Variational methods for overlapping and non-overlapping stochastic block models

Pierre Latouche

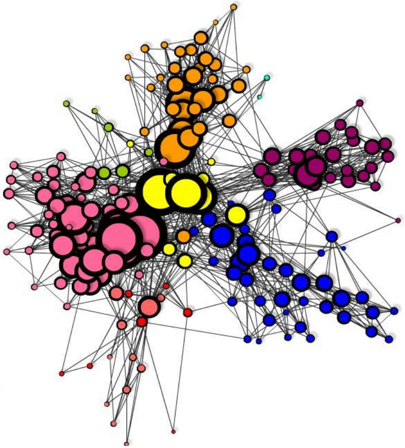
Université Paris 1 Panthéon-Sorbonne
Laboratoire SAMM

MSTGA 2012

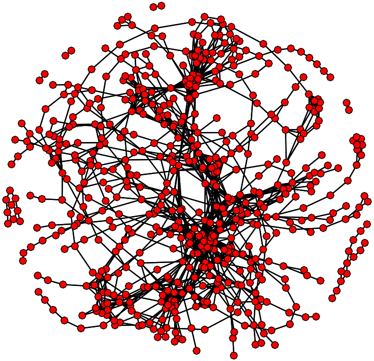
Contents

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

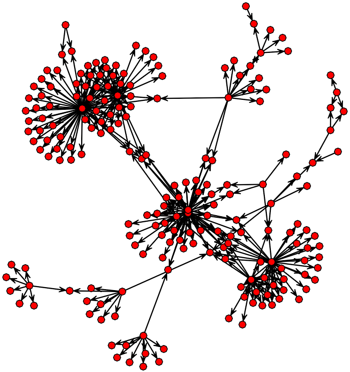


Sample of 250 blogs (nodes) with their links (edges) of the French political Blogosphere.



The metabolic network of bacteria *Escherichia coli* (Lacroix et al., 2006).

In Biology



Subset of the yeast transcriptional regulatory network (Milo et al., 2002).

Real networks

- **Properties** :
  - Sparsity : $m = \mathrm{O}(N)$
  - Existence of a giant component
  - Heterogeneity
  - Preferential attachment
  - Small world

$\hookrightarrow$ Topological structure (groups of vertices)

Real networks

- **Properties** :
  - Sparsity : $m = \mathrm{O}(N)$
  - Existence of a giant component
  - Heterogeneity
  - Preferential attachment
  - Small world

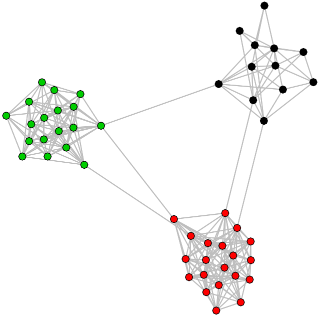
$\hookrightarrow$ Topological structure (groups of vertices)

Graph clustering

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

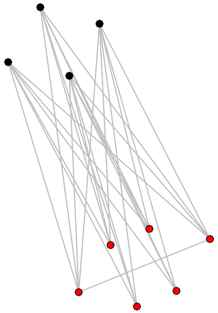
Graph clustering

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



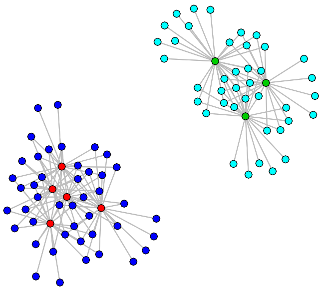
Graph clustering

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



Graph clustering

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



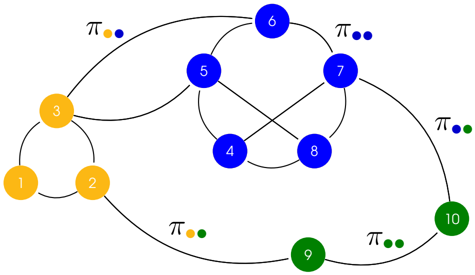
Stochastic Block Model (SBM)

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

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

- A mixture model for graphs :

$$X_{ij} \sim \sum_{k=1}^{K}\sum_{l=1}^{K} \alpha_k \alpha_l \mathcal{B}(\pi_{kl})$$



π
6
5
7
3
4
8
1
2
10
9

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(\mathbf{X} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi}) = \log \{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi})\}$
    $\hookrightarrow K^N$ 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)

Variational EM
Daudin et al. (2008)

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(\mathbf{X} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi}) = \log \{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Pi})\}$
    $\hookrightarrow$ K^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)

Variational EM
Daudin et al. (2008)

Maximum likelihood estimation

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

Problem

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

Variational EM

Daudin et al. (2008)

Model selection

Criteria

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

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

ICL

Biernacki et al. (2000) $\hookrightarrow$ Daudin et al. (2008)

Variational Bayes EM $\hookrightarrow$ $ILvb$

Latouche et al. (2012)

Model selection

Criteria

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

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

ICL

Biernacki et al. (2000) $\hookrightarrow$ Daudin et al. (2008)

Variational Bayes EM $\hookrightarrow ILvb$

Latouche et al. (2012)

Bayesian framework

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

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

Decomposition

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

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

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

Variational Bayes EM
Latouche et al. (2009)

E-step

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

M-step

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

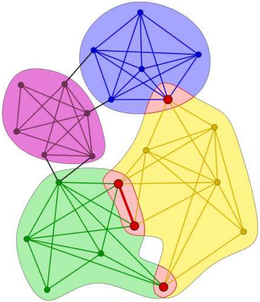
A new model selection criterion : ILvb
Latouche et al. (2012)

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

ILvb

$$IL_{vb} = \log \left\{ \frac{\Gamma(\sum_{k=1}^{K} n_k^0) \prod_{k=1}^{K} \Gamma(n_k)}{\Gamma(\sum_{k=1}^{K} n_k) \prod_{k=1}^{K} \Gamma(n_k^0)} \right\} + \sum_{k \leq l}^{K} \log \left\{ \frac{\Gamma(\eta_{kl}^0 + \zeta_{kl}^0)\Gamma(\eta_{kl})\Gamma(\zeta_{kl})}{\Gamma(\eta_{kl} + \zeta_{kl})\Gamma(\eta_{kl}^0)\Gamma(\zeta_{kl}^0)} \right\} - \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik}$$

Contents



Palla et al. (2006)

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 :

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

Overlapping Stochastic Block model (OSBM)

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

$$\mathbf{Z}_i \sim \prod_{k=1}^{K} \mathcal{B}(Z_{ik};\ \alpha_k) = \prod_{k=1}^{K} \alpha_k^{Z_{ik}}(1-\alpha_k)^{1-Z_{ik}}$$

Overlapping Stochastic Block model (OSBM)

- Latouche et al. (2011)
- $\mathbf{X}\,|\,\mathbf{Z}$ edges drawn independently :

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

- $\mathbf{\Pi}_{\mathbf{Z}_i,\mathbf{Z}_j} = g\big(a_{\mathbf{Z}_i,\mathbf{Z}_j}\big)$
- $a_{\mathbf{Z}_i,\mathbf{Z}_j} = \underbrace{\mathbf{Z}_i^\intercal\, \mathbf{W}\, \mathbf{Z}_j}_{i \leftrightarrow j} + \underbrace{\mathbf{Z}_i^\intercal\, \mathbf{U}}_{i \rightarrow ?} + \underbrace{\mathbf{V}^\intercal\, \mathbf{Z}_j}_{? \rightarrow j} + \underbrace{W^*}_{\text{bias}}$
- $g(t) = 1/\left(1 + \exp(-t)\right)$ is the logistic function

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 : $\left\{ \boldsymbol{\alpha}, \tilde{\mathbf{W}} \right\}$

Contents

Bayesian framework

- **Conjugate prior distributions** :
  - $p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \text{Beta}(\alpha_k;\, \eta_k^0, \zeta_k^0)$
  - $p(\tilde{\mathbf{W}}^{\text{vec}}) = \mathcal{N}(\tilde{\mathbf{W}}^{\text{vec}};\, \tilde{\mathbf{W}}_0^{\text{vec}}, \mathbf{S}_0)$
- The vec operator : if

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

then

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

Bayesian framework

- $\mathbf{x}^\intercal \mathbf{A}\, \mathbf{y} = (\mathbf{y} \otimes \mathbf{x})^\intercal \mathbf{A}^{\text{vec}}$
- In practice : set $\tilde{\mathbf{W}}_0^{\text{vec}} = \mathbf{0}$ and $\mathbf{S}_0 = \frac{\mathbf{I}}{\beta}$

Problem

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

Decomposition

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

where

$$\mathcal{L}(r) = \sum_{\mathbf{Z}} \int r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \log \left(\frac{p(\mathbf{X} \,|\, \mathbf{Z}, \tilde{\mathbf{W}}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}})}{r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})}\right) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}}$$

Lower bound

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

Problem

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

Local bound

- Use the bound of Jaakkola and Jordan (2000) for Bayesian logistic regression

$$\log p(\mathbf{X} \,|\, \mathbf{Z}, \tilde{\mathbf{W}}) \geq \log h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}), \forall\, \boldsymbol{\xi} \in \mathbb{R}^{N \times N}$$

where

$$\log h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) = \sum_{i \neq j}^{N} \Big\{ (X_{ij} - \frac{1}{2}) a_{\mathbf{Z}_i, \mathbf{Z}_j} - \frac{\xi_{ij}}{2} + \log g(\xi_{ij}) - \lambda(\xi_{ij})(a^2_{\mathbf{Z}_i, \mathbf{Z}_j} - \xi^2_{ij}) \Big\}$$

and

$$\lambda(\xi) = \frac{1}{4\xi} \tanh(\frac{\xi}{2}) = \frac{1}{2\xi} \{ g(\xi) - \frac{1}{2} \}$$

Lower Bound

$$\log p(\mathbf{X}) = \log \left\{ \sum_{\mathbf{Z}} \int p(\mathbf{X} \,|\, \mathbf{Z}, \tilde{\mathbf{W}}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}}) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}} \right\}$$
$$\geq \mathcal{L}(\boldsymbol{\xi})$$

where

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

ξ Transformation

Decomposition

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

where

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

Lower bound

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

Inference

Local optimization

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

E-step

- $r(Z_{ik}) = \mathcal{B}(Z_{ik};\, \tau_{ik})$

M-step

- $r(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \text{Beta}(\alpha_k;\, \eta_k^N, \zeta_k^N)$
- $r(\tilde{\mathbf{W}}^{vec}) = \mathcal{N}(\tilde{\mathbf{W}}^{\text{vec}};\, \tilde{\mathbf{W}}_N^{\text{vec}}, \mathbf{S}_N)$

Model selection

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

ILosbm

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

L_2 regularization

$$p(\tilde{\mathbf{W}}^{\text{vec}}) = \mathcal{N}(\tilde{\mathbf{W}}^{\text{vec}}; \mathbf{0}, \frac{\mathbf{I}}{\beta})$$

- β too small $\hookrightarrow$ overfit
- β too large $\hookrightarrow$ IL_{osbm} maximized for very large values of K

Question

Can we estimate β from the data ?

- **Conjugate prior distributions :**
  - $p(\tilde{\mathbf{W}}^{\text{vec}}) = \mathcal{N}(\tilde{\mathbf{W}}^{\text{vec}};\ \mathbf{0}, \frac{\mathbf{I}}{\beta})$
  - $p(\beta) = \text{Gamma}(\beta;\ a_0, b_0)$

Inference

- Use a variational Bayes EM algorithm to maximize:

$$\mathcal{L}(r;\, \boldsymbol{\xi}) =$$

$$\sum_{\mathbf{Z}} \int r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \beta) \log \Big(\frac{h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}}) p(\beta)}{r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \beta)} \Big) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}}\, d$$

- $r(\beta) = \text{Gamma}(\beta;\, a_N, b_N)$, where

$$a_N = a_0 + \frac{(K+1)^2}{2}$$

and

$$b_N = b_0 + \frac{1}{2}\text{Tr}\big(S_N + (\tilde{\mathbf{W}}_N^{\text{vec}})^{\intercal}\, \tilde{\mathbf{W}}_N^{\text{vec}}\, \big)$$

Criterion

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

Inference

- Use a variational Bayes EM algorithm to maximize:

$$\mathcal{L}(r;\, \boldsymbol{\xi}) =$$

$$\sum_{\mathbf{Z}} \int r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \beta) \log \Big(\frac{h(\mathbf{Z}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) p(\mathbf{Z} \,|\, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\tilde{\mathbf{W}}) p(\beta)}{r(\mathbf{Z}, \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \beta)}\Big) d\,\boldsymbol{\alpha}\, d\,\tilde{\mathbf{W}}\, d$$

- $r(\beta) = \mathrm{Gamma}(\beta;\, a_N, b_N)$, where

$$a_N = a_0 + \frac{(K+1)^2}{2}$$

and

$$b_N = b_0 + \frac{1}{2}\mathrm{Tr}\big(S_N + (\tilde{\mathbf{W}}_N^{\mathrm{vec}})^\intercal \, \tilde{\mathbf{W}}_N^{\mathrm{vec}}\big)$$

Criterion

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

ILosbm

$$
\begin{aligned}
IL_{osbm} = & \sum_{i \neq j}^{N} \left\{ \log g(\xi_{ij}) - \frac{\xi_{ij}}{2} + \lambda(\xi_{ij})\xi_{ij}^{2} \right\} \\
& + \sum_{k=1}^{K} \log \left\{ \frac{\Gamma(\eta_k^0 + \zeta_k^0)\Gamma(\eta_k^N)\Gamma(\zeta_k^N)}{\Gamma(\eta_k^0)\Gamma(\zeta_k^0)\Gamma(\eta_k^N + \zeta_k^N)} \right\} + \log \frac{\Gamma(a_N)}{\Gamma(a_0)} + a_0 \log b_0 \\
& + a_N(1 - \frac{b_0}{b_N} - \log b_N) + \frac{1}{2}(\tilde{\mathbf{W}}_N^{\text{vec}})^{\intercal} \, \mathbf{S}_N^{-1} \, \tilde{\mathbf{W}}_N^{\intercal} \\
& + \frac{1}{2} \log | \, \mathbf{S}_N \, | - \sum_{i=1}^{N} \sum_{k=1}^{K} \{ \tau_{ik} \log \tau_{ik} + (1 - \tau_{ik}) \log(1 - \tau_{ik}) \} \, .
\end{aligned}
$$

Experiments on simulated data

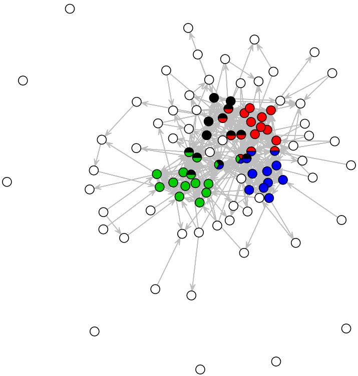
- **Two topological structures** :
  - Community structures (affiliation) :

$$\mathbf{W} = \begin{pmatrix} \boldsymbol{\lambda} & -\epsilon & \dots & -\epsilon \\ -\epsilon & \boldsymbol{\lambda} & & \vdots \\ \vdots & & \ddots & -\epsilon \\ -\epsilon & \dots & -\epsilon & \boldsymbol{\lambda} \end{pmatrix}$$

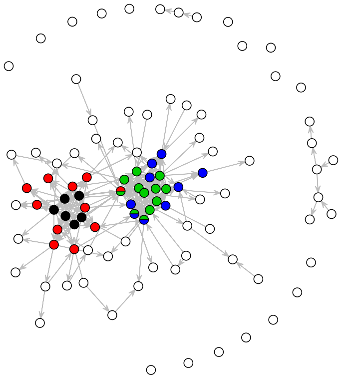
  - Community structures and stars :

$$\mathbf{W} = \begin{pmatrix} \boldsymbol{\lambda} & \boldsymbol{\lambda} & -\epsilon & \dots & \dots & \dots & -\epsilon \\ \boldsymbol{-\epsilon} & \boldsymbol{-\lambda} & -\epsilon & \dots & \dots & \dots & \vdots \\ \vdots & -\epsilon & \boldsymbol{\lambda} & \boldsymbol{\lambda} & -\epsilon & \dots & \vdots \\ \vdots & \vdots & \boldsymbol{-\epsilon} & \boldsymbol{-\lambda} & -\epsilon & \dots & \vdots \\ \vdots & \vdots & \vdots & -\epsilon & \ddots & -\epsilon & -\epsilon \\ \vdots & \vdots & \vdots & \vdots & -\epsilon & \boldsymbol{\lambda} & \boldsymbol{\lambda} \\ -\epsilon & \dots & \dots & \dots & \dots & \boldsymbol{-\epsilon} & \boldsymbol{-\lambda} \end{pmatrix}$$

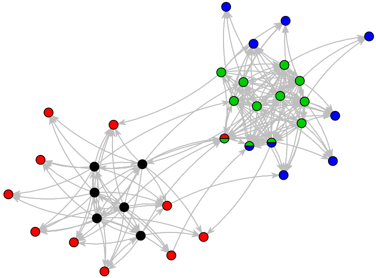
Community structures



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



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



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

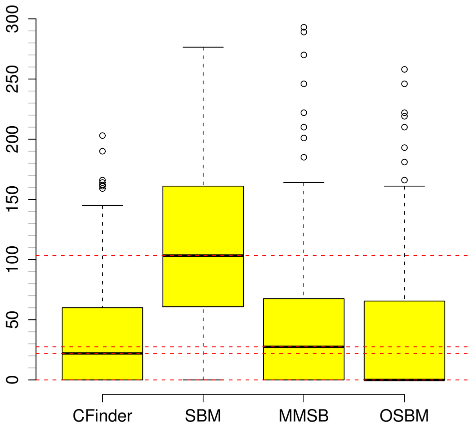
Experiments on simulated data

- $N = 100$
- $\lambda = 4$
- $\epsilon = 1$
- $W^* = -5.5$
- $\mathbf{U} = \mathbf{V} = \begin{pmatrix} \epsilon & \dots & \epsilon \end{pmatrix}$
- $\alpha_k = 0.25$
- $K = 4$
- 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)

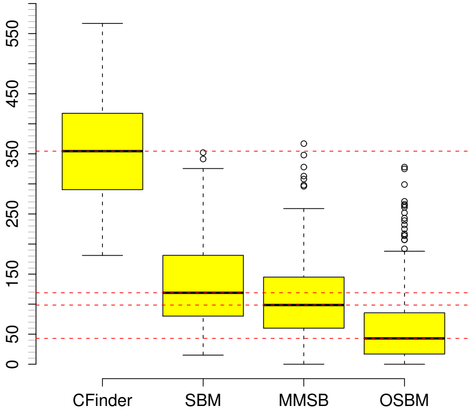
How to compare the methods ?

- CFinder and OSBM can deal with outliers ($\mathbf{Z}_i = \mathbf{0}$)
- SBM and MMSB are run with $K + 1$ 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.



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

Model selection

- Community structure
- $N = 100$
- $\epsilon = 1$
- $W^* = -5.5$
- $\alpha_k = 1/K$
- $K_{True} \in \{3, \ldots, 7\}$
- $K \in \{2, \ldots, 8\}$
- 100 simulations

Table: $K_{True} \backslash K_{IL_{osbm}} (p_{intra} \approx 0.92)$

	2	3	4	5	6	7	8
3	0	**99**	1	0	0	0	0
4	0	0	**99**	1	0	0	0
5	0	0	0	**93**	5	2	0
6	0	0	0	7	**64**	22	7
7	0	0	0	0	16	**47**	37

Results

Table: $K_{True} \backslash K_{IL_{osbm}} (p_{intra} \approx 0.62)$

	2	3	4	5	6	7	8
3	0	**99**	1	0	0	0	0
4	0	0	**85**	9	5	0	1
5	0	0	4	**53**	26	9	8
6	0	0	0	18	**34**	27	21
7	0	0	0	4	18	**30**	48

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 $K = 4$ clusters using OSBM. 196 vertices, 2864 edges.

Conclusion

- Computational cost : $\mathrm{O}(K^4N^2) \neq \mathrm{O}(K^2N^2)$
- New model selection criterion : ILosbm
- R package **OSBM** soon available on the CRAN
- Can be used to analyze SBM networks

References

- K. Nowicki and T.A.B. Snijders (2001), Estimation and prediction for stochastic blockstructures. 96, 1077-1087
- E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing (2008), Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9, 1981-2014
- J-J. Daudin, F. Picard et S. Robin (2008), A mixture model for random graphs. Statistics and Computing, 18, 2, 151-171
- P. Latouche, E. Birmelé, C. Ambroise (2011), Overlapping stochastic block models with application to the French political blogosphere network. Annals of Applied Statistics, 5, 1, 309-336
- P. Latouche, E. Birmelé, C. Ambroise (2012), Variational Bayesian inference and complexity control for stochastic block models. Statistical Modelling, 12, 1, 93-115