Analyse de réseaux sociaux à l'aide de modèles de graphe aléatoire

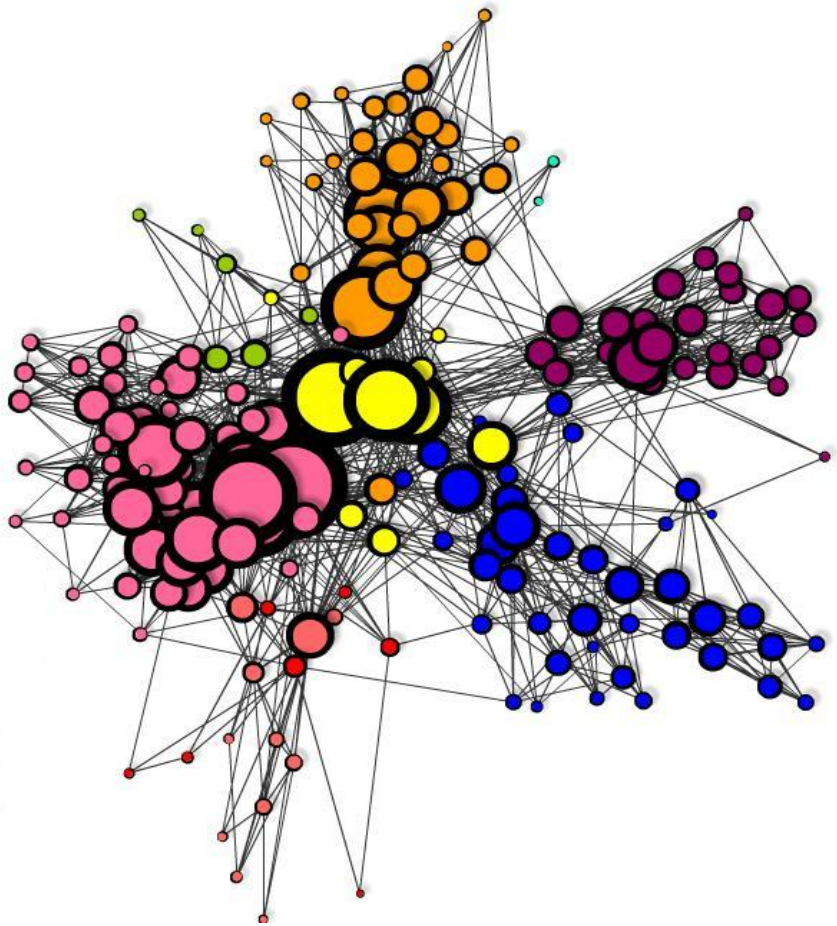
Pierre Latouche

Laboratoire Statistique et Génome

Journées MASHS, 23/06/2011

Real networks

- **Many scientific fields** :
  - World Wide Web
  - Biology, sociology, physics
- **Nature of data under study**:
  - Interactions between N objects
  - $\mathcal{O}(N^2)$ possible interactions
- **Network topology** :
  - Describes the way nodes interact, structure/function relationship



Sample of 250 blogs (nodes) with their links (edges) of the French political Blogosphere.

Real networks

- **Properties** :
  - Sparsity
  - Existence of a giant component
  - Heterogeneity
  - Preferential attachment
  - Small world

$\hookrightarrow$ Topological structure (groups of vertices)

Real networks

- **Properties** :
  - Sparsity
  - Existence of a giant component
  - Heterogeneity
  - Preferential attachment
  - Small world

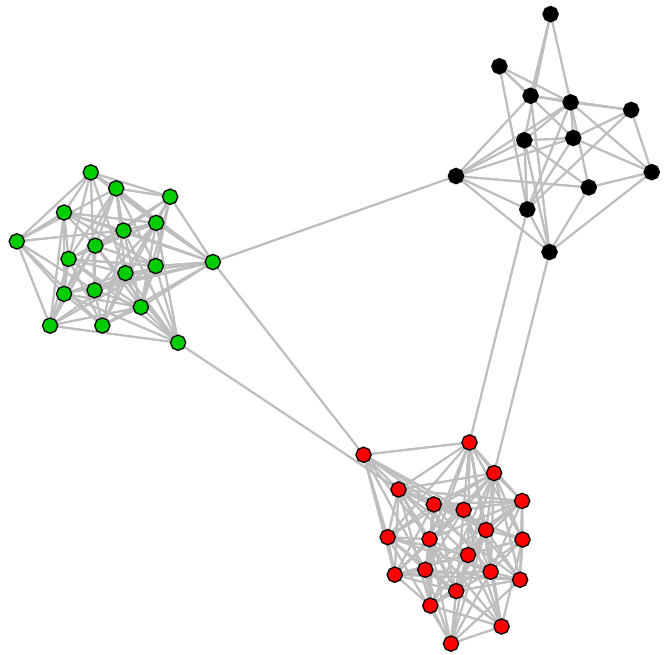
$\hookrightarrow$ Topological structure (groups of vertices)

Graph clustering

- **Existing methods look for** :
  - Community structure
  - Disassortative mixing
  - Heterogeneous structure

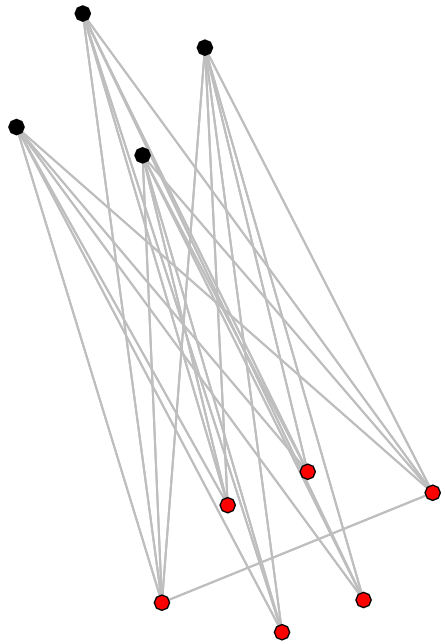
Graph clustering

- **Existing methods look for** :
  - Community structure
  - Disassortative mixing
  - Heterogeneous structure



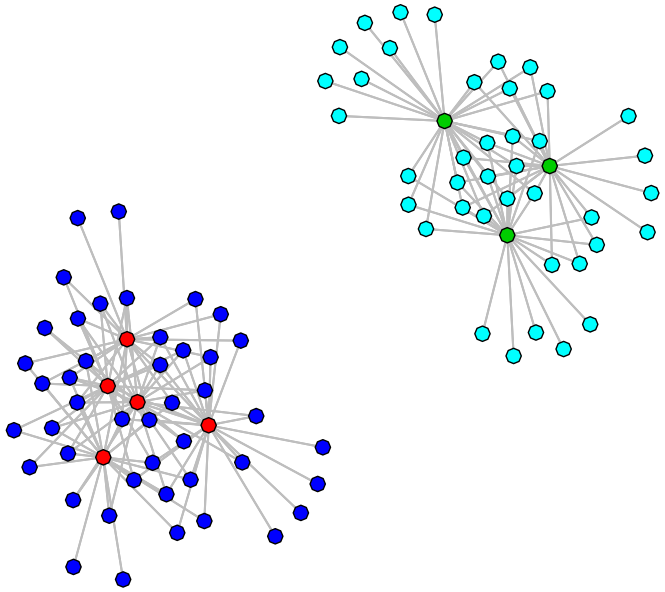
Graph clustering

- **Existing methods look for** :
  - Community structure
  - Disassortative mixing
  - Heterogeneous structure



Graph clustering

- **Existing methods look for** :
  - Community structure
  - Disassortative mixing
  - Heterogeneous structure



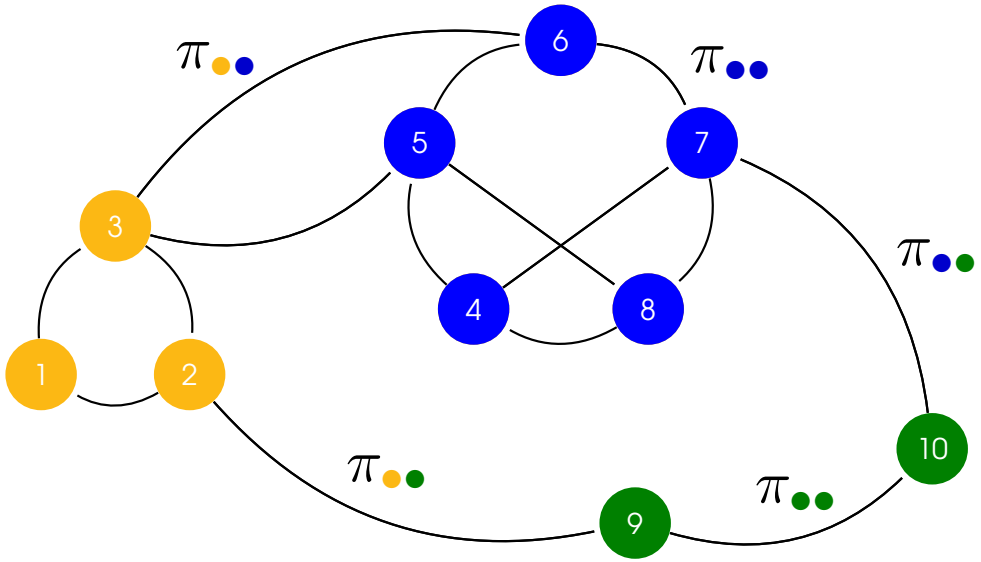
Stochastic Block Model (SBM)

- Nowicki and Snijders (2001)
  - Earlier work : Govaert et al. (1977)
- $\mathbf{Z}_i$ independent hidden variables :
  - $\mathbf{Z}_i \sim \mathcal{M}\Big(1,\, \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_Q)\Big)$
  - $Z_{iq} = 1$: vertex i belongs to class q
- $\mathbf{X} \,|\, \mathbf{Z}$ edges drawn independently :

$$X_{ij} | \{Z_{iq} Z_{jl} = 1\} \sim \mathcal{B}(\pi_{ql})$$

- A mixture model for graphs :

$$X_{ij} \sim \sum_{q=1}^{Q} \sum_{l=1}^{Q} \alpha_q \alpha_l \mathcal{B}(\pi_{ql})$$


$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
6
5
7
3
4
8
1
2
10
9

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Complete-data : $\log p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi})$
  - Observed-data : $\log p(\mathbf{X} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi}) = \log \{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi})\}$
    $\hookrightarrow$ Q^N terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$

Problem

$p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$ is not tractable (no conditional independence)

Model selection

Criteria

Since $\log p(\mathbf{X} \mid \boldsymbol{\alpha}, \boldsymbol{\Pi})$ is not tractable, we *cannot* rely on:

- $AIC = \log p(\mathbf{X} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Pi}}) - K$
- $BIC = \log p(\mathbf{X} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Pi}}) - \frac{K}{2} \log \frac{N(N-1)}{2}$

Contents

Introduction
- Real networks
- Graph clustering

Variational Bayesian inference and complexity control for stochastic block models
- Bayesian framework
- VBEM algorithm
- Model selection
- Experiments

Overlapping stochastic block models
- The model
- Inference and model selection
- Experiments

Bayesian framework

- **Conjugate prior distributions** :
  - $p\Big(\boldsymbol{\alpha} \,|\, \mathbf{n}^0 = \{n_1^0, \ldots, n_Q^0\} \Big) = \mathrm{Dir}(\boldsymbol{\alpha};\ \mathbf{n}^0)$
  - $p\Big(\boldsymbol{\Pi} \,|\, \boldsymbol{\eta}^0 = (\eta_{ql}^0), \boldsymbol{\zeta}^0 = (\zeta_{ql}^0) \Big) = \prod_{q \leq l} \mathrm{Beta}(\pi_{ql};\ \eta_{ql}^0, \zeta_{ql}^0)$
- **Non informative Jeffreys prior** :
  - $n_q^0 = 1/2$
  - $\eta_{ql}^0 = \zeta_{ql}^0 = 1/2$

- $p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi} \,|\, \mathbf{X})$ not tractable

Decomposition

$$\log p(\mathbf{X}) = \mathcal{L}\left(q\right) + \mathrm{KL}\left(q(\cdot) \,||\, p(\cdot|\, \mathbf{X})\right)$$

where

$$\mathcal{L}(q) = \sum_{\mathbf{Z}} \int \int q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi})} \right\} d\,\boldsymbol{\alpha}\, d\,\boldsymbol{\Pi}$$

Factorization

$$q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi}) = q(\boldsymbol{\alpha}) q(\boldsymbol{\Pi}) q(\mathbf{Z}) = q(\boldsymbol{\alpha}) q(\boldsymbol{\Pi}) \prod_{i=1}^{N} q(\mathbf{Z}_i)$$

E-step

- $q(\mathbf{Z}_i) = \mathcal{M}(\mathbf{Z}_i;\ 1, \boldsymbol{\tau_i} = \{\tau_{i1}, \ldots, \tau_{iQ}\})$

M-step

- $q(\boldsymbol{\alpha}) = \mathrm{Dir}(\alpha;\ \mathbf{n})$
- $q(\boldsymbol{\Pi}) = \prod_{q \leq l}^{Q} \mathrm{Beta}(\pi_{ql};\ \eta_{ql}, \zeta_{ql})$

A new model selection criterion : ILvb

Latouche et al. (2011a)

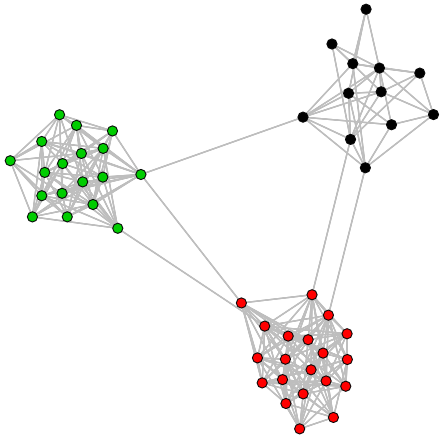
- $\log p(\mathbf{X}\,|Q) = \mathcal{L}\,(q) + \mathrm{KL}(...)$
- After convergence, use $\mathcal{L}\,(q)$ as an approximation of $\log p(\mathbf{X}\,|Q)$

ILvb

$$IL_{vb} = \log\left\{\frac{\Gamma(\sum_{q=1}^{Q} n_q^0)\prod_{q=1}^{Q}\Gamma(n_q)}{\Gamma(\sum_{q=1}^{Q} n_q)\prod_{q=1}^{Q}\Gamma(n_q^0)}\right\} + \sum_{q\leq l}^{Q}\log\left\{\frac{\Gamma(\eta_{ql}^0+\zeta_{ql}^0)\Gamma(\eta_{ql})\Gamma(\zeta_{ql})}{\Gamma(\eta_{ql}+\zeta_{ql})\Gamma(\eta_{ql}^0)\Gamma(\zeta_{ql}^0)}\right\} - \sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}\log\tau_{iq}$$

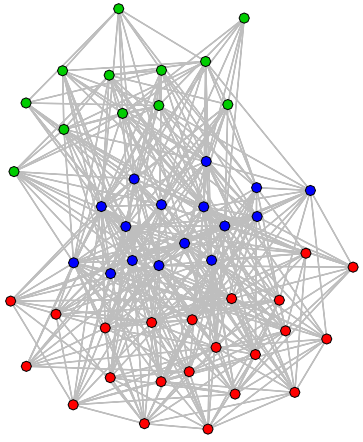
- **Two topological structures** :
  - Affiliation :

$$\mathbf{\Pi} = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon \\ \epsilon & \lambda & & \vdots \\ \vdots & & \ddots & \epsilon \\ \epsilon & \dots & \epsilon & \lambda \end{pmatrix}$$



  - Affiliation and a class of hubs :

$$\mathbf{\Pi} = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon & \lambda \\ \epsilon & \lambda & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \lambda & \dots & \dots & \dots & \lambda \end{pmatrix}$$



Experiments on simulated data

- $N = 50$
- $\lambda = 0.9$
- $\epsilon = 0.1$
- $\alpha_Q = 1/Q$
- $Q_{True} \in \{3, \ldots, 7\}$
- $Q \in \{1, \ldots, 7\}$
- 100 simulations
- 2 criteria : ICL (Biernacki et al. 2000, Daudin et al., 2008), lLvb

Affiliation networks

(a) $Q_{True} \backslash Q_{ICL}$

	2	3	4	5	6	7
3	0	100	0	0	0	0
4	0	0	100	0	0	0
5	0	0	23	**77**	0	0
6	0	1	28	59	**12**	0
7	0	8	49	42	1	**0**

(b) $Q_{True} \backslash Q_{ILvb}$

	2	3	4	5	6	7
3	0	100	0	0	0	0
4	0	0	100	0	0	0
5	0	0	0	**99**	1	0
6	0	0	4	23	**73**	0
7	0	2	14	44	27	**13**

Affiliation networks and a class of hubs

(c) $Q_{True} \backslash Q_{ICL}$

	2	3	4	5	6	7
3	0	**100**	0	0	0	0
4	0	0	**100**	0	0	0
5	0	0	12	**88**	0	0
6	0	0	19	59	**22**	0
7	0	3	29	56	12	**0**

(d) $Q_{True} \backslash Q_{ILvb}$

	2	3	4	5	6	7
3	0	**100**	0	0	0	0
4	0	0	**100**	0	0	0
5	0	0	2	**98**	0	0
6	0	0	1	29	**70**	0
7	0	0	3	34	45	**18**

Conclusion (1)

- variational Bayes to approximate $p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi} \mid \mathbf{X})$
- computational cost : $\mathrm{O}(Q^2 N^2)$
- model selection criterion : lLvb
- provides a relevant estimation of the number of classes
- implemented in a R package available on the CRAN : **mixer**

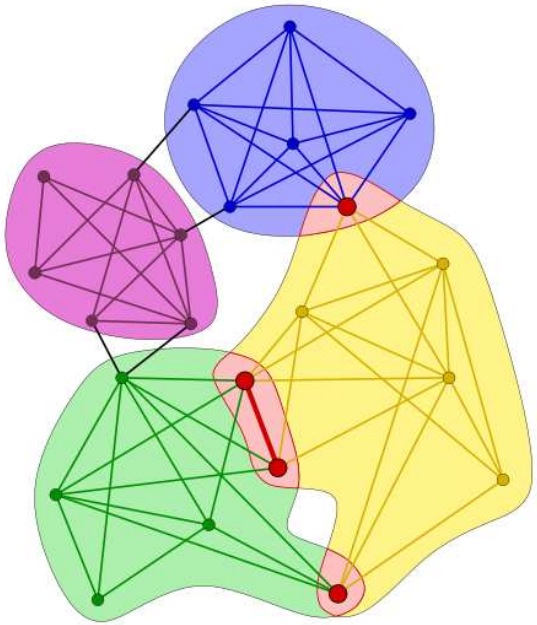
Contents

Introduction
- Real networks
- Graph clustering

Variational Bayesian inference and complexity control for stochastic block models
- Bayesian framework
- VBEM algorithm
- Model selection
- Experiments

Overlapping stochastic block models
- The model
- Inference and model selection
- Experiments

Overlaps in networks



Palla et al. (2005)

Problem

The stochastic block model (SBM) and most existing methods assume that each vertex belongs to a single class

Stochastic Block Model (SBM)

- Nowicki and Snijders (2001)
- $\mathbf{Z}_i$ independent hidden variables :

$$\mathbf{Z}_i \sim \mathcal{M}\Big(1,\ \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_Q)\Big)$$

Overlapping Stochastic Block model (OSBM)

- Latouche et al. (2011b)
- Z_{iq} independent hidden variables :

$$\mathbf{Z}_i \sim \prod_{q=1}^{Q} \mathcal{B}(Z_{iq};\, \alpha_q) = \prod_{q=1}^{Q} \alpha_q^{Z_{iq}} (1-\alpha_q)^{1-Z_{iq}}$$

Overlapping Stochastic Block model (OSBM)

- Latouche et al. (2011b)
- Z_{iq} independent hidden variables :

$$\mathbf{Z}_i \sim \prod_{q=1}^{Q} \mathcal{B}(Z_{iq};\ \alpha_q) = \prod_{q=1}^{Q} \alpha_q^{Z_{iq}} (1-\alpha_q)^{1-Z_{iq}}$$

- $\mathbf{X} \,|\, \mathbf{Z}$ edges drawn independently :

$$X_{ij} |\, \mathbf{Z}_i, \mathbf{Z}_j \sim \mathcal{B}\big(X_{ij};\ g(a_{\mathbf{Z}_i,\mathbf{Z}_j})\big)$$

- $g(t) = 1/\left(1+\exp(-t)\right)$ is the logistic function

$$a_{\mathbf{Z}_i,\mathbf{Z}_j} = \mathbf{Z}_i^\intercal\, \mathbf{W}\, \mathbf{Z}_j + \mathbf{Z}_i^\intercal\, \mathbf{U} + \mathbf{V}^\intercal\, \mathbf{Z}_j + W^*$$

OSBM

- $\tilde{\mathbf{Z}}_i = (\mathbf{Z}_i, 1)^\intercal$
- $\tilde{\mathbf{W}} = \begin{pmatrix} \mathbf{W} & \mathbf{U} \\ \mathbf{V}^\intercal & W^* \end{pmatrix}$
- $a_{\mathbf{Z}_i, \mathbf{Z}_j} = \tilde{\mathbf{Z}}_i^\intercal \, \tilde{\mathbf{W}} \, \tilde{\mathbf{Z}}_j$
- Parameter set : $\left\{ \boldsymbol{\alpha}, \tilde{\mathbf{W}} \right\}$

Bayesian framework

- **Conjugate prior distributions** :
  - $p(\boldsymbol{\alpha}) = \prod_{q=1}^{Q} \mathrm{Beta}(\alpha_q;\, \eta_q^0, \zeta_q^0)$
  - $p(\tilde{\mathbf{W}}^{\mathrm{vec}}) = \mathcal{N}(\tilde{\mathbf{W}}^{\mathrm{vec}};\, \tilde{\mathbf{W}}_0^{\mathrm{vec}}, \mathbf{S}_0)$
- The vec operator : if

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

then

$$\mathbf{A}^{\mathrm{vec}} = \begin{pmatrix} A_{11} \\ A_{21} \\ A_{12} \\ A_{22} \end{pmatrix}$$

Problem

$p(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}} \mid \mathbf{X})$ not tractable

q Transformation

Decomposition

$$\log p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p)$$

where

$$\mathcal{L}(q) = \sum_{\mathbf{Z}} \int q(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \log \Big(\frac{p(\mathbf{X} \mid \mathbf{Z}, \tilde{\mathbf{W}}) p(\mathbf{Z} \mid \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})} \Big) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}}$$

Lower bound

$$\log p(\mathbf{X}) \geq \mathcal{L}(q)$$

Problem

$\mathcal{L}(q)$ has a too complex form $\hookrightarrow$ no variational Bayes EM algorithm ??

Decomposition

$$\mathcal{L}(\boldsymbol{\xi}) = \mathcal{L}(q;\, \boldsymbol{\xi}) + \mathrm{KL}(q||p)$$

where

$$\mathcal{L}(q;\, \boldsymbol{\xi}) = \sum_{\mathbf{Z}} \int q(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \log \left(\frac{h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})} \right)$$

Lower bound

$$\log p(\mathbf{X}) \geq \mathcal{L}(\boldsymbol{\xi}) \geq \mathcal{L}(q;\, \boldsymbol{\xi})$$

Inference

Local optimization

- $\boldsymbol{\xi} = \text{argmax}_{\boldsymbol{\xi}} \mathcal{L}(q;\, \boldsymbol{\xi})$

E-step

- $q(Z_{iq}) = \mathcal{B}(Z_{iq};\, \tau_{iq})$

M-step

- $q(\boldsymbol{\alpha}) = \prod_{q=1}^{Q} \text{Beta}(\alpha_q;\, \eta_q^N, \zeta_q^N)$
- $q(\tilde{\mathbf{W}}^{vec}) = \mathcal{N}(\tilde{\mathbf{W}}^{\text{vec}};\, \tilde{\mathbf{W}}_N^{\text{vec}}, \mathbf{S}_N)$

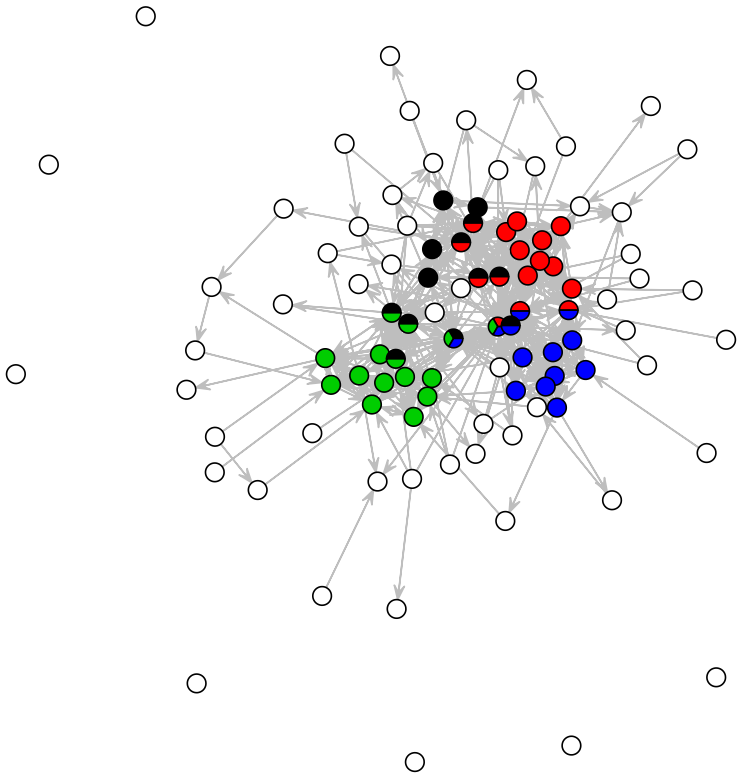
Model selection

- After convergence, use $\mathcal{L}(q;\, \boldsymbol{\xi})$ as an approximation of $\log p(\mathbf{X}\,|Q)$

ILosbm

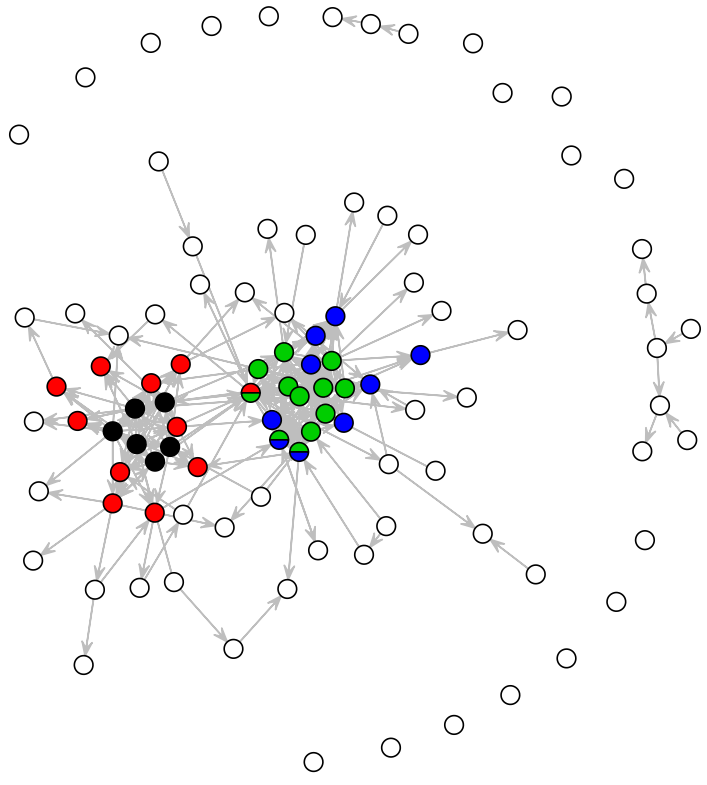
$$IL_{osbm} = \mathcal{L}(\hat{q};\, \hat{\boldsymbol{\xi}})$$

Community structures



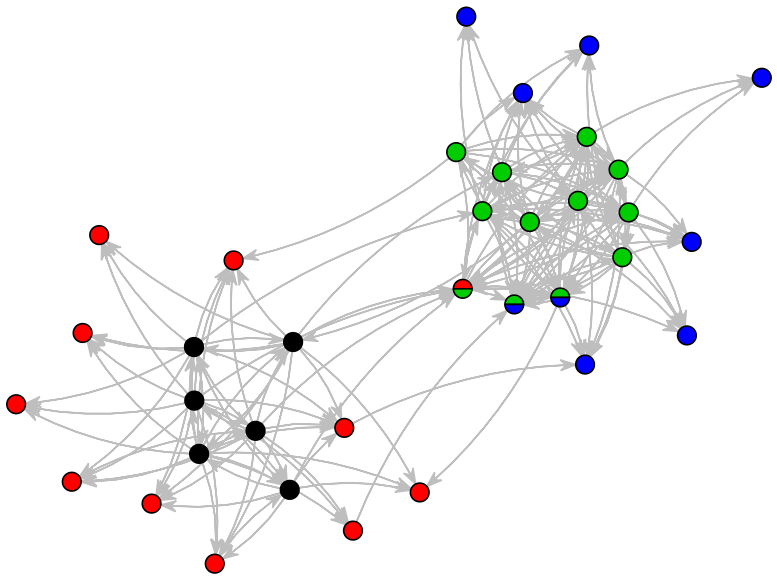
Example of an overlapping stochastic block model (OSBM) network with community structures.

Community structures and stars



Example of an overlapping stochastic block model (OSBM) network with community structures and stars.

Community structures and stars



Example of an overlapping stochastic block model (OSBM) network with community structures and stars.

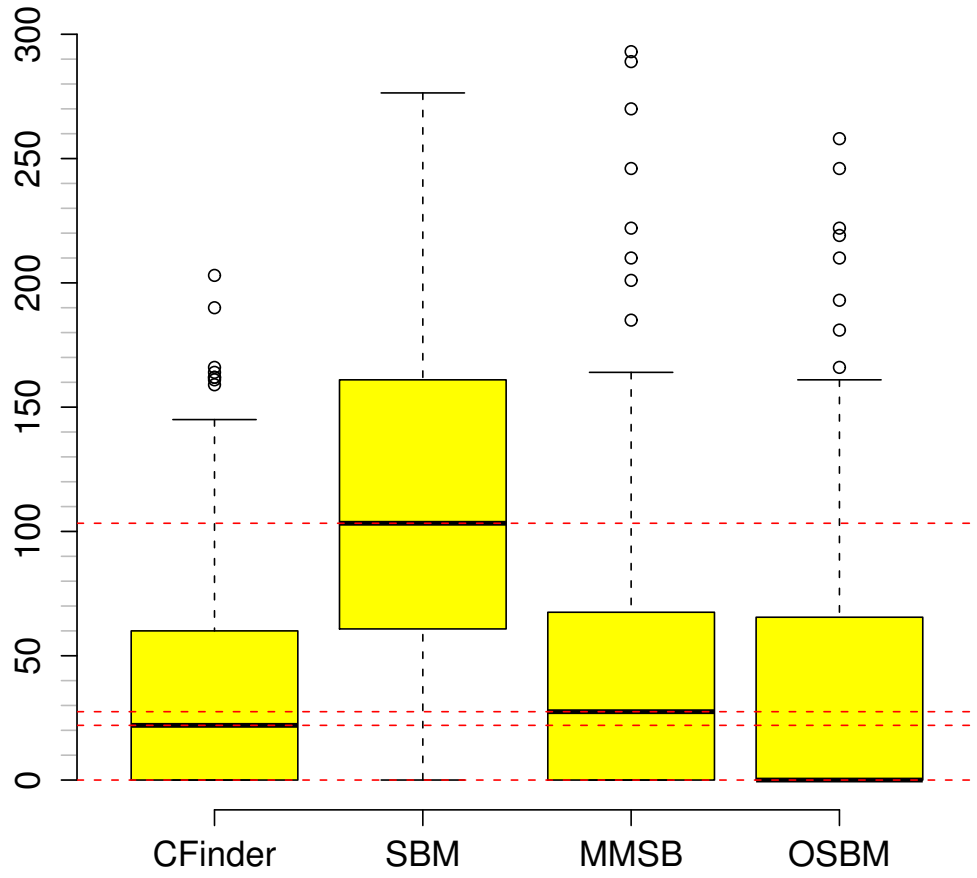
Experiments on simulated data

- $N = 100$
- $\lambda = 4$
- $\epsilon = 1$
- $W^* = -5.5$
- $\mathbf{U} = \mathbf{V} = \begin{pmatrix} \epsilon & \dots & \epsilon \end{pmatrix}$
- $\alpha_q = 0.25$
- $Q = 4$
- 100 simulations
- 4 graph clustering methods :
  - CFinder (Palla et al. 2006)
  - Stochastic Block Model (SBM)
  - Mixed Membership Stochastic Block Model (MMSB) (Airoldi et al. 2008)
  - Overlapping Stochastic Block Model (OSBM)

How to compare the methods ?

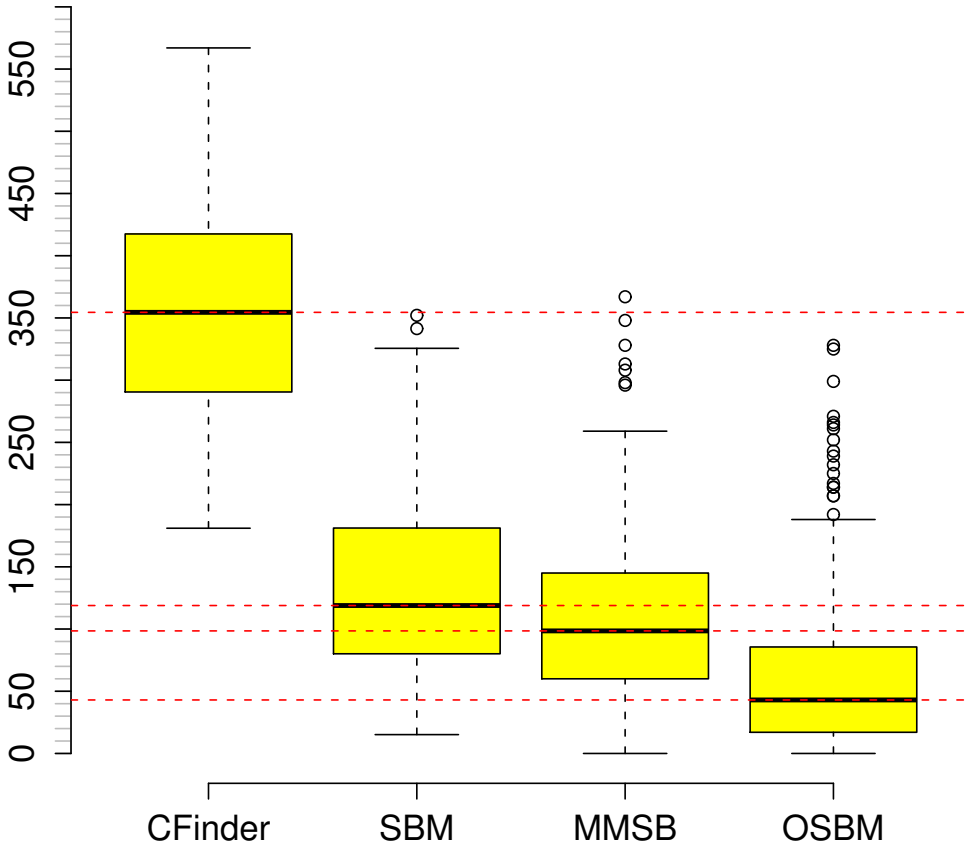
- CFinder and OSBM can deal with outliers ($\mathbf{Z}_i = \mathbf{0}$)
- SBM and MMSB are run with $Q + 1$ classes
  $\hookrightarrow$ identify the class of outliers
- Compute $\mathbf{P} = \mathbf{Z}\,\mathbf{Z}^\intercal$ and $\hat{\mathbf{P}} = \hat{\mathbf{Z}}\hat{\mathbf{Z}}^\intercal$:
  - invariant to column permutations of $\mathbf{Z}$ and $\hat{\mathbf{Z}}$
  - number of shared clusters between each pair of vertices
- Compute L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$

Community structures



L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ over the 100 samples of networks with community structures for CFinder, SBM, MMSB and OSBM.

Community structures and stars



L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ over the 100 samples of networks with community structures for CFinder, SBM, MMSB and OSBM.

The French blogosphere network

	UMP	UDF	liberal	PS	analysts	others
cluster 1	30 + 3	0 + 1	0	0	0 + 1	0
cluster 2	2 + 3	29 + 1	0	0	1 + 3	0
cluster 3	0	0	24	0	1 + 1	0
cluster 4	0	0 + 2	0	40	0 + 4	1
outliers	5	1	1	17	5	30

Classification of the blogs into $Q = 4$ clusters using OSBM. 196 vertices, 2864 edges.

Conclusion (2)

- A new random graph model : the overlapping stochastic block model (OSBM)
- Frequentist and Bayesian inference procedures
- Computational cost : $\mathrm{O}(Q^4 N^2)$
- New model selection criterion : ILosbm
- R package **OSBM** soon available on the CRAN

Papers

- K. Nowicki, T.A.B. Snijders. *Estimation and prediction for stochastic blockstructures*, Journal of the American Statistical Association, 2001.
- G. Palla, I. Derenyi, I. Farkas, T. Vicsek. *Uncovering the overlapping community structure of complex networks in nature and society*, Nature, 2005.
- J. Daudin, F. Picard, S. Robin. *A mixture model for random graphs*, Statistics and Computing, 2008.
- E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing. *Mixed membership stochastic blockmodels*, Journal of Machine Learning Research, 2008.
- P. Latouche, E. Birmelé, C. Ambroise. *Bayesian methods for graph clustering*, Advances in data handling and business intelligence, 2009.
- P. Latouche, E. Birmelé, C. Ambroise. *Variational Bayesian inference and complexity control for stochastic block models*, Statistical Computing, 2011.
- P. Latouche, E. Birmelé, C. Ambroise. *Overlapping stochastic block models with application to the French political blogosphere*, Annals of Applied Statistics, 2011.