Stochastic block models for random graphs

Catherine MATIAS

CNRS, Laboratoire Statistique & Génome, Évry
(Soon: Laboratoire de Mathématiques et Modélisation d'Évry (LaMME))

http://stat.genopole.cnrs.fr/∼cmatias

Outline

Brief introduction to stochastic block models

Parameter estimation and node clustering

- Identifiability
- Parameter estimation
- 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
- . . .

We would like to cluster the nodes into groups.

Mixture model 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.

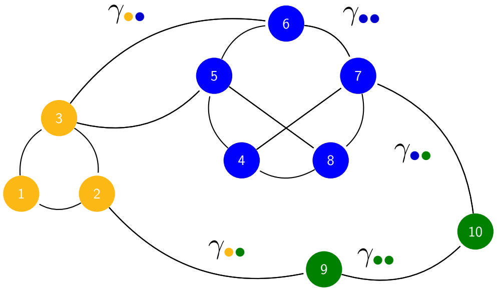
Advantages

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

Motivation/Justification: Szemerédi regularity Lemma [Szemerédi 78]

Every large enough graph can be divided into subsets of about the same size so that the edges between different subsets behave almost randomly.

Stochastic block model (binary graphs)

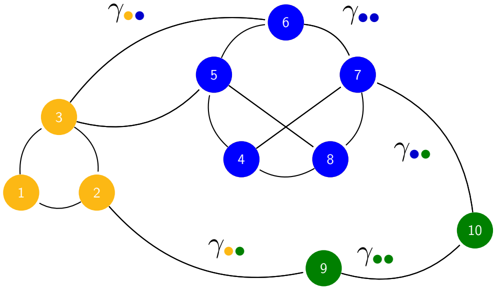


$$n = 10, Z_{5\bullet} = 1$$
$$Y_{12} = 1, Y_{15} = 0$$

Binary case (parametric model with $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$)

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

Stochastic block model (weighted graphs)



$n = 10, Z_{5\bullet} = 1$

$Y_{12} \in \mathbb{R}, Y_{15} = 0$

Weighted case (parametric model with $\theta = (\boldsymbol{\pi}, \boldsymbol{p}, \boldsymbol{\gamma})$)

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

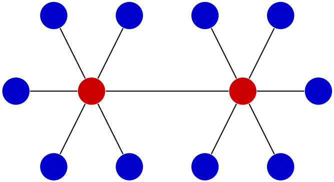
$$\mu_{Z_i Z_j}(\cdot) = p_{Z_i Z_j} f(\cdot, \gamma_{Z_i Z_j}) + (1 - p_{Z_i Z_j})\delta_0(\cdot)$$

(Assumption: f has continuous cdf at zero).

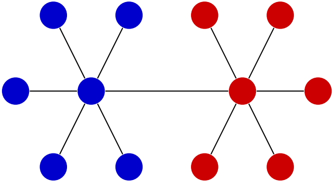
SBM clustering vs other clusterings

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 γ has only 2 parameters

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

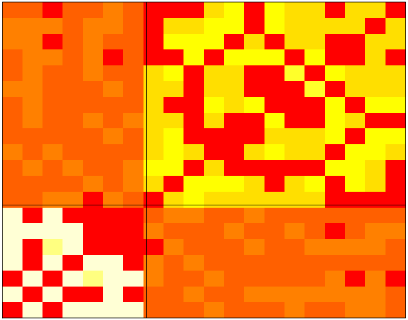
- Affiliation + $\alpha \gg \beta \implies$ community detection (cliques clustering).

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. Models bi-partite graphs.

Latent block models I

LBM notation

- Observations: array $\mathbf{Y}_{n,m} := \{Y_{ij}\}_{1 \le i \le n, 1 \le j \le m}$ with $Y_{ij} \in \mathcal{Y}$,
- $K \ge 1$ and $L \ge 1$ number of row and column groups, respectively.
- Groups prior distributions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ over $\mathcal{K} = \{1, \ldots, K\}$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_L)$ over $\mathcal{L} = \{1, \ldots, L\}$, such that $\sum_k \pi_k = \sum_l \rho_l = 1$.
- Latent variables $\mathbf{Z}_n := Z_1, \ldots, Z_n$ iid $\sim \boldsymbol{\pi}$ over $\mathcal{K}$ 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 : $n = m, \mathcal{K} = \mathcal{L}$, $Z_i = W_i$ for all $1 \leq i \leq n$ and $\boldsymbol{\pi} = \boldsymbol{\rho}$.

- Connectivity parameters $\boldsymbol{\gamma} = (\gamma_{kl})_{(k,l)\in\mathcal{K}\times\mathcal{L}}$,
- Conditional on $\{Z_i, W_j\}$, random variables $\{Y_{ij}\}$ are independent, with distribution

$$Y_{ij}|Z_i = k, W_j = l \sim f(\cdot;\gamma_{kl}).$$

Outline

Brief introduction to stochastic block models

Parameter estimation and node clustering

- Identifiability
- Parameter estimation
- Clustering convergence results

Outline

Parameter's identifiability

Problem

- Obviously, the model may only be identifiable up to a permutation on the group's labels.
- But whether one may uniquely recover the parameter up to label switching is a delicate task.

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].

Stating the identifiability problem

Identifiability if $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Rightarrow \theta_1 = \theta_2$ (injectivity of the map $\theta \mapsto \mathbb{P}_\theta$).

On a simple example: binary SBM with two groups

Parameters: $\boldsymbol{\pi} = (\pi_1, 1 - \pi_1)$ and $\boldsymbol{\gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{12} & \gamma_{22} \end{pmatrix}$.

Write down equations such as:

$$\mathbb{E}_\theta(Y_{ij}) = \sum_{1 \le k, l \le K} \pi_k \pi_l \gamma_{kl}, \quad \mathbb{E}_\theta(Y_{ij} Y_{il}) = \sum_{1 \le k \le l} \pi_k \Big(\sum_l \pi_l \gamma_{kl}\Big)^2 \dots$$

We end up with a set of polynomial equations and want to know if the solution is unique.

$\longrightarrow$ We rely on a general method for identifying parameters of latent structure models with many observed variables [Allman *et al.* 09].

A general method I

Toy model $\mathcal{M}(r;\kappa_1,\kappa_2,\kappa_3)$

$$\begin{tikzcd} & Z \arrow[dl] \arrow[d] \arrow[dr] & \\ X_1 & X_2 & X_3 \end{tikzcd}$$

- Z is a latent variable with r states, $Z \sim \boldsymbol{\pi} = (\pi_1, \ldots, \pi_r)$;
- For $j = 1, 2, 3$, X_j is observed with κ_j states ;
- $\{X_1, X_2, X_3\}$ independent conditional on Z and $X_j | Z = i \sim \boldsymbol{p}_{i,j} = (p_{i,j}(1), \ldots, p_{i,j}(\kappa_j))$.
- The distribution of (X_1, X_2, X_3) is the multivariate mixture $\mathbb{P}_\theta(X_1 = u, X_2 = v, X_3 = w) = \sum_{i=1}^r \pi_i p_{i,1}(u) p_{i,2}(v) p_{i,3}(w)$.
- Goal: Recover the parameters $\pi_i, p_{i,j}(u)$ from the mixture $\mathbb{P}_\theta$ (up to label swapping).

A general method II

Kruskal's result

For stochastic matrices M_j of size $r \times \kappa_j$ and a vector $\boldsymbol{\pi}$ of size r, define the three-way table $[\boldsymbol{\pi}, M_1, M_2, M_3]$ of size $\kappa_1 \times \kappa_2 \times \kappa_3$ by

$$[\boldsymbol{\pi}, M_1, M_2, M_3]_{u,v,w} = \sum_{i=1}^{r} \pi_i \, M_1(i,u)\, M_2(i,v)\, M_3(i,w).$$

The Kruskal rank, $\mathrm{rank}_K\, M$, of a matrix M, is the largest number I such that every set of I rows of M are independent.

Theorem [Kruskal 76]

Let $I_j = rank_K M_j$. If $I_1 + I_2 + I_3 \geq 2r + 2$, then $[\boldsymbol{\pi}, M_1, M_2, M_3]$ uniquely determines the M_j and $\boldsymbol{\pi}$, up to simultaneous permutation of the rows.

- Why 3 variates ? Because otherwise: matrix product.

A general method III

Corollary

The parameters of the model $\mathcal{M}(r;\kappa_1,\kappa_2,\kappa_3)$ are generically identifiable, up to label swapping.

Applications of this result [Allman *et al.* 09]

Parameters' identifiability in many different models

- Binary SBM with 2 classes;
- Multivariate Bernoulli mixtures;
- Multivariate non parametric mixtures;
- (Finite state space) HMMs;
- HMMs with non parametric emission distribution [Gassiat *et al.* 13].

Outline

Parameter estimation I

Parameter estimation issue

- `em` algorithm not feasible because latent variables are not independent conditional on observed ones.
  Ex (SBM) : $\mathbb{P}(\{Z_i\}_i | \{Y_{ij}\}_{i,j}) \neq \prod_i \mathbb{P}(Z_i | \{Y_{ij}\}_{i,j})$
- Alternatives:
  - Gibbs sampling or Variational approximation to `em`.
  - Composite likelihood approaches for affiliation valued graphs [Ambroise & Matias 10];

About LBM case

- Variational methods for binary, Gaussian or Poisson data arrays [Govaert & Nadif 03, Govaert & Nadif 08, Govaert & Nadif 10].
- Bayesian framework and Gibbs sampling for binary and Gaussian data [Wyse & Friel 12]
- `sem` Gibbs approach (for categorical data) [Keribin *et al.* 13].

Parameter estimation II

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.

Binary or Weighted Affiliation SBM [Ambroise & Matias 10] I

Models (Binary/weighted)

- $\{Z_i\}_{1 \le i \le n}$ i.i.d. latent vectors $Z_i = (Z_{i1}, \dots, Z_{iK}) \sim \mathcal{M}(1, \boldsymbol{\pi})$;
- Conditional on $\{Z_i\}$'s, the Y_{ij} are independent;
- Binary case:

$$Y_{ij} \sim \begin{cases} \mathcal{B}(\gamma_{\text{in}}) & \text{if } Z_i = Z_j \\ \mathcal{B}(\gamma_{\text{out}}) & \text{if } Z_i \neq Z_j. \end{cases}$$

- Weighted case:

$$Y_{ij} \sim \begin{cases} p_{\text{in}} f(\cdot, \gamma_{\text{in}}) + (1 - p_{\text{in}})\delta_0(\cdot) & \text{if } Z_i = Z_j \\ p_{\text{out}} f(\cdot, \gamma_{\text{out}}) + (1 - p_{\text{out}})\delta_0(\cdot) & \text{if } Z_i \neq Z_j. \end{cases}$$

Binary or Weighted Affiliation SBM [Ambroise & Matias 10] II

Composite likelihood idea - Weighted case

- The present edges $Y_{ij} \neq 0$ follow a mixture distribution

$$Y_{ij}|Y_{ij} \neq 0 \sim \{\textstyle\sum_{q=1}^{Q} \pi_q^2 p_{\text{in}}\} f(Y_{ij};\gamma_{\text{in}}) + \{\textstyle\sum_{q\neq\ell} \pi_q \pi_\ell p_{\text{out}}\} f(Y_{ij};\gamma_{\text{out}})$$

- Parameters of a mixture of two continuous distributions are in general identifiable.
- We form a composite log-likelihood

$$\mathcal{L}_X^{\mathsf{c}}(\boldsymbol{\theta}) = \tfrac{1}{n(n-1)} \sum_{i<j} \log[\alpha_{\text{in}} f(Y_{ij};\gamma_{\text{in}}) + \alpha_{\text{out}} f(Y_{ij};\gamma_{\text{out}})].$$

- If this converges to $\mathbb{E}[\log(\alpha_{\text{in}} f(Y_{ij};\gamma_{\text{in}}) + \alpha_{\text{out}} f(Y_{ij};\gamma_{\text{out}})]$ then we can estimate the parameters with

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}_X^{\mathsf{c}}(\boldsymbol{\theta}).$$

Binary or Weighted Affiliation SBM [Ambroise & Matias 10] III

Moment methods idea - Binary case

- Same idea does not apply directly in the Binary case, because $Y_{ij} \sim$ mixture of Bernoulli. Not identifiable !
- However, mixtures of 3-variate Bernoulli distributions are identifiable (in many cases).
- Develop same methodology with $\mathcal{L}^{\mathsf{c}}_X(\boldsymbol{\pi}, \alpha, \beta) = \frac{1}{n(n-1)(n-2)} \sum_{(i,j,k)\in\mathcal{I}_3} \log \mathbb{P}(Y_{ij}, Y_{ik}, Y_{jk})$.
- For the two approaches to be valid, we need to know whether the composite log-likelihoods converge.

Binary or Weighted Affiliation SBM [Ambroise & Matias 10] IV

Notation

- $\underline{\mathfrak{i}} = (i_1, \ldots, i_k)$ a k-tuple of nodes,
- $\mathbb{Y}^{\underline{\mathfrak{i}}} = (Y_{i_1 i_2}, \ldots, Y_{i_1 i_k}, Y_{i_2 i_3}, \ldots, Y_{i_{k-1} i_k})$ the vector of $p = \binom{k}{2}$ r.v. induced by the nodes $\underline{\mathfrak{i}}$,
- $g : \mathcal{Y}^p \to \mathbb{R}^s$, a function and
$$\widehat{m}_g = \frac{(n-k)!}{n!} \sum_{\underline{\mathfrak{i}} \in \mathcal{I}^k} g(\mathbb{Y}^{\underline{\mathfrak{i}}}) \quad \text{and} \quad m_g = \mathbb{E}(g(\mathbb{Y}^{(1,\ldots,k)})).$$

Theorem

For any $k, s \geq 1$ and $p = \binom{k}{2}$ and any measurable function $g : \mathcal{Y}^p \to \mathbb{R}^s$ such that $\mathbb{E}(\|g(\mathbb{Y}^{(1,\ldots,k)})\|^2) < +\infty$, the estimator $\hat{m}_g$ is consistent

$$\hat{m}_g \underset{n\to\infty}{\to} m_g \text{ almost surely,}$$

as well as asymptotically normal $\sqrt{n}(\hat{m}_g - m_g) \rightsquigarrow_{n\to\infty} \mathcal{N}(0, \Sigma_g)$.

Outline

Brief introduction to stochastic block models

Parameter estimation and node clustering

- Identifiability
- Parameter estimation
- Clustering convergence results

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 $p(\mathbf{Z}|\mathbf{Y};\gamma)$ 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)

Results from [Mariadassou & Matias 13]

- 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.

Equivalent configurations in SBM or LBM

- Label switching corresponds to $\mathbb{P}_{(\sigma(\boldsymbol{\pi}),\sigma(\boldsymbol{\gamma}))} = \mathbb{P}_{(\boldsymbol{\pi},\boldsymbol{\gamma})}$ for any permutation σ of $\{1,\dots,K\}$;
- In classical mixtures, identifiability requires that $\gamma_q \neq \gamma'_q$ for any $q \neq q'$;
- In SBM or LBM, one may have $\gamma_{ql} = \gamma_{q'l}$ for some $q \neq q'$;
- Then, if the matrix $\boldsymbol{\gamma}$ has symmetries, we may have $\sigma(\boldsymbol{\gamma}) = \boldsymbol{\gamma}$ with the model still identifiable if $\boldsymbol{\pi}$ has non equal entries. Namely $\mathbb{P}_{(\boldsymbol{\pi},\sigma(\boldsymbol{\gamma}))} = \mathbb{P}_{(\boldsymbol{\pi},\boldsymbol{\gamma})}$;
- As a consequence, the ratios between the posterior distributions at $(\mathbf{Z}_n, \mathbf{W}_m)$ and $\sigma^{-1}(\mathbf{Z}_n, \mathbf{W}_m)$ does not depend on data

$$\begin{aligned}
\mathbb{P}_{(\boldsymbol{\pi},\boldsymbol{\gamma})}(\mathbf{Z}_n, \mathbf{W}_m | \mathbf{Y}_{n,m}) &\propto \boldsymbol{\pi}(\mathbf{Z}_n, \mathbf{W}_m)\mathbb{P}_{\boldsymbol{\gamma}}(\mathbf{Y}_{n,m} | \mathbf{Z}_n, \mathbf{W}_m) \\
&\propto \boldsymbol{\pi}(\mathbf{Z}_n, \mathbf{W}_m)\mathbb{P}_{\sigma(\boldsymbol{\gamma})}(\mathbf{Y}_{n,m} | \mathbf{Z}_n, \mathbf{W}_m) \\
&\propto \frac{\boldsymbol{\pi}(\mathbf{Z}_n, \mathbf{W}_m)}{\boldsymbol{\pi}(\sigma^{-1}(\mathbf{Z}_n, \mathbf{W}_m))}\mathbb{P}_{(\boldsymbol{\pi},\boldsymbol{\gamma})}(\sigma^{-1}(\mathbf{Z}_n, \mathbf{W}_m) | \mathbf{Y}_{n,m}).
\end{aligned}$$

Conclusions

Modeling data

- SBM are natural and powerful models for handling networks data.
- Many variants, with overlapping groups or covariates. Data may be binary or weighted, sparse or not, directed or not . . . ;
- Natural generalisation of SBM for matrix data: LBM are handled in the same way.
- Model based clustering of the nodes of the graph (or the rows/columns of the array), that encompasses community detection approaches.

Theoretical results

- 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.

References III

[Gassiat *et al.* 13] E. Gassiat, A. Cleynen, S. Robin.
Finite state space non parametric Hidden Markov Models are in general identifiable.
arXiv:1306.4657, 2013.

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

[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.

References IV

[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.

[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.

References V

[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.

[Mariadassou & Matias 13] M. Mariadassou and C. Matias.
Convergence of the groups posterior distribution in latent or stochastic block models.
To appear in Bernoulli. hal-00713120, 2013.

[Szemerédi 78] Szemerédi, Endre.
Regular partitions of graphs.
Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), Colloq. Internat. CNRS, 260: 399-401, 1978.

References VI

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

[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.