Bayesian Methods for Graph Clustering

by

Pierre Latouche, Etienne Birmelé, and Christophe Ambroise

Research Report No. 17
September 2008

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
http://genome.jouy.inra.fr/ssb/

Bayesian Methods for Graph Clustering

Pierre Latouche PIERRE.LATOUCHE@GENOPOLE.CNRS.FR
Laboratoire Statistique et Genome
UMR CNRS 8071-INRA 1152-UEVE F-91000
Evry
France

Etienne Birmelé ETIENNE.BIRMELE@GENOPOLE.CNRS.FR
Laboratoire Statistique et Genome
UMR CNRS 8071-INRA 1152-UEVE F-91000
Evry
France

Christophe Ambroise CHRISTOPHE.AMBROISE@GENOPOLE.CNRS.FR
Laboratoire Statistique et Genome
UMR CNRS 8071-INRA 1152-UEVE F-91000
Evry
France

Editor:

Abstract

Networks are used in many scientific fields such as biology, social science, and information technology. They aim at modelling, with edges, the way objects of interest, represented by vertices, are related to each other. Looking for clusters of vertices, also called communities or modules, has appeared to be a powerful approach for capturing the underlying structure of a network. In this context, the Block-Clustering model has been applied on random graphs. The principle of this method is to assume that given the latent structure of a graph, the edges are independent and generated from a parametric distribution. Many EM-like strategies have been proposed, in a frequentist setting, to optimize the parameters of the model. Moreover, a criterion, based on an asymptotic approximation of the Integrated Classification Likelihood (ICL), has recently been derived to estimate the number of classes in the latent structure. In this paper, we show how the Block-Clustering model can be described in a full Bayesian framework and how the posterior distribution, of all the parameters and latent variables, can be approximated efficiently applying Variational Bayes (VB). We also propose a new non-asymptotic Bayesian model selection criterion. Using simulated and real data sets, we compare our approach to other strategies. We show that our Bayesian method is able to handle large networks and that our criterion is more robust than ICL.

1. Introduction

For the last few years, networks have been increasingly studied. Indeed, many scientific fields such as biology (Albert and Barabási (2002)), social science, and information technology, see those mathematical strutures as powerful tools to model the interactions between objects of

interest. Examples of data sets having such structures are friendship (Palla et al (2007)) and protein-protein interaction networks (Barabási and Oltvai (2004)), powergrids (Watts and Strogatz (1998)), and the Internet (Zanghi et al (2007)). In this context, a lot of attention has been paid on developing models to learn knowledge from the network topology. Many methods have been proposed, and in this work, we focus on statistical models that describe the way edges connect vertices.

A well known strategy consists in seeing a given network as a realization of a random graph model based on a mixture distribution (Frank and Harary (1982), Snijders and Nowicki (1997), Newman and Leicht (2007), Daudin et al (2008)). The method assumes that, according to its connection profile, each vertex belongs to a hidden class of a latent structure and that, given this latent structure, all the observed edges are independent and binary distributed. Many names have been proposed for this model, and in the following, it will be denoted MixNet, which is equivalent to the Block-Clustering model of Snijders and Nowicki (1997).

A key question is the estimation of the MixNet parameters. So far, the optimization procedures that have been proposed are based on heuristics (Newman and Leicht (2007)) or have been described in a frequentist setting (Daudin et al (2008)). Bayesian strategies have also been developed but are limited in a sense that they can not handle large networks. All those methods face the same difficulty. Indeed, the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi})$, of all the latent variables $\mathbf{Z}$ given the observed edges $\mathbf{X}$, can not be factorized. To tackle such problem, Daudin et al (2008) proposed a variational approximation of the posterior, which corresponds to a mean-field approximation. Online strategies have also been developed (Zanghi et al (2007)). They give biased estimates but are very efficient in terms of computational cost.

Another difficulty is the estimation of the number of classes in the mixture. Indeed, many criteria, such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) (Burnham and Anderson (2004)) are based on the likelihood $p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\pi})$ of the incomplete data set $\mathbf{X}$, which is intractable here. Therefore, Mariadassou and Robin (2007) derived a criterion based on an asymtotic approximation of the Integrated Classification Likelihood (also called Integrated Complete-data Likelihood). More details can be found in Biernacki et al (2000). They found that this criterion, that we will denote ICL for simplicity, was very accurate in most situations but tended to underestimate the number of classes when dealing with small graphs. We emphasize that ICL is currently the only model based criterion developed for MixNet.

In this paper, extending the work of Hofman and Wiggins (2008) who considered affiliation models, where only two types of edges exist (edges between nodes of the same class and edges between nodes of different classes), we propose a full Bayesian version of MixNet. Thus, after having presented MixNet and the frequentist approach of maximum likelihood estimation in Section 2, we introduce some prior distributions and describe the MixNet Bayesian probabilistic model in Section 3. We derive the model optimization equations using Variational Bayes and we propose a new criterion to estimate the number of classes. Finally, in Section 5, we carry out some experiments using simulated data sets and the metabolic network of *Escherichia coli* to compare both the number and the quality of the estimated clusters obtained with the ICL criterion and the variational frequentist strategy and our approach.

2. A Mixture Model for Networks

2.1 Model and Notations

We consider a binary random graph G, where V denotes a set of N fixed vertices and $\mathbf{X} = \{X_{ij}, (i,j) \in V^2\}$ is the set of all the random edges.

MixNet assumes that each vertex i belongs to an unknown class q among Q classes and the latent variable $\mathbf{Z_i}$ reflects our uncertainty as to which one that is:

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

where we denote $\boldsymbol{\alpha}$, the vector of class proportions. The edge probabilities are then given by:

$$X_{ij} | \{Z_{iq} Z_{jl} = 1\} \sim \mathcal{B}(X_{ij} | \pi_{ql}).$$

According to MixNet probabilistic model (Fig. 1), the latent variables in the set $\mathbf{Z} = (\mathbf{Z_1}, \ldots, \mathbf{Z_N})$ are *iid* and given this latent structure, all the edges are supposed to be independent. Thus, when considering an undirected graph without self loops, we obtain:

$$p(\mathbf{Z}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathcal{M}(\mathbf{Z_i};\ 1, \boldsymbol{\alpha}) = \prod_{i=1}^{N} \prod_{q=1}^{Q} \alpha_q^{Z_{iq}},$$

and

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\pi}) = \prod_{i<j} p(X_{ij}|\mathbf{Z_i}, \mathbf{Z_j}, \boldsymbol{\pi}) = \prod_{i<j}^{N} \prod_{q,l}^{Q} \mathcal{B}(X_{ij}|\pi_{ql})^{Z_{iq} Z_{jl}}.$$

In the case of a directed graph, the products over $i < j$ must be replaced by products over $i \neq j$. The edges X_{ii} must also be taken into account if the graph contains self-loops.

2.2 Maximum likelihood estimation

The likelihood $p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\pi})$ of the incomplete data set $\mathbf{X}$ can be obtained through the marginalization $p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\pi}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\pi})$. This summation involves Q^N terms and quickly becomes intractable. To tackle such problem, the well known Expectation-Maximization (EM) algorithm (Dempster et al (1977)) has been applied with success on a large variety of mixture models. Unfortunately, the E-step requires the calculation of the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi})$ which can not be factorized in the case of networks and need to be approximated (Daudin et al (2008)).

A first strategy consists in considering directly the predictions of $\mathbf{Z}$ rather than the full distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi})$. This Classification EM (CEM) approach has been subject to previous work (Zanghi et al (2007)). It gives biased estimates but is very efficient in terms of computational cost. Variational techniques have also been used to approximate the posterior with another distribution $q(\mathbf{Z})$. Thus, when using the Kullback-Leibler divergence $\mathrm{KL}(.\,||\,.)$, the log-likelihood of the incomplete data set is decomposed into two terms:

$$\ln p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\pi}) = \mathcal{L}\Big(q(.);\ \boldsymbol{\alpha}, \boldsymbol{\pi}\Big) + \mathrm{KL}\Big(q(.)\,||\,p(.|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi})\Big), \tag{1}$$

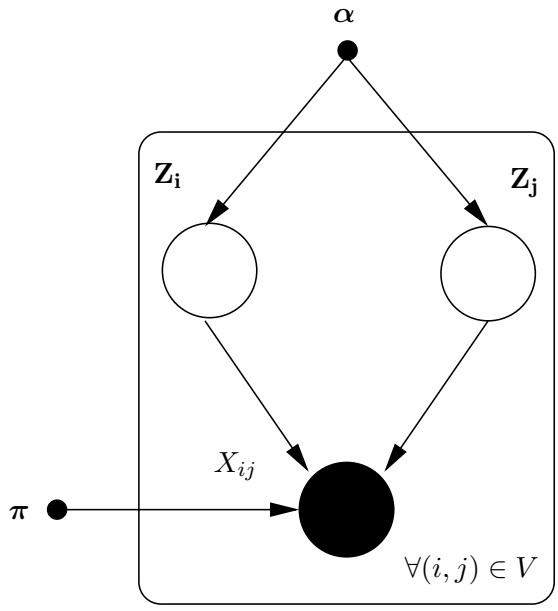

Figure 1 Graphical representation of the MixNet probabilistic model

where

$$\mathcal{L}\Big(q(.);\ \boldsymbol{\alpha}, \boldsymbol{\pi}\Big) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\{\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\pi})}{q(\mathbf{Z})}\}, \tag{2}$$

and

$$\mathrm{KL}\left(q(.) \,||\, p(.|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi})\right) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\{\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi})}{q(\mathbf{Z})}\}. \tag{3}$$

It can be easily verified that minimizing (3) is equivalent to maximizing the lower bound (2) of (1). To obtain a tractable algorithm, the variational methods assume that the distribution $q(\mathbf{Z})$ can be factorized such that:

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

This gives rise to a Variational EM procedure. Indeed, during the variational E-step, the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ are fixed and by optimizing (2), with respect to the distribution $q(\mathbf{Z})$, the algorithm looks for the optimal approximation of the posterior distribution. Conversely, during the Variational M-step, $q(\mathbf{Z})$ is fixed and the log-likelihood of the incomplete data set is optimized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$. This procedure is repeated until convergence.

3. Bayesian view of MixNet

3.1 Bayesian probabilistic model

We now show how MixNet can be described in a full Bayesian framework. To transform the MixNet frequentist probabilistic model, we first specify some prior distributions for the model parameters. To simplify the calculations, we use *conjugate* priors. Thus, since $p(\mathbf{Z_i}|\boldsymbol{\alpha})$ is a multinomial distribution, we choose a Dirichlet distribution for the mixing coefficients:

$$p\Big(\boldsymbol{\alpha}|\mathbf{n^0} = \{n_1^0, \ldots, n_Q^0\}\Big) = \mathrm{Dir}(\boldsymbol{\alpha};\, \mathbf{n^0}) = \frac{\Gamma(\sum_{q=1}^Q n_q^0)}{\Gamma(n_1^0)\ldots\Gamma(n_Q^0)} \prod_{q=1}^Q \alpha_q^{n_q^0-1},$$

where we denote n_q^0, the prior number of vertices in the q-th component of the mixture. In order to obtain a posterior distribution influenced primarily by the network data rather than the prior, small values have to be chosen. A typical choice is $n_q^0 = \frac{1}{2}, \forall q$. This leads to a non-informative Jeffreys prior distribution (Jeffreys (1946)). It is also possible to consider a uniform distribution on the $Q-1$ dimensional simplex by fixing $n_q^0 = 1, \forall q$.

Since $p(X_{ij}|\mathbf{Z_i}, \mathbf{Z_j}, \boldsymbol{\pi})$ is a Bernoulli distribution, we use Beta priors to model the connectivity matrix $\boldsymbol{\pi}$:

$$\begin{aligned} p\Big(\boldsymbol{\pi}|\boldsymbol{\eta^0} = (\eta_{ql}^0), \boldsymbol{\zeta^0} = (\zeta_{ql}^0)\Big) &= \prod_{q\leq l}^Q \mathrm{Beta}(\pi_{ql};\, \eta_{ql}^0, \zeta_{ql}^0) \\ &= \prod_{q\leq l}^Q \frac{\Gamma(\eta_{ql}^0 + \zeta_{ql}^0)}{\Gamma(\eta_{ql}^0)\Gamma(\zeta_{ql}^0)} \pi_{ql}^{\eta_{ql}^0-1}(1-\pi_{ql})^{\zeta_{ql}^0-1}, \end{aligned} \tag{4}$$

where η_{ql}^0 and ζ_{ql}^0 represent respectively the prior number of edges and non-edges connecting vertices of cluster q to vertices of cluster l. A common choice consists in setting $\eta_{ql}^0 = \zeta_{ql}^0 = 1, \forall q$. This gives rise to a uniform prior distribution. Note that if the graph G is directed, the products over $q \leq l$, must be replaced by products over q, l since $\boldsymbol{\pi}$ is no longer symmetric.

Thus, the model parameters are now seen as random variables, represented by circles in the Directed Acyclic Graph (DAG) Fig. 2. They depend on parameters $\mathbf{n^0}$, $\boldsymbol{\eta^0}$, and $\boldsymbol{\zeta^0}$ which are called *hyperparameters* in the Bayesian literature (MacKay (1992)). The joint distribution of the Bayesian probabilistic model is then given by:

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\mathbf{n^0}, \boldsymbol{\eta^0}, \boldsymbol{\zeta^0}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\pi})p(\mathbf{Z}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\mathbf{n^0})p(\boldsymbol{\pi}|\boldsymbol{\eta^0}, \boldsymbol{\zeta^0}).$$

For the rest of the paper, since the prior hyperparameters are fixed and in order to keep the notations simple, they will not be shown explicitly in the conditional distributions.

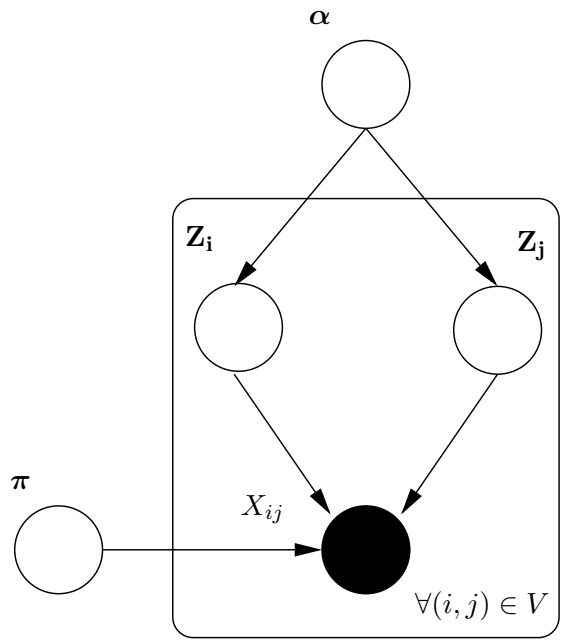

Figure 2 Directed acyclic graph representing the Bayesian view of the MixNet probabilistic model

3.2 Variational inference

The inference task consists in evaluating the posterior $p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\mathbf{X})$ of all the hidden variables (latent variables $\mathbf{Z}$ and parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$) given the observed edges $\mathbf{X}$. Unfortunately, under MixNet, this distribution is intractable. To overcome such difficulties, we follow the work of Attias (1999), Corduneanu and Bishop (2001), Svensén and Bishop (2004) on Bayesian mixture modelling and Bayesian model selection. Thus, we first introduce a factorized distribution:

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

and we use Variational Bayes to obtain an optimal approximation $q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})$ of the posterior. This framework is called the mean field theory in physics (Parisi (1988)). The Kullback-Leibler divergence enables us to decompose the log-marginal probability, usually called the model evidence or the log Integrated Observed-data Likelihood, and we obtain:

$$\ln p(\mathbf{X}) = \mathcal{L}\Big(q(.)\Big) + \mathrm{KL}\left(q(.) \,||\, p(.|\mathbf{X}\right), \tag{5}$$

where

$$\mathcal{L}\Big(q(.)\Big) = \sum_{\mathbf{Z}} \int\int q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \ln\{\frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})}\} d\boldsymbol{\alpha} d\boldsymbol{\pi}, \tag{6}$$

and

$$\mathrm{KL}\left(q(.)\,||\,p(.|\mathbf{X})\right) = -\sum_{\mathbf{Z}}\int\int q(\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi})\ln\{\frac{p(\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi}|\mathbf{X})}{q(\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi})}\}d\boldsymbol{\alpha}d\boldsymbol{\pi}. \tag{7}$$

Again, as for the variational EM approach (Section 2.2), minimizing (7) is equivalent to maximizing the lower bound (6) of (5). However, we now have a full variational optimization problem since the model parameters are random variables and we are looking for the best approximation $q(\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi})$ among all the factorized distributions. In the following, we call Variational Bayes E-step, the optimization of each distribution $q(\mathbf{Z_i})$ and Variational Bayes M-step, the approximations of the remaining factors. We derive the update equations in the case of an undirected graph G without self-loop.

3.2.1 Variational Bayes E-step

Proposition 1 *The optimal approximation at vertex i is:*

$$q(\mathbf{Z_i}) = \mathcal{M}(\mathbf{Z_i};\,1,\boldsymbol{\tau_i} = \{\tau_{i1},\ldots,\tau_{iQ}\}), \tag{8}$$

where τ_{iq} is the probability (responsability) of node i to belong to class q. It satisfies the fix point relation:

$$\tau_{iq} \propto e^{\psi(n_q)-\psi(\sum_{l=1}^{Q} n_l)}\prod_{j\neq i}^{N}\prod_{l=1}^{Q} e^{\tau_{jl}\left(\psi(\zeta_{ql})-\psi(\eta_{ql}+\zeta_{ql})+X_{ij}\left(\psi(\eta_{ql})-\psi(\zeta_{ql})\right)\right)}, \tag{9}$$

where $\psi(.)$ is the digamma *function.*

Proof According to Variational Bayes, the optimal distribution $q(\mathbf{Z_i})$ is given by:

$$\begin{aligned}
\ln q(\mathbf{Z_i}) &= \mathsf{E}_{\mathbf{Z}^{\setminus \mathbf{i}},\boldsymbol{\alpha},\boldsymbol{\pi}}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi})] + \mathrm{cst}\\
&= \mathsf{E}_{\mathbf{Z}^{\setminus \mathbf{i}},\boldsymbol{\pi}}[\ln p(\mathbf{X}|\mathbf{Z},\boldsymbol{\pi})] + \mathsf{E}_{\mathbf{Z}^{\setminus \mathbf{i}},\boldsymbol{\alpha}}[\ln p(\mathbf{Z}|\boldsymbol{\alpha})] + \mathrm{cst}\\
&= \mathsf{E}_{\mathbf{Z}^{\setminus \mathbf{i}},\boldsymbol{\pi}}[\sum_{i<j}\sum_{q,l} Z_{iq}Z_{jl}\Big(X_{ij}\ln\pi_{ql} + (1-X_{ij})\ln(1-\pi_{ql})\Big)]\\
&\qquad + \mathsf{E}_{\mathbf{Z}^{\setminus \mathbf{i}},\boldsymbol{\alpha}}[\sum_{i=1}^{N}\sum_{q=1}^{Q} Z_{iq}\ln\alpha_q] + \mathrm{cst}\\
&= \sum_{q=1}^{Q} Z_{iq}\Bigg(\mathsf{E}_{\alpha_q}[\ln\alpha_q] + \sum_{j\neq i}^{N}\sum_{l=1}^{Q}\tau_{jl}\Big(X_{ij}\big(\mathsf{E}_{\pi_{ql}}[\ln\pi_{ql}] - \mathsf{E}_{\pi_{ql}}[\ln(1-\pi_{ql})]\big)\\
&\qquad + \mathsf{E}_{\pi_{ql}}[\ln(1-\pi_{ql})]\Big)\Bigg) + \mathrm{cst}\\
&= \sum_{q=1}^{Q} Z_{iq}\Bigg(\psi(n_q) - \psi(\sum_{l=1}^{N} n_l) + \sum_{j\neq i}^{N}\sum_{l=1}^{Q}\tau_{jl}\Big(X_{ij}\big(\psi(\eta_{ql}) - \psi(\zeta_{ql})\big)\\
&\qquad + \psi(\zeta_{ql}) - \psi(\eta_{ql}+\zeta_{ql})\Big)\Bigg) + \mathrm{cst},
\end{aligned} \tag{10}$$

where $\mathbf{Z}^{\backslash \mathbf{i}}$ denotes the class of all nodes except node i. Moreover, to simplify the calculations, the terms that do not depend on Z_i have been absorbed into the constant. Taking the exponential of (10) and after normalization, we obtain the multinomial distribution (8). ■

The matrix $\boldsymbol{\tau}$ is obtained by iterating the relation (9) until convergence.

3.2.2 Variational Bayes M-step : optimization of $q(\boldsymbol{\alpha})$

Proposition 2 *The optimization of the lower bound with respect to $q(\boldsymbol{\alpha})$ produces a distribution with the same functional form as the prior $p(\boldsymbol{\alpha})$:*

$$q(\boldsymbol{\alpha}) = \mathrm{Dir}(\alpha;\, \mathbf{n}), \tag{11}$$

where n_q is the pseudo number of vertices in the q-th component of the mixture:

$$n_q = n_q^0 + \sum_{i=1}^{N} \tau_{iq}. \tag{12}$$

Proof According to Variational Bayes, the optimal distribution $q(\boldsymbol{\alpha})$ is given by:

$$\begin{aligned} \ln q(\boldsymbol{\alpha}) &= \mathsf{E}_{\mathbf{Z},\boldsymbol{\pi}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})] + \mathrm{cst} \\ &= \mathsf{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}|\boldsymbol{\alpha})] + \ln p(\boldsymbol{\alpha}) + \mathrm{cst} \\ &= \sum_{i=1}^{N}\sum_{q=1}^{Q} \tau_{iq} \ln \alpha_q + \sum_{q=1}^{Q} (n_q^0 - 1) \ln \alpha_q + \mathrm{cst} \\ &= \sum_{q=1}^{Q} \left(n_q^0 - 1 + \sum_{i=1}^{N} \tau_{iq} \right) \ln \alpha_q + \mathrm{cst}. \end{aligned} \tag{13}$$

Taking the exponential of (13) and after normalization, we obtain the Dirichlet distribution (11). ■

3.2.3 Variational Bayes M-step : optimization of $q(\boldsymbol{\pi})$

Proposition 3 *Again, the functional form of the prior $p(\boldsymbol{\pi})$ is conserved through the variational optimization:*

$$q(\boldsymbol{\pi}) = \prod_{q \le l}^{Q} \mathrm{Beta}(\pi_{ql} | \eta_{ql}, \zeta_{ql}), \tag{14}$$

where η_{ql} and ζ_{ql} represent respectively the pseudo number of edges and non-edges connecting vertices of cluster q to vertices of cluster l. For $q \neq l$, the hyperparameter η_{ql} is given by:

$$\eta_{ql} = \eta_{ql}^0 + \sum_{i \neq j}^{N} X_{ij} \tau_{iq} \tau_{jl}, \tag{15}$$

and $\forall q$*:*

$$\eta_{qq} = \eta_{qq}^0 + \sum_{i<j}^{N} X_{ij}\tau_{iq}\tau_{jq}, \tag{16}$$

Moreover, for $q \neq l$*, the hyperparameter* ζ_{ql} *is given by:*

$$\zeta_{ql} = \zeta_{ql}^0 + \sum_{i\neq j}^{N} (1 - X_{ij})\tau_{iq}\tau_{jl}, \tag{17}$$

and $\forall q$*:*

$$\zeta_{qq} = \zeta_{qq}^0 + \sum_{i<j}^{N} (1 - X_{ij})\tau_{iq}\tau_{jq}, \tag{18}$$

Proof According to Variational Bayes, the optimal distribution $q(\boldsymbol{\pi})$ is given by:

$$\begin{aligned}
\ln q(\boldsymbol{\pi}) &= \mathsf{E}_{\mathbf{Z},\boldsymbol{\alpha}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})] + \text{cst} \\
&= \mathsf{E}_{\mathbf{Z}}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\pi})] + \ln p(\boldsymbol{\pi}) + \text{cst} \\
&= \sum_{i<j}^{N}\sum_{q,l}^{Q} \tau_{iq}\tau_{jl}\Big(X_{ij}\ln\pi_{ql} + (1 - X_{ij})\ln(1 - \pi_{ql})\Big) \\
&\qquad + \sum_{q\leq l}^{Q}\Big((\eta_{ql}^0 - 1)\ln\pi_{ql} + (\zeta_{ql}^0 - 1)\ln(1 - \pi_{ql})\Big) + \text{cst} \\
&= \sum_{q<l}^{Q}\sum_{i\neq j}^{N} \tau_{iq}\tau_{jl}\Big(X_{ij}\ln\pi_{ql} + (1 - X_{ij})\ln(1 - \pi_{ql})\Big) \\
&\qquad + \sum_{q=1}^{Q}\sum_{i<j}^{N} \tau_{iq}\tau_{jq}\Big(X_{ij}\ln\pi_{qq} + (1 - X_{ij})\ln(1 - \pi_{qq})\Big) \\
&\qquad + \sum_{q\leq l}^{Q}\Big((\eta_{ql}^0 - 1)\ln\pi_{ql} + (\zeta_{ql}^0 - 1)\ln(1 - \pi_{ql})\Big) + \text{cst} \\
&= \sum_{q<l}^{Q}\left(\Big(\eta_{ql}^0 - 1 + \sum_{i\neq j}^{N}\tau_{iq}\tau_{jl}X_{ij}\Big)\ln\pi_{ql} + \Big(\zeta_{ql}^0 - 1 + \sum_{i\neq j}^{N}\tau_{iq}\tau_{jl}(1 - X_{ij})\Big)\ln(1 - \pi_{ql})\right) \\
&\qquad + \sum_{q=1}^{Q}\left(\Big(\eta_{qq}^0 - 1 + \sum_{i<j}^{N}\tau_{iq}\tau_{jq}X_{ij}\Big)\ln\pi_{qq} + \Big(\zeta_{qq}^0 - 1 + \sum_{i<j}^{N}\tau_{iq}\tau_{jq}(1 - X_{ij})\Big)\ln(1 - \pi_{qq})\right).
\end{aligned} \tag{19}$$

Taking the exponential of (19) and after normalization, we obtain the product of Beta distribution (14). ■

3.2.4 Lower bound

Proposition 4 *When computed just after the Variational Bayes M-step, most of the terms of the lower bound disappear. Only one term in $\boldsymbol{\tau}$ and the normalizing constants of the Dirichlet and Beta distributions remain:*

$$\mathcal{L}\Big(q(.)\Big) = \ln\{\frac{\Gamma(\sum_{q=1}^{Q} n_q^0)\prod_{q=1}^{Q}\Gamma(n_q)}{\Gamma(\sum_{q=1}^{Q} n_q)\prod_{q=1}^{Q}\Gamma(n_q^0)}\} + \sum_{q\leq l}^{Q}\ln\{\frac{\Gamma(\eta_{ql}^0+\zeta_{ql}^0)\Gamma(\eta_{ql})\Gamma(\zeta_{ql})}{\Gamma(\eta_{ql}+\zeta_{ql})\Gamma(\eta_{ql}^0)\Gamma(\zeta_{ql}^0)}\} - \sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}\ln\tau_{iq}. \tag{20}$$

The proof is given in the appendix.

3.3 Model selection

So far, we have seen that the Variational Bayes optimization of the lower bound leads to an approximation of the posterior of all the hidden variables, given the observed edges. However, we have not addressed yet the problem of estimating the number Q of classes in the mixture. Following Bishop (2006), we propose a Bayesian model selection criterion. First, note that the model evidence (5) depends on Q. Indeed, if the number of classes is changed, the all Variational Bayes optimization procedure presented in Section 3.2 is modified. To take this dependency into account, the model evidence can be written $p(\mathbf{X}|Q)$ and Bayes rule leads to the posterior:

$$p(Q|\mathbf{X}) \propto p(\mathbf{X}|Q)p(Q), \tag{21}$$

where Q is now seen as a random variable. If the prior $p(Q)$ is broad, maximizing (21) with respect to Q is equivalent to maximizing $p(\mathbf{X}|Q)$. However, since MixNet is a mixture model, for any given setting of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ there will be a total of $Q!$ parameters which lead to the same distribution over the edges. Moreover, as we saw previously, the model evidence is intractable and needs to be approximated. Thus, we propose to use the lower bound (20) and to add a term $\ln Q!$ to take the multimodality into account. In the case of networks, we emphasize that our work led to the first criterion based on a non-asymptotic approximation of the model evidence, also called Integrated Observed-data likelihood. When considering other types of mixture models, Biernacki et al (2000) showed that such criteria were very powerful to select the number of classes.

4. Iterative algorithm

Algorithm 1: Variational Bayes EM algorithm for undirected graphs without self-loop

`// INITIALIZATION`

Initialize $\boldsymbol{\tau}$ with a spectral clustering algorithm
$\mathbf{n^0} \leftarrow \mathbf{1_{1\times N}}$ (or $\frac{\mathbf{1_{1\times N}}}{2}$); $\eta^{\mathbf{0}} \leftarrow \mathbf{1_{Q\times Q}}$; $\zeta^{\mathbf{0}} \leftarrow \mathbf{1_{Q\times Q}}$;

`// OPTIMIZATION`

repeat
  `// M-step`
  for *q=1:Q* **do**
    $n_q \leftarrow n_q^0 + \sum_{i=1}^N \tau_{iq}$;
  end
  for *q=1:Q* **do**
    for *l=q:Q* **do**
      $\eta_{ql} \leftarrow \eta_{ql}^0 + \sum_{i\neq j}^N X_{ij}\tau_{iq}\tau_{jl}$;
      $\zeta_{ql} = \zeta_{ql}^0 + \sum_{i\neq j}^N (1 - X_{ij})\tau_{iq}\tau_{jl}$;
      if *q=l* **then**
        $\eta_{ql} \leftarrow \frac{\eta_{ql}}{2}$;
        $\zeta_{ql} \leftarrow \frac{\zeta_{ql}}{2}$;
      end
      else
        $\eta_{lq} \leftarrow \eta_{ql}$;
        $\zeta_{lq} \leftarrow \zeta_{ql}$;
      end
    end
  end
  `// E-step`
  repeat
    for *i=1:N* **do**
      for *q=1:Q* **do**
        $\tau_{iq} \leftarrow \prod_{j\neq i}^N \prod_{l=1}^Q e^{\tau_{jl}\left(\psi(\zeta_{ql})-\psi(\eta_{ql}+\zeta_{ql})+X_{ij}\left(\psi(\eta_{ql})-\psi(\zeta_{ql})\right)\right)}$;
        $\tau_{iq} \leftarrow e^{\psi(n_q)-\psi(\sum_{l=1}^Q n_l)}\tau_{iq}$;
      end
      Normalize $\boldsymbol{\tau_{i.}}$;
    end
  until $\boldsymbol{\tau}$ *converges*
until $\mathcal{L}\Big(q(.)\Big)$ *converges*

5. Experiments

We present some results of the experiments we carried out to assess our Bayesian version of MixNet and the model selection criterion we proposed in Section 3.3. Through all our experiments, we compared our approach to the work of Daudin et al (2008) who used ICL as a criterion to identify the number of classes in latent structures and the frequentist approach of variational EM, described in Section 2.2, to estimate the model parameters. We considered both synthetic data, generated according to known random graph models, and the metabolic network of bacteria *Escherichia coli*. In the first set of experiments, we used the synthetic graphs and we concentrated on analyzing the capacity of ICL and our criterion to retrieve the true number of classes in the latent structures. We recall that our criterion is the first criterion developed based on a non-asymptotic approximation of the model evidence, also called Integrated Observed-data likelihood. Finally, we applied our approach to the metabolic network and we analyzed the number of estimated classes and the learnt partitions.

5.1 Comparison of the criteria

In these experiments, we consider simple affiliation models where only two types of edges exist : edges between nodes of the same class and edges between nodes of different classes. Each type of edge has a given probability, respectively $\pi_{qq} = \lambda$ and $\pi_{ql} = \epsilon$. Following Mariadassou and Robin (2007) who showed that ICL tended to underestimate the number of classes in the case of small graphs, we generated graphs with only $n = 50$ vertices to analyze the robustness of our criterion. Moreover, to limit the number of free parameters, we studied the cases where $\lambda = 1 - \epsilon$ and $\lambda = \frac{1}{2} - \epsilon$ which correspond respectively to dense graphs (ADC[1] greater than 6) and sparser graphs (ADC smaller than 3.5). Thus, we considered seven affiliation models shown in Table 1. The differences between these models are related to their modular structure which varies from no structure to strong modular structure.

For each affiliation model, we analyzed graphs with $Q_{True} \in \{2, \ldots, 5\}$ classes mixed in the same proportions $\alpha_1 = \cdots = \alpha_{Q_{True}} = \frac{1}{Q_{True}}$. Thus, we studied a total of 28 graph models.

Table 1 Parameters of the seven affiliation models considered

Model	λ	ϵ
1	0.9	0.1
2	0.85	0.15
3	0.8	0.2
4	0.75	0.25
5	0.49	0.01
6	0.47	0.03
7	0.45	0.05

1. Average Degree of Connectivity

For each of these graph models, we simulated 100 networks. In order to estimate the number of classes in the latent structures, we applied our algorithm (Section 4) and the variational EM approach of Daudin et al (2008) on each network, for various numbers of classes $Q \in \{1, \ldots, 6\}$. Note that, we chose $n_q^0 = 1$, $\forall q \in \{1, \ldots, Q\}$ for the Dirichlet prior and $\eta_{ql}^0 = \zeta_{ql}^0 = 1, \forall (q, l) \in \{1, \ldots, Q\}^2$ for the Beta priors. We recall that such distributions correspond to uniform distributions. Like any optimization technique, the Bayesian and frequentist methods depend on the initialization. Thus, for each simulated network and each number of classes Q, we started the algorithms with five different initializations of $\boldsymbol{\tau}$ obtained using a spectral clustering method (Ng et al (2001)). Then, for the Bayesian algorithm, we used the criterion we proposed in Section 3.3 to select the best learnt model, whereas we used ICL in the frequentist approach. Finally, for each simulated network, we obtained two estimates Q_{ICL} and Q_{VB} of the number Q_{True} of latent classes by selecting $Q \in \{1, \ldots, 6\}$ for which the corresponding criteria were maximized.

In Table 2, we observe that for the most structured affiliation model, the two criteria always estimate correctly the true number of classes except when $Q_{True} = 5$. In this case, the Bayesian criterion performs better. Indeed, it has a percentage of accuracy of 95% against 87% for ICL.

Table 2 Confusion matrices for ICL and Bayesian (based on Variational Bayes) criteria. $\lambda = 0.9$, $\epsilon = 0.1$ and $Q_{True} \in \{2, \ldots, 5\}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	0	**100**	0	0
5	0	0	0	13	**87**	0

(a) $Q_{True} \backslash Q_{ICL}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	0	**100**	0	0
5	0	0	0	4	**95**	1

(b) $Q_{True} \backslash Q_{VB}$

These differences increase when considering less structured networks. For instance, in Table 3 and 4, when $Q_{True} = 5$, we notice that the percentage of accuracy of ICL falls down (respectively 29% and 3%) whereas the Bayesian criterion remains more stable (respectively 65% and 29%). Very similar remarks can be pointed out by looking at Table 5. Thus, when considering weaker and weaker modular structures, both criteria tend to underestimate the number of classes although the Bayesian criterion appears to be much more robust.

Table 3 Confusion matrices for ICL and Bayesian (based on Variational Bayes) criteria. $\lambda = 0.85$, $\epsilon = 0.15$ and $Q_{True} \in \{2, \ldots, 5\}$

	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

(a) $Q_{True} \backslash Q_{ICL}$

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

(b) $Q_{True} \backslash Q_{VB}$

Table 4 Confusion matrices for ICL and Bayesian (based on Variational Bayes) criteria. $\lambda = 0.8$, $\epsilon = 0.2$ and $Q_{True} \in \{2, \ldots, 5\}$

	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

(a) $Q_{True} \backslash Q_{ICL}$

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

(b) $Q_{True} \backslash Q_{VB}$

Table 5 Confusion matrices for ICL and Bayesian (based on Variational Bayes) criteria. $\lambda = 0.75$, $\epsilon = 0.25$ and $Q_{True} \in \{2, \ldots, 5\}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	1	**99**	0	0	0
4	1	28	50	**21**	0	0
5	30	51	19	0	**0**	0

(a) $Q_{True} \backslash Q_{ICL}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	1	**99**	0	0	0
4	0	13	45	**31**	11	0
5	9	42	38	8	**3**	0

(b) $Q_{True} \backslash Q_{VB}$

In Tables 6, 7, and 8, we consider sparser graphs (ADC smaller than 3.5). In general, for both criteria, we obtain less accurate predictions of the number of classes. However, as we saw for dense graphs, the Bayesian criterion remains more stable. This can be easily seen by looking, for example, at Table 8. Indeed, the Bayesian criteria has a percentage of accuracy of 39% against 14% for ICL.

Table 6 Confusion matrices for ICL and Bayesian (based on Variational Bayes) criteria. $\lambda = 0.49$, $\epsilon = 0.01$ and $Q_{True} \in \{2, \ldots, 5\}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	20	**80**	0	0
5	0	1	22	58	**19**	0

(a) $Q_{True} \backslash Q_{ICL}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	15	**84**	1	0
5	0	0	15	52	**31**	2

(b) $Q_{True} \backslash Q_{VB}$

We also used the Adjusted Rand Index (Hubert and Arabie (1985)) to evaluate the agreement between the true and estimated partitions. The computation of this index is based on a ratio between the number of node pairs belonging to the same and to different classes when considering the true par- tition and the estimated partition. Two identical partitions have an adjusted Rand index equal to 1. In the experiments we carried out, when the variational EM method and our algorithm were run on networks with the true number of latent classes, we obtained almost non-distinguishable Adjusted Rand Indices.

Table 7 Confusion matrices for ICL and Bayesian (based on Variational Bayes) criteria. $\lambda = 0.47$, $\epsilon = 0.03$ and $Q_{True} \in \{2, \ldots, 5\}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	4	46	**50**	0	0
5	1	18	59	19	**3**	0

(a) $Q_{True} \backslash Q_{ICL}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	0	**100**	0	0	0
4	0	2	34	**63**	1	0
5	0	8	51	36	**4**	1

(b) $Q_{True} \backslash Q_{VB}$

Table 8 Confusion matrices for ICL and Bayesian (based on Variational Bayes) criteria. $\lambda = 0.45$, $\epsilon = 0.05$ and $Q_{True} \in \{2, \ldots, 5\}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	3	**97**	0	0	0
4	2	24	60	**14**	0	0
5	12	56	30	2	**0**	0

(a) $Q_{True} \backslash Q_{ICL}$

	1	2	3	4	5	6
2	0	**100**	0	0	0	0
3	0	1	**99**	0	0	0
4	0	10	51	**39**	0	0
5	7	37	45	9	**2**	0

(b) $Q_{True} \backslash Q_{VB}$

Finally, we point out that we obtained almost the same results in this set of experiments by choosing uniform distributions ($n_q^0 = 1$, $\forall q \in \{1, \ldots, Q\}$) or Jeffreys distributions ($n_q^0 = \frac{1}{2}$, $\forall q \in \{1, \ldots, Q\}$) for the prior over the mixing coefficients.

5.2 The metabolic network of *Escherichia coli*

We apply the methodology described in this paper to the metabolic network of bacteria *Escherichia coli*. It is available at `http://pbil.univ-lyon1.fr/software/motus/`. In this network, there are 605 vertices which represent chemical reactions and a total number of 1782 edges. Two reactions are connected if a compound produced by the first one is a part of the second one (or vice-versa). We emphasize that, to the best of our knowledge, our inference method is currently the only Bayesian approach that can handle efficiently such large network. As in the previous Section, we considered uniform priors: we fixed $n_q^0 = 1$, $\forall q \in \{1, \ldots, Q\}$ for the Dirichlet prior and $\eta_{ql}^0 = \zeta_{ql}^0 = 1$, $\forall (q, l) \in \{1, \ldots, Q\}^2$ for the Beta priors. We compared our results with the work of Daudin et al (2008) on the same data set. They used ICL to estimate the number of latent classes and Variational EM to estimate MixNet parameters.

For the initialization and in order to obtain comparable results, we applied the same hierarchical clustering method they used in their experiments. Thus, for $Q \in \{1, \ldots, 40\}$, after having initialized $\boldsymbol{\tau}$, we ran our Bayesian algorithm and we computed the criterion proposed in Section 3.3. We repeated such procedure 60 times. The results are presented as boxplots in Figure 3. The criterion finds its maximum for $Q_{VB} = 22$ classes. Thus, for this particular real network, both criteria lead to almost the same estimates of the number of latent classes. Indeed, Daudin et al (2008) obtained $Q_{ICL} = 21$.

We also compared the learnt partitions in the Bayesian and in the frequentist approach. Figure 4 is a dot plot representation of the metabolic network after having applied the Bayesian algorithm for $Q_{VB} = 22$. Each vertex i was classified into the class for which τ_{iq} was maximal (Maximum A Posteriori estimate). We observe very similar patterns as in Daudin et al (2008). For instance, class 1 and 17 both have an expected probability of intra-connection equals to 1: $\mathsf{E}[\pi_{1-1}] = \mathsf{E}[\pi_{17-17}] = 1$ and so they correspond to cliques. Moreover, the expected probability of connection between class 1 and class 17 is also 1: $\mathsf{E}[\pi_{1-17}] = \mathsf{E}[\pi_{17-1}] = 1$. In other words, these two classes constitute in fact a single clique. However, that clique is split into two sub-cliques because of their different connectivities with vertices of classes 7 and 10.

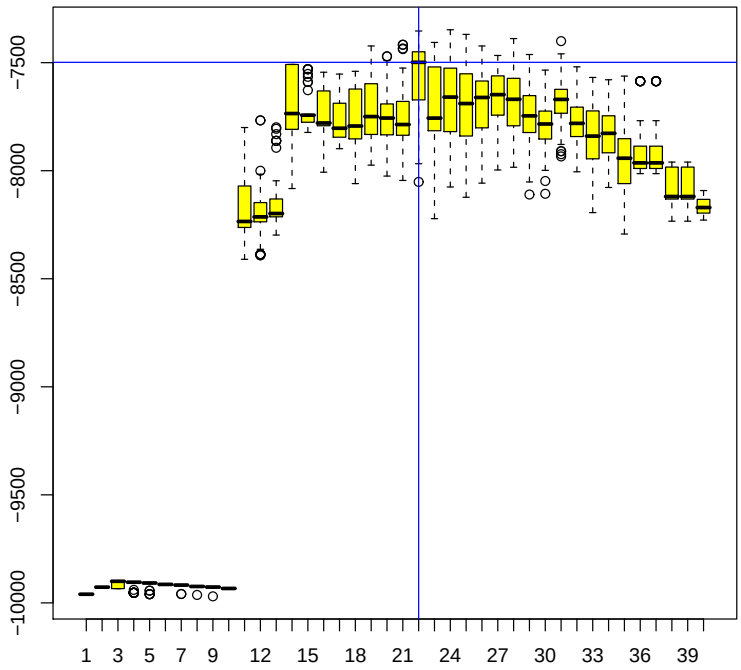

Figure 3 Boxplot representation (over 60 experiments) of the Bayesian criterion for $Q \in \{1, \ldots, 40\}$. The maximum is reached at $Q_{VB} = 22$

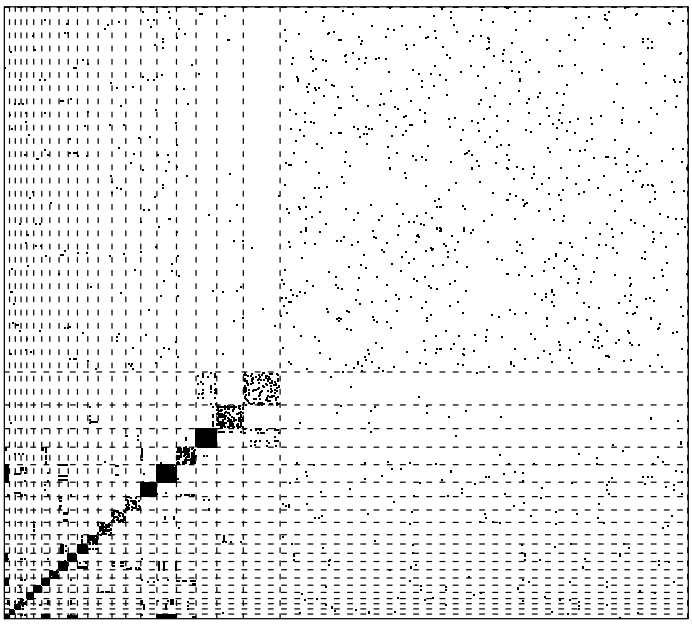

Figure 4 Dot plot representation of the metabolic network after classification of the vertices into $Q_{VB} = 22$ classes

6. Conclusion

In this paper, we showed how the MixNet model, also called the Block-Clustering model, could be described in a full Bayesian framework. Thus, we introduced priors over the model parameters and we developed a procedure, based on Variational Bayes, to approximate the posterior distribution of all the hidden variables given the observed edges. In this framework, we derived a new non-asymptotic Bayesian criterion to select the number of classes in latent structures. We found that our criterion was more robust than the criterion we denoted ICL in this paper and which is based on an asymptotic approximation of the Integrated Classification Likelihood. Indeed, by considering small networks and complex modular structures, we found that the percentage of accuracy of our criterion was always higher. Finally, using the metabolic network of *Escherichia coli*, we noticed that, contrary to the Bayesian methods that had been developed, our Bayesian strategy was able to handle large networks. For this particular network, we obtained almost the same results as the frequentist and ICL strategies. Overall, our Bayesian approach seems very promising for the investigation of rather small networks and/or based on complex structures.

Acknowledgments

The data has been provided by V. Lacroix and M.-F. Sagot (INRIA-Helix, INRIA, Lyon). The authors would like to thank G. Celeux (INRIA, Paris-sud univ) for his helpful remarks and suggestions.

References

Albert R, Barabási A (2002) Statistical mechanics of complex networks. Modern Physics 74:47–97

Attias H (1999) Inferring parameters and structure of latent variable models by variational bayes. In: Laskey K, Prade H (eds) Uncertainty in Artificial Intelligence : proceedings of the fifth conference, Morgan Kaufmann, pp 21–30

Barabási A, Oltvai Z (2004) Network biology : understanding the cell's functional organization. Nature Rev Genet 5:101–113

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Machine Intel 7:719–725

Bishop C (2006) Pattern recognition and machine learning. Springer-Verlag

Burnham K, Anderson D (2004) Model selection and multi-model inference : a practical information-theoric approach. Springer-Verlag

Corduneanu A, Bishop C (2001) Variational bayesian model selection for mixture distributions. In: Richardson T, Jaakkola T (eds) Proceedings eighth international conference on artificial intelligence and statistics, Morgan Kaufmann, pp 27–34

Daudin J, Picard F, Robin S (2008) A mixture model for random graph. Statistics and computing 18:1–36

Dempster A, Laird N, Rubin DB (1977) Maximum likelihood for incomplete data via the em algorithm. Journal of the Royal Statistical Society B39:1–38

Frank O, Harary F (1982) Cluster inference by using transitivity indices in empirical graphs. Journal of the American Statistical Association 77:835–840

Hofman J, Wiggins C (2008) A bayesian approach to network modularity. Physical review letters 100

Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2:193–218

Jeffreys H (1946) An invariant form for the prior probability in estimations problems. In: Proceedings of the Royal Society of London. Series A, vol 186, pp 453–461

MacKay D (1992) A practical bayesian framework for backpropagation networks. Neural Computation 4:448–472

Mariadassou M, Robin S (2007) Uncovering latent structure in networks. Tech. rep., INRA, SSB

Newman M, Leicht E (2007) Mixture models and exploratory analysis in networks. PNAS 104:9564–9569

Ng A, Jordan M, Weiss Y (2001) On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 14

Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. Nature 446:664–667

Parisi G (1988) Statistical field theory. Addison Wesley

Snijders T, Nowicki K (1997) Estimation and prediction for stochastic block-structures for graphs with latent block sturcture. Journal of Classification 14:75–100

Svensén M, Bishop C (2004) Robust bayesian mixture modelling. Neurocomputing 64:235–252

Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393:440–442

Zanghi H, Ambroise C, Miele V (2007) Fast online graph clustering via erdös-rényi mixture. Tech. rep., INRA, SSB

Appendix

The lower bound is given by

$$
\begin{aligned}
\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\alpha d\pi \\
&= \mathsf{E}_{\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi}}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi})] - \mathsf{E}_{\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi}}[\ln q(\mathbf{Z},\boldsymbol{\alpha},\boldsymbol{\pi})] \\
&= \mathsf{E}_{\mathbf{Z},\boldsymbol{\pi}}[\ln p(\mathbf{X}|\mathbf{Z},\boldsymbol{\pi})] + \mathsf{E}_{\mathbf{Z},\boldsymbol{\alpha}}[\ln p(\mathbf{Z}|\boldsymbol{\alpha})] + \mathsf{E}_{\boldsymbol{\alpha}}[\ln p(\boldsymbol{\alpha})] + \mathsf{E}_{\boldsymbol{\pi}}[\ln p(\boldsymbol{\pi})] \\
&\quad - \sum_{i=1}^{N} \mathsf{E}_{\mathbf{Z_i}}[\ln q(\mathbf{Z_i})] - \mathsf{E}_{\boldsymbol{\alpha}}[\ln q(\boldsymbol{\alpha})] - \mathsf{E}_{\boldsymbol{\pi}}[\ln q(\boldsymbol{\pi})] \\
&= \sum_{i<j}^{N}\sum_{q,l}^{Q} \tau_{iq}\tau_{jl}\left(X_{ij}\Big(\psi(\eta_{ql}) - \psi(\zeta_{ql})\Big) + \psi(\zeta_{ql}) - \psi(\eta_{ql}+\zeta_{ql})\right) \\
&\quad + \sum_{i=1}^{N}\sum_{q=1}^{Q} \tau_{iq}\Big(\psi(n_q) - \psi(\sum_{l=1}^{Q} n_l)\Big) + \ln\Gamma(\sum_{q=1}^{Q} n_q^0) - \sum_{q=1}^{Q}\ln\Gamma(n_q^0) \\
&\quad + \sum_{q=1}^{Q}\Big(n_q^0 - 1\Big)\Big(\psi(n_q) - \psi(\sum_{l=1}^{Q} n_l)\Big) + \sum_{q\le l}^{Q}\Bigg(\ln\Gamma(\eta_{ql}^0+\zeta_{ql}^0) \\
&\quad - \ln\Gamma(\eta_{ql}^0) - \ln\Gamma(\zeta_{ql}^0) + (\eta_{ql}^0 - 1)\Big(\psi(\eta_{ql}) - \psi(\eta_{ql}+\zeta_{ql})\Big) \\
&\quad + (\zeta_{ql}^0 - 1)\Big(\psi(\zeta_{ql}) - \psi(\eta_{ql}+\zeta_{ql})\Big)\Bigg) - \sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}\ln\tau_{iq} \\
&\quad - \ln\Gamma(\sum_{q=1}^{Q} n_q) + \sum_{q=1}^{Q}\ln\Gamma(n_q) - \sum_{q=1}^{Q}\Big(n_q - 1\Big)\Big(\psi(n_q) - \psi(\sum_{l=1}^{Q} n_l)\Big) \\
&\quad - \sum_{q\le l}^{Q}\Bigg(\ln\Gamma(\eta_{ql}+\zeta_{ql}) - \ln\Gamma(\eta_{ql}) - \ln\Gamma(\zeta_{ql}) \\
&\quad + (\eta_{ql} - 1)\Big(\psi(\eta_{ql}) - \psi(\eta_{ql}+\zeta_{ql})\Big) + (\zeta_{ql} - 1)\Big(\psi(\zeta_{ql}) - \psi(\eta_{ql}+\zeta_{ql})\Big)\Bigg) \\
&= \sum_{q<l}^{Q}\Bigg(\Big(\eta_{ql}^0 - \eta_{ql} + \sum_{i\neq j}^{N}\tau_{iq}\tau_{jl}X_{ij}\Big)\Big(\psi(\eta_{ql}) - \psi(\eta_{ql}+\zeta_{ql})\Big) \\
&\quad + \Big(\zeta_{ql}^0 - \zeta_{ql} + \sum_{i\neq j}^{N}\tau_{iq}\tau_{jl}(1 - X_{ij})\Big)\Big(\psi(\zeta_{ql}) - \psi(\eta_{ql}+\zeta_{ql})\Big)\Bigg) \\
&\quad + \sum_{q=1}^{Q}\Bigg(\Big(\eta_{qq}^0 - \eta_{qq} + \sum_{i<j}^{N}\tau_{iq}\tau_{jq}X_{ij}\Big)\Big(\psi(\eta_{qq}) - \psi(\eta_{qq}+\zeta_{qq})\Big) \\
&\quad + \Big(\zeta_{qq}^0 - \zeta_{qq} + \sum_{i<j}^{N}\tau_{iq}\tau_{jq}(1 - X_{ij})\Big)\Big(\psi(\zeta_{qq}) - \psi(\eta_{qq}+\zeta_{qq})\Big)\Bigg) \\
&\quad + \sum_{q=1}^{Q}\Big(n_q^0 - n_q + \sum_{i=1}^{N}\tau_{iq}\Big)\Big(\psi(n_q) - \psi(\sum_{l=1}^{Q} n_l)\Big) \\
&\quad + \ln\{\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)}\} + \sum_{q\le l}^{Q}\ln\{\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)}\} \\
&\quad - \sum_{i=1}^{N}\sum_{q=1}^{Q}\tau_{iq}\ln\tau_{iq}.
\end{aligned}
\tag{22}
$$