UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Community detection by stochastic block modelling

Auteur : Enrico Laddomada
Promoteur : Eugen Pircalabelu
Lecteur : Johan Segers
Année académique 2019-2020
Master en statistiques, orientation générale, finalité specialisée

Acknowledgment.

This thesis is dedicated to all the people that have helped me during the path to this thesis, in particular to my family as well as to all my friends.

Introduction

Stochastic Block Models (SBM) are models for analyzing data concerning networks. They have been broadly used in the last decade in various fields to analyze, for example, social networks (Bickel & Chen, 2009; Choi et al., 2012; Karrer & Newman, 2011; Nowicki & Snijders, 1997, 2001), financial networks (Bickel & Chen, 2009), political blogs networks (Karrer & Newman, 2011; Latouche et al., 2011), biological host-parasite networks (Mariadassou et al., 2010), genes networks (Latouche et al., 2011), telephone communication networks (Bickel & Chen, 2009), scientific articles citations networks (Barbillon et al., 2017) and books citations networks (Decelle et al., 2011).

Networks have become extremely important in the scientific literature, and their properties were widely studied and applied in different fields: for example, assortativity (Bickel & Chen, 2009) describes networks in which the density of connections between elements of the same group is higher than the density of connections across groups. As opposed to assortativity, disassortativity describes networks in which the density of connections between elements (or **nodes**) of the same group is lower than the density of connections across groups. Stochastic block models allow modelling networks displaying both assortativity and disassortativity, and complex patterns of connections between groups of nodes, ie. *communities*.

One of the first well-known probabilistic models for networks was the Erdös-Rényi graph model (Bondy & Murty, 2008; Choi et al., 2012; Daudin et al., 2008). This model has the downside of supposing that all nodes are statistically equivalent in terms of their **edges** (ie., their connections) distribution. The stochastic block model is in fact a generalisation of the Erdös-Rényi model, in which nodes can belong to different classes, with classes possibly yielding different statistical properties in terms of edge distribution. Therefore, the network structure, ie. the edges, can provide information on the classes to which nodes belong.

The main aim of stochastic block modelling is to classify nodes of a network as to their statistical properties in terms of connections. The "connection" can be any quantity measuring some kind of interaction between two nodes, for example the number/duration of the phone calls between two people, or the weblinks between two websites. The problem of classifying nodes of a network is also known as *community detection* or *clustering*.

For example, in order to analyze a network of political blogs, and split blogs according to their right-wing or left-wing leaning, it may not be necessary to have access to the content of the blog itself, but only to the list of weblinks between these blogs. Even without knowing the content of the blogs, we may still use the weblinks network to split the blogs in classes

corresponding roughly to political leaning. In the easiest, "totally-polarized" scenario, blogs with left-wing leaning only have weblinks to other blogs with left-wing leaning, and blogs with right-wing leaning only to other blogs with right-wing leaning. This scenario in which nodes are only connected to nodes of the same class is the easiest possible scenario. However, even in less polarized scenarios, information on the political leaning of the blogs can be found by exploiting the statistical (or *stochastic/non-deterministic*) properties of the classes of the network, and that is indeed the aim of stochastic block modelling.

The idea that nodes of a network can be classified according to their different connection properties is not new, and many other approaches for detecting communities exist. For example, nodes can be classified by the number of connections they display, ie. by their *degree*; however, only the degree may not be sufficient to classify correctly the nodes. Stochastic block modelling can classify nodes according to more sophisticated criteria. Figure 0.1 shows the result of a simulation from a stochastic block model, in which all classes have on average the same degree, however a clear structure of 4 classes appears, each class displaying different type of connections. Class 1 is slightly connected with class 2 and 4, but not at all with class 3; class 2 is slightly connected to class 1 and 3, but not to class 4, etc. Figure 0.2 shows a different way of representing nodes of a network, making evident very clear "blocks".

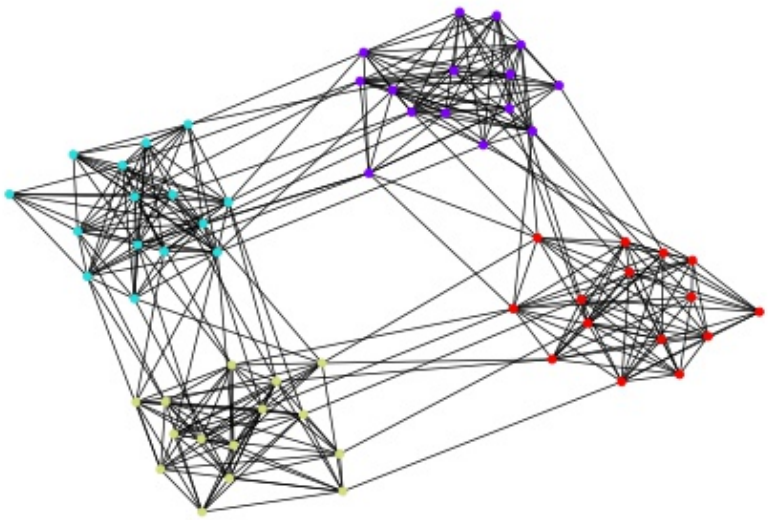


Figure 0.1: Result of a simulation of a network with 60 nodes and 4 classes (1-purple, 2-sky blue, 3-yellow, 4-red), using a binary undirected stochastic block model. For the simulation, the class weights, noted π_k for classes $k = 1, 2, 3$ and 4, were set to $\pi_k = \frac{1}{4}$, meaning uniform probability that a generated node belongs to any of the 4 classes. The probabilities of generating an edge between a node of class k and a node of class l, noted η_{kl}, were set to $\eta_{kk} = 0.7$ for $k = 1, 2, 3, 4$, and $\eta_{12} = \eta_{23} = \eta_{34} = \eta_{41} = 0.05$. Other edge-probabilities were set to 0.

Indeed, in most of the situations in which stochastic block modelling is used, classes are not directly observed, and can only be estimated from observing the network (Allman et al., 2009, 2011; Latouche et al., 2011; Nowicki & Snijders, 1997, 2001). This kind of scenarios in which the sample is incomplete has been broadly studied in the statistics literature for a long time,

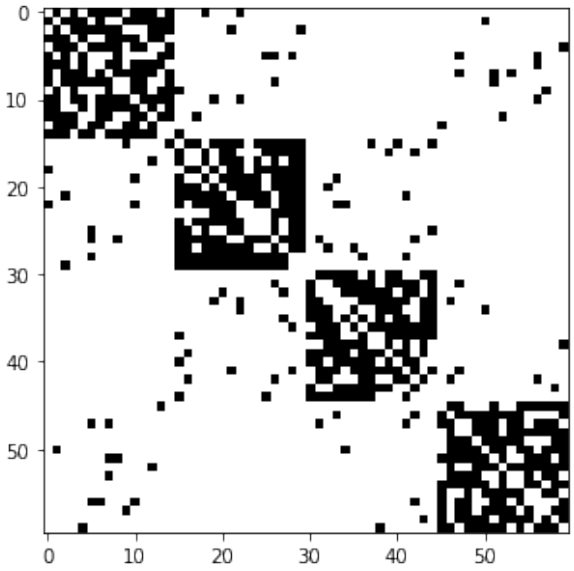


Figure 0.2: Graphical representation of the adjacency matrix of the same simulated network as in Figure 0.1, with black dots indicating the presence of an edge between two nodes. The block structure is evident.

because in such scenarios, classical methods of estimation such as the maximum-likelihood method are not directly applicable. However, the **Expectation-Maximization** algorithm (Barbillon et al., 2017; Bickel & Chen, 2009; Celisse et al., 2012; Dempster et al., 1977; Latouche et al., 2011; Mariadassou et al., 2010; Wu, 1983) introduced in the work of Dempster et al. (1977), yields a general iterative procedure converging to the maximum-likelihood estimator in many incomplete-data models. For the stochastic block models, the explicit formulas for performing this algorithm were found only for the very specific 2-class binary stochastic block model by Nowicki & Snijders (1997), but do not exist for more general stochastic block models. Still, **variational** methods are often used instead (Barbillon et al., 2017; Bickel & Chen, 2009; Celisse et al., 2012; Latouche et al., 2011; Mariadassou et al., 2010), as they yield an algorithm approximating well the Expectation-Maximization algorithm for general stochastic block models. Other methods for estimating communities that will be studied in this thesis include **profile likelihood** methods (Bickel et al., 2013; Choi et al., 2012; Karrer & Newman, 2011; Zhao et al., 2012), **Bayesian** methods (Nowicki & Snijders, 1997, 2001; Latouche et al., 2009), and **spectral** methods (Abbe, 2018; Amini et al., 2013; Lei & Rinaldo, 2015; Rohe et al., 2011). Other estimation methods based on stochastic block modelling were also developed in the literature, but will not be specifically studied in this thesis. These include belief-propagation methods (Decelle et al., 2011), moments-based methods (Bickel et al., 2013), pseudo-likelihood methods (Amini et al., 2013) and semi-definite programming methods (Amini & Levina, 2018).

Also, several extensions of the standard stochastic block model were developed in the literature. In this thesis, the degree-corrected SBM (Zhao et al., 2012; Karrer & Newman, 2011; Amini et al., 2013) will be studied. This model is much more realistic than the standard block model because it allows the presence of "hubs", ie. nodes having a much higher number of

connections with respect to the other nodes of the network. Other extensions exist, like the overlapping SBM (Latouche et al., 2011) allowing classes to overlap each other; the multiplex SBM (Barbillon et al., 2017) allowing to consider any finite number of binary networks; and the discrete relational data model (Nowicki & Snijders, 2001). These extensions will not be specifically studied in this thesis.

This thesis aims to evaluate if stochastic block models are suitable to detect communities in networks. In Chapter 1, stochastic block models will be formally defined from the theoretical point of view and some of their properties will be studied. In Chapter 2, focus will be given on identifiability properties of the stochastic block models. In Chapter 3, several methods for estimating communities will be explained: these are the expectation-maximization method, the variational expectation-maximisation method, the spectral methods, one Bayesian method based on Gibbs sampling, the profile-likelihood method and the modularity method. In Chapter 4, results of simulation studies will be presented in order to evaluate the performance of the variational method and of the spectral methods. In Chapter 5, three real networks will be analyzed by means of the variational method and of the spectral methods.

Contents

Chapter 1: The stochastic block models: definitions and properties

This chapter will start with the definition of the famous Erdös-Rényi model (Bondy & Murty, 2008; Choi et al., 2012; Daudin et al., 2008). Then, the binary stochastic block model (Allman et al., 2009, 2011; Choi et al., 2012; Daudin et al., 2008; Latouche et al., 2011; Nowicki & Snijders, 1997, 2001) will be defined and some of its properties will be examined in Section 1.2. This model can be thought of as the simplest stochastic block model. Indeed, in the literature, this same model is often called, simply, the stochastic block model, without further specification. However, in this thesis, a more general class of models, as introduced in the work of Mariadassou et al. (2010), will be defined. This class will be defined in Section 1.4. This more general definition will then allow us to define the Poisson block model (Karrer & Newman, 2011; Mariadassou et al., 2010; Zhao et al., 2012) (presented in Section 1.5) and the Gaussian block model (Mariadassou et al., 2010) (presented in Section 1.6). Two particular sub-models, the affiliation model (Abbe, 2018; Allman et al., 2011; Celisse et al., 2012; Decelle et al., 2011; Zanghi et al., 2008) and the symmetric model (Abbe, 2018) will be defined in Section 1.7. Some asymptotic properties of the block models will be discussed in Section 1.8. An extension of the stochastic block model, the degree-corrected stochastic block model (Amini et al., 2013; Karrer & Newman, 2011; Lei & Rinaldo, 2015; Zhang et al., 2014), will be defined in Section 1.9. Note that Section 1.3 is dedicated to a standard result concerning the connected components decomposition of graphs; readers familiar with graph theory may jump this section.

1.1 A preliminary model: the Erdös-Rényi model

Let $n \in \mathbb{N}$ denote the number of nodes in the graph. Nodes will be indexed by $\{1, 2, \dots, n\}$. We will also use the notation $[n]$ to indicate more shortly the index set $\{1, 2, \dots, n\}$; and the notation $[n]^2$ to denote the cartesian product $\{1, 2, \dots, n\} \times \{1, 2, \dots, n\}$ representing the set of all (ordered) pairs of nodes.

A (random) *graph adjacency matrix* matrix of dimension $n \times n$, denoted by $\boldsymbol{Y} = \{Y_{ij}\}_{i,j=1\dots n}$ is a binary matrix such that $Y_{ij} = 1$ means that there is an edge going from node i to node j, while $Y_{ij} = 0$ means that there is no such edge. Let $\mathcal{Y}$ denote the set of all possible binary matrices of dimension $n \times n$:

$$\mathcal{Y} := \mathbb{M}_{n\times n}(\{0,1\}) = \{n \times n \text{ matrices with entries in } \{0,1\}\}.$$

Two slightly different definitions of the Erdös-Rényi model exist, one for the **directed** model and one for the **undirected** model. The first definition will next be given, followed by the explanation of the minor differences between the two.

A **directed** Erdös-Rényi model (Bondy & Murty, 2008; Choi et al., 2012; Daudin et al., 2008) with n nodes and *edge-probability* parameter $\eta \in [0,1]$, denoted by $ER(n,\eta)$, is a generative model for the random matrix $\boldsymbol{Y}$, such that the following three hypotheses hereunder are valid. Remark that an instance of $\boldsymbol{Y}$ will be denoted by $\boldsymbol{y}$, with $\boldsymbol{y} = \{y_{ij}\}_{i,j=1\dots n}$.

The first hypothesis of the model is that the edge variables $\{Y_{ij}\}_{i,j=1\dots n}$ are mutually independent:

$$\{Y_{ij}\}_{i,j=1\dots n} \text{ are mutually independent.} \tag{1.1}$$

The second hypothesis of the model is that each edge variable $Y_{ij}, i \neq j$ follows a Bernoulli distribution of parameter η, whose probability mass function (p.m.f.) is denoted by $\mathcal{B}(\cdot|\eta)$:

$$Y_{ij} \sim \mathcal{B}(\cdot|\eta) \qquad \forall i,j = 1,...,n \;\; i \neq j, \tag{1.2}$$

therefore the event $\{Y_{ij} = 1\}$, indicating that an edge from node i to node j exists, happens with probability η.
Finally, the third and last hypothesis of the Erdös-Rényi model is that there are no self-edges:

$$Y_{ii} := 0 \qquad \forall i = 1,...,n. \tag{1.3}$$

However, this last hypothesis is often relaxed; if such is the case, the self-edge variables Y_{ii} are supposed to follow the same rules in (1.2) as other edge variables do.

The previous definition is useful in a scenario where for each couple of nodes $i \neq j$, two independent edge variables Y_{ij} and Y_{ji} are observed; therefore, this model is called the **directed** Erdös-Rényi model.
The **undirected** Erdös-Rényi model is used in scenarios where the relationship between each couple of nodes is symmetric. To define the undirected model, equations (1.1), (1.2) and (1.3) remain valid, and one simply adds the condition that $\boldsymbol{Y}$ is symmetric and relaxes the mutual independence condition (1.1) to $i \leq j$.

Equations (1.1), (1.2) and (1.3) imply that the directed Erdös-Rényi model is characterized by its probability mass function $f(\cdot) : \mathcal{Y} \to \mathbb{R}$, defined by:

$$\begin{aligned} f(\boldsymbol{y}) &:= \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \mathcal{B}(y_{ij}|\eta) \\ &= \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \eta^{y_{ij}}(1-\eta)^{1-y_{ij}} \quad \forall \boldsymbol{y} \in \mathcal{Y}. \end{aligned} \tag{1.4}$$

For the undirected Erdös-Rényi model, $f(\cdot)$ is given by:

$$\begin{aligned} f(\boldsymbol{y}) &:= \prod_{\substack{(i,j)\in[n]^2 \\ i<j}} \mathcal{B}(y_{ij}|\eta) \\ &= \prod_{\substack{(i,j)\in[n]^2 \\ i<j}} \eta^{y_{ij}}(1-\eta)^{1-y_{ij}} \quad \forall \boldsymbol{y} \in \mathcal{Y}. \end{aligned}$$

Note that the only difference between the two formulas is that the relation "$i \neq j$" for the directed model is replaced with "$i < j$" for the undirected model.

1.2 The binary stochastic block model

The binary stochastic block model (Allman et al., 2009, 2011; Choi et al., 2012; Daudin et al., 2008; Latouche et al., 2011; Nowicki & Snijders, 1997, 2001) generalises the Erdös-Rényi model, by allowing the presence of different classes of nodes.

Let $K \in \mathbb{N}$ denote the number of classes to which nodes belong. Classes will be labelled by $\{1, 2, ..., K\}$. The set of the labels, $\{1, 2, \dots, K\}$, will also be denoted by $[K]$. The notation $[K]^2$ will denote the cartesian product $\{1, 2, \dots, K\} \times \{1, 2, \dots, K\}$. We remind the reader that $\boldsymbol{Y}$ is the (random) *graph adjacency matrix*, ie. a binary matrix such that $Y_{ij} = 1$ means that there is an edge going from node i to node j, while $Y_{ij} = 0$ means that there is no such edge and $\mathcal{Y}$ denotes the set of all such matrices.

A (random) *class-membership vector* of length n, denoted by $\boldsymbol{Z} = \{Z_i\}_{i=1,\dots n}$, is a vector such that $Z_i = k$ means that node i belongs to class k. Let $\mathcal{Z} := [K]^n$ denote the set of all possible class-membership vectors of length n, ie. the cartesian product $\underbrace{\{1, 2, \dots, K\} \times \dots \times \{1, 2, \dots, K\}}_{n \text{ times}}$.

Let also Π_K denote the K-coordinates probability simplex, ie. the set of all vectors of length K whose coordinates are non-negative and sum exactly to 1:

$$\Pi_K := \{\boldsymbol{\pi} \in \mathbb{R}^K | \pi_k \geq 0 \quad \forall k = 1...K, \ \sum_{k=1}^{K} \pi_k = 1\}.$$

This represents the parameter space of a discrete probability distribution having K different outputs; or, in our case, K different classes of nodes. The parameter $\boldsymbol{\pi}$ will be used to model the "weight" we give to the K classes. For example, in a model with 2 classes having equal weight, we may have $\boldsymbol{\pi} = (0.5, 0.5)^T$, while if the first class has more weight than the second, we may have, for example, $\boldsymbol{\pi} = (0.8, 0.2)^T$. Exponent $(\cdot)^T$ denotes the transpose operation. In general, we will use in this thesis the convention that all vectors are column vectors.

Finally, let $[0, 1]^{K\times K}$ denote the set of all $K \times K$ square matrices $\boldsymbol{\eta}$ such that all scalar elements of $\boldsymbol{\eta}$, denoted by η_{kl} $k, l \in \{1, 2, ..., K\}$, are in the interval $[0, 1]$:

$$[0, 1]^{K\times K} := \{\boldsymbol{\eta} \in \mathbb{R}^{K\times K} | \ \ \eta_{kl} \in [0, 1] \quad \forall k, l \in \{1, 2, ..., K\}\},$$

where $\mathbb{R}^{K\times K}$ denotes the set of all squared $K \times K$ real matrices. $[0,1]^{K\times K}$ represents then the parameter space of the edge probabilities between classes k and l, where $k, l \in \{1, 2, ..., K\}$, ranging in the interval $[0, 1]$.

As for the Erdös-Rényi model, two slightly different definitions of the binary stochastic block model exist, one for the **directed** model and one for the **undirected** model. The first definition will next be given, followed by the explanation of the minor differences between the two.

1.2.1 Definition of the binary model

A **directed** binary stochastic block model (Latouche et al., 2011; Nowicki & Snijders, 2001) with n nodes, K classes and parameters $(\boldsymbol{\eta}, \boldsymbol{\pi}) \in [0,1]^{K\times K} \times \Pi_K$, denoted by $SBM(n, K, \boldsymbol{\eta}, \boldsymbol{\pi})$, where $\boldsymbol{\pi}$ is the vector containing the *class-weights*, or *class-probabilities* parameters π_k, and $\boldsymbol{\eta}$ is the matrix containing the *edge-probabilities* parameters $\eta_{kl} \in [0, 1]$, is a generative model for the couple of random arrays $(\boldsymbol{Y}, \boldsymbol{Z})$, such that the four hypotheses hereunder are valid. Remark that an instance of $(\boldsymbol{Y}, \boldsymbol{Z})$ will be denoted by $(\boldsymbol{y}, \boldsymbol{z})$, with $\boldsymbol{y} = \{y_{ij}\}_{i,j=1...n}$ and $\boldsymbol{z} = \{z_i\}_{i=1...n}$. Furthermore, the couple $(\boldsymbol{\pi}, \boldsymbol{\eta})$ will be also referred to as the *parametric couple* of the model. This parametric couple shall be considered as a single multidimensional parameter of dimension $K^2 + K$.

The first hypothesis of the directed binary stochastic block model is that the class-membership variables $\{Z_i\}_{i=1...n}$ are mutually independent and identically distributed one-trial multinomial variables with parameter $\boldsymbol{\pi}$, whose probability mass function is denoted by $Mult(\cdot|\boldsymbol{\eta})$:

$$\{Z_i\}_{i=1...n} \overset{iid}{\sim} Mult(\cdot|\boldsymbol{\pi}).$$

This is equivalent to supposing that the probability that $\boldsymbol{Z} = \boldsymbol{z}$ is defined by:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}) &= \mathbb{P}(Z_1 = z_1, ..., Z_n = z_n) \\
&= \prod_{i\in[n]} \mathbb{P}(Z_i = z_i) \\
&= \prod_{i\in[n]} Mult(z_i|\boldsymbol{\pi}) \\
&= \prod_{i\in[n]} \prod_{k\in[K]} \pi_k^{\mathbb{1}_{\{z_i=k\}}} \quad \forall \boldsymbol{z} \in \mathcal{Z},
\end{aligned} \tag{1.5}$$

where $\mathbb{1}_{\{z_i=k\}}$ denotes the indicator function defined by:

$$\mathbb{1}_{\{z_i=k\}} := \begin{cases} 1 & \text{if } z_i = k \\ 0 & \text{if } z_i \neq k, \end{cases}$$

taking value 1 if node i belongs to class k and 0 otherwise. The probability that a node belongs to class k is then given by π_k.
The second hypothesis of the model is that, conditionally on $\boldsymbol{Z}$, the edge variables $\{Y_{ij}\}_{i,j=1...n}$ are mutually independent:

$$\{Y_{ij}\}_{i,j=1...n} \text{ are mutually independent conditionally on } \boldsymbol{Z}. \tag{1.6}$$

The third hypothesis of the model is that, conditionally on the classes $\boldsymbol{Z}$, each edge variable $Y_{ij}, i \neq j$ follows a Bernoulli distribution of parameter η depending on the classes to which nodes i and j belong, ie.:

$$Y_{ij}|Z_i = k, Z_j = l \sim \mathcal{B}(\cdot|\eta_{kl}) \qquad \forall i, j = 1, ..., n \;\; i \neq j; \tag{1.7}$$

therefore the event $\{Y_{ij} = 1\}$, indicating that an edge from node i to node j exists, happens with probability η_{kl}, conditional that node i belongs to class k and node j belongs to class l. Finally, the fourth hypothesis of the model is that:

$$Y_{ii} := 0 \qquad \forall i = 1, ..., n, \tag{1.8}$$

ie. there are no self-edges. However, this last hypothesis may in the following text be relaxed for theoretical derivations; then the self-edge variables Y_{ii} are supposed to follow the same rules in (1.7) as other edge-variables do. If such is the case, it will be explicitly mentioned in the text.

To be precise from a theoretical point of view, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are measurable functions on a probability space $(\Omega, \mathcal{A}, \{\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\eta}}\}_{(\boldsymbol{\pi},\boldsymbol{\eta})\in\Pi_K\times[0,1]^{K\times K}})$, with: $\boldsymbol{Y} : (\Omega, \mathcal{A}) \mapsto (\mathcal{Y}, \mathscr{P}(\mathcal{Y}))$ and $\boldsymbol{Z} : (\Omega, \mathcal{A}) \mapsto (\mathcal{Z}, \mathscr{P}(\mathcal{Z}))$, where $\mathscr{P}(\mathcal{Y})$ denotes the powerset of $\mathcal{Y}$ and $\mathscr{P}(\mathcal{Z})$ the powerset of $\mathcal{Z}$.

The directed binary model depends on $K^2 + K - 1$ independent scalar parameters: K^2 independent edge-probability parameters $\{\eta_{kl}\}_{k,l\in\{1,...,K\}}$ in $\boldsymbol{\eta}$ and $K-1$ independent class-weight parameters in $\boldsymbol{\pi}$, due to the constraint $\sum_{k=1}^{K} \pi_k = 1$.

To define the undirected binary model (Allman et al., 2009, 2011; Choi et al., 2012; Daudin et al., 2008; Nowicki & Snijders, 1997), hypotheses (1.5), (1.6), (1.7) and (1.8) remain the same, except for the fact that we relax the mutual independence condition (1.6) to $i \leq j$, and we also add the conditions that $\boldsymbol{Y}$ and $\boldsymbol{\eta}$ are symmetric. The undirected binary model depends on $\frac{K(K+1)}{2} + K - 1$ independent scalar parameters: $\frac{K(K+1)}{2}$ independent parameters $\{\eta_{kl}\}_{k,l\in\{1,...,K\},k\leq l}$ in $\boldsymbol{\eta}$, due to its symmetry constraint, and $K-1$ independent parameters in $\boldsymbol{\pi}$, due to the constraint $\sum_{k=1}^{K} \pi_k = 1$. An undirected stochastic block model will also be denoted by $SBM(n, K, \boldsymbol{\pi}, \boldsymbol{\eta})$; the context will make clear if the directed or the undirected model is considered.

We also let z_{ik} denote the indicator function $\mathbb{1}_{\{z_i=k\}}$:

$$z_{ik} := \mathbb{1}_{\{z_i=k\}} = \begin{cases} 1 & \text{if } z_i = k \\ 0 & \text{if } z_i \neq k \end{cases} \qquad \forall i = 1, ..., n; \quad \forall k = 1, ..., K,$$

indicating 1 if node i belongs to class k and 0 otherwise. This shorter notation will enhance the readability of the formulas throughout this thesis.

Hypotheses (1.5), (1.6), (1.7) and (1.8) imply that the binary stochastic block model is characterized by its probability mass function $f_{\boldsymbol{Y},\boldsymbol{Z}}(\cdot) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$:

$$f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{z}) := f_{\boldsymbol{Z}}(\boldsymbol{z}) f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) \quad \forall (\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{Y} \times \mathcal{Z},$$

where $f_{\boldsymbol{Z}}(\cdot)$ is the marginal probability mass function of $\boldsymbol{Z}$ and $f_{\boldsymbol{Y}|\boldsymbol{Z}}(\cdot|\cdot)$ the probability mass function of $\boldsymbol{Y}$ conditional on $\boldsymbol{Z}$, both functions will next be defined.

For both the directed and the undirected binary model, $f_{\boldsymbol{Z}}(\cdot)$ is defined by:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) := \prod_{i\in[n]} Mult(z_i|\boldsymbol{\pi}) = \prod_{i\in[n]}\prod_{k\in[K]} \pi_k^{z_{ik}}. \tag{1.9}$$

For the directed binary model, $f_{\boldsymbol{Y}|\boldsymbol{Z}}(\cdot|\cdot)$ is defined by:

$$\begin{aligned} f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) &:= \prod_{\substack{(i,j)\in[n]^2\\ i\neq j}} \mathcal{B}(y_{ij}|\eta_{z_i z_j}) \\ &= \prod_{\substack{(i,j)\in[n]^2\\ i\neq j}} \prod_{(k,l)\in[K]^2} \mathcal{B}(y_{ij}|\eta_{kl})^{z_{ik}z_{jl}} \\ &= \prod_{\substack{(i,j)\in[n]^2\\ i\neq j}} \prod_{(k,l)\in[K]^2} (\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{1-y_{ij}})^{z_{ik}z_{jl}}. \end{aligned} \tag{1.10}$$

For the undirected binary model, $f_{\boldsymbol{Y}|\boldsymbol{Z}}(\cdot|\cdot)$ is defined by:

$$f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) := \prod_{\substack{(i,j)\in[n]^2\\ i<j}} \prod_{(k,l)\in[K]^2} (\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{1-y_{ij}})^{z_{ik}z_{jl}}. \tag{1.11}$$

Therefore, $f_{\boldsymbol{Y},\boldsymbol{Z}}(\cdot)$ is given, for the directed model, by:

$$f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) = \prod_{i\in[n]}\prod_{k\in[K]} \pi_k^{z_{ik}} \prod_{\substack{(i,j)\in[n]^2\\ i\neq j}} \prod_{(k,l)\in[K]^2} (\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{1-y_{ij}})^{z_{ik}z_{jl}}.$$

Similary, $f_{\boldsymbol{Y},\boldsymbol{Z}}(\cdot)$ is given, for the undirected model, by:

$$f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) = \prod_{i\in[n]}\prod_{k\in[K]} \pi_k^{z_{ik}} \prod_{\substack{(i,j)\in[n]^2\\ i<j}} \prod_{(k,l)\in[K]^2} (\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{1-y_{ij}})^{z_{ik}z_{jl}}.$$

Note that, again, the only difference between the two formulas is that the relation "$i \neq j$" for the directed model is replaced with "$i < j$" for the undirected model.

For simplifying notations, the notation $f(\cdot)$ will sometimes be used instead of $f_{\boldsymbol{Y},\boldsymbol{Z}}(\cdot)$ to denote the joint probability mass function of the couple $(\boldsymbol{Y},\boldsymbol{Z})$.

Remark that the binary stochastic block model is a generalisation of the Erdös-Rényi model. Indeed, if we consider the binary stochastic block model with only $K = 1$ class, the class-weight parameter $\boldsymbol{\pi}$ is trivially the scalar 1, and the edge-probability parameter is the scalar $\boldsymbol{\eta} = \eta$. We obtain then a model equivalent to the Erdös-Rényi model with parameter η.

Many authors (Allman et al., 2009, 2011; Latouche et al., 2011; Nowicki & Snijders, 1997, 2001) remark that very often, in practice, only $\boldsymbol{Y}$, ie. the graph of the network, can be observed. While $\boldsymbol{Z}$ is unobserved, ie. there is no information on the communities to which nodes belong. A scenario in which $\boldsymbol{Z}$ is unobserved is called *a-posteriori* block modelling (Karrer & Newman, 2011; Nowicki & Snijders, 2001). Scenarios in which one part of the sample is unobserved are also called *incomplete data* scenarios (Dempster et al., 1977). The *a-posteriori* block modelling scenario is opposed to the *a-priori* block modelling scenario, in which the vector $\boldsymbol{Z}$ is observed. A-priori block modelling scenarios are much less frequent than a-posteriori block modelling scenarios.

Indeed, the main aim of a-posteriori stochastic block modelling is estimating $\boldsymbol{Z}$ (Allman et al., 2009, 2011; Latouche et al., 2011; Nowicki & Snijders, 1997, 2001). In this thesis, we will assume in general that $\boldsymbol{Z}$ is unobserved, and focus will be given on methods that estimate $\boldsymbol{Z}$, ie. that estimate the communities of the network.

1.2.2 Sufficient statistics for the binary model

Simpler formulas for the probability mass function of the binary model were given by Nowicki & Snijders (1997). Let us define the following quantities:

$$n_k := \sum_{i\in[n]} z_{ik}$$

$$n_{kl} := \sum_{\substack{i,j\in[n]^2\\ i\neq j}} z_{ik}z_{jl} = \begin{cases} n_k(n_k-1) & \text{if } k=l \\ n_k n_l & \text{if } k\neq l \end{cases}$$

$$o_{kl} := \sum_{\substack{i,j\in[n]^2\\ i\neq j}} z_{ik}z_{jl}y_{ij}, \tag{1.12}$$

where n_k denotes the number of nodes belonging to class k, n_{kl} denotes the number of ordered couples (i,j), $i\neq j$ with node i belonging to class k and node j belonging to class l, and o_{kl} denotes the number of ordered couples (i,j) $i\neq j$ with node i belonging to class k and node j belonging to class l such that an edge goes from node i to node j.

Then, the marginal probability mass function of $\boldsymbol{Z}$ can be written as:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{i\in[n]}\prod_{k\in[K]} \pi_k^{z_{ik}} = \prod_{k\in[K]}\prod_{i\in[n]} \pi_k^{z_{ik}} = \prod_{k\in[K]} \pi_k^{\sum_{i\in[n]} z_{ik}} = \prod_{k\in[K]} \pi_k^{n_k}$$

and the p.m.f. of $\boldsymbol{Y}$ conditional on $\boldsymbol{Z}$ can be written, for the directed model, as:

$$\begin{aligned} f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) &= \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \prod_{(k,l)\in[K]^2} (\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{1-y_{ij}})^{z_{ik}z_{jl}} \\ &= \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \prod_{(k,l)\in[K]^2} \eta_{kl}^{z_{ik}z_{jl}y_{ij}}(1-\eta_{kl})^{z_{ik}z_{jl}(1-y_{ij})} \\ &= \prod_{(k,l)\in[K]^2} \eta_{kl}^{\sum_{(i\neq j)} z_{ik}z_{jl}y_{ij}}(1-\eta_{kl})^{\sum_{(i\neq j)} z_{ik}z_{jl}(1-y_{ij})} \\ &= \prod_{(k,l)\in[K]^2} \eta_{kl}^{o_{kl}}(1-\eta_{kl})^{n_{kl}-o_{kl}}. \end{aligned}$$

For the undirected model, this conditional probability mass function can be written as:

$$\begin{aligned} f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) &= \prod_{\substack{(i,j)\in[n]^2 \\ i<j}} \prod_{(k,l)\in[K]^2} (\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{1-y_{ij}})^{z_{ik}z_{jl}} \\ &= \prod_{(k,l)\in[K]^2} \eta_{kl}^{\sum_{(i<j)} z_{ik}z_{jl}y_{ij}}(1-\eta_{kl})^{\sum_{(i<j)} z_{ik}z_{jl}(1-y_{ij})} \\ &= \prod_{(k,l)\in[K]^2} \eta_{kl}^{o_{kl}/2}(1-\eta_{kl})^{(n_{kl}-o_{kl})/2}. \end{aligned}$$

The previous formulas have their equivalent on logarithmic scale:

$$\log f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) = \log f_{\boldsymbol{Z}}(\boldsymbol{z}) + \log f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z})$$

with

$$\log f_{\boldsymbol{Z}}(\boldsymbol{z}) = \sum_{k\in[K]} n_k \log \pi_k \tag{1.13}$$

and, for the directed model:

$$\log f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) = \sum_{(k,l)\in[K]^2} o_{kl}\log\eta_{kl} - o_{kl}\log(1-\eta_{kl}) + n_{kl}\log(1-\eta_{kl}). \tag{1.14}$$

For the undirected model, we simply divide the right hand side term of the last equality by a factor 2.

Because the probability mass function only depends on the statistics $\{n_k\}_{k\in[K]}$ and $\{o_{kl}\}_{(k,l)\in[K]^2}$, these contain all the information on the parameters of the model.
As such, by the Fisher-Neyman factorization theorem, we can affirm that the joint probability distribution of $(\boldsymbol{Y},\boldsymbol{Z})$ conditionally on these statistics does not depend on the parameters of the model.

1.2.3 Expected degree for the binary model

The degree of node i, denoted by d_i, is the number of edges starting from such a node (Bondy & Murty, 2008; Daudin et al., 2008; Karrer & Newman, 2011):

$$d_i := \sum_{j \in [n]} Y_{ij}.$$

Formulas for the expected value of d_i, that will be denoted by $\bar{d}$, will next be given. Note that the distribution of d_i, and by consequence its expected value, do not depend on i. We can then write:

$$\begin{aligned}
\bar{d} :&= \mathbb{E}(d_i) \\
&= \mathbb{E}(d_1) \\
&= \mathbb{E}(\sum_{j \in [n]} Y_{1j}) \\
&= \sum_{j \in [n]} \mathbb{E}(Y_{1j}) \\
&= n\mathbb{E}(Y_{12}),
\end{aligned}$$

where the fourth equality yields from linearity of the expectation, and the fifth equality holds because variables $\{Y_{1j}\}_{j \in [n]}$ have, by definition, the same distribution. We have assumed here that self-edges are allowed; otherwise, the multiplicative factor would be $n-1$ instead of n. Note that, for the binary model, $\mathbb{E}(Y_{ij}) = \mathbb{P}(Y_{ij} = 1)$. Furthermore, we can write (Daudin et al., 2008; Bickel et al., 2013):

$$\begin{aligned}
\mathbb{E}(Y_{ij}) &= \sum_{k,l \in [K]} \mathbb{E}(Y_{ij} | Z_i = k \text{ and } Z_j = l) \mathbb{P}(Z_i = k \text{ and } Z_j = l) \\
&= \sum_{k,l \in [K]} \eta_{kl} \pi_k \pi_l \\
&= \boldsymbol{\pi}^T \boldsymbol{\eta} \boldsymbol{\pi},
\end{aligned}$$

so that the expected degree of a node is given by the simple formula:

$$\bar{d} = n \boldsymbol{\pi}^T \boldsymbol{\eta} \boldsymbol{\pi}. \tag{1.15}$$

1.2.4 The binary model is an exponential family model

The exponential family is a wide class of parametric models with good statistical properties (Dempster et al., 1977; Wu, 1983). After reminding the general definition of exponential families and some of their basic properties, it will be shown that the binary stochastic block models are exponential family models.

A parametric model $\{\mathbb{P}_{\boldsymbol{\phi}}\}_{\boldsymbol{\phi} \in \Phi}$, where Φ denotes a d-dimensional convex set, is said to be a member of the regular exponential family (Dempster et al., 1977) if the probability mass function (or the probability density function, in the continuous scenario) has the following form:

$$f(\boldsymbol{x} | \boldsymbol{\phi}) = \frac{B(\boldsymbol{x}) e^{\boldsymbol{\phi}^T t(\boldsymbol{x})}}{A(\boldsymbol{\phi})} \qquad \forall \boldsymbol{\phi} \in \Phi, \tag{1.16}$$

where $\boldsymbol{x}$ belongs to an Euclidean space, e denotes Euler's number, $A(\boldsymbol{\phi}) \in \mathbb{R}_+^0$, with $\mathbb{R}_+^0$ denoting the set of the strictly positive real numbers, $B(\boldsymbol{x}) \in \mathbb{R}_+$, with $\mathbb{R}_+$ denoting the set of

non-negative real numbers, and $t(\boldsymbol{x}) \in \mathbb{R}^d$ is a vector of sufficient statistics.

Also, a more general class of models, the curved exponential family models (Wu, 1983) can be defined. Its definition is actually the same as (1.16), but it is supposed now that Φ is a curved manifold of a d-dimensional convex set Φ'; this means that parameters in $\boldsymbol{\phi} \in \Phi$ are subject to a constraint of the type $h(\boldsymbol{\phi}) = 0$ where h is a smooth function.

Many well-known families of distributions are regular exponential family models: exponential distributions, normal distributions, Bernoulli distributions, binomial distributions, Poisson distributions, etc. However, in order to see this, a different parametrisation than the "usual" one must be chosen. The parameter $\boldsymbol{\phi}$ is called the natural parameter of the exponential family (Dempster et al., 1977). One important property of the regular exponential families is that their log-likelihood function is concave. If the log-likelihood function is strictly concave, the maximum likelihood estimator has a unique solution, which is a very attractive property. Also, the Fisher information matrix for regular exponential families takes a simple form. This matrix can be defined as:

$$I(\boldsymbol{\phi}_0) := -\mathbb{E}_{\boldsymbol{\phi}_0}(D^2_{\boldsymbol{\phi}} \log f(\boldsymbol{x}|\boldsymbol{\phi})|_{\boldsymbol{\phi}_0}) \quad \forall \boldsymbol{\phi}_0 \in \Phi,$$

where $D^2_{\boldsymbol{\phi}} \log f(\boldsymbol{x}|\boldsymbol{\phi})$ is the Hessian of the function $\boldsymbol{\phi} \to \log f(\boldsymbol{x}|\boldsymbol{\phi})$, and $\mathbb{E}_{\boldsymbol{\phi}_0}(\cdot)$ denotes the expectation assuming that the parameter of the distribution is $\boldsymbol{\phi}_0$.
By abuse of notation, we will also use the notation $I(\boldsymbol{\phi}_0) = -\mathbb{E}_{\boldsymbol{\phi}_0}(D^2_{\boldsymbol{\phi}_0} \log f(\boldsymbol{x}|\boldsymbol{\phi}_0))$ in this section.

By denoting $a(\cdot) = \log(A(\cdot))$ and $b(\cdot) = \log(B(\cdot))$, we can express the logarithm of the probability mass function (or, in the continuous case, the logarithm of the probability density function), that we will call the log-mass function, in the following way:

$$\log f(\boldsymbol{x}|\boldsymbol{\phi}) = \boldsymbol{\phi}^T t(\boldsymbol{x}) - a(\boldsymbol{\phi}) + b(\boldsymbol{x}).$$

If we differentiate the log-mass function with respect to $\boldsymbol{\phi}$ we get:

$$\begin{aligned} D_{\boldsymbol{\phi}} \log f(\boldsymbol{x}|\boldsymbol{\phi}) &= D_{\boldsymbol{\phi}}(\boldsymbol{\phi}^T t(\boldsymbol{x}) - a(\boldsymbol{\phi}) + b(\boldsymbol{x})) \\ &= t(\boldsymbol{x}) - D_{\boldsymbol{\phi}} a(\boldsymbol{\phi}), \end{aligned}$$

and differentiating again:

$$\begin{aligned} D^2_{\boldsymbol{\phi}} \log f(\boldsymbol{x}|\boldsymbol{\phi}) &= D_{\boldsymbol{\phi}}(t(\boldsymbol{x}) - D_{\boldsymbol{\phi}} a(\boldsymbol{\phi})) \\ &= -D^2_{\boldsymbol{\phi}} a(\boldsymbol{\phi}), \end{aligned}$$

so that the Fisher information matrix is obtained in terms of the Hessian of $a(\cdot)$ as :

$$I(\boldsymbol{\phi}) = -\mathbb{E}_{\boldsymbol{\phi}}(D^2_{\boldsymbol{\phi}} \log f(\boldsymbol{x}|\boldsymbol{\phi})) = -\mathbb{E}_{\boldsymbol{\phi}}(-D^2_{\boldsymbol{\phi}} a(\boldsymbol{\phi})) = D^2_{\boldsymbol{\phi}} a(\boldsymbol{\phi}).$$

Another interesting property of the regular exponential families (proved in Appendix 7.2) is that the gradient of $a(\cdot)$ is equivalent to the expectation of $t(\boldsymbol{X})$:

$$D_{\boldsymbol{\phi}} a(\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{\phi}} t(\boldsymbol{X}),$$

which also implies that the maximum-likelihood estimator $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}(\boldsymbol{x})$, if it lies in the interior of the parameters space, satisfies the following relation:

$$\mathbb{E}_{\hat{\boldsymbol{\phi}}}(t(\boldsymbol{X})) = t(\boldsymbol{x}).$$

A last property of this family (proved in Appendix 7.3) is that the Hessian of $a(\cdot)$ is equivalent to the variance-covariance matrix of $t(\boldsymbol{X})$:

$$D^2_{\boldsymbol{\phi}} a(\boldsymbol{\phi}) = I(\boldsymbol{\phi}) = Var_{\boldsymbol{\phi}} t(\boldsymbol{X}),$$

and therefore:

$$D^2_{\boldsymbol{\phi}} \log f(\boldsymbol{x}|\boldsymbol{\phi}) = -I(\boldsymbol{\phi}) = -Var_{\boldsymbol{\phi}} t(\boldsymbol{x}).$$

Since the matrix $D^2_{\boldsymbol{\phi}} \log f(\boldsymbol{x}|\boldsymbol{\phi})$ is equal to the negative of the variance-covariance matrix of $t(\boldsymbol{X})$ (which is always positive semi-definite by well-known properties of the variance-covariance matrices), it is always negative semi-definite, and so the log-likelihood function is concave. If furthermore this matrix is negative definite, this implies that the log-likelihood is strictly convex and therefore the maximum likelihood estimator has a unique solution.

We will now show that the binary stochastic block model is a curved exponential family model. To see this, the parametrization introduced in the work of Bickel et al. (2013) can be used. Bickel et al. (2013) define the parameters $\boldsymbol{\omega} \in \mathbb{R}^{K-1}$ and $\boldsymbol{\nu} \in \mathbb{R}^{K \times K}$ as:

$$\begin{aligned} \omega_k &:= \log \frac{\pi_k}{1 - \sum_{l=1}^{K-1} \pi_l} = \log \frac{\pi_k}{\pi_K} \quad \forall k \in [K-1] \\ \nu_{kl} &:= \log \frac{\eta_{kl}}{1 - \eta_{kl}} \qquad\qquad \forall (k,l) \in [K]^2, \end{aligned}$$

where we remind the reader that $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$ are the usual parameters of the binary stochastic block model as introduced previously.

We will now show that the log-mass function of the binary stochastic block model:

$$\log f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) = \log f_{\boldsymbol{Z}}(\boldsymbol{z}) + \log f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z})$$

can be written as displaying the form of an exponential family log-mass function.

Indeed, in the next paragraph it will be proven that the marginal log-mass function of $\boldsymbol{Z}$ (for both the directed and undirected model) can be written as:

$$\log f_{\boldsymbol{Z}}(\boldsymbol{z}) = \Big(\sum_{k=1}^{K-1} n_k \omega_k\Big) - n \log\Big(1 + \sum_{k=1}^{K-1} e^{\omega_k}\Big), \tag{1.17}$$

(statistics n_k, n_{kl} and o_{kl} were defined in equation 1.12).

Furthermore, for the directed block model, it will be proven that the log-mass function of $\boldsymbol{Y}$ conditional on $\boldsymbol{Z}$ can be written as:

$$\log f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) = \sum_{k=1}^{K}\sum_{l=1}^{K} o_{kl}\nu_{kl} - n_{kl}\log(1+e^{\nu_{kl}}). \tag{1.18}$$

Similarly, for the undirected model, the log-mass function of $\boldsymbol{Y}$ conditional on $\boldsymbol{Z}$ can be written as:

$$\log f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) = \sum_{k=1}^{K}\sum_{l=1}^{K} \frac{o_{kl}}{2}\nu_{kl} - \frac{n_{kl}}{2}\log(1+e^{\nu_{kl}}). \tag{1.19}$$

Therefore, both the directed and the undirected binary stochastic block models are curved exponential family models, with natural parameters

$$\boldsymbol{\phi} = \{\{\omega_k\}_{k\in[K-1]}, \{\nu_{kl}\}_{k,l\in[K]^2}, \{\log(1+e^{\nu_{kl}})\}_{k,l\in[K]^2}\}.$$

For the directed model, $t(\boldsymbol{y},\boldsymbol{z})$ is given by:

$$t(\boldsymbol{y},\boldsymbol{z}) = \{\{n_k\}_{k\in[K-1]}, \{o_{kl}\}_{k,l\in[K]^2}, \{n_{kl}\}_{k,l\in[K]^2}\}.$$

For the undirected model, $t(\boldsymbol{y},\boldsymbol{z})$ is given by:

$$t(\boldsymbol{y},\boldsymbol{z}) = \{\{n_k\}_{k\in[K-1]}, \{\frac{o_{kl}}{2}\}_{k,l\in[K]^2}, \{\frac{n_{kl}}{2}\}_{k,l\in[K]^2}\},$$

and $a(\boldsymbol{\phi}) = n\log(1+\sum_{k=1}^{K-1} e^{\omega_k})$ for both the directed and the undirected model.

The proof of the formulas (1.17) and (1.18) is given below; the proof of formula (1.19) follows the same structure as the proof of formula (1.18).

Proof. For formula (1.17):

$$\begin{aligned}
\sum_{k=1}^{K-1} n_k\omega_k - n\log(1+\sum_{k=1}^{K-1} e^{\omega_k}) &= \sum_{k=1}^{K-1} n_k\omega_k - n\log(1+\sum_{k=1}^{K-1}\frac{\pi_k}{1-\sum_{l=1}^{K-1}\pi_l}) \\
&= \sum_{k=1}^{K-1} n_k\omega_k - n\log(1+\frac{\sum_{k=1}^{K-1}\pi_k}{1-\sum_{l=1}^{K-1}\pi_l}) \\
&= \sum_{k=1}^{K-1}(n_k\log(\frac{\pi_k}{1-\sum_{l=1}^{K-1}\pi_l})) - n\log(\frac{1}{1-\sum_{k=1}^{K-1}\pi_l}) \\
&= \sum_{k=1}^{K-1}(n_k\log\pi_k - n_k\log(1-\sum_{k=1}^{K-1}\pi_l)) + n\log(1-\sum_{k=1}^{K-1}\pi_l) \\
&= \sum_{k=1}^{K-1}(n_k\log\pi_k - n_k\log\pi_K) + n\log\pi_K \\
&= \sum_{k=1}^{K} n_k\log\pi_k.
\end{aligned}$$

This last term is indeed the log-mass function of $\boldsymbol{Z}$ (see equation 1.13). To obtain the last equality, the relation $\sum_{k=1}^{K} n_k = n$ was used.

For formula (1.18):

$$\begin{aligned}\sum_{k=1}^{K}\sum_{l=1}^{K} o_{kl}\nu_{kl} - n_{kl}\log(1+e^{\nu_{kl}}) &= \sum_{k=1}^{K}\sum_{l=1}^{K} o_{kl}(\log\eta_{kl} - \log(1-\eta_{kl})) - n_{kl}\log(1+\frac{\eta_{kl}}{1-\eta_{kl}}) \\ &= \sum_{k=1}^{K}\sum_{l=1}^{K} o_{kl}\log\eta_{kl} - o_{kl}\log(1-\eta_{kl}) - n_{kl}\log(1+\frac{\eta_{kl}}{1-\eta_{kl}}) \\ &= \sum_{k=1}^{K}\sum_{l=1}^{K} o_{kl}\log\eta_{kl} - o_{kl}\log(1-\eta_{kl}) + n_{kl}\log(1-\eta_{kl}).\end{aligned}$$

This last term is indeed the conditional log-mass of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ (see equation 1.14).

□

1.3 Connected components of a graph

Until now, a graph on n nodes has been represented as an $n \times n$ binary matrix $\boldsymbol{Y}$. However, the following equivalent, but formally different definition of a graph will be useful in this section. A graph is defined as the couple (V, E), where V denotes the vertex/node set, which was previously denoted by $[n]$. V^2 denotes the set of all the ordered couples of the network. The subset $E \subseteq V^2$ denotes then the edge set of the graph, which can be written in terms of $\boldsymbol{Y}$ as:

$$E := \{(i,j) \in V^2 \,|Y_{ij} = 1\}.$$

Such a definition is used for example in the work of Bondy & Murty (2008). It is clear that $\boldsymbol{Y}$ and E provide exactly the same information.
An important result concerning the structure of graphs will be stated in Theorem 1. Before stating such a result, a few other standard graph theory concepts must be defined.

For any vertex set V, the graph (V, V^2) is called the *complete graph* on V. In a complete graph, all nodes of the network are connected by an edge.

If two graphs (V, E_1) and (V, E_2) verify $E_1 \subseteq E_2$, we say that (V, E_1) is a *subgraph* of (V, E_2). Therefore, every graph is a subgraph of the complete graph.

Given m graphs (V_1, E_1), (V_2, E_2), ... , (V_m, E_m), the *union* of these graphs is defined as the graph (V, E) whose vertex set is the union of the individual vertex sets : $V = V_1 \cup V_2 ... \cup V_m$, and the edge set is the union of the individual edge sets: $E = E_1 \cup E_2 ... \cup E_m$. Two graphs (V_1, E_1) and (V_2, E_2) are said to be *disjoint* if $V_1 \cap V_2 = \emptyset$ and $E_1 \cap E_2 = \emptyset$.

A *path* is defined as a finite sequence of edges in E, $(i_1, j_1), (i_2, j_2), ..., (i_m, j_m)$ such that $j_1 = i_2$, $j_2 = i_3, ...$, $j_{m-1} = i_m$. A graph is said to be *connected* if for every pair of nodes i, j there exists a path going from i to j, ie. a path such that $i_1 = i$ and $j_m = j$.

Theorem 1 is a standard result in graph theory and shows that every graph is the union of pairwise disjoint connected graphs:

Theorem 1. *(Bondy & Murty, 2008). Every undirected graph can be decomposed as the union of pairwise disjoint connected graphs: such a decomposition is unique and yields the so-called connected components decomposition of the graph.*

This theorem will be useful in Section 1.8, in which some results on the asymptotic behaviour of networks shall be discussed.

1.4 The stochastic block model - general definition

In the binary stochastic block model, edges are represented as binary variables, therefore Bernoulli distributions are used to model them. However, when edges are not binary, but represent, for example, counts of some events (for example, the number of phone calls), or even continuous variables (for example, variables representing the duration of the phone call between two people), edges are *weighted* and must be modelled by other families of distributions. This brings us to the definition of a more general family of stochastic block models, as given in the work of Mariadassou et al. (2010), which includes the binary model as a particular case; this more general family also includes the Poisson block model (Section 1.5) and the Gaussian block model (Section 1.6).

For any one-dimensional parameter space $\Theta \subseteq \mathbb{R}$, let $\Theta^{K\times K}$ denote the matrix space consisting of the $K \times K$ matrices $\boldsymbol{\theta} = (\theta_{kl})_{k,l\in[K]}$ with entries θ_{kl} such that $\theta_{kl} \in \Theta$.

$$\Theta^{K\times K} := \{\boldsymbol{\theta} = \{\theta_{kl}\}_{k,l\in[K]} \mid \theta_{kl} \in \Theta \quad \forall k, l \in [K]\}.$$

This means that the matrix $\boldsymbol{\theta}$ is one instance of the space $\Theta^{K\times K} \subseteq \mathbb{R}^{K\times K}$. In a general directed stochastic block model of parameters $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in \Theta^{K\times K} \times \Pi_K$, three of the four hypotheses of the binary directed model, ie. hypotheses (1.5), (1.6) and (1.8) remain the same as for the binary model, while hypothesis (1.7) becomes:

$$Y_{ij}|Z_i = k, Z_j = l \sim g(\cdot|\theta_{kl}) \qquad \forall i, j = 1, ..., n \;\; i \neq j, \tag{1.20}$$

where $g(\cdot|\theta)$ is a probability mass function (or, in the continuous case, a probability density function) depending on the one-dimensional parameter $\theta \in \Theta$. More generally, θ may be d-dimensional, but such a generalisation will only be necessary for defining the Gaussian block model. In the binary model, the Poisson model, and other models used in practice, θ in indeed one-dimensional and represents the mean of the distribution $g(\cdot)$.
For example, in the binary block model of Section (1.2), we had that $g(\cdot)$ was the probability mass function of a Bernoulli distribution:

$$g(y_{ij}|\eta_{kl}) = \mathcal{B}(y_{ij}|\eta_{kl}) = (\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{1-y_{ij}})^{z_{ik}z_{jl}}.$$

While in a Poisson block model, $g(\cdot)$ will be the probability mass function of a Poisson distribution, denoted by $\mathcal{P}(\cdot)$:

$$g(y_{ij}|\lambda_{kl}) = \mathcal{P}(y_{ij}|\lambda_{kl}) = e^{-\lambda_{kl}}\frac{\lambda_{kl}^{y_{ij}}}{(y_{ij}!)}.$$

The probability mass function $f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{z})$ of a general stochastic block model can be written, as for the binary model, by the product of the functions $f_{\boldsymbol{Z}}(\boldsymbol{z})$ and $f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z})$:

$$f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) = f_{\boldsymbol{Z}}(\boldsymbol{z}) f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}).$$

The function $f_{\boldsymbol{Z}}(\boldsymbol{z})$ is defined as for the binary model (see equation 1.9):

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) := \prod_{i\in[n]} \prod_{k\in[K]} \pi_k^{z_{ik}}.$$

The function $f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z})$ is defined for a general directed model as:

$$\begin{aligned} f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) &:= \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} g(y_{ij}|\theta_{z_i z_j}) \\ &= \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \prod_{(k,l)\in[K]^2} g(y_{ij}|\theta_{kl})^{z_{ik}z_{jl}}. \end{aligned} \tag{1.21}$$

Note that if $g(\cdot)$ is the probability mass function of the Bernoulli distribution, (1.21) corresponds indeed to (1.10).
For a general undirected model, again, the function $f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z})$ is defined by simply replacing the relation "$i \neq j$" in the previous formula with "$i < j$":

$$f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) := \prod_{\substack{(i,j)\in[n]^2 \\ i<j}} g(y_{ij}|\theta_{z_i z_j}).$$

Therefore, $f_{\boldsymbol{Y},\boldsymbol{Z}}(\cdot)$ is given, for the general directed model, by:

$$\begin{aligned} f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) &= \prod_{i\in[n]} \prod_{k\in[K]} \pi_k^{z_{ik}} \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} g(y_{ij}|\theta_{z_i z_j}). \\ &= \prod_{i\in[n]} \prod_{k\in[K]} \pi_k^{z_{ik}} \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \prod_{k,l\in[K]^2} g(y_{ij}|\theta_{kl})^{z_{ik}z_{jl}}, \end{aligned} \tag{1.22}$$

which on the logarithmic scale yields:

$$\log f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) = \sum_{i\in[n]} \sum_{k\in[K]} z_{ik} \log \pi_k + \sum_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \sum_{(k,l)\in[K]^2} z_{ik} z_{jl} \log g(y_{ij}|\theta_{kl}). \tag{1.23}$$

For the general undirected model, $f_{\boldsymbol{Y},\boldsymbol{Z}}(\cdot)$ is given by:

$$\begin{aligned} f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) &= \prod_{i\in[n]} \prod_{k\in[K]} \pi_k^{z_{ik}} \prod_{\substack{(i,j)\in[n]^2 \\ i<j}} g(y_{ij}|\theta_{z_i z_j}). \\ &= \prod_{i\in[n]} \prod_{k\in[K]} \pi_k^{z_{ik}} \prod_{\substack{(i,j)\in[n]^2 \\ i<j}} \prod_{k,l\in[K]^2} g(y_{ij}|\theta_{kl})^{z_{ik}z_{jl}}, \end{aligned}$$

which on the logarithmic scale yields:

$$\log f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) = \sum_{i\in[n]}\sum_{k\in[K]} z_{ik}\log\pi_k + \sum_{\substack{(i,j)\in[n]^2\\ i<j}}\sum_{(k,l)\in[K]^2} z_{ik}z_{jl}\log g(y_{ij}|\theta_{kl}). \tag{1.24}$$

The notation $g_{kl}(y_{ij})$ instead of $g(y_{ij}|\theta_{kl})$ will be sometimes used to denote more shortly the probability mass function (or, in the continuous case, the density function) of $\boldsymbol{Y}_{ij}$ indexed by parameter θ_{kl}; if such is the case, it will be explicitly mentioned in the text.

1.5 The Poisson stochastic block model

In a Poisson SBM (Mariadassou et al., 2010), $\Theta = \mathbb{R}_+$, and $g(\cdot)$ is a Poisson distribution of parameter $\theta = \lambda \in \mathbb{R}_+$; therefore edges have integer weights. For a Poisson model, the probability mass function of $\boldsymbol{Y}$ conditional on $\boldsymbol{Z}$ is:

$$f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) = \prod_{\substack{(i,j)\in[n]^2\\ i\neq j}} \mathcal{P}(y_{ij}|\lambda_{z_iz_j}) \tag{1.25}$$

$$= \prod_{\substack{(i,j)\in[n]^2\\ i\neq j}} e^{-\lambda_{z_iz_j}}\frac{\lambda_{z_iz_j}^{y_{ij}}}{(y_{ij}!)} \quad \forall(\boldsymbol{y},\boldsymbol{z})\in\mathcal{Y}\times\mathcal{Z}, \tag{1.26}$$

where $\mathcal{Y} := \mathbb{M}_{n\times n}(\mathbb{N}_0)$ denotes now the set of all matrices whose elements are in the set {0,1,2...}, while the definition of $\mathcal{Z}$ is the as for the binary model.
Formula (1.25) is valid for the directed Poisson model; for the undirected Poisson model, again, we simply replace the relation "$i \neq j$" with "$i < j$":

$$f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) = \prod_{\substack{(i,j)\in[n]^2\\ i<j}} \mathcal{P}(y_{ij}|\lambda_{z_iz_j}) \quad \forall(\boldsymbol{y},\boldsymbol{z})\in\mathcal{Y}\times\mathcal{Z}. \tag{1.27}$$

The same developments that were done for computing the mean degree of the binary stochastic block model, (see equation 1.15), can also be done for the Poisson stochastic block model, to show that the mean degree in a Poisson stochastic block model is given by the formula:

$$\bar{d} := \mathbb{E}(\sum_{j\in[n]} Y_{ij}) = n\boldsymbol{\pi}^T\boldsymbol{\lambda}\boldsymbol{\pi}. \tag{1.28}$$

which is equivalent to formula (1.15), only with $\boldsymbol{\lambda}$ replacing $\boldsymbol{\eta}$. The Poisson model is very widely used. Indeed, many real-world networks are sparse, ie. have (very) few edges and many nodes, and can therefore be modelled by binary stochastic block models with small values of parameters η_{kl}. In such cases, it may be approximately equivalent to use a Poisson model of parameters $\lambda_{kl} = \eta_{kl}$. For example, Zhao et al. (2012) show that, for sparse networks, the Poisson profile likelihood estimator yields approximately the same results as the binary profile likelihood estimator (both of these estimators will be defined in Section 3.6).

1.6 The Gaussian stochastic block model

In a Gaussian SBM (Mariadassou et al., 2010), $\Theta = \mathbb{R} \times \mathbb{R}_+^0$, and $g(\cdot)$ is a Gaussian distribution of parameter $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$, so that:

$$f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) = \prod_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{y_{ij} - \mu_{z_i z_j}}{\sigma_{z_i z_j}})^2) \quad \forall (\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{Y} \times \mathcal{Z},$$

where $\mathcal{Y} = \mathbb{M}_{n\times n}(\mathbb{R})$ denotes now the set of all real matrices, and $\exp(\cdot)$ denotes the usual exponential function. (Note that in the Gaussian model, contrarily to the binary and the Poisson model, $f_{\boldsymbol{Y}|\boldsymbol{Z}}(\cdot)$ is not a probability mass function, but a density function with respect to the product measure of the Lebesgue measure and the discrete counting measure). This last formula is valid for the directed Gaussian model, for the undirected Gaussian model we simply replace the relation "$i \neq j$" with "$i < j$".

Such a model is not as widely used as the binary model and the Poisson model are, and will not be focused upon in rest of this thesis, but it is mentioned mainly as an additional example to show that the family of the stochastic block models is indeed very flexible.

1.7 The affiliation model and the symmetric model

The binary affiliation block model is a particular type of binary stochastic block model, discussed in several works (Allman et al., 2011; Celisse et al., 2012; Decelle et al., 2011; Zanghi et al., 2008). Such a model is also called the planted partition model (Decelle et al., 2011). It is a simpler version of the binary block model, depending only on the $K-1$ parameters $\boldsymbol{\pi}$ and 2 other parameters, p_{in} and p_{out}. The parameter p_{in} is the intra-class edge probability, ie. the probability of an edge between two nodes of the same class, while p_{out} is the inter-class edge probability, ie. the probability of an edge between two nodes of different classes. The binary affiliation model of parameters $(\boldsymbol{\pi}, p_{in}, p_{out})$ is then a binary stochastic block model with entries of the edge-probability matrix $\boldsymbol{\eta}$ being p_{in} on the diagonal and p_{out} off-diagonal:

$$\eta_{kl} = \begin{cases} p_{in} & \text{if } k = l \\ p_{out} & \text{if } k \neq l. \end{cases}$$

Similarly, a Poisson affiliation block model is a Poisson block model with matrix $\boldsymbol{\lambda}$ verifying:

$$\lambda_{kl} = \begin{cases} p_{in} & \text{if } k = l \\ p_{out} & \text{if } k \neq l. \end{cases}$$

In the Poisson affiliation model, p_{in} represents the mean weight of an edge between two nodes of the same class, while p_{out} represents the mean weight of an edge between two nodes of different classes.
By varying the values of the parameters of the affiliation model, different scenarios can be considered: if p_{in} is large and p_{out} is small, two nodes belonging to the same class are more likely to being connected than two nodes from different classes. The opposite is true if p_{in} is small and p_{out} is large. We will call an affiliation model *assortative* when $p_{in} > p_{out}$ and

disassortative when $p_{in} < p_{out}$.

A symmetric (Poisson or binary) block model (Abbe, 2018), denoted by $SSBM(n, K, p_{in}, p_{out})$, is defined as an affiliation model in which all classes have equal weights, ie. $\pi_k = 1/K$:

$$\theta_{kl} = \begin{cases} p_{in} & \text{if } k = l \\ p_{out} & \text{if } k \neq l, \end{cases}$$

$$\pi_k = \frac{1}{K} \qquad \forall k = 1, .., K. \tag{1.29}$$

Equations (1.15) and (1.28) (for respectively the binary model and the Poisson model) and simple algebraic manipulations imply that the mean degree in a symmetric model is given by:

$$\bar{d} = n\frac{p_{in} + (K-1)p_{out}}{K}. \tag{1.30}$$

1.8 Asymptotic behaviour of the block models

One important question that arises when studying block models is their asymptotic behaviour, ie. their behaviour when the number of nodes in the network tends to infinity.
We will focus on the binary model in this section. In order to study its asymptotic behaviour, different scenarios must be considered, depending on how the expected degree of the network behaves as n increases. We remind the reader that the expected degree for the binary model is given by (see equation 1.15):

$$\bar{d} = n\boldsymbol{\pi}^T\boldsymbol{\eta}\boldsymbol{\pi}.$$

In a first scenario, called the dense regime (Abbe, 2018), the expected degree of the nodes increases linearly with the size of the network:

$$\text{Dense regime:} \quad \bar{d} = cn, \tag{1.31}$$

where c is a constant; note that this is what happens when the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$ are fixed, ie. they do not change with n. The dense regime implies that as the size of the network grows, the amount of information in the network (ie., the number of edges) increases. A very different regime, called the sparse regime (Abbe, 2018), is one in which the degree is fixed:

$$\text{Sparse regime:} \quad \bar{d} = c, \tag{1.32}$$

where c is a constant. In such a regime, the network size has no effect on the amount of information in the network.
In a dense regime, increasing the sample size has a positive effect on the quality of the estimation; while in a sparse regime, increasing the sample does not improve the estimation. Both of these regimes are quite extreme. Statisticians are interested in intermediate regimes, that serve as good benchmarks (Abbe, 2018). In such regimes the degree increases when the network size increases, but not as fast as in the dense regime. An interesting intermediate scenario, from a theoretical point of view, is one in which the mean degree increases at a logarithmic rate with respect to the sample size:

$$\text{Intermediate regime:} \quad \bar{d} = s \log n, \tag{1.33}$$

where s is a constant that will be called the *sparsity/density* parameter. The higher s is, the denser is the network. Note that s is given in terms of the original parametric couple $(\boldsymbol{\pi}, \boldsymbol{\eta})$ as $s = \frac{\bar{d}}{\log n} = \frac{n}{\log n}\boldsymbol{\pi}^T\boldsymbol{\eta}\boldsymbol{\pi}$.
Very precise results on the asymptotic behaviour of networks were given in the work of Erdös & Rényi (1960) (as cited by Bondy & Murty (2008)), under the Erdös-Rényi model assumptions. The following result concerns networks in an intermediate regime and it states the conditions, in terms of s, for the network to be *connected* (in the sense given in Section 1.3) as n tends to infinity.

Theorem 2. *(Erdös & Rényi, 1960), as cited by (Bondy & Murty, 2008). Let s be a positive constant. Consider the sequence of undirected Erdös-Rényi models $ER(n, s\frac{\log n}{n})$, with expected degree $\bar{d} = s \log n$. Then the network is connected, ie. it has a unique connected component, with high probability if and only if $s > 1$.*

By saying that the network is connected with high probability, we mean that the probability that the network is connected tends to 1 as n tends to infinity. The proof of this theorem is based on branching processes and will not be given in this thesis. The condition that the network is connected is useful because if there are isolated nodes in the network, we can not hope to recover their class. Abbe (2018) states the possibility of generalizing Theorem 2 to the symmetric block model (defined in equation 1.29):

Conjecture *(Abbe, 2018) : Let s_{in} and s_{out} be two strictly positive constants, and K a positive integer constant representing the number of classes. Consider the sequence of undirected symmetric binary block models $SSBM(n, K, s_{in}\frac{\log n}{n}, s_{out}\frac{\log n}{n})$, with expected degree $\bar{d} = s \log n$, where s is defined by $s := \frac{s_{in}+(K-1)s_{out}}{K}$. Then the network is connected, ie. it has a unique connected component, with high probability if and only if $s > 1$.*

However, no proof of such a conjecture has been found.

1.9 The degree-corrected stochastic block model

One of the drawbacks of the block models as defined until now, is that it does not permit degree variation inside communities; however, in many real-world networks, there may be hubs, ie. nodes displaying a (much) higher number of edges than other members of the same class. In such scenarios, the fitting of a classical block model tends to simply divide nodes between high-degree nodes and low-degree nodes. However, this may lead to wrong and uninterpretable results. Therefore, the degree-corrected stochastic block model (DCSBM) has been introduced in the literature (Amini et al., 2013; Karrer & Newman, 2011; Lei & Rinaldo, 2015; Zhang et al., 2014) to model this kind of phenomenon.

We remind the reader that the binary stochastic block model supposes that edges, conditionally on the class membership, follow a Bernoulli distribution of parameter depending exclusively on the class of the nodes:

$$Y_{ij}|Z_i = k, Z_j = l \sim \mathcal{B}(\eta_{kl}) \qquad \forall i, j = 1, ..., n \quad i \neq j.$$

In the binary degree-corrected stochastic block model, one allows the parameter of the distribution to also depend on some individual-specific parameters β_i (Amini et al., 2013; Karrer & Newman, 2011; Lei & Rinaldo, 2015; Zhang et al., 2014):

$$Y_{ij}|Z_i = k, Z_j = l \sim \mathcal{B}(\beta_i \beta_j \eta_{kl}) \qquad \forall i, j = 1, ..., n \quad i \neq j.$$

The parameters vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)^T$ represents then the general tendency of nodes to be connected to the network. If β_i is large, node i will be more likely to create connections then if β_i was small. More specifically, the probability that an edge exists between node i and any other node is proportional to β_i.

In order to identify $\boldsymbol{\beta}$, some constraints must be added (Amini et al., 2013; Karrer & Newman, 2011; Lei & Rinaldo, 2015; Zhang et al., 2014); the most common one being that in each class, the average value of β_i is 1:

$$\sum_{i=1}^{n} \beta_i \mathbb{1}_{\{Z_i = k\}} = \sum_{i=1}^{n} \beta_i z_{ik} = 1 \quad \forall k = 1, 2, .., K. \tag{1.34}$$

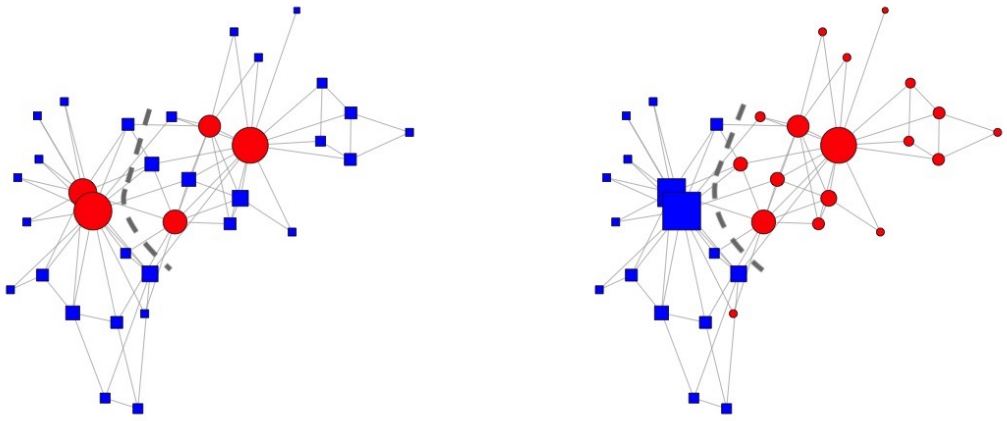
Note that this definition obviously generalizes the standard binary block model. Indeed a standard binary block model corresponds to a degree-corrected block model of constant parameters vector $\boldsymbol{\beta} = 1_n$, where 1_n denotes the vector $\underbrace{(1, ..., 1)}_{n \text{ times}}^T$.
Similarly, the degree-corrected Poisson model is defined by allowing the edge-weight distribution to depend on parameters β_i:

$$Y_{ij}|Z_i = k, Z_j = l \sim \mathcal{P}(\beta_i \beta_j \lambda_{kl}) \qquad \forall i, j = 1, ..., n \quad i \neq j.$$

It has been shown that, in several real-world situations, the degree-corrected block model outperforms the non-degree-corrected block model. For example, the well-known Zachary karate club network (Zachary, 1977) is a quite famous real-world example discussed in the literature. It is a network of a karate club of a university in the United States with 34 members that, after an internal dispute, was split in 2 clubs. For this club, the friendship network was firstly analyzed by Zachary (1977), who showed that the so-called Ford-Fulkerson algorithm splits perfectly the network in the 2 clubs, except for one member.

For this karate club network, Karrer & Newman (2011) showed that the degree-corrected SBM profile likelihood estimator (see Chapter 3) also splits perfectly the network, except for the same member that was misclassified by the algorithm used by Zachary, while the non degree-corrected SBM profile likelihood estimator performs very poorly, by just splitting members between those with many friendship ties and those with few friendship ties. These results are displayed in Figure (1.1). This network will be studied in Chapter 5.

Karrer & Newman (2011) also analyzed a network of 1222 blogs about the 2005 United States presidential election, and the web links between them, as measured on the same day in 2005. All these blogs were labelled by their known political leanings, and again, the degree-corrected block model performed well, producing a block structure yielding a mutual information (a formal definition of such a quantity can be found in Section 4.1) of 0.72 with respect to the true labels. Values of such an index range from 0 to 1, a value closer to 1 indicating a better



(a) Estimation with SBM profile likelihood (b) Estimation with DCSBM profile likelihood

Figure 1.1: *Estimation of the karate club network communities found using the standard (non degree-corrected) Poisson model profile likelihood method (a) and the degree-corrected model profile likelihood method (b). The size of each node is proportional to its degree; the colour and the shape of each node display its estimated group membership. The dashed line indicates the split observed in real life. Source: Karrer & Newman (2011).*

performance. The classical block model performed very poorly, yielding a mutual information of 0.0001. This network will also be studied in Chapter 5.

Chapter 2: Identifiability

This chapter investigates the identifiability properties of the stochastic block models. In Section 2.1, two different kinds of identifiability will be defined, strict identifiability and up-to-label-swapping identifiability. Identifiability properties of the binary models will be discussed in Section 2.2. In Section 2.3, these properties will be extended to the Poisson models. Finally, a result on the identifiability of the affiliation model will be given in Section 2.4.

2.1 Definitions : strict identifiability and up-to-label-swapping identifiability

The standard definition of identifiability for a discrete parametric model is given next. This property will be also referred to as **strict** identifiability. More details on this definition can be found in the works of Rothenberg (1971) and Lehmann & Casella (1998).

Definition *(Rothenberg, 1971; Lehmann & Casella, 1998): Let $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi}\in\Xi})$ be a discrete probability space, where $\Xi \subset \mathbb{R}^d$ is a parameter space. Let $f(\boldsymbol{x}|\boldsymbol{\xi})$ denote the probability mass function induced by the probability measure $\mathbb{P}_{\boldsymbol{\xi}}$. The parameter $\xi_0 \in \Xi$ is said to be (strictly) identifiable if the following condition is satisfied:*

$$f(\boldsymbol{x}|\boldsymbol{\xi}) = f(\boldsymbol{x}|\boldsymbol{\xi}_0) \quad \forall \boldsymbol{x} \in \mathcal{X} \implies \xi = \xi_0 \quad \forall \xi \in \Xi. \tag{2.1}$$

If such a condition is not satisfied, ξ_0 is said to be (strictly) unidentifiable.

Therefore, a parameter ξ_0 is identifiable if no other parameter yields the same distribution.

Unfortunately, for the stochastic block model, since only $\boldsymbol{Y}$ is observed in practice, unidentifiability occurs. Indeed, let us remind that the model supposes that K classes exist, and therefore the label set is $[K] = \{1, 2, ...K\}$, and the class-membership space of the whole network is $\mathcal{Z} = [K]^n$. However, the mapping from the classes to the labels is ambiguous if we only observe $\boldsymbol{Y}$. Indeed, we will show next that any parametric couple obtained by simply permuting the labels (in a sense that will be next explained) yields the same distribution of $\boldsymbol{Y}$. This property is intuitively trivial, but it will allow us to define later a more precise criterion than strict identifiability, that will be called **up-to-label-swapping** identifiability.

Let $Sym(K)$ denote the set of the permutations of the labels $\{1, ..., K\}$, ie. the set of the bijections, or one-to-one mappings, going from $\{1, ..., K\}$ to $\{1, ..., K\}$:

$$Sym(K) := \{\sigma : \{1, ..., K\} \to \{1, ..., K\} | \sigma \text{ bijective}\}. \tag{2.2}$$

Intuitively, permuting the labels corresponds to swapping the labels. For example, if we have 2 classes labelled by 2 colours, the first class being labelled in red and the second class in blue, we may decide to swap the labels and labelling then the first class in blue and the second class in red. Since labels do not have a meaning in themselves, such label swapping is totally legitimate. Note that such label swapping corresponds to the permutation σ defined by $\sigma(1) = 2$, $\sigma(2) = 1$. In general, the number of all possible permutations of the label set $\{1, 2, ...K\}$ is $K!$.
Let also $Id_K \in Sym(K)$ denote the trivial permutation defined by $Id_K(k) = k \quad \forall k \in \{1, ..., K\}$, and $\sigma \neq Id_K \in Sym(K)$ denote any other non-trivial permutation of the labels. For any parametric couple $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in \Theta^{K \times K} \times \Pi_K$, we define $(\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma) \in \Theta^{K \times K} \times \Pi_K$ as the parametric couple obtained by permuting, or *swapping*, the labels:

$$(\boldsymbol{\pi}_\sigma)_k := (\boldsymbol{\pi})_{\sigma(k)}$$
$$(\boldsymbol{\theta}_\sigma)_{kl} := (\boldsymbol{\theta})_{\sigma(k)\sigma(l)}.$$

Most often, we have that the parametric couple obtained by swapping the labels is different from the original parametric couple: $(\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma) \neq (\boldsymbol{\theta}, \boldsymbol{\pi})$. However, there exists a few parametric couples $(\boldsymbol{\theta}, \boldsymbol{\pi})$ such that $(\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma) = (\boldsymbol{\theta}, \boldsymbol{\pi})$ for a non-trivial permutation σ, implying that there exist some classes having both the same connectivity properties and the same weight.
Let $\mathcal{T} \subset \Theta^{K \times K} \times \Pi_K$ denote the subset of such parametric couples:

$$\mathcal{T} := \{(\boldsymbol{\theta}, \boldsymbol{\pi}) \in \Theta^{K \times K} \times \Pi_K \quad | \exists \sigma \neq Id_K \in Sym(K) \mid (\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma) = (\boldsymbol{\theta}, \boldsymbol{\pi})\}. \tag{2.3}$$

$\mathcal{T}$ is a negligible subset of the whole parameters space $\Theta^{K \times K} \times \Pi_K$; indeed, $\mathcal{T}$ has a null Lebesgue measure (Celisse et al., 2012).

Theorem 3. *Let us consider the probability space for the observed (directed or undirected) stochastic block model* $(\mathcal{Y}, \mathcal{P}(\mathcal{Y}), \{\mathbb{P}_{(\boldsymbol{\theta},\boldsymbol{\pi})}\}_{(\boldsymbol{\theta},\boldsymbol{\pi}) \in \Theta^{K \times K} \times \Pi_K})$. *Then, any parametric couple* $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in (\Theta^{K \times K} \times \Pi_K) \setminus \mathcal{T}$ *is strictly unidentifiable.*

Proof. The following proof concerns the general directed model. For the general undirected model, the structure of the proof is exactly the same. Let us consider a parametric couple $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in (\Theta^{K \times K} \times \Pi_K) \setminus \mathcal{T}$, and a (non-trivial) permutation of the labels $\sigma \neq Id_K \in Sym(K)$, so that $(\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma) \neq (\boldsymbol{\theta}, \boldsymbol{\pi})$. We will show that the marginal probability mass function of $\boldsymbol{Y}$ under $(\boldsymbol{\theta}, \boldsymbol{\pi})$ is exactly equal to this same probability mass function under $(\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma)$:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma) = f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) \quad \forall \boldsymbol{y} \in \mathcal{Y}, \tag{2.4}$$

which contradicts the definition of strict identifiability (see equation 2.1). Indeed, let $\sigma_{\mathcal{Z}}$ denote the $\mathcal{Z} \longrightarrow \mathcal{Z}$ bijection defined by:

$$(\sigma_{\mathcal{Z}}(\boldsymbol{z}))_{ik} := \boldsymbol{z}_{i\sigma(k)} \quad \forall \boldsymbol{z} \in \mathcal{Z}, i \in [n], k \in [K]. \tag{2.5}$$

This bijection simply swaps the class of each node in the network according to σ. Note that the partition of the network is unchanged. By using the general formula of the probability mass function of $(\boldsymbol{Y}, \boldsymbol{Z})$ given in equation (1.22), it is easily seen that the two joint probability mass functions $f_{\boldsymbol{Y},\boldsymbol{Z}}(\cdot, \cdot|\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma)$ and $f_{\boldsymbol{Y},\boldsymbol{Z}}(\cdot, \cdot|\boldsymbol{\theta}, \boldsymbol{\pi})$, are equal up to the application of $\sigma_{\mathcal{Z}}$:

$$
\begin{aligned}
f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\sigma_{\mathcal{Z}}(\boldsymbol{z})|\boldsymbol{\theta}_\sigma,\boldsymbol{\pi}_\sigma) &= \prod_{i\in[n]}\prod_{k\in[K]} \pi_{\sigma(k)}^{z_{i\sigma(k)}} \prod_{\substack{(i,j)\in[n]^2\\ i\neq j}} \prod_{(k,l)\in[K]^2} g(y_{ij}|\theta_{\sigma(k)\sigma(l)})^{z_{i\sigma(k)}z_{j\sigma(l)}} \\
&= \prod_{i\in[n]}\prod_{k\in[K]} \pi_{k}^{z_{ik}} \prod_{\substack{(i,j)\in[n]^2\\ i\neq j}} \prod_{(k,l)\in[K]^2} g(y_{ij}|\theta_{kl})^{z_{ik}z_{jl}} \\
&= f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\pi}) \qquad\qquad \forall \boldsymbol{y}\in\mathcal{Y}, \boldsymbol{z}\in\mathcal{Z}.
\end{aligned}
$$

Therefore, the marginal probability mass functions of $\boldsymbol{Y}$ under $(\boldsymbol{\theta},\boldsymbol{\pi})$ and under $(\boldsymbol{\theta}_\sigma,\boldsymbol{\pi}_\sigma)$ are equivalent:

$$
\begin{aligned}
f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}_\sigma,\boldsymbol{\pi}_\sigma) &= \sum_{\boldsymbol{z}\in\mathcal{Z}} f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta}_\sigma,\boldsymbol{\pi}_\sigma) \\
&= \sum_{\boldsymbol{z}\in\mathcal{Z}} f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\sigma_{\mathcal{Z}}(\boldsymbol{z})|\boldsymbol{\theta}_\sigma,\boldsymbol{\pi}_\sigma) \\
&= \sum_{\boldsymbol{z}\in\mathcal{Z}} f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\pi}) \\
&= f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\pi}) \qquad\qquad \forall \boldsymbol{y}\in\mathcal{Y},
\end{aligned}
$$

which contradicts the definition of strict identifiability, see (2.1).

□

For the undirected case, the proof is the same, provided that the relation "$i \neq j$" is replaced with "$i < j$".

Since parametric couples that are equivalent up-to-label-swapping yield the same distribution of $\boldsymbol{Y}$, another property, less strict than identifiability, must be sought. This property, called up-to-label-swapping identifiability, will be now rigorously defined and discussed.

Up-to-label-swapping identifiability

Up-to-label-swapping identifiability of the stochastic block models was thoroughly discussed in the works from Allman et al. (2009), Latouche et al. (2011) and Celisse et al. (2012). This concept is also used in Hidden Markov Model theory; for example in the work from Cappé et al. (2005). (In the works from Celisse et al. (2012) and Cappé et al. (2005), the expression label *switching* rather than label *swapping* is used).

For any two parametric couples of a stochastic block model $(\boldsymbol{\theta},\boldsymbol{\pi}) \in \Theta^{K\times K}\times\Pi_K$ and $(\boldsymbol{\theta}',\boldsymbol{\pi}') \in \Theta^{K\times K}\times\Pi_K$, let $(\boldsymbol{\theta}',\boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta},\boldsymbol{\pi})$ denote the *equivalence relation* defined by:

$$
(\boldsymbol{\theta}',\boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta},\boldsymbol{\pi}) \iff (\boldsymbol{\theta}',\boldsymbol{\pi}') = (\boldsymbol{\theta}_\sigma,\boldsymbol{\pi}_\sigma) \text{ for some permutation } \sigma\in Sym(K).
$$

When two parametric couples $(\boldsymbol{\theta}',\boldsymbol{\pi}')$ and $(\boldsymbol{\theta},\boldsymbol{\pi})$ satisfy the equivalence relation $(\boldsymbol{\theta}',\boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta},\boldsymbol{\pi})$, we say that $(\boldsymbol{\theta}',\boldsymbol{\pi}')$ and $(\boldsymbol{\theta},\boldsymbol{\pi})$ are *equivalent up-to-label-swapping*.

In Appendix 7.4, it is proven that the relation $\overset{swap}{\sim}$ is *reflexive*, *symmetric* and *transitive*. It is then called an *equivalence relation*. We can then denote by $[(\boldsymbol{\theta},\boldsymbol{\pi})]_{swap}$ the *equivalence class*, determined by $\overset{swap}{\sim}$, of the parametric couple $(\boldsymbol{\theta},\boldsymbol{\pi})$:

$$[(\boldsymbol{\theta}, \boldsymbol{\pi})]_{swap} := \{(\boldsymbol{\theta}', \boldsymbol{\pi}') \in \Theta^{K \times K} \times \Pi_K \mid (\boldsymbol{\theta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\theta}', \boldsymbol{\pi}')\},$$

which is the set of all parametric couples being equivalent, up to $\overset{swap}{\sim}$, to the parametric couple $(\boldsymbol{\theta}, \boldsymbol{\pi})$. (Note that the *equivalence classes* as they were just defined are not to be confused with the classes of the nodes). By general properties of the equivalence classes, it can also be proven (see Appendix 7.4) that the set of all distinct equivalence classes determined by $\overset{swap}{\sim}$ is a partition of $\Theta^{K \times K} \times \Pi_K$.

A rigorous definition of up-to-label-swapping identifiability can now be given. In Theorem 3 it has been proven that two couples of parameters $(\boldsymbol{\theta}, \boldsymbol{\pi})$, $(\boldsymbol{\theta}', \boldsymbol{\pi}')$ belonging to the same equivalence class, ie. such that $(\boldsymbol{\theta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\theta}', \boldsymbol{\pi}')$, yield the same distribution of $\boldsymbol{Y}$ (see equation 2.4). Therefore, we will say that a parametric couple $(\boldsymbol{\theta}, \boldsymbol{\pi})$, is identifiable up-to-label-swapping if no parametric couple outside of its equivalence class $[(\boldsymbol{\theta}, \boldsymbol{\pi})]_{swap}$ yields the same distribution of $\boldsymbol{Y}$:

Definition *(Latouche et al., 2011): Let us consider the probability space for the observed (directed or undirected) stochastic block model* $(\mathcal{Y}, \mathscr{P}(\mathcal{Y}), \{\mathbb{P}_{(\boldsymbol{\theta},\boldsymbol{\pi})}\}_{(\boldsymbol{\theta},\boldsymbol{\pi}) \in \Theta^{K \times K} \times \Pi_K})$. *The parametric couple* $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in \Theta^{K \times K} \times \Pi_K$ *is identifiable up-to-label-swapping if:*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}', \boldsymbol{\pi}') = f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) \quad \forall \boldsymbol{y} \in \mathcal{Y} \implies [(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap} = [(\boldsymbol{\theta}, \boldsymbol{\pi})]_{swap} \quad \forall (\boldsymbol{\theta}', \boldsymbol{\pi}') \in \Theta^{K \times K} \times \Pi_K.$$

In the following paragraphs, three theorems concerning the identifiability of the binary models will be given. Each of these theorems shows that, under conditions on the number of nodes n and the number of classes K, and the condition that all classes have strictly positive weights $\pi_k > 0$, the parameters of the binary model are identifiable except on a negligible subset of the parameter space. Let $\mathcal{S}_K \subset [0,1]^{K \times K} \times \Pi_K$ denote the parameter space in which all classes have strictly positive weights:

$$\mathcal{S}_K := \{(\boldsymbol{\eta}, \boldsymbol{\pi}) \in [0,1]^{K \times K} \times \Pi_K \mid \pi_k > 0 \ \forall k \in [K]\}.$$

Restricting the parameter space to $\mathcal{S}_K$ actually does not imply any real restriction, since if $\pi_k = 0$ for some class k we can simply consider the block model with a smaller value of K. The following theorems affirm that, under conditions on n and K, every parametric couple $(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \mathcal{S}_K$ is identifiable, except for a negligible subset of $\mathcal{S}_K$. When a property holds in a parameter space, except on a negligible subset of it, ie. a subset having null Lebesgue measure, we will say that the property holds *generically* (Celisse et al., 2012).

2.2 Identifiability for the binary model

Theorem 4 below concerns the undirected binary block model in the case where the number of classes is $K = 2$. It was the first result stating conditions on the identifiability of stochastic block models to appear in the literature. Theorem 5 also concerns the undirected model but it is much more general than the first, and it also considers the scenario in which the parameter $\boldsymbol{\pi}$ is known. Theorem 6 concerns both the directed and the undirected binary model.

Theorem 4. *(Allman et al., 2009) : Let us consider the undirected binary stochastic block model with $n \geq 16$ nodes, $K = 2$ classes, and parameter space $\mathcal{W}_K := \{(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \mathcal{S}_K \mid \eta_{11}, \eta_{12}, \eta_{22}$ are distinct $\}$. Then, the parametric couple $(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \mathcal{W}_K$ is identifiable, up to label swapping.*

Therefore, for a number of nodes bigger or equal to 16, the parameters of the binary model with 2 blocks are identifiable, except on the negligible subset $\{\eta_{11} = \eta_{12} = \eta_{22}\}$ (by negligible, we mean that it has null Lebesgue measure). It is natural to exclude parameters such that $\{\eta_{11} = \eta_{12} = \eta_{22}\}$, because such parameters actually mean that the two classes are equivalent. If such is the case, then any values of the weights parameters in $\boldsymbol{\pi}$ yield the same model, which implies unidentifiability.

Sketch of the proof of Theorem 4: A short sketch of the proof of Theorem 4, as given by Allman et al. (2009), is now given. More details on this proof can be found in the Appendix 7.5. The proof is divided in two steps. The first step of the proof consists in considering a network of 4 nodes, and defining the matrix B as the 16×64 matrix in which each row corresponds to one of the $2^4 = 16$ possible class assigments, and each column to one of the $2^{\binom{4}{2}} = 64$ possible subgraphs on 4 nodes, (each is a subgraph of the complete graph on 4 nodes, as defined in Section 1.3, which has $\binom{4}{2}$ edges), with each entry of B giving the probability of observing a subgraph conditional on a class assignment. The matrix B is shown to be full-rank. The author infers from this, in the second step, that for a network of 16 nodes, it is possible to find 3 edge-disjoint subgraphs of the complete graph on 16 nodes, denoted by G_1, G_2 and G_3, each G_i having $4\binom{4}{2} = 32$ edges, so that for each subgraph G_i, the matrix B_i, ie. the matrix with 2^{16} rows, each row corresponding to a class assignment, and $2^{4\binom{4}{2}} = 2^{32}$ columns corresponding to the subgraphs of G_i, and entries corresponding to probabilities of observing a subgraph conditional on a class assignments, is full-rank too. For constructing G_1, G_2 and G_3, we assume the 16 nodes were arranged in a 4×4 square grid. G_1 is found considering the 4 rows of such grid, each row yielding a fraction of 4 nodes. For each row, the complete graph on such fraction is considered. G_1 is then defined as the union of such 4 complete graphs. Similarly, G_2 is defined by considering the columns of the grid, and G_3 by considering the diagonals. The fact that matrices B_1, B_2 and B_3 are full-rank allows the author to conclude that the parameters of the model are identifiable up-to-label-swapping.

Theorem 5 below is much more general than Theorem 4. It states that if we restrict the parameter space to $\mathcal{S}_K$, stochastic block models are generically identifiable up-to-label-swapping, ie. identifiable up-to label-swapping everywhere except on a subspace of null Lebesgue measure.

Theorem 5. *(Allman et al., 2011) : Let us consider the undirected binary stochastic block model with K classes, n nodes, with $n \geq \left(K - 1 + \frac{(K+2)^2}{4}\right)^2$ if K is even (or $n \geq \left(K - 1 + \frac{(K+1)(K+3)}{4}\right)^2$ if K is uneven), and parameters space $\mathcal{V}_K := \mathcal{S}_K$. Then, (i) the parametric couple $(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \mathcal{S}_K$ is generically identifiable up-to-label-swapping. Furthermore, (ii) the result remains valid if $\boldsymbol{\pi}$ is fixed, ie. if we restrict the parameters space to $\mathcal{V}'_K := \mathcal{V}_K \cap \{\boldsymbol{\pi} = \boldsymbol{\pi}_0\}$.*

The structure of the proof of Theorem 5 is similar to that of Theorem 4, however it requires some additional technical steps that will not be discussed in this thesis.

Theorem 5 is made of two statements; the first (i) ensuring generic up-to-label-swapping identifiability on the parametric space $\mathcal{V}_K$, the second (ii) ensuring generic up-to-label-swapping

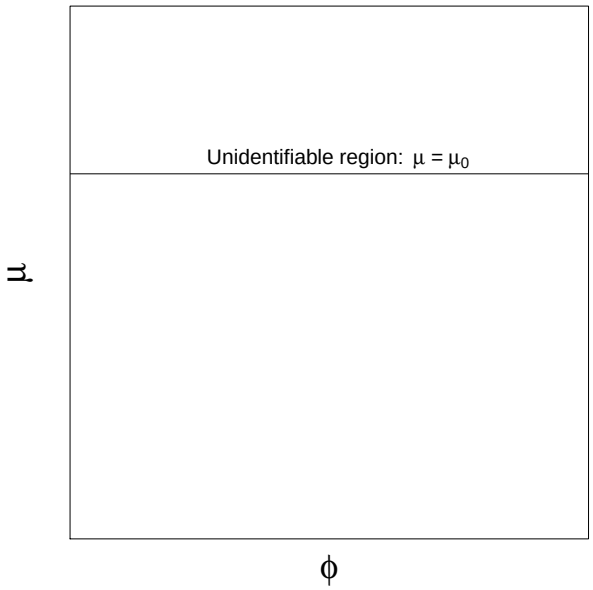


Figure 2.1: 2-*dimensional representation of a parameter space, with parameters* $\boldsymbol{\phi}$ *and* $\boldsymbol{\mu}$*, in which the unidentifiable region is the horiziontal line* $\boldsymbol{\mu} = \boldsymbol{\mu}_0$*. When* $\boldsymbol{\mu} = \boldsymbol{\mu}_0$*,* $\boldsymbol{\phi}$ *is totally unidentifiable. The whole parameter space is however generally identifiable because the region* $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ *has null Lebesgue measure in dimension 2 (intuitively, this means that the horizontal line has null area). Theorem 5 guarantees that this kind of scenario is impossible in the parameter space of the undirected binary SBM.*

identifiability when we restrict the parametric space to $\mathcal{V}_K \cap \{\boldsymbol{\pi} = \boldsymbol{\pi}_0\}$. The last one is an interesting property too because previous knowledge on parameter $\boldsymbol{\pi}$ may be available in practice (however, often it is not). Remark that the second statement is not a trivial result of the first, since '*fixing values of the* $\boldsymbol{\pi}$ *results in considering a subvariety of the full parameter space, which a-priori might be included in the subvariety of nonidentifiable parameters*' (Allman et al., 2011). A graphical representation of a parameter space which satisfies statement (i), but not statement (ii), is given in Figure 2.1. Then, a graphical representation of the parameter space of the undirected binary SBM is given in Figure 2.2.

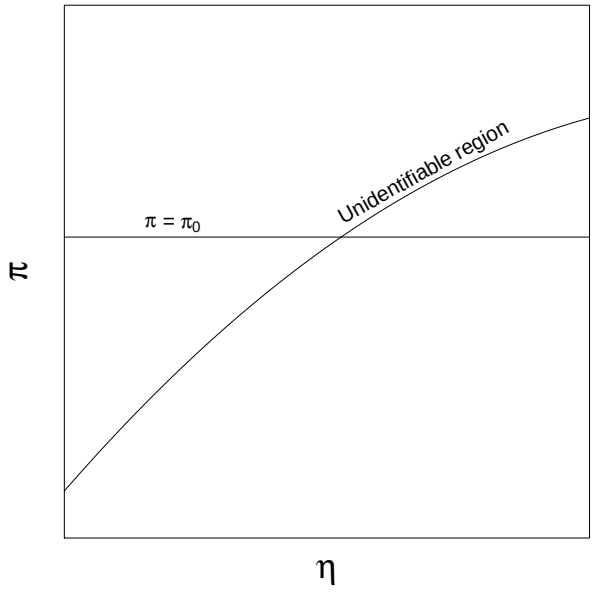


Figure 2.2: 2-*dimensional representation of the parameter space of the undirected binary SBM, with the horizontal axis representing values of* $\boldsymbol{\eta}$*, and the vertical axis representing values of* $\boldsymbol{\pi}$ *(these are arrays, and so they are not one-dimensional, but they are represented as if they were both one-dimensional). The unidentifiability region is represented as a curve. The parameter space is generally identifiable, and the subspace* $\boldsymbol{\pi} = \boldsymbol{\pi}_0$ *is generally identifiable for any choice of* $\boldsymbol{\pi}_0$*. Note that the unidentifiability region may also consists of several curves.*

Theorems 4 and 5 only concern undirected models. Theorem 6 concerns both directed and undirected models, and provides conditions on strict identifiability:

Theorem 6. *(Celisse et al., 2012) : Let us consider the (directed or undirected) binary stochastic block model with K classes, and n nodes, with $n \geq 2K$, and parameters space $\mathcal{U}_K := \{(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \mathcal{S}_K | \rho_1 < \rho_2 ... < \rho_K$, with $\boldsymbol{\rho} \in \mathbb{R}^K$ defined by $\boldsymbol{\rho} := \boldsymbol{\eta}\boldsymbol{\pi}\}$. Then, the parametric couple $(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \mathcal{U}_K$ is strictly identifiable.*

Vector $\boldsymbol{\rho} = \boldsymbol{\eta}\boldsymbol{\pi}$ has a very simple interpretation. The component ρ_k represents the expected value of the edge variable Y_{ij} $i \neq j$ conditionally on node i belonging to class k, ie. the probability that a node of class k is connected to any other edge. Indeed:

$$\begin{aligned}
\mathbb{P}(Y_{ij} = 1 | Z_i = k) &= \frac{\mathbb{P}(Y_{ij} = 1, Z_i = k)}{\mathbb{P}(Z_i = k)} \\
&= \frac{\sum_{l \in [K]} \mathbb{P}(Y_{ij} = 1, Z_i = k, Z_j = l)}{\mathbb{P}(Z_i = k)} \\
&= \frac{\sum_{l \in [K]} \mathbb{P}(Y_{ij} = 1 | Z_i = k, Z_j = l) \mathbb{P}(Z_i = k, Z_j = l)}{\mathbb{P}(Z_i = k)} \\
&= \frac{\sum_{l \in [K]} \mathbb{P}(Y_{ij} = 1 | Z_i = k, Z_j = l) \mathbb{P}(Z_i = k) \mathbb{P}(Z_j = l)}{\mathbb{P}(Z_i = k)} \\
&= \sum_{l \in [K]} \mathbb{P}(Y_{ij} = 1 | Z_i = k, Z_j = l) \mathbb{P}(Z_j = l) \\
&= \sum_{l \in [K]} \eta_{kl} \pi_l \\
&= \rho_k.
\end{aligned}$$

The ordering condition $\rho_1 < \rho_2 ... < \rho_K$ can be easily interpreted. Firstly, this condition implies that no rows of $\boldsymbol{\eta}$ are exactly equal, as in Theorem 4. Furthermore, it implies that if different classes have the same connectivity properties, then their weights must be different. Indeed, let us suppose that there exist some classes having the same connectivity properties, or in other words, that for the parameters matrix $\boldsymbol{\eta}$ there exists a non-trivial label swapping, denoted $\sigma^{\boldsymbol{\eta}}$, that does not change $\boldsymbol{\eta}$:

$$\exists \sigma^{\boldsymbol{\eta}} \neq Id_K \in Sym(K) \ | \ \boldsymbol{\eta}_{\sigma^{\boldsymbol{\eta}}} = \boldsymbol{\eta}, \tag{2.6}$$

(we can also suppose that $\sigma^{\boldsymbol{\eta}}$ is the label swapping such that $\boldsymbol{\eta}_{\sigma^{\boldsymbol{\eta}}} = \boldsymbol{\eta}$, having a maximum support, ie. such that a maximum number of labels is actually switched). Then, we have that:

$$\begin{aligned}
\boldsymbol{\rho}_{\sigma^{\boldsymbol{\eta}}} &:= (\boldsymbol{\eta}\boldsymbol{\pi})_{\sigma^{\boldsymbol{\eta}}} \\
&= \boldsymbol{\eta}_{\sigma^{\boldsymbol{\eta}}} \boldsymbol{\pi}_{\sigma^{\boldsymbol{\eta}}} \\
&= \boldsymbol{\eta}\boldsymbol{\pi}_{\sigma^{\boldsymbol{\eta}}}.
\end{aligned}$$

And so, we must have $\boldsymbol{\pi}_{\sigma^{\boldsymbol{\eta}}} \neq \boldsymbol{\pi}$, otherwise we would have $\boldsymbol{\rho}_{\sigma^{\boldsymbol{\eta}}} = \boldsymbol{\eta}\boldsymbol{\pi}_{\sigma^{\boldsymbol{\eta}}} = \boldsymbol{\eta}\boldsymbol{\pi} = \boldsymbol{\rho}$, which contradicts the ordering condition $\rho_1 < \rho_2 ... < \rho_K$ of the theorem.
The proof of Theorem 6 is based on some quite complex and clever linear algebra reasoning, using Vandermonde matrices.

As another proof of the relationships between strict and up-to-label-swapping identifiability, Theorem 6 implies Corollary 1, stating that if we relax the strict ordering constraint $\rho_1 < \rho_2 ... < \rho_K$ (but we still assume elements of $\boldsymbol{\rho}$ to be pairwise distinct), the model is still identifiable up-to-label-swapping.

Corollary 1. *Let us consider the (directed or undirected) binary stochastic block model with K classes, and n nodes, with $n \geq 2K$, and parameters space $\tilde{\mathcal{U}}_K := \{(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \mathcal{S}_K \mid$ the elements of the vector $\boldsymbol{\rho} := \boldsymbol{\eta}\boldsymbol{\pi} \in \mathbb{R}^K$ are pairwise distinct $\}$. Then, the parametric couple $(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \tilde{\mathcal{U}}_K$ is identifiable up-to-label-swapping.*

Proof. Let us consider two parametric couples $(\boldsymbol{\eta}, \boldsymbol{\pi}) \in \tilde{\mathcal{U}}_K$ and $(\bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\pi}}) \in \tilde{\mathcal{U}}_K$, such that

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\pi}) = f_{\boldsymbol{Y}}(\boldsymbol{y}|\bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\pi}}) \quad \forall \boldsymbol{y} \in \mathcal{Y}, \tag{2.7}$$

with $\boldsymbol{\rho} := \boldsymbol{\eta}\boldsymbol{\pi}$ and $\bar{\boldsymbol{\rho}} := \bar{\boldsymbol{\eta}}\bar{\boldsymbol{\pi}}$. By definition of $\tilde{\mathcal{U}}_K$, there exist permutations $\sigma \in Sym(K)$ and $\bar{\sigma} \in Sym(K)$ such that:

$$\rho_{\sigma(1)} < \rho_{\sigma(2)} < ... < \rho_{\sigma(K)},$$

$$\bar{\rho}_{\bar{\sigma}(1)} < \bar{\rho}_{\bar{\sigma}(2)} < ... < \bar{\rho}_{\bar{\sigma}(K)}.$$

Furthermore, it is trivial to verify that $\boldsymbol{\eta}_\sigma\boldsymbol{\pi}_\sigma = \boldsymbol{\rho}_\sigma$ and $\bar{\boldsymbol{\eta}}_{\bar{\sigma}}\bar{\boldsymbol{\pi}}_{\bar{\sigma}} = \bar{\boldsymbol{\rho}}_{\bar{\sigma}}$. As such, $(\boldsymbol{\eta}_\sigma, \boldsymbol{\pi}_\sigma) \in \mathcal{U}_K$ and $(\bar{\boldsymbol{\eta}}_{\bar{\sigma}}, \bar{\boldsymbol{\pi}}_{\bar{\sigma}}) \in \mathcal{U}_K$. Furthermore, we have that :

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\pi}) = f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\eta}_\sigma, \boldsymbol{\pi}_\sigma) \quad \forall \boldsymbol{y} \in \mathcal{Y} \tag{2.8}$$

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\pi}}) = f_{\boldsymbol{Y}}(\boldsymbol{y}|\bar{\boldsymbol{\eta}}_{\bar{\sigma}}, \bar{\boldsymbol{\pi}}_{\bar{\sigma}}) \quad \forall \boldsymbol{y} \in \mathcal{Y} \tag{2.9}$$

since $(\boldsymbol{\eta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\eta}_\sigma, \boldsymbol{\pi}_\sigma)$ and $(\bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\pi}}) \overset{swap}{\sim} (\bar{\boldsymbol{\eta}}_{\bar{\sigma}}, \bar{\boldsymbol{\pi}}_{\bar{\sigma}})$.
By equations (2.7), (2.8) and (2.9) one infers that :

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\eta}_\sigma, \boldsymbol{\pi}_\sigma) = f_{\boldsymbol{Y}}(\boldsymbol{y}|\bar{\boldsymbol{\eta}}_{\bar{\sigma}}, \bar{\boldsymbol{\pi}}_{\bar{\sigma}}) \quad \forall \boldsymbol{y} \in \mathcal{Y}. \tag{2.10}$$

As such, by Theorem 6, we conclude that $(\boldsymbol{\eta}_\sigma, \boldsymbol{\pi}_\sigma) = (\bar{\boldsymbol{\eta}}_{\bar{\sigma}}, \bar{\boldsymbol{\pi}}_{\bar{\sigma}})$. This finally implies that $[(\boldsymbol{\eta}, \boldsymbol{\pi})]_{swap} = [(\bar{\boldsymbol{\eta}}, \bar{\boldsymbol{\pi}})]_{swap}$ because equivalence classes partition the parameters space. □

Therefore, only parametric couples $(\boldsymbol{\eta}, \boldsymbol{\pi})$ yielding non-distinct elements of vector $\boldsymbol{\rho}$ are unidentifiable; this is a negligible subset of the parameters space.

2.3 Identifiability for the Poisson model

Theorems 4, 5, and 6 , and Corollary 1, all concerning the identifiability of the binary models, are easily extensible to the Poisson models. Indeed, for every $(\boldsymbol{Y}, \boldsymbol{Z})$ following a Poisson block model of parameters $(\boldsymbol{\pi}, \boldsymbol{\lambda})$, we can simply define the random matrix $\boldsymbol{Y}'$ as the matrix whose (i, j)-th entry is:

$$Y'_{ij} := \mathbb{1}_{\{Y_{ij}>0\}} = 1 - \mathbb{1}_{\{Y_{ij}=0\}}.$$

Then, $(\boldsymbol{Y}', \boldsymbol{Z})$ follows a binary block model of parameters $(\boldsymbol{\pi}, \boldsymbol{\eta})$, with $\eta_{kl} = 1 - e^{-\lambda_{kl}}$, as:

$$\begin{aligned}\mathbb{P}(Y'_{ij} = 1 | Z_i = k, Z_j = l) &= 1 - \mathbb{P}(Y'_{ij} = 0 | Z_i = k, Z_j = l) \\ &= 1 - \mathbb{P}(Y_{ij} = 0 | Z_i = k, Z_j = l) \\ &= 1 - e^{-\lambda_{kl}},\end{aligned}$$

and therefore one can identify $(\boldsymbol{\pi}, \boldsymbol{\lambda})$ by the distribution of $(\boldsymbol{Y}', \boldsymbol{Z})$.

For example, Theorem 6 can be extended to the Poisson model by the following statement:

Corollary 2. *Let us consider the (directed or undirected) Poisson stochastic block model with K classes and n nodes, with $n \geq 2K$, and parameters space $\{(\boldsymbol{\lambda}, \boldsymbol{\pi}) \in {\mathbb{R}_+}^{K \times K} \times \Pi_K \mid \rho_1 < \rho_2 < \ ... \ < \rho_K, \ \text{with } \boldsymbol{\rho} := \boldsymbol{\lambda}\boldsymbol{\pi}\}$. Then, the parametric couple $(\boldsymbol{\lambda}, \boldsymbol{\pi})$ is strictly identifiable.*

Remark that the parameter ρ_k represents, for the Poisson model, the average weight of an edge between a node of class k and any other node.
Similarly, extensions of Theorems 4 and 5 and Corollary 1 to the Poisson model can be deduced.

2.4 Identifiability for the affiliation model

We remind the reader that the binary affiliation model and the Poisson affiliation model (Section 1.7) are particular block models, both depending on parameters $(\boldsymbol{\pi}, p_{in}, p_{out})$, such that:

$$\theta_{ij} = \begin{cases} p_{in} & \text{if } i = j \\ p_{out} & \text{if } i \neq j. \end{cases}$$

Theorem 6 implies the following result on the identifiability of the parameters of the binary affiliation model:

Corollary 3. *Let us consider the (directed or undirected) binary affiliation model with K classes and n nodes, with $n \geq 2K$, and parameters space $\{(p_{in}, p_{out}, \boldsymbol{\pi}) \in [0,1] \times [0,1] \times \Pi_K \mid \pi_1 < \pi_2 < \ ... \ < \pi_K\}$. Then, the parametric triplet $(p_{in}, p_{out}, \boldsymbol{\pi})$ is strictly identifiable.*

Proof. By Theorem 6, it is sufficient to prove that coordinates of $\boldsymbol{\rho} = \boldsymbol{\eta}\boldsymbol{\pi}$ are strictly ordered, ie. $\rho_1 < \rho_2 < \ ... \ < \rho_K$, to prove strict identifiability. Indeed:

$$\begin{aligned}\rho_k &= \pi_k p_{in} + (1 - \pi_k) p_{out} \\ &= p_{out} + \pi_k (p_{in} - p_{out}) \quad \forall k = 1...K.\end{aligned}$$

But since π_k are strictly ordered and $p_{in} \neq p_{out}$, ρ_k are strictly ordered too.

□

Similarly, Corollary 2 implies the equivalent result for the Poisson affiliation model:

Corollary 4. *Let us consider the (directed or undirected) Poisson affiliation model with K classes and n nodes, with $n \geq 2K$, and parameters space $\{(p_{in}, p_{out}, \boldsymbol{\pi}) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \Pi_K \mid \pi_1 < \pi_2 < \ ... \ < \pi_K\}$. Then, the parametric triplet $(p_{in}, p_{out}, \boldsymbol{\pi})$ is strictly identifiable.*

The proof of Corollary 4 is almost identical to that of Corollary 3.

Chapter 3: Estimation methods

In this chapter, the maximum-likelihood estimator is discussed in Section 3.1, after which several methods for community detection are presented. The Expectation-Maximization algorithm (Barbillon et al., 2017; Bickel & Chen, 2009; Dempster et al., 1977; Latouche et al., 2011; Mariadassou et al., 2010; Nowicki & Snijders, 1997; Wu, 1983) (presented in Section 3.2) is an algorithm that has very good properties in terms of convergence to the maximum-likelihood estimator. It is extremely general, being theoretically applicable to any incomplete data model, however explicit formulas for its application have to be found for each particular model. Variational methods (Barbillon et al., 2017; Bickel & Chen, 2009; Celisse et al., 2012; Latouche et al., 2011; Mariadassou et al., 2010) (presented in Section 3.3) allow finding explicit formulas for an algorithm which approximates well the Expectation-Mazimization algorithm for stochastic block models. A Bayesian approach based on the Gibbs sampling method (Nowicki & Snijders, 1997; Latouche et al., 2009) will be presented in Section 3.4. Bayesian approaches have the advantage of yielding a criterion for selecting the number of classes of the block model. Spectral methods (Abbe, 2018; Amini et al., 2013; Lei & Rinaldo, 2015; Rohe et al., 2011) will then be illustrated in Section 3.5. These have the advantage of being faster than other methods, and some authors suggest using them as starting points for other iterative algorithms such as the Expectation-maximization algorithm; several variants of such methods will be introduced. Finally, the profile likelihood method (Bickel & Chen, 2009; Choi et al., 2012; Karrer & Newman, 2011; Zhao et al., 2012) and the modularity method (Bickel & Chen, 2009; Karrer & Newman, 2011; Zhao et al., 2012) will be illustrated; these two methods are different, but share similar properties; so, following the approach by Bickel & Chen (2009) and Zhao et al. (2012), they will both be presented at the same time in Section 3.6.

3.1 Maximum likelihood

Several methods have been studied in the literature for the estimation of the parameters of the SBM. This task is very hard because in practice $\boldsymbol{Z}$ is not observed. One of the standard methods for estimation is the maximum-likelihood technique. Only $\boldsymbol{Y} = \boldsymbol{y}$ being observed in practice, the likelihood function $\mathscr{L}$ is defined as:

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{y}) := f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\pi}).$$

It is also convenient to define the log-likelihood function ℓ:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{y}) := \log \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{y}).$$

The maximum likelihood estimator is then given by:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\pi}})_{ML} := \arg \max_{(\boldsymbol{\theta},\boldsymbol{\pi}) \in \Theta^{K \times K} \times \Pi_K} \ell(\boldsymbol{\theta}, \boldsymbol{\pi} | \boldsymbol{y}) = \arg \max_{(\boldsymbol{\theta},\boldsymbol{\pi}) \in \Theta^{K \times K} \times \Pi_K} \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\pi} | \boldsymbol{y}).$$

This approach is not directly applicable in practice for the SBM because an explicit formula for finding $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\pi}})_{ML}$ does not exist; numeric approaches are not effective because even computing ℓ for a single value of $(\boldsymbol{\theta}, \boldsymbol{\pi})$ requires to compute $|\mathcal{Z}| = K^n$ different terms, which is a huge number even for relatively small values of n and K. However, many more sophisticated methods are inspired from this standard method, starting from the expectation-maximization algorithm.

3.2 Expectation-Maximization

The expectation-maximization (EM) algorithm was proposed by Dempster et al. (1977) for the general framework of incomplete data models. In this general framework, the algorithm can be defined in two equivalent ways. Both will be given, the first because it corresponds to the first formulation given by Dempster et al. (1977) and because it will be used to show some interesting properties of the EM-algorithm for exponential families. The second definition, given in the works from Daudin et al. (2008) and Latouche et al. (2011), is offered because it will be used later in the section on variational methods. This algorithm is valid not only for the stochastic block model, but also for any incomplete-data model $(\boldsymbol{Y}, \boldsymbol{Z}) \in \mathcal{Y} \times \mathcal{Z}$ of parameter $\boldsymbol{\phi} \in \Phi$, where $\boldsymbol{Y}$ is observed and $\boldsymbol{Z}$ is unobserved.

Before giving the two equivalent definitions, let $w(\cdot|\boldsymbol{y}, \boldsymbol{\phi})$ denote the probability mass function of $\boldsymbol{Z}$ conditional on $\boldsymbol{Y}$, assuming parameter $\boldsymbol{\phi}$, ie.:

$$w(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\phi}) := \frac{f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\phi})}{f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi})} \quad \forall \boldsymbol{z} \in \mathcal{Z}, \forall \boldsymbol{y} \in \mathcal{Y}.$$

Furthermore, let $\mathcal{D}_{\mathcal{Z}}$ denote the space of the probability distributions on $\mathcal{Z}$. Given two any such probability distributions $q(\cdot) \in \mathcal{D}_{\mathcal{Z}}$ and $q'(\cdot) \in \mathcal{D}_{\mathcal{Z}}$, let finally $KL(q||q')$ denote their Kullback-Leibler divergence (Barbillon et al., 2017; Bickel & Chen, 2009; Celisse et al., 2012; Daudin et al., 2008; Latouche et al., 2011; Mariadassou et al., 2010):

$$KL(q||q') := \sum_{\substack{\boldsymbol{z} \in \mathcal{Z} \\ q(\boldsymbol{z})>0}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{q'(\boldsymbol{z})} \quad \forall q(\cdot), q'(\cdot) \in \mathcal{D}_{\mathcal{Z}}. \tag{3.1}$$

This quantity is well defined by using the conventions that any strictly positive real number divided by 0 equals $+\infty$, and that $\log(+\infty) = +\infty$. The Kullback-Leibler divergence can be interpreted as a distance between two distributions; however, it is not symmetric in the sense that $KL(q||q') \neq KL(q'||q)$ in general. The following properties of the Kullback-Leibler divergence (proved in Appendix 7.6) will be useful:

$$\begin{aligned} &KL(q||q') \geq 0 \quad \forall q(\cdot), q'(\cdot) \in \mathcal{D}_{\mathcal{Z}} \\ &KL(q||q') = 0 \iff q(\cdot) = q'(\cdot). \end{aligned} \tag{3.2}$$

Expectation-Maximization algorithm - First definition

For any couple of parameters $\boldsymbol{\phi}, \boldsymbol{\phi}' \in \Phi$, let $\mathcal{R}$ be defined as:

$$\mathcal{R}(\boldsymbol{\phi}'|\boldsymbol{\phi}) := \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\phi},\boldsymbol{y}} \log f(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{\phi}') = \sum_{\boldsymbol{z}\in\mathcal{Z}} \log f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\phi}') w(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\phi}) \quad \forall \boldsymbol{\phi}, \boldsymbol{\phi}' \in \Phi.$$

The Expectation-Maximization algorithm (Dempster et al., 1977; Wu, 1983) is defined then as the algorithm starting from an initial value $\boldsymbol{\phi}_{(0)}$ and then, given $\boldsymbol{\phi}_{(m)}$, computes iteratively $\boldsymbol{\phi}_{(m+1)}$ by the following steps:

$$\begin{aligned} &\text{Expectation-step: Compute } \mathcal{R}(\boldsymbol{\phi}|\boldsymbol{\phi}_{(m)}) \\ &\text{Maximization-step: Compute } \boldsymbol{\phi}_{(m+1)} = \arg\max_{\boldsymbol{\phi}\in\Phi}(\mathcal{R}(\boldsymbol{\phi}|\boldsymbol{\phi}_{(m)})). \end{aligned}$$

Expectation-Maximization Algorithm - Second definition

For any distribution $q(\cdot) \in \mathcal{D}_{\mathcal{Z}}$ and any parameter $\boldsymbol{\phi} \in \Phi$, let $\mathcal{J}(\cdot)$ be defined as:

$$\mathcal{J}(q, \boldsymbol{\phi}) := \ell(\boldsymbol{\phi}|\boldsymbol{y}) - KL(q||w(\cdot|\boldsymbol{y}, \boldsymbol{\phi})) \quad \forall q(\cdot) \in \mathcal{D}_{\mathcal{Z}}, \forall \boldsymbol{\phi} \in \Phi.$$

Properties of Kullback-Leibler divergence (see equation 3.2) imply that:

$$\begin{aligned} &KL(q||w(\cdot|\boldsymbol{y}, \boldsymbol{\phi})) \geq 0 \quad \forall q(\cdot) \in \mathcal{D}_{\mathcal{Z}} \\ &KL(q||w(\cdot|\boldsymbol{y}, \boldsymbol{\phi})) = 0 \iff q(\cdot) = w(\cdot|\boldsymbol{y}, \boldsymbol{\phi}), \end{aligned}$$

and therefore:

$$\arg\max_{q(\cdot)\in\mathcal{D}_{\mathcal{Z}}} \mathcal{J}(q, \boldsymbol{\phi}) = \arg\min_{q(\cdot)\in\mathcal{D}_{\mathcal{Z}}} KL(q||w(\cdot|\boldsymbol{y}, \boldsymbol{\phi})) = w(\cdot|\boldsymbol{y}, \boldsymbol{\phi}).$$

In the works of Daudin et al. (2008); Latouche et al. (2011), the EM-algorithm is defined as the algorithm starting from an initial value $\boldsymbol{\phi}_{(0)}$ and then, given $\boldsymbol{\phi}_{(m)}$, computes iteratively $\boldsymbol{\phi}_{(m+1)}$ by the following steps:

$$\begin{aligned} &\text{Expectation-step: compute } q_{(m+1)} := \arg\max_{q(\cdot)\in\mathcal{D}_{\mathcal{Z}}} \mathcal{J}(q, \boldsymbol{\phi}_{(m)}) = w(\cdot|y, \boldsymbol{\phi}_{(m)}) \\ &\text{Maximization-step: compute } \boldsymbol{\phi}_{(m+1)} := \arg\max_{\boldsymbol{\phi}\in\Phi} \mathcal{J}(q_{(m+1)}, \boldsymbol{\phi}). \end{aligned} \tag{3.3}$$

Equivalence of the two defnitions of the EM algorithm

Results from Dempster et al. (1977) imply that:

$$\arg\max_{\boldsymbol{\phi}\in\Phi} \mathcal{R}(\boldsymbol{\phi}|\boldsymbol{\phi}_{(m)}) = \arg\max_{\boldsymbol{\phi}\in\Phi} \mathcal{J}(w(\cdot|\boldsymbol{y}, \boldsymbol{\phi}_{(m)}), \boldsymbol{\phi}) \quad \forall \boldsymbol{\phi}_{(m)} \in \Phi.$$

And therefore, the two versions of the EM algorithm as given above are equivalent, yielding the same sequence $\{\boldsymbol{\phi}_{(m)}\}_{m=0,1,2,\ldots}$ for any starting $\boldsymbol{\phi}_{(0)}$.

EM algorithm characterization for the exponential families
In the case of exponential families models, whose complete log-mass function is:

$$\log f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\phi}) = \boldsymbol{\phi}^T t(\boldsymbol{y}, \boldsymbol{z}) - a(\boldsymbol{\phi}) + b(\boldsymbol{y}, \boldsymbol{z}),$$

$\mathcal{R}(\cdot|\cdot)$ is given by:

$$\begin{aligned}\mathcal{R}(\boldsymbol{\phi}'|\boldsymbol{\phi}) &:= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\phi},\boldsymbol{y}} \log f(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{\phi}') \\ &= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\phi},\boldsymbol{y}}(\boldsymbol{\phi}'^T t(\boldsymbol{y}, \boldsymbol{Z}) - a(\boldsymbol{\phi}') + b(\boldsymbol{y}, \boldsymbol{Z})) \\ &= \boldsymbol{\phi}'^T \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\phi},\boldsymbol{y}} t(\boldsymbol{y}, \boldsymbol{Z}) - a(\boldsymbol{\phi}') + \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\phi},\boldsymbol{y}} b(\boldsymbol{y}, \boldsymbol{Z}).\end{aligned}$$

Dempster et al. (1977) showed that the algorithm is equivalent to the following: start from an initial value $\boldsymbol{\phi}_{(0)}$ and then perform iteratively the steps:

Expectation-step: Compute $t_{(m)} := \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\phi}_{(m)},\boldsymbol{y}} t(\boldsymbol{y}, \boldsymbol{Z})$
Maximization-step: Compute $\boldsymbol{\phi}_{(m+1)} = \arg\max_{\boldsymbol{\phi}}(\boldsymbol{\phi}^T t_{(m)} - a(\boldsymbol{\phi}))$.

Therefore, $\boldsymbol{\phi}_{(n+1)}$ is the maximum-likelihood estimator for the complete data under the assumption that $t(\boldsymbol{y}, \boldsymbol{Z}) = t_{(m)}$. Properties of the exponential families discussed in Chapter 1 imply then that the Maximization-step is also equivalenty given by the following formula, provided that the maximum-likelihood estimator does not lie on the boundary of the parameters space:

Maximization-step: Compute $\boldsymbol{\phi}_{(m+1)}$ as the solution of $\mathbb{E}_{\boldsymbol{\phi}} t(\boldsymbol{Y}, \boldsymbol{Z}) = \boldsymbol{t}_{(m)}$.

In the general scenario of incomplete-data models, it has been proven that the likelihood increases or, at worst, is unchanged at each iteration of the algorithm. (Dempster et al., 1977, Theorem 1). Wu (1983) proved that for any exponential family model with compact natural parameters space, any accumulation point $\boldsymbol{\phi}^*$ of the sequence of estimates $\{\boldsymbol{\phi}_{(m)}\}_{m=1,2,\ldots}$, (ie. any point $\boldsymbol{\phi}^*$ such that a subsequence of $\{\boldsymbol{\phi}_{(m)}\}_{m=0,1,2,\ldots}$ converges to it) is a local maximum of the function $\boldsymbol{\phi} \to \ell(\boldsymbol{\phi}|\boldsymbol{y})$, regardless of the initial starting value. Furthermore, if the gradient $D_{\boldsymbol{\phi}'}\mathcal{R}(\boldsymbol{\phi}'|\boldsymbol{\phi})$ is continuous (which is true for an SBM), and the function $\boldsymbol{\phi} \to \ell(\boldsymbol{\phi}|\boldsymbol{y})$ is unimodal, the sequence $\{\boldsymbol{\phi}_{(m)}\}_{m=1,2,\ldots}$ converges to the unique global arg-maximum.

So, for the SBM, the expectation-maximization algorithm is guaranteed to find points that maximize locally the likelihood function; such points are the accumulation points of the sequence that the algorithm produces: however, no guarantee exists theoretically that such points are optimal globally. To have this guarantee, one should prove that the function $\boldsymbol{\phi} \to \ell(\boldsymbol{\phi}|\boldsymbol{y})$ is unimodal. Note that obviously this function can not be unimodal on the whole parameters space because of the up-to-label-swapping unidentifiability of the parameters, but it may be unimodal when restricting the parameters space to up-to-label swapping equivalence classes (see Chapter 2). Unfortunately, no general results have been found about the unimodularity of this function, even when restricting the parameters space. However, the likelihood function for the complete data model, ie. the model in which both $\boldsymbol{Y}$ and $\boldsymbol{Z}$) are observed, is indeed unimodular in the binary case (see proof in Appendix 7.10).

For the undirected binary stochastic block model, Nowicki & Snijders (1997) gave explicit formulas for both the Expectation step and the Maximization step in the case $K = 2$. Explicit formulas do not exist in general, but the variational algorithm, presented in the next

section, is approximately equivalent to the expectation-maximization algorithm: the variational algorithm follows the same steps as the expectation-maximization algorithm, if not for a slight difference in the Expectation step as was given in equation (3.3), in which rather than computing $q_{(m+1)}$, an approximation of $q_{(m+1)}$ is computed.

3.3 Variational expectation-maximization algorithm

In the previous section, we argued that explicit formulas for performing the EM-algorithm do not always exist; however a similar algorithm, the variational algorithm (Barbillon et al., 2017; Bickel & Chen, 2009; Celisse et al., 2012; Latouche et al., 2011; Mariadassou et al., 2010), will next be introduced. Such an algorithm is almost equivalent to the EM-algorithm, if not for the difference that in the variational algorithm, the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{y}$, $w(\cdot|\boldsymbol{y}, \boldsymbol{\phi})$, is approximated by a *completely factorized* distribution. Being more specific, by the *completely factorized* distribution which is closest to $w(\cdot|\boldsymbol{y}, \boldsymbol{\phi})$ in terms of KL-divergence.
A distribution $q(\cdot) \in \mathcal{Z}$ is said to be completely factorized if it can be written as the product of its marginals $\{q_i(\cdot)\}_{i=1,2...,n}$ (Mariadassou et al., 2010):

$$q(\boldsymbol{z}) = q(z_1, ..., z_n) = \prod_{i=1}^{n} q_i(z_i),$$

with

$$q_i(z_i) := \sum_{\mathbf{z}' \in \mathcal{Z} | z'_i = z_i} q(\mathbf{z}').$$

Let us denote by $\mathcal{D}_{\mathcal{Z}}^{fact} \subset \mathcal{D}_{\mathcal{Z}}$ the space of the completely factorized distributions on $\mathcal{Z}$. The variational algorithm is very similar to the EM-algorithm, with only a slightly different definition of the Expectation-step given in (3.3):

$$\text{Expectation-step (variational): compute } q_{(m+1)} := \arg \max_{q(\cdot) \in \mathcal{D}_{\mathcal{Z}}^{fact}} \mathcal{J}(q, \boldsymbol{\phi}_{(m)}) \approx w(\cdot|\boldsymbol{y}, \boldsymbol{\phi}_{(m)})$$

$$\text{Maximization-step: compute } \boldsymbol{\phi}_{(m+1)} := \arg \max_{\boldsymbol{\phi} \in \Phi} \mathcal{J}(q_{(m+1)}, \boldsymbol{\phi}).$$

Given any completely factorized distribution $q(\cdot) \in \mathcal{D}_{\mathcal{Z}}^{fact}$, each of its marginal distributions $q_i(\cdot)$ is by definition a one-trial multinomial distribution of parameters $\{\tau_{ik}^q\}_{k=1...K}$, with:

$$\tau_{ik}^q := q_i(k) = \mathbb{P}_q(Z_i = k) = \mathbb{E}_q(Z_{ik}), \tag{3.4}$$

where $\mathbb{P}_q$ and $\mathbb{E}_q$ represent respectively the probability measure and the expectation, assuming that the distribution of $\boldsymbol{Z}$ is q, and we remind that $Z_{ik} := \mathbb{1}_{\{Z_i = k\}}$. By definition, these parameters satisfy the constraints:

$$\sum_{k=1}^{K} \tau_{ik}^q = 1 \quad \forall i \in [n]. \tag{3.5}$$

These constraints will be useful later. Therefore, $q(\cdot) \in \mathcal{D}_{\mathcal{Z}}^{fact}$ is completely determined, and will be identified by $\boldsymbol{\tau}^q := \{\tau_{ij}^q\}_{(i,j) \in [K]^2}$. For notation purposes, from now on we will write $\boldsymbol{\tau}$ instead of $\boldsymbol{\tau}^q$.

For stochastic block models, the following explicit expressions for $\mathcal{J}(\cdot)$ are valid for any completely factorized distribution $q(\cdot) \in \mathcal{D}_{\mathcal{Z}}^{fact} \subset \mathcal{D}_{\mathcal{Z}}$ (Mariadassou et al., 2010; Latouche et al., 2011).
For the directed stochastic block model, the explicit formula for $\mathcal{J}(\cdot)$ is the following (the developments leading to such formula can be found in Appendix 7.9):

$$\begin{aligned}\mathcal{J}(q,\boldsymbol{\theta},\boldsymbol{\pi}) &= \mathcal{H}(q) + \sum_{\substack{\boldsymbol{z}\in\mathcal{Z}\\ q(\boldsymbol{z})>0}} q(\boldsymbol{z}) \log f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\phi}) \\ &= -\sum_{i\in[n]}\sum_{k\in[K]} \tau_{ik}\log\tau_{ik} + \sum_{i\in[n]}\sum_{k\in[K]} \tau_{ik}\log\pi_k + \sum_{\substack{(i,j)\in[n]^2\\ i\neq j}}\sum_{(k,l)\in[K]^2} \tau_{ik}\tau_{jl}\log g(y_{ij}|\theta_{kl})\end{aligned}$$

for all $(q,\boldsymbol{\theta},\boldsymbol{\pi}) \in \mathcal{D}_{\mathcal{Z}}^{fact} \times \Theta^{K\times K} \times \Pi_K$, where $\mathcal{H}(q)$ denotes the Shannon entropy of q :

$$\mathcal{H}(q) := -\sum_{\substack{\boldsymbol{z}\in\mathcal{Z}\\ q(\boldsymbol{z})>0}} q(\boldsymbol{z})\log q(\boldsymbol{z}). \tag{3.6}$$

For the undirected model, the formula for $\mathcal{J}(q,\boldsymbol{\theta},\boldsymbol{\pi})$ is the same as above, provided that the relation "$i \neq j$" is replaced with "$i < j$". We remind the reader that the parameter $\boldsymbol{\tau}$ depends on $q(\cdot) \in \mathcal{D}_{\mathcal{Z}}^{fact}$ (see equation (3.4)).

The Expectation step in equation (3.3) yields a maximization of $\mathcal{J}(\cdot)$ under the constraints given in (3.5), therefore the Lagrangian function, denoted by $\mathcal{J}_{Lagr}(\cdot)$, is for the directed model (Mariadassou et al., 2010; Latouche et al., 2011):

$$\begin{aligned}\mathcal{J}_{Lagr}(q,\boldsymbol{\theta},\boldsymbol{\pi},\{\zeta_i\}_{i\in[n]}) &= \mathcal{J}(q,\boldsymbol{\theta},\boldsymbol{\pi}) + \sum_{i=1}^{n}\zeta_i\Big(\sum_{k=1}^{K}\tau_{ik} - 1\Big) \\ &= -\sum_{i\in[n]}\sum_{k\in[K]} \tau_{ik}\log\tau_{ik} + \sum_{i\in[n]}\sum_{k\in[K]} \tau_{ik}\log\pi_k + \sum_{\substack{(i,j)\in[n]^2\\ i\neq j}}\sum_{(k,l)\in[K]^2} \tau_{ik}\tau_{jl}\log g(y_{ij}|\theta_{kl}) \\ &\quad + \sum_{i=1}^{n}\zeta_i\Big(\sum_{k=1}^{K}\tau_{ik} - 1\Big),\end{aligned}$$

where ζ_i, $i \in [n]$ represents the Lagrangian multiplier corresponding to the i-th constraint. This function shall be maximized. For the directed block model, the derivative of the Lagrangian function with respect to τ_{ik} is (Mariadassou et al., 2010):

$$\frac{\partial\mathcal{J}_{Lagr}}{\partial\tau_{ik}} = -\log\tau_{ik} - 1 + \log\pi_k + \sum_{\substack{j\in[n]\\ j\neq i}}\sum_{l\in[K]} \tau_{jl}(\log g(y_{ij}|\theta_{kl}) + \log g(y_{ji}|\theta_{lk})) + \zeta_i.$$

Therefore, by setting this quantity to 0, we find that the optimal parameter $\boldsymbol{\tau}^{(m+1)}$ must satisfy (Mariadassou et al., 2010):

$$\tau_{ik}^{(m+1)} = e^{\zeta_i - 1}\pi_k^{(m)} \prod_{\substack{j\in[n]\\ j\neq i}}\prod_{l\in[K]} \left(g_{kl}{}^{(m)}(y_{ij})g_{lk}{}^{(m)}(y_{ji})\right)^{\tau_{jl}^{(m+1)}}$$

$$\propto \pi_k^{(m)} \prod_{\substack{j\in[n]\\ j\neq i}} \prod_{l\in[K]} ({g_{kl}}^{(m)}(y_{ij}){g_{lk}}^{(m)}(y_{ji}))^{\tau_{jl}^{(m+1)}}, \tag{3.7}$$

where ${g_{kl}}^{(m)}(y_{ij}) := g(y_{ij}|\theta_{kl}^{(m)})$, and the multiplicative factor $e^{\zeta_i - 1}$does not depend on k.

For the undirected model, similarly, the derivative of the Lagrangian function with respect to τ_{ik} is:

$$\frac{\partial \mathcal{J}_{Lagr}}{\partial \tau_{ik}} = -\log \tau_{ik} - 1 + \log \pi_k + \sum_{\substack{j\in[n]\\ j\neq i}} \sum_{l\in[K]} \tau_{jl} \log g(y_{ij}|\theta_{kl}) + \zeta_i.$$

As such, the optimial parameter $\boldsymbol{\tau}^{(m+1)}$ must satisfy:

$$\tau_{ik}^{(m+1)} \propto \pi_k^{(m)} \prod_{\substack{j\in[n]\\ j\neq i}} \prod_{l\in[K]} ({g_{kl}}^{(m)}(y_{ij}))^{\tau_{jl}^{(m+1)}}, \tag{3.8}$$

where the multiplicative factor does not depend on k. For notation purposes, let $\upsilon(\cdot)$ denote the function in the right hand side term above:

$$\upsilon(\boldsymbol{\tau}^{(m+1)}) := \pi_k^{(m)} \prod_{\substack{j\in[n]\\ j\neq i}} \prod_{l\in[K]} ({g_{kl}}^{(m)}(y_{ij}))^{\tau_{jl}^{(m+1)}}.$$

Equation (3.8) can be written as $\boldsymbol{\tau}^{(m+1)} \propto \upsilon(\boldsymbol{\tau}^{(m+1)})$. This equation has no closed form solution but can be solved numerically by standard fixed point iteration; ie. computing iteratively $\upsilon^t(\boldsymbol{\tau})$ (we drop the superscript $(m+1)$ in $\boldsymbol{\tau}$ for now) for all $t = 1, 2, 3...$ and normalizing it so that $\sum_{j=1}^{K}(\upsilon^t(\boldsymbol{\tau}))_{ij} = 1$, until $\upsilon^t(\boldsymbol{\tau})$ converges to a fixed point (Mariadassou et al., 2010). The notation υ^t indicates here the functions defined by $\upsilon^1 := \upsilon$ and the recurrence relation $\upsilon^t := \upsilon \circ \upsilon^{t-1}$, where $\circ$ denotes the composition of two functions. The same reasoning is valid for the equation (3.7) concerning the directed model.

For the Maximization step, the optimal parameter arrays $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{\pi}^{(m+1)}$ satisfy, for both the directed and undirected model (Mariadassou et al., 2010):

$$\theta_{kl}^{(m+1)} = \arg\max_{\theta\in\Theta} \sum_{\substack{(i,j)\in[n]^2\\ i\neq j}} \tau_{ik}^{(m+1)}\tau_{jl}^{(m+1)} \log g(y_{ij}|\theta)$$

$$\pi_k^{(m+1)} = \frac{1}{n}\sum_{i\in[n]} \tau_{ik}^{(m+1)}.$$

For the binary model, both directed and undirected, the explicit formula for $\eta_{kl}^{(m+1)}$ is:

$$\eta_{kl}^{(m+1)} = \Big(\sum_{\substack{(i,j)\in[n]^2\\ i\neq j}} \tau_{ik}^{(m+1)}\tau_{jl}^{(m+1)} y_{ij}\Big)\Big(\sum_{\substack{(i,j)\in[n]^2\\ i\neq j}} \tau_{ik}^{(m+1)}\tau_{jl}^{(m+1)}\Big)^{-1}.$$

For the Poisson model, both directed and undirected, the formula for $\lambda_{kl}^{(m+1)}$ is the same:

$$\lambda_{kl}^{(m+1)} = \Big(\sum_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \tau_{ik}^{(m+1)} \tau_{jl}^{(m+1)} y_{ij} \Big) \Big(\sum_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \tau_{ik}^{(m+1)} \tau_{jl}^{(m+1)} \Big)^{-1}.$$

Mariadassou et al. (2010) also gives a general formula for computing $\boldsymbol{\theta}^{(m+1)}$ for $g(\cdot)$ belonging to any exponential family distribution.

3.3.1 One implementation of the variational expectation-maximization algorithm

One implementation of the variation algorithm has been developed in the R package `blockmodels`, which provides a relatively fast algorithm. In this implementation, the network is iteratively split in an increasing number of communities, starting from the trivial single-community model, to a K_{max}-community model. The maximum number of communities to detect, K_{max}, can be chosen by the user. At each iteration, the variational algorithm is launched, by using several starting points. The starting points needed for launching the variational-algorithm are obtained either by the results obtained at the previous iterations, either by the Laplacian matrix spectral decomposition of the network (see Section 3.5). Such implementation will be used in Chapters 4 and 5.

3.4 Bayesian method: Gibbs sampler for the binary model

A Bayesian approach for stochastic block modelling was studied in the context of discrete relational data models in the work of Nowicki & Snijders (2001). In such models, any finite number of binary relations between nodes can be considered; therefore the binary model, as defined in Section 1.2, corresponds to the simplest model considered.
Their approach aims at obtaining the posterior distribution $f(\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\pi}|\boldsymbol{y})$, supposing that, a-priori, $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$ are independent, with $\boldsymbol{\pi}$ following a $Dirichlet(T_1, ..., T_K)$ distribution:

$$\boldsymbol{\pi} \sim Dirichlet(T_1, ..., T_K),$$

where parameters T_k represent the prior information one may have about class proportions. If class membership was observed in a previous experiment, then T_k could be the previously observed size of class k Nowicki & Snijders (2001). If no previous information is available, the uniform $Dirichlet(1, 1, ..., 1)$ may be used. Alternatively, Latouche et al. (2009) mentions the non-informative Jeffrey's prior:

$$\boldsymbol{\pi} \sim Dirichlet(1/2, 1/2, ...1/2).$$

Parameters $\{\eta_{kl}\}_{(k,l)\in[K]^2}$ follow, a-priori, independent $Beta(E_1^{kl}, E_2^{kl})$ distributions, where E_1^{kl} and E_2^{kl} represent the prior information one may have about the presence/absence of edges between nodes of class k and l. E_1^{kl} is then the number of edges previously observed between nodes of class k and of class l in a previous experiment, and E_2^{kl} the number of previously observed non-edges. If no previous information is available, the uniform $Beta(1, 1)$ is sometimes used.
Nowicki & Snijders (2001) obtained thus explicit formulas for performing the Gibbs sampling algorithm, which is guaranteed to converge to the posterior distribution $f(\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\pi}|\boldsymbol{y})$.
Let d_{ik} denote the number of edges connecting node i to nodes of class k:

$$d_{ik} := \sum_{j=1}^{n} y_{ij} z_{jk},$$

$\tilde{n}_{kl}$ denote the maximum potential number of edges going from nodes of class k to nodes of class l:

$$\tilde{n}_{kl} := n_{kl} \quad \forall k, l \in [K]^2, k \neq l$$

$$\tilde{n}_{kk} := \frac{n_{kk}}{2} \quad \forall k \in [K],$$

and $\tilde{o}_{kl}$ denote the number of edges going from nodes of class k to nodes of class l:

$$\tilde{o}_{kl} := \frac{1}{1 + \mathbb{1}_{\{k=l\}}} o_{kl}.$$

The Gibbs sampling algorithm, as defined by Nowicki & Snijders (2001), is based on the following procedure: start from an initial $\boldsymbol{\pi}^{(0)}, \boldsymbol{\eta}^{(0)}, \boldsymbol{z}^{(0)}$, and repeat the following steps for a certain number of iterations, starting from $t = 1$:

> Step 1: draw $\boldsymbol{\pi}^{(t)}, \boldsymbol{\eta}^{(t)}$ from $f(\boldsymbol{\pi}, \boldsymbol{\eta} | \boldsymbol{y}, \boldsymbol{z}^{(t-1)})$, ie. the posterior distribution of $(\boldsymbol{\pi}, \boldsymbol{\eta})$ given $\boldsymbol{y}$ and $\boldsymbol{z}^{(t-1)}$.
>
> Step 2: For $i = 1...n$, draw $\boldsymbol{z}_i^{(t)}$ from $f(\boldsymbol{z}_i | \boldsymbol{y}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\eta}^{(t)}, z_1^{(t)}, ..., z_{i-1}^{(t)}, z_{i+1}^{(t-1)}, ..., z_n^{(t-1)})$, ie. the posterior distribution of $\boldsymbol{z}_i$ given $\boldsymbol{\pi}^{(t)}, \boldsymbol{\eta}^{(t)}$, $z_h^{(t)}$ for $h = 1, ..., i-1$ and $z_h^{(t-1)}$ for $h = i+1, ..., n$.

For step 1, since $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$ are independent in their prior distribution, they are also independent in their posterior distribution (Nowicki & Snijders, 2001). To see this, let us remind that independence implies that the distribution $f(\boldsymbol{\pi}, \boldsymbol{\eta})$ can be factorized in two terms each depending only on one of the two parameters; let us then write:

$$\begin{aligned} f(\boldsymbol{\pi}, \boldsymbol{\eta} | \boldsymbol{y}, \boldsymbol{z}) &= \frac{f(\boldsymbol{\pi}, \boldsymbol{\eta}) f(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{\pi}, \boldsymbol{\eta})}{f(\boldsymbol{z}, \boldsymbol{y})} \\ &= \frac{f(\boldsymbol{\pi}, \boldsymbol{\eta}) f(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) f(\boldsymbol{z} | \boldsymbol{\pi}, \boldsymbol{\eta})}{f(\boldsymbol{z}, \boldsymbol{y})} \\ &= \frac{f(\boldsymbol{\pi}, \boldsymbol{\eta}) f(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{\eta}) f(\boldsymbol{z} | \boldsymbol{\pi})}{f(\boldsymbol{z}, \boldsymbol{y})}, \end{aligned}$$

where the second equality is an application of the conditional probability rule, and the third equality is due to the fact that $f(\boldsymbol{z} | \boldsymbol{\pi}, \boldsymbol{\eta}) = f(\boldsymbol{z} | \boldsymbol{\pi})$, since the class-membership does not depend on the edge probability matrix $\boldsymbol{\eta}$, as can be seen by equation (1.9). Furthermore, $f(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) = f(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{\eta})$, since the network matrix does not depend on the class-probability vector $\boldsymbol{\pi}$, as can be seen by equations (1.10) and (1.11). Therefore, if $f(\boldsymbol{\pi}, \boldsymbol{\eta})$ can be factorized in two terms each depending only on one of the two parameters, the same is true for $f(\boldsymbol{\pi}, \boldsymbol{\eta} | \boldsymbol{y}, \boldsymbol{z})$. This implies that $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$ are a-posteriori independent.
Therefore $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$ can be drawn separately from their individual posterior distributions: for $\boldsymbol{\pi}$, Nowicki & Snijders (2001) show that this posterior distribution is a Dirichlet distribution of parameters $(n_k + T_k)_{k=1...K}$, and for η_{kl} it is a Beta distribution of parameters

$(E_1^{kl} + \tilde{o}_{kl}, E_2^{kl} + \tilde{n}_{kl} - \tilde{o}_{kl})\ \forall k \leq l.$

For step 2, Nowicki & Snijders (2001) show that the posterior distribution of $\boldsymbol{z}_i$ is given, for the undirected binary model, by the distribution defined by:

$$\mathbb{P}(z_i = k|\boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\pi}, \{z_j\}_{j\neq i}) = c\pi_k \prod_{h=1}^{K} \eta_{kh}^{d_{ih}} (1-\eta_{kh})^{n_h - d_{ih}},$$

where c is a normalizing constant that does not depend on k. At each step t, it is clear that values of d_{ik}, $\tilde{n}_k$ and $\tilde{o}_{kl}$ must be computed using $\boldsymbol{z}^{(t-1)}$. For the directed binary model, such posteriori distribution has a very similar form as for the directed model (Nowicki & Snijders, 2001).

Estimators generated by this algorithm are guaranteed to converge to the posterior distribution $f(\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\pi}|\boldsymbol{y})$. For testing if convergence has taken place, the algorithm can be run using several starting values, and data drawn after a burn-in period, for example data drawn after $t > 10000$, are compared. If the distribution of the data does not change too much depending on the starting value, it is likely that convergence to the limit distribution has taken place. (Nowicki & Snijders, 2001).

The Gibbs sampler also allows to integrate prior knowledge about classes to which specific nodes belong, by making minor changes to the steps of the algorithm (Nowicki & Snijders, 2001). However, it has the downside of being slow when compared to the other estimation methods; indeed, its applications are in practice limited to networks of the order of 100 nodes.

Concerning the choice of the parameters of the prior distribution of $\boldsymbol{\pi}$, if all classes have a-priori a similar weight, choosing $T_k = T$ is a good choice. Choosing small values for T is risky because there are more chances that the algorithm gets stuck in regions where sizes of some classes are close to 0 and corresponding elements of η are close to uniform distribution. It might then be better to choose larger values of T. Nowicki & Snijders (2001) suggest using $T = 100K$.
For the choice of parameters for the prior distribution of π_{kl}, larger values of E_1^{kl} (and/or lower values of E_2^{kl}) favor models with high probabilities of edges between class k and l (Nowicki & Snijders, 2001).

For selecting the number of the classes of the SBM, the ICL criterion is often used (Daudin et al., 2008; Barbillon et al., 2017; Biernacki et al., 2000; Mariadassou et al., 2010). This criterion is based on the Maximization, in the K-class model m_K, of the integrated complete-data likelihood :

$$\int_{\Theta^{K\times K}\times\Pi_K} f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\pi}, m_K) f(\boldsymbol{\theta}, \boldsymbol{\pi}|m_K) d\boldsymbol{\theta} d\boldsymbol{\pi}$$

where $f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\pi}, m_K)$ is the complete data likelihood under the K-class model and parameter $(\boldsymbol{\theta}, \boldsymbol{\pi})$, and $f(\boldsymbol{\theta}, \boldsymbol{\pi}|m_K)$ the prior distribution of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ under the K-class model. Mariadassou et al. (2010) show that, if $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are independent in their prior distribution, and a non-informative Jeffrey prior is chosen for $\boldsymbol{\pi}$, then maximizing the following quantity

is approximately equivalent to maximizing the integrated complete likelihood for the directed SBM:

$$\max_{\boldsymbol{\theta},\boldsymbol{\pi}\in\Theta^{K\times K}\times\Pi_K} \log f(\boldsymbol{y},\hat{\boldsymbol{z}}|\boldsymbol{\theta},\boldsymbol{\pi},m_K) - \frac{1}{2}\{K^2\log n(n-1) + (K-1)\log n\}$$

where $\hat{\boldsymbol{z}}$ is an estimator of $\boldsymbol{z}$.
Whereas for the undirected SBM, the approximated ICL criterion is given by (Daudin et al., 2008; Barbillon et al., 2017; Biernacki et al., 2000; Mariadassou et al., 2010):

$$\max_{\boldsymbol{\theta},\boldsymbol{\pi}\in\Theta^{K\times K}\times\Pi_K} \log f(\boldsymbol{y},\hat{\boldsymbol{z}}|\boldsymbol{\theta},\boldsymbol{\pi},m_K) - \frac{1}{2}\{\frac{K(K+1)}{2}\log\frac{n(n-1)}{2} + (K-1)\log n\}. \tag{3.9}$$

These criteria have the advantage of penalising models with too many classes.

3.5 Spectral methods

A class of algorithms, all of which are based on the spectral decomposition, will be presented in this section. Such a spectral decomposition is applied generally to the adjacency matrix $\boldsymbol{Y}$, or to other related matrices that will be defined later, such as the Laplacian matrix.
To show how this method works, an introductory example, as given by (Abbe, 2018), will be illustrated before giving some general results. Let us suppose that we are in the symmetric model scenario (section 1.29), with $K = 2$ classes, and that the number of nodes, n, is even. Without loss of generality, we can suppose that nodes from 1 to $\frac{n}{2}$ are in class 1 and nodes from $\frac{n}{2}+1$ to n are in class 2: $z_1 = 1, ..., z_{\frac{n}{2}} = 1, z_{\frac{n}{2}+1} = 2, ..., z_n = 2$. In this case, the expected value of the adjacency matrix $\boldsymbol{Y}$ is a $n \times n$ block matrix of the following form (supposing that self-edges are allowed):

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z}) = \begin{pmatrix} p_{in} & \cdots & p_{in} & p_{out} & .... & p_{out} \\ . & \cdots & . & . & .... & . \\ . & \cdots & . & . & .... & . \\ p_{in} & \cdots & p_{in} & p_{out} & .... & p_{out} \\ p_{out} & \cdots & p_{out} & p_{in} & .... & p_{in} \\ . & \cdots & . & . & .... & . \\ . & \cdots & . & . & .... & . \\ p_{out} & \cdots & p_{out} & p_{in} & .... & p_{in} \end{pmatrix}.$$

The matrix $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z})$ has three eigenvalues: the first eigenvalue is $n(p_{in}+p_{out})/2$ with multiplicity 1 and associated eigenvector 1_n; the second eigenvalue is $n(p_{in}-p_{out})/2$ with multiplicity 1 and associated eigenvector $((1_{n/2})^T, -(1_{n/2})^T)^T$; and finally the last eigenvalue is 0 with multiplicity $n-2$ (Abbe, 2018). So, the second eigenvector of $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z})$, $((1_{n/2})^T, -(1_{n/2})^T)^T$, provides all the information for clustering nodes (if nodes are permuted, eigenvalues are not affected, and corresponding eigenvectors are simply permuted as the nodes are).

The theorem in the following paragraph insures that in general, the eigen-decomposition of the matrix $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z})$ allows us to recover the classes of the nodes as we displayed in the previous example. Before stating the theorem, it is useful to give the general closed formula for $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z})$ in terms of parameters $\boldsymbol{\theta}$ and class-memberships $\boldsymbol{z}$. Indeed, by reminding ourselves of the definition of the SBM (section 1.2), we have that (Lei & Rinaldo, 2015; Rohe et al., 2011):

$$\mathbb{E}(Y_{ij}|\boldsymbol{z}) = \mathbb{E}(Y_{ij}|z_{ik}=1, z_{jl}=1) = \sum_{k',l' \in [K]} z_{ik'} z_{jl'} \theta_{k'l'} = (\boldsymbol{z\theta z}^T)_{ij}, \tag{3.10}$$

where $\boldsymbol{z}$ now denotes the $n \times K$ binary matrix with entries $\{z_{ik}\}_{i\in[n],k\in[K]}$. so that the i-th row of $\boldsymbol{z}$ contains one single 1 at the k-th column if and only if the i-th node belongs to the k-th class. While all other elements of the row are 0 (the symbol $\boldsymbol{z}$ denoted a slightly different object previously; still, the information it provides has not changed). As such, we can write:

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z}) = \boldsymbol{z\theta z}^T. \tag{3.11}$$

Theorem 7 will now be stated. Note that, for a matrix $\boldsymbol{X}$, and index subsets $\mathscr{S}_1$ and $\mathscr{S}_2$, we will write $\boldsymbol{X}_{\mathscr{S}_1 *}$ and $\boldsymbol{X}_{*\mathscr{S}_2}$ to denote the submatrix corresponding to the respective rows and columns corresponding to such indices.

Theorem 7. *(Lei & Rinaldo, 2015): Let the pair $(\boldsymbol{z}, \boldsymbol{\theta})$ parametrize a SBM with K communities, where $\boldsymbol{\theta}$ is of full rank. Let $\boldsymbol{UDU}^T$ be the eigen-decomposition of $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z}) = \boldsymbol{z\theta z}^T$. Then $\boldsymbol{U} = \boldsymbol{zX}$, where $\boldsymbol{X} \in \mathbb{R}^{K\times K}$ (i). Furthermore, $\boldsymbol{X}$ is of full rank and $||\boldsymbol{X}_{k*} - \boldsymbol{X}_{l*}||= \sqrt{n_k^{-1} + n_l^{-1}}$ (ii).*

Proof. Let $\boldsymbol{C}$ denote the diagonal matrix whose diagonal is $(\sqrt{n_1}, ..., \sqrt{n_K})^T$. Then:

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z}) = \boldsymbol{z\theta z}^T = \boldsymbol{zC}^{-1}\boldsymbol{C\theta CC}^{-1}\boldsymbol{z}^T = \boldsymbol{zC}^{-1}\boldsymbol{C\theta C}(\boldsymbol{zC}^{-1})^T; \tag{3.12}$$

it is straightforward to verify that $\boldsymbol{zC}^{-1}$ is orthonormal. Let $\boldsymbol{FDF}^T = \boldsymbol{C\theta C}$ be the eigen-decomposition of $\boldsymbol{C\theta C}$; so that:

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z}) = \boldsymbol{zC}^{-1}\boldsymbol{FDF}^T(\boldsymbol{zC}^{-1})^T = \boldsymbol{zC}^{-1}\boldsymbol{FD}(\boldsymbol{zC}^{-1}F)^T. \tag{3.13}$$

So, we can write $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z}) = \boldsymbol{UDU}^T$ where $\boldsymbol{U} = \boldsymbol{zC}^{-1}\boldsymbol{F}$; by writing $\boldsymbol{X} = \boldsymbol{C}^{-1}\boldsymbol{F}$, we obtain the first statement of the theorem. To obtain the second part, we can see that the rows of $\boldsymbol{C}^{-1}\boldsymbol{F}$ are perpendicular to each other, and the k-th row has length $\sqrt{1/n_k}$, by writing $\boldsymbol{C}^{-1}\boldsymbol{F}(\boldsymbol{C}^{-1}\boldsymbol{F})^T = \boldsymbol{C}^{-1}\boldsymbol{FF}^T\boldsymbol{C}^{-1} = (\boldsymbol{C}^{-1})^2$, the exponent operation ^{2} denoting the matrix multiplication of a matrix by itself. Note that the equalities $\boldsymbol{C}^{-1}\boldsymbol{F}(\boldsymbol{C}^{-1}\boldsymbol{F})^T = \boldsymbol{C}^{-1}\boldsymbol{FF}^T\boldsymbol{C}^{-1} = (\boldsymbol{C}^{-1})^2$ are due to the fact that $(\boldsymbol{C}^{-1})^T = \boldsymbol{C}^{-1}$ and $\boldsymbol{FF}^T$ is the identity matrix.

□

Therefore, the rows of $\boldsymbol{U}$ are obtained by rotating the rows of $\boldsymbol{z}$ by the invertible matrix $\boldsymbol{X}$; so, if (and only if) node i and node j are in the same class, so that the the ith line of $\boldsymbol{z}$ and the jth line of $\boldsymbol{z}$ are equal, then the ith line of $\boldsymbol{U}$ and the jth line of $\boldsymbol{U}$ are equal too.
Therefore, if $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z})$ was known, the eigenvectors correspoding to the K largest (in absolute value) eigenvalues could be used in order to split nodes according to their corresponding eigenvector. Since $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{z})$ is unknown, this same algorithm is applied to the observed $\boldsymbol{y}$, yielding an estimation of $\boldsymbol{U}$, denoted $\hat{\boldsymbol{U}}$.
A k-means algorithm is finally applied to the rows of $\hat{\boldsymbol{U}}$ to cluster the nodes (Lei & Rinaldo, 2015; Rohe et al., 2011).

Another very popular approach consists in using, in place of the adjacency matrix, the so-called standardized Laplacian matrix (Rohe et al., 2011; Amini et al., 2013). This matrix is

defined in the following way: let $\boldsymbol{\Delta}$ denote the degree matrix, ie. a diagonal matrix whose diagonal elements are the degrees of the nodes:

$$\Delta_{ii} := \sum_{j=1}^{n} y_{ij} \quad \forall i = 1, ..., n,$$

and $\Delta_{ij} := 0$ for $i \neq j$. Then, the standardized Laplacian matrix can be defined as:

$$\boldsymbol{L} := \boldsymbol{\Delta}^{-1/2} \boldsymbol{y} \boldsymbol{\Delta}^{-1/2}, \tag{3.14}$$

where $\Delta^{-1/2}$ denotes the square root of the matrix Δ^{-1}, meaning that $\Delta^{-1/2}$ is the $n \times n$ diagonal matrix, containing in its diagonal the scalars $\frac{1}{\sqrt{\Delta_{ii}}}$, $i = 1, ..., n$. In the following, we will refer to this last matrix as the Laplacian matrix. Equivalently, the difference between the identity marix and the Laplacian, $\boldsymbol{I} - \boldsymbol{L}$ can be used; indeed matrices $\boldsymbol{L}$ and $\boldsymbol{I} - \boldsymbol{L}$ have the same eigenvectors (Rohe et al., 2011).

The following theorem is almost equivalent to Theorem 7, with the Laplacian matrix substituting the adjacency matrix:

Theorem 8. *(Rohe et al., 2011): Let the pair $(\boldsymbol{z}, \boldsymbol{\theta})$ parametrize a SBM with K communities, where Θ is of full rank. Let $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{U}}^T$ be the eigen-decomposition of $\mathbb{E}(\boldsymbol{L}|\boldsymbol{z})$, where $\boldsymbol{L}$ denotes the (random) standardized Laplacian matrix as defined in 3.14. Then $\tilde{\boldsymbol{U}} = \boldsymbol{z}\tilde{\boldsymbol{X}}$, where $\tilde{\boldsymbol{X}} \in \mathbb{R}^{K \times K}$ (i). Furthermore, $\tilde{\boldsymbol{X}}$ is of full rank (ii).*

This theorem states that the rows of $\tilde{\boldsymbol{U}}$ (as the rows of $\boldsymbol{U}$ in Theorem 7) are obtained by rotating the rows of $\boldsymbol{z}$ by an invertible matrix; so, if (and only if) node i and node j are in the same class, so that the ith line of $\boldsymbol{z}$ and the jth line of $\boldsymbol{z}$ are equal, then the ith line of $\tilde{\boldsymbol{U}}$ and the jth line of $\tilde{\boldsymbol{U}}$ are equal too. The proof of Theorem 8 is not given here, but it is based on very similar arguments as those of Theorem 7.

In the context of degree-corrected block models, Lei & Rinaldo (2015) introduced a "spherical" spectral approach, in which the rows of $\hat{\boldsymbol{U}}$ are normalized so as to make their norm equal to 1; in such a way, one considers the directions that rows of $\hat{\boldsymbol{U}}$ take; this allows to isolate the effect of the degree-correction parameters $\{\beta_i\}_{i=1...n}$ (Lei & Rinaldo, 2015).
Furthermore, Amini et al. (2013) introduced a regularized spectral method, which in practice was found to give slightly better results than the standard spectral method in some particular cases in which the network is very sparse (Amini et al., 2013). Such a regularized approach consists in adding a perturbation to the adjacency matrix $\boldsymbol{y}$, ie. artificially adding edges having very small weights in order to make the network more connected; therefore, the matrix $\boldsymbol{y}_{reg}$ is defined as:

$$\boldsymbol{y}_{reg} := \boldsymbol{y} + \frac{\iota \bar{d}}{n} 1_n 1_n^T,$$

where $\bar{d} := \frac{\sum_{ij} y_{ij}}{2n}$ indicates the mean degree of the nodes and ι is a small constant.

All such spectral decomposition approaches have the advantage of being very fast, even when the size of the matrix is relatively large (let say, up to thousands of nodes); however, for very big matrices, such approaches can become slow too and require a large amount of memory.

3.6 Profile-likelihood estimator and modularity estimator

Other approaches for community detection are the profile likelihood method and the modularity method. The modularity method in not directly based on the stochastic block models, but is widely studied in the literature (Bickel & Chen, 2009; Karrer & Newman, 2011; Zhao et al., 2012). These two approaches share some common properties (Bickel & Chen, 2009). This section is mainly based on the works from Karrer & Newman (2011) and Zhao et al. (2012), which are both based on undirected model scenarios.

Note that in this section, it will be assumed that self-loops are allowed, ie. assumption (1.8) is relaxed. In this case, it is convenient to define Y_{ii} as twice the number of self-edges of node i.
Formulas will first be developed based on the standard, Poisson non degree-corrected model; and then on the degree-corrected SBM.

3.6.1 Profile-likelihood estimator and modularity estimator - Standard SBM

If we consider the class-membership vector $\boldsymbol{z}$ as fixed (as it was a parameter), by reminding ourselves of equation (1.27), the likelihood of the standard (Poisson) model can be written as (Karrer & Newman, 2011):

$$\begin{aligned}\mathscr{L}(\boldsymbol{\lambda}, \boldsymbol{z}) &= \prod_{\substack{(i,j)\in[n]^2 \\ i<j}} \frac{\lambda_{z_i z_j}^{y_{ij}}}{y_{ij}!} \exp(-\lambda_{z_i z_j}) \prod_{i\in[n]} \frac{(\frac{1}{2}\lambda_{z_i z_j})^{y_{ij}/2}}{(y_{ij}/2)!} \exp(-\frac{1}{2}\lambda_{z_i z_j}) \\ &= \frac{1}{\prod_{i<j} y_{ij}! \prod_i 2^{y_{ii}/2}(y_{ii}/2)!} \prod_{(k,l)\in[K]^2} \lambda_{kl}^{o_{kl}/2} \exp(-\frac{1}{2} n_k n_l \lambda_{kl}),\end{aligned}$$

where we remind that n_k and o_{kl} were defined in (1.12); for the Poisson model, o_{kl} represents the sum of the weights of the edges going from nodes of class k to nodes of class l. The log-likelihood is, neglecting constants and terms independent of $\boldsymbol{\lambda}$ and $\boldsymbol{z}$, equal to (Karrer & Newman, 2011):

$$\ell(\boldsymbol{\lambda}, \boldsymbol{z}) = \sum_{(k,l)\in[K]^2} (o_{kl} \log(\lambda_{kl}) - n_k n_l \lambda_{kl}). \tag{3.15}$$

The log-likelihood function $\ell(\cdot)$ can be maximized in two stages, first with respect to $\boldsymbol{\lambda}$ and then to $\boldsymbol{z}$. By differentiation of $\ell(\cdot)$ with respect to $\boldsymbol{\lambda}$, we find that the maximum-likelihood estimator $\hat{\boldsymbol{\lambda}}$ is given by:

$$\hat{\lambda}_{kl} = \frac{o_{kl}}{n_k n_l},$$

and the maximum value of $\ell(\cdot)$ reached by $\hat{\boldsymbol{\lambda}}$ is then given, by substituting $\hat{\boldsymbol{\lambda}}$ in equation

(3.15), by:

$$\begin{aligned}\ell(\hat{\boldsymbol{\lambda}}, \boldsymbol{z}) &= \sum_{(k,l)\in[K]^2} (o_{kl} \log(\frac{o_{kl}}{n_k n_l}) - n_k n_l \frac{o_{kl}}{n_k n_l}) \\ &= \sum_{(k,l)\in[K]^2} (o_{kl} \log(\frac{o_{kl}}{n_k n_l}) - o_{kl}) \\ &= \sum_{(k,l)\in[K]^2} (o_{kl} \log(\frac{o_{kl}}{n_k n_l})) - \sum_{(k,l)\in[K]^2} o_{kl} \\ &= \sum_{(k,l)\in[K]^2} (o_{kl} \log(\frac{o_{kl}}{n_k n_l})) - 2o,\end{aligned}$$

where $o := \frac{1}{2}\sum_{(k,l)\in[K]^2} o_{kl}$ is the sum of the weights of all edges in the network, which is independent of the parameters and the class-membership too, and so can be dropped.
We can then define the unnormalized log-likelihood as (Karrer & Newman, 2011):

$$\ell_{unn}(\boldsymbol{z}) = \sum_{(k,l)\in[K]^2} (o_{kl} \log(\frac{o_{kl}}{n_k n_l})). \tag{3.16}$$

The maximization of this quantity (called "profile likelihood modularity" by Bickel & Chen (2009)) with respect to the class-membership $\boldsymbol{z}$ yields an estimation of the class-membership, $\hat{\boldsymbol{z}}$.
An interpretation in terms of Kullback-Leibler divergence can be given to this quantity (Karrer & Newman, 2011). Indeed, by adding and dividing by constant factors, we can also write the alternative form:

$$\ell_{unn}(\boldsymbol{z}) = \sum_{(k,l)\in[K]^2} (\frac{o_{kl}}{2o} \log(\frac{o_{kl}/2o}{n_k n_l/n^2})). \tag{3.17}$$

For a given class-membership assignment, let X_1 be the group membership at a (randomly selected) end of one edge of the network, and X_2 the group assignment at the other end of the edge. The probability distribution of variables X_1 and X_2 is then given by $p_K(X_1 = k, X_2 = l) = p_K(k,l) = o_{kl}/2o$, which appears twice in (3.17). The remaining term in the denominator of the logarithm, $n_k n_l/n^2$, is equal to the expected value of the same probability in a network with the same class-membership assignment, but edges placed completely randomly (Karrer & Newman, 2011). If we call p_1 this second distribution, equation (3.17) can be rewritten as:

$$\ell_{unn}(\boldsymbol{z}) = \sum_{(k,l)\in[K]^2} (p_K(k,l) \log(\frac{p_K(k,l)}{p_1(k,l)})),$$

which corresponds indeed to the Kullback-Leibler divergence between distributions $p_K(\cdot)$ and $p_1(\cdot)$. As such, the class-membership yielding the distribution of edges most "distant" from a random distribution is the optimal one, according to the profile likelihood criterion (Karrer & Newman, 2011).

Another popular general criterion is given by the modularity, which in its general form is written as (Zhao et al., 2012; Karrer & Newman, 2011):

$$M(\boldsymbol{z}) = \sum_{(i,j)\in[n]^2} (y_{ij} - P_{ij}) \mathbb{1}_{\{z_i = z_j\}}. \tag{3.18}$$

Note that in the formula in the work from Karrer & Newman (2011), a multiplicative constant is added, however such constant has no influence on the maximization process. P_{ij} represents here the estimated probability of having an edge between node i and node j under some null model. If we choose the null model as being the trivial Erdös-Rényi model, for which P_{ij} is a constant which can be estimated by $P_{ij} = 2o/n^2$, we obtain, by plugging $P_{ij} = 2o/n^2$ in equation (3.18), what is called the Erdös-Rényi modularity (Zhao et al., 2012), denoted by M_{ER} :

$$M_{ER}(\boldsymbol{z}) = \sum_{k\in[K]} (o_{kk} - \frac{n_k^2}{n^2} 2o).$$

The estimator $\hat{\boldsymbol{z}}$ of the class-membership $\boldsymbol{Z}$ for, respectively, the profile-likelihood and the Erdös-Rényi modularity, is then given by:

$$\begin{aligned} \text{Profile-likelihood:} \quad & \arg\max_{\boldsymbol{z}\in\mathcal{Z}} \ell_{unn}(\boldsymbol{z}) \\ \text{ER modularity:} \quad & \arg\max_{\boldsymbol{z}\in\mathcal{Z}} M_{ER}(\boldsymbol{z}). \end{aligned}$$

Such an optimisation is an NP-hard problem (Zhao et al., 2012; Karrer & Newman, 2011); therefore, heuristic iterative algorithms are used to approximate the exact maximum. Two examples of such algorithms will be given at the end of the next section on the degree-corrected SBM.

3.6.2 Profile-likelihood estimator and modularity estimator - Degree-corrected SBM

For the degree-corrected (Poisson) SBM, similar developments as for the standard model can be done, and the likelihood function can be written as (Karrer & Newman, 2011):

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{z}) = \prod_{\substack{(i,j)\in[n]^2 \\ i<j}} \left(\frac{(\beta_i\beta_j\lambda_{z_iz_j})^{y_{ij}}}{y_{ij}!} \exp(-\beta_i\beta_j\lambda_{z_iz_j})\right) \prod_{i\in[n]} \left(\frac{(\frac{1}{2}\beta_i^2\lambda_{z_iz_j})^{y_{ij}/2}}{(y_{ij}/2)!} \exp(-\frac{1}{2}\beta_i^2\lambda_{z_iz_j})\right),$$

which can also be rewritten, by using the normalization constraints $\sum_{i=1}^n \beta_i z_{ik} = 1$ for $k \in [K]$ as:

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{z}) = \frac{1}{\prod_{i<j} y_{ij}! \prod_i 2^{y_{ii}/2}(y_{ii}/2)!} \left(\prod_{i\in[n]} \beta_i^{d_i}\right) \left(\prod_{(k,l)\in[K]^2} \lambda_{kl}^{o_{kl}/2} \exp(-\frac{1}{2}\lambda_{kl})\right)$$

and so the log-likelihood is, up to constants:

$$\ell(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{z}) = 2\sum_{i\in[n]} (d_i \log \beta_i) + \sum_{(k,l)\in[K]^2} (o_{kl} \log(\lambda_{kl}) - \lambda_{kl}).$$

For a fixed $\boldsymbol{z}$, if we differentiate $\ell(\cdot)$ with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$, we find that maximum-likelihood estimators $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\beta}}$ are given by:

$$\hat{\lambda}_{kl} = o_{kl} \quad \hat{\beta}_i = \frac{d_i}{\delta_{z_i}}$$

where δ_k denotes the sum of the degrees of nodes in class k, ie. $\delta_k := \sum_{i\in[n]} d_i z_{ik}$, and the maximal value of $\ell(\cdot)$ reached by plugging in $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\beta}}$ is then given by:

$$\begin{aligned}\ell(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \boldsymbol{z}) &= 2\sum_{i\in[n]}(d_i \log \frac{d_i}{\delta_{z_i}}) + \sum_{(k,l)\in[K]^2}(o_{kl}\log(o_{kl}) - o_{kl}) \\ &= 2\sum_{i\in[n]}(d_i \log \frac{d_i}{\delta_{z_i}}) + \sum_{(k,l)\in[K]^2}(o_{kl}\log(o_{kl})) - 2o.\end{aligned}$$

where, again, the last term $-2o$ can be dropped since it is not depending on the parameters or the class-membership. The first term can be rewritten as:

$$2\sum_{i\in[n]}(d_i \log \frac{d_i}{\delta_{z_i}}) = 2\sum_{i\in[n]} d_i \log d_i - 2\sum_{i\in[n]} d_i \log \delta_{z_i}.$$

Furthermore, it can be proven (see Appendix 7.11) that:

$$2\sum_{i\in[n]} d_i \log \delta_{z_i} = \sum_{(k,l)\in[K]^2} o_{kl} \log \delta_k \delta_l. \tag{3.19}$$

Therefore, the likelihood function $\ell(\cdot)$ can be rewritten as:

$$\begin{aligned}\ell(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}}, \boldsymbol{z}) &= 2\sum_{i\in[n]}(d_i \log d_i) - \sum_{(k,l)\in[K]^2} o_{kl}\log\delta_k\delta_l + \sum_{(k,l)\in[K]^2}(o_{kl}\log(o_{kl})) - 2o \\ &= \sum_{(k,l)\in[K]^2}(o_{kl}\log(\frac{o_{kl}}{\delta_k\delta_l})) + 2\sum_{i\in[n]}(d_i\log d_i) - 2o,\end{aligned}$$

so that, dropping terms independent of $\boldsymbol{z}$, the unnormalized log-likelihood function is defined as (Karrer & Newman, 2011):

$$\ell_{unn}(\boldsymbol{z}) = \sum_{(k,l)\in[K]^2}(o_{kl}\log(\frac{o_{kl}}{\delta_k\delta_l})).$$

Note that the only difference with respect to the formula for the standard model (see equation 3.16) is that δ_k and δ_l appear instead of n_k and n_l.
This quantity too can be interpreted in term of a Kullback-Leibler divergence; indeed, by adding and dividing by constants, we can write an alternative formula for $\ell_{unn}(\cdot)$:

$$\ell_{unn}(\boldsymbol{z}) = \sum_{(k,l)\in[K]^2}(\frac{o_{kl}}{2o}\log(\frac{o_{kl}/2o}{(\delta_k/2o)(\delta_l/2o)})).$$

This quantity corresponds to the Kullback-Leibler divergence between p_K as defined previously, and the expectation of the same probability in a model with the same class-membership

assignment, in which edges are placed randomly, but the degree of each node i is fixed to d_i this time (Karrer & Newman, 2011).

If we use the modularity formula in equation (3.18), with the null model being the trivial degree-corrected SBM with $K = 1$ class, for which P_{ij} can be estimated by $(d_i d_j / 2o)$, we obtain the so-called Girvan-Newman modularity (Zhao et al., 2012):

$$M_{GN}(\boldsymbol{z}) = \sum_{k \in [K]} (o_{kk} - \frac{\delta_k^2}{2o}).$$

The estimator $\hat{\boldsymbol{z}}$ of the class-membership $\boldsymbol{Z}$ for, respectively, the profile-likelihood and the Girvan-Newman modularity, is then given by:

$$\begin{aligned} \text{Profile-likelihood:} \quad & \arg\max_{\boldsymbol{z} \in \mathcal{Z}} \ell_{unn}(\boldsymbol{z}) \\ \text{GN modularity:} \quad & \arg\max_{\boldsymbol{z} \in \mathcal{Z}} M_{GN}(\boldsymbol{z}). \end{aligned}$$

For both the standard model and the degree-corrected model, the optimization of $\ell(\cdot)$ and $M(\cdot)$ over $\mathcal{Z}$ is a NP-hard problem; therefore in practice heuristic algorithms must be used to optimise ℓ_{unn}, M_{GN} and M_{ER} iteratively ; for example Karrer & Newman (2011) uses an heuristic algorithm inspired from the so-called Kernighan-Lin algorithm, which has the advantage of allowing to calculate rapidly the change in ℓ_{unn} when one node moves from one class to the other. While in the work from Zhao et al. (2012), an algorithm called tabu-search is used.

Consistency of the profile likelihood estimator

In the work of Bickel & Chen (2009), the consistency of the profile-likelihood estimator, and of the broader class of the modularity estimators, was studied.
For the standard, non-degree corrected model, they showed that if no rows of $\boldsymbol{\eta}$ are equal and all classes have strictly positive weight, it can be proved that when the expected degree $\bar{d}$ grows fast enough with the respect to the size of the network, n, the profile-likelihood estimator is a strongly consistent estimator. Being more specific, we have that if $\lim_{n\to\infty} \frac{\bar{d}_n}{\log n} = \infty$, then the probability that $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ are up-to-label-swapping equal, tends to 1. Similarly, weak consistency is guaranteed (ie. the probability that the fraction of misclassified nodes is bounded by ϵ tends to 1, for any arbitrary $\epsilon > 0$) under the weaker condition $\lim_{n\to\infty} \bar{d}_n = \infty$. Zhao et al. (2012) proved that this result can be extended to the degree-corrected SBM, and that both the Newman-Girvan modularity estimator and the Erdös-Rényi modularity estimators require stronger restrictions for consistency.

Chapter 4: Simulation studies

In this chapter, results of the simulation studies developed for this thesis will be presented and analyzed. Data are simulated using three different stochastic block models: the standard binary undirected symmetric model, the degree-corrected binary undirected symmetric model, and the Poisson undirected symmetric model (as defined in Sections 1.7 and 1.9). Spectral methods and the variational expectation-maximization method (as defined in Chapter 3) are tested, and their quality in detecting the communities of a network is discussed. In Section 4.1, the settings of two independent simulation studies will be explained, and the results of these simulations will be illustrated in Sections 4.2 and 4.3 respectively. Conclusions concerning the simulation studies will be given in Section 4.4.

4.1 Simulation setting

Data were simulated from both the binary undirected symmetric model and the Poisson undirected symmetric model (see Section 1.7 equation 1.29), ie. models in which all K classes have on average the same weight, $\frac{1}{K}$:

$$\pi_k = \frac{1}{K} \quad \forall k = 1...K, \tag{4.1}$$

and the expected value of the edge weight between two nodes i and j verifies:

$$\mathbb{E}(Y_{ij}|Z_i = k, Z_j = l) = \theta_{kl} = \begin{cases} p_{in} & \text{if } k = l \\ p_{out} & \text{if } k \neq l. \end{cases}$$

Furthermore, data were also generated from the degree-corrected version of the binary symmetric model (see Section 1.9), which has been shown to be more realistic than its standard, uncorrected version. We remind the reader that in the degree-corrected model, the edge-weight between node i and node j is on average proportional to the node-specific parameters β_i and β_j:

$$\mathbb{E}(Y_{ij}|Z_i = k, Z_j = l) = \beta_i\beta_j\theta_{kl} = \begin{cases} \beta_i\beta_j p_{in} & \text{if } k = l \\ \beta_i\beta_j p_{out} & \text{if } k \neq l. \end{cases}$$

For each simulation, parameters β_i, $i = 1, ..., n$ were independently sampled from the same power law of parameter 5; ie. the continuous probability law whose density is given by:

$$f(x) = 4x^{-5}\mathbb{1}_{\{x>1\}}. \tag{4.2}$$

Power laws are used to model the nodes degree distribution in a network, because they display a "tail" representing hubs of the network. A histogram of a simulation from the power law described in equation (4.2) can be seen in Figure 4.1.

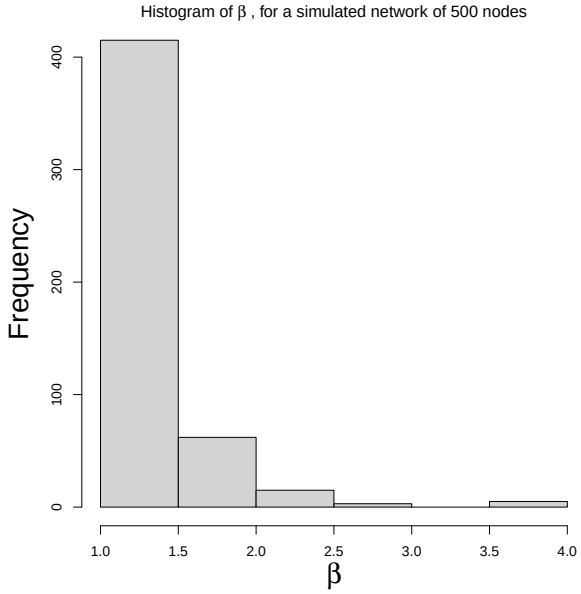


Figure 4.1: *A histogram of the values of parameters β_i, $i = 1, ...500$ for a network simulated from a degree-corrected stochastic block model with $n = 500$ nodes.*

The mean degree of the network, $\bar{d}$, was supposed to grow as in an intermediate regime, ie. logarithmically with respect to the network size, and controlled by the sparsity/density parameter s (see Section 1.8). Furthermore, the "out-in" ratio parameter is defined by $\gamma := \frac{p_{out}}{p_{in}}$, representing the ratio of the edge probability between nodes of the same class over the edge probability between nodes of different classes. This parameter will also be referred to as the *assortativity/disassortativity* parameter: the higher it is, the more the network is disassortative. The parameters n, K, γ and s allow then to specify exactly the models: specifically, by reminding ourselves of equations (1.30) and (1.33), we have that:

$$s = \frac{\bar{d}}{\log n} = \frac{n}{\log n} \frac{p_{in} + (K-1)p_{out}}{K}, \tag{4.3}$$

and, by definition :

$$\gamma = \frac{p_{out}}{p_{in}}. \tag{4.4}$$

Therefore, p_{in} and p_{out} are given by solving the system of equations:

$$\begin{cases} s = \frac{n}{\log n} \frac{p_{in}+(K-1)p_{out}}{K} \\ \gamma = \frac{p_{out}}{p_{in}} \end{cases}$$

whose solutions are

$$\begin{cases} p_{in} = \frac{Ks \log n}{n(\gamma(K-1)+1)} \\ p_{out} = \gamma \frac{Ks \log n}{n(\gamma(K-1)+1)}. \end{cases}$$

Because in a binary model we must have $p_{in} \leq 1$, parameters n, K, γ, s are subject to the following constraint:

$$\frac{Ks \log n}{n(\gamma(K-1)+1)} \leq 1,$$

or, equivalently:

$$\frac{Ks \, \frac{log(n)}{n} - 1}{(K-1)} \leq \gamma. \tag{4.5}$$

The lower bound for γ for the binary models is denoted by $\gamma_L^* := \frac{Ks \, \frac{log(n)}{n} - 1}{(K-1)}$.

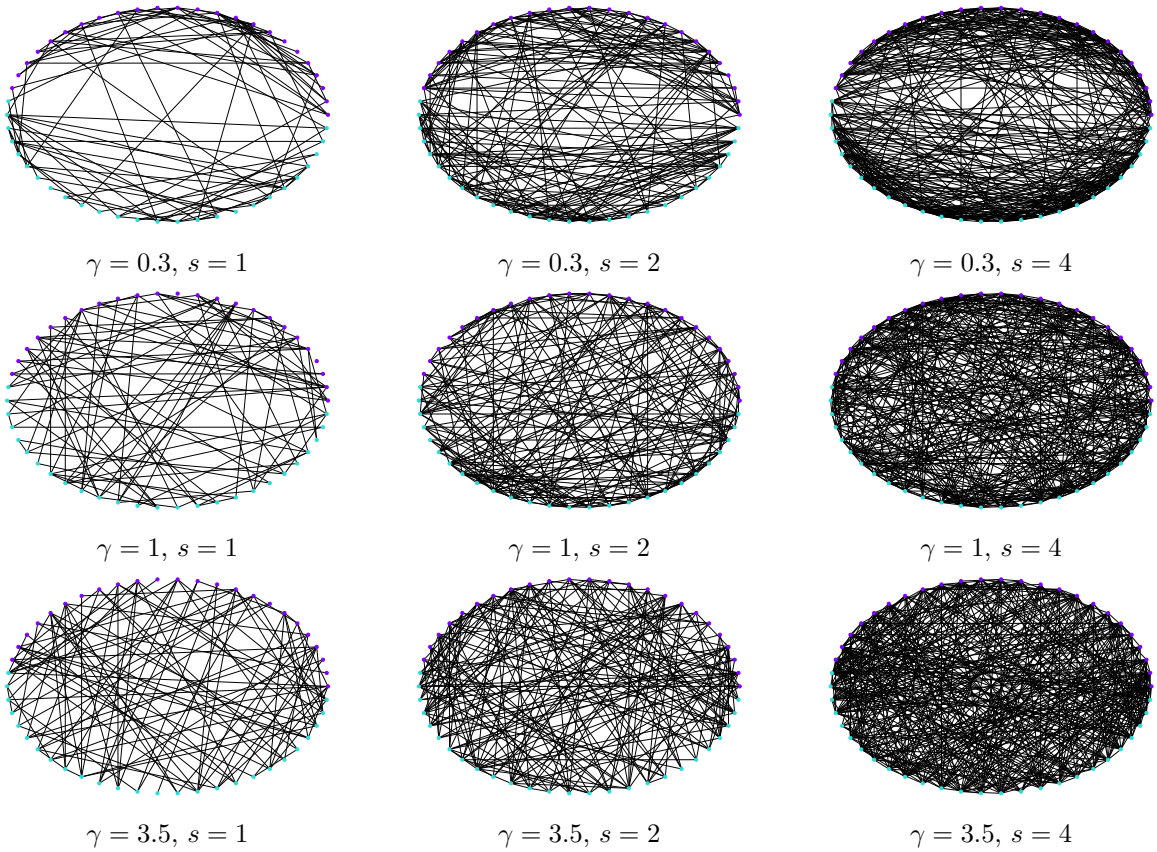


Figure 4.2: *Examples of networks simulated from a standard binary stochastic block model of parameters* $n = 50$ *nodes,* $K = 2$ *classes, in assortative scenarios* ($\gamma = 0.3$, *top row), boundary scenarios* ($\gamma = 0.1$, *middle row), disassortative scenarios* 3.5 ($\gamma = 3.5$, *bottom row). At each row, three different density scenarios are displayed: sparse scenarios* ($s = 1$, *left), moderately dense scenarios* ($s = 2$, *center), and dense scenarios* ($s = 4$, *right).*

In the first round of simulations (Section 4.2), the network size was set respectively to $n = 50$ and $n = 500$. The number of classes was set respectively to $K = 2$ and $K = 10$. The parameter s was set respectively to $s = 1$, $s = 2$ and $s = 4$. Finally, the last parameter,

the out-in ratio parameter, was set respectively to $\gamma = 0.1, \gamma = 0.3$, $\gamma = 1$, $\gamma = 3.5$ and $\gamma = 5$, representing respectively a strongly assortative scenario, a mildly assortative scenario, a boundary scenario in which $p_{in} = p_{out}$ (implying that one single community exists), a mildly disassortative scenario, and a strongly disassortative scenario. The boundary scenario should make the detection of the classes of the nodes very hard. Also, $s > 1$ is the condition for making the Erdös-Rényi network connected, meaning that there are no isolated components, with high probability as n tends to infinity (see Theorem 2). As such, $s = 1$ should be a hard scenario to detect communities, while $s = 4$ is an easier scenario, implying mean degrees of $4\log(50) \approx 15.5$ and $4\log(500) \approx 25$ for the case $n = 50$ and $n = 500$ respectively.

The performances of 5 different estimators were tested: the variational expectation-maximization algorithm (see Section 3.3), and four spectral methods. Two standard spectral methods based respectively on the Laplacian matrix and on the adjacency matrix, the regularized spectral method based on the Laplacian matrix (with regularization parameter set to $\alpha = 1$), and the spherical spectral method based on the adjacency matrix (see Section 3.5). For each configuration of parameters, 500 simulations were launched. For this first round of simulations, the number of classes/communities, ie. K, was supposed known.

Because for the first round of simulations, the number of communities was supposed known, a second smaller round of simulations was also launched (Section 4.3). In this second round of simulations, after performing estimation of several models corresponding to various values of K, the number of communities was estimated by using the Integrated Complete Likelihood, or ICL criterion (see equation 3.9). The configuration of the parameters for this second round of simulations was very similar as to the first round, but this time data were simulated only from the degree-corrected binary models, and K was set respectively to $K = 2$ and $K = 8$. Note that for this second round of simulations, data were not simulated from models with $K = 10$ communities, because this would have required too large a computation time in the estimation process. The setting of the other parameters was exactly the same as for the first round of simulations. For this second round of simulations, the variational expectation-maximization algorithm and three spectral methods were tested - the spectral method based on the Laplacian matrix, the regularized spectral method based on the Laplacian matrix, and the spherical spectral method based on the adjacency matrix. For this second round of simulation, only three spectral methods were tested because the first round of simulations proved only very minor differences among all spectral methods. Therefore, for each estimation method, $\hat{K}$ was set to several values, each value yielding a different estimate. Specifically, $\hat{K}$ was set to $\hat{K} = 1, 2, 3, 4$ when simulating data from a model with $K = 2$, while it was set to $\hat{K} = 1, 2, ..., 10$ when simulating data from a model with $K = 8$. Such choice in the setting of $\hat{K}$ was also dictated by computation time constraints, however it is not unrealistic if some previous information allows to exclude models with too many communities. The best estimate in term of the Integrated Complete Likelihood was selected. For this second round of simulations, 250 repetitions were launched for each configuration of parameters.

Note that for the binary model, for the configuration in which $n = 50$, $K = 8$, $s = 4$, the constraint given in equation (4.5) yields a lower bound for γ of approximately 0.21:

$$\gamma_L^* = \frac{Ks\,\frac{log(n)}{n} - 1}{(K-1)} = \frac{8 \cdot 4 \cdot \frac{log(50)}{50} - 1}{(8-1)} \approx 0.21. \tag{4.6}$$

As such, for binary models, the configuration in which $n = 50$, $K = 8$, $s = 4$, $\gamma = 0.1$ can not be tested, simply because there is no binary model with such a configuration of parameters. For the same kind of reason, the configuration in which $n = 50$, $K = 10$, $s = 4$, $\gamma = 0.1$ can not be tested neither.

All computations were performed using the *R* software. The functions `sampleSimpleSBM` from package `SBM` and `BlockModel.gen` from package `randnet` were used for generating data respectively from standard models and degree-corrected models. Furthermore, the functions `BM_bernoulli` and `BM_poisson` from the package `blockmodel` were used to perform the variational Expectation-Maximization algorithm for binary and Poisson data respectively; while the package `randnet` provided the functions `reg.SP` and `reg.SSP` for spectral methods.

Criteria for measuring the performance of an estimator: Agreement, Normalized Mutual Information and Rand Index

In this section, criteria will be defined for assessing the accuracy of a class-membership estimation, denoted by $\hat{\boldsymbol{z}} = (\hat{z}_1, ..., \hat{z}_n)^T$. For assessing the accuracy, different ways of measuring a "distance" between the estimated communities, as given by $\hat{\boldsymbol{z}}$, and the true communities, as given by $\boldsymbol{z}$, will de defined.
Ideally, a simple criterion would be to compute the mean number of nodes that are allocated to their true class; yielding the following criterion studied by Decelle et al. (2011), that we will call *overlap*:

$$overlap(\boldsymbol{z}, \hat{\boldsymbol{z}}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{z_i = \hat{z}_i\}}. \tag{4.7}$$

However, because of the unidentifiability of the classes, this quantity is not a suitable measure of the accuracy of the estimation. To see why such measure is not suitable, consider for example a network with $n = 8$ nodes, the first 4 nodes being in class 1 and the last 4 nodes in class 2; while $\hat{\boldsymbol{z}}$ allocates the first 4 nodes to class 2 and the last 4 nodes to class 1:

$$\boldsymbol{z}^0 = (1, 1, 1, 1, 2, 2, 2, 2) \quad \text{and} \quad \hat{\boldsymbol{z}}^0 = (2, 2, 2, 2, 1, 1, 1, 1).$$

The estimate $\hat{\boldsymbol{z}}$ recovers perfectly the group structure, except that labels are permuted. We can not hope from an estimation to do better than $\hat{\boldsymbol{z}}$ does, due to the unidentifiability of the labels; however, the overlap is 0:

$$overlap(\boldsymbol{z}^0, \hat{\boldsymbol{z}}^0) = \frac{1}{8} \sum_{i=1}^{8} \mathbb{1}_{\{z_i^0 = \hat{z}_i^0\}} = 0.$$

As such, different criteria must be found. One idea is computing the following quantity, that will be called *agreement* (Decelle et al., 2011):

$$agreement(\boldsymbol{z}, \hat{\boldsymbol{z}}) := \max_{\sigma \in Sym(K)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{z_i = \sigma(\hat{z}_i)\}}, \tag{4.8}$$

that considers all possible permutations of the labels of the estimated classes $\hat{\boldsymbol{z}}$, and then the better performance in terms of *overlap* is selected. This criterion is much better than the previous one; indeed, going back to the previous example, we have that

$$agreement(\boldsymbol{z}^0, \hat{\boldsymbol{z}}^0) = 1.$$

The agreement has however an important downside in practice: the number of label permutations, $K!$, grows very rapidly with K. So, for high values of K, it was not computed in this simulation. Remark that a normalized version of such agreement has also been defined in the literature (Decelle et al., 2011).

Another measure of the distance of $\hat{\boldsymbol{z}}$ from $\boldsymbol{z}$, used for example by Karrer & Newman (2011) and Amini et al. (2013), is given by the mutual information criterion. Indeed $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ yield each a probability distribution over the set of the class labels $[K] = \{1, ..., K\}$. The set of all probability distributions over the set of the class labels is denoted by $\mathcal{D}_{[K]}$. The true membership vector $\boldsymbol{z}$ yields the probability distribution $p_{\boldsymbol{z}} \in \mathcal{D}_{[K]}$ defined by $p_{\boldsymbol{z}}(k) := \frac{n_k}{n} \quad \forall k \in [K]$, where n_k is the number of nodes in class k; and $\hat{\boldsymbol{z}}$ yields the probability distribution $p_{\hat{\boldsymbol{z}}} \in \mathcal{D}_{[K]}$ defined by $p_{\hat{\boldsymbol{z}}}(k) := \frac{\hat{n}_k}{n} \quad \forall k \in [K]$, where $\hat{n}_k$ is the estimated number of nodes in class k.
Two probability distributions over the space $[K] \times [K]$ can be defined. Firstly, the joint probability distribution of $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$, that will be denoted by $p_{(\boldsymbol{z},\hat{\boldsymbol{z}})}$, defined by:

$$p_{(\boldsymbol{z},\hat{\boldsymbol{z}})}(k,l) := \frac{\sum_{i=1}^{n} z_{ik}\hat{z}_{il}}{n},$$

which considers the number of nodes that are allocated to class k by $\boldsymbol{z}$ and to class l by $\hat{\boldsymbol{z}}$; secondly, the product of the probability distributions $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$, that will be denoted by $p_{\boldsymbol{z}} \otimes p_{\hat{\boldsymbol{z}}}$, defined by:

$$(p_{\boldsymbol{z}} \otimes p_{\hat{\boldsymbol{z}}})(k,l) := \frac{n_k \hat{n}_l}{n^2}.$$

The (unnormalized) mutual information is defined by the Kullback-Leibler divergence (see equation 3.1) between the joint probability distribution $p_{(\boldsymbol{z},\hat{\boldsymbol{z}})}$ and the product probability distribution $p_{\boldsymbol{z}} \otimes p_{\hat{\boldsymbol{z}}}$:

$$MI(\boldsymbol{z}, \hat{\boldsymbol{z}}) := KL(p_{(\boldsymbol{z},\hat{\boldsymbol{z}})} || p_{\boldsymbol{z}} \otimes p_{\hat{\boldsymbol{z}}}) \tag{4.9}$$

which measures a 'distance' between the joint distribution and the product distribution. It is possible to normalize the previous quantity so that it always goes from 0 to 1, by dividing it by the sum of the Shannon entropies (see equation 3.6) of the two distributions; this yields the Normalized Mutual Information defined by:

$$NMI(\boldsymbol{z}, \hat{\boldsymbol{z}}) := \frac{2 \cdot KL(p_{(\boldsymbol{z},\hat{\boldsymbol{z}})} || p_{\boldsymbol{z}} \otimes p_{\hat{\boldsymbol{z}}})}{\mathcal{H}(p_{(\boldsymbol{z},\hat{\boldsymbol{z}})}) + \mathcal{H}(p_{\boldsymbol{z}} \otimes p_{\hat{\boldsymbol{z}}})}. \tag{4.10}$$

Another popular criterion is the Rand Index (Hubert & Arabie, 1985). To define this index, let yy denote the the number of (unordered) pairs of nodes (i, j) such that the true class of i and the true class of j are equal, and the estimated class of i and the estimated class of j are equal too:

$$\mathcal{G}_1 := |\{(i,j), i < j | z_i = z_j \text{ and } \hat{z}_i = \hat{z}_j\}|. \tag{4.11}$$

Let also $\mathcal{G}_2$ denote the number of false negatives, ie. the number of (unordered) pairs of nodes (i, j) such that the true class of i and the true class of j are different, and the estimated class of i and the estimated class of j are different too:

$$\mathcal{G}_2 := |\{(i,j), i < j | z_i \neq z_j \text{ and } \hat{z}_i \neq \hat{z}_j\}|. \tag{4.12}$$

Then, the Rand Index (Hubert & Arabie, 1985) is defined by:

$$RI(\boldsymbol{z}, \hat{\boldsymbol{z}}) := \frac{\mathcal{G}_1 + \mathcal{G}_2}{\binom{n}{2}}. \tag{4.13}$$

Therefore, the Rand Index is the number of pairs of nodes that either are both considered as being separated by $\boldsymbol{z}$ as well as by $\hat{\boldsymbol{z}}$; either are both considered as being put in the same cluster by $\boldsymbol{z}$ as well as by $\hat{\boldsymbol{z}}$; divided by the number of all possible pairs of nodes.
Hubert & Arabie (1985) also suggest using a normalized version of the Rand Index, the adjusted Rand Index:

$$ARI(\boldsymbol{z}, \hat{\boldsymbol{z}}) := \frac{RI(\boldsymbol{z}, \hat{\boldsymbol{z}}) - RI_{mean}}{RI_{max} - RI_{mean}}$$

where RI_{mean} and RI_{max} represent respectively the expected value and the maximal value of the Rand Index under the assumption that the sizes of the clusters of both $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$, respectively $\{n_k\}_{k\in[K]}$ and $\{\hat{n}_k\}_{k\in[K]}$, are fixed, and the nodes are randomly permuted.

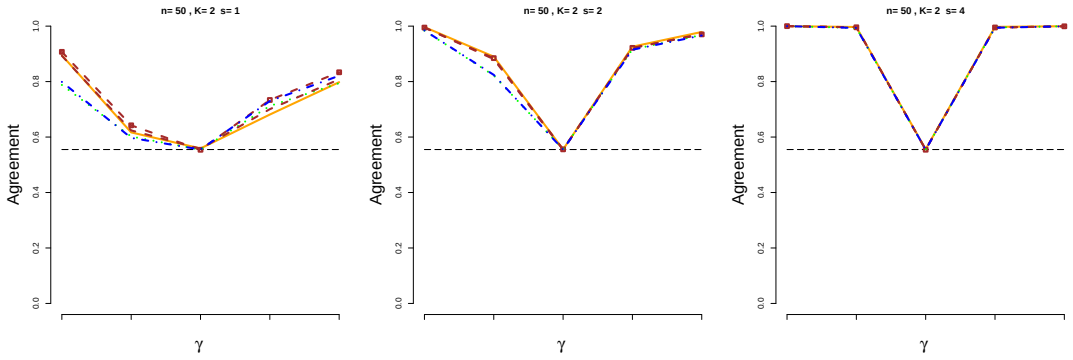
A reference estimator

The performances of estimators, as measured in terms of Mutual Information and the other criteria, are very sensible to the size of the network n and the number of classes K, and are therefore not directly comparable when varying n and K. Therefore, a reference estimator will be defined; the performance of such reference estimator will yield a term of comparison to assess the quality of the estimators when n and K change.

Such a reference estimator, that will be called the **uniform estimator**, allocates each of the n nodes to one of the K classes randomly and independently, with each class having the same probability weight. As such, the probability of assigning a node to its true class is $w = \frac{1}{K}$, ie. $w = 0.5$ for $K = 2$ and $w = 0.1$ for $K = 10$.

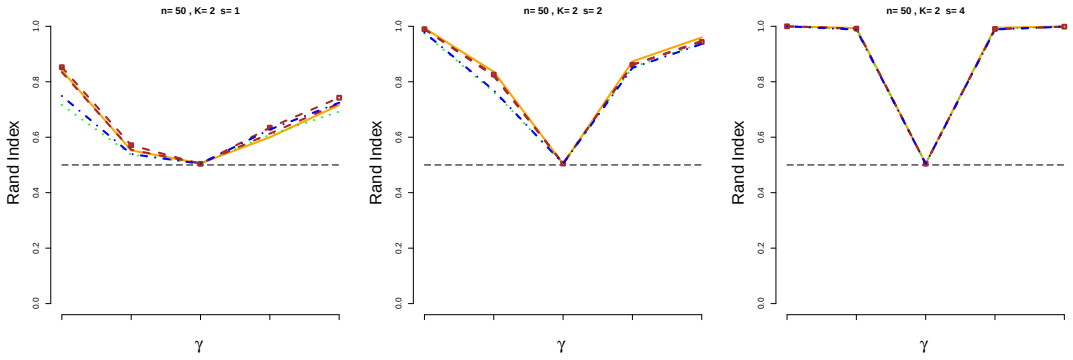
4.2 Results of the first round of simulations

Figures 4.3 to 4.8 display the results of the simulations from standard stochastic block models. These figures display the average quality of several estimators, as measured by the average Normalized Mutual Information, the average Rand Index, and for $K = 2$, also by the average Agreement, between their estimation of the class membership, and the true class membership. 500 simulations were launched, for all the combinations of parameters described above.
Figures 4.3 (displaying the Agreement), 4.4 (displaying the Rand Index) and 4.5 (displaying the Normalized Mutual Information), show that as the network becomes denser, the quality of the estimation increases for all estimators. Indeed, in dense scenarios the average mutual information is close to 1, meaning that we are very close to estimating perfectly the communities. In sparse scenarios, the mutual information is much lower. This can be explained by the fact that as the network gets sparser, it contains less information.
Furthermore, for $\gamma = 1$, ie. when $p_{in} = p_{out}$, we see that the estimators do not perform significantly better than the uniform estimator. This can be explained by the fact that when $\gamma = 1$, the network contains on average no information on the class structure, because all classes have exactly the same properties in terms of edge distributions.



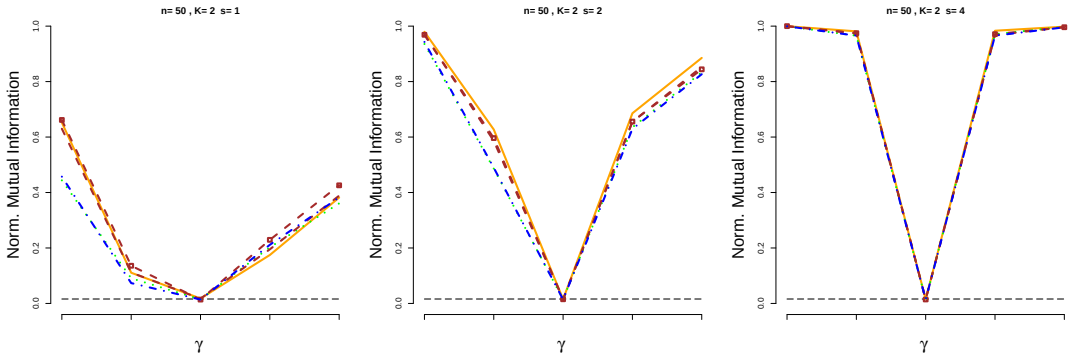
(a) Sparse (b) Moderately dense (c) Dense

Figure 4.3: *Average estimation quality, as measured by the Agreement, of several estimators over 500 simulations, from a binary SBM with $n = 50$, $K = 2$, in a sparse ($s = 1$), moderately dense ($s = 2$) and dense ($s = 4$) scenario, for five values of the out-in edge probability ratio γ: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*



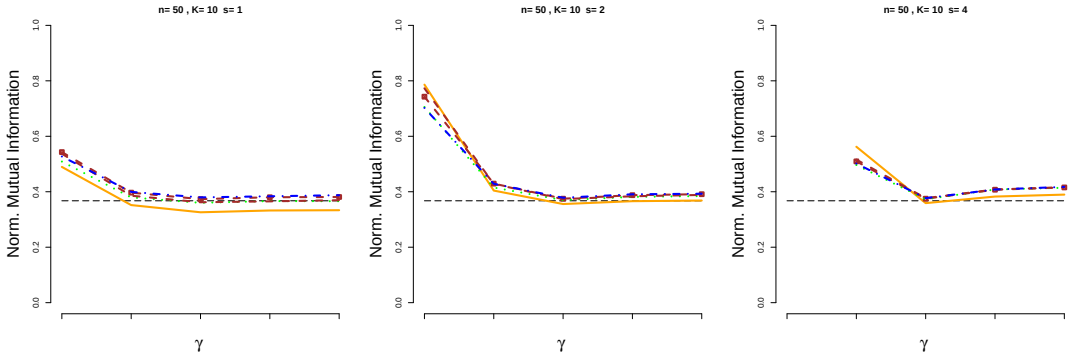
(a) Sparse (b) Moderately dense (c) Dense

Figure 4.4: *Average estimation quality, as measured by the Rand Index, of several estimators over 500 simulations from a binary SBM with $n = 50$, $K = 2$, in a sparse ($s = 1$), moderately dense ($s = 2$) and dense ($s = 4$) scenario, for five values of the out-in edge probability ratio γ: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*



(a) Sparse (b) Moderately dense (c) Dense

Figure 4.5: *Average estimation quality, as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a binary SBM with $n = 50$, $K = 2$, in a sparse (s=1), moderately dense (s=2) and dense (s=4) scenario, for five values of the out-in edge probability ratio γ: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator, allocating each node to its true class with probability $w = 0.5$ and $w = 0.75$ respectively.*



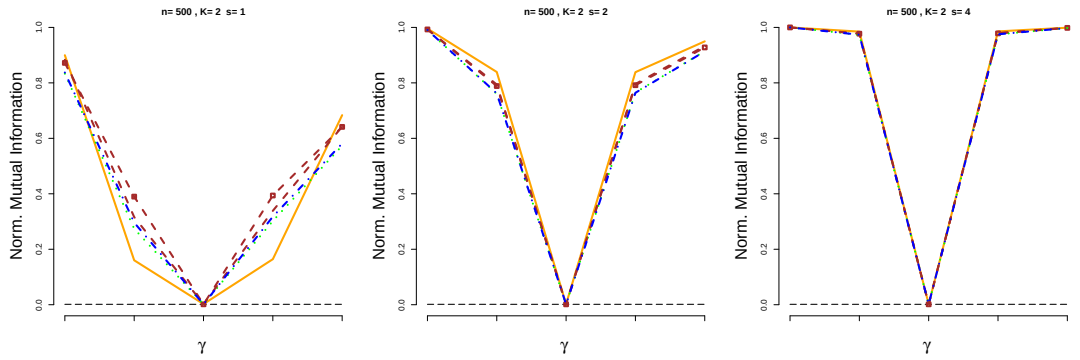
(a) Sparse (b) Moderately dense (c) Dense

Figure 4.6: *Average estimation quality, as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a binary SBM with $n = 50$, $K = 10$, in a sparse ($s = 1$), moderately dense ($s = 2$) and dense ($s = 4$) scenario, for five values of the out-in edge probability ratio γ: 0.1, 0.3, 1, 3.5 and 5 (except for the parameters configuration given by $n = 50$, $K = 10$, $s = 4$, for which the minimal tested value of γ was $\gamma = 0.3$, see equation (4.6)). Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*

Comparing Figures 4.5 and 4.6 shows that the quality of the estimation decreases when the number of classes increases, the mutual information being lower in Figure 4.6 than in Figure

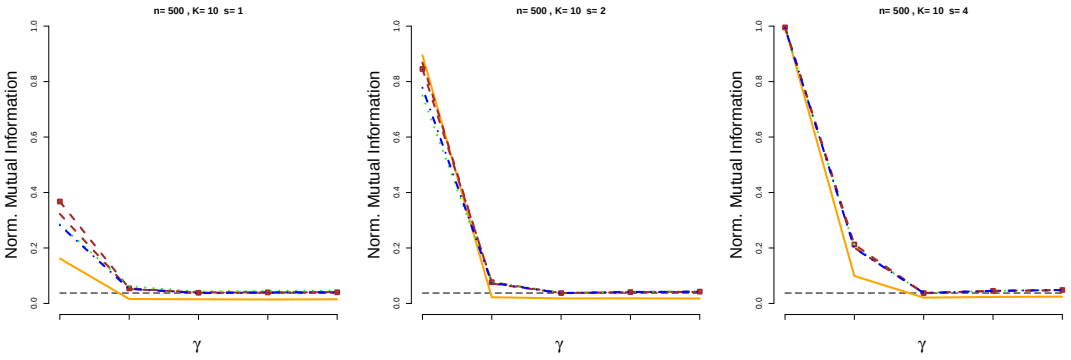
4.5; this is intuitive since as the number of classes increases, the size of the classes decreases (as the size of the network, n, is being kept constant), and so less information on the classes is available.
Furthermore, in the hardest scenario, ie. the scenario in which n is small, K is large and the network is sparse, the uniform estimator performs approximately as well as the other methods. This is explained by the fact that when K increases and n is constant, the average size of the clusters, n/K, tends to 0. Therefore, many communities will tend to be either empty or with very few edges, implying that an uniform random split will be close to the true split. In this scenario, the variational E-M method displays even a slightly lower mutual information than the uniform estimator. This is explained by the fact that when the network provides a low amount of information, the expectation-maximization algorithm tends to split the network in few large clusters and many small clusters, while spectral methods tend instead to produce balanced clusters.

Comparing Figures 4.5 and 4.7 shows that as the network increases in size, the quality of the estimation improves. This is normal, since it was assumed that the degree increases as the network size, n, grows. Such an improvement is however not spectacular, because it was assumed that the degree increases at a logarithmic rate with respect to the network size.



(a) Sparse (b) Moderately dense (c) Dense

Figure 4.7: *Average estimation quality, as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a binary SBM with* $n = 500$, $K = 2$, *in a sparse* ($s = 1$), *moderately dense* ($s = 2$) *and dense* ($s = 4$) *scenario, for five values of the out-in edge probability ratio* γ*: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*



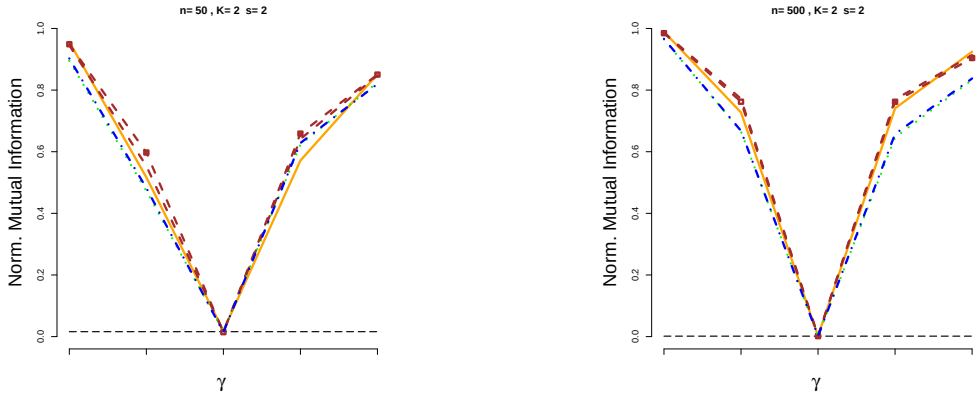
(a) Sparse (b) Moderately Dense (c) Dense

Figure 4.8: *Average estimation quality, as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a binary SBM with* $n = 500$*,* $K = 10$*, in a sparse* ($s = 1$)*, moderately dense* ($s = 2$) *and dense* ($s = 4$) *scenario, for five values of the out-in edge probability ratio* γ*: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*

By comparing Figures 4.6 and 4.8, the performances of the spectral estimators and the variational expectation-maximization estimator seem even, at first sight, to deteriorate when the network increases in size. However, the reference estimator displays a clear deterioration too.

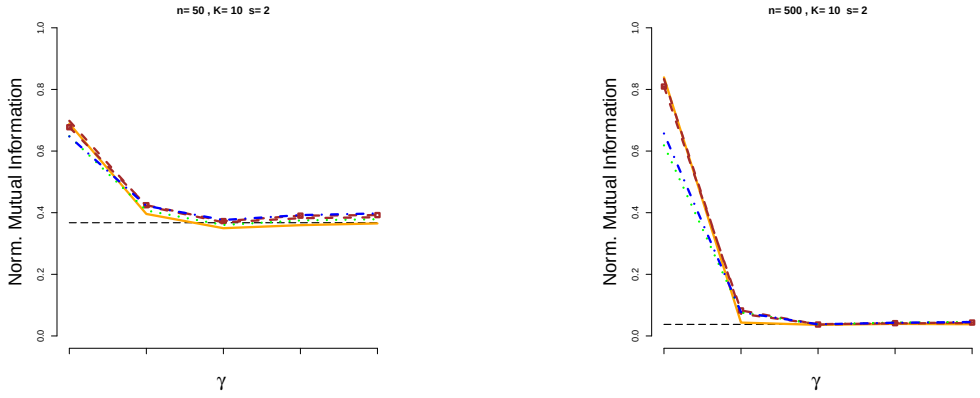
By looking at Figures 4.3 to 4.8, it appears that, in terms of normalized mutual information, the expectation-maximization algorithm performs at least as well the spectral methods when the network is moderately dense and moderately assortative, and K is not too large with respect to n. If the network gets disassortative, or very sparse, or K gets large with respect to n, both variational and spectral methods perform significantly worse, with the variational algorithm performing slightly worse than spectral methods. When classes are strongly distinguishable, all methods perform very well and no significative difference between methods appear. Among spectral methods, the regularized Laplacian spectral method performs at least as well, or slightly better than the other spectral methods.

Figures 4.9 and 4.10 show the results of the simulations from moderately dense degree-corrected binary block models. The performances of the estimators seem in general to be slightly worse if compared to those concerning simulations from the standard binary block model. For example, for the standard block model, Figure 4.6 displays, for the configuration $n = 50$, $K = 10$, $s = 2$, $\gamma = 0.1$, an average Normalized Mutual Information approaching 0.8 for all methods, while for the degree-corrected block model, for the same parameters configuration, Figure 4.10 displays a Normalized Mutual Information close to 0.7 for all methods. This phenomenon can be explained by the larger presence of hubs in networks simulated from the degree-corrected block models.



(a) $n = 50$ nodes (b) $n = 500$ nodes

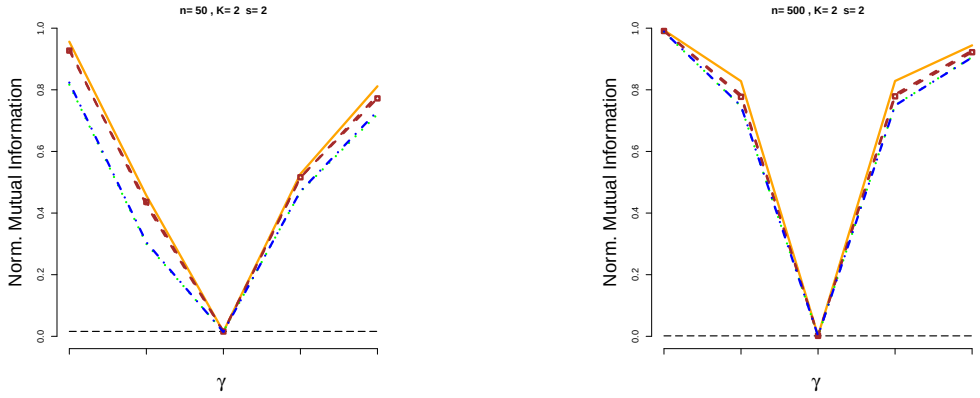
Figure 4.9: *Average estimation quality as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a moderately dense, $s = 2$, binary degree-corrected SBM with $K = 2$ communities, for a small sample size ($n = 50$ nodes) and a large simple size ($n = 500$ nodes) scenario, for five values of the out-in edge probability ratio γ: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*



(a) $n = 50$ nodes (b) $n = 500$ nodes

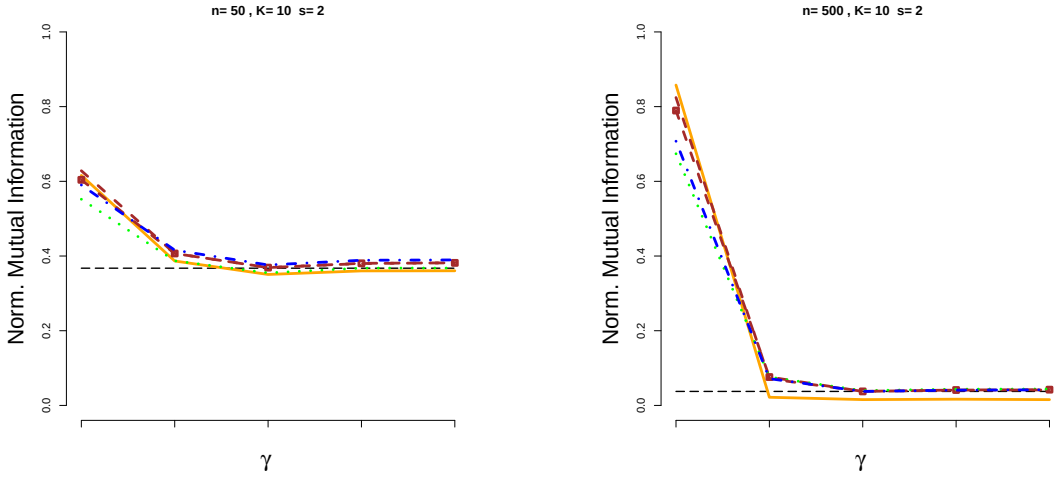
Figure 4.10: *Average estimation quality as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a moderately dense, $s = 2$, binary degree-corrected SBM with $K = 10$ communities, for a small sample size ($n = 50$ nodes) and a large simple size ($n = 500$ nodes) scenario, for five values of the out-in edge probability ratio γ: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*

Finally, results of the simulations from moderately dense Poisson block models are displayed in Figures 4.11 and 4.12. As for simulations from the degree-corrected models, it appears that the performances of the estimators slightly deteriorate with respect to simulations from the standard binary block model.



(a) $n = 50$ nodes (b) $n = 500$ nodes

Figure 4.11: *Average estimation quality as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a moderately dense, $s = 2$, Poisson SBM with $K = 2$ communities, for a small sample size ($n = 50$ nodes) and a large simple size ($n = 500$ nodes) scenario, for five values of the out-in edge probability ratio γ: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*

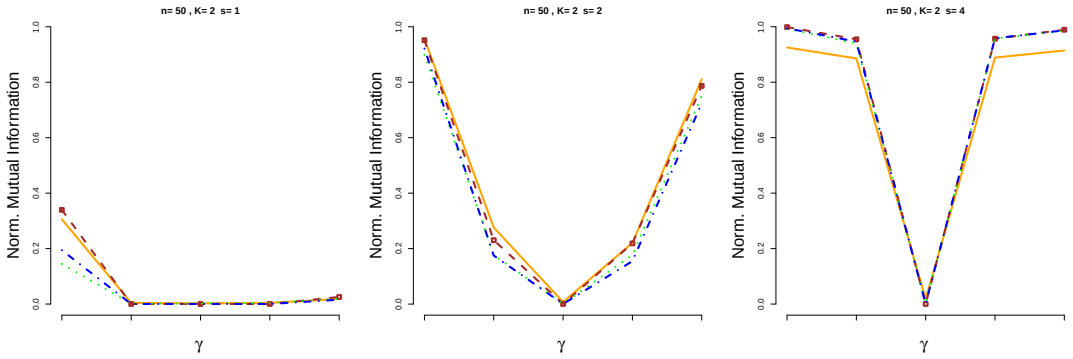


(a) $n = 50$ nodes (b) $n = 500$ nodes

Figure 4.12: *Average estimation quality as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a moderately dense, $s = 2$, Poisson SBM with $K = 10$ communities, for a small sample size ($n = 50$ nodes) and a large simple size ($n = 500$ nodes) scenario, for five values of the out-in edge probability ratio γ: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed lines: spectral estimators based on the Laplacian matrix, regularized and non-regularized (a square indicates the regularized version); green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix; black long-dashed line: the uniform estimator.*

4.3 Results of the second round of simulations

Some results of the second round of simulations are displayed in Figures 4.13 and 4.14. Generally, the results are very similar to the results of the first round of simulation, in terms of the effect of the parameters n, K, γ and s on the quality of the estimated models. The variational-expectation maximization algorithm seems now to perform slightly worse than the spectral methods, in the $K = 2$ dense scenario. In general, the results of the second round of simulations are comparable to the results of the first round of simulations. Therefore, the ICL criterion criterion seems to estimate soundly the number of communities.



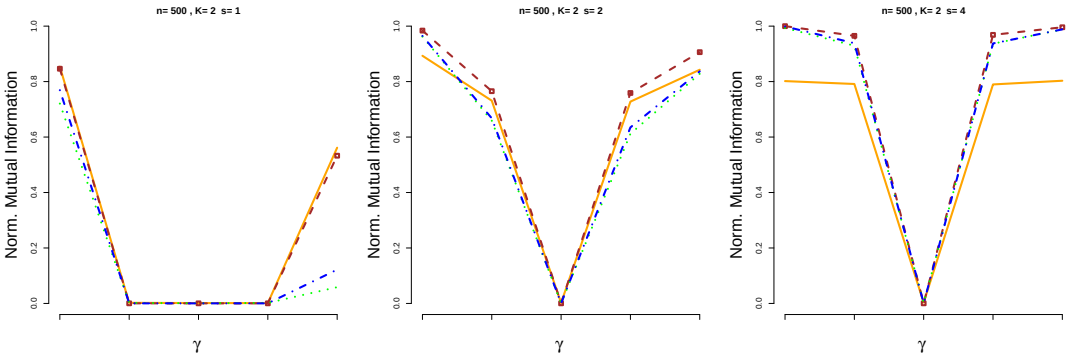
(a) Sparse (b) Moderately dense (c) Dense

Figure 4.13: *Average estimation quality as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a binary degree-corrected SBM with* $n = 50$, $K = 2$, *in a sparse* ($s = 1$), *moderately dense* ($s = 2$) *and dense* ($s = 4$) *scenario, for five values of the out-in edge probability ratio* γ*: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed line: regularized spectral estimator based on the Laplacian matrix; green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix. The number of communities to estimate was selected by the ICL criterion.*

4.4 Conclusions of the simulation study

In general, all methods tested in this simulation study have been shown to be able to estimate communities sufficiently well. As expected, the quality of the estimation is generally higher for smaller values of K and higher values of s and n. If the network is disassortative, the quality of the esimation is generally worse than when the network is assortative.

No large differences in the performance of the methods tested for this simulation study have appeared. In particular, only very minor differences between spectral methods have appeared. Also, the performance of the variational algorithm implementation used for this simulation has been proved to be comparable to spectral methods, even if in some scenarios, its performance deteriorates with respect to the spectral methods. The fact that no major difference between spectral methods and the variational algorithm have appeared may be explainable by the fact that the particular implementation of the variational algorithm which was used for this simulation study is based on spectral decomposition initialization. Among the tested spectral



(a) Sparse (b) Moderately dense (c) Dense

Figure 4.14: *Average estimation quality as measured by the Normalized Mutual Information, of several estimators over 500 simulations from a binary degree-corrected DSBM with* $n = 500$, $K = 2$, *in a sparse* ($s = 1$), *moderately dense* ($s = 2$) *and dense* ($s = 4$) *scenario, for five values of the out-in edge probability ratio* γ*: 0.1, 0.3, 1, 3.5 and 5. Orange continuous line: Variational E-M estimator; red dashed line: regularized spectral estimator based on the Laplacian matrix; green dotted line: spectral estimator based on the adjacency matrix; blue dashed-and-dotted line: spherical spectral estimator based on the adjacency matrix. The number of communities to estimate was selected by the ICL criterion.*

methods, the regularized spectral method based on the Laplacian matrix seems to perform slightly better than the other methods.

Chapter 5: Applications

In this chapter, three real datasets will be analyzed, and community detection will be performed by using the variational expectation-maximization method and the spectral methods (as defined in Chapter 3). The first dataset (Section 5.1) is a social network which was often studied in the literature (Bickel & Chen, 2009; Karrer & Newman, 2011; Zachary, 1977), the so-called Zachary karate club network. The second dataset (Section 5.2) is a network of political blogs concerning the 2004 United States presidential election, that was also analyzed in the works of Amini et al. (2013), Karrer & Newman (2011) and Zhao et al. (2012). The third dataset (Section 5.3) is unpublished and concerns a social network of students of the same school and of the same age.

5.1 Zachary karate club

The Zachary karate club dataset (Zachary, 1977) consists of an undirected friendship network of 34 members of a karate club of an university in the United States; the author took part to an observational study in which for three years he observed the interactions between club members in contexts such as attendance to the same tournaments, association in academic activities, interactions at the bar close to the karate room, etc., for a total of 8 different contexts. In the three years in which the author took part to the observational study, conflicts between club members led to the division of the club in two factions (one led by the instructor, called Mr. Hi by the author, the other led by an officer of the club) which eventually became two separate clubs. At the end of the observational study, the author counted the number of activities during which students interacted; therefore the strength of the relationship between two club members was measured by an integer number ranging from 0 to 8. Table 5.1 displays a part of the network. The complete 34×34 table can be found in the Appendix 7.12. Note here that the network table, as reported originally by Zachary (1977), displays a few inconsistencies. Indeed, the table is not exactly symmetrical. For example, the strength of the relationship between members 1 and 13 is uncertain, because the entry at row 1 and column 13 is 2, while the entry at row 13 and column 1 is 3. In total, however, only 7 entries out of $\binom{34}{2} = 578$ display similar inconsistencies. Furthermore, because only the binary version of the network (see Table 5.2) will be considered for the next analysis, these inconsistencies have no influence, except for the couple of members 23 and 34, for which it is uncertain if a relationship exists.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		4	5	3	3	3	3	2	2		2	3	2	3				2		2
2	4		6	3				4						5				1		2
3	5	6		3				4	5	1				3						
4	3	3	3					3					3	3						
5	3						2				3									
6	3						5				3						3			
7	3				2	5											3			
8	2	4	4	3																
9	2		5																	
10			1																	
11	2				3	3														
12	3																			
13	1			3																
14	3	5	3	3																
15																				
16																				
17						3	3													
18	2	1																		
19																				
20	2	2																		

Table 5.1: *The Karate club network weighted adjacency matrix - only the first 20 members out of 34 are displayed. The strength of the relationships is measured by an ordinal scale ranging from 0 to 8. Source: Zachary (1977). The complete 34 × 34 table can be found in the Appendix 7.12.*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		1	1	1	1	1	1	1	1		1	1	1	1				1		1
2	1		1	1				1						1						1
3	1	1		1				1	1					1						
4	1	1	1					1					1	1						
5	1						1				1									
6	1						1				1						1			
7	1				1	1											1			
8	1	1	1	1																
9	1		1																	
10																				
11	1				1	1														
12	1																			
13				1																
14	1	1	1	1																
15																				
16																				
17						1	1													
18	1																			
19																				
20	1	1																		

Table 5.2: *The Karate club network adjacency matrix, binary version - only the first 20 members out of 34 are displayed.*

The variational expectation-maximization algorithm and three different spectral methods - the regularized Laplacian matrix spectral method, the standard Adjacency matrix spectral method and the spherical Adjacency matrix method, were used to estimate the network structure. For all models, the number of communities to estimate was supposed to be $K = 2$.

All three spectral methods yield the same estimate, which corresponds exactly with the true factions, except only for member 9. The variational expectation-maximization algorithm performs poorly, by roughly splitting the club members according to their degree. Table 5.3 display these results. Note that the variational expectation-maximization method splits the members between those with a degree higher than 6 and a degree smaller or equal to 6. The performance of the spectral methods correspond exactly with the results obtainted by using the degree-corrected SBM profile likelihood estimator in the work of Karrer & Newman (2011), while the performance of the variational expectation-maximization method corresponds exactly with the results obtainted by using the standard SBM profile likelihood estimator (Karrer & Newman, 2011).

Member	Degree	Faction	Spectral methods estimate	Var-EM estimate
1	16	Mr. Hi	A	A
2	8	Mr. Hi	A	A
3	9	Mr. Hi	A	A
4	6	Mr. Hi	A	B
5	3	Mr. Hi	A	B
6	4	Mr. Hi	A	B
7	4	Mr. Hi	A	B
8	4	Mr. Hi	A	B
9	5	Mr. Hi	B	B
10	1	Officer	B	B
11	3	Mr. Hi	A	B
12	1	Mr. Hi	A	B
13	2	Mr. Hi	A	B
14	5	Mr. Hi	A	B
15	2	Officer	B	B
16	2	Officer	B	B
17	2	Mr. Hi	A	B
18	1	Mr. Hi	A	B
19	1	Officer	B	B
20	2	Mr. Hi	A	B
21	1	Officer	B	B
22	2	Mr. Hi	A	B
23	1	Officer	B	B
24	5	Officer	B	B
25	3	Officer	B	B
26	3	Officer	B	B
27	2	Officer	B	B
28	4	Officer	B	B
29	3	Officer	B	B
30	4	Officer	B	B
31	4	Officer	B	B
32	6	Officer	B	B
33	11	Officer	B	A
34	14	Officer	B	A

Table 5.3: *The Karate club network degree distribution, true factions and estimated factions. Estimation was performed using three spectral methods, (the regularized Laplacian matrix spectral method, the standard adjacency matrix spectral method, and the spherical adjacency matrix method), and the variational expectation-maximization method.*

5.2 Political blogs network

In this section, a political blogs network will be studied. This dataset was already analyzed in the works from Amini et al. (2013), Karrer & Newman (2011) and Zhao et al. (2012), in which profile likelihood, modularity and pseudo-likelihood methods were tested. The network comprises a total of 1222 blogs about the 2004 United States presidential election, and the web links between them, as measured on the same day in 2004. Note that these 1222 blogs are part of a bigger network of 1490 blogs; however, the remaining 268 blogs being isolated from the main component of the network, they are disregarded. The network dataset was originally compiled, and blogs were labeled as left-leaning or right-leaning, in the work from Adamic & Glance (2005) (as cited by Amini et al. (2013), Karrer & Newman (2011) and Zhao et al. (2012)). The complete dataset was downloaded from the R library `networkdata`. The network is balanced, with 758 blogs labelled as left-wing and 732 as right-wing. The undirected network was considered, so that an edge between two blogs means that at least one of the two blogs has a link to the other.

Figure 5.1 shows that the degree distribution is highly skewed; indeed, most of the blogs display less than 20 links; however, there are also blogs with over 200 and even 300 links.

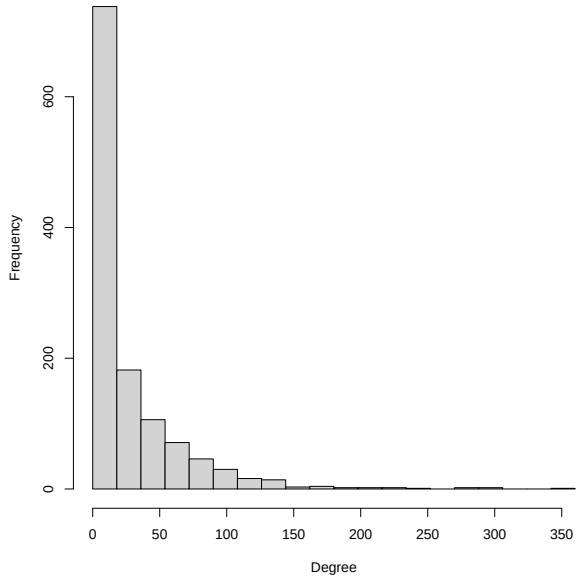


Figure 5.1: *Degree distribution for the 1222 blogs of the network.*

Table 5.4 shows that the edge density between left-wing blogs is approximately 0.04, while it is slightly lower for right-wing blogs. The edge density between blogs of opposite leanings is approximately 0.004.

	Left-wing	Right-wing
Left-wing	.0426	.0042
Right-wing	.0042	.0388

Table 5.4: *True edge probability matrix for the political blogs network.*

Table 5.5 shows that the estimation quality of the spherical spectral method is very good. The other methods perform much worse. In particular, the variational expectation-maximization estimator performs very poorly. The poor quality of most estimators may be explained by the large variability in the degree distribution.

	NMI	Rand Index	Agreement
VAR-EM	.00	.50	.52
Laplacian matrix, regularized	.24	.57	.69
Adjacency matrix	.18	.54	.64
Adjacency matrix, spherical	.71	.90	.95

Table 5.5: *Estimation quality of the political blogs classes, as measured by the normalized mutual information, the Rand Index and the Agreement, of the variational expectation-maximization algorithm and three different spectral estimation methods. The number of communities to detect was supposed to be* $K = 2$.

The estimated edge-probability matrix, displayed in Table 5.6 is also close to the true edge-probability matrix.

	Left-wing	Right-wing
Left-wing	.0475	.0033
Right-wing	.0033	.0369

Table 5.6: *Estimated edge-probability matrix for the political blogs network, via spherical adjacency matrix spectral decomposition.*

5.3 Students friendship network

In this section, a friendship network dataset of students will be studied. The dataset concerns 196 senior year students of the same secondary school in the same academic year. This school welcomes students coming from several countries.

This school is organised in "linguistic sections". Each student is enrolled in one linguistic section, generally corresponding to his/her mother tongue. Students are given some courses inside their linguistic section, ie. in their mother tongue, while other courses are taught in a different language. For example, students of the Italian section follow subjects such as Italian - First language, Mathematics, Biology, etc., in Italian. However, courses of other subjects such as History, Geography, Second language, Third language, Economics, Arts, Music, etc. are organised in groups of students coming from different linguistic sections (the number of such subjects increases as the pupil goes through the years). In these courses, instructors teach in one language between English, French, and German, which eventually serve as "lingua franca" facilitating friendship contacts among students of different sections. Indeed, learning at least one of these three languages is mandatory since the first year of primary school.

The friendship network data were gathered from a book, printed by the students at the end of their last academic year. Such book contains messages of students to each other. Each page of the book is devoted to one of the students, and includes all messages written to him/her by fellow students. All messages are signed with a signature such as name + first letter of the surname, or only the first name, or, seldomly, with a nickname, so that some messages could not be attributed to the true author.

In total, the book considered in this study contains data of 196 senior-year students and a total of 2764 messages. Of these 2764 messages, 249 messages were discarded: 51 of these were written by "external" students not in their senior-year. Therefore such messages were considered as irrelevant because outside of the network under study. For the other 198 messages, which were signed with nicknames, it was not possible to identify the author, who could be either an unidentified senior year student or an "external" student. The 2515 messages were summarized in a 196×196 binary table. Therefore, the accuracy of the network can be estimated to be over $\frac{2515}{2515+198} = 92\%$.

The undirected network of the senior year students as reported in the above described table was considered in this study. For each couple of students, an undirected edge was recorded when at least one of the two students sent a message to the other. The total number of recorded undirected edges is 1820.

The number of students in each linguistic section - German, English, Finnish, French, Italian, Dutch, Portuguese, Swedish, per gender, is reported in Table 5.7. Note that, because the French section includes many students, it is divided in three sections: French-A, French-B and French-C sections. Note also that there is no exact correspondance between linguistic sections and nationalities: for example, the English sections include many Irish students, the French and the Dutch sections include many Belgian students, the German section includes many Austrian students. Furthermore, there are few students who are in sections not corresponding with their mother tongue: for example, the French section "historically" includes several

Greek students (no specific Greek section is present in the school under study). However, for simplicity, we will often refer to the students of the French sections as the French students, to the students of the English sections as the English students, etc.

	DE	ENA	ENB	FI	FRA	FRB	FRC	IT	NL	PT	SW	Total
F	6	11	7	2	13	5	8	5	11	9	7	84
M	13	9	7	7	7	11	5	13	9	19	12	112
Total	19	20	14	9	20	16	13	18	20	28	19	196

Table 5.7: *Number of students in each linguistic section, per gender (M=male, F=female).*

Table 5.7 shows that the Finnish, Italian, Portuguese and German sections include a strong majority of boys, while the French-A section includes a strong majority of girls. The other sections display approximately a balance of boys and girls.

The average degree of the network, ie. the average number of friendship relationships, as measured by the bac-book messages, per student, is $\frac{2 \cdot 1820}{196} = 18.5$ (see Figure 5.2). As expected, the distribution of such degree is slightly right-skewed. No student has less than 2 friendship ties. 4 students have 2 or 3 friendship ties, and 3 students have 40 or more friendship ties.

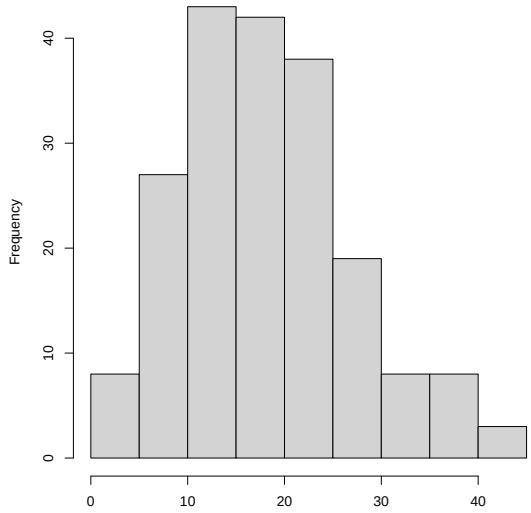


Figure 5.2: *Degree distribution of the students network.*

Table 5.8, displaying the edge probabilities between sections, shows that friendship relationships are generally much more frequent inside linguistic sections. In particular, the Italian (0.80), Portguese (0.65) and Swedish (0.68) sections display very high values of intra-section edge probability; while other sections as the Dutch section (0.47) and the French sections (between 0.12 and 0.51) display lower values. Furthermore, the links across the sections sharing the same language (French-A, French-B and French-C; English-A and English-B), are generally stronger than across other sections. Other across-sections edge probabilities are often below 0.05. Note that the northern countries display slightly higher values: indeed, the

	DE	ENA	ENB	FI	FRA	FRB	FRC	IT	NL	PT	SW
DE	.58	.04	.05	.04	.01	.02	.03	.05	.05	.03	.06
ENA	.04	.48	.27	.04	.04	.02	.02	.01	.07	.02	.04
ENB	.05	.27	.45	.06	.07	.05	.06	.05	.06	.05	.07
FI	.04	.04	.06	.42	.03	.01	.03	.01	.10	.02	.16
FRA	.01	.04	.07	.03	.40	.12	.21	.03	.03	.05	.03
FRB	.02	.02	.05	.01	.12	.23	.14	.00	.03	.02	.04
FRC	.03	.02	.06	.03	.21	.14	.51	.04	.06	.03	.08
IT	.05	.01	.05	.01	.03	.00	.04	.80	.04	.02	.03
NL	.05	.07	.06	.10	.03	.03	.06	.04	.47	.05	.09
PT	.03	.02	.05	.02	.05	.02	.03	.02	.05	.65	.02
SW	.06	.04	.07	.16	.03	.04	.08	.03	.09	.02	.68

Table 5.8: *Edge-probability matrix between linguistic sections.*

Finnish-Swedish edge-probability is 0.16, the Finnish-Dutch edge-probability is 0.10, and the Swedish-Dutch edge-probability is 0.09. This may be explained by the fact that students of the northern countries sections usually have excellent English skills. These skills are also due to the fact that most of them study English as their second language. The barrier between students of northern countries seems then to be lower than between students of other countries. Furthermore, the Finnish section is very small, including only 9 students; this may bring Finnish students to create more friendship ties with students from other sections.

The variational expectation-maximization method and three different spectral methods were used to estimate the community structure of the students network. For each method, several models for K ranging from 1 to 20 were tested, and the optimal value of K was selected using the ICL criterion. Such an optimal value is displayed in Table 5.9

	Number of communities
VAR-EM	11
Lap. matrix, reg.	6
Adj. matrix	8
Adj. matrix, sph.	7

Table 5.9: *Estimated number of communities by the variational expectation-maximization algorithm and three spectral methods.*

Each of the four models overlaps greatly with the 8 linguistics sections structure. Indeed, Table 5.10 shows high values of the criteria measuring the similarity between the estimated community structure and the linguistic sections structure.

	NMI	Rand Index	Agreement
VAR-EM	0.75	0.92	0.73
Lap. matrix, reg.	0.87	0.95	0.84
Adj. matrix	0.86	0.96	0.91
Adj. matrix, sph.	0.88	0.97	0.91

Table 5.10: *Normalized Mututal Information, Rand Index and agreement between the 8 linguistic sections and the community structure estimated by the variational expectation-maximization algorithm and three spectral methods.*

Tables 5.11 to 5.14 display the estimated students communities, in terms of the linguistic sections membership. The communities overlap sensibly with the linguistic sections. Indeed, Table 5.12 displays a 7-communities model corresponding almost perfectly to the 7 major linguistic sections. However, as already observed, the students from the Dutch, Swedish and Finnish sections seem to mix more than other students : in particular, the Dutch and the Swedish sections are considered part of a single community in the 6-communities model (Table 5.11), and there is no exclusively Finnish community in any model. This may be due to the fact that Finnish students make a very small group of only 9 students. Interestingly, the 8-communities model (Table 5.13) finds a single-student community, ie. community 3 : the student in this community is a girl of the Dutch section which is involved in many extra-academic school activities and is known by most of the students, displaying indeed more than 40 friendship ties, and is therefore a hub.

	DE	EN	FI	FR	IT	NL	PT	SW
Community 1	19	0	0	0	0	0	0	0
Community 2	0	34	2	1	0	0	0	0
Community 3	0	0	7	2	0	19	0	19
Community 4	0	0	0	46	0	1	0	0
Community 5	0	0	0	0	18	0	0	0
Community 6	0	0	0	0	0	0	28	0

Table 5.11: *Communities of the students network, as estimated by the Laplacian matrix regularized spectral decomposition method, in terms of their linguistic section.*

	DE	EN	FI	FR	IT	NL	PT	SW
Community 1	19	0	0	0	0	2	0	0
Community 2	0	34	3	1	0	0	0	0
Community 3	0	0	6	1	0	3	0	19
Community 4	0	0	0	46	0	0	0	0
Community 5	0	0	0	0	18	0	0	0
Community 6	0	0	0	0	0	0	28	0
Community 7	0	0	0	1	0	15	0	0

Table 5.12: *Communities of the students network, as estimated by the adjacency matrix spherical spectral decomposition method, in terms of their linguistic section.*

	DE	EN	FI	FR	IT	NL	PT	SW
Community 1	18	0	0	0	0	0	0	0
Community 2	0	32	2	1	0	0	0	0
Community 3	0	0	0	0	0	1	0	0
Community 4	0	1	3	47	0	0	1	0
Community 5	0	0	0	0	18	0	0	0
Community 6	1	1	3	0	0	18	0	0
Community 7	0	0	0	0	0	0	27	0
Community 8	0	0	1	1	0	1	0	19

Table 5.13: *Communities of the students network, as estimated by the adjacency matrix spectral decomposition method, in terms of their linguistic section.*

	DE	EN	FI	FR	IT	NL	PT	SW
Community 1	14	0	0	0	0	0	0	0
Community 2	0	25	1	1	0	0	0	0
Community 3	4	0	1	0	0	5	0	0
Community 4	0	0	0	33	0	0	0	0
Community 5	0	0	0	0	18	0	0	0
Community 6	0	1	4	0	0	12	0	1
Community 7	0	0	0	0	0	0	28	0
Community 8	0	0	3	0	0	2	0	13
Community 9	0	2	0	8	0	0	0	0
Community 10	0	0	0	7	0	0	0	2
Community 11	1	6	0	0	0	1	0	3

Table 5.14: *Communities of the students network, as estimated by the variational expectation-maximization method, in terms of their linguistic section.*

The model in Table 5.14 displays 11 communities, 7 of which correspond roughly to the 7 major linguistic sections; these are communities 1,2,4,5,6,7,8. Furthermore, communities 3, 9, 10 and 11 are "cliques" of students coming from different linguistic sections. Community 3 is made up chiefly of German and Dutch students. Community 9 is made chiefly of French and English students. Community 10 is made of French students and 2 Swedish students. Most of the students in community 9 wrote messages to each other concerning memories of parties organized in different clubs. Community 11 is a mix of students from northern Europe. Most of students in community 11 wrote messages to each other concerning memories of ski holidays spent together.

The four models themselves overlap significantly with each-other, in particular the three spectral methods, as can also be seen in Table 5.15.

It has been shown that sections play a very important role in the creation of social links between students; however, students coming from the sections of northern countries seem to mix more than students from other sections. This is particularly true for Finnish students, but also Dutch, English and Swedish students are likely to establish relationships outside of their

section. On the opposite, Italian and Portuguese students tend to create strong relationships inside their section, but much less outside of it.

	VAR-EM	Lap. matrix, reg.	Adj. matrix	Adj. matrix, sph.
VAR-EM	1.00	0.76	0.75	0.75
Lap. matrix, reg.	0.76	1.00	0.85	0.92
Adj. matrix	0.75	0.85	1.00	0.85
Adj. matrix, sph.	0.75	0.92	0.85	1.00

Table 5.15: *Similarity between estimated friendship communities, using the variational expectation-maximization algorithm, and three different spectral decomposition estimation methods, as measured by the Normalized Mutual Information.*

Conclusions and discussion

In the framework of this thesis, several properties of the stochastic block models were summarized and discussed. Simulations from these models were performed showing that the variational expectation-maximization method and the spectral methods are able to detect communities in a reliable way. Furthermore, these methods were used to detect communities in three different real-life networks of increasing complexity in terms of the size of the network and of the number of communities. The results of these studies have shown that, in general, these models are able to detect the communities in a sound way, confirming that stochastic block models are powerful tools for modelling networks. However, as emerged in the study of the two first real datasets, the variational expectation-maximization method may tend to produce inaccurate results due to the presence of hubs in the network, while its performance is comparable to spectral methods for networks in which few hubs are present. Indeed, both spectral methods and the variational expectation-maximization method were able to correctly detect communities of a relatively complex social network never studied before, made of several communities of students of different origin and linguistic background. It has also been shown that stochastic block modelling requires specific methods, which are often faced with computational challenges in terms of time and complexity. These computational challenges are also due to unidentifiability issues, and hinder the analysis of networks of larger size.

However, at the current state-of-the art, newer and more elaborated methods are being studied in order to detect communities of larger networks, such as methods based on semi-definite programming relaxations of the maximum-likelihood estimator (Amini & Levina, 2018) and pseudo-likelihood methods (Amini et al., 2013). These newer approaches would have the advantage of sensibly shortening the computational time, allowing to detect communities also on larger networks.

Appendix

7.1 Basic distributions

Multinomial distribution

This probability distribution generalizes the binomial distribution.

Let $n \in \mathbb{N}, k \in \mathbb{N}$ be two positive integers, with n being the number of trials and k the number of classes, and $\pi \in \mathbb{R}_+^k$ a k-dimesional vector of strictly positive numbers with $\sum_{i=1}^k \pi_i = 1$. A Multinomial distribution of parameters (n, π) is a discrete probability distribution, of probability-mass :

$$f(x_1, ..., x_k | n, \pi) = \frac{n!}{x_1!\, x_2! \cdots x_k!} \pi_1^{x_1} ... \pi_k^{x_k} \mathbb{1}_{\sum_i x_i = n}$$

for $\{x_i\}_{i=1...k}$ non-negative integers.

In the case where $n = 1$, this distribution describes a scenario in which X_i 's are binary variables of value 0 or 1, and events $\{X_i = 1\}_i$ are mutually exclusive, ie. $P(X_i = 1 \cap X_j = 1) = 0 \ \forall i \neq j$, with each of them having a probability π_i: $P(X_i = 1) = \pi_i$.
Multinomial distributions of parameter (n, π) will also be referred to as n-Multinomials. In practice, only 1-Multinomials distributions are of interest in this text.

Beta distribution

Let $\alpha \in \mathbb{R}_+$ and $\beta \in \mathbb{R}$ be two strictly positive real numbers. The Beta distribution of parameters (α, β) is defined by the density function :

$$f(x | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1} \cdot (1 - x)^{\beta - 1} \mathbb{1}_{x \in [0,1]}$$

where Γ is the standard Gamma function defined by $\Gamma(z) := \int_0^{+\infty} t^{z-1} e^{-t} dt$

Dirichlet distribution

Let $(\alpha_1, \alpha_2, ..., \alpha_k)^T \in \mathbb{R}_+^k$ be a vector of strictly positive real numbers. The Dirichlet distribution of parameters $(\alpha_1, \alpha_2, ..., \alpha_k)$ is defined by the density function :

$$f(x_1, x_2, ..., x_k | \alpha_1, ..., \alpha_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_k)} x_1^{\alpha_1} x_2^{\alpha_2} ... x_k^{\alpha_k}$$

for all $(x_1, x_2, ..., x_k)$ such that $\sum_{i=1}^{k} x_i = 1$, $x_i \geq 0 \quad \forall i \in \{1, 2, ..., k\}$, and where Γ indicates the standard Gamma function.

Bernoulli distribution

Let $\alpha \in [0, 1]$. A Bernoulli distribution of parameter α is a discrete probability distribution, of probability mass function:

$$f(x|\alpha) = \alpha^x (1 - \alpha)^{1-x} \mathbb{1}_{x \in \{0,1\}}$$

implying

$$\mathbb{P}(X = 1) = \alpha = 1 - P(X = 0).$$

7.2 Proof of the formula for the expectation of the sufficient statistics for an exponential family

In this section, it will be proved that, for an exponential family model, $\mathbb{E}_{\boldsymbol{\phi}} t_i(\boldsymbol{X}) = \frac{\partial}{\partial \phi_i} a(\boldsymbol{\phi})$ for any $i = 1, 2, ..., d$.

First, let remind that $\frac{\partial}{\partial \phi_i} \log f(\boldsymbol{x}|\boldsymbol{\phi}) = t_i(\boldsymbol{x}) - \frac{\partial}{\partial \phi_i} a(\boldsymbol{\phi})$.
Therefore, we have that :

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\phi}} t_i(\boldsymbol{X}) - \frac{\partial}{\partial \phi_i} a(\boldsymbol{\phi}) &= \int t_i(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} - \frac{\partial}{\partial \phi_i} a(\phi) \int f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \\
&= \int t_i(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} - \int \frac{\partial}{\partial \phi_i} a(\boldsymbol{\phi}) f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \\
&= \int (t_i(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} - \frac{\partial}{\partial \phi_i} a(\boldsymbol{\phi}) f(\boldsymbol{x}|\boldsymbol{\phi})) d\boldsymbol{x} \\
&= \int (t_i(\boldsymbol{x}) - \frac{\partial}{\partial \phi_i} a(\boldsymbol{\phi})) f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \\
&= \int (\frac{\partial}{\partial \phi_i} \log f(\boldsymbol{x}|\boldsymbol{\phi})) f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \\
&= \int \frac{\partial}{\partial \phi_i} \exp(\log f(\boldsymbol{x}|\boldsymbol{\phi})) d\boldsymbol{x} \\
&= \int \frac{\partial}{\partial \phi_i} f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \\
&= \frac{\partial}{\partial \phi_i} \int f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \\
&= \frac{\partial}{\partial \phi_i} 1 \\
&= 0,
\end{aligned}$$

where the permutation of the integral operator with the partial derivative operator is allowed by Leibniz regularity conditions.

7.3 Proof of the formula for the variance of the sufficient statistics for an exponential family

In Appendix 7.2, it has been shown that $\mathbb{E}_{\boldsymbol{\phi}} t_i(\boldsymbol{X}) = \frac{\partial}{\partial \phi_i} a(\boldsymbol{\phi})$ for any $i = 1, 2, ..., d$, implying that :

$$\frac{\partial}{\partial \phi_j} \frac{\partial}{\partial \phi_i} a(\boldsymbol{\phi}) = \frac{\partial}{\partial \phi_j} \mathbb{E}_{\boldsymbol{\phi}} t_i(\boldsymbol{X}). \tag{7.1}$$

Therefore, it is sufficient to prove that $\frac{\partial}{\partial \phi_j} \mathbb{E}_{\boldsymbol{\phi}} t_i(\boldsymbol{X}) = Cov(t_i(\boldsymbol{X}), t_j(\boldsymbol{X}))$. To see this :

$$
\begin{aligned}
\frac{\partial}{\partial \phi_j}\mathbb{E}_{\boldsymbol{\phi}} t_i(\boldsymbol{X}) &= \frac{\partial}{\partial \phi_j}\int f(\boldsymbol{x}|\boldsymbol{\phi}) t_i(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int \frac{\partial}{\partial \phi_j}(f(\boldsymbol{x}|\boldsymbol{\phi}) t_i(\boldsymbol{x})) d\boldsymbol{x} \\
&= \int t_i(\boldsymbol{x}) \frac{\partial}{\partial \phi_j} f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \\
&= \int t_i(\boldsymbol{x}) \frac{\partial}{\partial \phi_j} \exp(\log f(\boldsymbol{x}|\boldsymbol{\phi})) d\boldsymbol{x} \\
&= \int t_i(\boldsymbol{x}) \exp(\log f(\boldsymbol{x}|\boldsymbol{\phi})) \frac{\partial}{\partial \phi_j} \log f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \\
&= \int t_i(\boldsymbol{x}) \exp(\log f(\boldsymbol{x}|\boldsymbol{\phi}))(t_j(\boldsymbol{x}) - \frac{\partial}{\partial \phi_j} a(\phi)) d\boldsymbol{x} \\
&= \int t_i(\boldsymbol{x}) t_j(\boldsymbol{x}) \exp(\log f(\boldsymbol{x}|\boldsymbol{\phi})) d\boldsymbol{x} \;\; - \int t_i(\boldsymbol{x}) \exp(\log f(\boldsymbol{x}|\boldsymbol{\phi})) \frac{\partial}{\partial \phi_j} a(\phi)) d\boldsymbol{x} \\
&= \int t_i(\boldsymbol{x}) t_j(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\phi}) d\boldsymbol{x} \;\; - \frac{\partial}{\partial \phi_j} a(\phi)) \int t_i(\boldsymbol{x}) \exp(\log f(\boldsymbol{x}|\boldsymbol{\phi})) d\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{\phi}}(t_i(\boldsymbol{X}) t_j(\boldsymbol{X})) - (\mathbb{E}_{\boldsymbol{\phi}} t_j(\boldsymbol{X}))(\mathbb{E}_{\boldsymbol{\phi}} t_i(\boldsymbol{X})) \\
&= Cov(t_i(\boldsymbol{X}), t_j(\boldsymbol{X})),
\end{aligned}
$$

where the permutation of the integral operator with the partial derivative operator is allowed by Leibniz regularity conditions.

7.4 Proof of the properties of the up-to-label-swapping equivalence

Up-to-label-swapping equivalence, defined by:

$$(\boldsymbol{\theta}', \boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi}) \iff (\boldsymbol{\theta}', \boldsymbol{\pi}') = (\boldsymbol{\theta}_\sigma, \boldsymbol{\pi}_\sigma) \text{ for some permutation } \sigma \in Sym(K)$$

is a reflexive, symmetric and transitive relation, and so it is indeed an equivalence relation.

Proof. First, to prove reflexivity:

$$\text{Reflexivity} : (\boldsymbol{\theta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi})$$

just consider the identity permutation, denoted $Id \in Sym(K)$, defined by $Id(k) = k \quad \forall k \in [K]$, so that $(\boldsymbol{\theta}, \boldsymbol{\pi}) = (\boldsymbol{\theta}_{Id}, \boldsymbol{\pi}_{Id})$. Therefore $(\boldsymbol{\theta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\theta}_{Id}, \boldsymbol{\pi}_{Id}) = (\boldsymbol{\theta}, \boldsymbol{\pi})$

Now, to prove simmetry:

$$\text{Simmetry} : (\boldsymbol{\theta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\theta}', \boldsymbol{\pi}') \iff (\boldsymbol{\theta}', \boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi})$$

let prove the left-to-right implication $(\boldsymbol{\theta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\theta}', \boldsymbol{\pi}') \implies (\boldsymbol{\theta}', \boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi})$ (the proof of the right-to-left implication is basically the same). Suppose that $(\boldsymbol{\theta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\theta}', \boldsymbol{\pi}')$. Then, by definition $(\boldsymbol{\theta}, \boldsymbol{\pi}) = (\boldsymbol{\theta'}_\sigma, \boldsymbol{\pi'}_\sigma)$ for some permutation $\sigma \in Sym(K)$, and so $(\boldsymbol{\theta}_{\sigma^{-1}}, \boldsymbol{\pi}_{\sigma^{-1}}) =$

$(\boldsymbol{\theta}', \boldsymbol{\pi}')$, where σ^{-1} is the permutation inverse of σ . By definition of $\overset{swap}{\sim}$, this means that $(\boldsymbol{\theta}', \boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi})$.

Finally, to prove transitivity:

$$\text{Transitivity} : (\boldsymbol{\theta}', \boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi}) \text{ and } (\boldsymbol{\theta}'', \boldsymbol{\pi}'') \overset{swap}{\sim} (\boldsymbol{\theta}', \boldsymbol{\pi}') \implies (\boldsymbol{\theta}'', \boldsymbol{\pi}'') \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi})$$

let suppose that $(\boldsymbol{\theta}', \boldsymbol{\pi}') \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi})$ and $(\boldsymbol{\theta}'', \boldsymbol{\pi}'') \overset{swap}{\sim} (\boldsymbol{\theta}', \boldsymbol{\pi}')$. Then $(\boldsymbol{\theta}', \boldsymbol{\pi}') = (\boldsymbol{\theta}_{\sigma_1}, \boldsymbol{\pi}_{\sigma_1})$ for some $\sigma_1 \in Sym(K)$ and $(\boldsymbol{\theta}'', \boldsymbol{\pi}'') = (\boldsymbol{\theta}'_{\sigma_2}, \boldsymbol{\pi}'_{\sigma_2})$ for some $\sigma_2 \in Sym(K)$; then, by defining $\sigma = \sigma_2 \circ \sigma_1$, we have that $(\boldsymbol{\theta}'', \boldsymbol{\pi}'') = (\boldsymbol{\theta}_{\sigma}, \boldsymbol{\pi}_{\sigma})$. By definition of $\overset{swap}{\sim}$, this means that $(\boldsymbol{\theta}'', \boldsymbol{\pi}'') \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi})$.

□

It can also be very easily proved that the set of the distinct equivalence classes yield a partition

Proof. Let us consider two distinct equivalence classes $[(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap} \neq [(\boldsymbol{\theta}'', \boldsymbol{\pi}'')]_{swap}$; it is then sufficient to show that if there exists a couple $(\boldsymbol{\theta}, \boldsymbol{\pi})$ such that $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in [(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap}$ and $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in [(\boldsymbol{\theta}'', \boldsymbol{\pi}'')]_{swap}$, then we necessarily have $[(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap} = [(\boldsymbol{\theta}'', \boldsymbol{\pi}'')]_{swap}$.
Let us then suppose that

$$(\boldsymbol{\theta}, \boldsymbol{\pi}) \in [(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap} \tag{7.2}$$

and

$$(\boldsymbol{\theta}, \boldsymbol{\pi}) \in [(\boldsymbol{\theta}'', \boldsymbol{\pi}'')]_{swap}. \tag{7.3}$$

Let us consider any parametric couple $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\pi}}) \in [(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap}$. We have that $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\pi}}) \overset{swap}{\sim} (\boldsymbol{\theta}, \boldsymbol{\pi})$ and $(\boldsymbol{\theta}, \boldsymbol{\pi}) \overset{swap}{\sim} (\boldsymbol{\theta}'', \boldsymbol{\pi}'')$, and so by transitivity : $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\pi}}) \overset{swap}{\sim} (\boldsymbol{\theta}'', \boldsymbol{\pi}'')$; therefore $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\pi}}) \in [(\boldsymbol{\theta}'', \boldsymbol{\pi}'')]_{swap}$. This means that $[(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap} \subseteq [(\boldsymbol{\theta}'', \boldsymbol{\pi}'')]_{swap}$. Similarly, it can be shown that $[(\boldsymbol{\theta}'', \boldsymbol{\pi}'')]_{swap} \subseteq [(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap}$. So $[(\boldsymbol{\theta}', \boldsymbol{\pi}')]_{swap} = [(\boldsymbol{\theta}'', \boldsymbol{\pi}'')]_{swap}$.

□

7.5 Identifiability for 3-way contingency tables

Let us consider an r-class mixture model for 3 discrete variables, ie. a model in which 3 discrete variables X_1 (taking values in the set {1,2.., κ_1}), X_2 (taking values in the set {1,2.., κ_2}) and X_3 (taking values in the set {1,2.., κ_3}) are mutually independent conditional on a variable Z, taking values in $1, 2..., r$. So that :

$$\mathbb{P}(X_1 = u, X_2 = v, X_3 = w | Z = i) = \mathbb{P}(X_1 = u | Z = i)\mathbb{P}(X_2 = v | Z = i)\mathbb{P}(X_3 = w | Z = i)$$
$$\forall u \in \{1, 2.., \kappa_1\}, v \in \{1, 2.., \kappa_2\}, w \in \{1, 2.., \kappa_3\}, i \in \{1, 2..., r\}.$$

Z is supposed to be distributed according to the vector of parameters $\boldsymbol{\pi} = (\pi_1, ..., \pi_r)$. If we denote $\mathbb{P}_i(\cdot) = \mathbb{P}(\cdot | Z = i)$ we can write

$$\mathbb{P}(\cdot) = \sum_{i=1,..,r} \pi_i \mathbb{P}_i(\cdot)$$

which is a common form of describing a mixture model. The probability distribution of X_j conditional on $Z = i$ is specified by a vector $\boldsymbol{p}_{ij} \in [0, 1]^{\kappa_j}$. Entries of such a vector are denoted

by $\boldsymbol{p}_{ij}(l)$, with $\boldsymbol{p}_{ij}(l) = \mathbb{P}(X_j = l|Z = i)$. The parameters of the model are then $\boldsymbol{\pi}$, which is a r-dimentional vector, and $\mathbb{P}_1$, $\mathbb{P}_2$,..., $\mathbb{P}_r$, which are 3-dimensional $\kappa_1 \times \kappa_2 \times \kappa_3$ arrays :

$$\mathbb{P}_i = \boldsymbol{p}_{i1} \otimes \boldsymbol{p}_{i2} \otimes \boldsymbol{p}_{i3}$$

whose (l_1, l_2, l_3)- entry is $\boldsymbol{p}_{i1}(l_1)\boldsymbol{p}_{i2}(l_2)\boldsymbol{p}_{i3}(l_3)$.

For any matrix M, let us define the Kruskal rank (denoted $rank_K$) of a matrix as the maximum number n such that every subet of n rows of such matrix are linearly independent. Note that in general $rank(M) \geq rank_K(M)$. However, in the particular case where a matrix M of size $p \times q$ has rank p, it also has Kruskal rank p. For every $j = 1, 2, 3$, let M_j, be the $r \times \kappa_j$ matrix whose ith row is $\boldsymbol{p}_{ij} = \mathbb{P}(X_j = \cdot|Z = i)$. Then the following result holds:

Theorem *(Allman et al., 2011): Consider the model described above, parameterized by* $\boldsymbol{\pi}$, $\mathbb{P}_1$, $\mathbb{P}_2$,..., $\mathbb{P}_r$. *Suppose all entries of* $\boldsymbol{\pi}$ *are positive. For each* $j = 1, 2, 3$, *let* M_j *denote the* $r \times \kappa_j$ *matrix whose rows are* $\boldsymbol{p}_{ij}, i = 1, ...r$, *and let* I_j *denote its Kruskal rank. Then if*:

$$I_1 + I_2 + I_3 \geq 2r + 2$$

the parameters of the model are identifiable, up to label swapping.

The theorem is a direct application of an algebraic result from Kruskal (1976) (see Allman et al. (2011) or Kruskal (1976) for the exact statement of such an algebraic result).
The following corollary yields directly from the last theorem. Indeed, the condition on the sum of Kruskal ranks *"can be expressed through polynomial inequalities in the parameters, and thus holds generically"*. (Allman et al., 2011)

Corollary*: (Allman et al., 2011) The parameters of the model described above are generically identifiable, up to label swapping, provided*

$$\min(r, \kappa_1) + \min(r, \kappa_2) + \min(r, \kappa_3) \geq 2r + 2.$$

The assertion remains valid if, in addition, the class proportions $\{\pi_i\}_{1=1...r}$ *are held fixed and positive in the model.*

7.6 Proof of the basic properties of the Kullback-Leibler divergence

Let q and q' be 2 discrete probability distributions on a finite set Ω, $\sum_{x\in\Omega} q(x) = \sum_{x\in\Omega} q'(x) = 1$. The Kullback-Leibler divergence between q and q' is defined as:

$$KL(q(\cdot)||q'(\cdot)) := \sum_{\substack{x\in\Omega \\ q(x)>0}} q(x) \log \frac{q(x)}{q'(x)}$$

where the sum is well defined by using the conventions that any strictly positive real number divided by 0 equals $+\infty$, and that $\log(+\infty) = +\infty$. Then, the following properties hold:

$$\begin{aligned} &KL(q(\cdot)||q'(\cdot)) \geq 0 \\ &KL(q(\cdot)||q'(\cdot)) = 0 \iff q' = q \end{aligned}$$

Proof. If $x \in \Omega$ exists such that $q(x) > 0$ and $q'(x) = 0$, then:

$$KL(q(\cdot)||q'(\cdot)) := \sum_{\substack{x\in\Omega \\ q(x)>0}} q(x) \log \frac{q(x)}{q'(x)} = +\infty$$

If no $x \in \Omega$ exists such that $q(x) > 0$ and $q'(x) = 0$, (so every term in the sum is finite) but some $x \in \Omega$ exists such that $q'(x) > 0$ and $q(x) = 0$, then there exists a probability distribution q'' on Ω such that $q''(x) > q'(x)$ for all $x \in \Omega$ verifying $q(x) > 0$; and $\sum_{\substack{x\in\Omega \\ q(x)>0}} q''(x) = 1$; so we have that:

$$\begin{aligned} KL(q(\cdot)||q'(\cdot)) &:= \sum_{\substack{x\in\Omega \\ q(x)>0}} q(x) \log \frac{q(x)}{q'(x)} \\ &> \sum_{\substack{x\in\Omega \\ q(x)>0}} q(x) \log \frac{q(x)}{q''(x)} \\ &= -\sum_{\substack{x\in\Omega \\ q(x)>0}} q(x) \log \frac{q''(x)}{q(x)} \\ &\overset{(*)}{\geq} -\log \sum_{\substack{x\in\Omega \\ q(x)>0}} q(x) \frac{q''(x)}{q(x)} \\ &= -\log \sum_{\substack{x\in\Omega \\ q(x)>0}} q''(x) \\ &= -\log \sum_{\substack{x\in\Omega \\ q''(x)>0}} q''(x) \\ &= -\log 1 \\ &= 0 \end{aligned}$$

where $(*)$ holds from Jensen's inequality.

In the third case, $q'(x) > 0$ if and only if $q(x) > 0$, so every term in the sum is finite, and:

$$\begin{aligned}
KL(q(\cdot)||q'(\cdot)) &:= \sum_{\substack{x\in\Omega \\ q(x)>0}} q(x)\log\frac{q(x)}{q'(x)} \\
&= \sum_{\substack{x\in\Omega \\ q(x),q'(x)>0}} q(x)\log\frac{q(x)}{q'(x)} \\
&= -\sum_{\substack{x\in\Omega \\ q(x),q'(x)>0}} q(x)\log\frac{q'(x)}{q(x)} \\
&\overset{(*)}{\geq} -\log\sum_{\substack{x\in\Omega \\ q(x),q'(x)>0}} q(x)\frac{q'(x)}{q(x)} \\
&= -\log\sum_{\substack{x\in\Omega \\ q(x),q'(x)>0}} q'(x) \\
&= -\log\sum_{\substack{x\in\Omega \\ q'(x)>0}} q'(x) \\
&= -\log 1 \\
&= 0
\end{aligned}$$

where $(*)$ holds from Jensen's inequality. So, the Kullback-Leibler divergence is always non-negative. Furthermore, $(*)$ is an equality if and only if $\frac{q'}{q} = c$ in the sum, for some constant c, due to strict concavity of the logarithm. This implies that $q' = q$ since q' and q are probability distributions; so the Kullback-Leibler divergence is null if and only if $q' = q$. □

7.7 Proof of the formula for $\mathcal{J}$ in term of the Shannon entropy

$$
\begin{aligned}
\mathcal{J}(q,\boldsymbol{\phi}) &= \ell(\boldsymbol{\phi}) - KL(q||w(\cdot|\boldsymbol{y},\boldsymbol{\phi})) \\
&= \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi}) - \sum_{\substack{\boldsymbol{z}\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} [q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{w(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\phi}))}] \\
&= \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi}) + \sum_{\substack{z\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} [q(\boldsymbol{z}) \log w(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\phi})] - \sum_{\substack{\boldsymbol{z}\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} [q(\boldsymbol{z}) \log q(\boldsymbol{z})] \\
&= \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi}) + \sum_{\substack{z\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} [q(\boldsymbol{z}) \log \frac{f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\phi})}{f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi})}] + \mathcal{H}(q) \\
&= \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi}) - \sum_{\substack{z\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} [q(\boldsymbol{z}) \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi})] + \sum_{\substack{z\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} [q(\boldsymbol{z}) \log f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\phi})] + \mathcal{H}(q) \\
&= \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi}) - \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi})[\sum_{\substack{\boldsymbol{z}\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} q(\boldsymbol{z})] + \sum_{\substack{\boldsymbol{z}\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} [q(\boldsymbol{z}) \log f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\phi})] + \mathcal{H}(q) \\
&= \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi}) - \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\phi}) + \sum_{\substack{\boldsymbol{z}\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} [q(\boldsymbol{z}) \log f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\phi})] + \mathcal{H}(q) \\
&= \mathcal{H}(q) + \sum_{\substack{\boldsymbol{z}\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} q(\boldsymbol{z}) \log f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\phi}).
\end{aligned}
$$

7.8 Proof of the decomposition of the Shannon entropy, for completely factorisable distributions

In this section, it will be proved that the entropy of a totally factorized distribution q is decomposable in the sum of the entropies of its marginals $q_i(\cdot)$, ie:

$$\mathcal{H}(q) = \sum_{i\in[n]} \mathcal{H}(q_i)$$

where the entropies of $q(\cdot)$ and $q_i(\cdot)$ are defined by:

$$
\begin{aligned}
\mathcal{H}(q) &:= \sum_{(z_1...,z_n)\in[K]^n} q(z_1,...,z_n) \log q(z_1...,z_n) \\
\mathcal{H}(q_i) &:= \sum_{z_i\in[K]} q_i(z_i) \log q_i(z_i)
\end{aligned}
$$

We remind the reader that a totally factorized distribution satisfies:

$$q(z_1,...,z_n) = \prod_{i\in[n]} q_i(z_i).$$

We also remind the reader that marginals $q_i(\cdot)$ are defined by:

$$q_i(z_i) := \sum_{(z_1,...z_{i-1},z_{i+1}...,z_n)\in[K]^{n-1}} q(z_1,..,z_n)$$

Proof.

$$\begin{aligned}
\mathcal{H}(q) &:= \sum_{(z_1...,z_n)\in[K]^n} q(z_1...z_n)\log q(z_1,...,z_n) \\
&= \sum_{(z_1...,z_n)\in[K]^n} q_1(z_1)...q_n(z_n)\log q(z_1,...,z_n) \\
&= \sum_{(z_1...,z_n)\in[K]^n} \{q_1(z_1)...q_n(z_n)\sum_{i\in[n]}\log q_i(z_i)\} \\
&= \sum_{(z_1...,z_n)\in[K]^n}\sum_{i\in[n]} q_1(z_1)...q_n(z_n)\log q_i(z_i) \\
&= \sum_{i\in[n]}\sum_{(z_1...,z_n)\in[K]^n} q_1(z_1)...q_n(z_n)\log q_i(z_i) \\
&= \sum_{i\in[n]}\sum_{(z_1...,z_n)\in[K]^n} q(z_1,...,z_n)\log q_i(z_i) \\
&= \sum_{i\in[n]}\sum_{z_i\in[K]}\sum_{(z_1,...z_{i-1},z_{i+1}...,z_n)\in[K]^{n-1}} q(z_1,...,z_n)\log q_i(z_i) \\
&= \sum_{i\in[n]}\sum_{z_i\in[K]}\log q_i(z_i)\sum_{(z_1,...z_{i-1},z_{i+1}...,z_n)\in[K]^{n-1}} q(z_1,...,z_n) \\
&= \sum_{i\in[n]}\sum_{z_i\in[K]}\log q_i(z_i)q_i(z_i) \\
&= \sum_{i\in[n]}\mathcal{H}(q_i).
\end{aligned}$$

□

7.9 Proof of the formula by Mariadassou

It can be proven (see Appendix 7.7) that the function $\mathcal{J}(q,\boldsymbol{\phi})$ can be rewritten, for any distribution $q(\cdot)\in\mathcal{D}_{\mathcal{Z}}$ in terms of its Shannon entropy $\mathcal{H}(q)$, as:

$$\mathcal{J}(q,\boldsymbol{\phi}) = \mathcal{H}(q) + \sum_{\boldsymbol{z}\in\mathcal{Z}} q(\boldsymbol{z})\log f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\phi}) \tag{7.4}$$

where the Shannon entropy $\mathcal{H}(q)$ is defined as:

$$\mathcal{H}(q) := -\sum_{\substack{\boldsymbol{z}\in\mathcal{Z}\\ q(\boldsymbol{z})>0}} q(\boldsymbol{z})\log q(\boldsymbol{z}). \tag{7.5}$$

For a completely factorized distribution $q(\cdot)\in\mathcal{D}_{\mathcal{Z}}^{fact}$, the entropy of $q(\cdot)$ is given by the sum of the entropies of the marginals $\{q_i(\cdot)\}_{i=1...n}$ (see Appendix 7.8), so that:

$$\mathcal{H}(q) = \sum_{i\in[n]} \mathcal{H}(q_i) = -\sum_{i\in[n]} \sum_{k\in[K]} \tau_{ik} \log \tau_{ik} \quad \forall q \in \mathcal{D}_{\mathcal{Z}}^{fact}.$$

The log-mass function of a directed stochastic block model is given, under parameter $\boldsymbol{\phi} = (\boldsymbol{\pi}, \boldsymbol{\theta})$, by (see equation 1.23):

$$\log f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\phi}) = \sum_{i\in[n]} \sum_{k\in[K]} z_{ik} \log \pi_k + \sum_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \sum_{(k,l)\in[K]^2} z_{ik} z_{jl} \log g(y_{ij}|\theta_{kl}).$$

Therefore,

$$\begin{aligned}\sum_{\boldsymbol{z}\in\mathcal{Z}} q(\boldsymbol{z}) \log f(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\phi}) &= \sum_{i\in[n]} \sum_{k\in[K]} \mathbb{E}_q(\boldsymbol{Z}_{ik}) \log \pi_k + \sum_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \sum_{(k,l)\in[K]^2} \mathbb{E}_q(\boldsymbol{Z}_{ik}\boldsymbol{Z}_{jl}) \log g(y_{ij}|\theta_{kl}) \\ &= \sum_{i\in[n]} \sum_{k\in[K]} \tau_{ik} \log \pi_k + \sum_{\substack{(i,j)\in[n]^2 \\ i\neq j}} \sum_{(k,l)\in[K]^2} \tau_{ik}\tau_{jl} \log g(y_{ij}|\theta_{kl})\end{aligned}$$

where $\mathbb{E}_q(\boldsymbol{Z}_{ik}\boldsymbol{Z}_{jl}) = \mathbb{E}_q(\boldsymbol{Z}_{ik})\mathbb{E}_q(\boldsymbol{Z}_{jl}) = \tau_{ik}\tau_{jl}$ due to independence of the nodes classes. As such, we obtain that

$$\mathcal{J}(q, \boldsymbol{\theta}, \boldsymbol{\pi}) = \mathcal{H}(q) + \sum_{\substack{\boldsymbol{z}\in\mathcal{Z} \\ q(\boldsymbol{z})>0}} q(\boldsymbol{z}) \log f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\phi}).$$

7.10 Proof that the log-likelihood function is unimodular for the complete binary model

We define $\mathbf{H}(\xi_0) \in \mathbb{R}^{d^2}$ as the random matrix whose elements $\mathbf{H}_{kl}(\xi_0)$ are defined by :

$$\mathbf{H}_{kl}(\phi_0) := (\frac{\partial}{\partial\phi_k}\frac{\partial}{\partial\phi_l} \log f(\mathbf{X}|\phi))|_{\phi_0} \quad \forall (k,l) \in [d]^2 \quad \forall \phi_0 \in \Phi.$$

In order to distinguish elements of $\mathbf{H}$ corresponding to different types of parameters, $\mathbf{H}_{kl}(\phi_0)$ will also be noted as $\mathbf{H}_{\phi_k\phi_l}(\phi_0)$.

For the binary stochastic block model, if we use the parameters $(\boldsymbol{\omega}, \boldsymbol{\nu}) \in \mathcal{T}$, the log-mass function for the directed model is given by adding the right hand side terms of equation (1.17) and of equation (1.18):

$$\log f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\omega}, \boldsymbol{\nu}) = (\sum_{k=1}^{K-1} \omega_k n_k) - n \log(1 + \sum_{k=1}^{K-1} e^{\omega_k}) + \sum_{k=1}^{K}\sum_{l=1}^{K} \nu_{kl} o_{kl} - n_{kl} \log(1 + e^{\nu_{kl}}).$$

Therefore, the first derivatives of $f(\cdot)$ with respect to the parameters are given by:

$$\frac{\partial f}{\partial \omega_k} = n_k - n\frac{e^{\omega_k}}{1 + \sum_{l=1}^{K-1} e^{\omega_l}} = n_k - n\pi_k \quad \forall k \in [K-1]$$

$$\frac{\partial f}{\partial \nu_{kl}} = o_{kl} - n_{kl}\frac{e^{\nu_{kl}}}{1 + e^{\nu_{kl}}} = o_{kl} - n_{kl}\frac{\eta_{kl}/(1-\eta_{kl})}{1 + \eta_{kl}/(1-\eta_{kl})} = o_{kl} - n_{kl}\eta_{kl} \quad \forall (k,l) \in [K]^2.$$

The elements of the matrix $\mathbf{H}$ are then given by the second derivatives:

$$\begin{aligned}
\mathbf{H}_{\nu_{kl}\nu_{kl}} &= \frac{\partial}{\partial \nu_{kl}} \frac{\partial f}{\partial \nu_{kl}} \\
&= \frac{\partial}{\partial \nu_{kl}} (o_{kl} - n_{kl}\eta_{kl}) \\
&= -n_{kl} \frac{\partial}{\partial \nu_{kl}} \eta_{kl} \\
&= -n_{kl} \frac{\partial}{\partial \nu_{kl}} \frac{e^{\nu_{kl}}}{1+e^{\nu_{kl}}} \\
&= -n_{kl} \frac{e^{\nu_{kl}}(1+e^{\nu_{kl}}) - (e^{\nu_{kl}})^2}{(1+e^{\nu_{kl}})^2} = -n_{kl} \frac{e^{\nu_{kl}}}{1+e^{\nu_{kl}}} \frac{1}{1+e^{\nu_{kl}}} \\
&= -n_{kl} \frac{e^{\nu_{kl}}}{1+e^{\nu_{kl}}} (1 - \frac{e^{\nu_{kl}}}{1+e^{\nu_{kl}}}) = -n_{kl}\eta_{kl}(1-\eta_{kl}) \quad \forall (k,l) \in [K]^2,
\end{aligned}$$

while

$$\begin{aligned}
\mathbf{H}_{\omega_k\omega_k} &= \frac{\partial}{\partial \omega_k} \frac{\partial f}{\partial \omega_k} \\
&= \frac{\partial}{\partial \omega_k} (n_k - n\pi_k) \\
&= -n \frac{\partial}{\partial \omega_k} \pi_k \\
&= -n \frac{\partial}{\partial \omega_k} \frac{e^{\omega_k}}{1+\sum_{j=1}^{K-1} e^{\omega_j}} \\
&= -n \frac{e^{\omega_k}(1+\sum_{j=1}^{K-1} e^{\omega_j}) - (e^{\omega_k})^2}{(1+\sum_{j=1}^{K-1} e^{\omega_j})^2} \\
&= -n \frac{e^{\omega_k}}{1+\sum_{j=1}^{K-1} e^{\omega_j}} \frac{(1+\sum_{j=1}^{K-1} e^{\omega_j}) - e^{\omega_k}}{1+\sum_{j=1}^{K-1} e^{\omega_j}} \\
&= -n\pi_k(1-\pi_k) \qquad\qquad \forall k \in [K-1],
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{H}_{\omega_l\omega_k} &= \frac{\partial}{\partial \omega_l} \frac{\partial f}{\partial \omega_k} \\
&= \frac{\partial}{\partial \omega_l} (n_k - n\pi_k) \\
&= -n \frac{\partial}{\partial \omega_l} \pi_k \\
&= -n \frac{\partial}{\partial \omega_l} \frac{e^{\omega_k}}{1+\sum_{j=1}^{K-1} e^{\omega_j}} \\
&= -n \frac{-e^{\omega_k} e^{\omega_l}}{(1+\sum_{j=1}^{K-1} e^{\omega_j})^2} = n\pi_k\pi_l \qquad \forall (k,l) \in [K-1]^2 \quad k \neq l
\end{aligned}$$

all other elements being null:

$$\mathbf{H}_{\nu_{kl}\omega_m} = \frac{\partial^2 f}{\partial \nu_{kl}\partial \omega_m} = 0 \quad \forall (k,l) \in [K]^2 \quad m \in [K-1] \tag{7.6}$$

$$\mathbf{H}_{\nu_{kl}\nu_{k'l'}} = \frac{\partial^2 f}{\partial \nu_{kl}\partial \nu_{k'l'}} = 0 \quad \forall (k,l) \in [K]^2, (k',l') \in [K]^2 \quad (k,l) \neq (k',l'). \tag{7.7}$$

To resume, for every $(\boldsymbol{\omega}', \boldsymbol{\nu}') \in \mathcal{T}$, the matrix $\mathbf{H}$ is given, for the directed binary model, by:

$$\begin{aligned}
&\mathbf{H}_{\nu_{kl}\nu_{kl}}(\boldsymbol{\omega}', \boldsymbol{\nu}') = -\eta'_{kl}(1-\eta'_{kl})n_{kl} \quad \forall (k,l) \in [K]^2 \\
&\mathbf{H}_{\omega_k\omega_k}(\boldsymbol{\omega}', \boldsymbol{\nu}') = -n\pi'_k(1-\pi'_k) \quad \forall k \in [K-1] \\
&\mathbf{H}_{\omega_k\omega_l}(\boldsymbol{\omega}', \boldsymbol{\nu}') = n\pi'_k\pi'_l \quad \forall (k,l) \in [K-1]^2 \quad k \neq l
\end{aligned}$$

where $(\boldsymbol{\pi}', \boldsymbol{\eta}')$ denotes the parameters corresponding to $(\boldsymbol{\omega}', \boldsymbol{\nu}')$ in the familiar parametrization setting. All other elements of $\mathbf{H}$ are null.
For the undirected binary model, the formulas are the same except that we divide by 2 every element of $\mathbf{H}$.
By the previous formulas, $\mathbf{H}$ is a 2x2 diagonal blocks matrix with the first block on the diagonal, noted $\mathbf{H}_1 \in \mathbb{R}^{K^2}$ corresponding to indexes ν_{kl}, and the second block on the diagonal, noted $\mathbf{H}_2 \in \mathbb{R}^{K-1}$ corresponding to indexes ω_k, and null blocks outside of the diagonal.

$$\mathbf{H}(\boldsymbol{\omega}', \boldsymbol{\nu}') = \begin{bmatrix} \mathbf{H}_1(\boldsymbol{\omega}', \boldsymbol{\nu}') & 0 \\ 0 & \mathbf{H}_2(\boldsymbol{\omega}', \boldsymbol{\nu}') \end{bmatrix}.$$

It will now be proved that $-\mathbf{H}_1$ and $-\mathbf{H}_2$ are both positive definite. These results will imply that $-\mathbf{H}$ is positive definite too (this is due to the fact that the characteristic polynomial of a 2-blocks diagonal matrix is the product of the 2 characteristic polynomials); implying finally that $\mathbf{H}$ is negative definite. The matrix $-\mathbf{H}_1(\boldsymbol{\omega}', \boldsymbol{\nu}')$ is a diagonal matrix with strictly non-negative determinant, indeed, for the undirected model, the absolute value of the determinant is:

$$\begin{aligned}
\det -\mathbf{H}_1(\boldsymbol{\omega}', \boldsymbol{\nu}') &= \prod_{\substack{k,l \in [K]^2 \\ k \neq l}} -\eta'_{kl}(1-\eta'_{kl})n_{kl} \\
&= (-1)^{K(K-1)} \prod_{\substack{k,l \in [K]^2 \\ k \neq l}} \eta'_{kl}(1-\eta'_{kl})n_{kl} \\
&= \prod_{\substack{k,l \in [K]^2 \\ k \neq l}} \eta'_{kl}(1-\eta'_{kl})n_{kl} \\
&> 0
\end{aligned}$$

The last inequality holds provided that $n_{kl} > 0$ and $\eta'_{kl} \in (0,1) \quad \forall k,l \in [K]^2$. So -$\mathbf{H}_1$ is positive definite in the interior of the parameter space, provided that $n_{kl} > 0$. For the directed model, the development is the same, except that the product runs on $k < l$.
The second block is, for the directed model, of the form:

$$-\mathbf{H}_2(\boldsymbol{\omega}', \boldsymbol{\nu}') = \begin{bmatrix} n\pi'_1(1-\pi'_1) & -n\pi'_1\pi'_2 & -n\pi'_1\pi'_3 & . & . & . & -n\pi'_1\pi'_{K-1} \\ -n\pi'_2\pi'_1 & n\pi'_2(1-\pi'_2) & -n\pi'_2\pi'_3 & . & . & . & -n\pi'_2\pi'_{K-1} \\ -n\pi'_3\pi'_1 & -n\pi'_3\pi'_2 & n\pi'_3(1-\pi'_3) & . & . & . & -n\pi'_3\pi'_{K-1} \\ . & . & . & & & & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ -n\pi'_{K-1}\pi'_1 & -n\pi'_{K-1}\pi'_2 & -n\pi'_{K-1}\pi'_3 & . & . & . & n\pi'_{K-1}(1-\pi'_{K-1}) \end{bmatrix}$$

This is a well known type of matrix: it is the variance-covariance matrix of the $K-1$ first components of a Multinomial distribution with K classes and n trials. Theory insures that this matrix is positive definite provided that $\pi'_k \in (0,1)$. See for example Vandebril (2004). For the undirected model, again, we simply have to divide all elements of this matrix by the factor 2.

As such, $\mathbf{H}$ is negative definite in the interior of the parameters space, provided that all n_{kl} are strictly positive.

7.11 Proof of the formula by Karrer and Newman

Firstly, we remind the reader that $z_{ik} = 1$ if and only if node i belongs to class k, ie. $z_i = k$, and $z_{ik} = 0$ otherwise. Therefore $\sum_{l\in[K]} z_{il} = 1 \quad \forall i = 1, ..., n$.

The term $\log \delta_{z_i}$ can then be rewritten as :

$$\log \delta_{z_i} = \sum_{k\in[K]} z_{ik} \log \delta_k.$$

Then, d_i can be written as :

$$\begin{aligned} d_i &= \sum_{j\in[n]} y_{ij} \\ &= \sum_{j\in[n]} 1 \cdot y_{ij} \\ &= \sum_{j\in[n]} \sum_{l\in[K]} z_{jl} y_{ij} \\ &= \sum_{l\in[K]} \sum_{j\in[n]} y_{ij} z_{jl}. \\ &= \sum_{l\in[K]} d_{il}. \end{aligned}$$

So, the following equalitiy holds:

$$\begin{aligned} \sum_{l\in[K]} o_{kl} &= \sum_{l\in[K]} \sum_{(i,j)\in[n]^2} y_{ij} z_{ik} z_{jl} \\ &= \sum_{i\in[n]} z_{ik} \sum_{l\in[K]} \sum_{j\in[n]} y_{ij} z_{jl} \\ &= \sum_{i\in[n]} z_{ik} \sum_{l\in[K]} d_{il} \\ &= \sum_{i\in[n]} d_i z_{ik}. \end{aligned}$$

Similarly, it can be proven that $\sum_{k\in[K]} o_{kl} = \sum_{i\in[n]} d_i z_{il}$.

Therefore, we can write :

$$
\begin{aligned}
2\sum_{i\in[n]} d_i \log \delta_{z_i} &= 2\sum_{i\in[n]} d_i \sum_{k\in[K]} z_{ik} \log \delta_k \\
&= 2\sum_{i\in[n]} \sum_{k\in[K]} d_i z_{ik} \log \delta_k \\
&= 2\sum_{k\in[K]} \sum_{i\in[n]} d_i z_{ik} \log \delta_k \\
&= \sum_{k\in[K]} \sum_{i\in[n]} d_i z_{ik} \log \delta_k + \sum_{l\in[K]} \sum_{i\in[n]} d_i z_{il} \log \delta_l \\
&= \sum_{k\in[K]} \log \delta_k \sum_{i\in[n]} d_i z_{ik} + \sum_{l\in[K]} \log \delta_l \sum_{i\in[n]} d_i z_{il} \\
&= \sum_{k\in[K]} \log \delta_k \sum_{l\in[K]} o_{kl} + \sum_{l\in[K]} \log \delta_l \sum_{k\in[K]} o_{kl} \\
&= \sum_{k\in[K]} \sum_{l\in[K]} \log \delta_k o_{kl} + \sum_{l\in[K]} \sum_{k\in[K]} \log \delta_l o_{kl} \\
&= \sum_{k\in[K]} \sum_{l\in[K]} \log \delta_k o_{kl} + \log \delta_l o_{kl} \\
&= \sum_{(k,l)\in[K]^2} o_{kl} \log \delta_k \delta_l .
\end{aligned}
$$

7.12 Zachary karate club complete network

The complete Karate club network is displayed in the two following tables.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1		4	5	3	3	3	3	2	2		2	3	2	3			
2	4		6	3				4						5			
3	5	6		3				4	5	1				3			
4	3	3	3					3					3	3			
5	3						2				3						
6	3						5				3						3
7	3				2	5											3
8	2	4	4	3													
9	2		5														
10			1														
11	2				3	3											
12	3																
13	1			3													
14	3	5	3	3													
15																	
16																	
17						3	3										
18	2	1															
19																	
20	2	2															
21																	
22	2	2															
23																	
24																	
25																	
26																	
27																	
28			2														
29			2														
30																	
31		2							3								
32	2																
33			2						3						3	3	
34									4	2				3	2	4	

Table 7.1: *The Karate club network - part 1*

	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	2		2		2										2		
2	1		2		2									2			
3											2	2				3	
4																	
5																	
6																	
7																	
8																	
9														3		4	3
1																	2
11																	
12																	
13																	
14																	3
15																3	2
16																3	4
17																	
18																	
19																1	2
20																	1
21																3	1
22																	
23																2	
24									5		4		2			5	4
25									2		3				2		
26							5	2							7		
27													4				2
28							4	3									4
29															2		2
30							3			4						3	2
31																3	3
32								2	7			2				4	4
33		1		3		2	5						4	3	4		5
34		2	1	1		3	4			2	4	2	2	3	4	5	

Table 7.2: *The Karate club network - part 2*

List of Figures

List of Tables

References

Abbe, E. (2018). Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, *18*(177), 1–86.

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election. In *Proceedings of the WWW-2005 Workshop on the Webblogging Ecosystem.* New York: ACM.

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, *37*(6A), 3099–3132.

Allman, E. S., Matias, C., & Rhodes, J. A. (2011). Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference*, *141*(5), 1719–1736.

Amini, A. A., Chen, A., Bickel, P. J., & Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, *41*(4), 2097–2122.

Amini, A. A., & Levina, E. (2018). On semidefinite relaxations for the block model. *The Annals of Statistics*, *46*(1), 149–179.

Barbillon, P., Donnet, S., Lazega, E., & Bar-Hen, A. (2017). Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(1), 295–314.

Bickel, P. J., & Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, *106*(50), 21068–21073.

Bickel, P. J., Choi, D., Chang, X., & Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, *41*(4), 1922–1943.

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*, *22*(7), 719–725.

Bondy, J., & Murty, U. (2008). *Graph Theory.* Springer.

Cappé, O., Moulines, E., & Rydén, T. (2005). *Inference in hidden Markov models.* Springer.

Celisse, A., Daudin, J.-J., & Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, *6*, 1847–1899.

Choi, D. S., Wolfe, P. J., & Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, *99*(2), 273–284.

Daudin, J.-J., Picard, F., & Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, *18*(2), 173–183.

Decelle, A., Krzakala, F., Moore, C., & Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, *84*(6), 066106.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Erdös, P., & Rényi, A. (1960). On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, *7*, 623–641.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Karrer, B., & Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, *83*(1), 016107.

Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, *41*(3), 281–293.

Latouche, P., Birmelé, E., & Ambroise, C. (2009). Bayesian Methods for Graph Clustering. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in Data Analysis, Data Handling and Business Intelligence* (pp. 229–239). Springer.

Latouche, P., Birmelé, E., & Ambroise, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, *5*(1), 309–336.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Springer.

Lei, J., & Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, *43*(1), 215–237.

Mariadassou, M., Robin, S., & Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, *4*(2), 715–742.

Nowicki, K., & Snijders, T. A. (1997). Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, *14*(1), 75–100.

Nowicki, K., & Snijders, T. A. (2001). Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association*, *96*(455), 1077–1087.

Rohe, K., Chatterjee, S., & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, *39*(4), 1878–1915.

Rothenberg, T. J. (1971). Identification in Parametric Models. *Econometrica*, *39*(3), 577–591.

Vandebril, R. (2004). *Semiseparable matrices and the symmetric eigenvalue problem* (Unpublished doctoral dissertation). KULeuven.

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, *11*(1), 95–103.

Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, *33*(4), 452–473.

Zanghi, H., Ambroise, C., & Miele, V. (2008). Fast online graph clustering via Erdős–Rényi mixture. *Pattern Recognition*, *41*(12), 3592–3599.

Zhang, X., Wang, X., Zhao, C., Yi, D., & Xie, Z. (2014). Degree-corrected stochastic block models and reliability in networks. *Physica A: Statistical Mechanics and its Applications*, *393*, 553–559.

Zhao, Y., Levina, E., & Zhu, J. (2012). Consistency of community detetcion in networks under degree-corrected stochastic block models. *The Annals of Statistics*, *40*(4), 2266–2292.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**