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

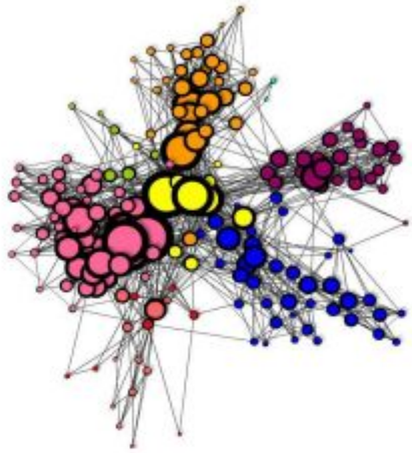
Pierre Latouche

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.

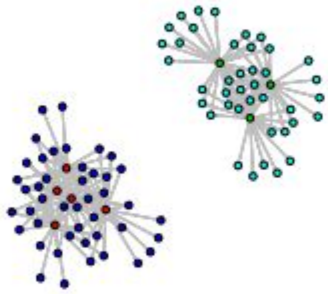
Real networks

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

$\hookrightarrow$ Topological structure (groups of vertices)

Graph clustering

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



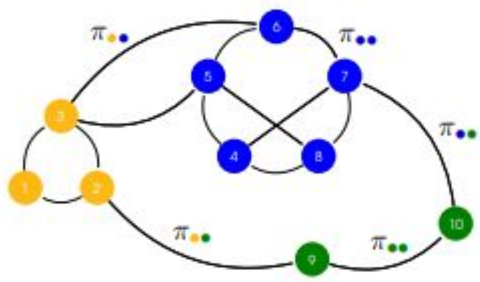
Stochastic Block Model (SBM)

- Nowicki and Snijders (2001)
  - Earlier work : Govaert et al. (1977)
- $\mathbf{Z}_i$ independent hidden variables :
  - $\mathbf{Z}_i \sim \mathcal{M}\Big(1,\ \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_Q)\Big)$
  - $Z_{iq} = 1$: 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 :

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



Bayesian framework

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

Variational Bayes EM
Latouche et al. (2009)

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

Decomposition

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

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

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

Variational Bayes EM
Latouche et al. (2009)

E-step

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

M-step

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

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

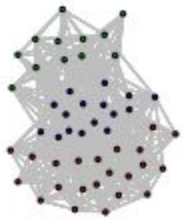
$$IL_{vb} = \log\left\{\frac{\Gamma(\sum_{q=1}^{Q} n_q^0)\prod_{q=1}^{Q}\Gamma(n_q)}{\Gamma(\sum_{q=1}^{Q} n_q)\prod_{q=1}^{Q}\Gamma(n_q^0)}\right\} + \sum_{q\leq l}^{Q}\log\left\{\frac{\Gamma(\eta_{ql}^0+\zeta_{ql}^0)\Gamma(\eta_{ql})\Gamma(\zeta_{ql})}{\Gamma(\eta_{ql}+\zeta_{ql})\Gamma(\eta_{ql}^0)\Gamma(\zeta_{ql}^0)}\right\}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

- **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 :

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



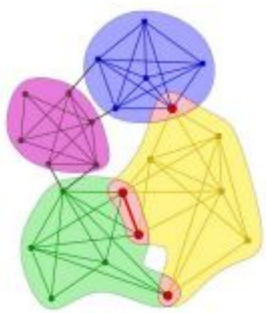
Experiments on simulated data

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

Conclusion (1)

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

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

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 :

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

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

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

Bayesian framework

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

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

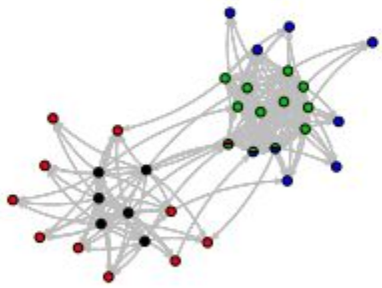
then

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

Problem

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

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 $d(\mathbf{P}, \hat{\mathbf{P}})$

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 : $O(Q^4 N^2)$
- New model selection criterion : ILosbm
- R package **OSBM** soon available on the CRAN

https://analyseshs.hypotheses.org/files/2011/09/latouche.pdf