

Bayesian Methods for Graph Clustering

Pierre Latouche, Etienne Birmelé, and Christophe Ambroise

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 data sets, we compare our approach to other strategies. We show that our criterion can outperform ICL.

Keywords Bayesian model selection · Block-clustering model · Integrated classification likelihood · Random graphs · Variational Bayes · Variational EM

1 Introduction

For the last few years, networks have been increasingly studied. Indeed, many scientific fields such as biology, social science, and information technology, see those mathematical structures as powerful tools to model the interactions between objects of interest. Examples of data sets having such structures are friendship and

Pierre Latouche (✉)

Laboratoire Statistique et Génome, UMR CNRS 8071-INRA 1152-UEVE, 91000 Evry, France,
e-mail: pierre.latouche@genopole.cnrs.fr

protein–protein interaction networks, powergrids, and the Internet. 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 (Snijders & Nowicki, 1997; Daudin, Picard, & Robin, 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 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. proposed a variational approximation of the posterior, which corresponds to a mean-field approximation.

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) 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 asymptotic approximation of the Integrated Classification Likelihood (also called Integrated Complete-data Likelihood). More details can be found in Biernacki, Celeux, and Govaert (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, we extend the work of Hofman and Wiggins (2008) who developed a variational Bayes algorithm to learn *affiliation models*. These are defined by only two probabilities of connection λ and ϵ . Given a network, it is assumed that the edges connecting nodes of the *same* class were generated with probability λ while edges connecting nodes of *different* classes were drawn with probability ϵ . The algorithm that they proposed can cluster the nodes and estimate the probabilities λ and ϵ very quickly. However, affiliation models can not characterize the complex topology of most real networks, which have the majority of their nodes with none or very few links and exhibit some hubs which make them locally dense (Daudin et al., 2008). Therefore, we propose an efficient Bayesian version of MixNet, which allows vertices to have different topological behaviors. Thus, after having presented MixNet in Sect. 2, we introduce some prior distributions and describe the MixNet Bayesian probabilistic model in Sect. 3. We derive the model optimization equations using Variational Bayes and we propose a new criterion to estimate the number of

classes. Finally, in Sect. 4, we carry out some experiments using simulated data sets to compare the number of the estimated clusters obtained with the ICL criterion and the variational frequentist strategy, and our approach.

An extended version of this paper with proofs of the results and more experiments is available (Latouche, Birmelé, & Ambroise, 2008).

2 A Mixture Model for Networks

We consider an undirected 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. We assume that G does not have any self loop. Therefore, the variables X_{ii} will not be taken into account.

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

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

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

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

Thus, contrary to affiliation models (Hofman & Wiggins, 2008), we consider a $Q \times Q$ matrix $\boldsymbol{\pi}$ of connection probabilities. Note that in the case of undirected networks, $\boldsymbol{\pi}$ is symmetric. The latent variables in the set $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_N)$ are *iid* and given this latent structure, all the edges are supposed to be independent. Thus, 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

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

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:

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

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. 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 $\boldsymbol{\pi}$:

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

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. Since $\boldsymbol{\pi}$ is symmetric, only the terms of the upper or lower triangular matrix have to be considered. This explains the product over $q \leq l$.

Thus, the model parameters are now seen as random variables. 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.

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) and Corduneanu and Bishop (2001) 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

$$\ln p(\mathbf{X}) = \mathcal{L}(q(\cdot)) + \text{KL}(q(\cdot) \parallel p(\cdot|\mathbf{X})), \tag{2}$$

where

$$\mathcal{L}(q(\cdot)) = \sum_{\mathbf{Z}} \iint q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})} \right\} d\boldsymbol{\alpha} d\boldsymbol{\pi}, \tag{3}$$

and

$$\text{KL}(q(\cdot) \parallel p(\cdot|\mathbf{X})) = - \sum_{\mathbf{Z}} \iint q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}) \ln \left\{ \frac{p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\mathbf{X})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\pi})} \right\} d\boldsymbol{\alpha} d\boldsymbol{\pi}. \tag{4}$$

Minimizing (4) is equivalent to maximizing the lower bound (3) of (2). 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 use a variational Bayes EM algorithm. 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 only in the case of an undirected graph G without self-loop. Our algorithm cycles through the E and M steps until convergence of the lower bound (11).

3.2.1 Variational Bayes E-Step

The optimal approximation at vertex i is

$$q(\mathbf{Z}_i) = \mathcal{M}(\mathbf{Z}_i; 1, \boldsymbol{\tau}_i = \{\tau_{i1}, \dots, \tau_{iQ}\}), \tag{5}$$

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

$$\tau_{iq} \propto e^{\psi(n_q) - \psi(\sum_{l=1}^Q n_l)} \prod_{j \neq i} \prod_{l=1}^Q e^{\tau_{jl} (\psi(\zeta_{ql}) - \psi(\eta_{ql} + \zeta_{ql}) + X_{ij} (\psi(\eta_{ql}) - \psi(\zeta_{ql})))}, \tag{6}$$

where $\psi(\cdot)$ is the *digamma* function. Given a matrix $\boldsymbol{\tau}^{old}$, the algorithm builds a new matrix $\boldsymbol{\tau}^{new}$ where each row satisfies (6). It then uses $\boldsymbol{\tau}^{new}$ to build a new matrix and so on. It stops when $\sum_{i=1}^N \sum_{q=1}^Q |\tau_{iq}^{old} - \tau_{iq}^{new}| < e$. A rather small values for e has to be chosen. In the experiments that we carried out, we chose $e = 10^{-14}$.

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

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

$$q(\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\alpha}; \mathbf{n}), \quad (7)$$

where $n_q = n_q^0 + \sum_{i=1}^N \tau_{iq}$ is the pseudo number of vertices in the q -th component of the mixture.

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

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

$$q(\boldsymbol{\pi}) = \prod_{q \leq l} \text{Beta}(\pi_{ql} | \eta_{ql}, \zeta_{ql}), \quad (8)$$

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

$$\eta_{ql} = \eta_{ql}^0 + \sum_{i \neq j}^N X_{ij} \tau_{iq} \tau_{jl}, \quad \text{and } \forall q : \eta_{qq} = \eta_{qq}^0 + \sum_{i < j}^N X_{ij} \tau_{iq} \tau_{jq}. \quad (9)$$

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

$$\zeta_{ql} = \zeta_{ql}^0 + \sum_{i \neq j}^N (1 - X_{ij}) \tau_{iq} \tau_{jl}, \quad \text{and } \forall q : \zeta_{qq} = \zeta_{qq}^0 + \sum_{i < j}^N (1 - X_{ij}) \tau_{iq} \tau_{jq}. \quad (10)$$

3.2.4 Lower Bound

The lower bound takes a simple form after the Variational Bayes M-step. Indeed, it only depends on the posterior probabilities τ_{iq} as well as the normalizing constants of the Dirichlet and Beta distributions:

$$\begin{aligned} \mathcal{L}(q(\cdot)) = \ln & \left\{ \frac{\Gamma(\sum_{q=1}^Q n_q^0) \prod_{q=1}^Q \Gamma(n_q)}{\Gamma(\sum_{q=1}^Q n_q) \prod_{q=1}^Q \Gamma(n_q^0)} \right\} \\ & + \sum_{q \leq l}^Q \ln \left\{ \frac{\Gamma(\eta_{ql}^0 + \xi_{ql}^0) \Gamma(\eta_{ql}) \Gamma(\xi_{ql})}{\Gamma(\eta_{ql} + \xi_{ql}) \Gamma(\eta_{ql}^0) \Gamma(\xi_{ql}^0)} \right\} - \sum_{i=1}^N \sum_{q=1}^Q \tau_{iq} \ln \tau_{iq}. \end{aligned} \tag{11}$$

3.3 Model Selection

We have not addressed yet the problem of estimating the number of classes in the mixture. Given a set of values of Q , our goal is to select Q^* which maximizes the log-probability of the observed edges $\ln p(\mathbf{X}|Q)$. Unfortunately, this quantity does not have any analytical expression. Indeed, for each value of Q , it involves integrating over all the hidden parameters as shown in Sect. 3.2. Nevertheless, it can be approximated using our Variational Bayes algorithm. Given a value of Q , the algorithm is used to maximize the lower bound (11). Meanwhile, the Kullback–Leibler divergence between the factorized and the unknown posterior distribution decreases. After convergence, although this distance can not be computed analytically, we expect it to be close to zero, and therefore, we can use the lower bound as an approximation of $\ln p(\mathbf{X}|Q)$. This procedure is repeated for the different values of Q considered.

Given a value of Q , since MixNet is a mixture model, for any given setting of the parameters α and π there will be a total of $Q!$ parameters which lead to the same distribution over the edges. These parameters would differ only through re-labelling of the components. In a frequentist framework, this redundancy is irrelevant since we only look for point estimates of the model parameters. In a Bayesian setting, however, we integrate over all possible parameter values. Since $p(\mathbf{X}|Q)$ is multimodal, variational techniques will tend to approximate the distribution in the neighborhood of one of the mode and ignore the others (see Bishop, 2006). Thus, when comparing different values of Q , we need to take this multimodality into account. As a consequence, we define a criterion by subtracting a term $\ln Q!$ from the lower bound (11) computed previously.

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 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 Sect. 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 to estimate the model parameters. We considered synthetic data, generated according to known random graph models and we concentrated on analyzing the capacity of ICL and our criterion to retrieve the true number of classes in the latent structures.

4.1 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 case where $\lambda = 1 - \epsilon$. Thus, we considered three affiliation models shown in Table 1.

For each affiliation model, we analyzed graphs with $Q_{True} \in \{2, \dots, 5\}$ classes mixed in the same proportions $\alpha_1 = \dots = \alpha_{Q_{True}} = \frac{1}{Q_{True}}$. Thus, we studied a total of 12 graph models.

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 and the variational EM approach of Daudin et al. (2008) on each network, for various numbers of classes $Q \in \{1, \dots, 6\}$. Note that, we chose $n_q^0 = 1, \forall q \in \{1, \dots, Q\}$ for the Dirichlet prior and $\eta_{ql}^0 = \zeta_{ql}^0 = 1, \forall (q, l) \in \{1, \dots, Q\}^2$ 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 τ obtained using a spectral clustering method (Ng, Jordan, & Weiss, 2001). Then, for the Bayesian algorithm, we used the criterion we proposed in Sect. 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 $Q \in \{1, \dots, 6\}$ 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 1 Parameters of the three affiliation models considered

Model	λ	ϵ
1	0.9	0.1
2	0.85	0.15
3	0.8	0.2

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

	1	2	3	4	5	6		1	2	3	4	5	6
2	0	100	0	0	0	0	2	0	100	0	0	0	0
3	0	0	100	0	0	0	3	0	0	100	0	0	0
4	0	0	0	100	0	0	4	0	0	0	100	0	0
5	0	0	0	13	87	0	5	0	0	0	4	95	1
	(a) $Q_{True} \setminus Q_{ICL}$							(b) $Q_{True} \setminus Q_{VB}$					

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

	1	2	3	4	5	6		1	2	3	4	5	6
2	0	100	0	0	0	0	2	0	100	0	0	0	0
3	0	0	100	0	0	0	3	0	0	100	0	0	0
4	0	0	1	98	1	0	4	0	0	0	98	2	0
5	0	0	10	61	29	0	5	0	0	1	29	65	5
	(a) $Q_{True} \setminus Q_{ICL}$							(b) $Q_{True} \setminus Q_{VB}$					

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

	1	2	3	4	5	6		1	2	3	4	5	6
2	0	100	0	0	0	0	2	0	100	0	0	0	0
3	0	0	100	0	0	0	3	0	0	100	0	0	0
4	0	0	14	86	0	0	4	0	0	5	94	1	0
5	0	17	36	44	3	0	5	0	4	18	43	29	6
	(a) $Q_{True} \setminus Q_{ICL}$							(b) $Q_{True} \setminus Q_{VB}$					

These differences increase when considering less structured networks. For instance, in Tables 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%). 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 stable.

In all the tables presented before, we did not specify what happens when $Q_{True} = 1$. Indeed, both techniques have always a 100% accuracy. We did not stipulate either what happens when $Q_{True} = 6$. In general, our results were very similar to what we obtained when considering $Q_{True} = 5$. We also used the Adjusted Rand Index (Hubert & 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 partition 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.

Moreover, we point out that we obtained almost the same results in this set of experiments by choosing uniform distributions ($n_q^0 = 1, \forall q \in \{1, \dots, Q\}$) or Jeffreys distributions ($n_q^0 = \frac{1}{2}, \forall q \in \{1, \dots, Q\}$) for the prior over the mixing coefficients. Finally, we compared the computational costs of the frequentist approach of variational EM and our Variational Bayes algorithm. Both are equal to $O(Q^2N^2)$. Analyzing a sparse network with 200 nodes takes about a minute, and about a hour for dense networks.

5 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 relevant 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. Overall, our Bayesian approach seems very promising for the investigation of rather small networks and/or based on complex structures.

References

- Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In K. B. Laskey, & H. Prade (Eds.), *Uncertainty in artificial intelligence: Proceedings of the fifth conference* (pp. 21–30). San Fransisco, CA: Morgan Kaufmann.
- Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7, 719–725.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.
- Corduneanu, A., & Bishop, C. M. (2001). Variational bayesian model selection for mixture distributions. In T. Richardson, & T. Jaakkola (Eds.), *Proceedings eighth international conference on artificial intelligence and statistics* (pp. 27–34). San Fransisco, CA: Morgan Kaufmann.
- Daudin, J., Picard, F., & Robin, S. (2008). A mixture model for random graph. *Statistics and Computing*, 18, 1–36.
- Hofman, J. M., & Wiggins, C. H. (2008). A bayesian approach to network modularity. *Physical Review Letters*, 100, 258701.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.
- Latouche, P., Birmelé, E., & Ambroise, C. (2008). *Bayesian methods for graph clustering*. Technical report, SSB.
- 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*. Technical report, INRA, SSB.
- Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14.
- Parisi, G. (1988). *Statistical field theory*. Reading MA: Addison Wesley.
- Snijders, T. A. B., & Nowicki, K. (1997). Estimation and prediction for stochastic block-structures for graphs with latent block structure. *Journal of Classification*, 14, 75–100.