Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 1, *pp.* 3–35

New consistent and asymptotically normal parameter estimates for random-graph mixture models

Christophe Ambroise and Catherine Matias

Université d'Évry Val d'Essonne, and Centre National de la Recherche Scientifique, Évry, France

[Received December 2010. Revised May 2011]

Summary. Random-graph mixture models are very popular for modelling real data networks. Parameter estimation procedures usually rely on variational approximations, either combined with the expectation–maximization (EM) algorithm or with Bayesian approaches. Despite good results on synthetic data, the validity of the variational approximation is, however, not established. Moreover, these variational approaches aim at approximating the maximum likelihood or the maximum *a posteriori* estimators, whose behaviour in an asymptotic framework (as the sample size increases to ∞) remains unknown for these models. In this work, we show that, in many different affiliation contexts (for binary or weighted graphs), parameter estimators based either on moment equations or on the maximization of some composite likelihood are strongly consistent and $\sqrt{n}$ convergent, when the number n of nodes increases to ∞. As a consequence, our result establishes that the overall structure of an affiliation model can be (asymptotically) caught by the description of the network in terms of its number of triads (order 3 structures) and edges (order 2 structures). Moreover, these parameter estimates are either explicit (as for the moment estimators) or may be approximated by using a simple EM algorithm, whose convergence properties are known. We illustrate the efficiency of our method on simulated data and compare its performances with other existing procedures. A data set of cross-citations among economics journals is also analysed.

Keywords: Composite likelihood; Mixture model; Random graph; Stochastic block model

1. Introduction

The analysis of network data appears in different scientific fields, such as social sciences, communication networks and many others, including a recent explosion in the field of molecular biology (with the study of metabolic networks, transcriptional regulatory networks and protein interactions networks). The literature is vast, and we refer for instance to Boccaletti *et al.* (2006), Goldenberg *et al.* (2010) and Kolaczyk (2009) for interesting introductions to networks.

Erdős and Rényi (1959) introduced one of the earliest and most studied random-graph models, in which binary random graphs are considered as a set of independent and identically distributed (IID) Bernoulli edge variables over a fixed set of nodes. This model is, however, too homogeneous to capture some important features of real networks, such as the presence of 'hubs', namely highly connected nodes. This lack of heterogeneity led to the introduction of mixture versions of the simple Erdős–Rényi model. So-called 'stochastic block models' (Daudin *et al.*, 2008; Frank and Harary, 1982; Holland *et al.*, 1983; Snijders and Nowicki, 1997) were

Address for correspondence: Catherine Matias, Laboratoire Statistique et Génome, 523, place des Terrasses de l'Agora, 91 000 Évry, France.
E-mail: catherine.matias@genopole.cnrs.fr

 1369–7412/12/74003

introduced in various forms, primarily in social sciences to study relational data. In this context, the nodes are partitioned into latent groups (blocks) characterizing the relationships between nodes. Block modelling thus refers to the particular structure of the adjacency matrix of the graph (i.e. the matrix containing the edge indicators). By reordering the nodes with respect to the groups that they belong to, this matrix exhibits blocks. Diagonal and off-diagonal blocks respectively represent intragroup and intergroup connections. Where blocks exhibit the same behaviour within their type (diagonal or off diagonal), we further obtain what we call an affiliation structure. Affiliation structures are parsimonious in the number of parameters that they use and may model many situations. For instance, affiliation models encompass both community structures and disassortative mixing (Newman and Leicht, 2007). In the first case (community structure) the intragroup connectivities are high whereas the intergroup connectivities are low. Disassortative mixing, rather, corresponds to high intergroup connectivities and low intragroup connectivities.

Many networks are or can be *weighted* (or in other words *valued*). Those weights are precious additional information on the graph and should be taken into account in their analysis. Well-known examples of weighted networks include airline traffic data between airports, co-authorship networks of scientists (Barrat *et al.*, 2004) or, when rather considering the corresponding adjacency matrix, financial correlation matrices (Laloux *et al.*, 1999). Whereas the two first examples correspond to sparse weighted networks, the last concerns dense (or complete) weighted graphs. Weighted networks are a way of integrating heterogeneous data and their analysis is thus of primary importance (Newman, 2004). Community detection (i.e. the problem of finding clusters of nodes with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters) has been widely considered in the context of weighted graphs (see for instance Fortunato (2010)). Whereas community detection methods are mainly algorithmic, another approach is to rely on generative models and random-graph mixtures. Stochastic block models for analysing random graphs with non-binary relationships between nodes have been considered either in the case of a finite number of possible relationships (Nowicki and Snijders, 2001) or for more general weighted graphs (Mariadassou *et al.*, 2010). Our approach builds on these latter references. We also point out the existence of *generalized block models* for valued networks (Ziberna, 2007; Doreian *et al.*, 2005) which, however, do not rely on a probabilistic model as we shall do here.

In this article we shall be interested in both binary and weighted random graphs and we shall focus on mixture models. We mention the existence of an increasing literature on two different related concepts: *mixed membership* (Airoldi *et al.*, 2008; Erosheva *et al.*, 2004) and *overlapping* (Latouche *et al.*, 2011a) stochastic block models for binary networks, in which nodes may belong to several classes. However, these models are beyond the scope of the present work.

Current parameter estimation procedures in random-graph mixture models rely on approximation of the likelihood, which is itself intractable owing to the presence of the non-observed groups. Either the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977) or Bayesian approaches are at the core of these strategies. Both rely on the computation of the distribution of the hidden node states, conditional on the observed edge variables. However, in the particular case of random-graph mixtures, the exact computation of this conditional distribution cannot be obtained, owing to its non-factorized form. Thus, approximate computations are made, leading to what is called 'variational' EM or Bayes strategies (Daudin *et al.*, 2008; Latouche *et al.*, 2011b; Picard *et al.*, 2009; Zanghi *et al.*, 2008). The major drawback of these methods is their relatively large computational time. Besides, even if these methods exhibit good behaviour on simulated data, they suffer from a lack of theoretical support. Indeed, two major features of these procedures still lack understanding. First, the quality of the variational approximation

is not known, and this approximation may even prevent convergence to local maxima of the likelihood (Gunawardana and Byrne, 2005). Second, the consistency of the maximum likelihood or of the maximum *a posteriori* estimators is still an open question in these models.

Here, we propose simple strategies for estimating the parameters of mixture random-graph models, in the particular affiliation case. The methods not only rely on established convergence results but are also simpler than variational approaches. By focusing on small structures (edges and triads) and treating these as if they were (but never assuming that they are) independent, we prove that we may recover the main features of an affiliation model. We adopt strategies that are based on either solving moment equations or maximizing a composite marginal likelihood. A composite marginal likelihood consists in the product of marginal distributions and may replace the likelihood in models with some dependence structure (see for instance Cox and Reid (2004) and Varin (2008)). In the weighted random-graphs case, our result shows that parameters may be estimated relying on a composite likelihood of univariate marginals. This is not so for binary random graphs, because parameters of mixtures of univariate Bernoulli distributions are not identifiable. However, parameters of mixtures of three-variate Bernoulli distributions are identifiable (see Allman *et al.* (2009), corollary 5). Thus, in the binary random-graph case, we develop moment or composite likelihood methods based on the marginals of triads, namely the three random variables (X_{ij}, X_{ik}, X_{jk}) induced by a set of three nodes (i, j, k).

Once the convergence of our estimators, let us say $\hat{\theta}_n$ to θ, has been established, the next question of interest concerns the order at which the discrepancy $\hat{\theta}_n - \theta$ converges to 0. We establish asymptotic normality results, thus obtaining rates of convergence of our procedures. This is in sharp contrast with existing methods and the first insight on the difficult issue of exhibiting (optimal) rates of convergence for parameter estimation procedures in these random-graph models. Indeed, a still open problem may be stated as follows: what is the parametric rate of convergence when observing $n(n-1)/2$ (non-independent) random variables over a set of n nodes, distributed according to a random-graph model? Is it $1/\sqrt{n}$ or $1/n$? In other words, the issue is whether the observation of these potentially $n(n-1)/2$ dependent edge variables over a set of n nodes enables existence of estimation procedures with rates of convergence of the order $1/n$ or rather $1/\sqrt{n}$. We obtain here theoretical results with rates of convergence of the order at least $1/\sqrt{n}$ (which might not be optimal). Moreover, in the *degenerate* case where the group proportions are equal, the rates of convergence increase to $1/n$. Our simulations seem also to indicate rates of convergence that are faster than $1/\sqrt{n}$, that might be due to degeneracies in the limiting variances of our central limit results, i.e. the fact that these variances might be 0.

Note that our results are of a very different nature from those recently obtained on clustering procedures for community detection in Bickel and Chen (2009), Rohe *et al.* (2011) and Choi *et al.* (2010). Indeed, those references establish that, under some conditions, the number of misclassified nodes (resulting from different algorithmic procedures) converges to 0 (as the number of nodes increases). Moreover, these results only concern the case of binary graphs and community detection, the latter being more restrictive than node clustering under an affiliation structure. However, we do not provide in this work any convergence result on the clustering procedure that we propose and rather focus on parameter estimation properties. Finally, note that Choi *et al.* (2010) also proposed convergence results on parameter estimates, but in a set-up of independent Bernoulli random variables, whereas in our context the random variables are not independent.

The paper is organized as follows. In Section 2, we state the various notations and present the general assumptions of our model as well as the main result: a law of large numbers and a central limit theorem for normalized sums of functions of variables over a k-tuple of nodes. Section 3 focuses on binary random graphs: after introducing the specific model for binary variables, we

present two different estimation procedures. The first (Section 3.1) relies on moment equations and is restrictive as it assumes that the group proportions π are known. Its main interest is that it gives some new light on the method that was proposed in the seminal article by Frank and Harary (1982). The second (Section 3.2) is more general (it does not assume known group proportions) and relies on composite likelihood. Section 4 presents the weighted random-graph model as well as the parameter estimation procedure, relying also on a composite likelihood approach. Whereas the first part of our work focuses on theoretical results about consistency of the parameter estimation procedures, the second part is dedicated to algorithmic issues as well as experiments. In Section 5, we present the implementation of the estimation procedures. Particular attention is paid to the problem of unravelling the latent structure of the model (Section 5.2). In Section 6, the performances of our procedures are illustrated on synthetic data and we also provide an analysis of a real data example. Finally, all the proofs are postponed until Appendix A.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Model and main result

We first give some notation that will be useful throughout this article. For any $Q \geqslant 1$, let $\mathcal{S}_Q$ denote the simplex $\{(\pi_1, \ldots, \pi_Q); \pi_i \geqslant 0; \Sigma_{i=1}^{Q} \pi_i = 1\}$ and $\mathcal{V}_Q = \{(v_1, \ldots, v_Q), v_i \in \{0, 1\}, \Sigma_{i=1}^{Q} v_i = 1\}$. For simplicity, in what follows we consider only undirected graphs with no self-loops. Easy generalizations may be done to handle directed graphs, with or without self-loops.

In this section, we define a general mixture model of random graphs in the following way. First, let $\{Z_i\}_{1 \leqslant i \leqslant n}$ be IID vectors $Z_i = (Z_{i1}, \ldots, Z_{iQ}) \in \mathcal{V}_Q$, following a multinomial distribution $\mathcal{M}(1, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q) \in \mathcal{S}_Q$. Random variable Z_i indicates to which group (among Q possibilities) node i belongs. These random variables are used to introduce heterogeneity in the random-graph model.

Next, the observations $\{X_{ij}\}_{1 \leqslant i < j \leqslant n}$ are indexed by the node pairs $\{i, j\}$ and take values in a general normed vector space $\mathcal{X}$ (in the next sections, $\mathcal{X} = \{0, 1\}$ or $\mathbb{N}$ or $\mathbb{R}^l$). We then assume that, conditional on the latent classes $\{Z_i\}_{1 \leqslant i \leqslant n}$, the random variables $\{X_{ij}\}_{1 \leqslant i < j \leqslant n}$ are independent. Moreover, the conditional distribution of X_{ij} depends only on Z_i and Z_j and has finite variance. The *general model* may thus be summarized in the following way:

$$\left.\begin{array}{c} \{Z_i\}_{1 \leqslant i \leqslant n} \text{ IID vectors in } \mathcal{V}_Q, \text{with distribution } \mathcal{M}(1, \boldsymbol{\pi}), \\ \{X_{ij}\}_{1 \leqslant i < j \leqslant n} \text{ observations in } \mathcal{X}, \\ \mathbb{P}(\{X_{ij}\}_{1 \leqslant i < j \leqslant n} | \{Z_i\}_{1 \leqslant i \leqslant n}) = \otimes_{1 \leqslant i < j \leqslant n} \mathbb{P}(X_{ij} | Z_i, Z_j), \\ \mathbb{E}(\|X_{i,j}\|^2 | Z_i, Z_j) < \infty. \end{array}\right\} \tag{1}$$

It may be worth noting that the variables $\{X_{ij}\}_{1 \leqslant i < j \leqslant n}$ are not independent in general, but we often make use of the fact that sets of non-adjacent edges induce independent random variables. More precisely, if $I, J \subset \{1, \ldots, n\}$ with $I \cap J = \emptyset$, then $\{X_{ij}\}_{(i,j) \in I^2}$ and $\{X_{ij}\}_{(i,j) \in J^2}$ are independent.

In the next sections, we shall focus on the particular affiliation mixture model, where the conditional distribution of an edge variable X_{ij} depends only on whether the end points i and j belong to the same group (i.e. $Z_i = Z_j$). We shall thus refer to the *affiliation structure* assumption

$$\mathbb{P}(X_{ij} | Z_i, Z_j) = \mathbb{P}(X_{ij} | \mathbf{1}_{Z_i = Z_j}), \tag{2}$$

where $\mathbf{1}_A$ is the indicator function of the set A.

Moreover, in the particular case of equal group proportions and affiliation structure, we shall observe some degeneracy phenomena. These are due to the fact that the distribution becomes invariant under permutation of the specific values of the node groups (see lemma 1 in Appendix A for more details). For later use, we thus also introduce the *equal group proportions* setting

$$\pi_q = 1/Q \qquad \text{for any } q \in \{1, \ldots, Q\}. \tag{3}$$

We now motivate the following developments. Under the affiliation structure assumption, the distribution of a single edge follows a two-components mixture of the form

$$X_{ij} \sim \gamma\, \mathbb{P}(X_{ij}|Z_i = Z_j) + (1-\gamma)\mathbb{P}(X_{ij}|Z_i \neq Z_j).$$

For weighted random graphs, we shall assume a parametric form for this absolutely continuous conditional distribution, namely $\mathbb{P}(X_{ij}|Z_i = Z_j) = \mathbb{P}_{\theta_{\text{in}}}(X_{ij})$ and $\mathbb{P}(X_{ij}|Z_i \neq Z_j) = \mathbb{P}_{\theta_{\text{out}}}(X_{ij})$. The vast majority of families of parametric absolutely continuous distributions give finite mixtures whose parameters are identifiable. This is equivalent to saying that $\mathbb{E}[\log\{\gamma\, \mathbb{P}_{\theta_{\text{in}}}(X_{12}) + (1-\gamma)\, \mathbb{P}_{\theta_{\text{out}}}(X_{12})\}]$ has a unique maximum at the true parameter value $(\theta_{\text{in}}, \theta_{\text{out}})$. This reasoning is at the core of maximum likelihood estimation and motivates the introduction of a *composite* log-likelihood

$$\mathcal{L}_X^{\text{compo}}(\theta) = \sum_{1 \leqslant i < j \leqslant n} \log\{\gamma\, \mathbb{P}_{\theta_{\text{in}}}(X_{ij}) + (1-\gamma)\, \mathbb{P}_{\theta_{\text{out}}}(X_{ij})\},$$

which is not the model likelihood as the random variables X_{ij} are not independent. Its usefulness to estimate the parameters relies on whether the renormalized criterion $\mathcal{L}_X^{\text{compo}}(\theta)/n(n-1)$ converges to the expectation $\mathbb{E}[\log\{\gamma\, \mathbb{P}_{\theta_{\text{in}}}(X_{12}) + (1-\gamma)\, \mathbb{P}_{\theta_{\text{out}}}(X_{12})\}]$. We shall prove below that the answer is yes and, thus, maximizing $\mathcal{L}_X^{\text{compo}}(\theta)$ with respect to θ is a good strategy.

In the binary random-graph case, however, the strategy must be modified because each random variable X_{ij} follows a mixture of univariate Bernoulli distributions whose parameters are not identifiable. We thus rather consider mixtures of three-variate Bernoulli distributions which appear to be sufficient to estimate the parameters consistently.

Thus, we are now interested more generally in the behaviour of empirical sums of functions of the random variables induced by a k-tuple of nodes. These empirical estimators are at the core of the estimation procedures that we shall later consider. For this, we introduce some more notation.

Define the set of nodes $\mathcal{I} = \{1, \ldots, n\}$ and the set of k distinct nodes $\mathcal{I}_k = \{(i_1, \ldots, i_k) \in \mathcal{I}^k; i_j \neq i_l \text{ for any } j \neq l\}$. ($\mathcal{I}_k$ is also the set of injective maps from $\{1, \ldots, k\}$ to $\mathcal{I} = \{1, \ldots, n\}$.) For any fixed integer $k \geqslant 1$, and any k-tuple of nodes $\mathbf{i} = (i_1, \ldots, i_k) \in \mathcal{I}_k$, we let $\mathbb{X}^{\mathbf{i}} = (X_{i_1 i_2}, \ldots, X_{i_1 i_k}, X_{i_2 i_3}, \ldots, X_{i_{k-1} i_k})$ be the vector of $p = \binom{k}{2}$ random variables induced by the k-tuple of nodes $\mathbf{i}$. Moreover, for any $s \geqslant 1$ and any measurable function $g : \mathcal{X}^p \to \mathbb{R}^s$, we let

$$\hat{m}_g = \frac{(n-k)!}{n!} \sum_{\mathbf{i} \in \mathcal{I}^k} g(\mathbb{X}^{\mathbf{i}}),$$
$$m_g = \mathbb{E}\{g(\mathbb{X}^{(1,\ldots,k)})\}.$$

Our first theorem establishes a strong law of large numbers as well as asymptotic normality of the estimator $\hat{m}_g$. As the random variables $\{X_{ij}\}$ are not independent, consistency (as well as asymptotic normality) of this empirical estimator is not trivial and must be established carefully.

Theorem 1. Under the assumptions of model (1), for any $k, s \geqslant 1$ and $p = \binom{k}{2}$ and any measurable function $g : \mathcal{X}^p \to \mathbb{R}^s$ such that $\mathbb{E}\{\|g(\mathbb{X}^{(1,\ldots,k)})\|^2\}$ is finite, the estimator $\hat{m}_g$ is consistent

$$\hat{m}_g \underset{n\to\infty}{\longrightarrow} m_g \qquad \text{almost surely,}$$

as well as asymptotically normal $\sqrt{n}(\hat{m}_g - m_g) \rightsquigarrow_{n\to\infty} \mathcal{N}(0, \Sigma_g)$. If we moreover assume an affiliation structure (2) with equal group proportions (3), then $\Sigma_g = 0$ and $n(\hat{m}_g - m_g)$ converges in distribution as $n \to \infty$.

Let us now give some comments about the previous result. First, an expression for the limiting covariance matrix Σ_g is given in the proof of the theorem. Such an expression is useful for instance in the construction of confidence intervals. However, although our estimators of the model parameters are derived from estimators of the form $\hat{m}_g$, here we did not obtain simple expressions for their limiting variance from an expression of Σ_g. Thus, rather than the exact form of the limiting distribution, we are more interested here in rates of convergence.

Theorem 1 states that the convergence of $\hat{m}_g$ to m_g happens with a rate at least $1/\sqrt{n}$. In the case where we consider an affiliation structure with equal group proportions, we prove that the limiting variance is null (i.e. $\Sigma_g = 0$), meaning that $\sqrt{n}(\hat{m}_g - m_g)$ converges in probability to 0. We then further prove that the sequence $n(\hat{m}_g - m_g)$ converges in distribution (to some non-Gaussian limit). Thus, in this degenerate case, the convergence of $\hat{m}_g$ happens at the faster rate $1/n$.

We shall see that consistency as well as rates of convergence are preserved in the estimation procedures that we deduce from moment estimators of the form $\hat{m}_g$. To our knowledge, this work is the first giving some insights about consistency and rates of convergence of parameter estimation procedures in random-graph mixture models.

In the next sections, we consider two particular instances of the mixture model that is defined in expression (1): the binary affiliation model (Section 3) and the weighted affiliation model (Section 4).

3. Binary affiliation model

In the case of binary random graphs, we observe binary random variables $\{X_{ij}\}_{1\leqslant i<j\leqslant n}$ indicating presence (1) or absence (0) of an edge between nodes i and j. The latent classes $\{Z_i\}_{1\leqslant i\leqslant n}$ are still distributed as IID multinomial vectors on $\mathcal{V}_Q$. Conditional on these latent classes $\{Z_i\}_{1\leqslant i\leqslant n}$, we assume that $\{X_{ij}\}_{1\leqslant i<j\leqslant n}$ are independent Bernoulli $\mathcal{B}(\cdot)$ random variables, with parameters depending on the node groups. More precisely, we restrict our attention to the affiliation structure model (2), where nodes connect differently whether they belong to the same group or not. We let

$$\forall q, l \in \{1, \ldots, Q\}, \qquad X_{ij} | Z_{iq} Z_{jl} = 1 \sim \begin{cases} \mathcal{B}(\alpha) & \text{if } q = l, \\ \mathcal{B}(\beta) & \text{if } q \neq l. \end{cases} \tag{4}$$

Here, α and β respectively are the intragroup and the intergroup connectivities and we let $p_{ql} = \alpha \mathbf{1}_{q=l} + \beta \mathbf{1}_{q\neq l}$, for any $1 \leqslant q, l \leqslant Q$. In what follows, we always assume that $\alpha \neq \beta$.

The whole parameter space is given by

$$\Pi = \{(\boldsymbol{\pi}, \alpha, \beta); \boldsymbol{\pi} \in \mathcal{S}_Q \cap (0,1)^Q, \alpha \in (0,1), \beta \in (0,1), \alpha \neq \beta\}.$$

We shall use the notation $b(x, p) = p^x (1-p)^{1-x}$ (where $x \in \{0, 1\}$ and $p \in [0, 1]$) for a Bernoulli density with respect to counting measure. Note that, in this set-up, the complete data log-likelihood is simply written

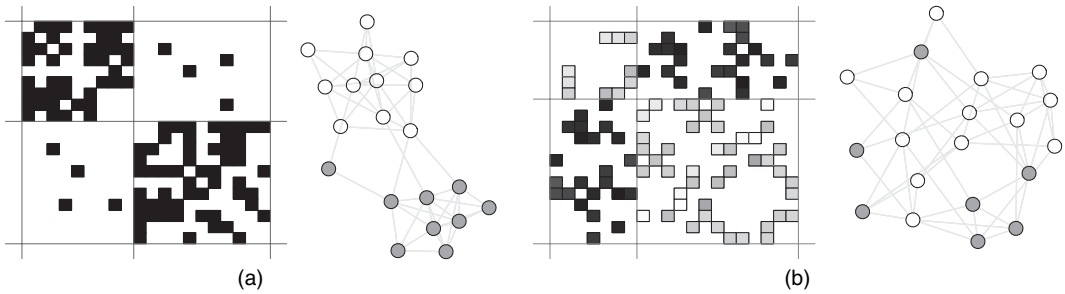

Fig. 1. Simulation of (a) binary and (b) Gaussian weighted graphs with two classes and 20 nodes: in each case, the picture displays the graph representations with vertices tinted according to classes, as well as the adjacency matrices where each entry X_{ij} of the matrix is the binary or weight value of the edge between vertex i and vertex j; the rows and columns of these matrices are organized according to the classes

$$\begin{aligned}\mathcal{L}_{X,Z}(\boldsymbol{\pi},\alpha,\beta) = \log\{\mathbb{P}_{\boldsymbol{\pi},\alpha,\beta}(\{X_{ij}\}_{1\leqslant i<j\leqslant n},\{Z_i\}_{1\leqslant i\leqslant n})\} &= \sum_{i=1}^{n}\sum_{q=1}^{Q} Z_{iq}\log(\pi_q)\\ &+ \sum_{1\leqslant i<j\leqslant n}\sum_{q=1}^{Q} Z_{iq}Z_{jq}\{X_{ij}\log(\alpha)+(1-X_{ij})\log(1-\alpha)\}\\ &+ \sum_{1\leqslant i<j\leqslant n}\sum_{1\leqslant q\neq l\leqslant Q} Z_{iq}Z_{jl}\{X_{ij}\log(\beta)+(1-X_{ij})\log(1-\beta)\}. \qquad (5)\end{aligned}$$

Fig. 1(a) displays an example of a binary random graph distributed according to this affiliation model.

3.1. Moment estimators in the binary affiliation model with known group proportions

The following approach based on moment equations was initially proposed by Frank and Harary (1982) to estimate the connectivity parameters α and β (as well as, in some cases, the number of groups Q). The core idea is simple: the moment equations corresponding to the distribution of a triplet (X_{ij}, X_{ik}, X_{jk}) give three equations which can be used to estimate the two parameters α and β, as soon as the group proportions (also appearing in these equations) are known. However, this method has not been thoroughly checked by Frank and Harary and may give rise to multiple solutions. Indeed, they did not discuss uniqueness of the solutions to the system of (non-linear) equations that they consider. This point has been partly discussed in Allman *et al.* (2011) and the estimation procedures proposed here are an echo to the identifiability results that were obtained there.

The following method applies only when the mixture proportions $\boldsymbol{\pi}$ are known. We develop in Section 5 an algorithmic procedure that iteratively estimates the group proportions in the first step, and the connectivity parameters (α,β) in a second step. This second step uses the method that we shall now describe.

First, we let $s_2 = \Sigma_q\,\pi_q^2$ and $s_3 = \Sigma_q\,\pi_q^3$. Then, we easily obtain the formula

$$\left.\begin{aligned} m_1 := \mathbb{E}(X_{ij}) &= s_2\alpha + (1-s_2)\beta,\\ m_2 := \mathbb{E}(X_{ij}X_{ik}) &= s_3\alpha^2 + 2(s_2-s_3)\alpha\beta + (1-2s_2+s_3)\beta^2,\\ m_3 := \mathbb{E}(X_{ij}X_{ik}X_{jk}) &= s_3\alpha^3 + 3(s_2-s_3)\alpha\beta^2 + (1-3s_2+2s_3)\beta^3. \end{aligned}\right\} \qquad (6)$$

Since any triplet (X_{ij}, X_{ik}, X_{jk}) takes finitely many states, its distribution is completely characterized by a finite number of its moments. In the binary affiliation mixture model context,

in fact only three different moments are induced by a triplet distribution. Thus, the previous three moment equations completely characterize the distribution of any triplet (X_{ij}, X_{ik}, X_{jk}). Note that looking at higher order motifs, namely at the distribution of a set of $p = \binom{k}{2}$ random variables over a set of k nodes for $k \geqslant 4$, would provide more equations but would also lead to more intricate methods (see for instance Allman *et al.* (2011)).

In Allman *et al.* (2011), the possible solutions (with respect to α and β) of this set of moment equations are examined. Their result distinguishes the equal group proportions case ($\pi_q = 1/Q, \forall 1 \leqslant q \leqslant Q$) where a degeneracy phenomenon takes place.

Theorem 2. (Allman *et al.*, 2011). If $m_2 \neq m_1^2$, then the π_qs are unequal and we can recover the parameters β and α via the rational formulae

$$\begin{aligned} \beta &= \frac{(s_3 - s_2 s_3)m_1^3 + (s_2^3 - s_3)m_2 m_1 + (s_3 s_2 - s_2^3)m_3}{(m_1^2 - m_2)(2s_2^3 - 3s_3 s_2 + s_3)}, \\ \alpha &= \frac{m_1 + (s_2 - 1)\beta}{s_2}. \end{aligned} \tag{7}$$

If $m_2 = m_1^2$, then the π_qs are equal and we have

$$\begin{aligned} \beta &= m_1 + \left(\frac{m_1^3 - m_3}{Q - 1}\right)^{1/3}, \\ \alpha &= Q m_1 + (1 - Q)\beta. \end{aligned} \tag{8}$$

As soon as s_2 and s_3 are known, by plugging estimators of the moments m_i into these equations, we obtain simple estimates for parameters α and β. We thus first introduce empirical moment estimators $\hat{m}_i$, which are defined by

$$\left.\begin{aligned} \hat{m}_1 &= \frac{1}{n(n-1)} \sum_{(i,j) \in \mathcal{I}_2} X_{ij}, \\ \hat{m}_2 &= \frac{1}{n(n-1)(n-2)} \sum_{(i,j,k) \in \mathcal{I}_3} X_{ij} X_{ik}, \\ \hat{m}_3 &= \frac{1}{n(n-1)(n-2)} \sum_{(i,j,k) \in \mathcal{I}_3} X_{ij} X_{ik} X_{jk}. \end{aligned}\right\} \tag{9}$$

Note that these estimators are all of the form $\hat{m}_g$ for some specific function g. Thus, their consistency is a consequence of theorem 1. We can then prove the following result.

Theorem 3. In the binary affiliation model specified by expressions (1) and (4), when the group proportions $\boldsymbol{\pi}$ are supposed to be known, we have the following results.

(a) When the π_qs are unequal, the estimators $(\hat{\alpha}, \hat{\beta})$ defined through expression (7) where the m_is are replaced by the $\hat{m}_i$s converge almost surely to (α, β). Moreover, the rate of this convergence is at least $1/\sqrt{n}$.
(b) When the π_qs are equal, the estimators $(\hat{\alpha}, \hat{\beta})$ defined through expression (8) where the m_is are replaced by the $\hat{m}_i$s converge almost surely to (α, β). Moreover, the rate of this convergence is at least $1/n$.

The performances of this method, combined with an iterative procedure to uncover the latent structure and to estimate the group proportions, are illustrated in Section 6.

3.2. M-estimators in the binary affiliation model

We shall now describe another parameter estimation procedure based on M-estimators (see for instance van der Vaart (1998), chapter 5), i.e. estimators maximizing some criterion (here, a composite likelihood). This procedure is more direct than the previous moments method that was developed in Section 3.1, as it does not assume a preliminary knowledge of the group proportions $\boldsymbol{\pi}$.

Let us recall that $\mathbb{X}^{(i,j,k)} = (X_{ij}, X_{ik}, X_{jk})$. The random vectors $\mathbb{X}^{(i,j,k)}$ form a set of non-independent, but identically distributed vectors, with distribution of each $\mathbb{X}^{(i,j,k)}$ given by the mixture

$$\mathbb{P}_{\boldsymbol{\pi},\alpha,\beta}(\mathbb{X}^{(1,2,3)}) = \sum_{1\leqslant q,l,m\leqslant Q} \pi_q \pi_l \pi_m \, b(X_{12}, p_{ql}) \, b(X_{13}, p_{qm}) \, b(X_{23}, p_{lm}),$$

where we recall that $p_{ql} = \alpha \, \mathbf{1}_{q=l} + \beta \, \mathbf{1}_{q\neq l}$. In this mixture, many components are in fact equal. Indeed, the components reduce to only four (when $Q=2$) or five (when $Q \geqslant 3$) different distributions. More precisely, we may write

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi},\alpha,\beta}(\mathbb{X}^{(1,2,3)}) = {} & \gamma_1 \, b(X_{12},\alpha) \, b(X_{13},\alpha) \, b(X_{23},\alpha) + \gamma_2 \, b(X_{12},\beta) \, b(X_{13},\beta) \, b(X_{23},\alpha) \\ & + \gamma_3 \, b(X_{12},\beta) \, b(X_{13},\alpha) \, b(X_{23},\beta) + \gamma_4 \, b(X_{12},\alpha) \, b(X_{13},\beta) \, b(X_{23},\beta) \\ & + \gamma_5 \, b(X_{12},\beta) \, b(X_{13},\beta) \, b(X_{23},\beta), \end{aligned} \tag{10}$$

where the five proportions $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_5) \in \mathcal{S}_5$ appearing in this mixture are related to the original proportions $\boldsymbol{\pi}$, by the following relationships:

$$\left.\begin{aligned} \gamma_1 &= \sum_{q=1}^{Q} \pi_q^3 = s_3, \\ \gamma_j &= \sum_{1\leqslant q\neq l\leqslant Q} \pi_q^2 \pi_l = s_2 - s_3, \text{ for } j \in \{2,3,4\}, \\ \gamma_5 &= \sum_{\substack{1\leqslant q,\,l,\,m\leqslant Q \\ |\{q,l,m\}|=3}} \pi_q \pi_l \pi_m = 1 - 3s_2 + 2s_3. \end{aligned}\right\} \tag{11}$$

When $Q=2$, the fifth proportion γ_5 is automatically equal to 0. Moreover, as soon as $Q \leqslant 3$, the set of equations (11) defines a one-to-one relation between $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$. However, when $Q > 3$, the parameter $\boldsymbol{\pi}$ is not uniquely defined from $\boldsymbol{\gamma}$ and is not identifiable from the mixture distribution (10).

We emphasize that distribution (10) is a constrained three-variate Bernoulli mixture. Parameter identifiability of such a distribution is further discussed below. However, we should already remark that, whereas parameters of mixture models may in general be identified only up to a permutation on the node labels, the constrained form of the mixture (10) has the following consequence: the parameters α and β will be exactly recovered as soon as the mixture components are identified from equation (10) and whatever the labelling of these mixture components is. Indeed, among the five unordered components of the mixture, only three of them will be the product of two identical one-dimensional distributions, times a different distribution. The parameter β is then the parameter appearing in exactly two marginals in any of those three components.

Let us consider as our criterion a composite marginal log-likelihood of the observations

$$\mathcal{L}_X^{\text{compo}}(\boldsymbol{\pi},\alpha,\beta) = \sum_{(i,j,k)\in\mathcal{I}_3} \log\{\mathbb{P}_{\boldsymbol{\pi},\alpha,\beta}(\mathbb{X}^{(i,j,k)})\}. \tag{12}$$

We stress that this quantity is not derived from the marginal of the model complete-data likelihood (expressed in equation (5)) and is simpler. It would be the log-likelihood of the observations if the triplets $\{\mathbb{X}^{(i,j,k)}\}_{(i,j,k)\in\mathcal{I}_3}$ were independent, which is obviously not so. We now define our estimators as

$$(\hat{\boldsymbol{\pi}}_n, \hat{\alpha}_n, \hat{\beta}_n) = \arg\max_{\boldsymbol{\pi},\alpha,\beta} \{\mathcal{L}_X^{\text{compo}}(\boldsymbol{\pi}, \alpha, \beta)\}. \tag{13}$$

Note that, according to the non-uniqueness of group proportions $\boldsymbol{\pi}$ corresponding to mixture proportions $\boldsymbol{\gamma}$, the maximum with respect to $\boldsymbol{\pi}$ in the above equation may not be unique. We also let $\hat{\boldsymbol{\gamma}}_n$ be defined from $\hat{\boldsymbol{\pi}}_n$ through expression (11).

Using theorem 1, the renormalized criterion (12) converges to a limit. The key point here is that, under an identifiability assumption on the model parameters, this limit is a function whose maximum is attained only at the true parameter value $(\boldsymbol{\gamma}, \alpha, \beta)$. Using classical results from M-estimators (van der Vaart, 1998; Wald, 1949), we can then obtain consistency and asymptotic normality of the estimators defined through equation (13). We thus need here to assume the identifiability of the model parameters.

Assumption 1. The parameters $\boldsymbol{\gamma}, \alpha$ and β of the model that is defined by equation (10) are identifiable. In other words, if there exist $(\boldsymbol{\pi}, \alpha, \beta)$ and $(\boldsymbol{\pi}', \alpha', \beta')$ such that for any $(x, y, z) \in \{0, 1\}^3$ we have

$$\mathbb{P}_{\boldsymbol{\pi},\alpha,\beta}(X_{12} = x, X_{13} = y, X_{23} = z) = \mathbb{P}_{\boldsymbol{\pi}',\alpha',\beta'}(X_{12} = x, X_{13} = y, X_{23} = z),$$

then $(\boldsymbol{\gamma}, \alpha, \beta) = (\boldsymbol{\gamma}', \alpha', \beta')$, where $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are defined through expression (11) as functions of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ respectively.

We now make comments on this assumption. We first mention that identifiability of all the parameters $(\boldsymbol{\pi}, \alpha, \beta)$ in the model that is defined by expressions (1) and (4), i.e. relying on the full distribution over $\cup_{n\geqslant 1} \{0, 1\}^{\binom{n}{2}}$ (comprising the marginal distributions of the random graphs over a set of n nodes, for any value of n), is a difficult issue, for which only partial results have been obtained in Allman *et al.* (2011). Surprisingly, the results under the affiliation assumption are more difficult to obtain than in the non-affiliation case. The question here is slightly different and we ask whether a triplet distribution (10) is sufficient to identify only α and β (as well as the corresponding proportions $\boldsymbol{\gamma}$). As already pointed out, distribution (10) is a constrained distribution from the larger class of three-variate Bernoulli mixtures. In the case of (unconstrained) finite mixtures of multivariate (or three-variate) Bernoulli distributions, although the models have been used for decades and were strongly believed to be identifiable (Carreira-Perpiñán and Renals, 2000), the rigorous corresponding result has been established only very recently and by using rather elaborate techniques (see Allman *et al.* (2009), corollary 5). Unfortunately, this latter result does not apply directly here. Although this might be difficult to establish, we strongly believe that $\boldsymbol{\gamma}$, α and β are identifiable from distribution (10) and we advocate that, from the simulations that we performed, it seems a reasonable assumption to make.

In what follows, we also restrict our attention to compact parameter spaces, as this greatly simplifies the proofs and is not much restrictive. Generalizations could be done at the cost of technicalities (see for instance van der Vaart (1998), chapter 5).

Assumption 2. Assume that there is some $\delta > 0$ such that the parameter space is restricted to $\Pi_\delta = \{(\boldsymbol{\pi}, \alpha, \beta) \in \Pi; \forall 1 \leqslant q \leqslant Q, \pi_q \geqslant \delta, \alpha \in [\delta, 1-\delta], \beta \in [\delta, 1-\delta]\}$.

We can then prove the following result.

Theorem 4. In the model defined by expressions (1) and (4), under assumptions 1 and 2, the estimators $(\hat{\gamma}_n, \hat{\alpha}_n, \hat{\beta}_n)$ defined by expression (13) are consistent, as the sample size n grows to ∞. Moreover, the rate of this convergence is at least $1/\sqrt{n}$ and increases to $1/n$ in the particular case of equal group proportions (3).

We now comment on this result. We prove that the rate of convergence of our estimators is at least $1/\sqrt{n}$. However, our simulations (see Section 6) seem to exhibit a faster rate, indicating that the limiting covariance matrix of the discrepancy $\sqrt{n}(\hat{\gamma}_n - \gamma, \hat{\alpha}_n - \alpha, \hat{\beta}_n - \beta)$ might be zero, even beyond the case of equal group proportions. Also, when $Q \leqslant 3$, a consequence of the above result is that the estimator of the group proportions $\hat{\boldsymbol{\pi}}_n$ defined through $\hat{\gamma}_n$ as the unique solution to the system of equations (11) is also consistent and converges with a rate at least $1/\sqrt{n}$.

As is always the case for mixture models, the (composite) log-likelihood (12) cannot be computed exactly (except for very small sample sizes). Approximate computation of the estimators in expression (13) can be done by using an EM algorithm (Dempster *et al.*, 1977). This procedure is presented in Section 5.1. It is known (Wu, 1983) that, under reasonable assumptions, the EM algorithm will give a solution converging to the estimators that are defined by expression (13), as the number of iterates grows to ∞.

4. Weighted random graphs

In this section, we focus on a particular instance of model (1) for weighted random graphs. The observations are random variables $\{X_{ij}\}_{1 \leqslant i < j \leqslant n}$ that are either equal to 0, indicating the absence of an edge between nodes i and j, or a non-null real number, indicating the weight of the corresponding edge. We still assume that, conditional on the latent structure $\{Z_i\}_{1 \leqslant i \leqslant n}$, the random variables $\{X_{ij}\}_{1 \leqslant i < j \leqslant n}$ are independent, and the distribution of each X_{ij} depends only on Z_i and Z_j. We now further specify the model by assuming the following form for this distribution:

$$\forall q, l \in \{1, \ldots, Q\}, \qquad X_{ij} | Z_{iq} Z_{jl} = 1 \sim p_{ql} f(\cdot, \theta_{ql}) + (1 - p_{ql})\, \delta_0(\cdot), \tag{14}$$

where $\{f(\cdot, \theta), \theta \in \Theta\}$ is a parametric family of distributions, δ_0 is the Dirac measure at 0 and $p_{ql} \in (0, 1]$ are sparsity parameters. We let $\mathbf{p} = \{p_{ql}\}$ and $\boldsymbol{\theta} = \{\theta_{ql}\}$. The conditional distribution of X_{ij} is thus a mixture of a Dirac distribution at zero accounting for non-present edges, with proportion given by the sparsity parameter $\mathbf{p}$ (which can be 1 in a complete weighted graph) and a parametric distribution with density f that gives the weight of present edges. We focus on two different sparsity structures:

(a) either the sparsity is constant across the graph, $p_{ql} = p, \forall 1 \leqslant q, l \leqslant Q$;
(b) or the sparsity parameters model an affiliation structure $p_{ql} = \alpha\, \mathbf{1}_{q=l} + \beta\, \mathbf{1}_{q \neq l}$ and we assume that $\alpha \neq \beta$.

We moreover assume that we know the sparsity structure type. In any case, the connectivity parameter $\boldsymbol{\theta}$ is assumed to take exactly two different values:

$$\forall q, l \in \{1, \ldots, Q\}, \theta_{ql} = \begin{cases} \theta_{\text{in}} & \text{if } q = l, \\ \theta_{\text{out}} & \text{if } q \neq l, \end{cases}$$

with $\theta_{\text{in}} \neq \theta_{\text{out}}$. For identifiability reasons, we also constrain the parametric family $\{f(\cdot, \theta), \theta \in \Theta\}$ such that any distribution in this set admits a continuous cumulative distribution function at zero. Indeed, if this were not so, it would not be possible to distinguish between a zero weight

and an absent edge. Note that this model satisfies the affiliation assumption given by equation (2). Here, the complete-data log-likelihood is simply written

$$\mathcal{L}_{X,Z}(\boldsymbol{\pi},\mathbf{p},\boldsymbol{\theta}) = \log[\mathbb{P}_{\boldsymbol{\pi},\mathbf{p},\boldsymbol{\theta}}(\{X_{ij}\}_{1\leqslant i<j\leqslant n},\{Z_i\}_{1\leqslant i\leqslant n})] = \sum_{i=1}^{n}\sum_{q=1}^{Q} Z_{iq}\log(\pi_q) + \sum_{1\leqslant i<j\leqslant n}\sum_{1\leqslant q,l\leqslant Q} Z_{iq}Z_{jl}(\mathbf{1}_{X_{ij}\neq 0}[\log\{f(X_{ij},\theta_{ql})\}+\log(p_{ql})] + \mathbf{1}_{X_{ij}=0}\log(1-p_{ql})). \tag{15}$$

We now give some examples of parametric families $\{f(\cdot,\theta),\theta\in\Theta\}$ that could be used as weights (or values) on the edges.

(a) *Example 1*: let $\theta=(\mu,\sigma^2)\in\mathbb{R}\times(0,\infty)$ and consider $f(\cdot,\theta)$ the density of the Gaussian distribution with mean μ and variance σ^2.
(b) *Example 2*: let $\theta\in(0,\infty)$ and consider $f(\cdot,\theta)$ the density (with respect to the counting measure) of the Poisson distribution, with parameter θ, truncated at zero. Namely,

$$\forall k\geqslant 1,\qquad f(k,\theta)=\frac{\theta^k}{k!}\{\exp(\theta)-1\}^{-1}.$$

In example 2, the Poisson distribution is truncated at zero because, as previously mentioned, it would not be possible to distinguish a zero-valued weight from an absent edge. Sparsity of the graph is modelled through the parameter $\mathbf{p}$ only and the density $f(\cdot,\theta)$ concerns weights on present edges.

Fig. 1 illustrates the difference between binary and weighted random-graph affiliation models. For example the weighted graph of Fig. 1 displays no binary affiliation structure: if the weights were truncated by using the function $x\to\mathbf{1}_{x\neq 0}$, we would not obtain that the two groups have different intragroup and intergroup connectivities. This means that classical community clustering algorithms would fail to find any meaningful structure on this type of graph.

To our knowledge, this model has never been proposed in this form in the literature. In particular, the closest form was given in Mariadassou *et al.* (2010) who did not introduce a possible Dirac mass at zero to enable sparsity of the graph.

We now describe our estimation procedure based on M-estimators and a composite likelihood criterion. We proceed in two steps and first estimate the sparsity parameter, relying on an induced binary random graph. In the second step, we plug in this estimator and focus on the connectivity parameters $\boldsymbol{\theta}$ by relying only on the present edges.

(a) *Estimating the sparsity parameter*: let us first consider the case where $p_{ql}=p$. Then, we naturally estimate the sparsity parameter p by

$$\hat{p}_n=\frac{2}{n(n-1)}\sum_{1\leqslant i<j\leqslant n}\mathbf{1}_{X_{ij}=0}.$$

The consistency, as well as the rate of convergence, of this estimator follows from theorem 1. In the case where the sparsity parameter rather satisfies $p_{ql}=\alpha\,\mathbf{1}_{q=l}+\beta\,\mathbf{1}_{q\neq l}$, with $\alpha\neq\beta$, we rely on the underlying binary random graph (obtained by setting $Y_{ij}=\mathbf{1}_{X_{ij}\neq 0}$) and apply the results of Sections 3.1 or 3.2 to estimate α and β consistently.
(b) *Estimating the connectivity parameter* $\boldsymbol{\theta}$: the *present* edges X_{ij} (where i and j are such that $X_{ij}\neq 0$) are non-independent random variables, distributed according to a simple univariate mixture model $\Sigma_{ql}\,\pi_q\pi_l p_{ql}\,f(\cdot,\theta_{ql})$. For classical distributions $f(\cdot,\theta)$, it is possible to estimate the connectivity parameters θ_{ql} of this univariate mixture directly. In fact, we prove that, as soon as the parameters $\{\theta_{ql}\}$ are uniquely identified from the mixture

$\Sigma_{ql}\,\pi_q\pi_l p_{ql}\; f(\cdot,\theta_{ql})$ and, for regular parametric families $\{f(\cdot,\theta),\theta\in\Theta\}$, a consequence of theorem 1 is that maximizing a composite likelihood of the set of present edge variables provides a consistent estimator of the parameters. Let us introduce the assumptions needed.

Assumption 3. The parameters of finite mixtures of the family of measures $\mathcal{F}=\{f(\cdot;\theta);\theta\in\Theta\}$ are identifiable (up to label swapping). In other words, for any integer $m\geqslant 1$, if

$$\sum_{i=1}^{m}\lambda_i f(\cdot,\theta_i)=\sum_{i=1}^{m}\lambda_i' f(\cdot,\theta_i')$$

then

$$\sum_{i=1}^{m}\lambda_i\,\delta_{\theta_i}(\cdot)=\sum_{i=1}^{m}\lambda_i'\,\delta_{\theta_i'}(\cdot).$$

Continuing examples 1 and 2, note that both the families of Gaussian and truncated Poisson densities satisfy assumption 3. More generally, a wide range of parametric families of densities on $\mathbb{R}$ satisfy assumption 3 (see section 3.1 in Titterington *et al.* (1985) for more details).

The next assumption deals with regularity conditions on the model. Note that this assumption could be weakened by using the concept of differentiability in quadratic mean (see for instance van der Vaart (1998)).

Assumption 4. The functions $\theta\mapsto f(\cdot,\theta)$ are twice continuously differentiable on Θ.

This assumption is only technical and not very restrictive. It requires the parameter set to be compact and could be weakened at the cost of some technicalities.

Assumption 5. Assume that there is some $\delta>0$ and some compact subset $\Theta_c\subset\Theta$ such that the parameter space is restricted to the set $\{(\boldsymbol{\pi},\mathbf{p},\boldsymbol{\theta});\forall 1\leqslant q\leqslant Q,\pi_q\geqslant\delta,\mathbf{p}\in[\delta,1-\delta],\boldsymbol{\theta}\in\Theta_c\}$.

Now, each *present* edge variable X_{ij} such that $X_{ij}\neq 0$ is distributed according to the mixture $\Sigma_{1\leqslant q,l\leqslant Q}\pi_q\pi_l p_{ql} f(\cdot;\theta_{ql})$. As there are only two different components in this mixture, we express it in the more convenient form

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi},\mathbf{p},\boldsymbol{\theta}}(X_{ij})&=\left(\sum_{q=1}^{Q}\pi_q^2 p_{qq}\right)f(X_{ij};\theta_{\text{in}})+\left(\sum_{1\leqslant q\neq l\leqslant Q}\pi_q\pi_l p_{ql}\right)f(X_{ij};\theta_{\text{out}})\\&:=\gamma_{\text{in}}\;f(X_{ij};\theta_{\text{in}})+\gamma_{\text{out}}\;f(X_{ij};\theta_{\text{out}}).\end{aligned}\tag{16}$$

We consider a *composite* log-likelihood of present edges defined by

$$\mathcal{L}_X^{\text{compo}}(\boldsymbol{\pi},\mathbf{p},\boldsymbol{\theta})=\sum_{1\leqslant i<j\leqslant n}\log\{\gamma_{\text{in}}\;f(X_{ij};\theta_{\text{in}})+\gamma_{\text{out}}\;f(X_{ij};\theta_{\text{out}})\}.\tag{17}$$

We stress that this quantity is not derived from the marginal of the model complete-data likelihood (expressed in equation (15)) and is simpler. We now define estimators as

$$\hat{\boldsymbol{\theta}}_n=\{\hat{\theta}_{\text{in}},\hat{\theta}_{\text{out}}\}=\arg\max_{\boldsymbol{\theta}}\{\mathcal{L}_X^{\text{compo}}(\boldsymbol{\pi},\hat{\mathbf{p}}_n,\boldsymbol{\theta})\},\tag{18}$$

where $\hat{\mathbf{p}}_n$ is a preliminary step estimator of $\mathbf{p}$. Note that, owing to the label swapping issue on the hidden states, we estimate the set of values $\{\theta_{\text{in}},\theta_{\text{out}}\}$ and cannot distinguish θ_{in} from θ_{out}. Section 5.2 deals further with this issue.

We can now prove the following theorem.

Theorem 5. In the model defined by expressions (1) and (14), under assumptions 3–5, the set of unordered M-estimators $\hat{\boldsymbol{\theta}}_n = \{\hat{\theta}_{\text{in}}, \hat{\theta}_{\text{out}}\}$ defined by equation (18) is consistent, as the sample size n grows to ∞. Moreover, the rate of this convergence is at least $1/\sqrt{n}$ and increases to $1/n$ in the particular case of equal group proportions (3).

The proof mainly relies on the consistency of the normalized criterion (17). This point is a direct consequence of theorem 1. Then, from the criterion consistency, the identifiability and the regularity assumptions, one can derive the consistency of the corresponding M-estimator from classical theory (van der Vaart, 1998; Wald, 1949).

As already noted in the case of theorem 4, our result establishes a rate of convergence that is equal at least to $1/\sqrt{n}$. The simulations (Section 6) seem to indicate that the rate may in fact be faster, which is something that may be due to the degeneracy of the limiting variance of $\sqrt{n}(\hat{\theta}_n - \theta)$.

As for M-estimators in the binary case, we shall approximate this maximum (composite) likelihood estimator by using an EM procedure (Dempster *et al.*, 1977) whose convergence properties are well established (Wu, 1983). In contrast with the procedure that was presented in Section 3.2 where we need to adapt the EM framework to our specific model, we rely here on the classical EM algorithm and thus do not recall it.

5. Algorithms

In this section, we provide tools to implement the procedures that were described previously, as well as a complement to the issue of recovering the latent structure of a graph.

5.1. Expectation–maximization algorithm with triplets

In this section we describe the EM algorithm that was developed to approximate the estimators defined by expression (13).

In what follows, each set of three nodes $\{i, j, k\}$ corresponds to an index $\mathbf{i}$ ranging over the set $\{1, \ldots, N\}$, where $N = n(n-1)(n-2)$ is the total number of triplets. We let $\mathbb{X}^{\mathbf{i}} = (X^{\mathbf{i},1}, X^{\mathbf{i},2}, X^{\mathbf{i},3})$ be one of the observed triplets (namely each $X^{\mathbf{i},j}$ for $1 \leqslant j \leqslant 3$ corresponds to some former random variable X_{st} for some $1 \leqslant s, t \leqslant n$) and $U_{\mathbf{i}} \sim \mathcal{M}(1, \boldsymbol{\gamma})$ is the vector encoding the corresponding hidden state, namely, $U_{\mathbf{i}} \in \mathcal{V}_5$. We also denote by $\tau_{\mathbf{i}k}$ the posterior probability of node triplet $\mathbf{i}$ being in state k, conditional on the observation $\mathbb{X}^{\mathbf{i}}$, namely $\tau_{\mathbf{i}k} = \mathbb{P}(U_{\mathbf{i}k} = 1 | \mathbb{X}^{\mathbf{i}})$, for $1 \leqslant k \leqslant 5$ and $1 \leqslant \mathbf{i} \leqslant N$. Moreover, we encode the fact that, conditional on the five different hidden states of U, each co-ordinate of $\mathbb{X}$ is distributed according to either $\mathcal{B}(\alpha)$ or $\mathcal{B}(\beta)$, using the notation

$$\delta_{jk} = (\delta^1_{jk}, \delta^2_{jk}) = (\mathbf{1}_{X^{\mathbf{i},j} | U_{\mathbf{i}k} = 1 \sim \mathcal{B}(\alpha)}, \mathbf{1}_{X^{\mathbf{i},j} | U_{\mathbf{i}k} = 1 \sim \mathcal{B}(\beta)}),$$

for all $1 \leqslant j \leqslant 3$, $1 \leqslant k \leqslant 5$ and any $1 \leqslant \mathbf{i} \leqslant N$. Note that δ_{jk} is deterministic and that $\delta_{jk} \in \mathcal{V}_2$. With this notation, we are in the situation where we consider a composite likelihood (12) of random vectors $\{\mathbb{X}^{\mathbf{i}}\}_{1 \leqslant \mathbf{i} \leqslant N}$ from the mixture of five different three-dimensional Bernoulli distributions, the latent classes being the random vectors $\{U_{\mathbf{i}}\}_{1 \leqslant \mathbf{i} \leqslant N}$.

The EM algorithm is intended to compute and optimize iteratively, with respect to $(\boldsymbol{\gamma}, \alpha, \beta)$, the function

$$Q\{(\boldsymbol{\gamma}, \alpha, \beta); (\boldsymbol{\gamma}^{(s)}, \alpha^{(s)}, \beta^{(s)})\} = \mathbb{E}_{\boldsymbol{\gamma}^{(s)}, \alpha^{(s)}, \beta^{(s)}}(\log[\mathbb{P}_{\boldsymbol{\gamma}, \alpha, \beta}(\{U_{\mathbf{i}}\}_{1 \leqslant \mathbf{i} \leqslant N}, \{\mathbb{X}^{\mathbf{i}}\}_{1 \leqslant \mathbf{i} \leqslant N}) | \{\mathbb{X}^{\mathbf{i}}\}_{1 \leqslant \mathbf{i} \leqslant N}]),$$

using the current value of the parameter $(\boldsymbol{\gamma}^{(s)}, \alpha^{(s)}, \beta^{(s)})$. If we let $\tau_{\mathbf{i}k}^{(s)} = \mathbb{P}_{\boldsymbol{\gamma}^{(s)},\alpha^{(s)},\beta^{(s)}}(U_{\mathbf{i}k} = 1 | \mathbb{X}^{\mathbf{i}})$, we can write

$$\begin{aligned} Q\{(\boldsymbol{\gamma}, \alpha, \beta); (\boldsymbol{\gamma}^{(s)}, \alpha^{(s)}, \beta^{(s)})\} = & \sum_{\mathbf{i}=1}^{N} \sum_{k=1}^{5} \tau_{\mathbf{i}k}^{(s)} \log(\gamma_k) + \sum_{\mathbf{i}=1}^{N} \sum_{k=1}^{5} \tau_{\mathbf{i}k}^{(s)} \sum_{j=1}^{3} \delta_{jk}^{1} \{ X^{\mathbf{i},j} \log(\alpha) \\ & + (1 - X^{\mathbf{i},j}) \log(1-\alpha) \} \\ & + \delta_{jk}^{2} \{ X^{\mathbf{i},j} \log(\beta) + (1 - X^{\mathbf{i},j}) \log(1-\beta) \}. \end{aligned} \tag{19}$$

Starting from an initial value $(\boldsymbol{\gamma}^{(1)}, \alpha^{(1)}, \beta^{(1)})$, the EM algorithm proceeds in two iterative steps. At iteration s, the E-step computes the posterior distribution of $U_{\mathbf{i}}$ conditional on $\mathbb{X}^{\mathbf{i}}$, namely

$$\tau_{\mathbf{i}k}^{(s)} = \mathbb{P}_{\boldsymbol{\gamma}^{(s)},\alpha^{(s)},\beta^{(s)}}(U_{\mathbf{i}k} = 1 | \mathbb{X}^{\mathbf{i}}) = \frac{\gamma_k^{(s)} \prod_{j=1}^{3} b(X^{\mathbf{i},j}, \delta_{jk}^{1} \alpha^{(s)} + \delta_{jk}^{2} \beta^{(s)})}{\sum_{l=1}^{5} \gamma_l^{(s)} \prod_{j=1}^{3} b(X^{\mathbf{i},j}, \delta_{jl}^{1} \alpha^{(s)} + \delta_{jl}^{2} \beta^{(s)})},$$

for every $1 \leqslant \mathbf{i} \leqslant N$ and every $1 \leqslant k \leqslant 5$. By using equation (19), we then obtain the value of $Q\{(\boldsymbol{\gamma}, \alpha, \beta); (\boldsymbol{\gamma}^{(s)}, \alpha^{(s)}, \beta^{(s)})\}$. In the M-step, this quantity is maximized with respect to $(\boldsymbol{\gamma}, \alpha, \beta)$ and the maximizer gives the next value of the parameter $(\boldsymbol{\gamma}^{(s+1)}, \alpha^{(s+1)}, \beta^{(s+1)})$. This step relies on the following equations:

$$\gamma_k^{(s+1)} = N^{-1} \sum_{\mathbf{i}=1}^{N} \tau_{\mathbf{i}k}^{(s)}, \qquad k = 1, 5,$$

$$\gamma_k^{(s+1)} = (3N)^{-1} \sum_{\mathbf{i}=1}^{N} \tau_{\mathbf{i}2}^{(s)} + \tau_{\mathbf{i}3}^{(s)} + \tau_{\mathbf{i}4}^{(s)}, \qquad k = 2, 3, 4,$$

$$\begin{aligned} \alpha^{(s+1)} = & \left\{ \sum_{\mathbf{i}=1}^{N} \tau_{\mathbf{i}1}^{(s)} (X^{\mathbf{i},1} + X^{\mathbf{i},2} + X^{\mathbf{i},3}) + \tau_{\mathbf{i}2}^{(s)} X^{\mathbf{i},3} + \tau_{\mathbf{i}3}^{(s)} X^{\mathbf{i},2} + \tau_{\mathbf{i}4}^{(s)} X^{\mathbf{i},1} \right\} \\ & \times \left(\sum_{\mathbf{i}=1}^{N} 3\tau_{\mathbf{i}1}^{(s)} + \tau_{\mathbf{i}2}^{(s)} + \tau_{\mathbf{i}3}^{(s)} + \tau_{\mathbf{i}4}^{(s)} \right)^{-1}, \end{aligned}$$

$$\begin{aligned} \beta^{(s+1)} = & \left\{ \sum_{\mathbf{i}=1}^{N} \tau_{\mathbf{i}2}^{(s)} (X^{\mathbf{i},1} + X^{\mathbf{i},2}) + \tau_{\mathbf{i}3}^{(s)} (X^{\mathbf{i},1} + X^{\mathbf{i},3}) + \tau_{\mathbf{i}4}^{(s)} (X^{\mathbf{i},2} + X^{\mathbf{i},3}) + \tau_{\mathbf{i}5}^{(s)} (X^{\mathbf{i},1} + X^{\mathbf{i},2} + X^{\mathbf{i},3}) \right\} \\ & \times \left(\sum_{\mathbf{i}=1}^{N} 2\tau_{\mathbf{i}2}^{(s)} + 2\tau_{\mathbf{i}3}^{(s)} + 2\tau_{\mathbf{i}4}^{(s)} + 3\tau_{\mathbf{i}5}^{(s)} \right)^{-1}. \end{aligned}$$

The sum over all the N possible triplets reduces in fact to a sum over eight different possible patterns for the values of $\mathbb{X}^{\mathbf{i}}$. Indeed, the posterior probabilities $\tau_{\mathbf{i}k}$ are constant across triplets with the same observed value.

5.2. Unravelling the latent structure

The general method that we develop in this section aims at recovering the latent structure $\{Z_i\}_{1 \leqslant i \leqslant n}$ on the graph nodes. Indeed, the procedures that were developed in the previous sections focus only on estimating the parameters and do not directly provide an estimate for the node groups.

We rely here on a simple method: we plug in the estimators that were obtained from the previous sections in the complete-data likelihood of the model (namely the likelihood of the observations $\{X_{ij}\}_{1 \leqslant i < j \leqslant n}$ and the latent classes $\{Z_i\}_{1 \leqslant i \leqslant n}$). As we do not have estimates of the mixture proportions $\boldsymbol{\pi}$, we simply remove this part from the expression of the complete-data

likelihood. Then, we simply maximize this criterion (which we call a classification likelihood) with respect to the latent structure $\{Z_i\}_{1\leqslant i\leqslant n}$. In a latter step, we then estimate the unknown proportions $\boldsymbol{\pi}$ by the frequencies that are observed on the estimated groups $\hat{Z}_i$.

5.2.1. Criterion in the binary case

In this set-up, we introduce a criterion $\mathcal{C}$, which is built on the complete-data likelihood, where we plugged in the estimators of α and β and removed the dependence on $\boldsymbol{\pi}$. This criterion is simply written

$$\begin{aligned}\mathcal{C}(\{Z_i\}_{1\leqslant i\leqslant n}) = &\sum_{1\leqslant i<j\leqslant n}\sum_{q=1}^{Q} Z_{iq}Z_{jq}\{X_{ij}\log(\hat{\alpha})+(1-X_{ij})\log(1-\hat{\alpha})\}\\ &+\sum_{1\leqslant i<j\leqslant n}\sum_{1\leqslant q\neq l\leqslant Q} Z_{iq}Z_{jl}\{X_{ij}\log(\hat{\beta})+(1-X_{ij})\log(1-\hat{\beta})\}.\end{aligned}$$

5.2.2. Criterion in the weighted case

Recall that the estimation procedure from Section 4 recovers only the set of unordered values $\{\theta_{\text{in}},\theta_{\text{out}}\}$. As we know these parameters up to permutation only, let $\{\hat{\theta}_1,\hat{\theta}_2\}$ be any choice of label for the corresponding estimators. We can consider two different criteria, denoted $\mathcal{C}^{1,2}$ and $\mathcal{C}^{2,1}$, as follows:

$$\begin{aligned}\mathcal{C}^{u,v}(\{Z_i\}_{1\leqslant i\leqslant n}) = &\sum_{\substack{1\leqslant i<j\leqslant n\\ 1\leqslant q\neq l\leqslant Q}} Z_{iq}Z_{jl}(\mathbf{1}_{X_{ij}\neq 0}[\log\{f(X_{ij};\hat{\theta}_u)\}+\log(\hat{p}_{ql})]+\mathbf{1}_{X_{ij}=0}\log(1-\hat{p}_{ql}))\\ &+\sum_{\substack{1\leqslant i<j\leqslant n\\ 1\leqslant q\leqslant Q}} Z_{iq}Z_{jq}(\mathbf{1}_{X_{ij}\neq 0}[\log\{f(X_{ij};\hat{\theta}_v)\}+\log(\hat{p}_{qq})]+\mathbf{1}_{X_{ij}=0}\log(1-\hat{p}_{qq})),\end{aligned}$$

where $\{u,v\}=\{1,2\}$. For each of these criteria, we can select the latent structure $\hat{Z}^{u,v}=(\hat{Z}_1,\ldots,\hat{Z}_n)^{u,v}$ maximizing it. Then, choosing the couple (u^*,v^*) maximizing the resulting quantity $\mathcal{C}^{u,v}(\hat{Z}^{u,v})$ seems to be an interesting strategy. We thus finally define our estimated latent structure $(\hat{Z}_1,\ldots,\hat{Z}_n)$ as $\hat{Z}^{u^*,v^*}$.

5.2.3. Iterative estimation of the latent structure

In any case (either binary or weighted), we propose to use an iterative procedure to compute the maximum $\hat{Z}$ of the criterion $\mathcal{C}(\{Z_i\}_i)$. Starting from an initial value $Z^{(1)}=(Z_1^{(1)},\ldots,Z_n^{(1)})$ of the latent structure, we iterate the following steps. At step s, we (uniformly) choose a node i_0 and select $Z_{i_0}^{(s+1)}$ as

$$Z_{i_0}^{(s+1)} = \arg\max_{1\leqslant q\leqslant Q}[\mathcal{C}(\{Z_i^{(s)}\}_{i\neq i_0}, Z_{i_0}=q)]$$

and let $Z_j^{(s+1)}=Z_j^{(s)}$ for $j\neq i_0$. At each time step, we increase the classification likelihood $\mathcal{C}(\{Z_i\}_{1\leqslant i\leqslant n})$ and thus the procedure eventually converges to a (local) maximum. By using different initial values $Z^{(1)}=(Z_1^{(1)},\ldots,Z_n^{(1)})$, we should finally find the global maximum. Once we estimated the latent groups $\hat{Z}_i$, we may obtain an estimate of the group proportions $\hat{\boldsymbol{\pi}}$ by simply taking the corresponding frequencies. The procedure is summarized in function `latent.structure` (graph, parameters):

input: observed graph and parameter values;
output: latent structure and group proportions.

Start from latent structure $\{Z_i\}$:
while convergence is not attained,
  choose node i_0 and
replace Z_{i_0} with $\arg\max_q[\mathcal{C}(\{Z_i\}_{i\neq i_0}, Z_{i_0}=q)]$.
Compute group proportions $\boldsymbol{\pi}$ from $\{Z_i\}$.

5.3. Description of complete algorithm

The following algorithm describes the procedures for analysing binary or weighted random graphs. We introduce a variable 'method' which can take three different values: 'moments' or 'tripletEM' in the binary case and 'weighted' for weighted random graphs. In the weighted case, we moreover use a second variable called 'sparsity' to indicate whether we estimate a global sparsity parameter p ('sparsity=global') or two parameters α and β from an affiliation structure ('sparsity=affiliation'). The performances of the procedures proposed in the current section are illustrated in the following section.

If method='moments' *then*
  compute $\hat{m}_i$, $i=1,2,3,$ from expression (9).
    Initialization—
  start from latent structure $\{Z_i\}$ with proportions $\boldsymbol{\pi}$ and compute s_2 and s_3:
  while convergence is not attained,
    update parameters—
  if $\text{abs}\,(\hat{m}_2-\hat{m}_1^2)<\varepsilon$ *then*
    compute α,β through expression (8);
  otherwise
    compute α,β through expression (7);
    update latent structure—
  apply `latent.structure` to the current parameter values.

If method = 'tripletEM' *then*
  estimate α,β from the EM algorithm with triplets (Section 5.1),
  apply `latent.structure` to the parameter values.

If method = 'weighted' *then*—
  sparsity parameters—
  transform weights X_{ij} into binary variables $Y_{ij}=\mathbf{1}_{X_{ij}\neq 0}$;
  if sparsity='global' *then*
    compute $\hat{p}=\{2/n(n-1)\}\Sigma_{i<j}Y_{ij}$;
  otherwise
  estimate α,β from the EM algorithm with triplets (Section 5.1)—
    connectivity parameters—
  estimate $\{\theta_{\text{in}},\theta_{\text{out}}\}$ from the EM algorithm with present edges (Section 4)—
    latent structure—
  apply `latent.structure` to the parameter values.

6. Numerical experiments

We carried out a simulation study to examine the bias and variance of the estimators proposed. In the binary affiliation model, we also compared the performance of our proposal with the variational EM (VEM) strategy that was proposed by Daudin *et al.* (2008). Note that Gibbs sampling has already been compared with VEM strategies in Zanghi *et al.* (2010) and gave

very similar results. Note also that the weighted affiliation model that is proposed here is original and we thus cannot compare our results in this case with any other existing implemented method.

6.1. Binary affiliation model: simulations set-up

In these experiments, we assumed that edges are distributed according to the binary affiliation model that was described in Section 3. The data were generated in various settings, with the number of groups $Q \in \{2, 5\}$ and the number of vertices $n \in \{20, 50, 100, 500, 1000\}$. For each of these cases, we created three settings corresponding to models with different ratios of intragroup and intergroup connectivity parameters (Fig. 2). Moreover, we considered two different cases: equal or free group proportions π.

In each of these settings, we applied three different methods for estimating the model parameters: the moment method (corresponding to Section 3.1), the triplet EM method (corresponding to Section 3.2) and the VEM strategy proposed by Daudin *et al.* (2008), which we adapted to constrain it to an affiliation structure. The results for equal or free group proportions were similar and we thus present only the equal group proportions case.

Fig. 3 shows the estimated density (over 100 graphs simulations) of the estimators $\hat{\alpha}$ and $\hat{\beta}$ for the three algorithms and the three models for graphs with 500 vertices. We see that for a given model the three methods produce estimators with similar densities. In particular, the estimators of α and β seem to have little or no bias and the variances are of the same order of magnitude for the three estimation methods. As the behaviours of the estimators of α and β are comparable over all the simulations, we focus the discussion on the estimation of the parameter α.

Figs 4(a)–4(c) display the estimations of α averaged over 100 graph simulations as a function of the number of graph vertices on a log-scale. For all three models, we see that all the algorithms produce unbiased estimation when the number of vertices is sufficiently large. In addition to the asymptotically unbiased estimation, we observe agreement in the sign of the bias among all algorithms, when the graphs are small. For example, when estimating α in model 1 where $(\alpha = 0.3, \beta = 0.03)$, all methods underestimate α and overestimate β.

To compare the dispersion of all the estimators, we consider their empirical standard deviation computed over 100 simulations. Figs 4(d)–4(f) show the evolution of the logarithm of the empirical standard deviation of $\hat{\alpha}$ when the size of the graphs grows from 20 vertices up to 1000 vertices. We see a linear dependence between the logarithm of the graph size and the log-standard-

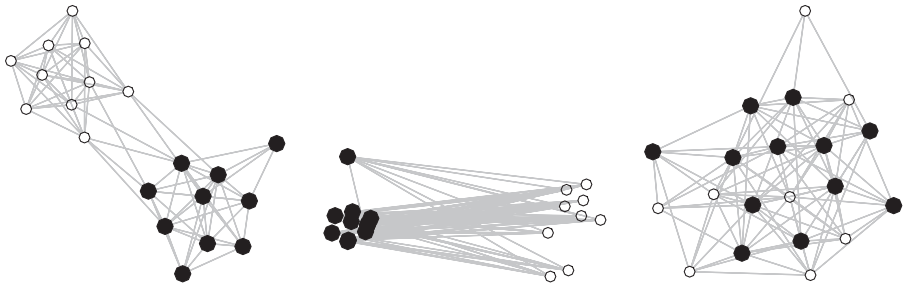

Fig. 2. (a) Low intergroup connectivity and strong intragroup connectivity (model 1; $\alpha = 0.3$ and $\beta = 0.03$), (b) strong intergroup connectivity and low intragroup connectivity (model 2; $\alpha = 0.03$ and $\beta = 0.3$) and (c) model without structure close to the Erdős–Rényi random-graph model (model 3; $\alpha = 0.55$ and $\beta = 0.45$): the figure displays an example with $Q = 2$ groups

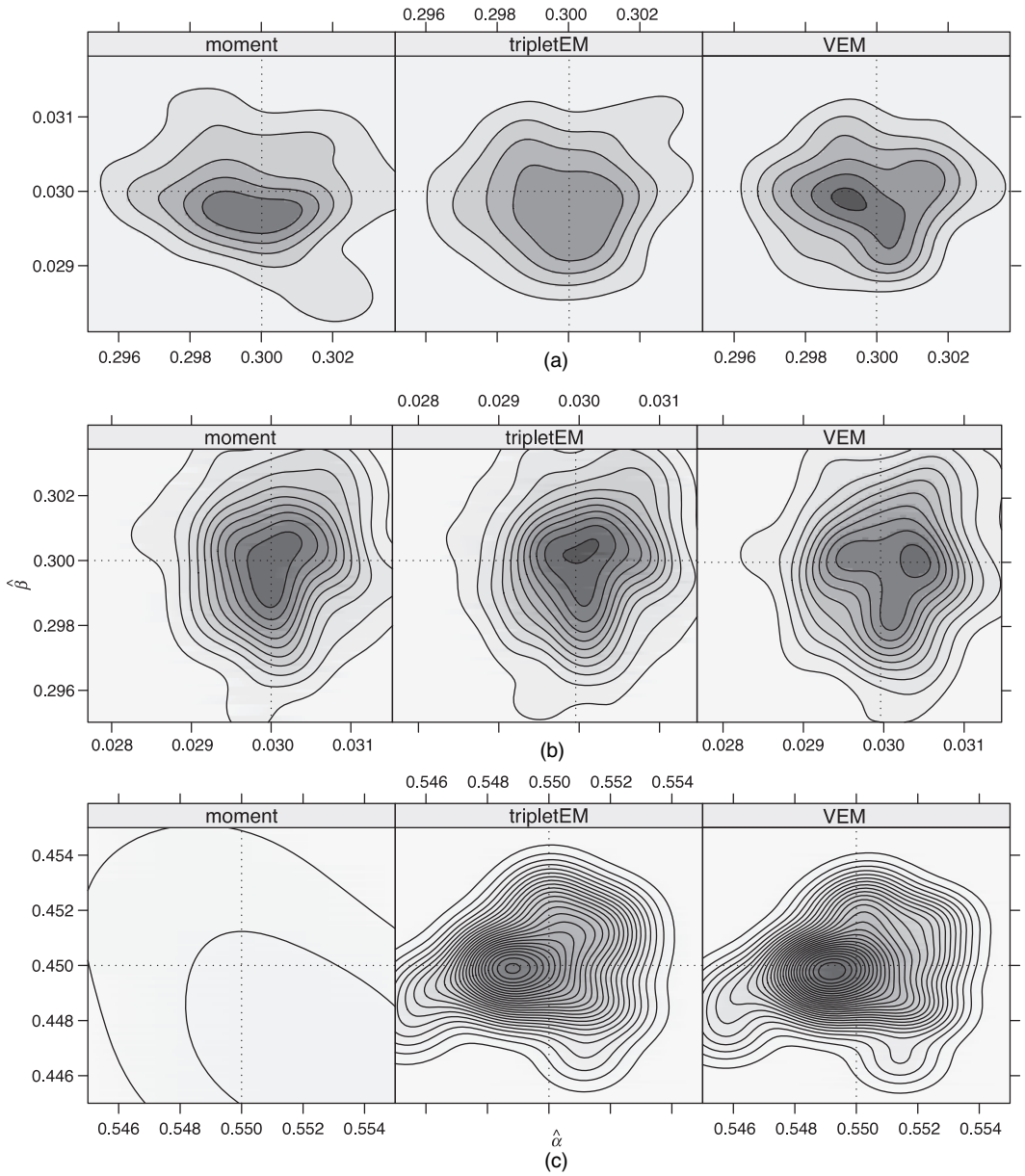

Fig. 3. Empirical joint distribution of the estimators $\hat{\alpha}$ and $\hat{\beta}$, computed over 100 simulations of graphs with 500 vertices, $Q = 2$ groups and equal group proportions (, true values of α and β): (a) model 1; (b) model 2; (c) model 3

deviation. The slope of the lines is about -1, which indicates that the standard deviation decreases with rate of the order $1/n$ (where n is the number of vertices of the graph). The differences between the intercepts relate to constant factors driving the relationships between all rates of convergence. When $Q = 2$, we observe very similar intercepts for all methods, both for model 1 and for model 2. When $Q = 5$, the VEM algorithm appears to converge faster but the orders of the standard deviations remain comparable among all estimation methods. For

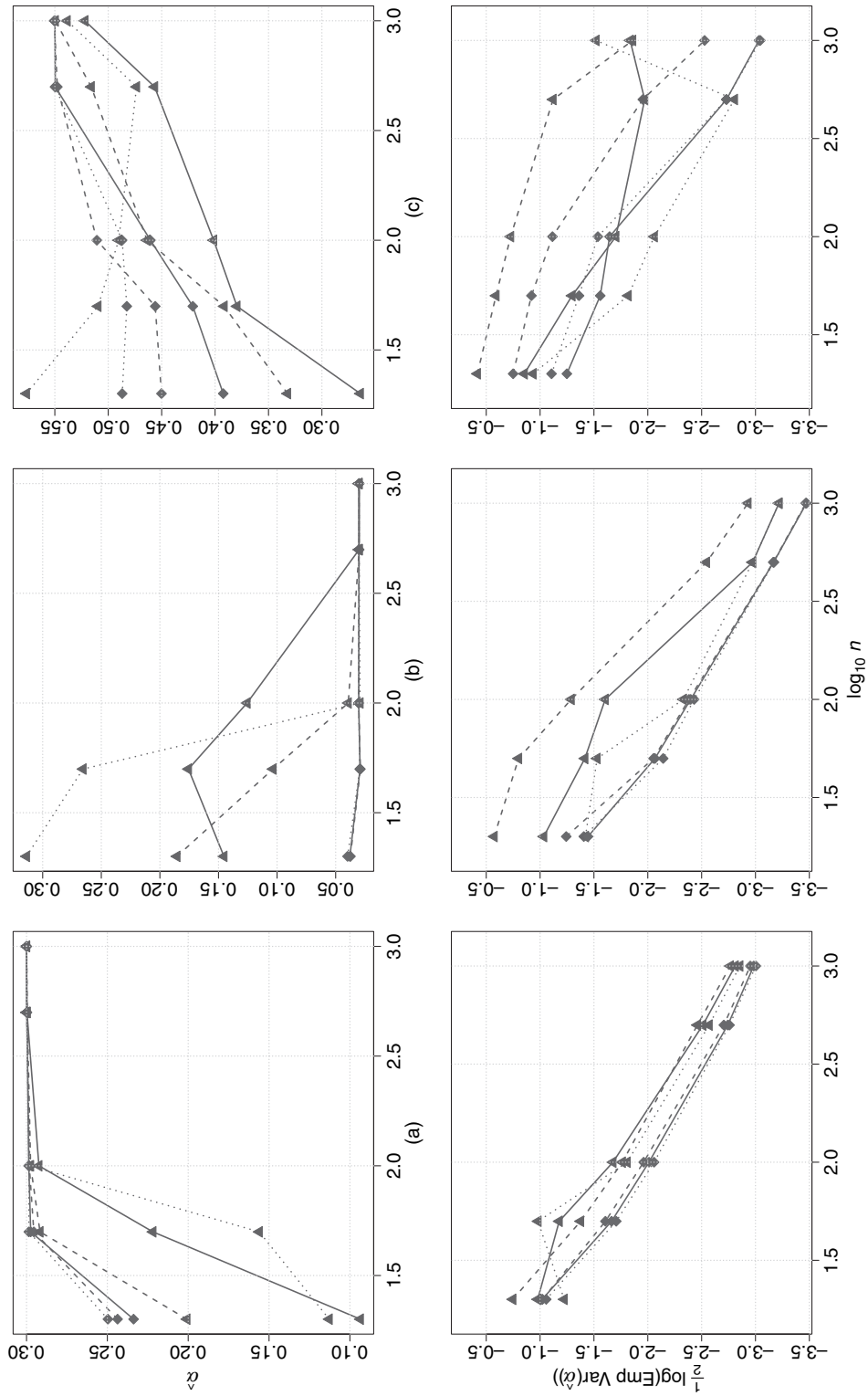

Fig. 4. (a)–(c) $\hat{\alpha}$ and (d)–(f) the logarithm of its empirical standard deviation averaged over 100 graph simulations, as functions of the number of graph vertices on a log-scale for equal group proportions (··◆··, $Q = 2$, VEM; ··▲··, $Q = 5$, VEM; -◆-, $Q = 2$, moment; -▲-, $Q = 5$, moment; —◆—, $Q = 2$, tripletEM; —▲—, $Q = 5$, tripletEM): (a), (d) model 1; (b), (e) model 2; (c), (f) model 3

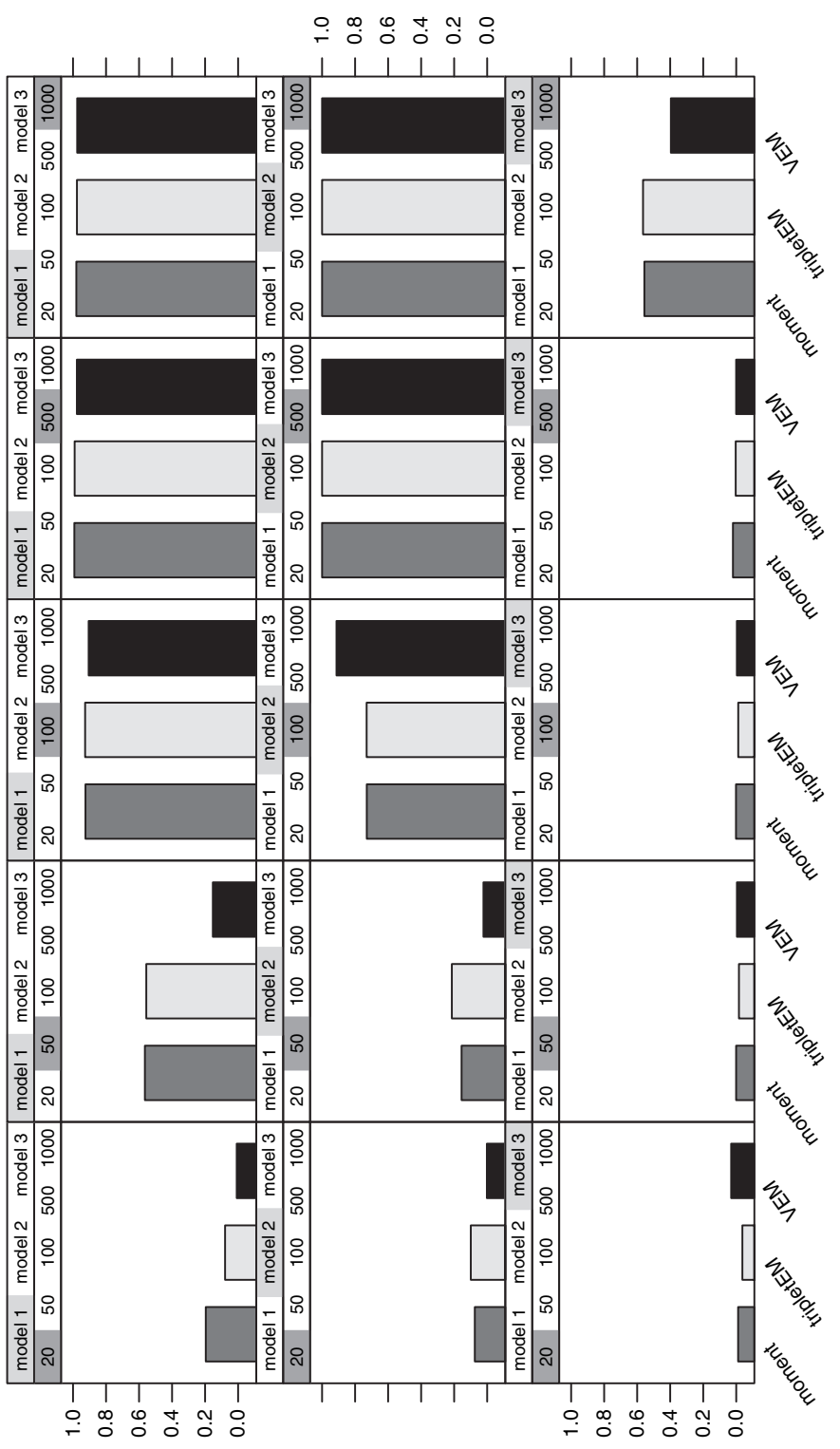

Fig. 5. Evaluation of the Rand index to compare the true and estimated latent structure of the binary affiliation graph

model 3, the moment-based estimations have greater dispersion but still decrease with the same rate.

We use the adjusted Rand index (Hubert and Arabie, 1985) to evaluate the agreement between the estimated and the true latent structure. The computation of the Rand index is based on a ratio between the number of node pairs belonging to the same and to different classes when considering the actual latent structure *versus* the estimated structure. This index lies between 0 and 1, two identical latent structures having an adjusted Rand index equal to 1. Fig. 5 displays the Rand index for the three models and five different graph sizes. It appears that the three algorithms allow a reasonable recovery of the latent structure, for models 1 and 2, when the graphs considered have more than 100 vertices. As expected, the larger the number of nodes, the better the recovery of the latent structure that we observe. We also note that our proposed strategy for recovering the latent structure performs as well as or better than the variational approach in all cases.

The previous experiments show that the two estimation procedures proposed in this work behave as well as or better than the variational-based algorithm, both for the parameter estimation and for the recovery of the latent structure. Note also that the moment-based method does not depend on any sort of initialization, since it relies on the analytical resolution of a simple system based on triads (order 3 structures).

6.2. Weighted affiliation model: simulations set-up

In the following experiments, we use a sparsity parameter constant across the graph and non-missing edges are distributed according to a Gaussian model as described in Section 4, with different means μ_{in} and μ_{out} and equal variance σ^2. The intricacy of a model is inversely related to the 'distance' between the parameters θ_{in} and θ_{out}. We use the Mahalanobis distance $\Delta = |(\mu_{\text{in}} - \mu_{\text{out}})/\sigma|$. Three models are considered with different levels of intricacy: we fix $\mu_{\text{in}} = 2$ and $\mu_{\text{out}} = 1$; thus $\Delta = |(\mu_{\text{in}} - \mu_{\text{out}})/\sigma| = 1/\sigma$, which takes the values $\Delta = 10$ (model A), $\Delta = 2$ (model B) and $\Delta = 1$ (model C). We fix the number of groups $Q = 2$, take equal group proportions and consider various numbers of vertices $n \in \{20, 100, 500, 1000\}$.

We computed bias and empirical standard deviations over 100 simulations. As illustrated by Fig. 6(a) in the case of $\hat{\mu}_{\text{in}}$, the method recovers the parameters with no bias, except for model C where a small bias occurs due to the high level of intricacy of the model. Fig. 6(b) displays the evolution of the logarithm of the empirical standard deviation of $\hat{\mu}_{\text{in}}$ when the size of the graphs grows from 20 vertices up to 1000 vertices. As for the binary affiliation model estimators, we observe a linear dependence between the logarithm of the graph size and the log-standard-deviation, the slope of the lines lying in $[-\frac{1}{2}, -1]$.

Fig. 6(b) displays the Rand index for the three different models (A,B,C) and four different graph sizes. When graphs have more than 100 nodes, recovery of the hidden structure is almost perfect in all situations as previously observed in the binary case.

The previous experiments show that, when dealing with weighted affiliation graphs, the estimation of the parameters and of the graph latent structure can be efficiently achieved considering only edges (order 2 structures).

6.3. Cross-citations of economics journals

We illustrate the difference between weighted and binary models for graph clustering by using a real data example. We consider cross-citations of 42 economics journals over the years 1995–1997 (Pieters and Baumgartner, 2002). The raw data correspond to a weighted non-symmetric graph where vertices are journals and directed edges the number of citations from one journal

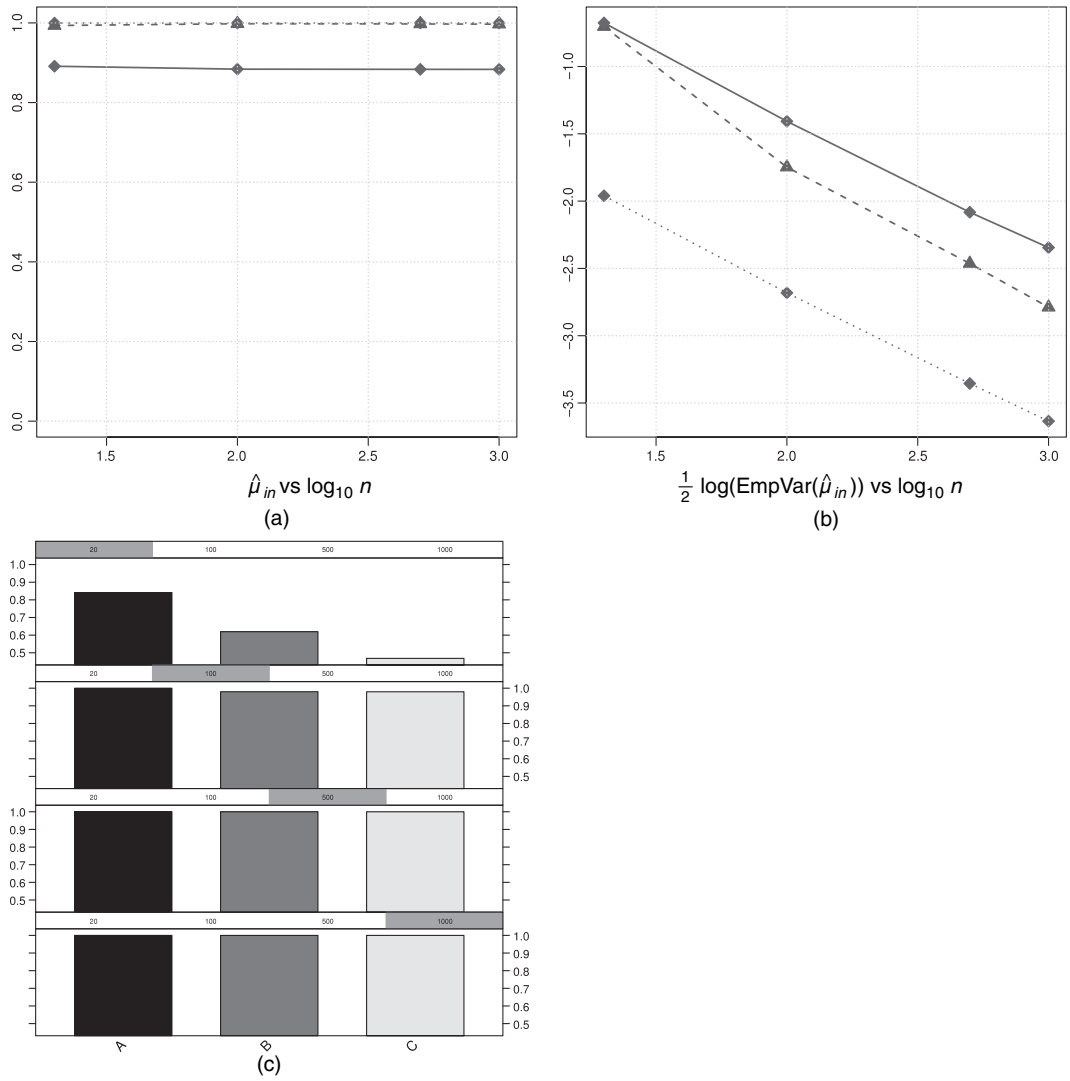

Fig. 6. Evaluation of the estimation of the parameters of a weighted graph: (a) $\hat{\mu}_{\text{in}}$ and (b) logarithm of its empirical standard deviation, both as functions of the number of graph vertices, expressed on a log-scale (each estimation is averaged over 100 graph simulations; ··♦··, model A; ▲, model B; –♦–, model C) and (c) Rand index computation comparing the true latent structure with the estimated one for the three models (columns A, B and C) and four graph sizes (rows $n = 20$,100, 500,1000)

to another. We first take the mean value of citations between each pair of journals (leading to a symmetric adjacency matrix) and work with its normalized Laplacian. Fig. 7 displays the affiliation matrices structured according to a partition in four classes. Clustering based on the binary model and on the weighted model (respectively Fig. 7(a) and Fig. 7(b)) exhibit very different cluster structures. The binary model finds classes which tend to be homogeneous in terms of probability of intragroup and intergroup connections, whereas the weighted model finds classes which are homogeneous in terms of intragroup and intergroup connection weights. This distinction results in completely different interpretations.

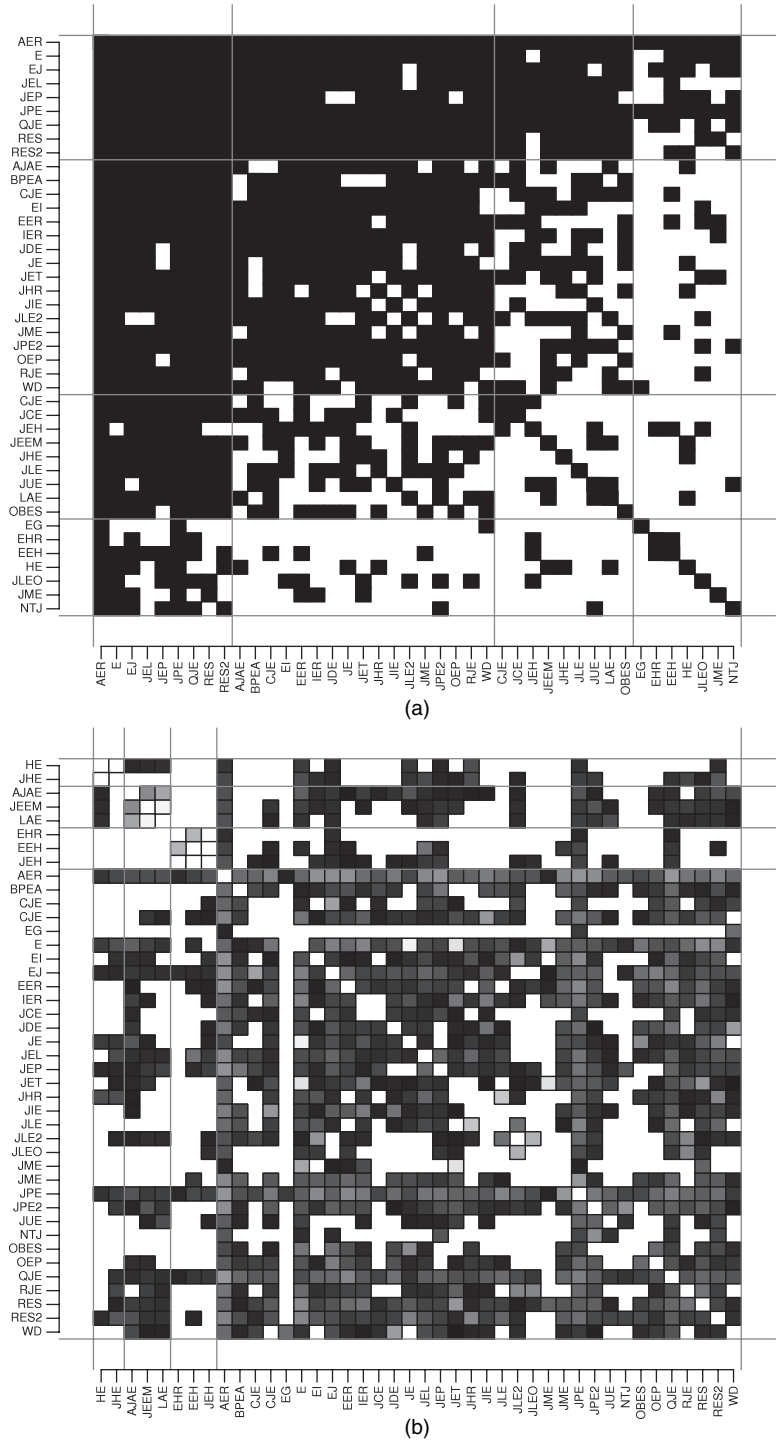

Fig. 7. Matrices of cross-citations between 42 economics journals with rows and columns reorganized according to groups found by the binary random-graph mixture model (a) compared with groups found with the weighted random-graph mixture model (b)

The binary model finds two groups of nodes which are strongly connected within their groups but also with nodes from the other groups. It also exhibit two other smaller classes with low intragroup connectivity and nodes that preferentially link to the first class which plays the role of a reference class. Indeed the first class (top left) found by the binary model is composed of journals with high impact factors: the *American Economic Review*, AER, *Econometrica*, E, the *Journal of Economic Literature*, JEL, the *Journal of Economic Perspectives*, JEP, the *Journal of Political Economy*, JPE, the *Quarterly Journal of Economics*, QJE, the *Review of Economic Studies*, RES, and the *Review of Economics and Statistics*, RES2.

The result produced by the weighted model shows a main class of strongly interconnected journals and three smaller classes of journals, which weakly cross-cite each other:

(a) *class 1* (health), *Health Economics*, HE, and the *Journal of Health Economics*, JHE;
(b) *class 2* (natural resources), the *Journal of Agricultural Economics*, AJAE, *Land Economics*, LAE, and the *Journal of Environmental Economics and Management*, JEEM;
(c) *class 3* (economic history), *Exploration of Economic History*, EEH, the *Journal of Economic History*, JEH, and the *Economic History Review*, EHR.

Each of these three classes is composed of journals that are dedicated to similar topics (respectively health, natural resources and economic history). They preferentially cite journals from the first class which contains journals with less specific topics.

Acknowledgements

The authors thank Jérome Dedecker for helpful insights concerning this work as well as the Associate Editor and two referees for their remarks leading to considerable improvements to the manuscript. The authors have been supported by the French Agence Nationale de la Recherche under grant Networks Motifs ANR-08-BLAN-0304-01.

Appendix A: Proofs

A.1. Proof of theorem 1

To facilitate the reading of the proof, we decompose it into several stages.

A.1.1. Preliminaries

We fix $k, s \geqslant 1$ and $p = \binom{k}{2}$. Recall that $\mathcal{V}_Q$ is the set of Q-size vectors such that, for any $v = (v_1, \ldots, v_Q) \in \mathcal{V}_Q$, we have $v_i \in \{0, 1\}$ and $\sum_{i=1}^{Q} v_i = 1$. We also let $\mathcal{Q} = \{1, \ldots, Q\}$. We then consider the set

$$\mathcal{Z} = \left\{ z \in \mathcal{V}_Q^{\mathbb{N}}; \forall\, \mathbf{q} = (q_1, \ldots, q_k) \in \mathcal{Q}^k, \frac{(n-k)!}{n!} n_{\mathbf{q}} := \frac{(n-k)!}{n!} \sum_{\mathbf{i} \in \mathcal{I}_k} \prod_{l=1}^{k} z_{i_l q_l} \underset{n \to \infty}{\longrightarrow} \prod_{l=1}^{k} \pi_{q_l} \right\}.$$

Moreover, we let $N_{\mathbf{q}} = \Sigma_{\mathbf{i} \in \mathcal{I}_k} \Pi_{l=1}^{k} Z_{i_l q_l}$. The strong law of large numbers gives the almost sure convergence, as $n \to \infty$, of $\{(n-k)!/n!\} N_{\mathbf{q}}$ to $\Pi_{l=1}^{k} \pi_{q_l}$. This implies that $\mathbb{P}(\{Z_n\}_{n \geqslant 1} \in \mathcal{Z}) = 1$.

A.1.2. Consistency of $\hat{m}_g$

We first introduce the conditional mean of $g(\mathbb{X}^{\mathbf{i}})$ given that the hidden groups at position $\mathbf{i}$ are given by $\mathbf{q}$

$$m_g(\mathbf{q}) = \mathbb{E}\left\{ g(\mathbb{X}^{\mathbf{i}}) \,\middle|\, \prod_{l=1}^{k} Z_{i_l q_l} = 1 \right\}.$$

Using the equalities

$$\begin{aligned} &\forall \mathbf{i} \in \mathcal{I}_k, \ \sum_{\mathbf{q} \in \mathcal{Q}^q} \prod_{l=1}^{k} Z_{i_l q_l} = 1, \\ &m_g = \sum_{\mathbf{q} \in \mathcal{Q}^k} \left(\prod_{l=1}^{k} \pi_{q_l} \right) m_g(\mathbf{q}), \end{aligned} \tag{20}$$

we may write the decomposition

$$\begin{aligned} \hat{m}_g - m_g &= \frac{(n-k)!}{n!} \sum_{\mathbf{q} \in \mathcal{Q}^k} \sum_{\mathbf{i} \in \mathcal{I}_k} \prod_{l=1}^{k} Z_{i_l q_l} \, g(\mathbb{X}^{\mathbf{i}}) - \sum_{\mathbf{q} \in \mathcal{Q}^k} \left(\prod_{l=1}^{k} \pi_{q l} \right) m_g(\mathbf{q}) \\ &= \sum_{\mathbf{q} \in \mathcal{Q}^k} \left[\frac{(n-k)!}{n!} \sum_{\mathbf{i} \in \mathcal{I}_k} \left(\prod_{l=1}^{k} Z_{i_l q_l} \right) \{ g(\mathbb{X}^{\mathbf{i}}) - m_g(\mathbf{q}) \} + m_g(\mathbf{q}) \left\{ \frac{(n-k)!}{n!} N_{\mathbf{q}} - \prod_{l=1}^{k} \pi_{q_l} \right\} \right]. \end{aligned} \tag{21}$$

To establish the consistency of $\hat{m}_g$, we rely on a conditioning argument. Let A be the event $\limsup_{n \to \infty} |\hat{m}_g - m_g| = 0$. We then have

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{E}\{\mathbf{1}_A | \{Z_n\}_{n \geqslant 1}\}]. \tag{22}$$

Now, conditional on $\{Z_n\}_{n \geqslant 1} = z$, the random variables $\{\mathbb{X}^{\mathbf{i}}; \mathbf{i} \in \mathcal{I}_k, \Pi_{l=1}^{k} z_{i_l q_l} = 1\}$ form an $n_{\mathbf{q}}$-sample of independent and identically distributed random variables. Letting B be the event

$$\limsup_{n \to \infty} \left[\frac{1}{N_{\mathbf{q}}} \left| \sum_{\mathbf{i} \in \mathcal{I}_k; \Pi_{l=1}^{k} Z_{i_l q_l} = 1} \{ g(\mathbb{X}^{\mathbf{i}}) - m_g(\mathbf{q}) \} \right| \right] = 0,$$

the strong law of large numbers yields that, for any $z \in \mathcal{Z}$,

$$\mathbb{E}\{\mathbf{1}_B | \{Z_n\}_{n \geqslant 1} = z\} = 1.$$

Conditional on $\{Z_n\}_{n \geqslant 1} = z \in \mathcal{Z}$, we may thus rewrite the decomposition (21) as

$$\hat{m}_g - m_g = \sum_{\mathbf{q} \in \mathcal{Q}^k} \left[\frac{(n-k)!}{n!} n_{\mathbf{q}} \times \frac{1}{n_{\mathbf{q}}} \sum_{\mathbf{i} \in \mathcal{I}_k; \Pi_{l=1}^{k} z_{i_l q_l} = 1} \{ g(\mathbb{X}^{\mathbf{i}}) - m_g(\mathbf{q}) \} + m_g(\mathbf{q}) \left\{ \frac{(n-k)!}{n!} n_{\mathbf{q}} - \prod_{l=1}^{k} \pi_{q_l} \right\} \right],$$

which establishes that, for any $z \in \mathcal{Z}$, we have $\mathbb{E}\{\mathbf{1}_A | \{Z_n\}_{n \geqslant 1} = z\} = 1$. Coming back to equation (22), we thus obtain

$$\mathbb{P}\{ \lim_{n \to \infty} (\hat{m}_g) = m_g \} = 1.$$

A.1.3. Asymptotic normality of $\hat{m}_g$.
We now prove a central limit result for $\sqrt{n}(\hat{m}_g - m_g)$. First, the central limit theorem applied to the Q-size vector $\Sigma_{i=1}^{n} (Z_i - \boldsymbol{\pi})/\sqrt{n}$ gives the convergence

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \boldsymbol{\pi}) \rightsquigarrow \mathcal{N}(0, \Sigma), \qquad \text{as } n \to \infty, \tag{23}$$

where $\Sigma_{qq} = \pi_q(1 - \pi_q)$ and $\Sigma_{ql} = -\pi_q \pi_l$ when $q \neq l$.

Now, consider the second term appearing on the right-hand side of decomposition (21). To establish a central limit theorem for $N_{\mathbf{q}}$, we decompose the sum of products

$$\sum_{\mathbf{i} \in \mathcal{I}_k} \prod_{l=1}^{k} Z_{i_l q_l} = \sum_{\mathbf{i} \in \mathcal{I}_k} \prod_{l=1}^{k} (Z_{i_l q_l} - \pi_{q_l} + \pi_{q_l})$$

into sums of products of centred terms $Z_{i_l q_l} - \pi_{q_l}$. This leads to

$$\frac{(n-k)!}{n!} N_{\mathbf{q}} - \prod_{l=1}^{k} \pi_{q_l} = \sum_{u=1}^{k} \sum_{L \subset \{1, \ldots, k\}; |L| = u} \frac{(n-u)!}{n!} \left(\prod_{l \notin L} \pi_{q_l} \right) \sum_{\mathbf{i} \in \mathcal{I}_L} \prod_{l \in L} (Z_{i_l q_l} - \pi_{q_l}),$$

where $|L|$ denotes the cardinality of the set L and $\mathcal{I}_L$ denotes the set of injective maps from L to $\mathcal{I} = \{1, \ldots, n\}$. In this expression, the leading term (obtained for singleton sets L, i.e. when $u = 1$) gives the rate of convergence in the central limit theorem. In other words,

$$\left\{\frac{(n-k)!}{n!}N_{\mathbf{q}}-\prod_{l=1}^{k}\pi_{q_l}\right\}\sqrt{n}=\sum_{l=1}^{k}\left(\prod_{j\neq l}\pi_{q_j}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_{iq_l}-\pi_{q_l})$$
$$+\sum_{u=2}^{k}\sum_{L\subset\{1,\ldots,k\};|L|=u}\frac{\sqrt{n}(n-u)!}{n!}\left(\prod_{l\notin L}\pi_{q_l}\right)\sum_{\mathbf{i}\in\mathcal{I}_L}\prod_{l\in L}(Z_{i_lq_l}-\pi_{q_l}). \tag{24}$$

The first term on the right-hand side of equation (24) converges to a linear combination of the co-ordinates of an $\mathcal{N}(0,\Sigma)$ vector, whereas the second term converges to 0. Indeed, for any value $u\geqslant 2$ and any set L of cardinality u, we may write

$$\frac{\sqrt{n}(n-u)!}{n!}\sum_{\mathbf{i}\in\mathcal{I}_L}\prod_{l\in L}(Z_{i_lq_l}-\pi_{q_l})=\frac{1}{\sqrt{n}(n-1)\ldots(n-u+1)}\sum_{\mathbf{i}\in\mathcal{I}_L}\prod_{l\in L}(Z_{i_lq_l}-\pi_{q_l})$$

which converges to 0. Thus we obtain

$$\sqrt{n}\sum_{\mathbf{q}\in\mathcal{Q}^k}m_g(\mathbf{q})\left\{\frac{(n-k)!}{n!}N_{\mathbf{q}}-\prod_{k=1}^{l}\pi_{q_l}\right\}=\sum_{\mathbf{q}\in\mathcal{Q}^k}m_g(\mathbf{q})\left\{\sum_{l=1}^{k}\frac{\prod_{j\neq l}\pi_{q_j}}{\sqrt{n}}\sum_{i=1}^{n}(Z_{iq_l}-\pi_{q_l})+R_{n,\mathbf{q}}\right\},$$

where $R_{n,\mathbf{q}}=o_P(1)$ are negligible terms converging in probability to 0, as $n\to\infty$. According to expression (23), we obtain that

$$\sqrt{n}\sum_{\mathbf{q}\in\mathcal{Q}^k}m_g(\mathbf{q})\left\{\frac{(n-k)!}{n!}N_{\mathbf{q}}-\prod_{l=1}^{k}\pi_{q_l}\right\}\underset{n\to\infty}{\rightsquigarrow}\sum_{\mathbf{q}\in\mathcal{Q}^k}m_g(\mathbf{q})\sum_{l=1}^{k}\left(\prod_{j\neq l}\pi_{q_j}\right)W_{q_l},$$

where $W=(W_1,\ldots,W_Q)\sim\mathcal{N}(0,\Sigma)$.

To obtain a central limit theorem for $\hat{m}_g$ it now suffices to prove that the first term on the right-hand side of equation (21) is negligible, when scaled by the rate of convergence $\sqrt{n}$. Indeed, we may write this term as

$$\tilde{R}_n=\sum_{\mathbf{q}\in\mathcal{Q}^k}\left\{\frac{(n-k)!}{(n-1)!}\right\}^{1/2}\left\{\frac{(n-k)!}{n!}N_{\mathbf{q}}\right\}^{1/2}\frac{1}{\sqrt{N_{\mathbf{q}}}}\sum_{\mathbf{i}\in\mathcal{I}_k;\Pi_l Z_{i_lq_l}=1}\{g(\mathbb{X}^{\mathbf{i}})-m_g(\mathbf{q})\},$$

which satisfies, for any $k\geqslant 2$, any $\varepsilon>0$ and any $z\in\mathcal{Z}$,

$$\mathbb{P}(|\tilde{R}_n|\geqslant\varepsilon|\{Z_n\}_{n\geqslant 1}=z)\underset{n\to\infty}{\longrightarrow}0.$$

Using dominated convergence, we also have $\mathbb{P}(|\tilde{R}_n|\geqslant\varepsilon)\to_{n\to\infty}0$, for any $\varepsilon>0$. Now, going back to equation (21), we finally obtain

$$\sqrt{n}(\hat{m}_g-m_g)\underset{n\to\infty}{\rightsquigarrow}\sum_{\mathbf{q}\in\mathcal{Q}^k}m_g(\mathbf{q})\sum_{l=1}^{k}\left(\prod_{j\neq l}\pi_{q_j}\right)W_{q_l}\sim\mathcal{N}(0,\Sigma_g).$$

A.1.4. Expression for the limiting variance Σ_g

The computation of the variance Σ_g could be done by using the above expression, but this leads to tedious formulae. A simpler expression of the limiting variance is obtained in the following way. We prove that $U_n\sqrt{n}:=(\hat{m}_g-m_g)\sqrt{n}$ has a bounded third-order moment. This is sufficient to claim that Σ_g can be obtained as the limiting variance of $U_n\sqrt{n}$.

First, since non-adjacent edges form independent variates, it is easy to see that we have

$$\mathbb{E}(\|U_n\sqrt{n}\|^3)\leqslant\left\{\frac{(n-k)!}{\sqrt{n}(n-1)!}\right\}^3\sum_{\mathbf{i},\mathbf{j},\mathbf{k}\,\mathbf{i}\cap\mathbf{j}\cap\mathbf{k}\neq\emptyset}\mathbb{E}\{\|g(\mathbb{X}^{\mathbf{i}})-m_g\|\|g(\mathbb{X}^{\mathbf{j}})-m_g\|\|g(\mathbb{X}^{\mathbf{k}})-m_g\|\},$$

where $\mathbf{i}\cap\mathbf{j}$ stands for the intersection of $\mathbf{i}$ and $\mathbf{j}$ viewed as index sets (instead of k-tuples). The above sum contains at most $k^2n\{(n-1)\ldots(n-k+1)\}^3$ terms, which are bounded (there are finitely many of them).

Thus this quantity converges to 0 as $n \to \infty$. Moreover,

$$\mathrm{var}(U_n\sqrt{n}) = \left\{ \frac{(n-k)!}{\sqrt{n}(n-1)!} \right\}^2 \sum_{\mathbf{i},\mathbf{j};\mathbf{i}\cap\mathbf{j}\neq\emptyset} \mathrm{cov}\{g(\mathbb{X}^{\mathbf{i}}), g(\mathbb{X}^{\mathbf{j}})\}.$$

This sum may be decomposed according to the cardinality of the set $\mathbf{i}\cap\mathbf{j}$. It is then easy to see that the dominating term is obtained when $|\mathbf{i}\cap\mathbf{j}|=1$, whereas the other terms converge to 0, namely

$$\mathrm{var}(U_n\sqrt{n}) = \left\{ \frac{(n-k)!}{\sqrt{n}(n-1)!} \right\}^2 \sum_{\mathbf{i},\mathbf{j};\mathbf{i}\cap\mathbf{j}=1} \mathrm{cov}\{g(\mathbb{X}^{\mathbf{i}}), g(\mathbb{X}^{\mathbf{j}})\} + o(1).$$

To describe all the possible configurations where $|\mathbf{i}\cap\mathbf{j}|=1$, we may fix the first index $\mathbf{i}$ to $(1,\ldots,k)$ and let the second index $\mathbf{j}$ describe the set of indices where some position s takes one of the values $\{1,\ldots,k\}$ (corresponding to the intersection $\mathbf{i}\cap\mathbf{j}$) and, at any other position, there is some value in $\{k+1,\ldots,n\}$. For any $s,t\in\{1,\ldots,k\}$, we thus let $\mathbf{e}_s^t \in \mathcal{I}_k$ satisfying $\mathbf{e}_s^t(s)=t$ and $\mathbf{e}_s^t(j)\in\{k+1,\ldots,n\}$ for any $j\neq s$. With this notation, we obtain

$$\Sigma_g = \lim_{n\to\infty}\{\mathrm{var}(\sqrt{n}U_n)\} = \sum_{s=1}^{k}\sum_{t=1}^{k} \mathrm{cov}\{g(\mathbb{X}^{(1,\ldots,k)}), g(\mathbb{X}^{\mathbf{e}_s^t})\}.$$

In the case of an affiliation structure with equal group proportions, we could prove from this expression that $\Sigma_g = 0$ (using for instance the results of lemma 1 that are presented below). Anyway this will be a consequence of the following developments.

A.1.5. Degenerate case

We now finish this proof by considering the specific case where we have an affiliation structure (2) and equal group proportions (3). Coming back to equation (21), we write $\hat{m}_g - m_g = T_1 + T_2$ where

$$T_1 = \sum_{\mathbf{q}\in\mathcal{Q}^k} \frac{(n-k)!}{n!} \sum_{\mathbf{i}\in\mathcal{I}_k} \left(\prod_{l=1}^{k} Z_{i_l q_l} \right) \{ g(\mathbb{X}^{\mathbf{i}}) - m_g(\mathbf{q}) \},$$

$$T_2 = \sum_{\mathbf{q}\in\mathcal{Q}^k} m_g(\mathbf{q}) \left\{ \frac{(n-k)!}{n!} N_{\mathbf{q}} - \prod_{l=1}^{k} \pi_{q_l} \right\}.$$

We first deal with the second term T_2. According to equation (24), we have

$$T_2 = \sum_{\mathbf{q}\in\mathcal{Q}^k} m_g(\mathbf{q}) \sum_{l=1}^{k} \frac{1}{Q^{k-1}} \frac{1}{n} \sum_{i=1}^{n} (Z_{iq_l} - \pi_{q_l}) + \sum_{u=2}^{k} \sum_{L\subset\{1,\ldots,k\};|L|=u} \frac{(n-u)!}{n!} \frac{1}{Q^{k-u}} \sum_{\mathbf{i}\in\mathcal{I}_L} \prod_{l\in L} (Z_{i_l q_l} - \pi_{q_l}) := T_{2,1} + T_{2,2}.$$

We now prove that the first term on the right-hand side of this equality, namely $T_{2,1}$, is 0. This result relies on the following lemma, stating that the model is invariant under a permutation of the values of the node groups.

Lemma 1. Under the assumptions and notation of theorem 1, assuming moreover expressions (2) and (3), for any $\sigma\in\mathcal{S}_Q$ the set of permutations of $\mathcal{Q}$, we have

$$(\{Z_i\}_{1\leqslant i\leqslant n}, \{X_{ij}\}_{1\leqslant i<j\leqslant n}) \overset{\mathrm{d}}{=} (\{\sigma(Z_i)\}_{1\leqslant i\leqslant n}, \{X_{ij}\}_{1\leqslant i<j\leqslant n}),$$

where $=^{\mathrm{d}}$ means equality in distribution. As a consequence, for any value $\mathbf{q}\in\mathcal{Q}^k$, the conditional expectation $m_g(\mathbf{q})$ is constant along the orbit (induced by $\mathcal{S}_Q$) of the point $\mathbf{q}$, i.e. the set of values $\{m_g\{\sigma(\mathbf{q})\}; \sigma\in\mathcal{S}_Q\}$ is a singleton for any fixed $\mathbf{q}\in\mathcal{Q}^k$.

Indeed, according to expressions (1)–(3), and using that any permutation σ is a one-to-one application, we have

$$\begin{aligned}\mathbb{P}(\{Z_i\}_{1\leqslant i\leqslant n}, \{X_{ij}\}_{1\leqslant i<j\leqslant n}) &= \prod_{i=1}^{n} \mathbb{P}(Z_i) \prod_{1\leqslant i<j\leqslant n} \mathbb{P}(X_{ij}|\mathbf{1}_{Z_i=Z_j}) \\ &= \frac{1}{Q^n} \prod_{1\leqslant i<j\leqslant n} \mathbb{P}(X_{ij}|\mathbf{1}_{\sigma(Z_i)=\sigma(Z_j)}) = \mathbb{P}(\{\sigma(Z_i)\}_{1\leqslant i\leqslant n}, \{X_{ij}\}_{1\leqslant i<j\leqslant n}).\end{aligned}$$

As a consequence, for any $\sigma \in \mathcal{S}_Q$ and any value $\mathbf{q} \in \mathcal{Q}^k$, the conditional expectation $m_g\{\sigma(\mathbf{q})\}$ satisfies

$$m_g\{\sigma(\mathbf{q})\} = \mathbb{E}\{g(\mathbb{X}^{(1,\ldots,k)})|(Z_1,\ldots,Z_k) = \sigma(\mathbf{q})\} = \mathbb{E}\{g(\mathbb{X}^{(1,\ldots,k)})|(\sigma(Z_1),\ldots,\sigma(Z_k)) = \sigma(\mathbf{q})\} = m_g(\mathbf{q}).$$

Thus the set of values $\{m_g\{\sigma(\mathbf{q})\}; \sigma \in \mathcal{S}_Q\}$ is reduced to a singleton. This finishes the proof of lemma 1.

Now, going back to the term $T_{2,1}$, the set $\mathcal{Q}^k$ may be partitioned into the disjoint union of the orbits induced by $\mathcal{S}_Q$, namely $\mathcal{Q}^k = \cup_{\mathcal{O}\text{orbit}}\mathcal{O}$, with $\mathbf{q} \to m_g(\mathbf{q})$ being constant on each orbit $\mathcal{O}$. We let $m_{g,\mathcal{O}}$ denote the value of the function $\mathbf{q} \to m_g(\mathbf{q})$ on the orbit $\mathcal{O}$. Then we write

$$T_{2,1} = \frac{1}{nQ^{k-1}} \sum_{\mathcal{O}\text{orbit}} m_{g,\mathcal{O}} \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{\mathbf{q}\in\mathcal{O}} (Z_{iq_l} - \pi_{q_l}).$$

For each orbit $\mathcal{O}$ and any position $l \in \{1,\ldots,k\}$, if we fix some $\mathbf{q} \in \mathcal{O}$, then we argue that $\mathcal{O}$ contains all the points of the form $(q_1,\ldots,q_{l-1},j,q_{l+1},\ldots,q_k)$ for any $1 \leqslant j \leqslant Q$. Indeed, all these points are images of $\mathbf{q}$ by the simple transpositions $(q_l\ j)$. Thus, the sum $\Sigma_{\mathbf{q}\in\mathcal{O}}(Z_{iq_l} - \pi_{q_l})$ contains the sum $\Sigma_{q_l\in\mathcal{Q}}(Z_{iq_l} - \pi_{q_l})$, which is 0. This proves that $T_{2,1} = 0$ and thus

$$n(\hat{m}_g - m_g) = n(T_1 + T_{2,2}) = \sum_{\mathbf{q}\in\mathcal{Q}^k} \frac{(n-k)!}{(n-1)!} N_{\mathbf{q}}^{1/2} \frac{1}{N_{\mathbf{q}}^{1/2}} \sum_{\substack{\mathbf{i}\in\mathcal{I}_k \\ \Pi_{l=1}^k Z_{i_l q_l} = 1}} \{g(\mathbb{X}^{\mathbf{i}}) - m_g(\mathbf{q})\}$$
$$+ \frac{1}{Q^{k-2}} \sum_{q,l\in\mathcal{Q}, q\neq l} \frac{1}{n-1} \sum_{1\leqslant i\neq j\leqslant n} (Z_{iq} - \pi_q)(Z_{jl} - \pi_l) + o(1),$$

where, as in the non-degenerate case, we argued that the terms in $T_{2,2}$ involving sets L with cardinality $u \geqslant 3$ are negligible. We then obtain that, for $k = 2$, we have

$$n(\hat{m}_g - m_g) \underset{n\to\infty}{\rightsquigarrow} \frac{1}{Q} \sum_{q,l\in\mathcal{Q}} V_{ql} + \sum_{q,l\in\mathcal{Q}, q\neq l} \left(W_q W_l + \frac{1}{Q^2} \right),$$

where, for any $1 \leqslant q,l \leqslant Q$, the random variables V_{ql} are independent, with distribution $\mathcal{N}[0, \text{var}\{g(X_{12})|Z_{1q}Z_{2l} = 1\}]$ and $W = (W_1,\ldots,W_Q)$ is independent from the V_{ql}s, with distribution $\mathcal{N}_Q(0,\Sigma)$, and in the equal group proportions case Σ simplifies to $\Sigma_{ql} = -1/Q^2$ when $q \neq l$ and $\Sigma_{qq} = (Q-1)/Q^2$.

In the same way, whenever $k \geqslant 3$, all the terms appearing in T_1 are now negligible and we obtain

$$n(\hat{m}_g - m_g) \underset{n\to\infty}{\rightsquigarrow} \sum_{q,l\in\mathcal{Q}, q\neq l} \left(W_q W_l + \frac{1}{Q^2} \right).$$

A.2. Proof of theorem 3

Following the proof of theorem 1, we can easily write a joint central limit theorem for the triplet $(\hat{m}_1, \hat{m}_2, \hat{m}_3)$, namely

$$\sqrt{n} \begin{pmatrix} \hat{m}_1 - m_1 \\ \hat{m}_2 - m_2 \\ \hat{m}_3 - m_3 \end{pmatrix} \underset{n\to\infty}{\rightsquigarrow} \mathcal{N}_3(0, V),$$

with some covariance matrix V. Thus, we can apply a delta method (see for instance van der Vaart (1998), chapter 3) to the estimators $\hat{\beta} = \phi(\hat{m}_1, \hat{m}_2, \hat{m}_3)$ and $\hat{\alpha} = \psi(\hat{m}_1, \hat{m}_2, \hat{m}_3)$ where the functions ϕ and ψ are differentiable. This gives the convergence of the estimators $(\hat{\alpha}, \hat{\beta})$ and guarantees the same rates of convergence for $\hat{\alpha}$ and $\hat{\beta}$ as for the $\hat{m}_i$s.

A.3. Proof of theorem 4

Following the classical proof of Wald (1949) (see also van der Vaart (1998)), we may obtain the almost sure convergence of $(\hat{\gamma}_n, \hat{\alpha}_n, \hat{\beta}_n)$ to the true value of the parameter $(\gamma^*, \alpha^*, \beta^*)$, provided that the parameter space is compact and the three following assumptions are satisfied:

(a) convergence of the criterion

$$l_n(\boldsymbol{\pi},\alpha,\beta) := \frac{1}{n(n-1)(n-2)} \sum_{(i,j,k)\in\mathcal{I}_3} \log\{\mathbb{P}_{\boldsymbol{\pi},\alpha,\beta}(X_{ij},X_{ik},X_{jk})\}$$
$$\underset{n\to\infty}{\longrightarrow} H\{(\boldsymbol{\pi},\alpha,\beta);(\boldsymbol{\pi}^*,\alpha^*,\beta^*)\} := \mathbb{E}_{\boldsymbol{\pi}^*,\alpha^*,\beta^*}[\log\{\mathbb{P}_{\boldsymbol{\pi},\alpha,\beta}(X_{12},X_{13},X_{23})\}],$$

$\mathbb{P}_{\boldsymbol{\pi}^*,\alpha^*,\beta^*}$ almost surely;

(b) identification of the parameter $(\boldsymbol{\gamma},\alpha,\beta)$

$$H\{(\boldsymbol{\pi},\alpha,\beta);(\boldsymbol{\pi}^*,\alpha^*,\beta^*)\} \leqslant H\{(\boldsymbol{\pi}^*,\alpha^*,\beta^*);(\boldsymbol{\pi}^*,\alpha^*,\beta^*)\},$$

with equality if and only if $(\boldsymbol{\gamma},\alpha,\beta)=(\boldsymbol{\gamma}^*,\alpha^*,\beta^*)$, where $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$ are related through equation (10);

(c) uniform equicontinuity of the family of functions $(\boldsymbol{\pi},\alpha,\beta)\to l_n(\boldsymbol{\pi},\alpha,\beta)$, namely, for any $\varepsilon>0$, there is some $\nu>0$ such that, for all $n\geqslant 1$ and as soon as $\|(\boldsymbol{\pi},\alpha,\beta)-(\boldsymbol{\pi}',\alpha',\beta')\|_\infty\leqslant\nu$, we have $|l_n(\boldsymbol{\pi},\alpha,\beta)-l_n(\boldsymbol{\pi}',\alpha',\beta')|\leqslant\varepsilon$.

Item (a) follows from theorem 1, whereas (b) follows from Jensen's inequality and identifiability of the parameters, i.e. assumption 1. Let us now establish (c). We fix for the moment some $\nu>0$ and consider $\eta=(\boldsymbol{\pi},\alpha,\beta)$ and $\eta'=(\boldsymbol{\pi}',\alpha',\beta')$ such that $\|\eta-\eta'\|_\infty\leqslant\nu$. We recall that $(X_{ij},X_{ik},X_{jk})=\mathbb{X}^{(i,j,k)}$. We then write

$$\begin{aligned}|\log\{\mathbb{P}_\eta(Z_{iq}Z_{jl}Z_{km}=1,\mathbb{X}^{(i,j,k)})\}-\log\{\mathbb{P}_{\eta'}(Z_{iq}Z_{jl}Z_{km}=1,\mathbb{X}^{(i,j,k)})\}|&\\ \leqslant|\log(\pi_q)-\log(\pi'_q)|+|\log(\pi_l)-\log(\pi'_l)|+|\log(\pi_m)-\log(\pi'_m)|&\\ +|\log\{\mathbb{P}_\eta(\mathbb{X}^{(i,j,k)}|Z_{iq}Z_{jl}Z_{km}=1)\}-\log\{\mathbb{P}_{\eta'}(\mathbb{X}^{(i,j,k)}|Z_{iq}Z_{jl}Z_{km}=1)\}|.&\end{aligned}$$

The second term on the right-hand side of this inequality may be bounded as follows:

$$\begin{aligned}|\log\{\mathbb{P}_\eta(\mathbb{X}^{(i,j,k)}|Z_{iq}Z_{jl}Z_{km}=1)\}-\log\{\mathbb{P}_{\eta'}(\mathbb{X}^{(i,j,k)}|Z_{iq}Z_{jl}Z_{km}=1)\}|&\\ \leqslant 3\max\{|\log(\alpha)-\log(\alpha')|,|\log(1-\alpha)-\log(1-\alpha')|,|\log(\beta)-\log(\beta')|,&\\ |\log(1-\beta)-\log(1-\beta')|\}.&\end{aligned}$$

We now make use of the fact that we restricted our attention to the parameter space Π_δ, in which all the parameters are lower bounded by δ (assumption 2). Moreover, for any $x,y>0$, we may use $|\log(x)-\log(y)|\leqslant|x-y|/\min(x,y)$. This finally leads to

$$|\log\{\mathbb{P}_\eta(Z_{iq}Z_{jl}Z_{km}=1,\mathbb{X}^{(i,j,k)})\}-\log\{\mathbb{P}_{\eta'}(Z_{iq}Z_{jl}Z_{km}=1,\mathbb{X}^{(i,j,k)})\}|\leqslant 6\delta^{-1}\nu.$$

Now, we obtain

$$\begin{aligned}\mathbb{P}_\eta(\mathbb{X}^{(i,j,k)})&=\sum_{qlm}\mathbb{P}_\eta(Z_{iq}Z_{jl}Z_{km}=1,\mathbb{X}^{(i,j,k)})\leqslant\exp(6\delta^{-1}\nu)\sum_{qlm}\mathbb{P}_{\eta'}(Z_{iq}Z_{jl}Z_{km}=1,\mathbb{X}^{(i,j,k)})\\&=\exp(6\delta^{-1}\nu)\,\mathbb{P}_{\eta'}(\mathbb{X}^{(i,j,k)}),\end{aligned}$$

and thus

$$\log\{\mathbb{P}_\eta(\mathbb{X}^{(i,j,k)})\}\leqslant 6\nu/\delta+\log\{\mathbb{P}_{\eta'}(\mathbb{X}^{(i,j,k)})\}.$$

As this inequality is symmetric with respect to η and η', we further obtain

$$|\log\{\mathbb{P}_\eta(\mathbb{X}^{(i,j,k)})\}-\log\{\mathbb{P}_{\eta'}(\mathbb{X}^{(i,j,k)})\}|\leqslant 6\nu/\delta.$$

Finally,

$$|l_n(\eta)-l_n(\eta')|\leqslant\frac{1}{n(n-1)(n-2)}\sum_{i,j,k}|\log\{\mathbb{P}_\eta(\mathbb{X}^{(i,j,k)})\}-\log\{\mathbb{P}_{\eta'}(\mathbb{X}^{(i,j,k)})\}|\leqslant 6\nu/\delta,$$

which establishes assumption (c).

To obtain further the rates of convergence of the estimators, one usually proceeds to a Taylor series expansion of the derivative $\partial l_n(\boldsymbol{\pi}^*,\alpha^*,\beta^*)$ near the estimator $(\hat{\boldsymbol{\pi}}_n,\hat{\alpha}_n,\hat{\beta}_n)$. Write

$$0=\partial l_n(\hat{\boldsymbol{\pi}}_n,\hat{\alpha}_n,\hat{\beta}_n)=\partial l_n(\boldsymbol{\pi}^*,\alpha^*,\beta^*)+\{(\hat{\boldsymbol{\pi}}_n,\hat{\alpha}_n,\hat{\beta}_n)-(\boldsymbol{\pi}^*,\alpha^*,\beta^*)\}\,\partial^2 l_n(\tilde{\boldsymbol{\pi}}_n,\tilde{\alpha}_n,\tilde{\beta}_n),$$

where $(\tilde{\boldsymbol{\pi}}_n, \tilde{\alpha}_n, \tilde{\beta}_n)$ is some point between $(\hat{\boldsymbol{\pi}}_n, \hat{\alpha}_n, \hat{\beta}_n)$ and $(\boldsymbol{\pi}^*, \alpha^*, \beta^*)$. Applying theorem 1 to the quantity $\partial l_n(\boldsymbol{\pi}^*, \alpha^*, \beta^*)$, we obtain its almost sure convergence to $\mathbb{E}_{\boldsymbol{\pi}^*,\alpha^*,\beta^*}[\partial \log\{\mathbb{P}_{\boldsymbol{\pi}^*,\alpha^*,\beta^*}(X_{12}, X_{13}, X_{23})\}] = 0$, as well as the asymptotic normality

$$\sqrt{n}\, \partial l_n(\boldsymbol{\pi}^*, \alpha^*, \beta^*) \underset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0, J).$$

Now, at a fixed point $(\boldsymbol{\pi}, \alpha, \beta)$, the Hessian matrix $\partial^2 l_n(\boldsymbol{\pi}, \alpha, \beta)$ converges from theorem 1 to $\mathbb{E}_{\boldsymbol{\pi}^*,\alpha^*,\beta^*}[\partial^2 \log\{\mathbb{P}_{\boldsymbol{\pi},\alpha,\beta}(X_{12}, X_{13}, X_{23})\}]$. Combining the almost sure convergence of $(\hat{\boldsymbol{\pi}}_n, \hat{\alpha}_n, \hat{\beta}_n)$ to $(\boldsymbol{\pi}^*, \alpha^*, \beta^*)$, with uniform equicontinuity of the family of functions $(\boldsymbol{\pi}, \alpha, \beta) \to \partial^2 l_n(\boldsymbol{\pi}, \alpha, \beta)$ (the proof is similar to point (c) above and is therefore omitted), we obtain the almost sure convergence

$$\partial^2 l_n(\tilde{\boldsymbol{\pi}}_n, \tilde{\alpha}_n, \tilde{\beta}_n) \underset{n\to\infty}{\rightsquigarrow} \mathbb{E}_{\boldsymbol{\pi}^*,\alpha^*,\beta^*}[\partial^2 \log\{\mathbb{P}_{\boldsymbol{\pi}^*,\alpha^*,\beta^*}(X_{12}, X_{13}, X_{23})\}] := -K.$$

If the Fisher information matrix K is invertible, we obtain

$$\{(\hat{\boldsymbol{\pi}}_n, \hat{\alpha}_n, \hat{\beta}_n) - (\boldsymbol{\pi}^*, \alpha^*, \beta^*)\}\sqrt{n} \underset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0, K^{-1}JK^{-1}).$$

In this case, the inverse of the limiting variance is known as Godambe information (Varin, 2008). Its form is due to the fact that $K^{-1} \neq J$ in general, resulting in a loss of efficiency of the estimators. In cases where K is not invertible, or when $J = 0$, the rate of convergence of the estimators is faster than $1/\sqrt{n}$. In particular, when the group proportions are equal, we know from theorem 1 that $n\partial l_n(\boldsymbol{\pi}^*, \alpha^*, \beta^*)$ converges in distribution and then the rate of convergence of $(\hat{\boldsymbol{\pi}}_n, \hat{\alpha}_n, \hat{\beta}_n)$ is at least $1/n$.

A.4. Proof of theorem 5

The proof of theorem 5 follows the scheme that was described in the proof of theorem 4. We denote by $(\boldsymbol{\pi}^*, \mathbf{p}^*, \boldsymbol{\theta}^*)$ the true value of the parameter and by $\mathbb{P}^*$ and $\mathbb{E}^*$ the corresponding probability and expectation. First, we establish the consistency of the normalized composite likelihood (point (a)). According to theorem 1, we have, for any fixed value of $(\boldsymbol{\pi}, \mathbf{p}, \boldsymbol{\theta})$,

$$\frac{2}{n(n-1)} \sum_{1\leqslant i<j\leqslant n} \mathbf{1}_{X_{ij}\neq 0} \log\{\mathbb{P}_{\boldsymbol{\pi},\mathbf{p},\boldsymbol{\theta}}(X_{ij})\} \underset{n\to\infty}{\to} \mathbb{E}^*[\mathbf{1}_{X_{12}\neq 0} \log\{\mathbb{P}_{\boldsymbol{\pi},\mathbf{p},\boldsymbol{\theta}}(X_{12})\}]$$
$$:= H\{(\boldsymbol{\pi}, \mathbf{p}, \boldsymbol{\theta}); (\boldsymbol{\pi}^*, \mathbf{p}^*, \boldsymbol{\theta}^*)\}, \qquad \mathbb{P}^* \text{ almost surely.}$$

Here, we need to deal with the fact that we use a random value for $\mathbf{p}$ (a preliminary step estimate) in the definition of $\hat{\boldsymbol{\theta}}$. It is thus necessary to prove that this convergence happens uniformly with respect to $\mathbf{p}$. But this will be a consequence of point (c) below. Combining this with the almost sure convergence of $\hat{\mathbf{p}}_n$ to the true value $\mathbf{p}^*$ (this is either a consequence of theorem 1 when $\mathbf{p} = p$ is constant, or a consequence of Sections 3.1 and 3.2 when $\mathbf{p} = (\alpha, \beta)$), we obtain

$$\frac{2}{n(n-1)} \mathcal{L}_X^{\text{compo}}(\boldsymbol{\pi}, \hat{\mathbf{p}}_n, \boldsymbol{\theta}) \underset{n\to\infty}{\longrightarrow} H\{(\boldsymbol{\pi}, \mathbf{p}, \boldsymbol{\theta}); (\boldsymbol{\pi}^*, \mathbf{p}^*, \boldsymbol{\theta}^*)\}, \qquad \mathbb{P}^* \text{ almost surely.}$$

Moreover, we assumed that $f(\cdot, \theta)$ has a continuous cumulative distribution function and the distribution of a present edge is given by equation (16), so we have

$$H\{(\boldsymbol{\pi}, \mathbf{p}, \boldsymbol{\theta}); (\boldsymbol{\pi}^*, \mathbf{p}^*, \boldsymbol{\theta}^*)\} = \int_x \log\{\gamma_{\text{in}}\, f(x; \theta_{\text{in}}) + \gamma_{\text{out}}\, f(x; \theta_{\text{out}})\}\{\gamma^*_{\text{in}}\, f(x; \theta^*_{\text{in}}) + \gamma^*_{\text{out}}\, f(x; \theta^*_{\text{out}})\}\, \mathrm{d}x,$$

where $(\gamma_{\text{in}}, \gamma_{\text{out}})$ as well as $(\gamma^*_{\text{in}}, \gamma^*_{\text{out}})$ are defined through $(\boldsymbol{\pi}, \mathbf{p})$ and $(\boldsymbol{\pi}^*, \mathbf{p}^*)$ respectively. Thus, the difference

$$H\{(\boldsymbol{\pi}^*, \mathbf{p}^*, \boldsymbol{\theta}^*); (\boldsymbol{\pi}^*, \mathbf{p}^*, \boldsymbol{\theta}^*)\} - H\{(\boldsymbol{\pi}, \mathbf{p}, \boldsymbol{\theta}); (\boldsymbol{\pi}^*, \mathbf{p}^*, \boldsymbol{\theta}^*)\}$$

is a Kullback–Leibler divergence between two mixture distributions of the form (16). This entails positivity of this difference. Moreover, assumption 3 ensures that the difference is 0 if and only if

$$\gamma_{\text{in}}\delta_{\theta_{\text{in}}} + \gamma_{\text{out}}\delta_{\theta_{\text{out}}} = \gamma^*_{\text{in}}\delta_{\theta^*_{\text{in}}} + \gamma^*_{\text{out}}\delta_{\theta^*_{\text{out}}},$$

which establishes point (b), up to a permutation on the label parameters $\{\text{in}, \text{out}\}$. Finally, the proof of point (c) follows the same lines as in the proof of theorem 4, and uses the continuity of the map $\theta \mapsto f(\cdot, \theta)$, which is a consequence of assumption 4.

To obtain further the rates of convergence of our estimators, we proceed exactly as we did in the proof of theorem 4.

References

Airoldi, E., Blei, D., Fienberg, S. and Xing, E. (2008) Mixed-membership stochastic block-models. *J. Mach. Learn. Res.*, **9**, 1981–2014.

Allman, E., Matias, C. and Rhodes, J. (2009) Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, **37**, no. 6A, 3099–3132.

Allman, E., Matias, C. and Rhodes, J. (2011) Parameters identifiability in random graph mixture models. *J. Statist. Planng Inf.*, **141**, 1719–1736.

Barrat, A., Barthélemy, M., Pastor-Satorras, R. and Vespignani, A. (2004) The architecture of complex weighted networks. *Proc. Natn. Acad. Sci. USA*, **101**, 3747–3752.

Bickel, P. and Chen, A. (2009) A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natn. Acad. Sci. USA*, **106**, 21068–21073.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D. (2006) Complex networks: structure and dynamics. *Phys. Rep.*, **424**, 175–308.

Carreira-Perpiñán, M. A. and Renals, S. (2000) Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neur. Computn*, **12**, 141–152.

Choi, D. S., Wolfe, P. J. and Airoldi, E. M. (2010) Stochastic blockmodels with growing number of classes. *Technical Report ArXiv:1011.4644*.

Cox, D. R. and Reid, N. (2004) A note on pseudolikelihood constructed from marginal densities. *Biometrika*, **91**, 729–737.

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

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

Doreian, P., Batagelj, V. and Ferligoj, A. (2005) *Generalized Blockmodeling*. Cambridge: Cambridge University Press.

Erdős, P. and Rényi, A. (1959) On random graphs: I. *Publ. Math. Debrecen*, **6**, 290–297.

Erosheva, E., Fienberg, S. and Lafferty, J. (2004) Mixed-membership models of scientific publications. *Proc. Natn. Acad. Sci. USA*, **97**, 11885–11892.

Fortunato, S. (2010) Community detection in graphs. *Phys. Rep.*, **486**, 75–174.

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

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

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

Holland, P., Laskey, K. and Leinhardt, S. (1983) Stochastic blockmodels: some first steps. *Socl Netwrks*, **5**, 109–137.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classificn*, **2**, 193–218.

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

Laloux, L., Cizeau, P., Bouchaud, J.-P. and Potters, M. (1999) Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, **83**, 1467–1470.

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

Latouche, P., Birmelé, E. and Ambroise, C. (2011b) Variational bayesian inference and complexity control for stochastic block models. *Statist. Modlling*, to be published.

Mariadassou, M., Robin, S. and Vacher, C. (2010) Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Statist.*, **4**, 715–742.

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

Newman, M. E. J. and Leicht, E. A. (2007) Mixture models and exploratory analysis in networks. *Proc. Natn. Acad. Sci. USA*, **104**, 9564–9569.

Nowicki, K. and Snijders, T. (2001) Estimation and prediction for stochastic blockstructures. *J. Am. Statist. Ass.*, **96**, 1077–1087.

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

Pieters, F. and Baumgartner, H. (2002) Who talks to whom? intra- and inter- disciplinary communication of economics journals. *J. Econ. Lit.*, **40**, 483–509.

Rohe, K., Chatterjee, S. and Yu, B. (2011) Spectral clustering and the high-dimensional stochastic block model. *Ann. Statist.*, to be published.

Snijders, T. A. B. and Nowicki, K. (1997) Estimation and prediction for stochastic block-models for graphs with latent block structure. *J. Classificn*, **14**, 75–100.
Titterington, D., Smith, A. and Makov, U. (1985) *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.
van der Vaart, A. W. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.
Varin, C. (2008) On composite marginal likelihoods. *AStA Adv. Statist. Anal.*, **92**, 1–28.
Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 595–601.
Wu, C. (1983) On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103.
Zanghi, H., Ambroise, C. and Miele, V. (2008) Fast online graph clustering via Erdős Rényi mixture. *Pattn Recogn*, **41**, 3592–3599.
Zanghi, H., Picard, F., Miele, V. and Ambroise, C. (2010) Strategies for online inference of model-based clustering in large and growing networks. *Ann. Appl. Statist.*, **4**, 687–714.
Ziberna, A. (2007) Generalized blockmodeling of valued networks. *Socl Netwrks*, **29**, 105–126.