Modeling heterogeneity in random graphs: a selective review

Catherine Matias, Stéphane Robin

Modeling heterogeneity in random graphs: a selective review

Catherine Matias* & Stéphane Robin†‡

February 18, 2014

Abstract

We present a selective review on modeling heterogeneity in random graphs. We focus on state space models and more particularly on stochastic block models and their extensions that have undergone major developments in the last five years.

1 Introduction

Network analysis arises in many fields of application such as biology, sociology, ecology, industry, internet, etc. Random graphs represent a natural way to describe how individuals or entities interact. An interaction network consists in a graph where each node represents an individual and an edge exists between two nodes if the two corresponding individuals interact in some way. Interaction may refer to social relationships, molecular bindings, wired connexion or web hyperlinks, depending on the context. Such interactions can be directed or not, binary (when only the presence or absence of an edge is recorded) or weighted (when a value is associated with each observed edge).

A huge literature exists on random graphs and we refer the interested reader e.g. to the recent book by Kolaczyk [2009] for a general and statistical approach to the field. A survey of statistical networks models appeared some years ago in Goldenberg et al. [2010] and more recently in Channarond [2013], Snijders [2011]. In the present review, we focus on model-based methods for detecting heterogeneity in random graphs and clustering the nodes into homogeneous groups with respect to (w.r.t.) their connectivities. This apparently specific focus still covers a quite large literature. The field has undergone so many developments in the past few years that there already exist other interesting reviews of the field and that the present one is supposed to be complementary to those. In particular, we mention the complementary review by Daudin [2011] focusing on binary graphs and the one by Leger et al. [2013] that reviews methods and algorithms for detecting homogeneous subsets of vertices in binary or weighted and directed or undirected graphs.

The literature on statistical approaches to random graphs is born in the social science community, where an important focus is given on properties such as *transitivity* ('the friends of your friends are likely to be your friends'), that is measured through clustering indexes, and to *community detection* that aims at detecting sets of nodes that share a large number of connections. However, a more general approach may be taken to analyze networks and clusters can be defined as a set of nodes that share the same connectivity behavior. Thus, we would like to stress that there is a fundamental difference between general nodes clustering methods

*Laboratoire de Mathématiques et Modélisation d'Évry, Université d'Évry Val d'Essonne, UMR CNRS 8071, USC INRA, Évry, France. email:catherine.matias@genopole.cnrs.fr

†AgroParisTech, UMR 518 Mathématiques et Informatique Appliquées, Paris 5ème, France.

‡INRA, UMR 518 Mathématiques et Informatique Appliquées, Paris 5ème, France. email:robin@agroparistech.fr

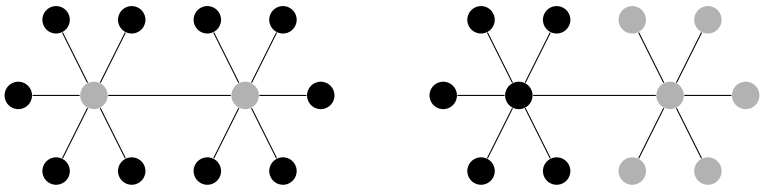

Figure 1: A toy example of community detection (on the right) and a more general clustering approach based on connectivity behavior (on the left) applied to the same graph with 2 groups. The two groups inferred are represented by colors (grey/black).

and community detection, the latter being a particular case of the former. For instance general nodes clustering methods could put all the hubs (i.e. highly connected nodes) together in one group and all peripheral nodes into another group, while such clusters could never be obtained from a community detection approach. This is illustrated in the toy example from Figure 1.

In the present review, we do not discuss general clustering methods, but only those that are model-based. In particular, there is an abundant literature on spectral clustering methods or other modularities-based methods for clustering networks and we only mention those for which some performance results have been established within heterogenous random graphs models. More precisely, a large part of this review focuses on the stochastic block model (SBM) and we discuss non model-based clustering procedures only when these have been studied in some way within SBM.

Exponential random graph models also constitutes a broad class of network models, which is very popular in the social science community (see Robins et al. [2007] for an introduction). Such models have been designed to account for expected social behaviors such as the transitivity mentioned earlier. In some way, they induce some heterogeneity, but they are not generative and the identification of some unobserved structure is not their primary goal. For the sake of simplicity and homogeneity, we decide not to discuss these models in this review.

The review is organized as follows. Section 2 is a general introduction to state space models for random graphs, that covers all model-based clustering methods. The beginning of this section (Section 2.1) introduces the properties shared by all those models. Then Section 3 focuses more precisely on the stochastic block model, which assumes that the nodes are clustered into a finite number of different latent groups, controlling the connectivities between the nodes from those groups. In Section 4, we review the various generalizations that have been proposed to SBM and in Section 5, we briefly conclude this review on discussing the next challenges w.r.t. modeling heterogeneity in random graphs.

Note that most of the following results are stated for undirected binary or weighted random graphs with no self-loops. However easy generalizations may often be obtained for directed graphs, with or without self-loops.

2 State space models for random graphs

2.1 Common properties of state space models

Let us first describe the set of observations at stake. When dealing with random graphs, we generally observe an adjacency matrix $\{Y_{ij}\}_{1\leq i,j\leq n}$ characterizing the relations between n distinct individuals or nodes. This can either be a binary matrix ($Y_{ij} \in \{0,1\}$ indicating presence or absence of each possible edge) or a vector valued table ($Y_{ij} \in \mathbb{R}^s$ being a – possibly multivariate – *weight* or any value characterizing the relation between nodes i, j). The graph may be directed or undirected (in which case $Y_{ij} = Y_{ji}$ for any $1 \leq i, j \leq n$), it may either admit self-loops or not ($Y_{ii} = 0$ for any $1 \leq i \leq n$).

State space models generally assume the existence of a latent random variable, whose value

characterizes the distribution of the observation. In the specific case of random graphs, observations are the random variables Y_{ij} that characterize the relation between nodes i, j. Note that assuming the existence of a latent variable for each relation (i, j) would not make advantage of the graph structure on the observations. Thus, one usually assumes that the latent variables rather characterize the nodes behaviors and each observation Y_{ij} will have its distribution characterized through the two different latent variables at nodes i and j. We thus assume that there exist some independent latent random variables $\{Z_i\}_{1\leq i\leq n}$ being indexed by the set of nodes. Moreover, conditional on the Z_i's, the observations $\{Y_{ij}\}_{1\leq i,j\leq n}$ are independent and the distribution of each Y_{ij} only depends on Z_i and Z_j. This general framework is considered in Bollobás et al. [2007], where the topological properties of such random graphs are studied extensively from a probabilistic point of view. Note that this model results in a set of observations $\{Y_{ij}\}_{1\leq i,j\leq n}$ that are not independent anymore. In fact, the dependency structure induced on the Y's is rather complex as will be seen below.

Now, we will distinguish two different cases occurring in the literature: the latent random variables Z_i may either take finitely many values denoted by $\{1, \ldots, Q\}$, or being continuous and belong to some latent space, e.g. $\mathbb{R}^q$ or $[0, 1]$. In both cases (finite or continuous), the network characteristics are summarized through a low dimensional latent space. The first case corresponds to the stochastic block model (SBM) and will be reviewed in extension below (see Section 3). The second case will be dealt with in Sections 2.2 and 2.3.

Before closing this section, we would like to explain some issues appearing when dealing with parameter estimation that are common to all these models. Indeed, as in any state space model, likelihood may not be computed except for small sample sizes, as it requires summing over the set of possible latent configurations that is huge. Note that some authors circumvent this computational problem by considering the latent variables as model parameters and computing a likelihood conditional on these latent variables. Then the observations are independent and the corresponding likelihood has a very simple form with nice statistical properties. As a counterpart, the maximization with respect to those latent parameters raises new issues, as it results in a discrete optimization problem with associated combinatorial complexity. Besides the resulting dependency structure on the observations is then different.

However, the classical answer to maximum likelihood computation with latent variables lies in the use of the expectation-maximization (`em`) algorithm [Dempster et al., 1977]. Though, the `e`-step of the `em` algorithm may be performed only when the distribution of the latent variables Z_i, conditional on the observations Y_{ij}, may be easily computed. This is the case for instance in classical finite mixture models (namely when observations are associated with respectively independent latent variables) as well as in models with more complex dependency structure such as hidden Markov models [HMM, see for instance Ephraim and Merhav, 2002, Cappé et al., 2005] or more general conditional random fields, where the distribution of the latent Z_i's conditional on the observed Y_{ij}'s is explicitely modeled [Lafferty et al., 2001, Sutton and McCallum, 2012]. In the case of random graphs where latent random variables are indexed by the set of nodes while observations are indexed by pairs of nodes, this is not the case anymore and the distribution of the Z_i's conditional on the Y_{ij}'s is not tractable. The reason for this complexity is explained in Figure 2. In this figure, the left panel reminds that the latent variables $\{Z_i\}$ are first drawn independently and that the observed variables $\{Y_{ij}\}$ are then also drawn independently, conditional on the $\{Z_i\}$ with distribution that only depends on their *parent* variables [see Lauritzen, 1996, for details on graphical models]. The *moralization step* shows that the parents are not independent anymore when conditioning on their common offspring. Indeed, if $p(Z_i, Z_j, Y_{ij}) = p(Z_i)p(Z_j)p(Y_{ij}|Z_i, Z_j)$, then $p(Z_i, Z_j|Y_{ij}) = p(Z_i, Z_j, Y_{ij})/p(Y_{ij})$ can not be factorized anymore ('parents get married'). The right panel gives the resulting joint conditional distribution of the $\{Z_i\}$ given the $\{Y_{ij}\}$, which is a clique. This dependency structure prevents any factorization, as opposed to models such as HMM or tree-based models where this structure

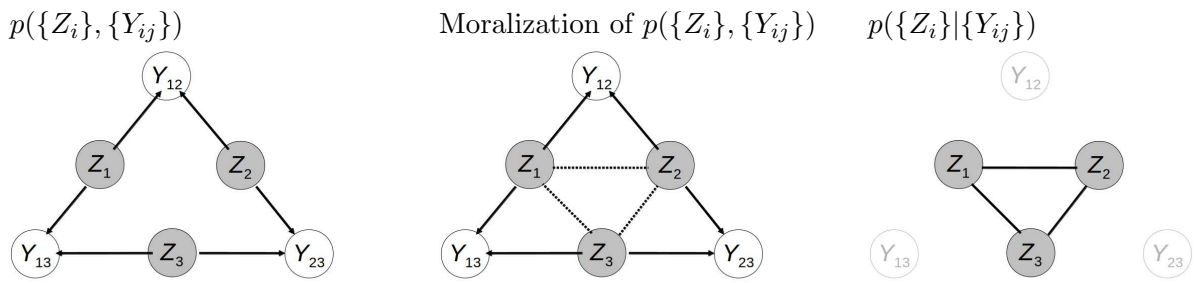

Figure 2: Graphical representation of the dependency structure in state-space models for graphs. Left: state-space model as a directed probabilistic graphical model. Center: Moralization of the graph. Right: conditional distribution of the latent variables as an undirected probabilistic graphical model. Legend: filled white: observed variables, filled gray: latent variables, light gray lines: conditioning variables.

is tree-shaped.

Thus, `em` algorithm does not apply and other strategies need to be developed for parameter inference. These may be classified into two main types: Monte Carlo Markov chains (`mcmc`) strategies [Gilks et al., 1995] and variational approaches [Jaakkola, 2000]. The former methods aim at sampling from the true conditional distribution of the latent variables conditional on observed ones (e.g. relying on a Gibbs sampler) and suffer from low computational efficiency. In fact, these methods are limited to network sizes of the order of a few hundreds of nodes. The variational approaches result in `em`-like algorithms in which an approximate conditional distribution of the latent variables conditional on observed ones is used. They are more efficient and may handle larger data sets (up to some thousands of nodes). In general, variational approaches suffer from a lack of convergence of the parameter estimates with respect to sample size [Gunawardana and Byrne, 2005]. But in the specific case of random graphs state space models, they appear to be surprisingly accurate. The reason for this will be given, at least for SBM, in Section 3.3 below. All the methods for parameter inference in SBM will be described in Section 3.2.

2.2 Latent space models (for binary graphs)

Latent space models have been developed in the context of binary graphs only. In Hoff et al. [2002], the latent space $\mathbb{R}^q$ represents a *social space* where the proximity of the actors induces a higher probability of connection in the graph. Thus, only relative positions in this latent space are relevant for the model. More precisely, the model is defined for binary random graphs and allows for covariate vectors $\mathbf{x}_{ij}$ on each relation (i, j). Two different parametrization have been proposed in Hoff et al. [2002] to deal with undirected and directed graphs, respectively. For undirected graphs, the probability of connection between nodes i, j is parametrized through a logistic regression model

$$\text{logit}(\mathbb{P}(Y_{ij} = 1|Z_i, Z_j, \mathbf{x}_{ij})) = \frac{\mathbb{P}(Y_{ij} = 1|Z_i, Z_j, \mathbf{x}_{ij})}{1 - \mathbb{P}(Y_{ij} = 1|Z_i, Z_j, \mathbf{x}_{ij})} = \alpha + \beta^\intercal \mathbf{x}_{ij} - \|Z_i - Z_j\|,$$

where $\|\cdot\|$ denotes Euclidean norm in latent space $\mathbb{R}^q$, the model parameters are α, β and $u^\intercal$ denotes the transpose of vector u. Note that the Euclidean norm could be replaced by any kind of distance. In the case of directed networks, the distance is replaced by the scalar product

$Z_i^\intercal Z_j$, normalized by the length $\|Z_i\|$ of vector Z_i. Thus, the model becomes

$$\text{logit}(\mathbb{P}(Y_{ij} = 1 | Z_i, Z_j, \mathbf{x}_{ij})) = \alpha + \beta^\intercal \mathbf{x}_{ij} + \frac{Z_i^\intercal Z_j}{\|Z_i\|}.$$

Note that in both cases, the latent variables $\{Z_i\}$ might be recovered only up to rotation, reflection and translation as these operations would induce *equivalent configurations*. Whether these restrictions are sufficient for ensuring the uniqueness of these latent vectors has not been investigated to our knowledge. Also note that in the model proposed here, the latent positions Z_i's are considered as model parameters. Thus, the total number of parameters is $nq - q(q+1)/2 + 2$ (including α and β), which can be quite large unless q is small.

Hoff et al. [2002] consider a Bayesian setting by putting prior distributions on α, β and the Z_i's and rely on `mcmc` sampling to do parameter inference. The authors first compute a likelihood that has a very simple form (since latent variables are considered as parameters, the observations are i.i.d.) and argue that this likelihood is convex w.r.t. the distances and may thus be first optimized w.r.t. these. Then, a multidimensional scaling approach enables to identify an approximating set of positions $\{Z_i\}$ in $\mathbb{R}^q$ fitting those distances. These estimates $\hat{Z}_i$ form an initialization for the second part of the procedure. Indeed, in a second step, the authors use an acceptance-rejection algorithm to sample from the posterior distribution of $(\alpha, \beta, \{Z_i\}_{1 \leq i \leq n})$ conditional on the observations. Note that with this latent space model, the nodes of the graph are not automatically clustered into groups as it is the case when the latent space is finite. For this reason, Handcock et al. [2007], proposed to further model the latent positions through a finite mixture of multivariate Gaussian distributions, with different means and spherical covariance matrices. Two procedures are proposed for parameter inference: either a two-stage maximum likelihood method, where the first stage estimates the latent positions as in Hoff et al. [2002] (relying on a simple-form likelihood), while the second one is an `em` procedure with respect to the latent clusters, conditionally on the estimated latent positions; or a Bayesian approach based on `mcmc` sampling. Besides, the number of clusters may be determined by relying on approximate conditional Bayes factors.

Latent space models have been generalized into *random dot product graphs*. Introduced in Nickel [2006] and Young and Scheinerman [2007], these models assume that each vertex is associated with a latent vector in $\mathbb{R}^q$ and the probability that two vertices are connected is then given by a function g of the dot product of their respective latent vectors. Three different versions of the model have been proposed in Nickel [2006], who shows that in at least two of those models, the resulting graphs obey a power law degree distribution, exhibit clustering, and have a low diameter. In the model further studied in Young and Scheinerman [2007], each coordinate of those latent vectors Z_i is drawn independently and identically from the distribution $q^{-1/2}\mathcal{U}([0,1])^\alpha$, namely for any $1 \leq k \leq q$, the coordinate $Z_i(k)$ equals $U_k^\alpha / \sqrt{q}$ where $U_1, \ldots, U_q$ are i.i.d. with uniform distribution $\mathcal{U}([0,1])$ on $[0,1]$ and $\alpha > 1$ is some fixed parameter. Moreover, the probability of connection between two nodes i, j is exactly the dot product of corresponding latent vectors $Z_i^\intercal Z_j$. Interestingly, the one-dimensional ($q = 1$) version of this model corresponds to a graphon model (see next section and Section 4.5) with function $g(u,v) = (uv)^\alpha$. To our knowledge, parameter inference has not been dealt with in the random dot product models (namely inferring α and in some models the parametric link function g). However, Tang et al. [2013] have proposed a method to consistently estimate latent positions (up to an orthogonal transformation), relying on the eigen-decomposition of $(AA^\intercal)^{1/2}$, where A is the adjacency matrix of the graph. We mention that the authors also provide classification results in a supervised setting where latent positions are labeled and a training set is available (namely latent positions and their labels are observed). However their convergence results concern a model of i.i.d. observations where latent variables are considered as fixed parameters.

Before closing this section, we mention that the problem of choosing the dimension q of the latent space has not been the focus of much attention. We already mentioned that the number

of parameters can become quite large with q. In practice, people seem to use $q = 2$ or 3 and heuristically compare the resulting fit (taking into account the number of parameters in each case). However the impact of the choice of q has not been investigated thoroughly. Moreover, as already mentioned, the parameters' identifiability (with fixed q) has not been investigated in any of the models described above.

2.3 Other state space models

Models with different state spaces have also been proposed. In Daudin et al. [2010], the latent variables $\{Z_i\}$ are supposed to belong to the simplex within $\mathbb{R}^Q$ (namely, $Z_i = (Z_{i1}, \ldots, Z_{iQ})$ with all $Z_{iq} > 0$ and $\sum_q Z_{iq} = 1$). As for the inference, the latent positions in the simplex are considered as fixed and maximum likelihood is used to estimate both the positions and the connection probabilities. Note that, because the Z_i's are defined in a continuous space, the optimization problem with respect to the Z_i's is manageable. Conversely, as they are considered as parameters, the Z_i's have to be accounted for in the penalized criterion to be used for the selection of the dimension Q. This model can be viewed as a continuous version of the stochastic block model (that will be extensively discussed in the next section): in the stochastic block model the Z_i's would be required to belong to the vertices of the simplex.

The graphon model (or W-graph) is a another popular model in the probability community as it can be viewed as a limit for dense graphs [Lovász and Szegedy, 2006]. This model states that nodes are each associated with hidden variables U_i, all independent and uniformly distributed on $[0, 1]$. A graphon function $g : [0, 1]^2 \mapsto [0, 1]$ is further defined and the binary edges $(Y_{ij})_{1 \leq i < j \leq n}$ are then drawn independently conditional on the U_i's, such that

$$\mathbb{P}(Y_{ij} = 1 | U_i, U_j) = g(U_i, U_j). \tag{1}$$

The connections between this model and the stochastic block model will be discussed in Section 4.5.

3 Stochastic block model (binary or weighted graphs)

In this section, we review the many results that have been recently obtained in the stochastic block model.

3.1 Notation

Let us start by recalling some notation. We consider a random graph on a set $V = \{1, \ldots, n\}$ of n nodes, defined as follows. Let $\mathbf{Z} := \{Z_1, \ldots, Z_n\}$ be i.i.d. random variables taking values in a finite set $\{1, \ldots, Q\}$ that are latent (namely unobserved), with some distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$. Alternatively, each Z_i may be viewed as a size-Q vector $Z_i = (Z_{i1}, \ldots, Z_{iQ})$ with entries in $\{0, 1\}$ that sum up to one, whose distribution is multinomial $\mathcal{M}(1, \boldsymbol{\pi})$. The observations consist in the set $\mathbf{Y} := \{Y_{ij}\}_{(i,j) \in \mathcal{I}}$ of random variables Y_{ij} belonging to some space $\mathcal{Y}$, that characterize the relations between each pair of nodes i and j. When the graph is undirected with no self-loops, the index set is $\mathcal{I} = \{(i, j); 1 \leq i < j \leq n\}$, while it is equal to $\mathcal{I} = \{(i, j); 1 \leq i \neq j \leq n\}$ for directed graphs with no self-loops. Easy generalizations are obtained when authorizing self-loops. The distribution of the set of variables $\mathbf{Y} = \{Y_{ij}\}_{(i,j) \in \mathcal{I}}$ is as follows: conditional on $\mathbf{Z} = \{Z_i\}_{1 \leq i \leq n}$, the Y_{ij}'s are independent and the distribution of each variable Y_{ij} only depends on Z_i and Z_j. We let $F(\cdot; \gamma_{Z_i Z_j})$ denote this distribution where $\boldsymbol{\gamma} = (\gamma_{q\ell})_{1 \leq q, \ell \leq Q}$ is called the connectivity parameter. Note that this matrix is symmetric when

modeling undirected graphs. In conclusion, SBM is characterized by the following

$$\begin{aligned}
&\cdot\ \mathbf{Z} = Z_1, \dots, Z_n \text{ i.i.d. latent random variables with distribution } \boldsymbol{\pi} \text{ on } \{1, \dots, Q\},\\
&\cdot\ \mathbf{Y} = \{Y_{ij}\}_{(i,j)\in\mathcal{I}} \text{ set of observations in } \mathcal{Y}^{\mathcal{I}},\\
&\cdot\ \mathbb{P}(\mathbf{Y}|\mathbf{Z}) = \otimes_{(i,j)\in\mathcal{I}} \mathbb{P}(Y_{ij}|Z_i, Z_j) \text{ (conditional independence)},\\
&\cdot\ \forall (i,j) \in \mathcal{I} \text{ and } \forall 1 \le q, \ell \le Q, \text{ we have } Y_{ij}|Z_i = q, Z_j = \ell \sim F(\cdot; \gamma_{q\ell}).
\end{aligned} \tag{2}$$

Now we distinguish binary versions of the model from weighted ones. Binary SBMs were introduced in the early eighties [Frank and Harary, 1982, Holland et al., 1983], while weighted versions of the model appeared only much later [Mariadassou et al., 2010, Jiang et al., 2009, Ambroise and Matias, 2012]. In the binary SBM, the distribution of Y_{ij} conditional on Z_i, Z_j is simply Bernoulli $\mathcal{B}(\gamma_{Z_i Z_j})$. Namely

$$\forall y \in \{0, 1\}, \quad F(y; \gamma) = \gamma^y (1 - \gamma)^{1-y}. \tag{3}$$

Generalizing the model to weighted graphs, we consider that the distribution of Y_{ij} conditional on Z_i, Z_j is any type of distribution that depends only on Z_i, Z_j. More precisely, it is useful to restrict to parametric distributions, such as Poisson, Gaussian, etc. However, considering for instance a Gaussian distribution would induce a complete graph, which is not desirable in most applications. Thus, it makes sense to consider instead a mixture from a Dirac mass at zero modeling absent edges, with any parametric distribution that models the strength or weight of present edges [Ambroise and Matias, 2012]. For identifiability reasons, this latter distribution is restricted to have a cumulative distribution function (cdf) continuous at zero. In other words, we let

$$\forall y \in \mathcal{Y}, \quad F(y; \gamma) \sim \gamma^1 G(\cdot, \gamma^2) + (1 - \gamma^1)\delta_0(\cdot), \tag{4}$$

where the connectivity parameter γ has now two coordinates $\gamma = (\gamma^1, \gamma^2)$ with $\gamma^1 \in [0, 1]$ and $G(\cdot, \gamma^2)$ is the conditional distribution on the weights (or intensity of connection), constrained to have a continuous cdf at zero. When all the γ^1_{ql} are equal to 1, the graph is complete. The particular case where $G(\cdot, \gamma^2_{ql})$ is the Dirac mass at point 1 corresponds to binary SBM. Weighted SBM may for instance consider G to be truncated Poisson, or a (multivariate) Gaussian, etc. Also note that in the non binary case, the model may be simplified by assuming γ^1_{ql} is constant (equal to some fixed γ^1), so that connectivity is constant throughout the different groups and only intensity varies.

In what follows, the whole parameter of the SBM distribution is denoted by $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\gamma})$.

Note that a particular case of SBM is obtained when considering an *affiliation* structure. Namely, the connectivity parameter $\boldsymbol{\gamma}$ takes only two different values, depending whether $q = \ell$ or not. In other words,

$$\forall 1 \le q, \ell \le Q, \quad \gamma_{q\ell} = \begin{cases} \gamma_{\text{in}} & \text{when } q = \ell, \\ \gamma_{\text{out}} & \text{when } q \neq \ell. \end{cases} \tag{5}$$

Note also that when considering a binary affiliation SBM and assuming moreover that $\gamma_{\text{in}} \gg \gamma_{\text{out}}$, the clustering structure induced by the model corresponds exactly to clustering based on the graph topology (namely searching for sets of nodes that almost form a clique). That corresponds to community detection [Fortunato, 2010]. As already mentioned, unconstrained SBM induces a node clustering that is much more general than community detection.

To conclude this section, we discuss Szemerédi's Lemma [Szemerédi, 1978] as a potential motivation for SBM. Indeed, this lemma roughly states that every large enough graph can be divided into subsets of about the same size so that the edges between different subsets behave almost randomly. To state this more precisely, we introduce some definitions. A simple graph is an undirected graph with no-self-loops, nor multi-edges. For any finite simple graph $G = (V, E)$,

let $X, Y \subset V$ be disjoint subsets of vertices; $e(X, Y)$ the number of edges from vertices of X to vertices of Y and

$$\rho(X, Y) = \frac{e(X, Y)}{|X||Y|},$$

the edge density between X and Y (here $|X|$ denotes the cardinality of X). Note that $0 \leq \rho(X, Y) \leq 1$.

Definition 1. *Let $\epsilon > 0$. A pair (A, B) of disjoint subsets of V is called ϵ-regular if for any $X \subset A$ and any $Y \subset B$ with $|X| \geq \epsilon|A|$ and $|Y| \geq \epsilon|B|$ we have*

$$|\rho(X, Y) - \rho(A, B)| \leq \epsilon.$$

In other words, when ϵ is small, an ϵ-regular pair of node subsets is such that edges are almost uniformly distributed between those subsets.

Definition 2. *A partition $\{V_0, V_1, \ldots, V_K\}$ of V is called ϵ-regular if*

i) $|V_0| \leq \epsilon|V|$

ii) $|V_1| = |V_2| = \cdots = |V_K|$

iii) All but at most ϵK^2 of the pairs (V_i, V_j) for $1 \leq i, j \leq K$ are ϵ-regular.

We are now ready to state Szemerédi's Lemma.

Lemma 1 (Szemerédi's Lemma 1978)**.** *Given any $\epsilon > 0, m \geq 1$ there exists some $M = M(m, \epsilon)$ such that every simple graph with number of vertices larger than M admits an ϵ-regular partition $\{V_0, \ldots, V_K\}$ for some $K \in [m; M]$.*

Note that the lemma is not of a probabilistic nature. Moreover, K can be arbitrary large (bounded by M).

3.2 Parameter estimation

Parameters' identifiability. We start this section by discussing identifiability of the parameters in SBM. As in any mixture model, the parameters of SBM may be recovered only up to a permutation on the groups labels. This is known as *identifiability up to label switching*. However, the issue of whether this restriction was or not sufficient to ensure the parameters' identifiability has been overlooked in the literature for a long time. In fact, the question was first solved only recently in the particular case of binary (undirected) SBM with only $Q = 2$ groups in Theorem 7 from Allman et al. [2009]. It was later fully solved in Allman et al. [2011] for (undirected) SBM, both in binary and weighted cases, including parametric and non-parametric conditional distributions on the weights. Note that Celisse et al. [2012] provide another identifiability result valid for (directed or undirected) binary SBM.

Parameter estimation versus clustering. It is important to note that at least two different approaches may be considered when dealing with graphs under SBM. The first one is to estimate the SBM parameters first, or at the same time as the nodes clusters. The second one is to cluster the nodes first (with no information on the SBM parameters) and then recover the SBM parameters through these estimated clusters. The latter is less related to SBM since generally, the clustering is done without using the model and is based on the graph structure. In Section 3.3 , we discuss these methods when their theoretical properties within SBM have been discussed. As for the rest of the current section, we focus on the first approach (namely parameter estimation in SBM).

MCMC approaches. Snijders and Nowicki [1997] have developed `mcmc` methods for Bayesian parameter estimation procedures in binary SBM. We mention that these authors first considered a maximum likelihood method, that is limited to graphs with up to 20 or 30 vertices for binary SBM with $Q = 2$ groups. Then they proposed a Gibbs sampler for Bayesian estimation of the parameters. More precisely, given current values $(\mathbf{Z}^{(t)}, \boldsymbol{\theta}^{(t)})$ of both the latent groups and the SBM parameter, the algorithm samples

- $\boldsymbol{\theta}^{(t+1)} = (\boldsymbol{\pi}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)})$ from the posterior distribution given the complete data $(\mathbf{Z}^{(t)}, \mathbf{Y})$, namely by using

$$\mathbb{P}(\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi} | \mathbf{Z}^{(t)}, \mathbf{Y}) \propto \mu_{\boldsymbol{\pi}}(\boldsymbol{\pi}) \prod_{i=1}^{n} \prod_{q=1}^{Q} \pi_q^{Z_{iq}^{(t)}},$$

$$\mathbb{P}(\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma} | \mathbf{Z}^{(t)}, \mathbf{Y}) \propto \mu_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) \prod_{(i,j)\in\mathcal{I}} \prod_{1\leq q,\ell\leq Q} F(Y_{ij}; \gamma_{q\ell})^{Z_{iq}^{(t)} Z_{j\ell}^{(t)}},$$

  where $\propto$ stands for "proportional to" and $\mu_{\boldsymbol{\pi}}, \mu_{\boldsymbol{\gamma}}$ are the prior distributions on $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ respectively;

- For $i = 1$ to n, sample $Z_i^{(t+1)}$ from its posterior distribution given $(\mathbf{Y}, Z_1^{(t+1)}, \ldots, Z_{i-1}^{(t+1)}, Z_{i+1}^{(t)}, \ldots, Z_n^{(t)}, \boldsymbol{\theta}^{(t+1)})$. To do this step, one uses the formula

$$\mathbb{P}(Z_i = q | \mathbf{Y}, \{Z_j\}_{j\neq i}, \boldsymbol{\theta}) \propto \pi_q \prod_{j;(i,j)\in\mathcal{I}} \prod_{\ell=1}^{Q} F(Y_{ij}; \gamma_{q\ell})^{Z_{j\ell}}. \tag{6}$$

These results on Bayesian estimation with a Gibbs sampler have been extended in Nowicki and Snijders [2001] to handle directed graphs, an arbitrary number of classes (restricted to $Q = 2$ earlier) and a finite number of values for each relation Y_{ij}. In practice, those Bayesian methods are restricted to small sample sizes (graphs with up to few hundred nodes). However, very recent attempts have been made to develop heuristic algorithms with performances equivalent to exact `mcmc` procedures but much lower running time [Peixoto, 2014].

Variational approximations. As already explained in Section 2.1, the `em` algorithm may not be performed exactly in SBM due to the intricate form of the conditional distribution of the groups given the data. The natural solution in such a case is to replace this conditional distribution by its best approximation within a reduced class of distributions with simpler form. This leads to what is called a variational approximation to the maximum likelihood computation. Let us explain this in more details. The data log-likelihood $\mathcal{L}_{\mathbf{Y}}(\boldsymbol{\theta})$ may decomposed as follows

$$\mathcal{L}_{\mathbf{Y}}(\boldsymbol{\theta}) := \log \mathbb{P}(\mathbf{Y}; \boldsymbol{\theta}) = \log \mathbb{P}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta}) - \log \mathbb{P}(\mathbf{Z}|\mathbf{Y}; \boldsymbol{\theta})$$

and by taking on both sides the expectation with respect to some distribution $\mathbb{Q}$ acting only on $\mathbf{Z}$, we get

$$\mathcal{L}_{\mathbf{Y}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta})) + \mathcal{H}(\mathbb{Q}) + \mathcal{KL}(\mathbb{Q} \| \mathbb{P}(\mathbf{Z}|\mathbf{Y}; \boldsymbol{\theta})), \tag{7}$$

where $\mathcal{H}(\mathbb{P})$ is the entropy of distribution $\mathbb{P}$ and $\mathcal{KL}(\mathbb{P}\|\mathbb{Q})$ is the Kullback-Leibler divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$. Starting from this relation, `em` algorithm is an iterative procedure based on the iteration of the two following steps. Starting from current parameter value $\theta^{(t)}$, we do

- `e`-step: maximise the quantity $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta}^{(t)})) + \mathcal{H}(\mathbb{Q})$ with respect to $\mathbb{Q}$. From (7), since $\mathcal{L}_{\mathbf{Y}}(\boldsymbol{\theta}^{(t)})$ does not depend on $\mathbb{Q}$, this is equivalent to minimizing $\mathcal{KL}(\mathbb{Q}\|\mathbb{P}(\mathbf{Z}|\mathbf{Y}; \boldsymbol{\theta}^{(t)}))$ with respect to $\mathbb{Q}$. The optimal solution is thus given by the conditional distribution $\mathbb{P}(\mathbf{Z}|\mathbf{Y}; \boldsymbol{\theta}^{(t)})$ for current parameter value $\boldsymbol{\theta}^{(t)}$;

- `m`-step: keeping now $\mathbb{Q}$ fixed, maximize the quantity $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta})) + \mathcal{H}(\mathbb{Q})$ with respect to $\boldsymbol{\theta}$ and update the parameter value $\boldsymbol{\theta}^{(t+1)}$ to this maximiser. As $\mathbb{Q}$ does not involve the parameter $\boldsymbol{\theta}$, this is equivalent to maximizing the conditional expectation $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta}))$ w.r.t. $\boldsymbol{\theta}$. Note that here, with our choice of $\mathbb{Q}$, this quantity is the conditional expectation $\mathbb{E}(\log \mathbb{P}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta})|\mathbf{Y}, \theta^{(t)})$ w.r.t. $\boldsymbol{\theta}$. Moreover, this will automatically increase the log-likelihood $\mathcal{L}_{\mathbf{Y}}(\boldsymbol{\theta})$.

When the true distribution $\mathbb{P}(\mathbf{Z}|\mathbf{Y})$ is intractable (e.g. when it can not be factorized in any way), the exact solution from `e`-step may not be computed. Instead, going back to (7) and using that the Kullback-Leibler divergence term is positive, we obtain the following lower bound

$$\mathcal{L}_{\mathbf{Y}}(\boldsymbol{\theta}) \geq \mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta})) + \mathcal{H}(\mathbb{Q}). \tag{8}$$

In the variational approximation, instead of computing the exact solution at `e`-step, we rather search for an optimal solution within a restricted class of distributions, e.g. within the class of factorized distributions

$$\mathbb{Q}(\mathbf{Z}) = \prod_{i=1}^{n} \mathbb{Q}(Z_i),$$

and the `m`-step is unchanged, with $\mathbb{Q}$ fixed to the previous solution. In the case of SBM, taking $\mathbb{Q}$ within the class of factorized distributions:

$$\mathbb{Q}(\mathbf{Z}) = \prod_{i=1}^{n} \mathbb{Q}(Z_i) = \prod_{i=1}^{n} \prod_{q=1}^{Q} \tau_{iq}^{Z_{iq}},$$

where $\tau_{iq} = \mathbb{Q}(Z_i = q)$ (with $\sum_q \tau_{iq} = 1$ for all i), the solution to the `e`-step at the current parameter value $\boldsymbol{\theta}$, within the above class distributions satisfies the following fixed point relation (see Proposition 5 in Daudin et al. [2008] in the binary case and Section 4.2 in Mariadassou et al. [2010] for weighted graphs):

$$\tau_{iq} \propto \pi_q \prod_{j;(i,j)\in\mathcal{I}} \prod_{\ell=1}^{Q} [f(Y_{ij}; \gamma_{q\ell})]^{\tau_{j\ell}}.$$

The resulting approximation is sometimes called a *mean field approximation* because, when considering the (conditional) distribution of Z_i, all other Z_j's are set to their respective (conditional) means τ_{iq}. A link can be made with `mcmc` techniques applied to SBM, which most often rely on a Gibbs sampler. In this Gibbs sampling step, the Z_i are iteratively sampled conditionally on the observed variables, the current parameters and the other Z_j's, that is with distribution given by (6). In the variational framework, the probability τ_{iq} is the best possible approximation of $\mathbb{P}(Z_i = q|\mathbf{Y}, \boldsymbol{\theta})$ in terms of Kullback-Leibler divergence, within the set of factorized distributions.

By doing this variational approximation, we only optimize the lower bound on the right hand-side of (8) with respect to θ and have no guarantee of approximating the maximum likelihood estimator. In fact, as already mentioned in Section 2.1, these variational approximations are known to be non convergent in regular cases [Gunawardana and Byrne, 2005]. More precisely, in general the `e`-step approximation prevents convergence to a local maxima of the likelihood. However, it appears that for SBM, both empirical and theoretical results ensure that these procedures are pretty accurate (see paragraph "Asymptotic properties" below as well as Section 3.3 for more details).

In Daudin et al. [2008], a frequentist variational approach is developed in the context of binary SBM, while the method is extended to weighted graphs in Mariadassou et al. [2010]. In Picard et al. [2009], the variational procedure is applied in the context of binary SBM to

different biological networks, such as a transcriptional regulatory network, a metabolic network, a cortex network, etc. A Bayesian version of the variational approximation appears in Latouche et al. [2012a] for binary graphs (see also Latouche [2010] for more details) and more recently in Aicher et al. [2013] for weighted ones. The aim is now to approximate the posterior distribution $\mathbb{P}(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{Y})$ of both the groups and the parameters, given the data, a task realized within the class of factorized distributions. Online extensions of the (frequentist) variational approach have been given in Zanghi et al. [2008, 2010a] for binary SBM and more general weighted graphs where the conditional distribution of the edges, given the nodes groups belongs to the exponential family.

Note that efficient implementations of the variational approach for binary or weighted graphs are available in several softwares and packages: `MixNet` [Picard et al., 2009] for binary graphs, `WMixnet` [Leger, 2014] for both binary and weighted graphs, `Mixer` [R package, Ambroise et al., 2013] for binary SBM and `OSBM` for overlapping extensions [R package, Latouche et al., 2012b].

Moment methods. The very first work on stochastic block models is due to Frank and Harary [1982] where parameter estimation is based on a moment method. The article considers only a binary affiliation model and aims at estimating the parameters γ_{in} and γ_{out} under the assumption that group proportions $\boldsymbol{\pi}$ are known. Note that this method has some drawbacks and has been both discussed and extended in Section 3.1 from Ambroise and Matias [2012]. A similar approach based on moment methods has been taken in Bickel et al. [2011] for a more general model, namely the graphon model described below in Section 4.5.

Pseudo or composite likelihood methods. Besides considering moment methods for binary affiliation graphs, Ambroise and Matias [2012] propose two different composite likelihood approaches that are suited for binary and weighted affiliation graphs, respectively. In the weighted affiliation case, their approach relies on optimizing a criteria that would correspond to the the log-likelihood of the Y_{ij}'s in a model where these variables were independent. As for the binary affiliation case, considering the same criteria is not possible because in such a model the Y_{ij}'s are a mixture of two Bernoulli distributions (one for out-group connections and the other for in-group ones), whose parameters cannot be identified. However, by looking at triplets (Y_{ij}, Y_{jk}, Y_{ik}) and considering these as independent, they obtain a multivariate Bernoulli distribution whose parameters can be identified [Allman et al., 2009]. They prove convergence results justifying this pseudo-likelihood approach (see below) and exhibit a good accuracy on simulations. In the same way, Amini et al. [2013] proposed a pseudo-likelihood approach for binary graphs. Starting with an initial configuration (namely nodes groups assignment), they consider the random variables $\mathbf{b}_i = (b_{iq})_{1\leq q\leq Q}$, where b_{iq} is the number of connections of node i to nodes within class q for the given assignment of nodes classes. They consider a pseudo (or composite) likelihood of these variables (namely doing as if the $\mathbf{b}_i$'s were independent) and optimize this criterion with an `em`-algorithm. At the end of this `em` run, the resulting clustering on the nodes is used as a starting point for the next `em` run. Thus, the `em` procedure is itself iterated many times until convergence. The method is shown to perform well on simulations (but no theoretical result supports this approach w.r.t. parameter estimation).

Ad-hoc methods: nodes degrees. In Channarond et al. [2012], a method based on the degree distribution of the nodes is explored for (undirected) binary SBM. More precisely, by letting $\bar{\gamma}_q = \sum_{l=1}^{Q} \pi_l \gamma_{ql}$ be the probability of connection of a node given that it belongs to class q, the separability assumption ensures that all the $\bar{\gamma}_q, 1 \leq q \leq Q$ are distinct. Under this assumption, Channarond et al. propose a method to estimate the parameters, based only on the nodes degrees. The approach is a generalization of a proposal by Snijders and Nowicki [1997]

in the case of $Q = 2$ nodes (see Section 5 in that reference). Theoretical results associated with this method will be discussed in Section 3.3. Note that since the method only relies on the nodes degrees, it is very fast and may easily handle very large graphs.

Asymptotic properties for parameter estimates. Very few convergence results have been obtained concerning the different parameter estimation procedures developed in SBM. Concerning the variational estimates, as previously said these estimators were not expected to be convergent, in the sense that infinitely many iterations of the variational algorithm would not necessarily lead to a local maximum of the likelihood. However empirical results exhibit the accuracy of these variational estimates [Gazal et al., 2012]. This latter reference studies in fact the empirical convergence (from an asymptotic perspective, as the size of the graph increases) of three different procedures: the (frequentist) variational estimate, a belief propagation estimate and the variational Bayes estimator. The nice convergence properties of the variational estimator might be the joint consequence of two different facts:

1. the variational procedure approximates a local maxima of the likelihood, as the number of iterations increases;

2. the maximum likelihood estimator is convergent to the true parameter value, as the size of the graph increases.

Point (1) is in fact a consequence of some kind of degeneracy of the model, where asymptotically the conditional distribution of the latent variables $\mathbf{Z}$ given the observed ones $\mathbf{Y}$ is a product of Dirac masses, thus a factorized distribution and the variational approximation turns out to be asymptotically exact. We discuss this point further in Section 3.3.
Now, point (2) has been established in Celisse et al. [2012] under some assumptions. More precisely, for binary (possibly directed) SBM, the authors prove that maximum likelihood and variational estimators of the group connectivities $\boldsymbol{\gamma}$ are consistent (as the size of the graph increases). However, they can not establish the convergence of the same estimators for the groups proportions $\boldsymbol{\pi}$ without the additional assumption that the estimators of $\boldsymbol{\gamma}$ converge at rate faster than $\sqrt{\log n}/n$ (where n is the number of nodes). Note that such an assumption is not harmless since the rates of convergence of those estimates are still unknown. Moreover, up to the logarithmic term, the rate required here is $1/n$ and not $1/\sqrt{n}$, which would correspond to the parametric rate for an amount of n^2 (independent) data. The fact that the group proportion parameters $\boldsymbol{\pi}$ and the connectivity parameters $\boldsymbol{\gamma}$ fundamentally play a different role in SBM occurs in many problems, as for instance for the model selection issue (see Section 3.4).
In the affiliation SBM, Ambroise and Matias [2012] obtained convergence results for the moment estimators they proposed in the binary case as well as for the maximum composite likelihood estimators developed in both binary and weighted cases. Note that in this reference, the authors establish rates of convergence of the procedures. Surprisingly, the rates obtained there are not of the order $1/n$ but $1/\sqrt{n}$ instead. More precisely, in the more general affiliation cases, they establish an asymptotic normality result with non degenerate limiting covariance matrix ensuring that the estimators converge at the usual $1/\sqrt{n}$ parametric rate (for an amount of n "independent" data). Then, in very specific subcases (e.g. equal group proportions), the limiting covariance degenerates and rate of convergence increases to $1/n$. The issue of whether these rates are or not optimal in this context remains open.

We conclude this section by mentioning that all the above asymptotic results are established only in a dense regime where the number of edges in the graph increases as n grows to infinity. Other setups such as letting Q fixed but $\boldsymbol{\gamma} = \boldsymbol{\gamma}_n$ goes to zero, or letting $Q = Q_n$ increase to infinity and $\boldsymbol{\gamma}$ fixed need to be further investigated. In the next section, we discuss asymptotic properties of some clustering procedures. Note that procedures that asymptotically correctly

recover the nodes groups (e.g. with large probability, w.l.p.) and base their parameter estimation on these estimated groups will automatically be consistent (e.g. w.l.p.). We refer to Theorem 4.1 in Channarond et al. [2012] for a formal proof of such a result.

3.3 Clustering

Before starting this section, let us recall that clustering within SBM is not limited to community detection. The latter corresponds to the very special case of an affiliation model, with additional constraint that intra-group connectivity γ_{in} should be larger than outer-group connectivity γ_{out}. Many methods have been proposed to cluster the nodes within SBM, among which we distinguish maximum a posteriori (MAP) estimation based on the groups posterior distribution given the data, from other methods. In the first approach, parameters are estimated first (or at the same time as the clusters) and one considers the properties of the resulting posterior distribution $\mathbb{P}(\mathbf{Z}|\mathbf{Y};\hat{\theta})$ at an estimated parameter value $\hat{\theta}$, while in the second approach, the clusters are estimated first (without relying on parameter inference) and then parameters estimators are naturally obtained from these clusters.

Maximum a posteriori. As previously mentioned, Celisse et al. [2012] studied the behavior of the maximum likelihood and the variational estimators in (binary) SBM. To this aim, they have studied the posterior distribution of the groups, given the data. These authors establish two different results. The first one [Theorem 3.1 in Celisse et al., 2012] states that at the true parameter value, the groups posterior distribution converges to a Dirac mass at the actual value of groups configuration (controlling also the corresponding rate of convergence). This result is valid only at the true parameter value and not an estimated one. The second result they obtain on the convergence of the groups posterior distribution [Proposition 3.8 in Celisse et al., 2012] is valid at an estimated parameter value, provided this estimator converges at rate at least n^{-1} to the true value. Note that we already discussed rates of convergence for SBM parameters and the latter property has not been proved for any estimator yet. The article Mariadassou and Matias [In press] is more dedicated to the study of the groups posterior distribution in any binary or weighted graph (their results being in fact valid for the more general latent block model described in Section 4.2). The authors study this posterior for any parameter value in the neighborhood of the true value, thus requiring only consistency of a parameter estimator. They establish sufficient conditions for the groups posterior distribution to converge (as the size of the data increases) to a Dirac mass located at the actual (random) groups configuration. These conditions highlight the existence of particular cases in SBM, where some *equivalent configurations* exist, and exact recovery of the latent groups is not possible. They also give results in a sparse regime when the proportion of non-null entries in the data matrix converges to zero.

Note that those results of convergence of the groups posterior distribution to a (product of) Dirac mass at the actual groups configurations explains the accuracy of the variational approximation. Indeed, in an asymptotic setup, the approximation in the variational method is correct.

Other methods. Many other methods have been used to do clustering of a graph and we only consider those for which their properties have been studied under SBM assumption. Among those methods, there are some based on modularities, some based on the nodes degrees and we also mention some properties of spectral clustering in a framework related to SBM.

In Bickel and Chen [2009], the authors show that groups estimates based on the use of different modularities are consistent in the sense that with probability tending to one, these recover the original groups of a binary SBM. Quoting Bickel and Chen [2009], *the Newman-Girvan*

modularity measures the fraction of edges on the graph that connect vertices of the same type (i.e. within-community edges) minus the expected value of the same quantity on a graph with the same community divisions but random connections between the vertices. This modularity is clearly designed for community detection purposes. The authors also introduce a *likelihood modularity*, that is a profile likelihood, where the nodes groups are considered as parameters. Under a condition that is quite difficult to understand and whose consequences remain unclear (see Condition I in that reference), they establish consistency of the clustering procedures that rely on these modularities. In particular, it is unlikely that Condition I is satisfied for Newman-Girvan modularity in a non affiliation SBM. Moreover, the likelihood modularity (or profile likelihood) is computed through a stochastic search over the node labels. In practice, this might raise some issues that are not discussed by the authors.

In the specific case of a binary graph with $Q = 2$ nodes groups, Snijders and Nowicki [1997] proved that the groups could be recovered exactly with probability tending to 1, by using a method based on the nodes degrees (see Section 5 in that reference). Channarond et al. [2012] generalized the method to binary graphs with any number of groups. Under the separability assumption already mentioned above, the authors establish a bound on the probability of misclassification of at least one node.

Rohe et al. [2011] propose a classification algorithm based on spectral clustering that achieves vanishing classification error rate under a model called binary SBM. In fact, it is worth noticing that the setup is the one of *independent* Bernoulli random observations: latent groups are viewed as parameters instead of random variables. This is strictly different from SBM. In the same vein, Rohe and Yu [2012] are concerned with a framework in which nodes of a binary graph belong to two groups: a *receiver* group and a *sender* group. This is a refinement of standard SBM, which assumes equal sender and receiver groups, and is motivated by the study of directed graphs. They generalize the results of Rohe et al. [2011] to a framework called *stochastic co-block model*. Here again, this would be a generalization of SBM, except that they consider that the edges are independent random variables. The results from Rohe et al. [2011], Rohe and Yu [2012] allow the number of groups to grow with network size (i.e. nodes number) but require that node degrees increase nearly linearly with this size, an assumption that can be restrictive. Note that Choi et al. [2012] provided results for likelihood-based clustering in a sparser setup where nodes degrees increase poly-logarithmically w.r.t. the number of nodes. But here again, the setup is the one of independent Bernoulli random variables and not SBM. Also note that the complexity of the spectral clustering algorithm is $O(n^3)$, which makes its practical use computationally demanding for large networks, even though faster approximate versions exist.

3.4 Model selection

As in most discrete state-space models, the number of classes Q is unknown in general and needs to be estimated. Few model selection criteria have been proposed up to now to address this question. Nowicki and Snijders [2001] introduced an information measure and a posterior entropy parameter that both can be used to evaluate the reliability of the clustering, but with no explicit model selection procedure as for the choice of Q.
For undirected binary graphs, Daudin et al. [2008] derived an ICL-like (integrated complete likelihood) criterion. The ICL criterion has been first proposed by Biernacki et al. [2000] in the context of mixture models and is the same as the BIC (Bayesian information criterion) with an additional penalty term which corresponds to the entropy of the conditional distribution $\mathbb{P}(\mathbf{Z}|\mathbf{Y})$. Daudin et al. used the term $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\theta}))$ from the lower bound (8) as a proxy for the expectation of the complete log-likelihood. Interestingly, they end up with a two-term

penalty with the form

$$\frac{1}{2}\left((Q-1)\log n + \frac{Q(Q+1)}{2}\log\frac{n(n-1)}{2}\right),$$

where the first term refers to $\boldsymbol{\pi}$ and the second to $\boldsymbol{\gamma}$. This form reminds that the relevant sample size for the estimation of the group proportions π_q is the number of nodes, whereas it is the number of edges as for the estimation of the connection probabilities $\gamma_{q\ell}$. The same rates are observed in a simulation study of Gazal et al. [2012].

Latouche et al. [2012a] and Côme and Latouche [2013] elaborated on this approach in the context of variational Bayes inference, and proposed both a BIC and an ICL criterion. In this context, the Laplace approximation involved in the classical BIC and ICL criterion is not needed and the corresponding integral can be computed in an exact manner. Note that no formal proof of the consistency of these criteria with respect to the estimation of Q exist.

For SBM, it is most often observed that the difference between ICL and BIC is almost zero. Reminding that this difference corresponds to the conditional entropy of $\mathbf{Z}$ given $\mathbf{Y}$, this is consistent with the fact that $\mathbb{P}(\mathbf{Z}|\mathbf{Y})$ concentrates around one unique point [Mariadassou and Matias, In press, Celisse et al., 2012].

More recently, Channarond et al. [2012] proposed a criterion which does not rely on the likelihood or on some approximation of it, but only on the distribution of gaps between the ordered degrees of the nodes. This criterion is proved to be consistent.

4 Extensions of SBM

Several extensions of the SBM have been proposed in the literature with different aims. We present some of them in this section. Default notation are those of the standard SBM defined in (2).

4.1 Overlapping groups

As mentioned earlier, SBM is often used for clustering purposes, that is to assign individuals (nodes) to groups. As most clustering methods, the standard SBM assumes that each individual of the population under study belongs to one single group, as each hidden state Z_i has a multinomial distribution $\mathcal{M}(1;\boldsymbol{\pi})$ over $\{1,\ldots,Q\}$. This assumption may seem questionable, especially when analyzing social networks where an individual may play a different role in each of its relationship with other individuals. Two main alternatives have been proposed to overcome this limitation of SBM.

Airoldi et al. [2008] proposed a mixed-membership model where each node i possesses its own unknown probability vector $\boldsymbol{\pi}_i = (\pi_{i1},\ldots\pi_{iQ})$. For each link, nodes i and j first choose their group membership $Z_{i\to j}$ and $Z_{j\to i}$ for this precise link, according to their respective probability vectors $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$. The value of the link Y_{ij} is then sampled according to distribution $F(\cdot;\gamma_{Z_{i\to j},Z_{j\to i}})$. The number of hidden variables involved in this model is fairly large (there are $n(Q-1)$ independent π_{iq}'s and $2n^2$ independent $Z_{i\to j}$'s) and a `mcmc` strategy is proposed to achieve the inference. A similar mixed-membership model was proposed by Erosheva et al. [2004] in the context of a simple mixture model, without network structure.

Latouche et al. [2011] proposed an overlapping version of SBM in which individuals may belong simultaneously to any subset of classes. The group membership vector $Z_i = (Z_{iq})_{1\le q\le Q}$ is drawn as a set of independent Bernoulli variables with respective probabilities π_q (which are not required to sum to 1). Thus Z_i can take 2^Q different values, meaning that one node can belong to zero, one, two and up to Q classes. In the binary version of the overlapping SBM, the link Y_{ij} is then present with logit-probability $Z_i^\intercal W Z_j + Z_i^\intercal U + V^\intercal Z_j + W^*$, where the matrix W, the vectors U and V and the scalar W^* have to be inferred, as well as the π_q's. This model

involves $Q + (Q+1)^2$ parameters, which is much less than its mixed-membership counterpart. The authors propose a variational approach for their estimation and for the inference of the membership vectors $\{Z_i\}_{1\leq i\leq n}$.

4.2 Bipartite graphs

Some networks depict interactions or relationships between two distinct types of entities, such as authors and journals, chemical compounds and reactions, hosts species and parasites species, etc. In such networks the link has most often an asymmetric meaning, such as 'published an article in', 'contributes to' or 'is contaminated by'. In such networks, no link between nodes of the same type can exist. When considering n nodes of the first type and m nodes of the second type, the adjacency matrix (Y_{ij}) is rectangular with n rows and m columns. The SBM model can rephrase in an asymmetric way, denoting $\{Z_i\}_{1\leq i\leq n} \in \{1,\dots Q\}^n$ the memberships of the row nodes and $\{W_j\}_{1\leq j\leq m} \in \{1,\dots,K\}^m$ the memberships of the column nodes. All membership variables are drawn independently with multinomial distribution $\mathcal{M}(1;\boldsymbol{\pi})$ for the Z_i's and $\mathcal{M}(1;\boldsymbol{\rho})$ for the W_j's. Links $\{Y_{ij}\}$ are then drawn independently conditional on $\{Z_i, W_j\}_{1\leq i\leq n, 1\leq j\leq m}$ and with respective distribution $F(\cdot;\gamma_{Z_i,W_j})$.

This model is actually a latent block model (LBM) first proposed by Govaert and Nadif [2003] in the context of bi-clustering to infer simultaneously Q row groups and K column groups. The same authors proposed a variational approximation for the parameter inference. More recently, Keribin et al. [2012] proposed a model selection criterion in a variational Bayes context and Mariadassou and Matias [In press] proved the convergence of the conditional distribution of the memberships toward the true ones.

4.3 Degree-corrected block model

The regular SBM model assumes that the expected connectivity of a node only depends on the group it belongs to. Indeed, in many situations, beside their group membership, some nodes may be likely to be more connected than others because of their individual specificities. Karrer and Newman [2011] extended the Poisson-valued SBM taking

$$Y_{ij}|Z_i = q, Z_j = \ell \sim \mathcal{P}(\gamma_{q\ell}\kappa_i\kappa_j),$$

where $\gamma_{q\ell}$ plays the same role as in the Poisson-valued SBM and κ_i controls the expected degree of node i. A similar model is considered in Mørup and Hansen [2009], who proposed a relaxation strategy for the inference of the parameters. Zhu et al. [2013] proposed an oriented version of this model and use `mcmc` for the parameter inference. An asymmetric version is considered in Reichardt et al. [2011], generalizing LBM in the same way. Yan et al. [2012] proposed a likelihood-ratio test for the comparison of the degree corrected block model with the regular SBM and Zhao et al. [2012] provided a general characterization of the community detection criteria that will provide a consistent classification under the degree-corrected SBM.

4.4 Accounting for covariates

The existence or the value of the links between individuals can sometimes be partially explained by observed covariates. Accounting for such covariates is obviously desirable to better understand the network structure. It is first important to distinguish if the covariates are observed at the node level (e.g. $\mathbf{x}_i$ = (age, sex) of individual i) or at the edge level (e.g. $\mathbf{x}_{ij}$ = (genetic similarity, spatial distance) between species i and j). Node covariates $\mathbf{x}_i = (x_{i1},\dots,x_{ip})$ are sometimes transformed into edge covariates taking, e.g., $\mathbf{x}_{ij} = (|x_{i1}-x_{j1}|,\dots,|x_{ip}-x_{jp}|)$. Pattison and Robins [2007] propose a brief review of how such an information can be incorporated in a random graph model in absence of hidden structure. For example, in presence of edge

covariates, a simple logistic regression model can be fitted as

$$(Y_{ij})_{i,j} \text{ independent}, \qquad \text{logit}\left(\mathbb{P}\{Y_{ij}=1\}\right) = \mathbf{x}_{ij}^\intercal \boldsymbol{\beta},$$

where $\mathbf{x}_{ij}$ is the vector of covariates and $\boldsymbol{\beta}$ the vector of regression parameters. The statistical inference of this model raises no specific issue.

We now focus on how covariates can be used to enrich SBM. For node covariates $\{\mathbf{x}_i\}$, Tallberg [2005] proposed a multinomial probit model where the distribution of the hidden class Z_i depends on the vector of covariates $\mathbf{x}_i$

$$(Z_i)_i \text{ independent}, \quad Z_i \sim \mathcal{M}(1; \boldsymbol{\pi}(\mathbf{x}_i)).$$

This model states that the covariates act on the edge value through the membership of the nodes. In this context, the author proposed a Bayesian inference approach for which some full conditional distributions can be derived in a close form.

In presence of edge covariates $\{\mathbf{x}_{ij}\}$, Mariadassou et al. [2010] proposed to combine them with the hidden structure using the generalized linear model framework. In the Poisson case, this leads to

$$Y_{ij}|Z_i = q, Z_j = \ell \sim \mathcal{P}(\exp(\mathbf{x}_{ij}^\intercal \boldsymbol{\beta} + \gamma_{q\ell})).$$

A similar Gaussian model is considered in Zanghi et al. [2010b]. In both cases, the proposed inference method relies on a variational approximation. Mariadassou et al. [2010] argue that the term related to the hidden state $\gamma_{q\ell}$ measures the heterogeneity in the network that is not explained by the regression term $\mathbf{x}_{ij}^\intercal \boldsymbol{\beta}$. A similar interpretation is given by Choi et al. [2012], who first apply a regression step and then perform SBM inference on the residuals.

4.5 SBM as a graphon model

It can be seen that SBM is a graphon model, as defined (1), where the function g is block-wise constant, the blocks being rectangular with respective dimension $\pi_q \times \pi_\ell$ and height $\gamma_{q\ell}$. Spliting the interval $[0,1]$ into Q sub-intervals with respective widths $\pi_1, \ldots, \pi_Q$, the hidden state Z_i of SBM is simply the number of the sub-interval into which the corresponding U_i falls.

The inference of the function g has received few attention until now. Chatterjee [2012] first proposed a direct estimation of the $p_{ij} = g(U_i, U_j)$ based on a low rank decomposition of the adjacency matrix. More recently, several articles [Airoldi et al., 2013, Wolfe and Olhede, 2013, Latouche and Robin, 2013, Olhede and Wolfe, 2013] proposed to infer g through the parameter of a SBM, which can be seen as a discrete approximation of the graphon.

It is important to note that the graphon model suffers a strong and intrinsic identifiability problem as composing g with any measure-preserving function from the unit interval to itself will result in the same model. This issue is accounted for in all the papers cited above, but the interpretability of the resulting function is still questionable. As shown in Diaconis and Janson [2008], subgraphs (also called 'motifs') frequency are invariant to such transformation and therefore characterize a W graph. The inference of this frequencies is addressed by Latouche and Robin [2013].

4.6 Network evolution

More and more attention is paid to evolutionary networks, that is network in which the value of edges may vary along time. In this this setting, the data at hand consists in $\{Y_{ij}(t)\}$ where, typically, t belongs to a finite set of observation times. The set of nodes is kept fixed here.

Recently, different proposals have been made to deal with groups structure within dynamic networks. We mention here only the models connected with SBM. DuBois et al. [2013] suppose that the membership of the nodes are kept fixed and that the edges values evolve according to a conditional Markov jump process. In the model considered by Fu et al. [2009] both the

memberships and the conditional connection probabilities evolve along time. Xu and Hero III [2013] also propose a dynamic version of SBM where both node memberships and edges evolve along time. In this paper, the parameter inference relies on a extended Kalman filter algorithm.

5 Some perspectives

In this last section, we try to briefly underline what we think are the next challenges in the modeling of heterogeneity for networks.

First, it is important to develop scalable methods that are able to handle very large graphs. Recently, Vu et al. [2013] proposed a model for clustering in very large networks: they handled a dataset with more than 131,000 nodes and 17 billions edge variables. They consider discrete-valued networks, possibly with covariates, and assuming independence of each dyad (namely Y_{ij} in the case of undirected edges and (Y_{ij}, Y_{ji}) in the case of directed edges), conditional on the nodes groups. They rely on the variational approximation of the `em` algorithm but replace the `e`-step by a minorization-maximization algorithm. As a result, this increases the lower bound in the variational approximation instead of maximizing it at each step. However, the large network they handle is still very sparse (with about only 840,798 non zero edges) and the number of groups they use $Q = 5$ is small with respect to the sample size. Indeed, there is still room for improvement to make model-based clustering methods scalable to very large networks.

Second, we feel that there is a great need in terms of statistical methods for the analysis of evolutionary networks. Indeed many technologies now give access to follow-up observations of social or biological networks. A strong attention should be paid to both proper dynamical modeling and their associate inference. As shown throughout this paper, due to the network structure, statistical models for statistical network suffer from very intricate dependency structures between the nodes. Accounting for dependency along time will obviously make it even more complex. The conception of both statistically valid and computationally efficient inference methods is an interesting challenge.

Finally, statistical properties of the models and procedures should be further studied from a theoretical point of view. Asymptotic results have recently been established but finite sample properties would also be welcome. Much attention has been paid to the dense case and sparser setups still need to be studied. Validation of the procedures can not be limited to simulations and theoretical studies will help better understand the models and thus design new inference methods.

References

C. Aicher, A. Jacobs, and A. Clauset. Adapting the stochastic block model to edge-weighted networks. ICML Workshop on Structured Learning (SLG), 2013.

E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 692–700. 2013.

E. Allman, C. Matias, and J. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 2009.

E. Allman, C. Matias, and J. Rhodes. Parameters identifiability in random graph mixture models. *Journal of Statistical Planning and Inference*, 141(5):1719–1736, 2011.

C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random graph mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):3–35, 2012.

C. Ambroise, G. Grasseau, M. Hoebeke, P. Latouche, V. Miele, and F. Picard. mixer: Routines for the analysis (unsupervised clustering) of networks using mixtures of Erdös-Rényi random graphs, 2013. R package version 1.7 — For new features, see the 'Changelog' file (in the package source). `http://cran.r-project.org/web/packages/mixer/`.

A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.*, 41(4):2097–2122, 2013.

P. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *PNAS*, 106(50):21068–21073, 2009.

P. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Ann. Statist.*, 39(5):2280–2301, 2011.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel.*, 22(7):719–725, 2000.

B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Rand. Struct. Algo.*, 31(1):3–122, 2007.

O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005.

A. Celisse, J.-J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model. *Electron. J. Statist.*, 6:1847–1899, 2012.

A. Channarond. *Recherche de structure dans un graphe aléatoire: modèles à espace latent*. PhD thesis, Université Paris 11, Orsay, 2013.

A. Channarond, J.-J. Daudin, and S. Robin. Classification and estimation in the Stochastic Blockmodel based on the empirical degrees. *Electron. J. Stat.*, 6:2574–2601, 2012.

S. Chatterjee. Matrix estimation by Universal Singular Value Thresholding. Technical report, arXiv:1212.1247, 2012.

D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–84, 2012.

E. Côme and P. Latouche. Model selection and clustering in stochastic block models with the exact integrated complete data likelihood. Technical report, arXiv:1303.2962, 2013.

J.-J. Daudin. A review of statistical models for clustering networks with an application to a PPI network. *Journal de la Société Française de Statistique*, 152(2):111–125, 2011.

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Stat. Comput.*, 18 (2):173–183, 2008.

J.-J. Daudin, L. Pierre, and C. Vacher. Model for heterogeneous random networks using continuous latent variables and an application to a tree–fungus network. *Biometrics*, 66(4): 1043–1051, 2010.

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

P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica*, 28:33–61, 2008.

C. DuBois, C. Butts, and P. Smyth. Stochastic blockmodeling of relational event dynamics. *J. Machine Learning Res., Workshop & Conference Proc.*, 31:238–46, 2013.

Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48(6): 1518–1569, 2002. Special issue on Shannon theory: perspective, trends, and applications.

E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *PNAS*, 97(22):11885–11892, 2004.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75 – 174, 2010.

O. Frank and F. Harary. Cluster inference by using transitivity indices in empirical graphs. *J. Amer. Statist. Assoc.*, 77(380):835–840, 1982.

W. Fu, L. Song, and E. P. Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th annual international conference on machine learning*, pages 329–36. ACM, 2009.

S. Gazal, J.-J. Daudin, and S. Robin. Accuracy of variational estimates for random graph mixture models. *Journal of Statistical Computation and Simulation*, 82(6):849–862, 2012.

W. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice (Chapman & Hall/CRC Interdisciplinary Statistics)*. Chapman and Hall/CRC, 1995.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, 2010.

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

A. Gunawardana and W. Byrne. Convergence theorems for generalized alternating minimization procedures. *J. Mach. Learn. Res.*, 6:2049–2073, 2005.

M. Handcock, A. Raftery, and J. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–54, 2007.

P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.*, 97(460):1090–98, 2002.

P. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: some first steps. *Social networks*, 5:109–137, 1983.

T. S. Jaakkola. Tutorial on variational approximation methods. In *In Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press, 2000.

Q. Jiang, Y. Zhang, and M. Sun. Community detection on weighted networks: A variational Bayesian method. In Z.-H. Zhou and T. Washio, editors, *Advances in Machine Learning*, volume 5828 of *Lecture Notes in Computer Science*, pages 176–190. Springer Berlin Heidelberg, 2009.

B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83:016107, 2011.

C. Keribin, V. Brault, G. Celeux, G. Govaert, et al. Model selection for the binary latent block model. In *20th International Conference on Computational Statistics*, 2012.

E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

P. Latouche. *Modèles de graphes aléatoires à structure cachée pour l'analyse des réseaux*. PhD thesis, Université d'Évry val d'Essonne, 2010.

P. Latouche and S. Robin. Bayesian model averaging of stochastic block models to estimate the graphon function and motif frequencies in a W-graph model. Technical report, arXiv:1310.6150, 2013.

P. Latouche, E. Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat.*, 5(1):309–336, 2011.

P. Latouche, E. Birmelé, and C. Ambroise. Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115, 2012a.

P. Latouche, E. Birmelé, and C. Ambroise. OSBM: Overlapping stochastic blockmodel, 2012b. R package. `http://stat.genopole.cnrs.fr/logiciels/osbm/`.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

J.-B. Leger. Wmixnet: Software for clustering the nodes of binary and valued graphs using the stochastic block model. Technical report, arXiv:1402.3410, 2014. `http://www.agroparistech.fr/mia/productions:logiciel:wmixnet`.

J.-B. Leger, C. Vacher, and J.-J. Daudin. Detection of structurally homogeneous subsets in graphs. *Statistics and computing*, pages 1–18, 2013.

L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933 – 957, 2006.

M. Mariadassou and C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, In press.

M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Stat.*, 4(2):715–42, 2010.

M. Mørup and L. K. Hansen. Learning latent structure in complex networks. In *NIPS Workshop on Analyzing Networks and Learning with Graphs*, 2009.

C. Nickel. *Random dot product graphs: A model for social networks*. PhD thesis, Johns Hopkins University, 2006.

K. Nowicki and T. Snijders. Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.*, 96(455):1077–1087, 2001.

S. C. Olhede and P. J. Wolfe. Network histograms and universality of blockmodel approximation. Technical report, arXiv:1312.5306, 2013.

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

T. P. Peixoto. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Phys. Rev. E*, 89:012804, 2014.

F. Picard, V. Miele, J.-J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics*, 10:1–11, 2009.

J. Reichardt, R. Alamino, and D. Saad. The interplay between microscopic and mesoscopic structures in complex networks. *PloS one*, 6(8):e21282, 2011.

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p^*) models for social networks. *Social networks*, 29(2):173–191, 2007.

K. Rohe and B. Yu. Co-clustering for directed graphs: the stochastic co-blockmodel and a spectral algorithm. Technical report, arXiv:1204.2296, 2012.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block model. *Ann. Statist*, 39(4):1878–1915, 2011.

T. A. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37:129–151, 2011.

T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification*, 14(1):75–100, 1997.

C. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4), 2012.

E. Szemerédi. Regular partitions of graphs. In *Problèmes combinatoires et théorie des graphes*, volume 260 of *Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976*, pages 399–401, 1978.

C. Tallberg. A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1):1–23, 2005.

M. Tang, D. Sussman, and C. Priebe. Universally consistent vertex classification for latent positions graphs. *Ann. Statist.*, 41(3):1406–1430, 2013.

D. Vu, D. Hunter, and M. Schweinberger. Model-based clustering of large networks. *Ann. Appl. Stat.*, 7(2):613–1248, 2013.

P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. Technical report, arXiv:1309.5936, 2013.

K. S. Xu and A. Hero III. Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In A. Greenberg, W. Kennedy, and N. Bos, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 7812 of *Lecture Notes in Computer Science*, pages 201–210. Springer Berlin Heidelberg, 2013.

X. Yan, C. Rohilla Shalizi, J. E. Jensen, F. Krzakala, C. Moore, L. Zdeborova, P. Zhang, and Y. Zhu. Model selection for degree-corrected block models. Technical report, arXiv:1207.3994, 2012.

S. J. Young and E. R. Scheinerman. Random dot product graph models for social networks. In A. Bonato and F. R. Chung, editors, *Algorithms and Models for the Web-Graph*, volume 4863 of *Lecture Notes in Computer Science*, pages 138–149. Springer Berlin Heidelberg, 2007.

H. Zanghi, C. Ambroise, and V. Miele. Fast online graph clustering via Erdős Rényi mixture. *Pattern Recognition*, 41(12):3592–3599, 2008.

H. Zanghi, F. Picard, V. Miele, and C. Ambroise. Strategies for online inference of model-based clustering in large and growing networks. *Ann. Appl. Stat.*, 4(2):687–714, 2010a.

H. Zanghi, S. Volant, and C. Ambroise. Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, 31:830–836, 2010b.

Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.*, 40(4):2266–2292, 2012.

Y. Zhu, X. Yan, and C. Moore. Oriented and degree-generated block models: generating and inferring communities with inhomogeneous degree distributions. *Journal of Complex Networks*, 2013.