OVERLAPPING STOCHASTIC BLOCK MODELS WITH APPLICATION TO THE FRENCH POLITICAL BLOGOSPHERE[1]

BY PIERRE LATOUCHE, ETIENNE BIRMELÉ AND CHRISTOPHE AMBROISE

University of Evry

Complex systems in nature and in society are often represented as networks, describing the rich set of interactions between objects of interest. Many deterministic and probabilistic clustering methods have been developed to analyze such structures. Given a network, almost all of them partition the vertices into *disjoint* clusters, according to their connection profile. However, recent studies have shown that these techniques were too restrictive and that most of the existing networks contained overlapping clusters. To tackle this issue, we present in this paper the Overlapping Stochastic Block Model. Our approach allows the vertices to belong to multiple clusters, and, to some extent, generalizes the well-known Stochastic Block Model [Nowicki and Snijders (2001)]. We show that the model is generically identifiable within classes of equivalence and we propose an approximate inference procedure, based on global and local variational techniques. Using toy data sets as well as the French Political Blogosphere network and the transcriptional network of *Saccharomyces cerevisiae*, we compare our work with other approaches.

1. Introduction. Networks have been extensively studied ever since the work of Moreno (1934). They are used in many scientific fields to represent the interactions between objects of interest. For instance, in Biology, regulatory networks can describe the regulation of genes with transcriptional factors [Milo et al. (2002)], while metabolic networks focus on representing pathways of biochemical reactions [Lacroix, Fernandes and Sagot (2006)]. In the social sciences, networks are commonly used to represent relational ties between actors [Snijders and Nowicki (1997); Nowicki and Snijders (2001)].

In this context, many deterministic and probabilistic clustering methods have been used to acquire knowledge from the network topology. As shown in Newman and Leicht (2007), most of these techniques seek specific structures in networks. Thus, some models look for community structure where vertices are partitioned into classes such that vertices of a class are mostly connected to vertices of the same class [Hofman and Wiggins (2008)]. They are particularly suitable for the analysis of affiliation networks [Latouche, Birmelé and Ambroise (2009)]. Most existing community discovery algorithms are based on the modularity score of

---

Girvan and Newman (2002). However, Bickel and Chen (2009) showed that these algorithms were (asymptotically) biased and that using modularity scores could lead to the discovery of an incorrect community structure, even for large graphs. The model of Handcock, Raftery and Tantrum (2007) which extends Hoff, Raftery and Handcock (2002) is an alternative approach. Vertices are clustered depending on their positions in a continuous latent space. They proposed a Bayesian inference procedure, based on Markov Chain Monte Carlo (MCMC), which is implemented in the R package latentnet [Krivitsky and Handcock (2009)], as well an asymptotic BIC criterion. Other models look for disassortative mixing in which vertices mostly connect to vertices of different classes. They are commonly used to analyze bipartite networks [Estrada and Rodriguez-Velazquez (2005)] which are present in many applications. For more details, see Newman and Leicht (2007).

The Stochastic Block Model (SBM) can uncover heterogeneous structures in a large variety of networks [Latouche, Birmelé and Ambroise (2009)]. Originally developed in the social sciences, SBM is a probabilistic generalization [Fienberg and Wasserman (1981); Holland, Laskey and Leinhardt (1983)] of the method described in White, Boorman and Breiger (1976). Given a network, it assumes that each vertex belongs to a latent class among Q classes and uses a $Q \times Q$ connectivity matrix $\mathbf{\Pi}$ to describe the connection probabilities [Frank and Harary (1982)]. No assumption is made on $\mathbf{\Pi}$ such that SBM is a very flexible model. In particular, it can be used, among others, to look for community structure and disassortative mixing. Many inference methods have been employed to estimate the SBM parameters. They all face the same problem. Indeed, contrary to Gaussian mixture models or other usual mixture models, the posterior distribution $p(\mathbf{Z}|\mathbf{X})$, of all the hidden label variables, given the observation $\mathbf{X}$, cannot be factorized due to conditional dependency. Nowicki and Snijders (2001) proposed a Bayesian probabilistic approach. Their algorithm is implemented in the software BLOCKS, which is part of the package StoCNET [Boer et al. (2006)]. It uses Gibbs sampling to approximate the posterior distributions and leads to accurate a posteriori estimates. Two model based criteria have been proposed to choose the optimal value of Q. Thus, Daudin, Picard and Robin (2008) used an ICL criterion, based on a Laplace approximation of the Integrated Classification Likelihood, while Latouche, Birmelé and Ambroise (2009) used a nonasymptotic approximation of the marginal likelihood. For an extensive discussion on statistical network models and blockmodel selection, we refer to Goldenberg et al. (2010).

A drawback of existing graph clustering techniques is that they all partition the vertices into disjoint clusters, while lots of objects in real world applications typically belong to multiple groups or communities. For instance, many proteins, so-called *moonlighting proteins*, are known to have several functions in the cells [Jeffery (1999)], and actors might belong to several groups of interests [Palla et al. (2005)]. Thus, a graph clustering method should be able to uncover overlapping clusters. This issue has received growing attention in the last few years, starting with an algorithmic approach based on small complete sub-graphs developed by

Palla et al. (2005) and implemented in the software CFinder [Palla et al. (2006)]. They defined a k-clique community as a union of all k-cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent[2] k-cliques. Given a network, their algorithm first locates all cliques and then identifies the communities using a clique–clique overlap matrix [Everett and Borgatti (1998)]. By construction, the resulting communities can overlap. In order to select the optimal value of k, the authors suggested a global criterion which looks for a community structure as highly connected as possible. Small values of k lead to a giant community which smears the details of a network by merging small communities. Conversely, when k increases, the communities tend to become smaller, more disintegrated, but also more cohesive. Therefore, they proposed a heuristic which consists in running their algorithm for various values of k and then to select the lowest value such that no giant community appears.

More recent work [Airoldi et al. (2008)] proposed the Mixed Membership Stochastic Block model (MMSB) which has been used with success to analyze networks in many applications [Airoldi et al. (2007); Airoldi et al. (2006)]. They used variational techniques to estimate the model parameters and proposed a criterion to select the number of classes. As detailed in Heller, Williamson and Ghahramani (2008), mixed membership models, as Latent Dirichlet Allocation [Blei, Ng and Jordan (2003)], are flexible models which can capture partial membership [Griffiths and Ghahramani (2005); Heller and Ghahramani (2007)], in the form of attribute-specific mixtures. In MMSB, a mixing weight vector $\boldsymbol{\pi}_i$ is drawn from a Dirichlet distribution for each vertex in the network, π_{iq} being the probability of vertex i to belong to class q. The edge probability from vertex i to vertex j is then given by $p_{ij} = \mathbf{Z}_{i \to j}^{\top} \mathbf{B} \mathbf{Z}_{i \leftarrow j}$, where $\mathbf{B}$ is a $Q \times Q$ matrix of connection probabilities similar to the $\boldsymbol{\Pi}$ matrix in SBM. The vector $\mathbf{Z}_{i \to j}$ is sampled from a multinomial distribution $\mathcal{M}(1, \boldsymbol{\pi}_i)$ and describes the class membership of vertex i in its relation toward vertex j. By symmetry, the vector $\mathbf{Z}_{i \leftarrow j}$ is drawn from a multinomial distribution $\mathcal{M}(1, \boldsymbol{\pi}_j)$ and represents the class membership of vertex j in its relation toward vertex i. Thus, depending on its relations with other vertices, each vertex can belong to different classes and, therefore, MMSB can be viewed as allowing overlapping clusters. However, the limit of MMSB is that it does not produce edges which are themselves influenced by the fact that some vertices belong to multiple clusters. Indeed, for every pair (i, j) of vertices, only a single draw of $\mathbf{Z}_{i \to j}$ and $\mathbf{Z}_{i \leftarrow j}$ determines the probability p_{ij} of an edge, all the other class memberships of vertex i and j toward other vertices in the network do not play a part. In this paper we present a complementary approach which tackles this issue.

Fu and Banerjee (2008) modeled overlapping clusters on Q components by characterizing each individual i by a latent $\{0, 1\}$ vector z_i of length Q drawn

---

[2]Two k-cliques are adjacent if they share $k - 1$ vertices.

from independent Bernoulli distributions. The ith row of the data matrix then only depends on z_i. In the underlying clustering structure, i belongs to the components corresponding to a 1 in z_i. Nevertheless, the proposed model needs Q parameters for each individual and supposes independence between rows and columns of the data matrix, which is not the case when looking for network structures.

In this paper we propose a new model for generating networks, depending on $(Q + 1)^2 + Q$ parameters, where Q is the number of components in the mixture. A latent $\{0, 1\}$-vector of length Q is assigned to each vertex, drawn from products of Bernoulli distributions whose parameters are not vertex-dependent. Each vertex may then belong to several components, allowing overlapping clusters, and each edge probability depends only on the components of its endpoints.

In Section 2 we briefly review the stochastic block model introduced by Nowicki and Snijders (2001). In Section 3 we present the overlapping stochastic block model and we show in Section 4 that the model is identifiable within classes of equivalence. In Section 5 we propose an EM-like algorithm to infer the parameters of the model. Finally, in Section 6 we compare our work with other approaches using simulated data and two real networks. We show the efficiency of our model to detect overlapping clusters in networks.

2. The stochastic block model. In this paper we consider a directed binary random graph $\mathcal{G}$ represented by an $N \times N$ binary adjacency matrix $\mathbf{X}$. Each entry X_{ij} describes the presence or absence of an edge from vertex i to vertex j. We assume that $\mathcal{G}$ does not have any self loop, and, therefore, the variables X_{ii} will not be taken into account. The Stochastic Block Model (SBM) introduced by Nowicki and Snijders (2001) associates to each vertex of a network a latent variable $\mathbf{Z}_i$ drawn from a multinomial distribution:

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

where $\boldsymbol{\alpha}$ denotes the vector of class proportions. As in other standard mixture models, the vector $\mathbf{Z}_i$ sees all its components set to zero except one such that $Z_{iq} = 1$ if vertex i belongs to class q. The model then verifies

$$(2.1) \qquad \sum_{q=1}^{Q} Z_{iq} = 1 \qquad \forall i \in \{1, \ldots, N\}$$

and

$$(2.2) \qquad \sum_{q=1}^{Q} \alpha_q = 1.$$

Finally, the edges of the network are drawn from a Bernoulli distribution:

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

where $\mathbf{\Pi}$ is a $Q \times Q$ matrix of connection probabilities. According to this model, the latent variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ are i.i.d. and given this latent structure, all the edges are supposed to be independent. Note that SBM was originally described in a more general setting, allowing any discrete relational data. However, as explained previously, we concentrate in the following on binary edges only.

3. The overlapping stochastic block model. In order to allow each vertex to belong to multiple classes, we relax the constraints (2.1) and (2.2). Thus, for each vertex i of the network, we introduce a latent vector $\mathbf{Z}_i$, of Q independent Boolean variables $Z_{iq} \in \{0, 1\}$, drawn from a multivariate Bernoulli distribution:

$$(3.1) \qquad \mathbf{Z}_i \sim \prod_{q=1}^{Q} \mathcal{B}(Z_{iq}; \alpha_q) = \prod_{q=1}^{Q} \alpha_q^{Z_{iq}} (1 - \alpha_q)^{1 - Z_{iq}}.$$

We point out that $\mathbf{Z}_i$ can also have all its components set to zero which is a useful feature in practice as described in Sections 3.2 and 6. The edge probabilities are then given by

$$X_{ij}|\mathbf{Z}_i, \mathbf{Z}_j \sim \mathcal{B}(X_{ij}; \mathrm{g}(a_{\mathbf{Z}_i, \mathbf{Z}_j})) = e^{X_{ij} a_{\mathbf{Z}_i, \mathbf{Z}_j}} \mathrm{g}(-a_{\mathbf{Z}_i, \mathbf{Z}_j}),$$

where

$$(3.2) \qquad a_{\mathbf{Z}_i, \mathbf{Z}_j} = \mathbf{Z}_i^{\top} \mathbf{W} \mathbf{Z}_j + \mathbf{Z}_i^{\top} \mathbf{U} + \mathbf{V}^{\top} \mathbf{Z}_j + W^*,$$

and $\mathrm{g}(x) = (1 + e^{-x})^{-1}$ is the logistic sigmoid function. $\mathbf{W}$ is a $Q \times Q$ real matrix, whereas $\mathbf{U}$ and $\mathbf{V}$ are Q-dimensional real vectors. The first term in the right-hand side of (3.2) describes the interactions between the vertices i and j. If i belongs only to class q and j only to class l, then only one interaction term remains ($\mathbf{Z}_i^{\top} \mathbf{W} \mathbf{Z}_j = W_{ql}$). However, as illustrated in Table 1, the model can take more complex interactions into account if one or both of these two vertices belong to multiple classes (Figure 1). Note that the second term in (3.2) does not depend on $\mathbf{Z}_j$. It models the overall capacity of vertex i to connect to other vertices. By symmetry, the third term represents the global tendency of vertex j to receive an edge. These two parameters $\mathbf{U}$ and $\mathbf{V}$ are related to the sender/receiver effects δ_i and γ_j in the Latent Cluster Random Effects Model (LCREM) of Krivitsky et al. (2009). However, contrary to LCREM, $\delta_i = \mathbf{Z}_i^{\top} \mathbf{U}$ and $\gamma_j = \mathbf{V}^{\top} \mathbf{Z}_j$ depend on the classes. In other words, two different vertices sharing the same classes will have exactly the same sender/receiver effects, which is not the case in LCREM. Finally, we use the scalar W^* as a bias, to model sparsity.

If we associate to each latent variable $\mathbf{Z}_i$ a vector $\widetilde{\mathbf{Z}}_i = (\mathbf{Z}_i, 1)^{\top}$, then (3.2) can be written

$$(3.3) \qquad a_{\mathbf{Z}_i, \mathbf{Z}_j} = \widetilde{\mathbf{Z}}_i^{\top} \widetilde{\mathbf{W}} \widetilde{\mathbf{Z}}_j,$$

where

$$\widetilde{\mathbf{W}} = \begin{pmatrix} \mathbf{W} & \mathbf{U} \\ \mathbf{V}^{\top} & W^* \end{pmatrix}.$$

TABLE 1
The values of $a_{\mathbf{Z}_i, \mathbf{Z}_j}$ in functions of $\mathbf{Z}_i$ (rows) and $\mathbf{Z}_j$ (columns) for an overlapping stochastic block model with $Q = 2$

	(0, 0)	(1, 0)	(0, 1)	(1, 1)
(0, 0)	W^*	$V_1 + W^*$	$V_2 + W^*$	$V_1 + V_2 + W^*$
(1, 0)	$U_1 + W^*$	$W_{11} + U_1 + V_1 + W^*$	$W_{12} + U_1 + V_2 + W^*$	$W_{11} + W_{12} + U_1 + V_1 + V_2 + W^*$
(0, 1)	$U_2 + W^*$	$W_{21} + U_2 + V_1 + W^*$	$W_{22} + U_2 + V_2 + W^*$	$W_{21} + W_{22} + U_2 + V_1 + V_2 + W^*$
(1, 1)	$U_1 + U_2 + W^*$	$W_{11} + W_{21} + U_1 + U_2 + V_1 + W^*$	$W_{12} + W_{22} + U_1 + U_2 + V_2 + W^*$	$W_{11} + W_{12} + W_{21} + W_{22} + U_1 + U_2 + V_1 + V_2 + W^*$

The $\tilde{Z}_{i(Q+1)}$'s can be seen as random variables drawn from a Bernoulli distribution with probability $\alpha_{Q+1} = 1$. Thus, one way to think about the model is to consider that all the vertices in the graph belong to a $(Q + 1)$th cluster which is overlapped by all the other clusters. In the following, we will use (3.3) to simplify the notation.

Finally, given the latent structure $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_N\}$, all the edges are supposed to be independent (see Figure 2). Thus, when considering directed graphs without self-loop, the Overlapping Stochastic Block Model (OSBM) is defined through the following distributions:

$$(3.4) \qquad p(\mathbf{Z}|\boldsymbol{\alpha}) = \prod_{i=1}^N \prod_{q=1}^Q \alpha_q^{Z_{iq}} (1 - \alpha_q)^{1 - Z_{iq}}$$
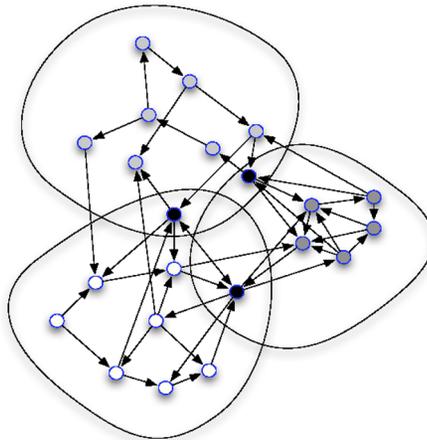

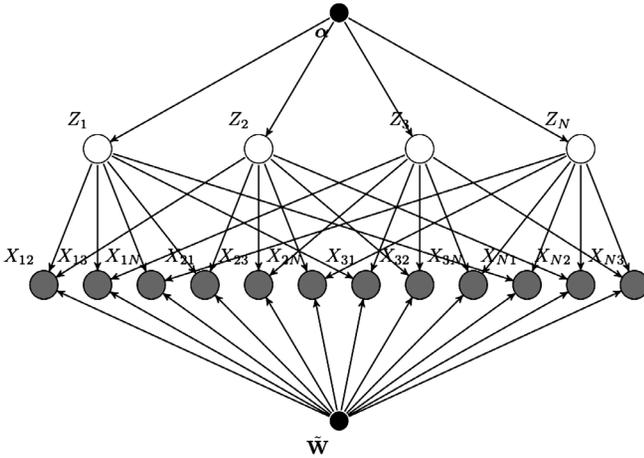FIG. 1. *Example of a directed graph with three overlapping clusters.*

FIG. 2. *Graphical representation of the overlapping stochastic block model.*

and

$$p(\mathbf{X}|\mathbf{Z}, \widetilde{\mathbf{W}}) = \prod_{i \neq j}^{N} e^{X_{ij} a_{\mathbf{Z}_i, \mathbf{Z}_j}} g(-a_{\mathbf{Z}_i, \mathbf{Z}_j}).$$

3.1. *Modeling sparsity.* As explained in Airoldi et al. (2008), real networks are often sparse[3] and it is crucial to distinguish the two sources of noninteraction. Sparsity might be the result of the rarity of interactions in general, but it might also indicate that some class (*intra* or *inter*) connection probabilities are close to zero. For instance, social networks (see Section 6.2) are often made of communities where vertices are mostly connected to vertices of the same community. This corresponds to classes with high *intra* connection probabilities and low *inter* connection probabilities. In (3.2) we can notice that W^* appears in $a_{\mathbf{Z}_i, \mathbf{Z}_j}$ for every pair of vertices. Therefore, W^* is a convenient parameter to model the two sources of sparsity. Indeed, low values of W^* result from the rarity of interactions in general, whereas high values signify that sparsity comes from the classes (parameters in $\mathbf{W}$, $\mathbf{U}$ and $\mathbf{V}$).

3.2. *Modeling outliers.* When applied on real networks, graph clustering methods often lead to giant classes of vertices having low output and input degrees [Daudin, Picard and Robin (2008); Latouche, Birmelé and Ambroise (2009)]. These classes are usually discarded and the analysis of networks focus on more highly structured classes to extract useful information. The product of Bernoulli distributions (3.4) provides a natural way to encode these "outliers." Indeed, rather

---

[3]The corresponding adjacency matrices contain mainly zeros.

than using giant classes, OSBM uses the null component such that $\mathbf{Z}_i = \mathbf{0}$ if vertex i is an outlier and should not be classified in any class.

4. Identifiability. Before looking for an optimization procedure to estimate the model parameters, given a sample of observations (a network), it is crucial to verify whether OSBM is identifiable. A theorem of Allman, Matias and Rhodes (2009) lies at the core of the results presented in this section.

If we denote $\mathcal{F}(\Theta) = \{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$, a family of models we are interested in, the classical definition of identifiability requires that for any two different values $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, the corresponding probability distributions $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\boldsymbol{\theta}'}$ are different.

4.1. *Correspondence with (nonoverlapping) stochastic block models.* Let Θ_{OSBM} be the parameter space of the family of OSBMs with Q classes:

$$\Theta_{\mathrm{OSBM}} = \{(\boldsymbol{\alpha}, \widetilde{\mathbf{W}}) \in [0, 1]^Q \times \mathbb{R}^{(Q+1)^2}\}.$$

Each $\boldsymbol{\theta}$ in Θ_{OSBM} corresponds to a random graph model which is defined by the distribution $p(\mathbf{X}|\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$. The aim of this Section is to characterize whether there exists any relation between two different parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ in Θ_{OSBM}, leading to the same random graph model.

We consider the (nonoverlapping) Stochastic Block Model (SBM) introduced by Nowicki and Snijders (2001). The model is defined by a set of classes $\mathcal{C}$, a vector of class proportions $\boldsymbol{\gamma} = \{\gamma_{\mathbf{C}}\}_{\mathbf{C} \in \mathcal{C}}$ verifying $\sum_{\mathbf{C} \in \mathcal{C}} \gamma_{\mathbf{C}} = 1$, and a matrix of connection probabilities $\boldsymbol{\Pi} = (\Pi_{\mathbf{C},\mathbf{D}})_{\mathbf{C},\mathbf{D} \in \mathcal{C}^2}$. Note that they are an infinite number of ways to represent and encode the classes. For simplicity, a common choice is to set $\mathcal{C} = \{1, \ldots, Q\}$ and possibly $\mathcal{C} = \{\mathbf{C} \in \{0, 1\}^Q, \sum_{q=1}^{Q} C_q = 1\}$, for a model with Q classes. The random graphs are drawn as follows. First, the class of each vertex is sampled from a multinomial distribution with parameters $(1, \boldsymbol{\gamma})$. Thus, each vertex i belongs only to one class, and that class is $\mathbf{C}$ with probability $\gamma_{\mathbf{C}}$. Second, the edges are drawn independently from each other from Bernoulli distributions, the probability of an edge (i, j) being $\Pi_{\mathbf{C},\mathbf{D}}$, if i belongs to class $\mathbf{C}$ and j to class $\mathbf{D}$.

Let Θ_{SBM} be the parameter space of the family of SBMs with 2^Q classes:

$$\Theta_{\mathrm{SBM}} = \left\{(\boldsymbol{\gamma}, \boldsymbol{\Pi}) \in [0, 1]^{2^Q} \times [0, 1]^{2^{2Q}}, \sum_{\mathbf{C} \in \mathcal{C}} \gamma_{\mathbf{C}} = 1\right\}.$$

Considering that each possible value of the vectors $\mathbf{Z}_i$'s in an OSBM with Q classes encodes a class in a SBM with 2^Q classes (i.e., $\mathcal{C} = \{0, 1\}^Q$) yields a natural function:

$$\phi : \begin{array}{c} \Theta_{\mathrm{OSBM}} \to \Theta_{\mathrm{SBM}} \\ (\boldsymbol{\alpha}, \widetilde{\mathbf{W}}) \to (\boldsymbol{\gamma}, \boldsymbol{\Pi}) \end{array},$$

where

$$\gamma_{\mathbf{C}} = \prod_{q=1}^{Q} \alpha_q^{C_q} (1 - \alpha_q)^{1-C_q} \qquad \forall \mathbf{C} \in \{0, 1\}^Q,$$

and

$$\Pi_{\mathbf{C},\mathbf{D}} = g(\mathbf{C}^\top \mathbf{W} \mathbf{D} + \mathbf{C}^\top \mathbf{U} + \mathbf{V}^\top \mathbf{D} + W^*)$$

$$\forall (\mathbf{C}, \mathbf{D}) \in \{0, 1\}^Q \times \{0, 1\}^Q.$$

Let $\mathcal{G}_N$ denote the set of probability measures on the graphs of N vertices. The OSBM of parameter $\boldsymbol{\theta}$ in Θ_{OSBM} and the SBM of parameter $\phi(\boldsymbol{\theta})$ in Θ_{SBM} clearly induce the same measure μ in $\mathcal{G}_N$. Thus, denoting by $\psi(\boldsymbol{\gamma}, \boldsymbol{\Pi})$ the probability measure in $\mathcal{G}_N$ induced by the SBM of parameter $(\boldsymbol{\gamma}, \boldsymbol{\Pi})$, the problem of identifiability is to characterize the relations between parameters $\boldsymbol{\theta} \in \Theta_{\text{OSBM}}$ and $\boldsymbol{\theta}' \in \Theta_{\text{OSBM}}$ such that $\psi(\phi(\boldsymbol{\theta})) = \psi(\phi(\boldsymbol{\theta}'))$:

$$\Theta_{\text{OSBM}} \rightarrow \Theta_{\text{SBM}} \rightarrow \mathcal{G}_N,$$
$$\boldsymbol{\theta} = (\boldsymbol{\alpha}, \widetilde{\mathbf{W}}) \overset{\phi}{\rightarrow} (\boldsymbol{\gamma}, \boldsymbol{\Pi}) \overset{\psi}{\rightarrow} \mu.$$

The identifiability of SBM was studied by Allman, Matias and Rhodes (2009), who showed that the model is generically identifiable up to a permutation of the classes. In other words, except in a set of parameters which has a null Lebesgue measure, two parameters imply the same random graph model if and only if they differ only by the ordering of the classes. Therefore, the main theorem of Allman, Matias and Rhodes (2009) implies the following result.

THEOREM 4.1. *There exists a set $\Theta_{\text{SBM}}^{\text{bad}} \subset \Theta_{\text{SBM}}$ of null Lebesgue measure such that, for every $(\boldsymbol{\gamma}, \boldsymbol{\Pi})$ and $(\boldsymbol{\gamma}', \boldsymbol{\Pi}')$ not in $\Theta_{\text{SBM}}^{\text{bad}}$, $\psi(\boldsymbol{\gamma}, \boldsymbol{\Pi}) = \psi(\boldsymbol{\gamma}', \boldsymbol{\Pi}')$ if and only if there exists a function P_ν such that $(\boldsymbol{\gamma}', \boldsymbol{\Pi}') = P_\nu((\boldsymbol{\gamma}, \boldsymbol{\Pi}))$, where:*

- *ν is a permutation on $\{0, 1\}^Q$,*
- *$\boldsymbol{\gamma}'_{\mathbf{C}} = \boldsymbol{\gamma}_{\nu(\mathbf{C})}, \forall \mathbf{C} \in \{0, 1\}^Q$,*
- *$\boldsymbol{\Pi}'_{\mathbf{C},\mathbf{D}} = \Pi_{\nu(\mathbf{C}),\nu(\mathbf{D})}, \forall (\mathbf{C}, \mathbf{D}) \in \{0, 1\}^Q \times \{0, 1\}^Q.$*

Thus, studying the generical identifiability of the OSBM is equivalent to characterizing the parameters of Θ_{OSBM} verifying $\phi(\boldsymbol{\theta}') = P_\nu(\phi(\boldsymbol{\theta}))$ for some permutation ν on $\{0, 1\}^Q$.

4.2. *Permutations and inversions.* As in the case of the SBM, reordering the Q classes of the OSBM and doing the corresponding modification in $\boldsymbol{\alpha}$ and $\widetilde{\mathbf{W}}$ does not change the generative random graph model. Indeed, let σ be a permutation on $\{1, \ldots, Q\}$ and let P_σ denote the function corresponding to the permutation σ of the classes. Then, $(\boldsymbol{\alpha}', \widetilde{\mathbf{W}}') = P_\sigma(\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$ is defined by

$$\alpha'_q = \alpha_{\sigma(q)} \qquad \forall q \in \{1, \ldots, Q\},$$

and

$$\widetilde{\mathbf{W}}'_{q,l} = \widetilde{\mathbf{W}}_{\sigma(q),\sigma(l)} \qquad \forall (q,l) \in \{1, \ldots, Q+1\}^2.$$

Now, let ν be the permutation of $\{0, 1\}^Q$ defined by

$$\nu(\mathbf{C}) = (C_{\sigma(1)}, \ldots, C_{\sigma(Q)}) \qquad \forall \mathbf{C} \in \{0, 1\}^Q.$$

It is then straightforward to see that, for every parameter θ in Θ_{OSBM} and every permutation σ, $\phi(P_\sigma(\theta)) = P_\nu(\phi(\theta))$, where P_ν is defined in Theorem 4.1.

There is another family of operations in Θ_{OSBM} which does not change the generative random graph model, which we call inversions. They correspond to exchanging the labels 0 to 1 and vice versa on some of the coordinates of the Z_i's. To give an intuition, consider a parameter $\theta = (\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$ in Θ_{OSBM}. Let us generate graphs under the probability measure in $\mathcal{G}_N$ induced by θ and consider only the first coordinate of the Z_i's. If we denote by "cluster 1" the vertices whose Z_i's have a 1 as first coordinate, the graph sampling procedure consists in sampling the set "cluster 1" and then drawing the edges conditionally on that information. Note that it would be equivalent to sample the vertices which are not in "cluster 1" and to draw the edges conditionally on that information. Thus, there exists an equivalent reparametrization where the 1's in the first coordinate correspond to the vertices which are not in "cluster 1." This is the parameter θ' obtained from θ by an inversion of the first coordinate.

Let $\mathbf{A}$ be any vector of $\{0, 1\}^Q$. We define the *A-inversion* $I_{\mathbf{A}}$ as follows:

$$I_{\mathbf{A}} : \begin{array}{c} \Theta_{\mathrm{OSBM}} \to \Theta_{\mathrm{OSBM}} \\ (\boldsymbol{\alpha}, \widetilde{\mathbf{W}}) \to (\boldsymbol{\alpha}', \widetilde{\mathbf{W}}') \end{array},$$

where

$$\alpha'_j = \begin{cases} 1 - \alpha_j, & \text{if } A_j = 1, \\ \alpha_j, & \text{otherwise} \end{cases} \qquad \forall j \in \{1, \ldots, Q\},$$

and

$$\widetilde{\mathbf{W}}' = \mathbf{M}_{\mathbf{A}}^\top \widetilde{\mathbf{W}} \mathbf{M}_{\mathbf{A}}.$$

The matrix $\mathbf{M}_{\mathbf{A}}$ is defined by

$$\mathbf{M}_{\mathbf{A}} = \begin{pmatrix} I - 2\,\mathrm{diag}(\mathbf{A}) & \mathbf{A} \\ 0 \quad \cdots \quad 0 & 1 \end{pmatrix},$$

with $\mathrm{diag}(\mathbf{A})$ being the $Q \times Q$ diagonal matrix whose diagonal is the vector $\mathbf{A}$.

PROPOSITION 4.1. *For every* $\mathbf{A} \in \{0, 1\}^Q$, *let* ν *be the permutation of* $\{0, 1\}^Q$ *defined by*

$$\forall \mathbf{C} \in \{0, 1\}^Q \qquad \nu(\mathbf{C})_i = \begin{cases} 1 - C_i, & \text{if } A_i = 1, \\ C_i, & \text{otherwise.} \end{cases}$$

Then, for every $\boldsymbol{\theta}$ in Θ_{OSBM},

$$\phi(I_{\mathbf{A}}(\boldsymbol{\theta})) = P_\nu(\phi(\boldsymbol{\theta})),$$

where P_ν is defined in Theorem 4.1.

PROOF. Consider $\boldsymbol{\theta} \in \Theta_{\mathrm{OSBM}}$ and $\mathbf{A} \in \{0, 1\}^Q$ and define $(\boldsymbol{\gamma}, \boldsymbol{\Pi}) = \phi(\boldsymbol{\theta})$ and $(\boldsymbol{\gamma}', \boldsymbol{\Pi}') = \phi(I_{\mathbf{A}}(\boldsymbol{\theta}))$. It is straightforward to verify that

$$\boldsymbol{\gamma}'_{\mathbf{C}} = \boldsymbol{\gamma}_{\nu(\mathbf{C})} \qquad \forall \mathbf{C} \in \{0, 1\}^Q.$$

Moreover, since $M_{\mathbf{A}} \begin{pmatrix} \mathbf{C} \\ 1 \end{pmatrix} = \begin{pmatrix} \nu(\mathbf{C}) \\ 1 \end{pmatrix}$, it follows that

$$\begin{aligned}
\boldsymbol{\Pi}'_{\mathbf{C},\mathbf{D}} &= g\left((\mathbf{C}^\top \quad 1) \, M_{\mathbf{A}}^\top \widetilde{\mathbf{W}} M_{\mathbf{A}} \begin{pmatrix} \mathbf{D} \\ 1 \end{pmatrix} \right) \\
&= g\left((\nu(\mathbf{C})^\top \quad 1) \, \widetilde{\mathbf{W}} \begin{pmatrix} \nu(\mathbf{D}) \\ 1 \end{pmatrix} \right) \\
&= \boldsymbol{\Pi}_{\nu(\mathbf{C}),\nu(\mathbf{D})}.
\end{aligned}$$

Therefore, $\phi(I_A(\boldsymbol{\theta})) = P_\nu(\phi(\boldsymbol{\theta}))$. $\square$

4.3. *Identifiability.* Let us define the following equivalence relation:

$$\boldsymbol{\theta} \sim \boldsymbol{\theta}' \qquad \text{if } \exists \sigma, \mathbf{A} | \boldsymbol{\theta}' = I_{\mathbf{A}}(P_\sigma(\boldsymbol{\theta})).$$

To be convinced that it is an equivalence relation, note that

$$I_{\mathbf{A}} \circ P_\sigma = P_\sigma \circ I_{\sigma^{-1}(\mathbf{A})}.$$

Consider the set of equivalence classes for the relation $\sim$. It follows that:

- Two parameters in the same equivalence class induce the same measure in $\mathcal{G}_N$.
- Each equivalence class contains a parameter $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$ such that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_Q \leq \frac{1}{2}$. Moreover, if the α_i's are all distinct and strictly lower than $\frac{1}{2}$, there is a unique such parameter in the equivalence class.

We are now able to state our main theorem about identifiability, that is, that the model is generically identifiable up to the equivalence relation $\sim$.

THEOREM 4.2. *For every $\boldsymbol{\alpha} \in]0, 1[^Q$, let $\beta \in \mathbb{R}^Q$ be the vector defined by $\beta_k = -\ln(\frac{\alpha_k}{1-\alpha_k})$, for every k.*

Define $\Theta_{\mathrm{OSBM}}^{\mathrm{bad}}$ as the set of parameters $(\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$ such that one of the following conditions holds:

- *there exists $1 \leq k \leq Q$ such that $\alpha_k = 0$ or $\alpha_k = 1$ or $\alpha_k = \frac{1}{2}$,*
- *there exist $1 \leq k, l \leq Q$ such that $\alpha_k = \alpha_l$,*
- *there exist $\mathbf{C}, \mathbf{D} \in \{0, 1\}^Q \times \{0, 1\}^Q$ such that $\sum_k \beta_k C_k = \sum_k \beta_k D_k$,*

• $\phi(\boldsymbol{\alpha}, \widetilde{\mathbf{W}}) \in \Theta_{\mathrm{SBM}}^{\mathrm{bad}}$, *set of null measure given by Theorem* 4.1.

Then $\Theta_{\mathrm{OSBM}}^{\mathrm{bad}}$ *has a null Lebesgue measure on* Θ_{OSBM} *and*

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in (\Theta_{\mathrm{OSBM}} \setminus \Theta_{\mathrm{OSBM}}^{\mathrm{bad}})^2 \qquad \phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}') \iff \boldsymbol{\theta} \sim \boldsymbol{\theta}'.$$

PROOF. $\Theta_{\mathrm{OSBM}}^{\mathrm{bad}}$ is the union of a finite number of hyperplanes or spaces which are isomorphic to hyperplanes. Therefore, $\mu(\Theta_{\mathrm{OSBM}}^{\mathrm{bad}}) = 0$.   □

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$, $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \widetilde{\mathbf{W}}')$, $\phi(\boldsymbol{\theta}) = (\boldsymbol{\gamma}, \boldsymbol{\Pi})$ and $\phi(\boldsymbol{\theta}') = (\boldsymbol{\gamma}', \boldsymbol{\Pi}')$. As ϕ is constant on each equivalence class and as $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are not in $\Theta_{\mathrm{OSBM}}^{\mathrm{bad}}$, we can assume that $0 < \alpha_1 < \cdots < \alpha_k < \frac{1}{2}$ and $0 < \alpha_1' < \cdots < \alpha_k' < \frac{1}{2}$. Proving the theorem is then equivalent to proving that $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

As $\phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}')$, Theorem 4.1 ensures that there exists a permutation $\nu : \{0, 1\}^Q \to \{0, 1\}^Q$ such that

$$\begin{cases} \gamma_{\mathbf{C}}' = \gamma_{\nu(\mathbf{C})} & \forall \mathbf{C}, \\ \Pi_{\mathbf{C},\mathbf{D}}' = \Pi_{\psi(\mathbf{C}),\psi(\mathbf{D})} & \forall \mathbf{C}, \mathbf{D}. \end{cases}$$

Then, in particular,

$$(4.1) \qquad \begin{aligned} & \left\{ \prod_k \alpha_k^{C_k} (1 - \alpha_k)^{1-C_k}, \mathbf{C} \in \{0, 1\}^Q \right\} \\ & = \left\{ \prod_k (\alpha_k')^{C_k} (1 - \alpha_k')^{1-C_k}, \mathbf{C} \in \{0, 1\}^Q \right\}. \end{aligned}$$

The minima of those two sets as well as the second lowest values are equal, that is,

$$\prod_k \alpha_k = \prod_k \alpha_k' \quad \text{and} \quad \left(\prod_{k \le Q-1} \alpha_k \right)(1 - \alpha_Q) = \left(\prod_{k \le Q-1} \alpha_k' \right)(1 - \alpha_Q').$$

Dividing those equations term by term yields $\frac{\alpha_Q}{1-\alpha_Q} = \frac{\alpha_Q'}{1-\alpha_Q'}$ and finally $\alpha_Q = \alpha_Q'$. Dividing all terms by $\alpha_Q^{C_Q}(1 - \alpha_Q)^{1-C_Q}$ in (4.1), by induction, it follows that

$$(4.2) \qquad \boldsymbol{\alpha} = \boldsymbol{\alpha}'.$$

Now, for any $\mathbf{C} \in \{0, 1\}^Q$, the fact that $\gamma_{\mathbf{C}}' = \gamma_{\nu(\mathbf{C})}$ can be written as

$$\prod_k \alpha_k^{C_k} (1 - \alpha_k)^{1-C_k} = \prod_k \alpha_k^{\nu(C)_k} (1 - \alpha_k)^{1-\nu(C)_k},$$

$$\sum_k C_k \ln\left(\frac{\alpha_k}{1 - \alpha_k} \right) + \sum_k \ln(1 - \alpha_k) = \sum_k \nu(C_k) \ln\left(\frac{\alpha_k}{1 - \alpha_k} \right) + \sum_k \ln(1 - \alpha_k),$$

$$\sum_k \beta_k C_k = \sum_k \beta_k \nu(C)_k.$$

Since $\boldsymbol{\theta} \notin \Theta_{\text{OSBM}}^{\text{bad}}$, this implies that $\nu(\mathbf{C}) = \mathbf{C}$. As it is true for every $\mathbf{C}$, ν is in fact the identity function.

Therefore, for every $\mathbf{C}, \mathbf{D}$, $\Pi_{\mathbf{C},\mathbf{D}} = \Pi'_{\mathbf{C},\mathbf{D}}$, that is,

$$\sum_{q,l} w_{ql} c_q d_l + \sum_q u_q c_q + \sum_l v_l d_l + w^* = \sum_{q,l} w'_{ql} c_q d_l + \sum_q u'_q c_q + \sum_l v'_l d_l + w'^*.$$

Applying it for $\mathbf{C} = \mathbf{D} = 0$ implies $W^* = W'^*$.

Applying it for $\mathbf{D} = 0$ and $\mathbf{C} = \boldsymbol{\delta}_q$, where $\boldsymbol{\delta}_q$ is the vector having a 1 on the qth coordinate and 0's elsewhere yields $u_q + W^* = u'_q + W'^*$ and, thus, $u_q = u'_q$.

By symmetry, $\mathbf{C} = 0$ and $\mathbf{D} = \boldsymbol{\delta}_l$ implies $v_l = v'_l$.

Finally, $\mathbf{C} = \boldsymbol{\delta}_q$ and $\mathbf{D} = \boldsymbol{\delta}_l$ gives $W_{ql} = W'_{ql}$.

Thus,

$$(4.3) \qquad\qquad\qquad \widetilde{\mathbf{W}} = \widetilde{\mathbf{W}}'.$$

By equations (4.2) and (4.3), we have $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

5. Statistical inference. Given a network, our aim in this section is to estimate the OSBM parameters.

The log-likelihood of the observed data set is defined through the marginalization: $p(\mathbf{X}|\boldsymbol{\alpha}, \widetilde{\mathbf{W}}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$. This summation involves 2^{NQ} terms and quickly becomes intractable. To tackle this issue, the Expectation–Maximization (EM) algorithm has been applied on many mixture models. However, the E-step requires the calculation of the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \widetilde{\mathbf{W}})$ which cannot be factorized in the case of networks [see Daudin, Picard and Robin (2008) for more details]. In order to obtain a tractable procedure, we present some approximations based on global and local variational techniques.

5.1. *The q-transformation.* Given a distribution $q(\mathbf{Z})$, the log-likelihood of the observed data set can be decomposed using the Kullback–Leibler divergence $\text{KL}(\cdot\|\cdot)$:

$$(5.1) \qquad \ln p(\mathbf{X}|\boldsymbol{\alpha}, \widetilde{\mathbf{W}}) = \mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}) + \text{KL}(q(\cdot)\|p(\cdot|\mathbf{X}, \boldsymbol{\alpha}, \widetilde{\mathbf{W}})),$$

where

$$(5.2) \qquad \mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\alpha}, \widetilde{\mathbf{W}})}{q(\mathbf{Z})} \right\}$$

and

$$(5.3) \qquad \text{KL}(q(\cdot)\|p(\cdot|\mathbf{X}, \boldsymbol{\alpha}, \widetilde{\mathbf{W}})) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \widetilde{\mathbf{W}})}{q(\mathbf{Z})} \right\}.$$

The maximum $\ln p(\mathbf{X}|\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$ of the lower bound $\mathcal{L}$ (5.2) is reached when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \widetilde{\mathbf{W}})$. Thus, if the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \widetilde{\mathbf{W}})$ was tractable, the

optimizations of $\mathcal{L}$ and $\ln p(\mathbf{X}|\boldsymbol{\alpha}, \widetilde{\mathbf{W}})$, with respect to $\boldsymbol{\alpha}$ and $\widetilde{\mathbf{W}}$, would be equivalent. However, in the case of networks, $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\alpha}, \widetilde{\mathbf{W}})$ cannot be calculated and $\mathcal{L}$ cannot be optimized over the entire space of $q(\mathbf{Z})$ distributions. Thus, we restrict our search to the class of distributions which satisfy

$$(5.4) \qquad q(\mathbf{Z}) = \prod_{i=1}^{N} q(\mathbf{Z}_i),$$

with

$$q(\mathbf{Z}_i) = \prod_{q=1}^{Q} \mathcal{B}(Z_{iq}; \tau_{iq}) = \prod_{q=1}^{Q} \tau_{iq}^{Z_{iq}} (1 - \tau_{iq})^{1-Z_{iq}}.$$

Each τ_{iq} is a variational parameter which corresponds to the posterior probability of node i to belong to class q. As for the vector $\boldsymbol{\alpha}$, the vectors $\boldsymbol{\tau}_i = \{\tau_{i1}, \ldots, \tau_{iQ}\}$ are not constrained to lie in the $(Q-1)$-dimensional simplex.

PROPOSITION 5.1. [*Proof in* Latouche, Birmelé and Ambroise (2010), *Appendix* A]. *The lower bound of the observed data log-likelihood is given by*

$$\mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}) = \sum_{i \neq j}^{N} \{X_{ij} \tilde{\boldsymbol{\tau}}_i^\top \widetilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]\}$$

$$(5.5) \qquad\qquad + \sum_{i=1}^{N} \sum_{q=1}^{Q} \{\tau_{iq} \ln \alpha_q + (1 - \tau_{iq}) \ln(1 - \alpha_q)\}$$

$$- \sum_{i=1}^{N} \sum_{q=1}^{Q} \{\tau_{iq} \ln \tau_{iq} + (1 - \tau_{iq}) \ln(1 - \tau_{iq})\}.$$

Unfortunately, since the logistic sigmoid function is nonlinear, $\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]$ in (5.5) cannot be computed analytically. Thus, we need a second level of approximation to optimize the lower bound of the observed data set.

5.2. ξ-transformation.

PROPOSITION 5.2 [Proof in Latouche, Birmelé and Ambroise (2010) in Appendix A]. *Given a variational parameter* ξ_{ij}, $\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]$ *satisfies*

$$(5.6) \qquad \begin{aligned} &\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})] \\ &\qquad \geq \ln g(\xi_{ij}) - \frac{(\tilde{\boldsymbol{\tau}}_i^\top \widetilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \xi_{ij})}{2} - \lambda(\xi_{ij})\big(\mathrm{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[(\tilde{\mathbf{Z}}_i^\top \widetilde{\mathbf{W}} \tilde{\mathbf{Z}}_j)^2] - \xi_{ij}^2\big). \end{aligned}$$

Eventually, a lower bound of the first lower bound can be computed:

$$(5.7) \qquad \ln p(\mathbf{X}|\boldsymbol{\alpha}, \widetilde{\mathbf{W}}) \geq \mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}) \geq \mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}, \boldsymbol{\xi}),$$

where

$$
\begin{aligned}
\mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}, \boldsymbol{\xi}) = \sum_{i \neq j}^{N} &\left\{ \left(X_{ij} - \frac{1}{2} \right) \tilde{\boldsymbol{\tau}}_i^\top \widetilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \ln \mathrm{g}(\xi_{ij}) - \frac{\xi_{ij}}{2} \right. \\
&\left. - \lambda(\xi_{ij}) \big(\mathrm{Tr}(\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{E}}_i \widetilde{\mathbf{W}} \boldsymbol{\Sigma}_j) + \tilde{\boldsymbol{\tau}}_j^\top \widetilde{\mathbf{W}}^\top \widetilde{\mathbf{E}}_i \widetilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j - \xi_{ij}^2 \big) \right\} \\
&+ \sum_{i=1}^{N} \sum_{q=1}^{Q} \{ \tau_{iq} \ln \alpha_q + (1 - \tau_{iq}) \ln(1 - \alpha_q) \} \\
&- \sum_{i=1}^{N} \sum_{q=1}^{Q} \{ \tau_{iq} \ln \tau_{iq} + (1 - \tau_{iq}) \ln(1 - \tau_{iq}) \}.
\end{aligned}
$$

The resulting variational EM algorithm (see Algorithm 1) alternatively computes the posterior probabilities $\boldsymbol{\tau}_i$ and the parameters $\boldsymbol{\alpha}$ and $\widetilde{\mathbf{W}}$ maximizing

$$\max_{\boldsymbol{\xi}} \mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}, \boldsymbol{\xi}).$$

The optimization equations are given in Latouche, Birmelé and Ambroise (2010), Appendix B.

The computational cost of the algorithm is equal to $O(N^2 Q^4)$. For comparison the computational cost of the methods proposed by Daudin, Picard and Robin (2008) and Latouche, Birmelé and Ambroise (2009) for (nonoverlapping) SBM is equal to $O(N^2 Q^2)$. Analyzing a sparse network with 100 nodes takes about ten seconds on a dual core, and about a minute for dense networks.

For all the experiments we present in the following section, set $\sigma^2 = 0.5$ and we used the Ascendant Hierarchical Classification (AHC) algorithm implemented in the R package "mixer" which is available at the following: http://cran.r-project.org/web/packages/mixer.

6. Experiments. We present some results of the experiments we carried out to assess OSBM. Throughout our experiments, we compared our approach to SBM (the nonoverlapping version of OSBM), the Mixed Membership Stochastic Block model (MMSB) of Airoldi et al. (2008), and the work of Palla et al. (2005), implemented in the software (Version 2.0.1) CFinder [Palla et al. (2006)].

In order to perform inference in SBM, we used the variational Bayes algorithm of Latouche, Birmelé and Ambroise (2009) which approximates the posterior distribution over the latent variables and model parameters, given the edges. We computed the Maximum A Posteriori (MAP) estimates and obtained the class membership vectors $\mathbf{Z}_i$. We recall that SBM assumes that each vertex belongs to a single

Algorithm 1 Overlapping stochastic block model for directed graphs without self loop.

// INITIALIZATION

Initialize $\boldsymbol{\tau}$ with an Ascendant Hierarchical Classification algorithm Sample $\widetilde{\mathbf{W}}$ from a zero mean σ^2 spherical Gaussian distribution

// OPTIMIZATION

repeat

    // ξ-transformation

    for $(i, j) \in V$ **do**

        $\xi_{ij} \leftarrow \sqrt{\operatorname{Tr}(\widetilde{\mathbf{W}}^\top \widetilde{\mathbf{E}}_i \widetilde{\mathbf{W}} \boldsymbol{\Sigma}_j) + \tilde{\boldsymbol{\tau}}_j^\top \widetilde{\mathbf{W}}^\top \widetilde{\mathbf{E}}_i \widetilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j}$

    end

    // M-step

    for $q = 1 : Q$ **do**

        $\alpha_q \leftarrow \frac{\sum_{i=1}^N \tau_{iq}}{N}$

    end

    Optimize $\mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}, \boldsymbol{\xi})$ with respect to $\widetilde{\mathbf{W}}$, with a gradient based optimization algorithm [e.g., quasi-Newton method of Broyden et al. (1970)]

    // E-step

    repeat

        for $i = 1 : N$ **do**

            Optimize $\mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}, \boldsymbol{\xi})$ with respect to $\boldsymbol{\tau}_i$, with a box constrained ($\tau_{iq} \in [0, 1]$) gradient based optimization algorithm [e.g., Byrd method Byrd et al. (1995)]

        end

    until $\boldsymbol{\tau}$ *converges*

until $\mathcal{L}(q; \boldsymbol{\alpha}, \widetilde{\mathbf{W}}, \boldsymbol{\xi})$ *converges*

---

class and, therefore, each vector $\mathbf{Z}_i$ has all its components set to zero except one, such that $Z_{iq} = 1$ if vertex i is classified into class q. For OSBM, we relied on the variational approximate inference procedure described in Section 5 and computed the MAP estimates. Contrary to SBM, each vertex can belong to multiple clusters and, therefore, the vectors $\mathbf{Z}_i$ can have multiple components set to one. As described in Section 1, MMSB can also be viewed as allowing overlapping clusters. For more details, we refer to Airoldi et al. (2008). In order to estimate the MMSB mixing weight vectors $\boldsymbol{\pi}_i$, we used the collapsed Gibbs sampling approach implemented in the R package lda [Chang (2010)]. We then converted each vector $\boldsymbol{\pi}_i$ into a binary membership vector $\mathbf{Z}_i$ using a threshold t. Thus, for $\pi_{iq} \geq t$, we set $Z_{iq} = 1$ and $Z_{iq} = 0$ otherwise. In all the experiments we carried out, we defined $t = 1/Q$ and we found that for higher values MMSB tended to behave like SBM. Finally, we considered CFinder which is a widely used algorithmic approach to uncover overlapping communities. As described in Section 1, CFinder looks for k-clique communities where each k-clique community is a union of all k-cliques (complete sub-graphs of size k) that can be reached from each other through a se-

ries of adjacent k-cliques. The algorithm first locates all cliques and then identifies the communities and overlaps between communities using a clique–clique overlap matrix [Everett and Borgatti (1998)]. Vertices that do not belong to any k-clique are seen as outliers and not classified.

Contrary to OSBM (and CFinder), SBM and MMSB cannot deal with outliers. Therefore, to obtain fair comparisons between the approaches, when OSBM was run with Q classes, SBM and MMSB were run with $Q + 1$ classes and we identified the class of outliers. In practice, this can easily be done since this class contains most of the vertices of the network having low output and input degrees.

The code implementing all the experiments is available upon request.

6.1. *Simulations.* In this set of experiments we generated two types of networks using the OSBM generative model. In Section 6.1.1 we sampled networks with community structures (Figure 3), where vertices of a community are mostly connected to vertices of the same community. To limit the number of free parameters, we considered the $Q \times Q$ real matrix $\mathbf{W}$:

$$(6.1) \qquad \mathbf{W} = \begin{pmatrix} \lambda & -\varepsilon & \ldots & -\varepsilon \\ -\varepsilon & \lambda & & \vdots \\ \vdots & & \ddots & -\varepsilon \\ -\varepsilon & \ldots & -\varepsilon & \lambda \end{pmatrix}.$$
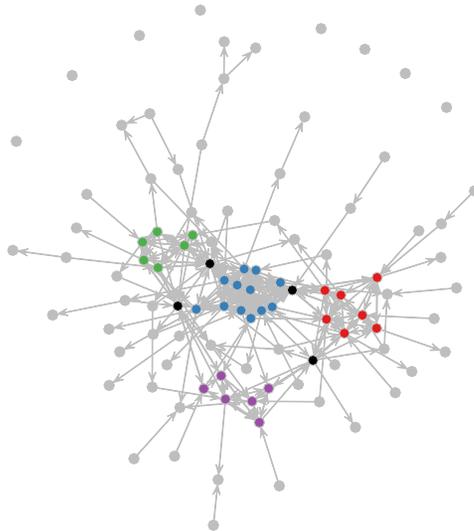

FIG. 3. *Example of a network with community structures. Overlaps are represented in black and outliers in gray.*
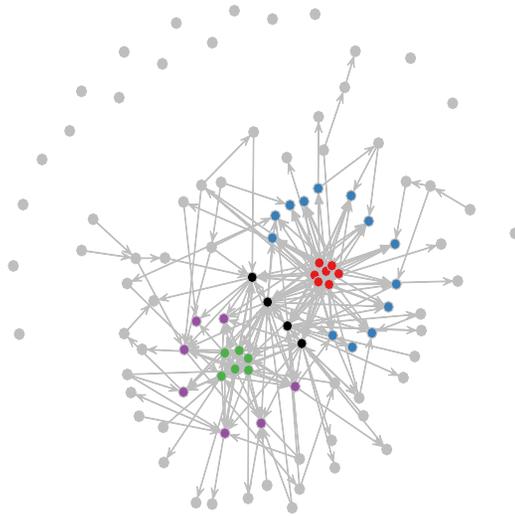
FIG. 4.   *Example of a network with community structures and stars. Overlaps are represented in black and outliers in gray.*

In Section 6.1.2 we generated networks with more complex topologies, using the matrix $\mathbf{W}$:

$$(6.2) \qquad \mathbf{W} = \begin{pmatrix} \lambda & \lambda & -\varepsilon & \dots & \dots & \dots & -\varepsilon \\ -\boldsymbol{\varepsilon} & -\lambda & -\varepsilon & \dots & \dots & \dots & \vdots \\ \vdots & -\varepsilon & \lambda & \lambda & -\varepsilon & \dots & \vdots \\ \vdots & \vdots & -\boldsymbol{\varepsilon} & -\lambda & -\varepsilon & \dots & \vdots \\ \vdots & \vdots & \vdots & -\varepsilon & \ddots & -\varepsilon & -\varepsilon \\ \vdots & \vdots & \vdots & \vdots & -\varepsilon & \lambda & \lambda \\ -\varepsilon & \dots & \dots & \dots & \dots & -\boldsymbol{\varepsilon} & -\lambda \end{pmatrix}.$$

In these networks, if class i is a community and has therefore a high *intra* connection probability, then its vertices also highly connect to vertices of class $i + 1$ which itself has a low *intra* connection probability. Such star patterns (Figure 4) often appear in transcription networks, as shown in Section 6.3, and protein–protein interaction networks.

For these two sets of experiments, we used the Q-dimensional real vectors $\mathbf{U}$ and $\mathbf{V}$:

$$(6.3) \qquad \mathbf{U} = \mathbf{V} = (\varepsilon \quad \dots \quad \varepsilon),$$

and we set $Q = 4$, $\lambda = 4$, $\varepsilon = 1$, $W^* = -5.5$. Moreover, for the vector $\boldsymbol{\alpha}$ of class probabilities, we set $\alpha_q = 0.25, \forall q \in \{1, \dots, Q\}$. We generated 100 networks with $N = 100$ vertices and for each of these networks, we clustered the vertices using CFinder, SBM, MMSB and OSBM. Finally, we used a criterion similar to the

one proposed by Heller and Ghahramani (2007); Heller, Williamson and Ghahramani (2008) to compare the true $\mathbf{Z}$ and the estimated $\hat{\mathbf{Z}}$ clustering matrices. Thus, for each network and each method, we computed the L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ where $\mathbf{P} = \mathbf{Z}\mathbf{Z}^\top$ and $\hat{\mathbf{P}} = \hat{\mathbf{Z}}\hat{\mathbf{Z}}^\top$. These two $N \times N$ matrices are invariant to column permutations of $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ and compute the number of shared clusters between each pair of vertices of a network. Therefore, $d(\mathbf{P}, \hat{\mathbf{P}})$ is a good measure to determine how well the underlying cluster assignment structure has been discovered. Since CFinder depends on a parameter k (size of the cliques), for each simulated network, we ran the software for various values of k and selected $\hat{k}$ for which the L_2 distance was minimized. Note that this choice of k tends to overestimate the performances of CFinder compared to the other approaches. Indeed, in practice, when analyzing a real network, k needs to be estimated (see Section 6.2), while $\mathbf{P}$ is unknown. OSBM was run with Q classes, whereas SBM and MMSB were run with $Q + 1$ classes. For both SBM and MMSB, and each generated network, after having identified the class of outliers, we set the latent vectors of the corresponding vertices to zero (null component). The L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ was then computed exactly as described previously.

6.1.1. *Networks with community structures.* The results that we obtained are presented in Table 2 and in Figure 5. We can observe that CFinder, MMSB and OSBM lead to very accurate estimates $\hat{\mathbf{Z}}$ of the true clustering matrix $\mathbf{Z}$. For most networks, they retrieve the clusters and overlaps perfectly, although CFinder and MMSB appear to be slightly biased. Indeed, while the median of the L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ over the 100 samples is null for OSBM, it is equal to 22 for CFinder and 27.5 for MMSB. Since CFinder is an algorithmic approach, and not a probabilistic model, it does not classify a vertex v_i if it does not belong to any k-cliques of a k-clique community. Conversely, OSBM is more flexible and can take the random nature of the network into account. Indeed, the edges are assumed to be drawn randomly, and, given each pair of vertices, OSBM deciphers whether or not they are likely to belong to the same class, depending on their connection profiles. There-

TABLE 2
Comparison of CFinder, SBM, MMSB and OSBM in terms of the L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ over the 100 samples of networks with community structures

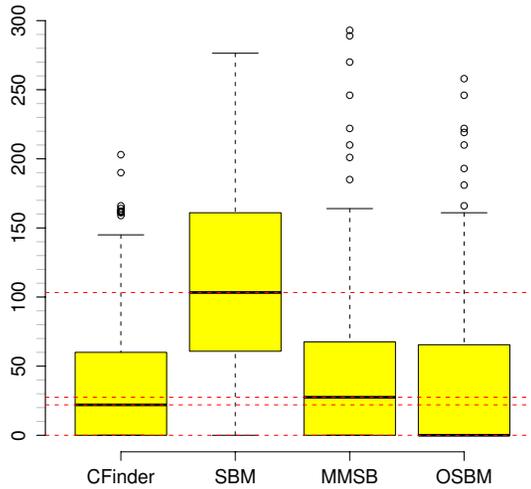|	Mean	Median	Min	Max
CFinder	43.53	22	**0**	**203**
SBM	116.46	103.3	**0**	321
MMSB	53.76	27.5	**0**	293
OSBM	**41.83**	**0**	**0**	258

FIG. 5. L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ over the 100 samples of networks with community structures, for CFinder, SBM, MMSB and OSBM. Measures how well the underlying cluster assignment structure has been retrieved.

fore, OSBM can predict that v_i belongs to a class q, although it does not belong to any k-cliques. Overall, we found that MMSB retrieves the clusters well but often misclassifies some of the overlaps. Thus, if a given vertex belongs to several clusters, it tends to be classified by MMSB into only one of them. Nevertheless, the results clearly illustrate that MMSB improves over SBM, which cannot retrieve any of the overlapping clusters. It should also be noted that CFinder has fewer outliers (Figure 5) than MMSB and OSBM and appears to be slightly more stable when looking for overlapping community structures in networks.

6.1.2. *Networks with community structures and stars.* In this set of experiments we considered networks with more complex topologies. As shown, in Table 3 and in Figure 6, the results of CFinder dramatically degrade while those of OSBM remain more stable. Indeed, the median of the L_2 distances $d(\mathbf{P}, \hat{\mathbf{P}})$ over the 100 samples is equal to 43 for OSBM, while it is equal to 354.5 for CFinder. This can be easily explained since CFinder only looks for community structures of adjacent k-cliques, and cannot retrieve classes with low *intra* connection probabilities. Conversely, OSBM uses a $Q \times Q$ real matrix $\mathbf{W}$ and two real vectors $\mathbf{U}$ and $\mathbf{V}$ of size Q to model the *intra* and *inter* connection probabilities. No assumption is made on these matrix and vectors such that OSBM can take heterogeneous and complex topologies into account. As for CFinder, the results of MMSB degrade, although they remain better than SBM. As for the previous Section, MMSB retrieves the clusters well but misclassifies the overlaps more frequently when considering networks with community structures and stars.

TABLE 3
Comparison of CFinder, SBM, MMSB and OSBM in terms of the L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ over the 100 samples of networks with community structures and stars

	Mean	Median	Min	Max
CFinder	362.07	354.5	181	567
SBM	134.68	118.87	15.14	352.09
MMSB	119.01	98.5	**0**	367
OSBM	**77**	**43**	**0**	**328**

6.2. *French political blogosphere.* We consider the French political blogosphere network and we focus on a subset of 196 vertices connected by 2864 edges. The data consists of a single day snapshot of political blogs automatically extracted on the 14th of October 2006 and manually classified by the "Observatoire Présidentielle project" [Zanghi, Ambroise and Miele (2008)]. Nodes correspond to hostnames and there is an edge between two nodes if there is a known hyperlink from one hostname to another. The four main political parties which are present in the data set are the UMP (french "republican"), UDF ("moderate" party), liberal party (supporters of economic-liberalism) and PS (french "democrat"). Therefore, we applied our algorithm with $Q = 4$ clusters and we obtained the results presented in Figure 7 and Table 4.

First, we notice that the clusters we found are highly homogeneous and correspond to the well-known political parties. Thus, cluster 1 contains 35 blogs among
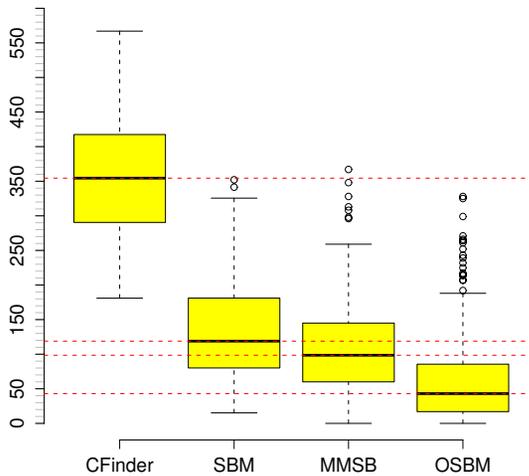

FIG. 6. *L_2 distance $d(\mathbf{P}, \hat{\mathbf{P}})$ over the 100 samples of networks with community structures and stars, for CFinder, SBM, MMSB and OSBM. Measures how well the underlying cluster assignment structure has been retrieved.*

	UMP	UDF	liberal	PS	analysts	others
cluster 1	30 + 3	0 + 1	0	0	0 + 1	0
cluster 2	2 + 3	29 + 1	0	0	1 + 3	0
cluster 3	0	0	24	0	1 + 1	0
cluster 4	0	0 + 2	0	40	0 + 4	1
outliers	5	1	1	17	5	30

FIG. 7.   *Classification of the blogs into $Q = 4$ clusters using OSBM. The entry (i, j) of the matrix describes the number of blogs associated to the jth political party (column) and classified into cluster i (row). Each entry distinguishes blogs which belong to a unique cluster from overlaps (single membership blogs + overlaps). The last row corresponds to the null component.*

which 33 are associated to UMP, while cluster 2 contains 39 blogs among which 30 are related to UDF. Similarly, it follows that cluster 3 corresponds to the liberal party and cluster 4 to PS. We found nine overlaps. Thus, three blogs associated to UMP belong to both clusters 1 (UMP) and 2 (UDF). This is a result we expected since these two political parties are known to have some relational ties. Moreover,

TABLE 4
The estimated $\widetilde{\mathbf{W}}$ matrix for the classification of the blogs into $Q = 4$ clusters using OSBM. The 4×4 matrix on the top left-hand side represents the $\mathbf{W}$ matrix, while the vectors on the top right-hand side and bottom left-hand side represent the vectors $\mathbf{U}$ and $\mathbf{V}^{\top}$ respectively. The remaining term corresponds to the bias. The diagonal of $\mathbf{W}$ indicates that blogs have a heavy tendency to connect to blogs of the same class. Blogs of cluster 1 (UMP) have also a positive tendency to connect to blogs of clusters 2 (UDF) and 3 (liberal party). Conversely, blogs of cluster 4 (PS), representing the left wing, are more isolated in the network

3.89	0.17	0.54	−0.70	−0.70
0.17	2.47	−0.40	−0.84	0.40
0.55	−0.40	4.43	−0.85	−0.38
−0.70	−0.84	−0.85	1.66	0.87
−0.70	0.40	−0.38	0.87	−3.60

a blog associated to UDF belongs to both clusters 1 (UMP) and 4 (PS), while another UDF blog belongs to clusters 2 (UDF) and 4 (PS). This can be easily understood since UDF is a moderate party. Therefore, it is not surprising to find UDF blogs with links with the two biggest political parties in France, representing the left and right wings. Very interestingly, among the nine overlaps we found, four of them correspond to blogs of political analysts. Thus, a blog overlaps clusters 1 (UMP) and 4 (PS). Another one overlaps clusters 2 (UDF), 3 (liberal party) and 4 (PS). Finally, the two last blogs of political analysts overlap clusters 2 (UDF) and 4 (PS).

We ran CFinder and we used the criterion [Palla et al. (2005)] they proposed to select k (see Section 1). Thus, we ran the software for various values of k and we found $\hat{k} = 7$. Lower values lead to giant components which smear the details of the network. Conversely, for higher values, the communities start disintegrating. Using $\hat{k}$, we uncovered 11 clusters which correspond to sub-clusters of the clusters we found using OSBM. For instance, cluster 3 (liberal party) was split into two clusters, whereas cluster 4 (PS) was split into three. Indeed, while OSBM predicted that the connection profiles of these sub-clusters were very similar and therefore should be merged, CFinder could not uncover any k-clique community, that is, a union of *fully* connected sub-graphs of size k, containing these sub-clusters. Note that using CFinder, we retrieved the overlaps uncovered by our algorithm. CFinder did not classify 95 blogs.

We also clustered the blogs of the network using MMSB and SBM. As previously, for both models, we used $Q + 1$ clusters and we identified the class of outliers. The results of MMSB are presented in Figure 8. Overall, we can notice that MMSB led to similar clusters as OSBM, although cluster 4 is less homogeneous in MMSB than in OSBM. We found eight overlaps using MMSB and we emphasize that five of them correspond exactly to the one found with our approach. Thus, the model retrieved two among the three UMP blogs overlapping clusters 1 (UMP) and 2 (UDF). Moreover, MMSB uncovered the UDF blog overlapping clusters 1 (UMP) and 4 (PS), as well as the blog of political analysts overlapping clusters 2 (UDF), 3 (liberal party) and 4 (PS). It also retrieved the blog of political analysts overlapping clusters 1 (UMP) and 4 (PS). Finally, the results of SBM are presented in Figure 9. Again, the clusters found by this approach are very similar to the one uncovered by OSBM. However, because SBM does not allow each vertex to belong to multiple clusters, it misses a lot of information in the network. In particular, while some of the blogs of political analysts are viewed as overlaps by OSBM, because of their relational ties with the different political parties, they are all classified into a single cluster by SBM.

6.3. *Saccharomyces cerevisiae transcription network.* We consider the yeast transcriptional regulatory network described in Milo et al. (2002) and we focus on a subset of 197 vertices connected by 303 edges. Nodes of the network correspond to operons, and two operons are linked if one operon encodes a transcriptional

	UMP	UDF	liberal	PS	analysts	others
cluster 1	27 + 2	0 + 2	0	0	1 + 1	0
cluster 2	2 + 2	29 + 1	0	0 + 1	3 + 2	0
cluster 3	0	0	25	0	1 + 2	0
cluster 4	0	0 + 1	0	30 + 1	0 + 2	1
cluster 5	9	1	0	26	3	30

FIG. 8. *Classification of the blogs into Q = 5 clusters using MMSB. The entry (i, j) of the matrix describes the number of blogs associated to the jth political party (column) and classified into cluster i (row). Each entry distinguishes blogs which belong to a unique cluster from overlaps (single membership blogs + overlaps). Cluster 5 corresponds to the class of outliers.*

factor that directly regulates the other operon. The network is made of three regulation patterns, each one of them having its own regulators and regulated operons. Therefore, using $Q = 6$ clusters, we applied our algorithm and we obtained the results in Table 5.

	UMP	UDF	liberal	PS	analysts	others
cluster 1	37	0	1	0	0	2
cluster 2	1	31	0	0	1	0
cluster 3	0	0	24	0	1	0
cluster 4	0	0	0	26	0	0
cluster 5	2	1	0	31	9	29

FIG. 9. *Classification of the blogs into Q = 5 clusters using SBM. The entry (i, j) of the matrix describes the number of blogs associated to the jth political party (column) and classified into cluster i (row). Cluster 5 corresponds to the class of outliers.*

TABLE 5
Classification of the operons into $Q = 6$ clusters. Operons in bold belong to multiple clusters

Cluster	Size	Operons
1	2	STE12 TEC1
2	33	YBR070C MID2 YEL033W SRD1 TSL1 RTS2 PRM5 YNL051W PST1 YJL142C **SSA4** YGR149W SPO12 YNL159C SFP1 YHR156C YPS1 YPL114W HTB2 MPT5 SRL1 DHH1 **TKL2** PGU1 YHL021C RTA1 WSC2 GAT4 YJL017W TOS11 YLR414C BNI5 YDL222C
3	2	MSN4 MSN2
4	32	CPH1 **TKL2** HSP12 SPS100 MDJ1 GRX1 SSA3 ALD2 GDH3 GRE3 HOR2 ALD3 SOD2 ARA1 HSP42 YNL077W **HSP78** GLK1 DOG2 HXK1 RAS2 **CTT1** HSP26 TPS1 TTR1 HSP104 GLO1 **SSA4** PNC1 MTC2 YGR086C **PGM2**
5	2	YAP1 SKN7
6	19	YMR318C **CTT1** TSA1 CYS3 ZWF1 HSP82 TRX2 GRE2 SOD1 AHP1 YNL134C **HSP78** CCP1 TAL1 DAK1 YDR453C TRR1 LYS20 **PGM2**

First, we notice that clusters 1, 3 and 5 contain only two operons each. These operons correspond to hubs which regulate respectively the nodes of clusters 2, 4 and 6, all having a very low *intra* connection probability. To analyze our results, we used GOToolBox [Martin et al. (2004)] on each cluster. This software aims at identifying statistically over-represented terms of the Gene Ontology (GO) in a gene data set. We found that the clusters correspond to well-known biological functions. Thus, the nodes of cluster 2 are regulated by STE12 and TEC1 which are both involved in the response to glucose limitation, nitrogen limitation and abundant fermentable carbon source. Similarly, MSN4 and MSN2 regulate the nodes of cluster 4 in response to different stress such as freezing, hydrostatic pressure and heat acclimation. Finally, the nodes of cluster 6 are regulated by YAP1 and SKN7 in the presence of oxygen stimulus. Our algorithm was able to uncover two overlapping clusters (operons in bold in Table 5). Interestingly, contrary to the other operons of clusters 2, 4 and 6, which are all regulated by operons of a single cluster (clusters 1, 3 or 5), these overlaps correspond to co-regulated operons. Thus, SSA4 and TKL2 belong to clusters 2 and 4 since they are co-regulated by (STE12, TEC1) and (MSN4 and MSN2). Moreover, HSP78, CTT1 and PGM2 belong to clusters 4 and 6 since they are co-regulated by (MSN4, MSN2) and (YAP1, SKN7). It should also be noted that OSBM did not classify 112 operons which all have very low output and input degrees.

Because the network is sparse, we obtained very poor results with CFinder. Indeed, the network contains only one 3-clique and no k-clique for $k > 3$. Therefore, for $k = 2$, all the operons were classified into a single cluster and no biological information could be retrieved. For $k = 3$, only three operons were classified into a single class and for $k > 3$ no operon was classified.

As previously, we ran MMSB and SMB with $Q + 1$ clusters and we identified the class of outliers. Both approaches retrieved the six clusters found by OSBM. However, we emphasize that, contrary to the political blogoshpere network, MMSB did not uncover any overlap in the yeast transcriptional regulatory network.

As in Section 6.1, these results clearly illustrate the capacity of OSBM to retrieve overlapping clusters in networks with complex topological structures. In particular, in situations where networks are not made of community structures, while the results of CFinder dramatically degrade or cannot even be interpreted, OSBM seems particularly promising.

7. Conclusion. In this paper we proposed a new random graph model, the Overlapping Stochastic Block Model, which can be used to retrieve overlapping clusters in networks. We used global and local variational techniques to obtain a tractable lower bound of the observed log-likelihood and we defined an EM like procedure which optimizes the model parameters in turn. We showed that the model is identifiable within classes of equivalence and we illustrated the efficiency of our approach compared to other methods, using simulated data and real networks. Since no assumption is made on the matrix $\mathbf{W}$ and vectors $\mathbf{U}$ and $\mathbf{V}$ used to characterize the connection probabilities, the model can take very different topological structures into account and seems particularly promising for the analysis of networks. In the experiment section we set the number Q of classes using a priori information we had about the networks. However, in future works, we believe it is crucial to develop a model selection criterion to estimate the number of classes automatically from the topology. We will also investigate introducing some priors over the model parameters to work in a full Bayesian framework.

SUPPLEMENTARY MATERIAL

Supplement: Appendix (DOI: 10.1214/10-AOAS382SUPP; .pdf). Describe how global and local variational techniques can be used to obtain a tractable lower bound. Introduce the optimization equations for the inference procedure.

REFERENCES

AIROLDI, E., BLEI, D., XING, E. and FIENBERG, S. (2006). Mixed membership stochastic block models for relational data with application to protein–protein interactions. In *Proceedings of the International Biometrics Society Annual Meeting*. Montréal, Québec, Canada.

AIROLDI, E., BLEI, D., FIENBERG, S. and XING, E. (2007). Mixed membership analysis of high-throughput interaction studies: Relational data. Available at ArXiv e-prints.

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

BICKEL, P. and CHEN, A. (2009). A non parametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BLEI, D., NG, A. and JORDAN, M. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BOER, P., HUISMAN, M., SNIJDERS, T., STEGLICH, C., WICHERS, L. and ZEGGELINK, E. (2006). StOCNET: An open software system for the advanced statistical analysis of social networks, Version 1.7.

BROYDEN, C., FLETCHER, R., GOLDFARB, D. and SHANNO, D. F. (1970). BFGS method. *J. Inst. Math. Appl.* **6** 76–90.

BYRD, R. H., LU, P., NOCEDAL, J. and ZHU, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16** 1190–1208. MR1346301

CHANG, J. (2010). The lda package Version, 1.2.

DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Statist. Comput.* **18** 173–183. MR2390817

ESTRADA, E. and RODRÍGUEZ-VELÁZQUEZ, J. A. (2005). Spectral measures of bipartivity in complex networks. *Phys. Rev. E (3)* **72** 046105. MR2202758

EVERETT, M. and BORGATTI, S. (1998). Analyzing clique overlap. *Connections* **21** 49–61.

FIENBERG, S. and WASSERMAN, S. (1981). Categorical data analysis of single sociometric relations. *Soc. Methodol.* **12** 156–192.

FRANK, O. and HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *J. Amer. Statist. Assoc.* **77** 835–840. MR0686407

FU, Q. and BANERJEE, A. (2008). Multiplicative mixture models for overlapping clustering. In *Proceedings of the IEEE International Conference on Data Mining* 791–796. Pisa, Italy.

GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826. MR1908073

GOLDENBERG, A., ZHENG, A., FIENBERG, S. and AIROLDI, E. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

GRIFFITHS, T. and GHAHRAMANI, Z. (2005). Infinite latent feature models and the Indian buffet process. *Adv. Neural Inform. Process. Syst.* **18** 475–482.

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300

HELLER, K. and GHAHRAMANI, Z. (2007). A nonparametric Bayesian approach to modeling overlapping clusters. In *Proceedings of the 11th International Conference on AI and Statistics*. San Juan, Puerto Rico.

HELLER, K., WILLIAMSON, S. and GHAHRAMANI, Z. (2008). Statistical models for partial membership. In *Proceedings of the 25th International Conference on Machine Learning* 392–399. Helsinki, Finland.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

HOFMAN, J. and WIGGINS, C. (2008). A Bayesian approach to network modularity. *Phys. Rev. Lett.* **100** 258701.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088

JEFFERY, C. (1999). Moonlighting proteins. *Trends Biochem. Sci.* **24** 8–11.

KRIVITSKY, P. and HANDCOCK, M. (2009). The latentnet package, Version 2.1-1.

KRIVITSKY, P., HANDCOCK, M., RAFTERY, A. and HOFF, P. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* **31** 204–213.

LACROIX, V., FERNANDES, C. and SAGOT, M.-F. (2006). Motif search in graphs: Application to metabolic networks. *Trans. Comput. Biol. Bioinform.* **3** 360–368.

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2009). *Advances in Data Analysis*, *Data Handling*, *and Business Intelligence*, *Bayesian Methods for Graph Clustering* 229–239. Springer, Berlin, Heidelberg.

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2010). Supplement A to "Overlapping stochastic block models with application to the French blogosphere network." DOI: 10.12.14/10-AOAS382SUPP.

MARTIN, D., BRUN, C., REMY, E., MOUREN, P., THIEFFRY, D. and JACQ, B. (2004). GOToolBox: Functional analysis of gene datasets based on Gene Ontology. *Genome Biol.* **5**.

MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, D., CHKLOVSKII, D. and ALON, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298** 824–827.

MORENO, J. (1934). *Who Shall Survive?*: *A New Approach to the Problem of Human Interrelations*. Nervous and Mental Disease Publishing, Washington, DC.

NEWMAN, M. and LEICHT, E. (2007). Mixture models and exploratory analysis in networks. *Proc. Natl. Acad. Sci. USA* **104** 9564–9569.

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255

PALLA, G., DERENYI, I., FARKAS, I. and VICSEK, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435** 814–818.

PALLA, G., DERENYI, I., FARKAS, I. and VICSEK, T. (2006). CFinder the community/cluster finding program, Version 2.0.1.

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block sturcture. *J. Classification* **14** 75–100. MR1449742

WHITE, H., BOORMAN, S. and BREIGER, R. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *Amer. J. Soc.* **81** 730–780.

ZANGHI, H., AMBROISE, C. and MIELE, V. (2008). Fast online graph clustering via Erdös–Renyi mixture. *Pattern Recognition* **41** 3592–3599.

LABORATOIRE STATISTIQUE ET GÉNOME
UMR CNRS 8071, INRA 1152
UNIVERSITY OF EVRY
91000 EVRY
FRANCE
E-MAIL: pierre.latouche@genopole.cnrs.fr
           etienne.birmele@genopole.cnrs.fr
           christophe.ambroise@genopole.cnrs.fr