A Stochastic Block Model for Multilevel Network

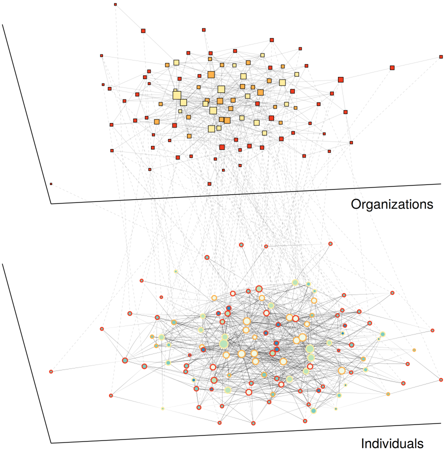
Saint-Clair Chabert-Liddell†
Joint work with P. Barbillon†, S. Donnet† & E. Lazega*

JdS 2021
10 June 2021

† UMR INRAe/AgroParisTech, MIA Paris
Université Paris-Saclay
* Institut d'Études Politiques de Paris

Motivation Dataset

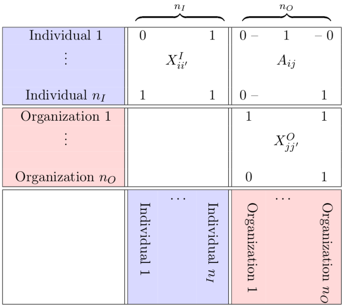
Economic and social networks in a television trade fair [1].



- Economic network: 109 organizations signing deals (undirected interactions)
- Represented on the trade fair by individuals
- Social network: 128 individuals sharing advice (directed interactions)

[1] Brailly (2016)

Objective of this work



Data :

X^I Interactions between individuals

X^O Interactions between organizations

A Affiliations 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 $\mathrm{X} = \{X^I, X^O\}$ given A
- Evaluate the influence of the inter-organizational level on the inter-individual level

Outline

Modeling

Inference

Model Selection

Simulation Studies

Application to Television Program Trade Fair

Modeling of a Multilevel SBM

Stochastic Block Model (SBM)[a]

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

[a] Snijders and Nowicki (1997)

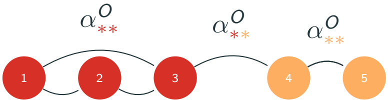
Modeling of a Multilevel SBM

Inter-organizational Level

- n_O organizations into Q_O blocks
- Latent variables are independent
- $Z_j^O = I \Leftrightarrow j \in I, \quad I \in \{1, \ldots, Q_O\}$

$$\mathbb{P}(Z_j^O = I) = \pi_I^O$$

Modeling of a Multilevel SBM



Inter-organizational Level

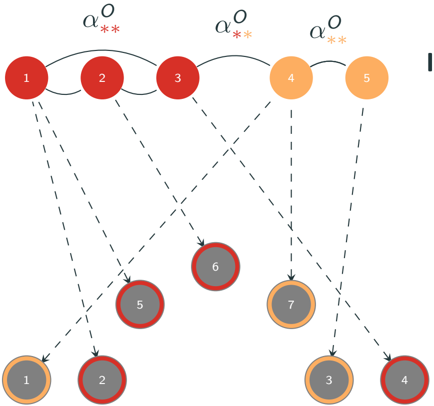
- n_O organizations into Q_O blocks
- Latent variables are independent
- $Z^O_j = I \Leftrightarrow j \in I, \quad I \in \{1, \ldots, Q_O\}$

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

- Connections are independent given the latent variables

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

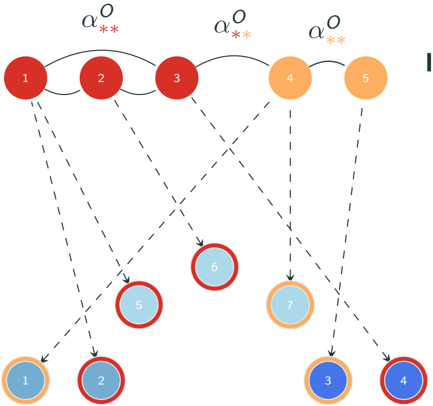
Modeling of a Multilevel SBM



Inter-individual Level

- n_I individuals into Q_I blocks
- The block of an individual depends on the block of her/his organization

Modeling of a Multilevel SBM



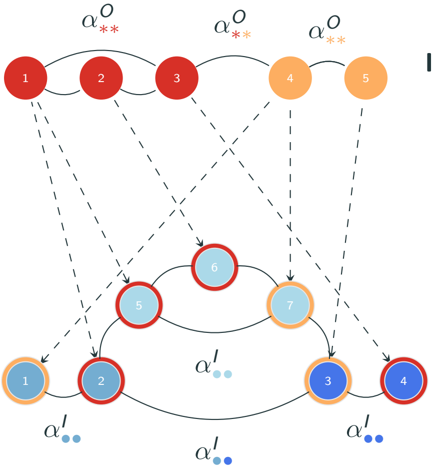
Inter-individual Level

- n_I individuals into Q_I blocks
- The block of an individual depends on the block of her/his organization

- $Z_i^I = k \Leftrightarrow i \in k, k \in \{1, \dots, Q_I\}$

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

Modeling of a Multilevel SBM



Inter-individual Level

- n_I individuals into Q_I blocks
- The block of an individual depends on the block of her/his organization

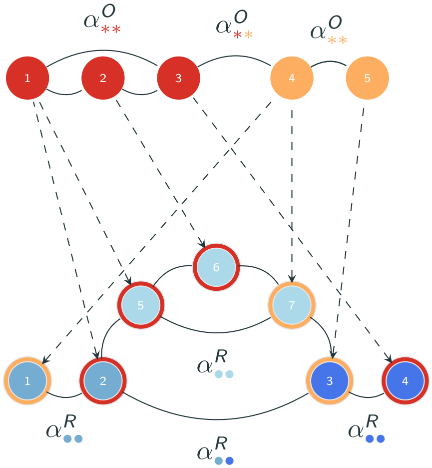
- $Z^I_i = k \Leftrightarrow i \in k, k \in \{1, \ldots, Q_I\}$

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

- Connections are independent given the latent variables

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

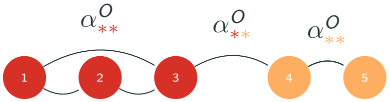
Independence Between Levels



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

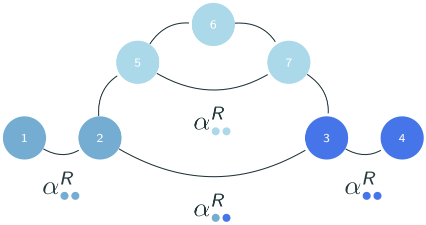
$$\mathcal{L}(X^I, X^O|A) = \mathcal{L}(X^I)\mathcal{L}(X^O)$$

Independence Between Levels



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

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



- Each level of the multilevel network is a SBM with $\pi^I = \gamma_{\cdot 1}$
- Organizational structure has no influence on the connections 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 $2Q_O$ organizations contain one individual or more.

Proof derived from Celisse et al. (2012)

Outline

Maximum Likelihood Inference

Objective Joint clustering of $\mathrm{Z} = \{Z^I, Z^O\}$ 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 $\mathcal{L}(\mathrm{Z}|\mathrm{X})$ also intractable

Solution Variational approach of the EM algorithm

Daudin et al. (2008)

Variational EM

Maximise a lower bound of the observed data likelihood

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

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

VEM algorithm
2 steps iterative algorithm

- VE Maximise $\mathcal{I}_\theta(\mathcal{R}(\mathsf{Z}))$ w.r.t. $\mathcal{R}(\mathsf{Z})$
- M Maximise $\mathcal{I}_\theta(\mathcal{R}(\mathsf{Z}))$ w.r.t. θ

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

Penalized criterion for choosing the number of blocks

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

- 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 γ

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

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

$$\max_{Q_I} ICL^I_{SBM}(Q_I) + \max_{Q_O} ICL^O_{SBM}(Q_O) < \max_{\{Q_I, Q_O\}} ICL_{Multilevel}(Q_I, Q_O)$$

Outline

Modeling

Inference

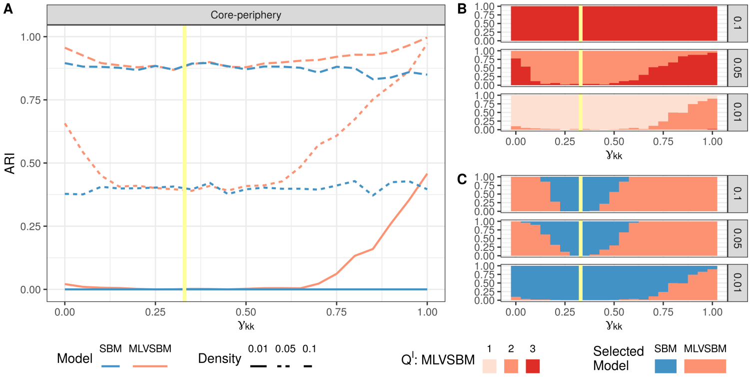
Model Selection

Simulation Studies

Application to Television Program Trade Fair

Simulation Studies

- Strong dependence between levels (γ_{kk} far from $1/3$) helps recover the structure of the inter-individual level with the information of the inter-organizational level.
- ICL tends to select model of small size $\Longrightarrow$ Good for testing the interdependence.



$$\alpha^I = Density * \begin{pmatrix} 3 & 3 & 1 \\ 3 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \alpha^O = \begin{pmatrix} .5 & .1 & .1 \\ .1 & .5 & .1 \\ .1 & .1 & .5 \end{pmatrix}, \gamma = \begin{pmatrix} \delta & 1-\frac{\delta}{2} & 1-\frac{\delta}{2} \\ 1-\frac{\delta}{2} & \delta & 1-\frac{\delta}{2} \\ 1-\frac{\delta}{2} & 1-\frac{\delta}{2} & \delta \end{pmatrix}$$

Outline

Modeling

Inference

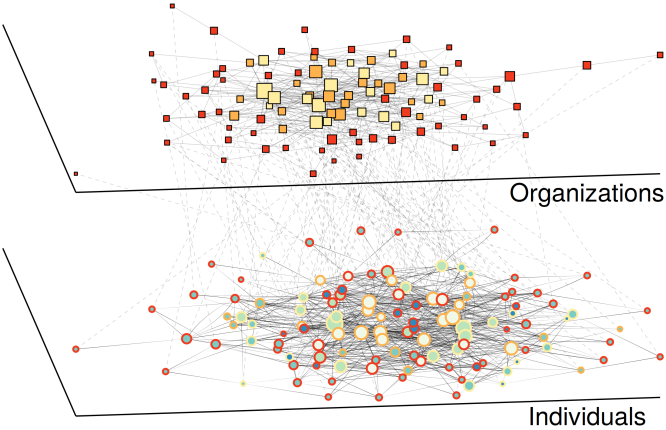
Model Selection

Simulation Studies

Application to Television Program Trade Fair

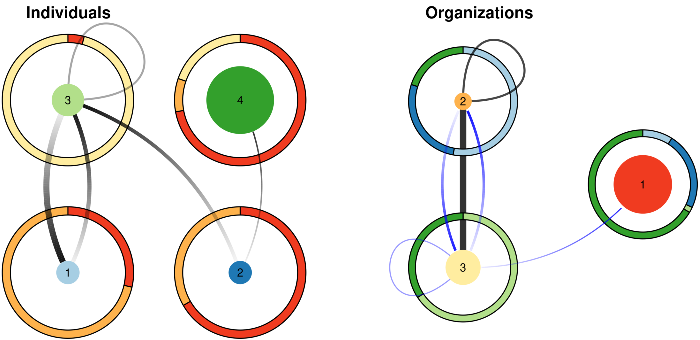
Application to a Television Program Trade Fair Dataset[5]

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



[5]Brailly (2016)

Dataset analysis



- Levels are interdependent, $(Q^I, Q^O) = (4, 3)$
- Core-periphery structure on X^O
- Mainly inter-block connections for X^I (except block 3, sub-group of sellers)
- Intra-block connection between individuals do not replicate the intra-block connections of their organizations (block 2 and 3)

- Paper: CL. et al. (2021) in CSDA
  `doi:10.1016/j.csda.2021.107179`
- `R` package available on CRAN and at
  `https://chabert-liddell.github.io/MLVSBM/`
    - Simulation and inference of multilevel networks
    - Handling of missing data on X^I and/or X^O
    - Prediction on missing dyads, missing links and spurious links
    - Extend to multi-affiliation datasets

Any question? saint-clair.chabert-liddell@agroparistech.fr

Thank you for your attention!

References i

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence 22*(7), 719–725.

Brailly, J. (2016). Dynamics of networks in trade fairs—a multilevel relational approach to the cooperation among competitors. *Journal of Economic Geography 16*(6), 1279–1301.

Celisse, A., J.-J. Daudin, L. Pierre, et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics 6*, 1847–1899.

CL., S.-C., P. Barbillon, S. Donnet, and E. Lazega (2021). A stochastic block model approach for the analysis of multilevel networks: An application to the sociology of organizations. *Computational Statistics & Data Analysis 158*, 107179.

Daudin, J.-J., F. Picard, and S. Robin (2008). A mixture model for random graphs. *Statistics and computing 18*(2), 173–183.

Snijders, T. A. and K. Nowicki (1997, jan). Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification 14*(1), 75–100.

Inference algorithm

- Variational method are very sensible to initialization
- Initialization done by clustering obtained from SBM on each level
- Cluster merging and splitting with HAC to initialize model with neighbour size on $\mathbb{N}^2$

Model size known

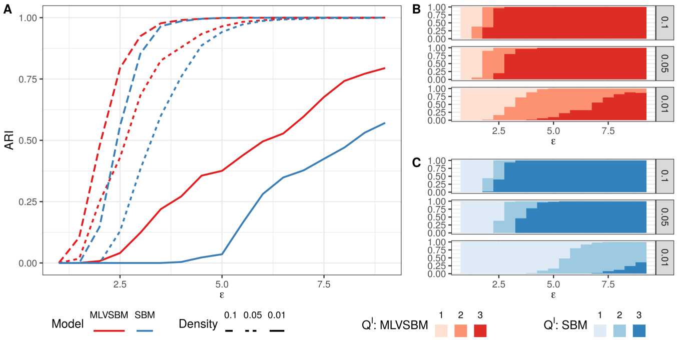
- 3 steps algorithm : known size –> neighbours –> known size
- Keep the model with the highest variational bound

Model size unknown

- Greedy algorithm to select the number of clusters
- Each step select the best model on each neighbour size
- Keep the model with the highest ICL

Simulation Studies

- Information from the inter-organizational level helps recover the structure of the inter-individual level



$$\alpha^{I} = Density * \begin{pmatrix} \epsilon & \epsilon & \mathbf{1} \\ \epsilon & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix} \quad \alpha^{O} = \begin{pmatrix} \mathbf{.5} & \mathbf{.1} & \mathbf{.1} \\ \mathbf{.1} & \mathbf{.5} & \mathbf{.1} \\ \mathbf{.1} & \mathbf{.1} & \mathbf{.5} \end{pmatrix} \quad \gamma = \begin{pmatrix} \mathbf{.8} & \mathbf{.1} & \mathbf{.1} \\ \mathbf{.1} & \mathbf{.8} & \mathbf{.1} \\ \mathbf{.1} & \mathbf{.1} & \mathbf{.8} \end{pmatrix}$$

Link Prediction

- The social network and the economic network are interdependent.
- Inter-organizational level helps predicting links on the inter-individual level.