Bayesian Methods for Graph Clustering

P. Latouche, E. Birmelé

Laboratoire "Statistique et Génome" (UMR CNRS 8071, INRA 1152)

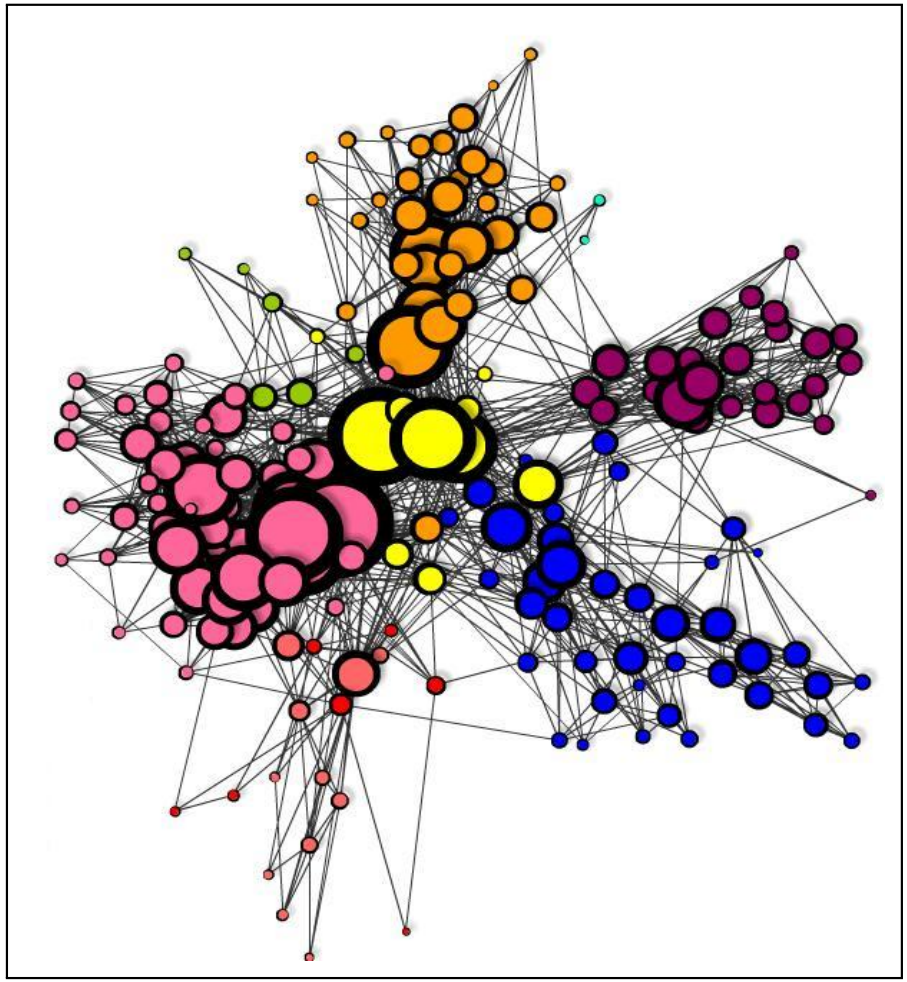
Journées MAS, August 2008

Outline

Outline

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.

Estimation of random graph models

Random graph models

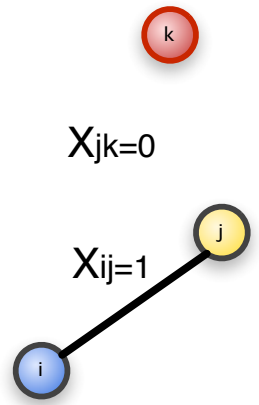
- Random graph approaches (Govaert 1977, Frank and Harary 1982, Handcock 2006, Newman and Leicht 2007, Hofman and Wiggins 2008).
- The Block-Clustering model (Snijders and Nowicki 1997).
- Erdös-Rényi Mixture Model for Network (MixNet; Daudin et al 2008).

Estimation of the model parameters

- Bayesian strategies cannot handle large networks.
- Maximum likelihood strategies.

The MixNet probabilistic model

- **Origin**
  - model developed by J. Daudin et al. (2008),
  - ER model generalization,
  - application fields: biology, internet, social network...
- **Modelling connection heterogeneity**
  - hyp.: there exists a hidden structure with Q classes,
  - $\mathbf{Z} = (\mathbf{Z}_i)_i$, $Z_{iq} = \mathbb{I}\{i \in q\}$ are indep. hidden variables,
  - $\boldsymbol{\alpha} = \{\alpha_q\}$, the *prior* proportions of groups,
  - $(\mathbf{Z}_i) \sim \mathcal{M}(1, \boldsymbol{\alpha})$.
- **Distribution of X**
  - Conditional distribution:
    $X_{ij}|\{Z_{iq}Z_{j\ell} = 1\} \sim \mathcal{B}(\pi_{ql})$
    where $\mathcal{B}()$ is the Bernoulli distribution,
  - $X_{ij}|Z$ are independant.
  - Marginal distribution: $X_{ij} \sim \sum_{q\ell} \alpha_q \alpha_\ell \mathcal{B}(\pi_{ql})$,



Maximum likelihood estimation

- **Likelihood(s) of the model**:
  $\longrightarrow$ Observed data : $p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\pi}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\pi})$.
  $\longrightarrow$ Complete data : $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\pi})$.
  $\longrightarrow$ EM-like strategies require the knowledge of $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi})$.

Problem

In our case, $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi})$ is not tractable (no conditional independence).

Variational method

Daudin et. al, 2008

Decomposition and variational EM

$$\ln p(\mathbf{X}|\alpha,\pi) = \mathcal{L}\Big(q(.);\ \alpha,\pi\Big) + \mathrm{KL}\left(q(.)\ ||\ p(.|\mathbf{X},\alpha,\pi)\right),$$

where

$$\mathcal{L}\Big(q(.);\ \alpha,\pi\Big) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\{\frac{p(\mathbf{X},\mathbf{Z}|\alpha,\pi)}{q(\mathbf{Z})}\},$$

and

$$\mathrm{KL}\left(q(.)\ ||\ p(.|\mathbf{X},\alpha,\pi)\right) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\{\frac{p(\mathbf{Z}|\mathbf{X},\alpha,\pi)}{q(\mathbf{Z})}\}.$$

Approximation

$$q(\mathbf{Z}) = \prod_{i=1}^{N} q(\mathbf{Z_i}) = \prod_{i=1}^{N} \mathcal{M}(\mathbf{Z_i};\ 1, \tau_\mathbf{i}).$$

Estimation of the number of classes

Criteria

Since $p(\mathbf{X}|\alpha, \pi)$ is not tractable, many criteria cannot be used:

1. Akaike Information Criterion: $AIC = \ln p(\mathbf{X}|\alpha_{\mathbf{ML}}, \pi_{\mathbf{ML}}) - M$.
2. Bayesian Information Criterion: $BIC = \ln p(\mathbf{X}|\alpha_{\mathbf{MAP}}, \pi_{\mathbf{MAP}}) - \frac{1}{2} M \ln n$.

Integrated Classification Likelihood (ICL)

1. Following the work of Biernacki et al. (2000), Mariadassou and Robin (2007) used a criterion based on an asymptotic approximation of the Integrated Classification Likelihood (ICL).
2. $ICL = \max_{\alpha,\pi} \ln p(\mathbf{X}, \tilde{\mathbf{Z}}|\alpha, \pi) - \frac{1}{2}\left(\frac{Q(Q+1)}{2} \ln\left(n(n+1)\right) - (Q-1)\ln(n)\right)$

Outline

Conjugate prior distributions

- **Mixing coefficients**: $\alpha \sim \mathit{Dirichlet}(\alpha;\, \mathbf{n^0})$
  $\longrightarrow$ $\mathbf{n^0} = (n_1^0, \ldots, n_Q^0)$.
  $\longrightarrow$ n_q^0 is the prior number of vertices in class q.
- **Connectivity matrix**: $\pi \sim \prod_{q,l}^{Q} \mathrm{Beta}(\pi_{ql};\, \eta_{ql}^0, \zeta_{ql}^0)$
  $\longrightarrow$ η_{ql}^0 is the prior number of edges connecting vertices of class q to vertices of class l.
  $\longrightarrow$ ζ_{ql}^0 is the prior number of *non*-edges connecting vertices of class q to vertices of class l.

Variational Bayes

Decomposition

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

where

$$\mathcal{L}\Big(q(.)\Big) = \sum_{\mathbf{Z}} \int \int q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \ln\{\frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})}\} d\boldsymbol{\alpha} d\boldsymbol{\pi},$$

and

$$\mathrm{KL}\left(q(.) \,||\, p(.|\mathbf{X})\right) = -\sum_{\mathbf{Z}} \int \int q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \ln\{\frac{p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\mathbf{X})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})}\} d\boldsymbol{\alpha} d\boldsymbol{\pi}.$$

Variational Bayes

Factorization

$$q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}) = q(\boldsymbol{\alpha})q(\boldsymbol{\pi})q(\mathbf{Z}) = q(\boldsymbol{\alpha})q(\boldsymbol{\pi}) \prod_{i=1}^{N} q(\mathbf{Z_i}).$$

Optimization

1. $\ln \tilde{q}(\mathbf{Z_i}) = \mathrm{E}_{\mathbf{Z}\setminus \mathbf{i}, \boldsymbol{\alpha}, \boldsymbol{\pi}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})] + \text{cste.}$
2. $\ln \tilde{q}(\boldsymbol{\alpha}) = \mathrm{E}_{\mathbf{Z}, \boldsymbol{\pi}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})] + \text{cste.}$
3. $\ln \tilde{q}(\boldsymbol{\pi}) = \mathrm{E}_{\mathbf{Z}, \boldsymbol{\alpha}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})] + \text{cste.}$

Optimization

Variational Bayes E-step

$$q(\mathbf{Z_i}) = \mathcal{M}(\mathbf{Z_i};\ 1, \tau_\mathbf{i} = \{\tau_{i1}, \ldots, \tau_{iQ}\}).$$

Variational Bayes M-step (1)

$$q(\boldsymbol{\alpha}) = \mathrm{Dir}(\alpha;\ \mathbf{n}),$$

where $n_q = n_q^0 + \sum_{i=1}^N \tau_{iq}$.

Variational Bayes M-step (2)

$$q(\boldsymbol{\pi}) = \prod_{q,l}^{Q} \mathrm{Beta}(\pi_{ql} | \eta_{ql}, \zeta_{ql}),$$

where $\eta_{ql} = \eta_{ql}^0 + \sum_{i \neq j}^N X_{ij} \tau_{iq} \tau_{jl}$ and $\zeta_{ql} = \zeta_{ql}^0 + \sum_{i \neq j}^N (1 - X_{ij}) \tau_{iq} \tau_{jl}$.

Model selection

- **The model evidence depends on Q.**
- **Bayes' rule leads to $p(Q|\mathbf{X}) \propto p(\mathbf{X}|Q)p(Q)$.**
- **If $p(Q)$ is broad, maximizing $p(Q|\mathbf{X})$ is equivalent to maximizing $p(\mathbf{X}|Q)$.**
- **Since $p(\mathbf{X}|Q)$ is intractable we propose to use the lower bound $\mathcal{L}\Big(q(.)\Big)$ and to add a term $\ln Q!$ to take the multimodality into account.**
- **First non-asymptotic criterion based on an approximation of the model evidence.**

Outline

Experiments on affiliation models

- **Probability of intra-connection :** λ**.**
- **Probability of inter-connection :** ϵ**.**
- **Number of vertices :** $n = 50$**.**
- **For each graph model (**$\lambda + \epsilon = 1$**) and for each number of classes** $Q_{True} \in \{2, 3, 4, 5\}$**, we generated** 100 **graphs**.
- 5 **initializations using spectral clustering techniques**.
- **Select the best number of estimated classes according to each criterion**.

Affiliation model (1)

a) Q_{True}/Q_{ICL}

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	1	**98**	1	0
5	0	0	10	61	**29**	0

b) Q_{True}/Q_{VB}

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	0	**98**	2	0
5	0	0	1	29	**65**	5

Table: $\lambda = 0.85$ and $\epsilon = 0.15$.

Affiliation model (2)

a) Q_{True}/Q_{ICL}

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	14	**86**	0	0
5	0	17	36	44	**3**	0

b) Q_{True}/Q_{VB}

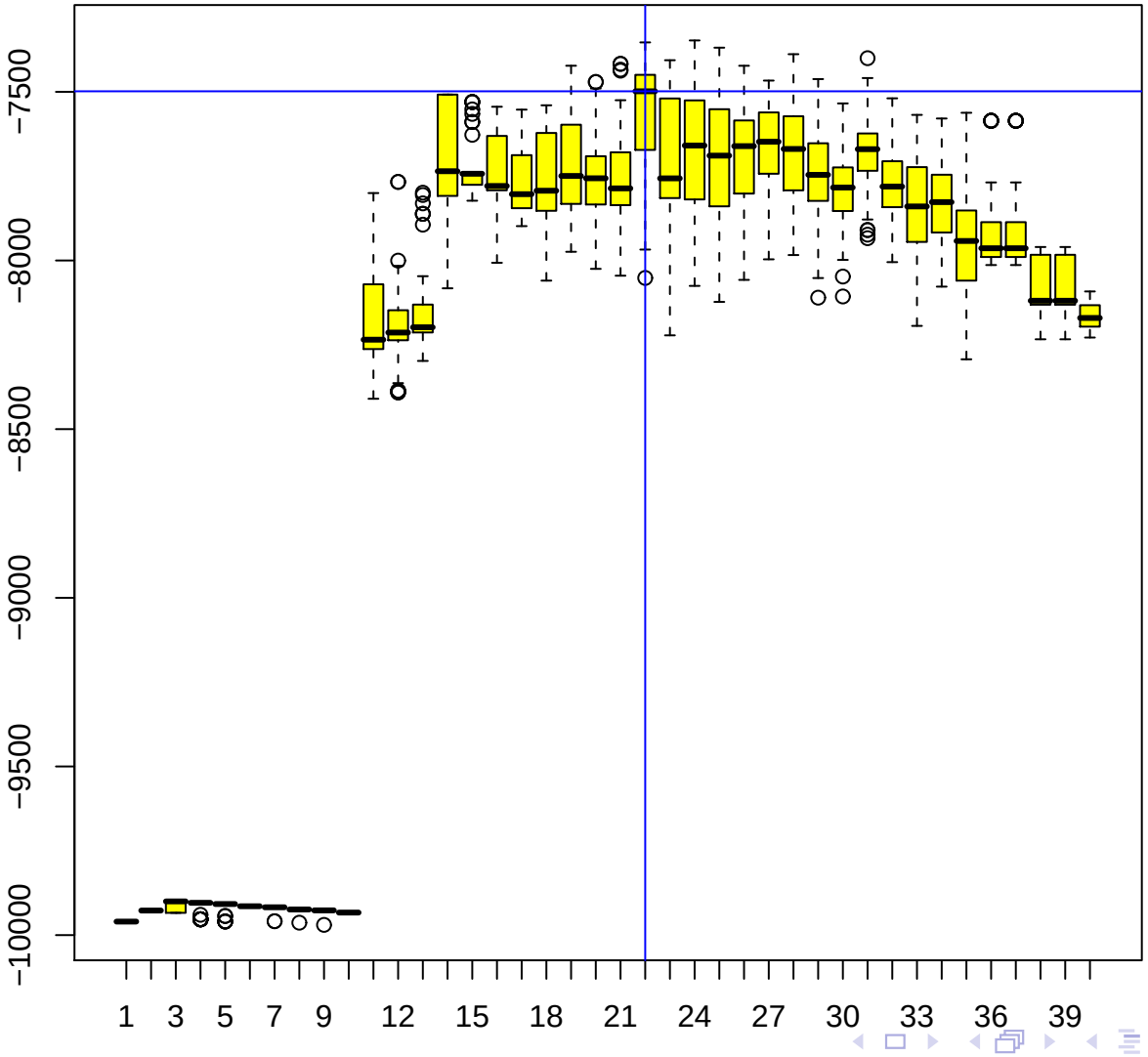
	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	5	**94**	1	0
5	0	4	18	43	**29**	6

Table: $\lambda = 0.8$ and $\epsilon = 0.2$.

Metabolic network of E. coli

- 593 **vertices**.
- 1782 **edges**.
- 60 **initializations**.
- **Compute the lower bound**.

Model selection



In summary

- **Flexibility of MixNet** :
  - MixNet is a probabilistic model which captures features of real-networks,
  - It considers classes of connectivity.
- **Bayesian framework**
  - Estimate of the number of classes more robust,
  - Can handle large graphs.
- **References** :
  - Daudin J-J., Picard F., Robin S. (2008) , A mixture model for random graphs, Statistic and Computing
  - Zanghi, H, Ambroise, C. and Miele, V. (to appear), Fast online Graph Clustering via Erdös-Rényi Mixture, Pattern Recognition
  - J.M. Hofman and C.H. Wiggins (2008), A bayesian approach to network modularity, Physical review letters
- **Softwares** :
  - MixNet, a C^{++} code (V. Miele) `http://stat.genopole.cnrs.fr/software/mixnet`
  - MixNet, a R wrapper of MixNet C^{++} code (available on demand).