# A Stochastic Block Model for Multilevel Networks

Saint-Clair Chabert-Liddell

Joint work with S. Donnet and P. Barbillon

UMR INRAe/AgroParisTech, MIA Paris

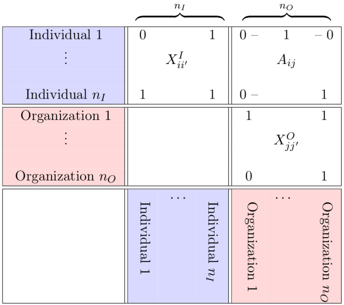
Séminaire des doctorants du LMO

3 June 2020

# Outline

1. Modelling
2. Inference
3. Model Selection
4. Simulation Studies
5. Application to Real Dataset

# Motivation Data Set




**Data :**

- $X^I$ Interaction between individuals (advice . . . )
- $X^O$ Interaction between organizations (contract . . . )
- $A$ Affiliation of the individuals to the organizations
  $A_{ij} = 1$ if $i$ is affiliated to $j$
  Only one affiliation per individual

## Objectives

- Joint probabilistic model on $\mathbf{X} = \{X^I, X^O\}$ given A
- Evaluate the influence of the inter-organizational level on the inter-individual level

# Outline

# Modelling of a Multilevel SBM


## Stochastic Block Model (SBM)[a]

[a] Snijders and Nowicki, 1997

- Mixture model for graphs
- Latent variables on vertices
- Model heterogeneity of connection

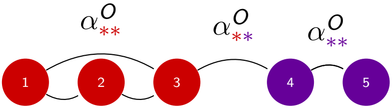
# Modelling of a Multilevel SBM


## Inter-organizational Level

- $n_O$ laboratories into $Q_O$ clusters
- Latent variables are independent
- $Z_j^O = l \Leftrightarrow j \in l, \quad l \in \{1, \dots, Q_O\}$

$$\mathbb{P}(Z_j^O = l) = \pi_l^O$$

# Modelling of a Multilevel SBM




## Inter-organizational Level

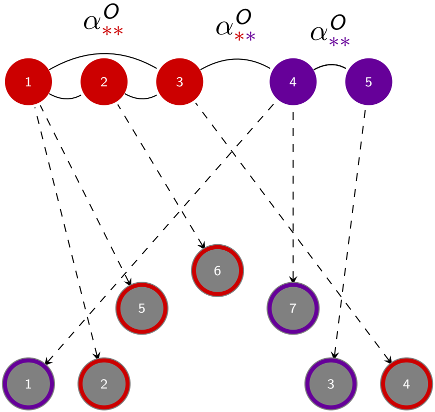
- $n_O$ laboratories into $Q_O$ clusters
- Latent variables are independent
- $Z^O_j = I \Leftrightarrow j \in I, \quad I \in \{1, \dots, Q_O\}$

$$\mathbb{P}(Z^O_j = I) = \pi^O_I$$

- Connectivity is independent given the latent variables

$$\mathbb{P}(X^O_{jj'} = 1 | Z^O_j = I, Z^O_{j'} = I') = \alpha^O_{II'}$$

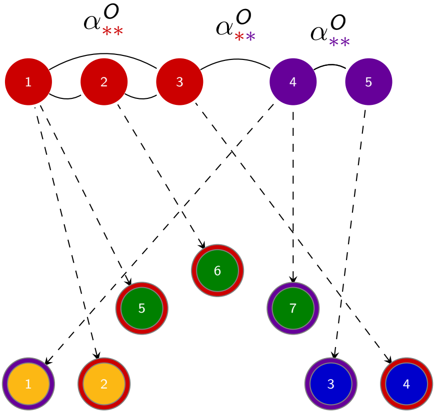
# Modelling of a Multilevel SBM




## Inter-individual Level

- $n_I$ researchers into $Q_I$ clusters
- A researcher's cluster depends on his laboratory's cluster

# Modelling of a Multilevel SBM

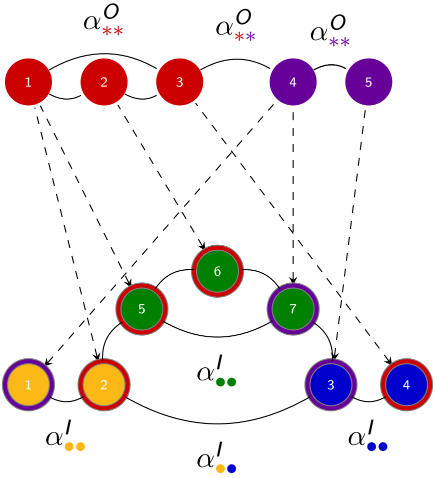



## Inter-individual Level

- $n_I$ researchers into $Q_I$ clusters
- A researcher's cluster depends on his laboratory's cluster
- $Z_i^I = k \Leftrightarrow i \in k, k \in \{1, \dots, Q_I\}$

$$\mathbb{P}(Z_i^I = k | A_i = j, Z_j^O = l) = \gamma_{kl}$$

# Modelling of a Multilevel SBM




## Inter-individual Level

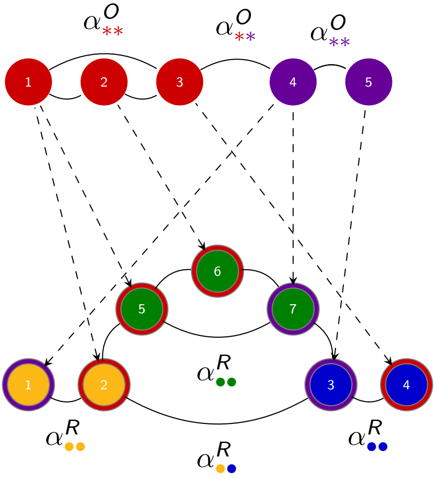
- $n_I$ researchers into $Q_I$ clusters
- A researcher's cluster depends on his laboratory's cluster
- $Z^I_i = k \Leftrightarrow i \in k, k \in \{1, \dots, Q_I\}$

$$\mathbb{P}(Z^I_i = k | A_i = j, Z^O_j = l) = \gamma_{kl}$$

- Connectivity is independent given the latent variables

$$\mathbb{P}(X^I_{ii'} = 1 | Z^I_i = k, Z^O_i = k) = \alpha^I_{kk'}$$

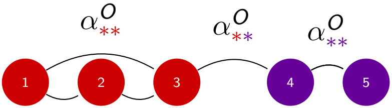
# Independence Between Levels

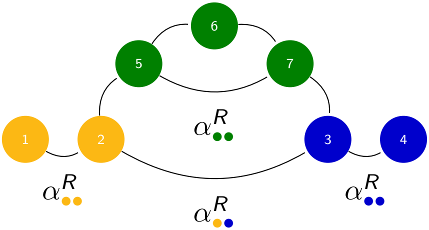



- $\pi^O$ is a probability vector
- Each column of $\gamma$ as well
- If $\gamma_{kl} = \gamma_{kl'} \quad \forall l, l'$

$$\mathcal{L}(X^I, X^O | A) = \mathcal{L}(X^I)\mathcal{L}(X^O)$$

# Independence Between Levels




- $\pi^O$ is a probability vector
- Each column of $\gamma$ as well
- If $\gamma_{kI} = \gamma_{kI'} \quad \forall I, I'$

$$\mathcal{L}(X^I, X^O | A) = \mathcal{L}(X^I)\mathcal{L}(X^O)$$

- Each level of the multilevel network is a SBM with $\pi^I = \gamma_{\cdot 1}$
- Organisational structure has no influence on the connectivity of individuals

# Identifiability

**Proposition**

The multilevel model is identifiable up to label switching under the following assumptions:

(i) All coefficients of $\alpha^O \cdot \pi^O$ are distinct

(ii) All coefficients of $\alpha^I \cdot \gamma \cdot \pi^O$ are distinct

(iii) $n_I \geq 2Q_I$

(iv) $n_O \geq \max\{2Q_O, Q_I + Q_O - 1\}$

(v) At least $2Q_I$ organizations contain one individual or more.

Proof derived from Celisse et al., 2012

# Outline

# Maximum Likelihood Inference

Objective: Joint clustering of $\mathbf{Z} = \{Z^I, Z^O\}$ and estimates of $\theta = \{\pi^O, \gamma, \alpha^O, \alpha^I\}$

Method: Maximum likelihood of the observed data

Idea: Calculate the complete likelihood and integrate on the latent variables

Problem: Intractable, sum of $Q_R^{n_R} \times Q_L^{n_L}$ terms

Solution: EM algorithm

Problem: $\mathcal{L}(\mathbf{Z}|\mathbf{X})$ also intractable

Solution: Variational approach of the EM algorithm

Daudin et al., 2008

# Variational EM

Maximise a lower bound of the observed data likelihood

$$
\begin{aligned}
\ell_\theta(\mathbf{X}) &\geq \ell_\theta(\mathbf{X}) - KL\left(\mathcal{R}(\mathbf{Z}) \| \mathbb{P}_\theta(\mathbf{Z}|\mathbf{X})\right) \\
&= \mathbb{E}_{\mathcal{R}}\left[\ell_\theta(\mathbf{X}, \mathbf{Z})\right] + \mathcal{H}\left(\mathcal{R}(\mathbf{Z})\right) \\
&= \mathcal{I}_\theta(\mathcal{R}(\mathbf{Z}))
\end{aligned}
$$

$\mathcal{R}(\mathbf{Z})$ is a mean-field approximation of $\mathbf{Z}|\mathbf{X}$
$\mathcal{H}$ is the entropy

## VEM algorithm

2 steps iterative algorithm

VE Maximise $\mathcal{I}_\theta(\mathcal{R}(\mathbf{Z}))$ w.r.t. $\mathcal{R}(\mathbf{Z})$

M Maximise $\mathcal{I}_\theta(\mathcal{R}(\mathbf{Z}))$ w.r.t. $\theta$

# Parameters update

## VE-Step : variational parameters

$$\widehat{\tau^O_{jl}} \propto \pi^O_l \prod_{i,k} {\gamma_{kl}}^{A_{ij}\widehat{\tau^I_{ik}}} \prod_{j' \neq j} \prod_{l'} \varphi(X^O_{jj'}, \alpha^O_{ll'}, \widehat{\tau^O_{j'l'}})$$

$$\widehat{\tau^I_{jl}} \propto \prod_{j,l} {\gamma_{kl}}^{A_{ij}\widehat{\tau^O_{jl}}} \prod_{i' \neq i} \prod_{k'} \varphi(X^I_{ii'}, \alpha^I_{kk'}, \widehat{\tau^I_{i'k'}})$$

$\tau^I_{ik} = \mathbb{P}_{\mathcal{R}}(Z^I_i = k) \quad \tau^O_{jl} = \mathbb{P}_{\mathcal{R}}(Z^O_j = l)$
$\varphi(X, \alpha, \tau) = \left(\alpha^X (1-\alpha)^{1-X}\right)^\tau$

## M-step : model parameters

$$\widehat{\pi^O_l} = \frac{1}{n_O} \sum_j \widehat{\tau^O_{jl}}$$

$$\hat{\alpha^I_{kk'}} = \frac{\sum_{i' \neq i} \widehat{\tau^I_{ik}} \widehat{\tau^I_{i'k'}} X^I_{ii'}}{\sum_{i' \neq i} \widehat{\tau^I_{ik}} \widehat{\tau^I_{i'k'}}}$$

$$\widehat{\alpha^O_{ll'}} = \frac{\sum_{j' \neq j} \widehat{\tau^O_{jl}} \widehat{\tau^O_{j'l'}} X^O_{jj'}}{\sum_{j' \neq j} \widehat{\tau^O_{jl}} \widehat{\tau^O_{j'l'}}}$$

$$\widehat{\gamma_{kl}} = \frac{\sum_{i,j} A_{ij} \widehat{\tau^I_{ik}} \widehat{\tau^O_{jl}}}{\sum_{i,j} A_{ij} \widehat{\tau^O_{jl}}}$$

# Outline

# Model Selection for the number of clusters

## Penalized criterion for choosing the number of clusters

$$
\begin{aligned}
ICL_{Multilevel}(Q_I, Q_O) = \max_{\theta} \ell_\theta(X^I, X^O, \hat{Z}^I, \hat{Z}^O | A) \\
\underbrace{-\frac{1}{2}\frac{Q_I(Q_I+1)}{2}\log\frac{n_I(n_I-1)}{2}}_{\alpha^I} - \underbrace{\frac{Q_O(Q_I-1)}{2}\log n_I}_{\gamma} \\
\underbrace{-\frac{1}{2}\frac{Q_O(Q_O+1)}{2}\log\frac{n_O(n_O-1)}{2}}_{\alpha^O} - \underbrace{\frac{Q_O-1}{2}\log n_O}_{\pi^O}
\end{aligned}
$$

- Step-wise procedure with relevant local initialization of VEM to optimise the ICL

Biernacki et al., 2000

# Model selection for independence

- ICL can be used to state on the independence between levels
- New penality term for $\gamma$

$$\text{pen}_\gamma = \frac{Q_I - 1}{2} \log n_I$$

- $ICL_{ind}(Q_I, Q_O) = ICL^I_{SBM}(Q_I) + ICL^O_{SBM}(Q_O)$
- We decide that levels are interdependent if

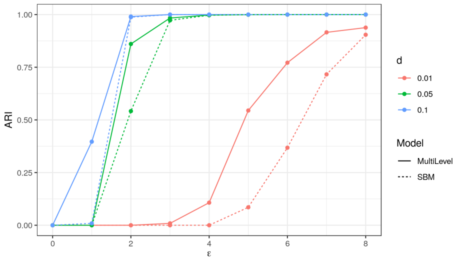
$$ICL_{ind}(Q_I, Q_O) < ICL_{Multilevel}(Q_I, Q_O)$$

# Outline

# Simulation Studies

$$Q_O = 3 \quad Q_I = 3 \quad n_O = 20 * Q_I \quad n_I = 3 * n_O \quad \pi^O = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

$$\alpha^I = d * \begin{bmatrix} 1+\epsilon & 1 & 1 \\ 1 & 1+\epsilon & 1 \\ 1 & 1 & 1+\epsilon \end{bmatrix} \alpha^O = \begin{bmatrix} .5 & .1 & .1 \\ .1 & .5 & .1 \\ .1 & .1 & .5 \end{bmatrix} \qquad \gamma = \begin{bmatrix} .8 & .1 & .1 \\ .1 & .8 & .1 \\ .1 & .1 & .8 \end{bmatrix}$$

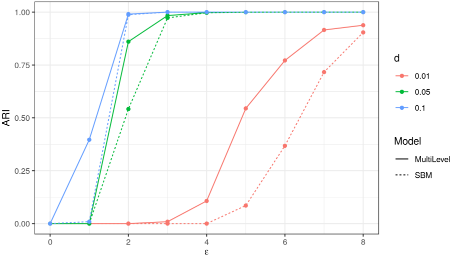


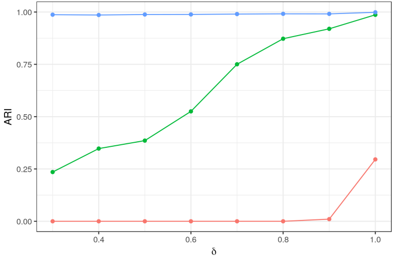

50 simulations on each dot

# Simulation Studies

$$Q_O = 3 \quad Q_I = 3 \quad n_O = 20 * Q_I \quad n_I = 3 * n_O \quad \pi^O = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

$$\alpha^I = d * \begin{bmatrix} 1+2 & 1 & 1 \\ 1 & 1+2 & 1 \\ 1 & 1 & 1+2 \end{bmatrix} \alpha^O = \begin{bmatrix} .5 & .1 & .1 \\ .1 & .5 & .1 \\ .1 & .1 & .5 \end{bmatrix} \quad \gamma = \begin{bmatrix} \delta & 1-\frac{\delta}{2} & 1-\frac{\delta}{2} \\ 1-\frac{\delta}{2} & \delta & 1-\frac{\delta}{2} \\ 1-\frac{\delta}{2} & 1-\frac{\delta}{2} & \delta \end{bmatrix}$$



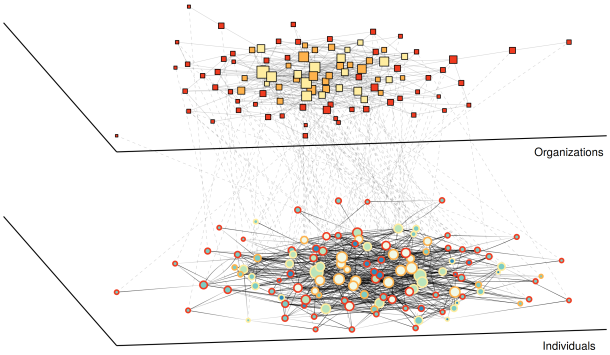




50 simulations on each dot

# Outline

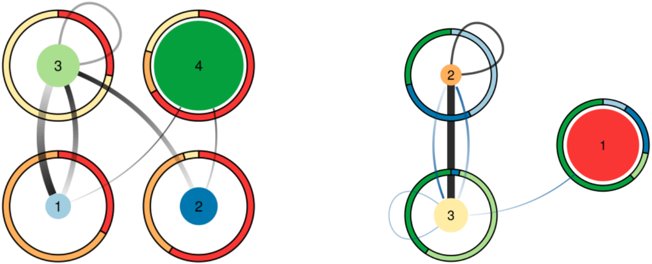
# Application to a Television Program Trade Fair Dataset[4]

128 individuals with directed interactions (advice) and 109 organizations with undirected interactions (deal).




[4] Brailly, 2016

# Dataset analysis




- 4 blocks of individuals and 3 blocks of organizations
- Levels are interdependent
- The structure of connection between individuals do not replicate the structure of connections between organizations

- Preprint available on arXiv: `https://arxiv.org/abs/1910.10512`
- `R` package available at `https://chabert-liddell.github.io/MLVSBM/`
  - Simulates and infers multilevel networks
  - Includes handling of missing data on $X^I$ and $X^O$
  - Prediction on missing dyads, missing links and spurious links
  - Works with multi-affiliation datasets

# References

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.

Brailly, Julien (2016). "Dynamics of networks in trade fairs—A multilevel relational approach to the cooperation among competitors". In: *Journal of Economic Geography* 16.6, pp. 1279–1301.

Celisse, Alain, Jean-Jacques Daudin, Laurent Pierre, et al. (2012). "Consistency of maximum-likelihood and variational estimators in the stochastic block model". In: *Electronic Journal of Statistics* 6, pp. 1847–1899.

Daudin, J-J, Franck Picard, and Stéphane Robin (2008). "A mixture model for random graphs". In: *Statistics and computing* 18.2, pp. 173–183.

Snijders, Tom A.B. and Krzysztof Nowicki (Jan. 1997). "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure". In: *Journal of Classification* 14.1, pp. 75–100. ISSN: