mathbb{1}\{e_j = k\}.$$

- Let $\mathbf{s}_i(e) = (s_{i1}(e), \ldots, s_{iK}(e))$.
- Given $\mathbf{c}$, $\{s_{i1}(e), \ldots, s_{iK}(e)\}$ are mutually independent random variables.
- $\mathbf{s}_i$ and $\mathbf{s}_j$ are not independent

- Let R be the $K \times K$ confusion matrix with (k, l)-th entry given by

$$R_{kl} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{e_i = k, c_i = l\} .$$

- Let R be the $K \times K$ confusion matrix with (k, l)-th entry given by

$$R_{kl} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{e_i = k, c_i = l\}.$$

- For each node i, conditioned on the labels $\mathbf{c} = (c_1, \ldots, c_n)$ with $c_i = l$, we have that

$$s_{ik}(e) \sim N(P_{lk}, \Lambda_{lk})$$

where $P_{lk} = nR_k.B._l$ and $\Lambda_{lk} = nR_k.\Sigma._l$.

- Let R be the $K \times K$ confusion matrix with (k, l)-th entry given by

$$R_{kl} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{e_i = k, c_i = l\}.$$

- For each node i, conditioned on the labels $\mathbf{c} = (c_1, \ldots, c_n)$ with $c_i = l$, we have that

$$s_{ik}(e) \sim N(P_{lk}, \Lambda_{lk})$$

where $P_{lk} = nR_{k\cdot}B_{\cdot l}$ and $\Lambda_{lk} = nR_{k\cdot}\Sigma_{\cdot l}$.

-

$$\begin{aligned} p(\mathbf{s_i}|\pi, P, \Lambda) &= \sum_{l=1}^{K} \pi_l \, p(\mathbf{s_i}|P, \Lambda, c_i = l) \\ &= \sum_{l=1}^{K} \pi_l \prod_{k=1}^{K} p(s_{ik}|P_{lk}, \Lambda_{lk}, c_i = l) \end{aligned}$$

Pseudo-likelihood

We can write the log pseudo-likelihood (up to a constant) as

$$\ell_{PL}(\pi, P, \Lambda; \{\mathbf{s_i}\}) = \sum_{i=1}^{n} \log \left(\sum_{l=1}^{K} \pi_l \prod_{k=1}^{K} \frac{1}{\sqrt{\Lambda_{lk}}} \exp\left\{ \frac{-(s_{ik}(e) - P_{lk})^2}{2\Lambda_{lk}} \right\} \right)$$

We use EM algorithm (Expectation-Maximization)

Pseudo-likelihood - EM algorithm

Input: Initial labeling e.
Output: Estimate $\widehat{c}$
Repeat T times:

1. Compute $\widehat{\pi}_l = \dfrac{n_l(e)}{n}$, $\widehat{R} = diag(\widehat{\pi}_1, \ldots, \widehat{\pi}_K)$, $\widehat{P}_{lk} = n\widehat{R}_k.\widehat{B}._l$ and $\widehat{\Lambda}_{lk} = n\widehat{R}_k.\widehat{\Sigma}._l$
2. Compute the block sums $\mathbf{s}_1, \cdots, \mathbf{s}_n$.
    3. Estimate the probabilities for node labels by $\widehat{\pi}_{il} = \mathbb{P}_{_{PL}}(c_i = l|\mathbf{s}_i)$.
    4. Update the parameters values $\widehat{\pi}_l$, $\widehat{P}_{lk}$ and $\widehat{\Lambda}_{lk}$.
    5. Return to step (3) until convergence.
6. Update the labels by $e_i = \arg\max_{l=1} \widehat{\pi}_{il}$ and return to (1).
7. Return $\widehat{c} = e$.

Pseudo-likelihood algorithm - EM algorithm

Input: Initial labeling e.

Output: Estimate $\widehat{c}$

Repeat T times:

1 Compute $\widehat{\pi}_l = \dfrac{n_l(e)}{n}$, $\widehat{R} = diag(\widehat{\pi}_1, \ldots, \widehat{\pi}_K)$, $\widehat{P}_{lk} = n\widehat{R}_k.\widehat{B}._l$ and $\widehat{\Lambda}_{lk} = n\widehat{R}_k.\widehat{\Sigma}._l$

2 Compute the block sums $\mathbf{s}_1, \cdots, \mathbf{s}_n$.

3 Estimate the probabilities for node labels by $\widehat{\pi}_{il} = \mathbb{P}_{PL}(c_i = l|\mathbf{s}_i)$.

4 Update the parameters values $\widehat{\pi}_l$, $\widehat{P}_{lk}$ and $\widehat{\Lambda}_{lk}$.

5 Return to step (3) until convergence.

6 Update the labels by $e_i = \arg\max_{l=1} \widehat{\pi}_{il}$ and return to (1).

7 Return $\widehat{c} = e$.

Consistency results

How does the choice of the initial labeling e affect $\widehat{c}(e)$ in one step of the algorithm?

Consistency results

How does the choice of the initial labeling e affect $\widehat{c}(e)$ in one step of the algorithm?

- The initial labeling e matches c on a fixed (but unknown) fraction of nodes γ;

Consistency results

How does the choice of the initial labeling e affect $\widehat{c}(e)$ in one step of the algorithm?

- The initial labeling e matches c on a fixed (but unknown) fraction of nodes γ;
- Overall error:

$$L(\widehat{c}, c) = \min_{\phi \in \Phi_K} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\widehat{c}_i \neq \phi(c_i)\},$$

Consistency results

How does the choice of the initial labeling e affect $\widehat{c}(e)$ in one step of the algorithm?

- The initial labeling e matches c on a fixed (but unknown) fraction of nodes γ;
- Overall error:

$$L(\widehat{c}, c) = \min_{\phi \in \Phi_K} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\widehat{c}_i \neq \phi(c_i)\},$$

- Consider the mean matrix B and the variance matrix Σ as

$$B_{kl} = \begin{cases} \mu_1, & \text{if } k = l \\ \mu_2, & \text{if } k \neq l \end{cases} \qquad \text{and} \qquad \Sigma_{kl} = \sigma^2$$

Consistency results

- Cerqueira, A., & Levina, E. (2023). A pseudo-likelihood approach to community detection in weighted networks. arXiv preprint arXiv:2303.05909.

Theorem

Assume that $\pi_1 = \cdots = \pi_K = \frac{1}{K}$. *Consider the initial labeling* $e \in \mathcal{E}_\gamma$.*For* $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma^2 > 0$ *and* $\gamma \in (0,1)$, $\gamma \neq \frac{1}{K}$, *we have that for any* $\epsilon > 0$

$$\mathbb{E}\left[\sup_{\widehat{\mu_1}, \widehat{\mu_2}, \widehat{\sigma}^2 \in \widehat{P}_{\mu_1,\mu_2,\sigma^2}} L(\widehat{c}(e), c)\right] \leq (K-1)\exp\left\{-\frac{1}{4}\frac{(\gamma K - 1)^2}{K(K-1)^2}\frac{n(\mu_1-\mu_2)^2}{\sigma^2}\right\}.$$

Consistency results

- Xu, M., Jog, V., & Loh, P. L. (2020). Optimal rates for community estimation in the weighted stochastic block model. The Annals of Statistics, 48(1), 183-204.
- Optimal rate of misclustering error

$$\mathbb{E}L(\widehat{c}, c) \geq \exp\left(-(1 + o(1))\frac{n}{K}\frac{(\mu_1 - \mu_2)^2}{4\sigma^2}\right).$$

Consistency results

- Xu, M., Jog, V., & Loh, P. L. (2020). Optimal rates for community estimation in the weighted stochastic block model. The Annals of Statistics, 48(1), 183-204.
- Optimal rate of misclustering error

$$\mathbb{E}L(\widehat{c}, c) \geq \exp\left(-(1 + o(1))\frac{n}{K}\frac{(\mu_1 - \mu_2)^2}{4\sigma^2}\right) .$$

$$n\,\frac{(\mu_{1n} - \mu_{2n})^2}{\sigma_n^2} \to \infty \qquad \text{when} \qquad n \to \infty \, .$$

Consistency results

- Xu, M., Jog, V., & Loh, P. L. (2020). Optimal rates for community estimation in the weighted stochastic block model. The Annals of Statistics, 48(1), 183-204.
- Optimal rate of misclustering error

$$\mathbb{E}L(\widehat{c}, c) \geq \exp\left(-(1+o(1))\frac{n}{K}\frac{(\mu_1-\mu_2)^2}{4\sigma^2}\right).$$

$$n\,\frac{(\mu_{1n}-\mu_{2n})^2}{\sigma_n^2} \to \infty \qquad \text{when} \qquad n \to \infty\,.$$

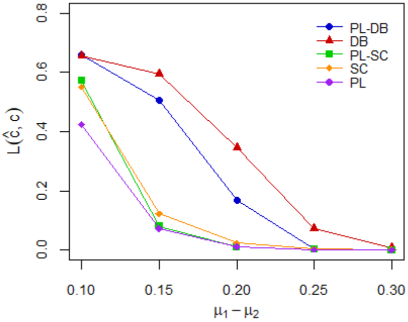
Remark: *Unbalanced case*

$$\text{If } \frac{n}{\sigma_n^2} \to \infty \quad \text{in such a way that} \quad \frac{n\,(\mu_{1n}-\mu_{2n})^2}{\sigma_n^2} \to \infty$$

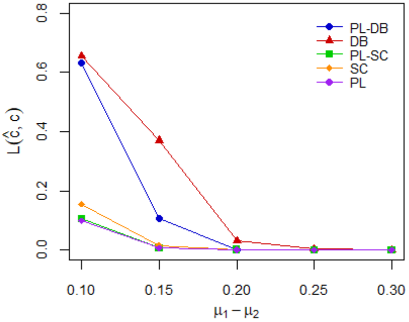
Simulations:

Setting:

$$B = \begin{bmatrix} \mu_1 & \mu_2 & \mu_2 \\ \mu_2 & \mu_1 & \mu_2 \\ \mu_2 & \mu_2 & \mu_1 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{bmatrix}.$$



(a) $n = 500$ and $\pi = (1/3, 1/3, 1/3)$

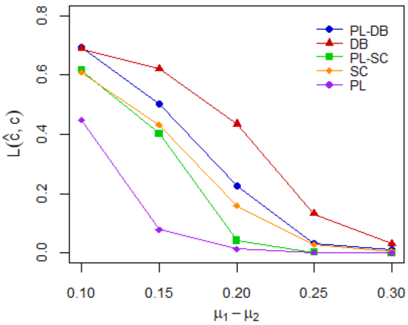


(b) $n = 1000$ and $\pi = (1/3, 1/3, 1/3)$

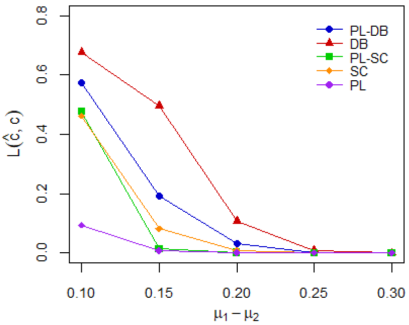
Simulations:

Setting:

$$B = \begin{bmatrix} \mu_1 & \mu_2 & \mu_2 \\ \mu_2 & \mu_1 & \mu_2 \\ \mu_2 & \mu_2 & \mu_1 \end{bmatrix} \qquad \textit{and} \qquad \Sigma = \begin{bmatrix} 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{bmatrix}.$$



(a) $n = 500$ and $\pi = (0.2, 0.5, 0.3)$



(b) $n = 1000$ and $\pi = (0.2, 0.5, 0.3)$

Application:

- Data consisting of resting state fMRI brain images of 54 schizophrenic patients and 69 healthy patients
- Total of 264 nodes
- The edge weights represent functional connectivity between the brain regions, measured by Fisher-transformed correlations between the time series of blood oxygenation levels at the corresponding regions
- We average the 69 weighted networks corresponding to healthy patients using the weighted network average method of Levin et al. (2022)

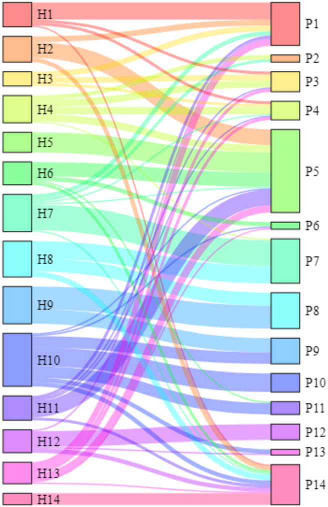
Application:

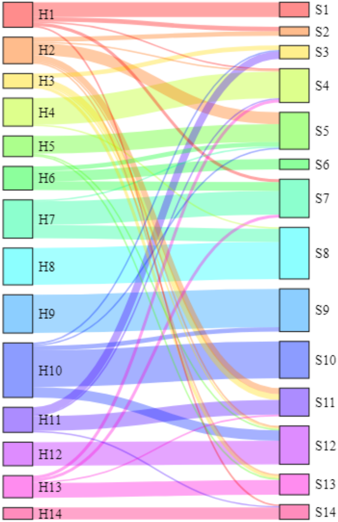
- Power parcellation

Region	Function	Nodes	Region	Function	Nodes
P1	Sensory/somatomotor Hand	30	P8	Fronto-pariental Task Control	25
P2	Sensory/somatomotor Mouth	5	P9	Salience	18
P3	Cingulo-opercular Task Control	14	P10	Subcortical	13
P4	Auditory	13	P11	Ventral attention	9
P5	Default mode	58	P12	Dorsal attention	11
P6	Memory retrieval	5	P13	Cerebellar	4
P7	Visual	31	P14	Uncertain	28

- We estimated 14 communities for both populations by the PL algorithm using SC as the initial value

Application:





Final Remarks:

- We considered homogenous models, where all within-communities edge distributions are the same, and between-communities edge distributions are also the same;
- We could incorporate edge distributions with a point mass at zero;

Thank you