# Stochastic block models for random graphs (and some variants)

Catherine MATIAS

CNRS, Laboratoire Statistique & Génome, Évry

http://stat.genopole.cnrs.fr/∼cmatias






# Outline

Stochastic block models (and variants)

Parameter estimation and node clustering

Convergence results

# Data: Biological networks

## Different networks types

- Protein-protein interaction networks (PPI),
- Metabolic networks
- Genes co-expression networks
- Genes regulation networks
- ...

## Some challenges

- Analyse big data sets, noisy data,
- Identify structures (topological patterns, cliques, nodes groups, etc),
- Compare networks between different species,
- Modelling evolution of these networks,
- ...

# Some models for biological networks

## Some existing models, advantages and drawbacks

- Erdös-Rényi, simple and mathematically well-understood, too homogeneous;
- Models based on degree distribution, scale-free property, only a partial descriptor of the graph, greedy numerical simulations with fixed-degrees models ;
- Generative processes (like preferential attachment), dynamic model, depends on parameters (initialisation, stop, ...), can we caracterize the result?
- Exponential random graph : see previous talk from Nial Friel !
- ...

We would like to cluster the nodes into groups.

# Mixture models approach

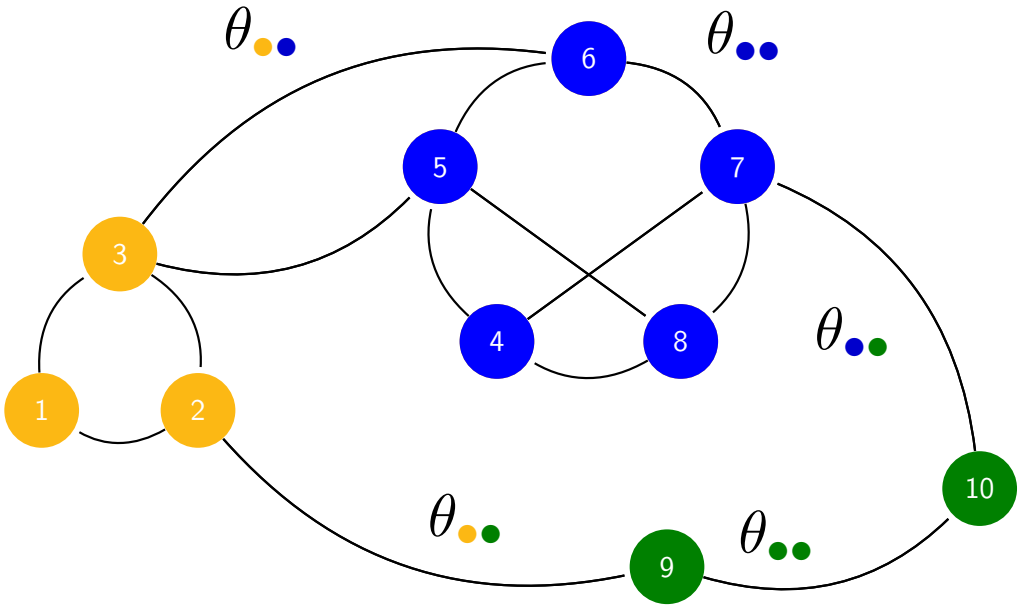
## Idea: probability model based clustering

Assume that the nodes of the graph belong to unobserved groups, that describe their connectivity to the other nodes.

## Advantages

- Induces heterogeneity in the data, keeping it simple,
- Clustering of the nodes groups induced by the model,
- Model encompasses the community detection framework.

# Stochastic block model (binary graphs)

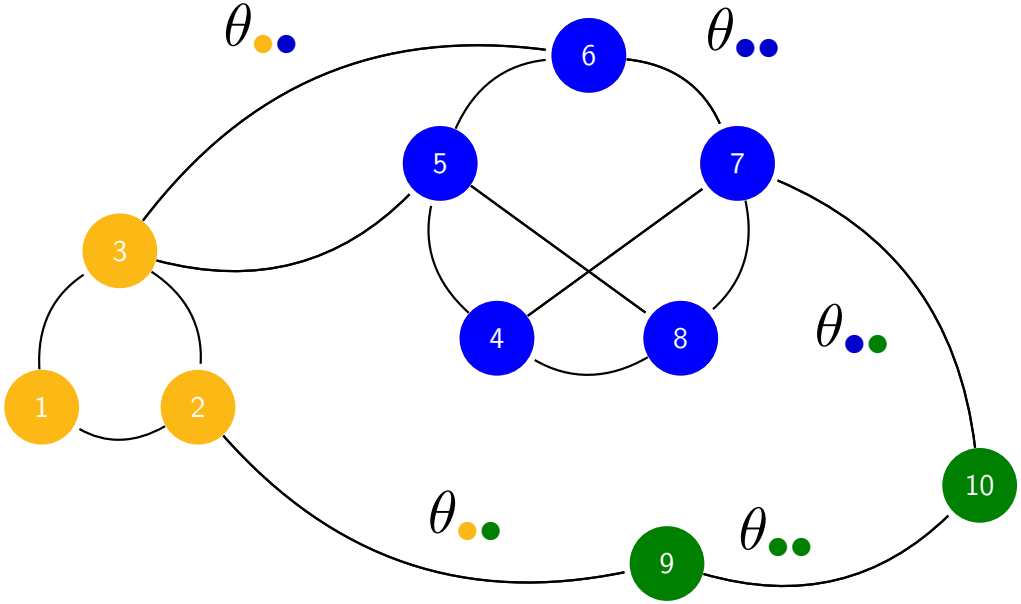



$$n = 10, Z_{5\bullet} = 1$$
$$X_{12} = 1, X_{15} = 0$$

## Binary case

- $Q$ groups (=colors ●●●).
- $\{Z_i\}_{1 \leq i \leq n}$ i.i.d. vectors $Z_i = (Z_{i1}, \ldots, Z_{iQ}) \sim \mathcal{M}(1, \boldsymbol{\pi})$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$ groups proportions. $Z_i$ not observed (latent).
- Observations: presence/absence of an edge $\{X_{ij}\}_{1 \leq i < j \leq n}$,
- Conditional on $\{Z_i\}$'s, the r.v. $X_{ij}$ are independent $\mathcal{B}(\theta_{Z_i Z_j})$.

# Stochastic block model (weighted graphs)




$$n = 10, Z_{5\bullet} = 1$$
$$X_{12} \in \mathbb{R}, X_{15} = 0$$

## Weighted case

- Observations: weights $X_{ij}$ , where $X_{ij} = 0$ or $X_{ij} \in \mathbb{R}^s \setminus \{0\}$,
- Conditional on the $\{Z_i\}$'s, the random variables $X_{ij}$ are independent with distribution

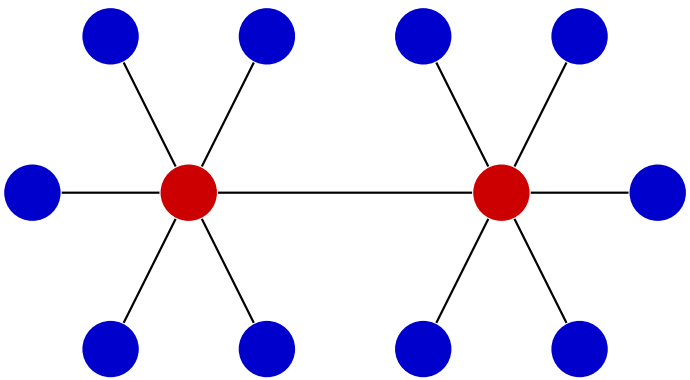
$$\mu_{Z_i Z_j}(\cdot) = p_{Z_i Z_j} f(\cdot, \theta_{Z_i Z_j}) + (1 - p_{Z_i Z_j})\delta_0(\cdot)$$

(Assumption: $f$ has continuous cdf at zero).

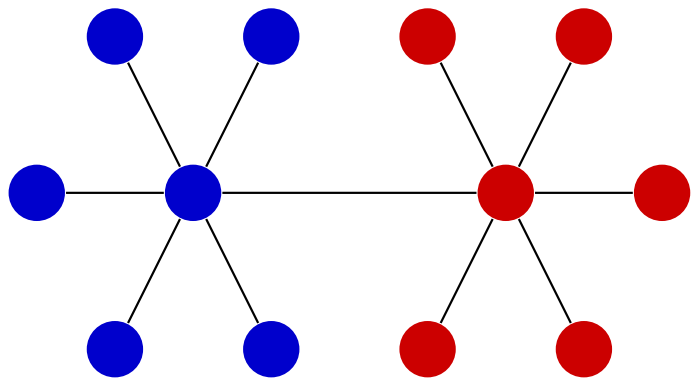
# SBM clustering vs other clusterings

## SBM clustering

- Nodes clustering induced by the model reflects a common connectivity behaviour;
- Many clustering methods try to group nodes that belong to the same clique (ex: community detection)
- Toy example



SBM cluster



Clustering based on cliques

# Particular cases and generalisations

## Particular cases

- Affiliation model: connectivity matrix $\boldsymbol{\theta}$ has only 2 parameters

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha & \dots & \beta \\ \vdots & \ddots & \vdots \\ \beta & \dots & \alpha \end{pmatrix} \qquad \alpha \neq \beta$$

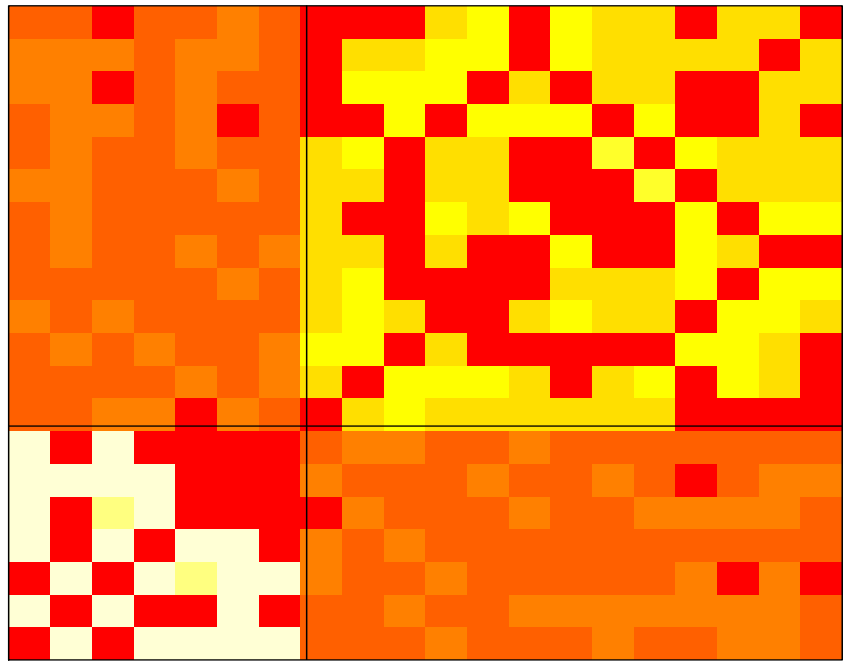
- Affiliation + $\alpha \gg \beta \implies$ community detection (cliques clustering).

## Generalisations

- Overlapping groups [Latouche *et al.* 11, Airoldi *et al.* 08] for binary graphs;
- Adding covariates [Zanghi *et al.* 10b] ;
- Latent block models (LBM), for array data.

# From SBM to LBM

- A graph is encoded through its adjacency matrix.
- Clustering the nodes corresponds to simultaneous and identical clustering of the rows and columns.



Generalise this to non square array data, without constraining identical rows and columns groups.

# Latent block models I

## LBM notation

- Observations: array $\mathbf{X}_{n,m} := \{X_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq m}$ with $X_{ij} \in \mathcal{X}$,
- $Q \geq 1$ and $L \geq 1$ number of row and column groups, respectively.
- Groups prior distributions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$ over $\mathcal{Q} = \{1, \ldots, Q\}$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_L)$ over $\mathcal{L} = \{1, \ldots, L\}$, such that $\sum_q \pi_q = \sum_l \rho_l = 1$.
- Latent variables $\mathbf{Z}_n := Z_1, \ldots, Z_n$ iid $\sim \boldsymbol{\pi}$ over $\mathcal{Q}$ and $\mathbf{W}_m := W_1, \ldots, W_m$ i.i.d. $\sim \boldsymbol{\rho}$ over $\mathcal{L}$.

# Latent block models II

## Two models in the same framework

- 2 cases occur

  LBM : $\{Z_i\}_{1\leq i\leq n}$ and $\{W_j\}_{1\leq j\leq m}$ independent.
  SBM : $n = m$, $\mathcal{Q} = \mathcal{L}$, $Z_i = W_i$ for all $1 \leq i \leq n$ and $\boldsymbol{\pi} = \boldsymbol{\rho}$.

- Connectivity parameters $\boldsymbol{\theta} = (\theta_{ql})_{(q,l)\in\mathcal{Q}\times\mathcal{L}}$,
- Conditional on $\{Z_i, W_j\}$, random variables $\{X_{ij}\}$ are independent, with distribution

$$X_{ij}|Z_i = q, W_j = l \sim f(\cdot; \theta_{ql}).$$

# Outline

# Parameter estimation

## Existing identifiability results

- Undirected SBM, binary or weighted [Allman *et al.* 09, Allman *et al.* 11],
- Directed and binary SBM [Celisse *et al.* 12],
- Overlapping SBM [Latouche *et al.* 11],
- Binary LBM [Keribin *et al.* 13]

## Parameter estimation issue

- `em` algorithm not feasible because latent variables are not independent conditional on observed ones.
  Ex (SBM) : $\mathbb{P}(\{Z_i\}_i|\{X_{ij}\}_{i,j}) \neq \prod_i \mathbb{P}(Z_i|\{X_{ij}\}_{i,j})$
- Alternatives: Gibbs sampling or Variational approximation to `em`.

# Parameter estimation for SBM I

## Methods

- Gibbs sampling [Nowicki & Snijders 01], small graphs only (max $n = 200$ nodes).
- Variational `em`, binary case [Daudin *et al.* 08, Picard *et al.* 09];
- Variational Bayes `em`, binary [Latouche *et al.* 12];
- Variational `em` for valued graphs [Mariadassou *et al.* 10], with covariates [Zanghi *et al.* 10b] and online versions [Zanghi *et al.* 08, Zanghi *et al.* 10a];
- Moments methods and composite likelihood for affiliation valued graphs [Ambroise & Matias 10];
- ...

# Variational approximation (SBM case) I

## (Observed) Likelihood decomposition

Let $q(\mathbf{Z})$ be any distribution over $\mathcal{Q}^n$. We have

$$\log p(\mathbf{X}|\theta) = \mathcal{L}_{ML}(q;\theta) + KL(q(\cdot)\|p(\cdot|\mathbf{X},\theta)),$$

with $KL$ is the Kullback-Leibler divergence between $q(\mathbf{Z})$ and the true posterior $p(\mathbf{Z}|\mathbf{X},\theta)$, namely

$$KL(q(\cdot)\|p(\cdot|\mathbf{X},\theta)) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})}$$

$$\text{and} \quad \mathcal{L}_{ML}(q;\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})}.$$

So that

$$\log p(\mathbf{X}|\theta) \geq \mathcal{L}_{ML}(q;\theta),$$

with equality iff $q(\mathbf{Z})$ is the true posterior $p(\mathbf{Z}|\mathbf{X},\theta)$.

# Variational approximation (SBM case) II

## Variational principle

- Maximise the lower bound $\mathcal{L}_{ML}(q;\theta)$ (with respect to $q, \theta$) instead of the unavailable likelihood $\log p(\mathbf{X}|\theta)$.
- Maximising $\mathcal{L}_{ML}(q;\theta)$ with respect to $q$ is equivalent to minimising the Kullback-Leibler divergence $KL(q(\cdot)\|p(\cdot|\mathbf{X},\theta))$.
- Restricting $q$ to product distributions over $\mathcal{Q}^n$, you look for the best product law that approximates the true posterior $p(\mathbf{Z}|\mathbf{X},\theta)$.
- Note that $\mathcal{L}_{ML}(q;\theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\theta) + c(q)$,
  For fixed value $q$, the quantity $\mathcal{L}_{ML}(q;\cdot)$ represents an expectation of the complete log-likelihood $\log p(\mathbf{X},\mathbf{Z}|\theta)$ under distribution $q$ (instead of the true posterior $p(\mathbf{Z}|\mathbf{X},\theta)$ in `em` algorithm).

# Parameter estimation (follow.)

## Parameters estimation for LBM

- Variational methods for binary, Gaussian or Poisson data arrays [Govaert & Nadif 03, Govaert & Nadif 08, Govaert & Nadif 10].
- `sem` Gibbs approach (for categorical data) [Keribin *et al.* 13].

## Model selection

- Maximal likelihood is not available (thus neither AIC or BIC),
- ICL criterion is used [Daudin *et al.* 08, Keribin *et al.* 13].
- MCMC approach to select number of LBM groups [Wyse & Friel 12].

## Node clustering

Automatically performed by the previous algorithms.

# Outline

# Convergence issues

## Why does the variational approximation work?

- The variational approximation appears to be efficient, both for LBM and SBM.
- Variational approximation does not converge unless the true posterior $p(\mathbf{Z}|\mathbf{X};\theta)$ is degenerate [Gunawardana & Byrne 05].

## Remaining issues

- What is the (asymptotic) behaviour of the groups posterior distribution ? Is it degenerate?
- Is variational approximation somehow equivalent to `em` approach ?
- Does maximum likelihood converge in this setting anyway?

# Maximum likelihood and variational approach

## Results from [Celisse *et al.* 12] in SBM case

- Variational `em` is asymptotically equivalent to classical `em` for SBM.
- Maximum likelihood is convergent in this setup.

# Convergence of the groups posterior distribution (LBM or SBM) [Mariadassou & Matias 12]

- In general, the groups posterior distribution converges to a Dirac mass (when $n, m \to \infty$).
- However, when there exist equivalent configurations (=nodes groups inducing the same likelihood), the posterior converges to a mixture of Dirac located at these configurations.
- In some cases -in particular affiliation-, the number of equivalent configurations is larger than the number of label switching configurations.
- When there are equivalent configurations, the posterior converges to a Dirac mass at the configuration with largest prior.

# Conclusions on SBM and LBM

- Simple and efficient models for graphs (square adjacency matrices) or array data.
- Model based clustering of the nodes of the graph (or the rows/columns of the array), that encompasses community detection approaches.
- Many variants, with overlapping groups or covariates. Data may be binary or weighted, sparse or not, directed or not ...
- Convergence results are difficult to obtain but some exist.
- Variational `em` approximations provide good practical results but tend to depend on initialisation: there is room for improvement !

# References I



[Airoldi *et al.* 08] E.M. Airoldi, D.M. Blei, S.E. Fienberg and E.P. Xing.
Mixed Membership Stochastic Blockmodels.
*J. Mach. Learn. Res.*, 9:1981-2014, 2008.


[Allman *et al.* 09] E.S. Allman, C. Matias and J.A. Rhodes.
Identifiability of parameters in latent structure models with many observed variables.
*Ann. Statist.*, 37(6A):3099-3132, 2009.


[Allman *et al.* 11] E.S. Allman, C. Matias and J.A. Rhodes.
Parameter identifiability in a class of random graph mixture models.
*J. Statist. Planning and Inference*, 141(5):1719-1736, 2011.

# References II



[Ambroise & Matias 10] C. Ambroise and C. Matias.
New consistent and asymptotically normal estimators for random graph mixture models.
*Journal of the Royal Statistical Society: Series B*, 74(1):3-35, 2012.


[Celisse *et al.* 12] A. Celisse, J.-J. Daudin, and L. Pierre.
Consistency of maximum-likelihood and variational estimators in the stochastic block model.
Electron. J. Statist., 6:1847-1899, 2012.


[Daudin *et al.* 08] J.-J. Daudin, F. Picard, and S. Robin.
A mixture model for random graphs.
*Stat. Comput.*, 18(2):173–183, 2008.


[Govaert & Nadif 03] G. Govaert and M. Nadif.
Clustering with block mixture models.
*Pattern Recognition*, 36(2):463–473, 2003.

# References III



[Govaert & Nadif 08] G. Govaert and M. Nadif.
Block clustering with Bernoulli mixture models: Comparison of different approaches.
*Computational Statistics and Data Analysis*, 52(6):3233–3245, 2008.


[Govaert & Nadif 10] G. Govaert and M. Nadif.
Latent block model for contingency table.
*Communications in Statistics - Theory and Methods*, 39(3): 416–425, 2010.


[Gunawardana & Byrne 05] Gunawardana and Byrne.
Convergence Theorems for Generalized Alternating Minimization Procedures.
*JMLR*, 6:2049–2073, 2005.

# References IV



[Keribin *et al.* 13] C. Keribin, V. Brault, G. Celeux and G. Govaert.
Estimation and selection for the latent block model on categorical data.
*INRIA Research report 8264*, 2013.


[Latouche *et al.* 11] P. Latouche, E. Birmelé and C. Ambroise.
Overlapping Stochastic Block Models With Application to the French Political Blogosphere.
*Annals of Applied Statistics*, 5(1):309-336, 2011.


[Latouche *et al.* 12] P. Latouche, E. Birmelé and C. Ambroise.
Variational Bayesian Inference and Complexity Control for Stochastic Block Models.
*Statistical Modelling*, 12(1):93-115, 2012.

# References V



[Mariadassou & Matias 12] M. Mariadassou and C. Matias.
Convergence of the groups posterior distribution in latent or stochastic block models.
*hal-00713120*, 2012.


[Mariadassou *et al.* 10] M. Mariadassou, S. Robin, C. Vacher.
Uncovering Latent Structure in Valued Graphs: A Variational Approach.
*Annals of Applied Statistics*, 4:2,715–742, 2010.


[Nowicki & Snijders 01] K. Nowicki and T.A.B. Snijders.
Estimation and prediction for stochastic block structures.
*JASA* 96(455):1077-1087, 2001.

# References VI



[Picard *et al.* 09] F. Picard, V. Miele, J-J. Daudin, L. Cottret and S. Robin.
Deciphering the connectivity structure of biological networks using MixNet.
*BMC Bioinformatics*, 10:1-11, 2009.


[Wyse & Friel 12] J. Wyse and N. Friel
Block clustering with collapsed latent block models.
*Stat Comput* 22:415–428, 2012.


[Zanghi *et al.* 08] H. Zanghi, C. Ambroise and V. Miele.
Fast Online Graph Clustering via Erdős Rényi Mixture.
*Pattern Recognition*, 41(12):3592-3599, 2008.


[Zanghi *et al.* 10a] H. Zanghi, F. Picard, V. Miele, and C. Ambroise.
Strategies for online inference of network mixture.
*Annals of applied statistics*, 4(2):687-714, 2010.

# References VII


[Zanghi *et al.* 10b] H. Zanghi, S. Volant and C. Ambroise.
Clustering based on random graph model embedding vertex features.
*Pattern Recognition Letters* 31(9):830-836, 2010.