Uncovering Latent Structure in Valued Graphs: A Variational Approach

by

Mahendra Mariadassou and Stéphane Robin

Research Report No. 10
October 2007

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

Uncovering Latent Structure in Valued Graphs: A Variational Approach

M. Mariadassou[1], S. Robin[1]

UMR 518 AgroParisTech/INRA Applied Mathematics and Computer Sciences
16, rue C. Bernard, F-75005 Paris, France

mariadas@agroparistech.fr

Abstract

As more and more network-structured datasets are available, the statistical analysis of valued graphs has become a common place. Looking for a latent structure is one of the many strategies used to better understand the behavior of a network. Several methods already exist for the binary case.

We present a model-based strategy to uncover groups of nodes in valued graphs. This framework can be used for a wide span of parametric random graphs models. Variational tools allow us to achieve approximate maximum likelihood estimation of the parameters of these models. We provide a simulation study showing that our estimation method performs well over a broad range of situations.

Keywords: Latent structure, Mixture model, Random graph, Valued graph, Variational method

1 Introduction

Datasets presenting a network structure are more and more studied in many different domains such as sociology, energy, communication or biology (Albert and Barabási (2002)). Statistical tools are therefore needed to analyze the structure of these networks, in order to understand there properties or behavior. A strong attention has been paid to the study of various topological characteristics such as degree distribution, clustering coefficient, diameter (see e.g. Barabási and Albert (1999), Newman et al. (2002)). These characteristics are useful to describe networks but not sufficient to understand its whole structure. For this latter purpose, a proper probabilistic model is desirable (see Pattison and Robins (2007) for a review).

Looking for groups of edges having similar connection profiles seems a natural way to capture an underlying structure of the network (Getoor and Diehl (2004), Newman et al. (2002), ...), sometimes refereed to as community structure (Girvan and Newman (2002); Newman (2004b) and Newman (2004a) for an extension to weigthed graphs). This turns into an un-supervised classification (or clustering) problem which requires efficient estimation algorithms since the datasets at hand are getting ever larger.

Several model-based approaches have been proposed, mostly for binary networks, *i.e.* when the only information is the presence or absence of the edges. In this framework, the stochastic block model proposed by Nowicki and Snijders (2001) is the reference at this time. This model is valid for both directed and undirected graphs but the Bayesian estimation strategy leads to strong limitations in the network size (limited to 200 nodes). Recently, Daudin et al. (2007) proposed a variational approach (see Jaakkola (2000)) to perform approximate maximum likelihood estimation of the parameters. This approach turns out to be much more efficient in terms of computation (Picard et al. (2007)). Methods not referring to a specific random graph model have also been proposed. For example, spectral graph clustering (von Luxburg et al. (2007)) aims at detecting classes of nodes with strong within connectivity.

Binary information only describes the topology of the network, but does not account for the intensity of the interactions between the nodes. However, power, communication, social or biological networks are often valued. The intensity of an edge may typically indicate the amount of energy transported from one node to another, the number of passengers or the number of co-publications. A statistical model accounting for these intensities is needed to analyze such networks.

In this paper, we propose a general mixture model describing the connection intensities between nodes spread among a certain number of classes (Section 2). A variational approach to get an optimal, in a sense to be defined, approximation of the likelihood is then presented in Section 3. In Section 4 we give a general estimation algorithm and derive some explicit formulas for the most popular distributions. The quality of the estimates is studied on synthetic data in Section 5.

2 Mixture Model

2.1 Model and Notations

Nodes. Consider a graph with n nodes, labeled in $\{1, \ldots, n\}$. In our model, the nodes are distributed among Q groups so that each node i is associated to a random vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iQ})$, with Z_{iq} being 1 if node i belongs to group q and 0 otherwise. The $\{\mathbf{Z}_i\}$ are supposed to be independent identically distributed observations from a multinomial distribution:

$$\{\mathbf{Z}_i\}_i \text{ i.i.d. } \sim \mathcal{M}(1; \boldsymbol{\alpha}) \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$ $(\sum_q \alpha_q = 1)$.

Edges. Each edge from a node i to a node j is associated to a random variable X_{ij}, coding for the strength of the edge. Conditionally to the group of each node, or equivalently knowing the $\{\mathbf{Z}_i\}$, the edges are supposed to be independent. Knowing group q of node i and group ℓ of node j, X_{ij} is distributed as $f(\cdot, \theta_{q\ell}) := f_{q\ell}(\cdot)$, where $f_{\theta_{q\ell}}$ is a probability distribution

known up to a finite dimensional parameter $\theta_{q\ell}$.

$$X_{ij}|i \in q, j \in \ell \sim f(\cdot, \theta_{q\ell}) := f_{q\ell}(\cdot) \tag{2}$$

Up to a relabeling of the classes, the model is identifiable and completely specified by both the mixture proportions $\boldsymbol{\alpha}$ and the connectivity matrix $\boldsymbol{\theta} = (\theta_{q\ell})_{q,\ell=1\ldots Q}$. We denote $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$ the parameter of the model.

Directed and undirected graphs. This modeling can be applied to both directed and undirected graphs. In the directed version, the variables X_{ij} and X_{ji} are supposed to be independent conditionally to the groups to which nodes i and j belong. This hypothesis is not always realistic since, for example, the traffic from i to j is likely to be correlated to the traffic from j to i. A way to account for such a dependency is to consider an undirected graph with edges labeled with the bivariate variables $\{(X_{ij}, X_{ji})\}_{1\leq i<j\leq n}$. All the results presented in this paper are valid for directed graphs. The results for undirected graphs can easily be derived and are only briefly mentioned.

2.2 Classical Distributions

We examine some classical distributions entering the framework of model.

Bernoulli. In some situations such as co-authorship or social networks, the only available information is the presence or absence of the edge. X_{ij} is then supposed to be Bernoulli distributed:

$$X_{ij}|i \in q, j \in \ell \sim \mathcal{B}(\pi_{q\ell}).$$

It is equivalent to the stochastic block model of Nowicki and Snijders (2001) or Daudin et al. (2007).

Multinomial. In a social network, X_{ij} may inform about the nature of the existing relation: professional, family, friend, etc. The X_{ij}s can then be modeled by multinomial variables:

$$X_{ij}|i \in q, j \in \ell \sim \mathcal{M}(1; \mathbf{p}_{q\ell}).$$

The parameter $\theta_{q\ell}$ to estimate is the vector of probability $\mathbf{p}_{q\ell} = \left(p^1_{q\ell}, \ldots, p^m_{q\ell}\right)$, m being the number of possible labels.
In directed random graphs, this setting allows to account for some dependency between symmetric edges X_{ij} and X_{ji}. We only need to consider the equivalent undirected graphs where edge (i, j) is labeled with the couple (X_{ij}, X_{ji}). $m = 4$ different labels can the be observed: $(0, 0)$ if no edge exists, $(1, 0)$ for $i \to j$, $(1, 1)$ for $i \leftarrow j$ and $(1, 1)$ for $i \leftrightarrow j$.

Poisson. In a co-authorship network, edges may be valued with the numbers of articles two authors co-published. X_{ij} can then be supposed to be Poisson distributed:

$$X_{ij}|i \in q, j \in \ell \sim \mathcal{P}(\lambda_{q\ell}).$$

$\theta_{ql} := \lambda_{q\ell}$ is the mean number of common papers (or mean interaction) between an author from group q and one from group ℓ.

Gaussian. Traffic networks describe the intensity of the traffic between nodes. The airport network is a typical example where the edges are valued according to the number of passenger travelling from airport i to airport j. The intensity X_{ij} of the traffic can be assumed to be Gaussian:

$$X_{ij}|i \in q, j \in \ell \sim \mathcal{N}(\mu_{q\ell}, \sigma^2_{q\ell}), \qquad \theta_{q\ell} = (\mu_{q\ell}, \sigma^2_{q\ell}).$$

Bivariate Gaussian. The correlation between symmetric edges X_{ij} and X_{ji} can be accounted for, considering the undirected valued graph where edge (i, j) is valued by (X_{ij}, X_{ji}), which is assumed to be Gaussian. Denoting $\mathbf{X}_{ij} = [X_{ij}\ X_{ji}]'$:

$$\mathbf{X}_{ij}|i \in q, j \in \ell \sim \mathcal{N}(\boldsymbol{\mu}_{q\ell}, \boldsymbol{\Sigma}_{q\ell}) \qquad \theta_{q\ell} = (\boldsymbol{\mu}_{q\ell}, \boldsymbol{\Sigma}_{q\ell}).$$

Linear regression. In case of real valued edges, the linear Gaussian model allows to include covariates. Denoting $\mathbf{y}_{ij}$ the $p \times 1$ vector of covariates describing edge (i, j), we set

$$X_{ij}|i \in q, j \in \ell \sim \mathcal{N}\left(\mathbf{y}'_{ij}\boldsymbol{\beta}_{q\ell}, \sigma^2_{q\ell}\right) \qquad \theta_{q\ell} = (\boldsymbol{\beta}_{q\ell}, \sigma_{q\ell}).$$

Covariates can also be involved when edges are integer valued (e.g. Bernoulli or Poisson) using the generalized linear model framework.

3 Likelihood and Variational Inference

We now address the estimation of the parameter $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$. We show that the standard maximum likelihood approach can not be applied to our model and propose an alternative strategy based on a variational approach.

3.1 Likelihoods

Let $\mathbf{X}$ denote the set of all edges: $\mathbf{X} = \{X_{ij}\}_{i,j=1..n}$, and $\mathbf{Z}$ the set of all indicator variables for nodes: $\mathbf{Z} = \{\mathbf{Z}_i\}_{i=1,n}$. In the mixture model literature (McLahan and Peel (2000)) $(\mathbf{X}, \mathbf{Z})$ is referred to as the complete dataset, while $\mathbf{X}$ is referred to as the incomplete dataset.

Proposition 1 *The log-likelihood of the complete dataset is*

$$\log \mathbb{P}(\mathbf{Z}, \mathbf{X}) = \sum_i \sum_q Z_{iq} \log \alpha_q + \sum_{i \neq j} \sum_{q,\ell} Z_{iq} Z_{j\ell} \log f_{q\ell}(X_{ij}).$$

Proof. This comes directly from (1) and (2) and from the decomposition $\log \mathbb{P}(\mathbf{Z}, \mathbf{X}) = \log \mathbb{P}(\mathbf{Z}) + \log \mathbb{P}(\mathbf{X}|\mathbf{Z})$. ■

The likelihood of the incomplete dataset can be obtained by summing $\mathbb{P}(\mathbf{Z}, \mathbf{X})$ over all possible $\mathbf{Z}$'s: $\mathbb{P}(\mathbf{X}) = \sum_{\mathbf{Z}} \mathbb{P}(\mathbf{Z}, \mathbf{X})$ This summation involves Q^n terms and quickly becomes intractable. The popular E-M algorithm (Dempster et al. (1977)), widely used in mixture problems, allows to maximize $\log \mathbb{P}(\mathbf{X})$ without calculating it. The E-step relies on the calculation of the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$: $\mathbb{P}(\mathbf{Z}|\mathbf{X})$. Unfortunately, in the case of network structured data, the strong dependency between edges makes the calculation of this conditional distribution out of reach.

Undirected graphs. Proposition 1 also holds for undirected graphs, replacing the sum over $i \neq j$ by a sum over $i < j$. This is also true for Propositions 2 and 4 given below.

3.2 Variational Inference

We propose to use an approximate maximum likelihood strategy based on a variational approach (see Jordan et al. (1999) or the tutorial by Jaakkola (2000)). This strategy is for example used in Govaert and Nadif (2005) for a biclustering problem. We consider a lower bound of the log-likelihood of the incomplete dataset:

$$\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma}) = \log \mathbb{P}(\mathbf{X}; \boldsymbol{\gamma}) - KL(R_{\mathbf{X}}(\cdot), \mathbb{P}(\cdot|\mathbf{X}; \boldsymbol{\gamma})) \tag{3}$$

where KL denotes the Kullback-Leibler divergence and $R_{\mathbf{X}}$ stands for some distribution on $\mathbf{Z}$. Classical properties of the Kullback-Leibler divergence ensure that $\mathcal{J}$ has a unique maximum $\log \mathbb{P}(\mathbf{X}; \boldsymbol{\gamma})$ and which is reached for $R_{\mathbf{X}}(\mathbf{Z}) = \mathbb{P}(\mathbf{Z}|\mathbf{X})$. In other words, if $\mathbb{P}(\mathbf{Z}|\mathbf{X}; \boldsymbol{\gamma})$ was tractable, the maximization of $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ would be equivalent to the maximization of $\log \mathbb{P}(\mathbf{X}; \boldsymbol{\gamma})$.

In our case, $\mathbb{P}(\mathbf{Z}|\mathbf{X}; \boldsymbol{\gamma})$ can not be calculated and the maximization of $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$ with respect to $R_{\mathbf{X}}$ can not be achieved without some restrictions on $R_{\mathbf{X}}$. Therefore, we limit our search to the class of completely factorized distributions:

$$R_{\mathbf{X}}(\mathbf{Z}) = \prod_i h(\mathbf{Z}_i, \boldsymbol{\tau}_i) \tag{4}$$

where h denotes the multinomial distribution, $\boldsymbol{\tau}_i$ stands for a vector of probabilities: $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iQ})$ (with $\sum_q \tau_{iq} = 1$). The $\boldsymbol{\tau}_i$s must be thought of as variational parameters to be optimized to fit $\mathbb{P}(\mathbf{Z}|\mathbf{X})$ to $\mathbb{P}(\mathbf{Z}|\mathbf{X}; \boldsymbol{\gamma})$ at best; they depend on the observed data $\mathbf{X}$. Since we set some restrictions on the form of $R_{\mathbf{X}}$, $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$ is a lower bound of $\log \mathbb{P}(\mathbf{X})$.

Proposition 2 *For factorized distributions* (4)*, we have*

$$\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma}) = -\sum_i \sum_q \tau_{iq} \log \tau_{iq} + \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_{i \neq j} \sum_{q,\ell} \tau_{iq} \tau_{j\ell} \log f_{q\ell}(X_{ij}).$$

Proof. As shown in Jaakkola (2000), $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$ can be rewritten as

$$\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma}) = \mathcal{H}(R_{\mathbf{X}}) + \sum_{\mathbf{Z}} R_{\mathbf{X}}(\mathbf{Z}) \log \mathbb{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma}) \tag{5}$$

where $\mathcal{H}(\cdot)$ denotes the entropy of a distribution. For factorized distributions, the entropy is additive over the coordinates so $\mathcal{H}(R_{\mathbf{X}}) = \sum_i \mathcal{H}(h(\cdot, \boldsymbol{\tau}_i)) = -\sum_i \sum_q \tau_{iq} \log \tau_{iq}$. The calculation of second term of (5) directly derives from the combination of Proposition 1:

$$\sum_{\mathbf{Z}} R_{\mathbf{X}}(\mathbf{Z}) \log \mathbb{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma}) = \sum_i \sum_q \mathbb{E}_{R_{\mathbf{X}}}(Z_{iq}) \log \alpha_q + \sum_{i \neq j} \sum_{q,\ell} \mathbb{E}_{R_{\mathbf{X}}}(Z_{iq} Z_{j\ell}) \log f_{q\ell}(X_{ij})$$

where $\mathbb{E}_{R_{\mathbf{X}}}$ denotes the expectation with respect to distribution $R_{\mathbf{X}}$. The result follows using (4). ■

Other form for $R_{\mathbf{X}}$. Considering a broader class of distribution $R_{\mathbf{X}}$ would provide a better lower bound of $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$. Since the second term of (5) only involves $\mathbb{E}_{R_{\mathbf{X}}}(Z_{iq})$ and $\mathbb{E}_{R_{\mathbf{X}}}(Z_{iq} Z_{j\ell})$, it seems natural to consider distributions like

$$R_{\mathbf{X}}(\mathbf{Z}) = \prod_{i<j} h(\mathbf{Z}_i, \mathbf{Z}_j | \boldsymbol{\tau}_{\mathbf{ij}})$$

where $\boldsymbol{\tau}_{ij}$ is a square probability matrix of size Q. Such distributions can account for interactions between pairs of nodes. Unfortunately, for such distributions, $\mathcal{H}(R_{\mathbf{X}})$ has no simple expression and its exact computation has complexity $\mathcal{O}(Q^n)$, so that the benefits are outweighted by the increase of the computational burden.

3.3 Exponential Bounds

First order approximation. A lower bound of the likelihood of the observed edges $\mathbb{P}(\mathbf{X})$ can be derived from a general exponential inequality given in Leisink and Kappen (2001): $\forall x, \mu \quad e^x \geq e^\mu (1 + x - \mu)$. When applied to $x = \log \mathbb{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma})$, it gives for any function $\mu(\cdot)$:

$$\mathbb{P}(\mathbf{X}; \boldsymbol{\gamma}) = \sum_{\mathbf{Z}} e^{\log \mathbb{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma})} \geq \sum_{\mathbf{Z}} e^{\mu(\mathbf{Z})} \left[1 + \log \mathbb{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma}) - \mu(\mathbf{Z})\right]$$

Suppose μ is linear: $\mu(\mathbf{Z}) = \sum_{i,q} \beta_{iq} Z_{iq}$ and denote $Q_\mu(\mathbf{Z}) = \exp[\mu(\mathbf{Z})]/C_\mu$, where C_μ is the normalizing constant such that $Q_\mu(\cdot)$ is a probability distribution function, we get:

$$\mathbb{P}(\mathbf{X}; \boldsymbol{\gamma}) \geq C_\mu (1 - \log C_\mu) + C_\mu \sum_{\mathbf{Z}} Q_\mu(\mathbf{Z}) \log \mathbb{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma}) + \mathcal{H}(Q_\mu) \tag{6}$$

The problem is then to optimize this lower bound with respect to parameters β_{iq} and C. Note that the right hand side of (6) is similar to $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$ given in (5). It can be shown that the optimization of these two quantities are actually equivalent; analytic formula connecting the optimal $(\{\beta_{iq}\}, C_\mu)$ and the optimal $\{\tau_{iq}\}$ can be derived (Mariadassou (2006)).

Third order approximation. Since our strategy is to minimize a lower bound of $\log \mathbb{P}(\mathbf{X})$, we are not interested in second order development of the exponential, which may provide either lower or upper bounds. In Yedidia et al. (2005), such an approximation is used and referred to as Bethe method.
A tighter lower bound may be derived from a third order exponential inequality such as:

$$\forall x, \mu, \nu \quad e^x \geq e^\nu (1 + x - \nu) + e^\mu \left((1 + \nu - \mu) \frac{(x-\nu)^2}{2} + \frac{(x-\nu)^3}{6} \right).$$

Again, this may be applied to $x = \log \mathbb{P}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma})$ for which optimal functions $\mu(\mathbf{Z})$ and $\nu(\mathbf{Z})$ have to be determined. This optimization problem cannot be solved in a general way. It can be achieved if $\mu(\mathbf{Z})$ and $\nu(\mathbf{Z})$ are supposed to be linear and equal up to an additive constant. However, each step of the optimization process has then complexity $\mathcal{O}(n^6 Q^6)$ and the improvement of the bound has the same order of magnitude as the computer numerical precision.

4 Parameter Estimation

4.1 Estimation Algorithm

As explained in Section 3.2, the maximum likelihood estimator of $\boldsymbol{\gamma}$ is

$$\widehat{\boldsymbol{\gamma}}_{ML} = \arg\max_{\boldsymbol{\gamma}} \log \mathbb{P}(\mathbf{X}; \boldsymbol{\gamma}) = \arg\max_{\boldsymbol{\gamma}} \max_{R_\mathbf{X}} \mathcal{J}(R_\mathbf{X}, \boldsymbol{\gamma}).$$

In the variational framework, we restrict the last optimization problem to factorized distributions. The estimate we propose is hence

$$\widehat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \max_{R_\mathbf{X} \text{ factorized}} \mathcal{J}(R_\mathbf{X}, \boldsymbol{\gamma})$$

The simultaneous optimization with respect to both $R_\mathbf{X}$ and $\boldsymbol{\gamma}$ is still too difficult so we adopt the following iterative strategy. Denoting by $R_\mathbf{X}^{(n)}$ and $\boldsymbol{\gamma}^{(n)}$ the estimates after step n, we compute

$$\left\{ \begin{array}{rcl} R_\mathbf{X}^{(n)} & = & \arg \max_{R_\mathbf{X} \text{ factorized}} \mathcal{J}(R_\mathbf{X}, \boldsymbol{\gamma}^{(n)}) \\ \boldsymbol{\gamma}^{(n+1)} & = & \arg\max_{\boldsymbol{\gamma}} \mathcal{J}(R_\mathbf{X}^{(n+1)}, \boldsymbol{\gamma}) \end{array} \right. \tag{7}$$

The two next sections are dedicated to each of these steps.

Initialisation step. The optimization procedure (7) only ensures the convergence toward a local optimum so the choice of the starting point for $\boldsymbol{\gamma}$ or $R_\mathbf{X}$ is crucial to avoid local optima. This choice is difficult but, to our experience, hierarchical clustering seems to be a good strategy to get an initial value for $R_\mathbf{X}$.

4.2 Optimal Approximate Conditional Distribution $R_{\mathbf{X}}$

We consider here the optimization of $\mathcal{J}$ with respect to $R_{\mathbf{X}}$. For a given value of $\boldsymbol{\gamma}$, we denote $\widehat{\boldsymbol{\tau}}$ the variational parameter defining the distribution $\widehat{R}_{\mathbf{X}} = \arg \max_{R_{\mathbf{X}} \text{ factorized}} \mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$.

Proposition 3 *For a given $\boldsymbol{\gamma}$, the optimal variational parameter $\widehat{\boldsymbol{\tau}}$ satisfies*

$$\widehat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{\ell} [f_{q\ell}(X_{ij}) f_{\ell q}(X_{ji})]^{\widehat{\tau}_{j\ell}} .$$

Proof. We maximize $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$ such as given in Proposition 2 subject to the condition that, for all i, the τ_{iq}s must sum to 1. The derivative with respect to τ_{iq} of the quantity to be maximized is

$$-\log \tau_{iq} - 1 + \log \alpha_q + \sum_{j \neq i} \sum_{\ell} \tau_{j\ell} \left[\log f_{q\ell}(X_{ij}) + \log f_{\ell q}(X_{ji})\right] + L_i$$

where L_i denotes the ith Lagrange multiplier. The result follows. ■

We obtain here a fixed point relation that can be related to a mean field approximation (see Jaakkola (2000)). We get $\widehat{\boldsymbol{\tau}}$ simply by iterating this relation until convergence.

Undirected graphs. For an undirected graph, $\widehat{\boldsymbol{\tau}}$ satisfies $\widehat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{\ell} [f_{q\ell}(X_{ij})]^{\widehat{\tau}_{j\ell}}$.

4.3 Parameter Estimates

We now have to maximize $\mathcal{J}$ with respect to $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$ for a given distribution $R_{\mathbf{X}}$.

Proposition 4 *For a given distribution $R_{\mathbf{X}}$ characterized by a variational parameter $\boldsymbol{\tau}$, the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are given by*

$$\widehat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq}, \qquad \widehat{\theta}_{q\ell} = \arg\max_{\theta_{q\ell}} \sum_{i \neq j} \tau_{iq} \tau_{j\ell} \log f(X_{ij}; \theta_{q\ell}).$$

Proof. We maximize $\mathcal{J}(R_{\mathbf{X}}, \boldsymbol{\gamma})$ such as given in Proposition 2 subject to $\sum_q \alpha_q = 1$. ■

Classical distributions. Proposition 4 can be applied to any distribution f. Table 1 gives the parameter estimates for the model listed in Section 2.2. The estimates of the mean parameter for the Poisson ($\lambda_{q\ell}$) and Gaussian ($\mu_{q\ell}$) distributions are the same as the estimate of the probability $\pi_{q\ell}$ in the Bernoulli case. The results displayed in this table are all straightforward. Note that all the estimates are weighted versions of the intuitive ones.

Table 1: Estimates of $\theta_{q\ell}$ for some classical distributions. Notations are defined in Section 2.2. $\kappa_{q\ell}$ stands for $1/\sum_{i\neq j}\tau_{iq}\tau_{j\ell}$. $\mathbf{W}_{q\ell}$ is the diagonal matrix with diagonal term $\tau_{iq}\tau_{j\ell}$. # param. is the number of independent parameters in the case on directed graph, except for the bivariate Gaussian only defined for a non oriented graph.

Distribution	Estimate			# param.
Bernoulli	$\widehat{\pi}_{q\ell}$	$=$	$\kappa_{q\ell}\sum_{i\neq j}\tau_{iq}\tau_{j\ell}X_{ij}$	Q^2
Multinomial	$\widehat{p}^k_{q\ell}$	$=$	$\kappa_{q\ell}\sum_{i\neq j}\tau_{iq}\tau_{j\ell}\mathbb{I}(X_{ij}=k)$	$(m-1)Q^2$
Gaussian	$\widehat{\sigma}^2_{q\ell}$	$=$	$\kappa_{q\ell}\sum_{i\neq j}\tau_{iq}\tau_{j\ell}(X_{ij}-\widehat{\mu}_{q\ell})^2$	Q^2
Bivariate Gaussian	$\widehat{\boldsymbol{\mu}}_{q\ell}$	$=$	$\kappa_{q\ell}\sum_{i\neq j}\tau_{iq}\tau_{j\ell}\mathbf{X}_{ij}$	$Q(Q+1)$
	$\widehat{\boldsymbol{\Sigma}}_{q\ell}$	$=$	$\kappa_{q\ell}\sum_{i\neq j}\tau_{iq}\tau_{j\ell}\left(\mathbf{X}_{ij}-\widehat{\boldsymbol{\mu}}_{q\ell}\right)\left(\mathbf{X}_{ij}-\widehat{\boldsymbol{\mu}}_{q\ell}\right)'$	$\frac{3}{2}Q(Q+1)$
Linear regression	$\widehat{\boldsymbol{\beta}}_{q\ell}$	$=$	$\left(\mathbf{Y}'\mathbf{W}^{-1}_{q\ell}\mathbf{Y}\right)^{-1}\mathbf{Y}'\mathbf{W}^{-1}_{q\ell}\mathbf{X}$	pQ^2
	$\widehat{\sigma}^2_{q\ell}$	$=$	$\kappa_{q\ell}\sum_{i\neq j}\tau_{iq}\tau_{j\ell}\left(X_{ij}-\mathbf{y}'_{ij}\widehat{\beta}_{q\ell}\right)^2$	Q^2

Exponential family. The optimal $\boldsymbol{\theta}$ is not explicit in the general case, but has a simpler form if the distribution f belongs to the exponential family.

Proposition 5 *If the distribution f belongs to the exponential family and if $\boldsymbol{\theta}$ is the natural parameter:*

$$f(x;\boldsymbol{\theta}) = \exp\left[\boldsymbol{\Psi}(x)'\boldsymbol{\theta} - A(\boldsymbol{\theta})\right]$$

the optimal $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \left(\nabla_{\boldsymbol{\theta}}A\right)^{-1}\left[\sum_{i\neq j}\tau_{iq}\tau_{j\ell}\boldsymbol{\Psi}(X_{ij})'\right]$$

Proof. According to Proposition 4, we look for $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}\sum_{i\neq j}\tau_{iq}\tau_{j\ell}\boldsymbol{\Psi}(X_{ij})'\boldsymbol{\theta} - A(\boldsymbol{\theta})$. The derivative of the quantity to be maximized with respect to $\boldsymbol{\theta}$ has to be null:

$$\sum_{i\neq j}\tau_{iq}\tau_{j\ell}\boldsymbol{\Psi}(X_{ij})' - \nabla_{\boldsymbol{\theta}}A(\boldsymbol{\theta}) = \mathbf{0}$$

and the result follows. ■

4.4 Choice of the Number of Groups

In practice the number of classes is unknown and should be estimated. Many criterion, such as BIC or AIC (Burnham and Anderson (1998)), are based on the likelihood of observed data $\mathbb{P}(\mathbf{X})$, which can not be calculated here. Therefore, we propose a Bayesian model selection criterion based in the Integrated Classification Likelihood (ICL) criterion developed by Biernacki et al. (2000). The ICL criterion is an approximation of the complete-data integrated likelihood defined such that:

$$\log \mathbb{P}(\mathbf{X}, \mathbf{Z}|m_Q) = \int \log \mathbb{P}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\gamma}, m_Q) g(\boldsymbol{\gamma}|m_Q) d\boldsymbol{\gamma},$$

where $\log \mathbb{P}(\mathbf{X}, \mathbf{Z}|\boldsymbol{\gamma}, m_Q)$ is the complete-data likelihood of model m_Q with Q classes.

Proposition 6 *For a model m_Q with Q classes where $\boldsymbol{\theta}$ involves P_Q independent parameters, the ICL criterion is :*

$$ICL(m_Q) = \max_{\boldsymbol{\gamma}} \log \mathbb{P}(\mathbf{X}, \widetilde{\mathbf{Z}}|\boldsymbol{\gamma}, m_Q) - \frac{1}{2}\left\{P_Q \log[n(n-1)] - (Q-1)\log(n)\right\}.$$

Proof. The derivation of this criterion is described in Daudin et al. (2007) in the case of a binary non oriented graph. The only difference lies in the number of parameters P_Q which is $Q(Q+1)/2$ for the Bernoulli or Poisson model, but may differ in the general case. ■

Note that the penalty term $-\frac{1}{2}\left\{P_Q \log[n(n-1)] - (Q-1)\log(n)\right\}$ is similar to the one of BIC, where the log term refers to number of data. In the case of graphs, the number of data is n (i.e. the number of nodes) for the vector of proportions $\boldsymbol{\alpha}$ ($Q-1$ independent parameters), whereas it is $n(n-1)$ (i.e. the number of edges) for parameter $\boldsymbol{\theta}$.

5 Simulation Study

5.1 Quality of the estimates

Simulation parameters. We considered undirected networks of size $n = 100$ and 500 with $Q = 3$ classes. To study balanced and unbalanced proportions, we set $\alpha_q \propto a^q$, with $a = 1,\ 0.5,\ 0.2$. $a = 1$ gives uniform proportions, while $a = 0.2$ gives very unbalanced proportions: $\boldsymbol{\alpha} = (80.6\%,\ 16.1\%,\ 3.3\%)$. We finally considered symmetric connection intensities λ_{pq}, setting $\lambda_{pp} = \lambda'$ for all p and $\lambda_{pq} = \lambda'\gamma$ for $p \neq q$. Parameter γ controls the difference between within class and between class connection intensities ($\gamma = 0.1,\ 0.5,\ 0.9,\ 1.5$) while λ' is set so that the mean connection intensity λ ($\lambda = 2, 5$) depends neither on γ nor a. γ close to one makes the distinction between the classes difficult. γ larger than one makes the within class connectivities less intense than the between ones. We expect the fitting to be rather easy for the combination $\{n = 500, a = 1, \lambda = 5, \gamma = 0.1\}$ and rather difficult for $\{n = 100, a = 0.2, \lambda = 2, \gamma = 0.9\}$.

Simulations and Computations. For each combination of the parameters, we simulated $S = 100$ random graphs according to the corresponding mixture model. We fitted the parameters using the algorithm described in Section 4. To solve the identifiability problem of the classes, we systematically ordered them in descending estimated proportion order: $\widehat{\alpha}_1 \geq \widehat{\alpha}_2 \geq \widehat{\alpha}_3$. For each parameter, we calculated the estimated Root Mean Squared Error (RMSE):

$$RMSE(\widehat{\alpha}_p) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\alpha_p}^{(s)} - \alpha_p\right)^2}, \qquad RMSE(\widehat{\lambda}_{pq}) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\lambda_{pq}}^{(s)} - \lambda_{pq}\right)^2},$$

where the superscript (s) labels the estimates obtained in simulation s. We also calculated the mean posterior entropy

$$H = \frac{1}{S}\sum_{s}\left(-\sum_{i}\sum_{q}\tau_{iq}^{(s)}\ln\tau_{iq}^{(s)}\right),$$

which gives us the degree of uncertainty of the classification.

Results. Figure 1 (resp. 2) gives the RMSE for the proportion α_q (resp. connection intensities λ_{pq}). As expected, the $RMSE$ is lower when n is larger. The parameters affecting the $RMSE$ are mainly a and γ, whereas λ has nearly no effect. The departures observed for α_1 and α_3 in the balanced case ($a = 1.0$) are due to the systematic reordering of the proportions.

Since the graph is undirected, $\lambda_{pq} = \lambda_{qp}$, so only non redundant parameters are considered in Figure 2. The overall quality of the estimates is satisfying, especially for the diagonal terms λ_{qq}. The within intensity parameter of the smallest class λ_{33} is the most difficult to estimate. The worst case corresponds to a small graph ($n = 100$) with very unbalanced classes ($a = 0.2$) for parameter λ_{12}. In this case, the algorithm is unable to distinguish the two larger classes (1 and 2), so that the estimates extra-diagonal term $\widehat{\lambda}_{12}$ is close to the diagonal ones $\widehat{\lambda}_{11}$ and $\widehat{\lambda}_{22}$, whereas its true value is up to ten times smaller.

Figure 3 gives the mean entropy. Not surprisingly, the most influent parameter is γ: when γ is close to 1, the classes are almost indistinguishable. For small graphs ($n = 100$), the mean intensity λ has almost no effect. Because of the identifiability problem already mentioned, we did not consider the classification error rate.

5.2 Model Selection

We considered an undirected graph of size $n = 50, 100, 500$ and 1000 with $Q^\star = 3$ classes. We considered the combination $\{a = 0.5, \lambda = 2, \gamma = 0.5\}$ which turned out to be a medium case (see Section 5.1) and computed ICL for Q ranging from 1 to 10 (from 1 to 5 for $n = 1000$) before selecting the Q maximizing ICL. We repeated this for $S = 100$ simulations.

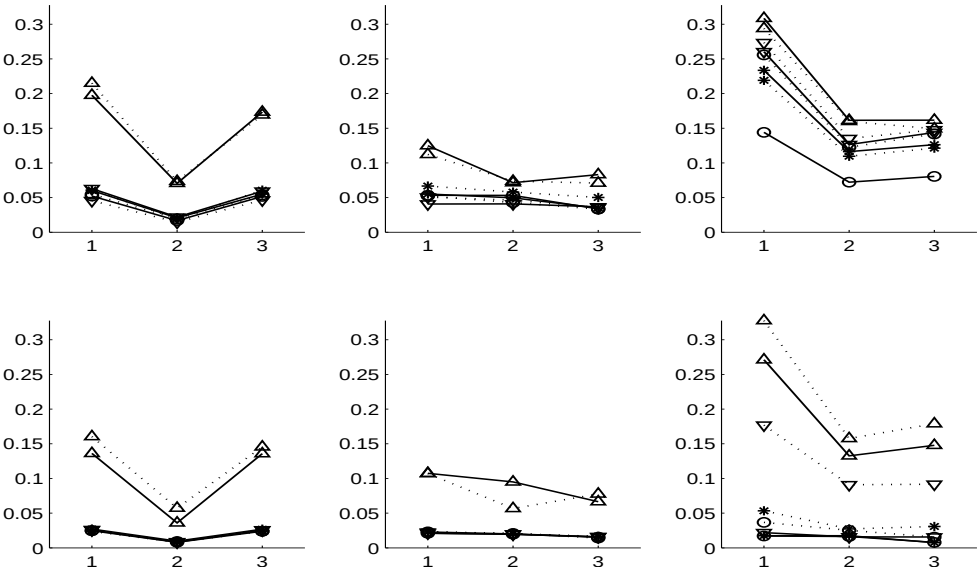

Figure 1: *RMSE* of the estimates $\widehat{\alpha}_q$. The x-axis refers to $\alpha_1, \alpha_2, \alpha_3$. Top: $n = 100$, bottom: $n = 500$, From left to right: $a = 1, 0.5, 0.2$. Solid line: $\lambda = 5$, dashed line: $\lambda = 2$. Symbols depend on γ: $\circ = 0.1$, $\nabla = 0.5$, $\triangle = 0.9$, $* = 1.5$.

Table 2: Frequency (in %) at which Q is selected for various sizes n.

	n			
Q	50	100	500	1000
2	82	7	0	0
3	17	90	100	100
4	1	3	0	0

Figure 4 gives ICL as a function of Q, while Table 2 returns the frequency with which each Q is selected. As soon as n is larger than 100, ICL almost always selects the correct number of classes; For smaller graphs ($n = 50$), it tends to underestimate it. The proposed criterion is thus highly efficient.

References

Albert, R. and A. L. Barabási (2002). Statistical mechanics of complex networks. *R. Modern Physics 74*(1), 47–97.

Barabási, A. L. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*, 509–512.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model

for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel. 22*(7), 719–725.

Burnham, K. P. and R. A. Anderson (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Wiley: New-York.

Daudin, J.-J., F. Picard, and S. Robin (2007). A mixture model for random graphs. *under review*.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B 39*, 1–38.

Getoor, L. and C. P. Diehl (2004). Link mining: A survey. *SIGKDD Explor. 7*(2), 3–12.

Girvan, M. and M. E. J. Newman (2002). Community strucutre in social and biological networks. *Proc. Natl. Acad. Sci. USA 99*(12), 7821–6.

Govaert, G. and M. Nadif (2005). An EM algorithm for the block mixture model. *IEEE Trans. Pattern Anal. Machine Intel. 27*(4), 643–7.

Jaakkola, T. (2000). *Advanced mean field methods: theory and practice*, Chapter Tutorial on variational approximation methods. MIT Press.

Jordan, M. I., Z. Ghahramani, T. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning 37*(2), 183–233.

Leisink, M. A. R. and H. J. Kappen (2001). A tighter bound for graphical models. *Neural Computation 13*(9), 2149–2171.

Mariadassou, M. (2006). Estimation paramétrique dans le modèle ERMG. Master's thesis, Université Paris XI / Ecole Nationale Supérieure.

McLahan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley.

Newman, M. E. J. (2004a). Analysis of weighted networks. *Phys. Rev. E*(70), 056131.

Newman, M. E. J. (2004b). Fast algorithm for detecting community structure in networks. *Phys. Rev. E*(69), 066133.

Newman, M. E. J., D. J. Watts, and S. H. Strogatz (2002). Random graph models of social networks. *PNAS 99*, 2566–2572.

Nowicki, K. and T. Snijders (2001). Estimation and prediction for stochastic block-structures. *J. Amer. Statist. Assoc. 96*, 1077–87.

Pattison, P. E. and G. L. Robins (2007). *Handbook of Probability Theory with Applications*, Chapter Probabilistic Network Theory. Sage Publication.

Picard, F., J.-J. Daudin, V. Miele, M. Mariadassou, and S. Robin (2007). A novel framework for random graph models with heterogeneous connectivity structure. *submitted*.

von Luxburg, U., M. Belkin, and O. Bousquet (2007). Consistency of spectral clustering. *AS. to appear*, `www.cse.ohio-state.edu/~mbelkin/TRspectral.pdf`.

Yedidia, J., W. Freeman, and Y. Weiss (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Information Theory 15*(7), 2282–312.

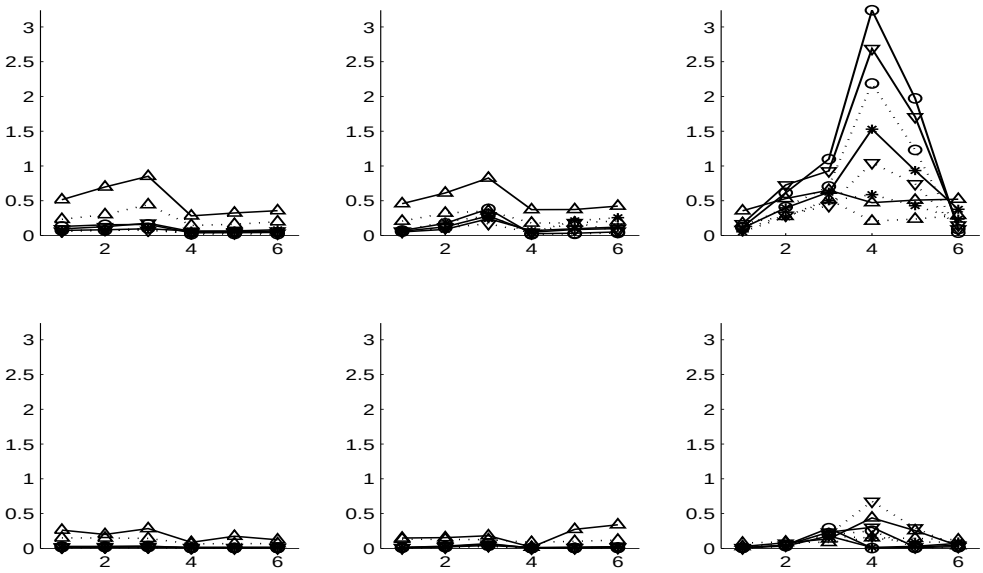

Figure 2: $RMSE$ of the estimates $\widehat{\lambda}_{pq}$. The x-axis refers to $\lambda_{11}, \lambda_{22}, \lambda_{33}, \lambda_{12}, \lambda_{13}, \lambda_{23}$. Same legend as Figure 1.

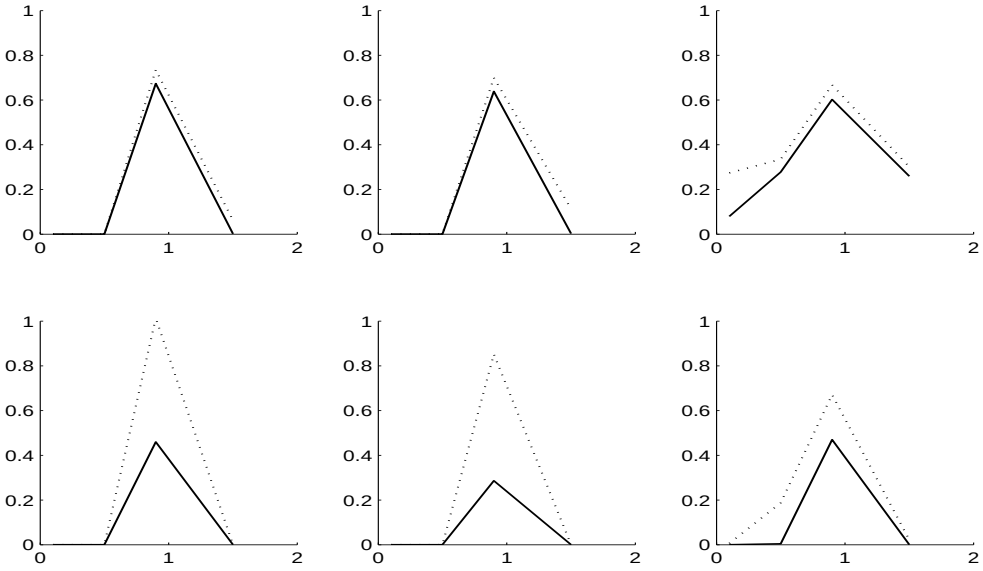

Figure 3: Mean (normalized) entropy H/n as a function of γ. Top: $n = 100$, bottom: $n = 500$, From left to right: $a = 1, 0.5, 0.2$. Solid line: $\lambda = 5$, dashed line: $\lambda = 2$.

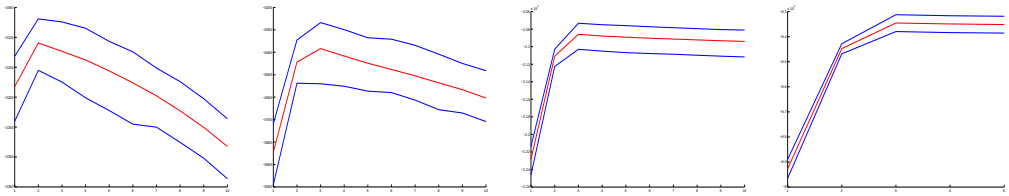

Figure 4: Mean ICL and 90% confidence interval as a function of Q. From left to right: $n = 50, n = 100, n = 500, n = 1000$.