IDENTIFICATION AND ESTIMATION OF STOCHASTIC BLOCKMODELS

By Koen Jochmans

Toulouse School of Economics, University of Toulouse Capitole

Revised on January 27, 2022

This paper contains new identification results for undirected weighted stochastic blockmodels. They are sharper than the ones available to date and the arguments underlying them are constructive. A nonparametric estimation framework that is computationally attractive is presented and the associated distribution theory is derived. Numerical experiments are reported on.

1. Introduction. The stochastic blockmodel provides a parsimonious way to incorporate latent heterogeneity in the analysis of network data. The original application of Holland, Laskey and Leinhardt [1983] concerned the binary decision of edge formation between two nodes and generalizes the Erdős-Rényi random-graph model. In the latter model, the edges are formed independently with a common probability. In the former, the set of nodes is partitioned into a finite set of latent communities, and the link probability between two nodes depends on the communities that they belong to. The latent block structure has since been used to study general (discrete or continuous) outcomes generated from pairwise interaction (Hoff, Raftery and Handcock 2002), thereby extending the applicability of the stochastic blockmodel to weighted graphs.

Under the assumption that the number of latent communities is known, Allman, Matias and Rhodes [2009; 2011] obtained a set of identification results for the remaining unknowns of the model, i.e., the distribution of the communities and the conditional distributions of the edge weights given the community membership of the nodes. Here and later identification is to be

AMS 2000 subject classifications: 62H30, 62G05.
Keywords and phrases: network, mixture model, random graph, stochastic blockmodel.

understood to be up to an arbitrary labelling of the latent communities (a well-known indeterminancy of the model). The approach of Allman, Matias and Rhodes [2009; 2011] is not constructive, in the sense that it does not suggest an approach to estimation.

In this paper we present a new approach to identification of stochastic blockmodels. It differs from the existing ones in three important respects. First, it covers all components of the model, including the number of latent communities. Second, it yields results that are considerably sharper than those of Allman, Matias and Rhodes [2009; 2011]. For example, in the classic setting with binary edge weights and two latent communities (as in Snijders and Nowicki 1997) they rely on complete subgraphs on 16 nodes. In contrast, we achieve identification from incomplete subgraphs involving no more than 4 nodes. Third, our approach is constructive, and we build on it to propose nonparametric estimators.

Our estimation procedure is computationally attractive. It is built around a joint (approximate) diagonalization step tailored to our setup. This type of routine has found applicability in the estimation of (multivariate) mixture models elsewhere [Bonhomme, Jochmans and Robin 2016a;b]. Here we use it as an auxiliary estimator in the construction of our main estimators of the components of the model. Moreover, once it has been computed, our estimators of the stochastic blockmodel are least-squares estimators and are thus immediate to compute. We present an estimator of the distribution of the communities as well as a generic estimator of linear functionals of the conditional distributions. The latter covers (cumulative) distribution functions, their moments, and probability mass functions, for example. We also give results for a kernel estimator for conditional densities for the case where edge weights are continuous.

Limit theory for these estimators is presented under an asymptotic scheme where the number of nodes in the network, n, goes to infinity, assuming that the number of communities is known. Under weak regularity conditions they converge in distribution to correctly-centered normal random variables at the rate $n^{-1/2}$. This rate equally applies to density estimation, as the

smoothing bias is small relative to the standard deviation. Undersmoothing is not needed to achieve this result.

It is instructive to place these findings against the existing alternative estimation routines. It is well known that (parametric) maximum-likelihood [Snijders and Nowicki 1997] and Bayesian estimation [Nowicki and Snijders 2001] are infeasible except in small-sized networks [Mariadassou, Robin and Vacher 2010]. The literature has explored variational alternatives that are succesful at alleviating much of the computational burden [Daudin, Picard and Robin 2008]. However, only few theoretical guarantees for variational estimation have been obtained, and these are restricted to the case with binary edge weights; see Celisse, Daudin and Pierre [2012] and Bickel et al. [2013].

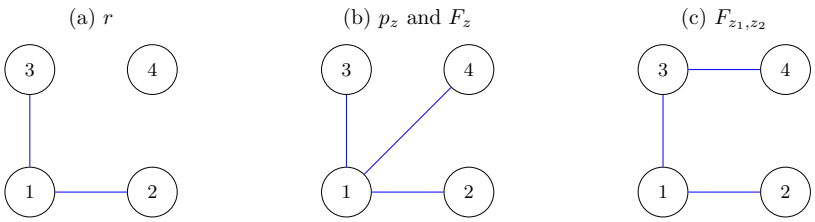
2. Stochastic blockmodel. Consider a graph involving n nodes where the set of nodes is partitioned into r latent communities, labelled $1, \ldots, r$. Each node is first assigned to a community independently according to some probability distribution $\boldsymbol{p} = (p_1, \ldots, p_r)'$. The community of node i is recorded in the latent variable Z_i. Thus, $\mathbb{P}(Z_i = z) = p_z > 0$ for each $1 \leq z \leq r$, and is equal to zero otherwise. Next, each pair of nodes $i \neq j$ draws a real-valued weighted edge $X_{i,j}$ from a distribution that depends on the communities they belong to,

$$F_{z_1,z_2}(x) := \mathbb{P}(X_{i,j} \leq x | Z_i = z_1, Z_j = z_2).$$

The $X_{i,j}$ are independent conditional on the community indicators (Z_i, Z_j). Because edges are undirected it is natural to work under a presumption of symmetry, where $F_{z_1,z_2} = F_{z_2,z_1}$ for all (z_1, z_2). This setup reduces to the standard (unweighted) stochastic blockmodel when the $X_{i,j}$ are binary random variables.

In the next section we present arguments that allow to nonparametrically identify the components of the stochastic blockmodel, i.e., the number of latent communities r, their distribution $\boldsymbol{p}$, and the conditional distributions of the edge weights F_{z_1,z_2} for all $z_1 < z_2$. As always, we will only be able to recover the distributions up to an arbitrary permutation of the latent

Fig 1: Subgraph configurations used for identification



communities. This indeterminancy is well understood and we will leave it implicit in the sequel unless by doing so there would be a risk of causing confusion.

3. Identification from 4 nodes. We first show that it is possible to constructively identify the stochastic blockmodel from the distribution of edge weights of subgraphs involving as little as 4 nodes. To show this we let

$$F_z(x) := \mathbb{P}(X_{i,j} \leq x | Z_i = z)$$

and impose the following condition.

ASSUMPTION 1. *The functions $F_1, \ldots, F_r$ are linearly independent.*

Rank conditions as this one arise frequently in the analysis of multivariate latent-variable models.

We will prove the following theorem and specialize it further afterwards. The three types of subgraphs that we will use in the proof are pictured in Figure 1, with nodes labelled 1, 2, 3, and 4. They are (a) two-star graphs; (b) three-star graphs; and (c) path graphs on 4 nodes. We do not rely on complete subgraphs.

THEOREM 1. *Suppose that Assumption 1 holds. Then,*

(i) The number of communities r is identified from subgraphs of three nodes;

(ii) The distribution of communities, $\boldsymbol{p}$, as well as expectations of the form

$$\varphi_{z_1,z_2} := \mathbb{E}(\varphi(X_{i,j}) | Z_i = z_1, Z_j = z_2)$$

are identified up to a common permutation of the latent communities from subgraphs of four nodes.

Consider, first, the subgraph in Figure 1(a), involving the edges between the three nodes $1, 2, 3$. Observe that the edge weights $X_{1,2}, X_{1,3}$ are independent conditional on Z_1. Furthermore, their (unconditional) distribution factors as

$$\sum_{z=1}^{r} p_z \, F_z \otimes F_z$$

This is a bivariate finite-mixture model. It follows from the work of Kwon and Mbakop [2019, Propositions 2.1 and 2.3] on mixture models that r is identified under Assumption 1. See also Kasahara and Shimotsu [2014] for related results.

Next look at the graph in Figure 1(b). The edge weights in this three-star graph—$X_{1,2}, X_{1,3}, X_{1,4}$—are again independent conditional on Z_1. Hence, their tri-variate distribution again factors as the multivariate finite mixture

$$\sum_{z=1}^{r} p_z \, F_z \otimes F_z \otimes F_z.$$

From this, the identification of $F_1, \ldots, F_r$ and $\boldsymbol{p} = (p_1, \ldots, p_r)'$ (up to an arbitrary but common ordering) follows from Bonhomme, Jochmans and Robin [2016b, Corollaries 1 and 2]. See also Allman, Matias and Rhodes [2009] and Bonhomme, Jochmans and Robin [2016a] for similar results in related settings.

Finally, Assumption 1 implies that there exists a finite integer l such that the $l \times r$ matrix $\boldsymbol{G}$,

$$(\boldsymbol{G})_{l',z} := \mathbb{E}(\alpha_{l'}(X_{i,j})|Z_i = z),$$

for a set of transformation functions $\alpha_1, \ldots, \alpha_l$, has full column rank. One choice for these transformation functions would be $\alpha_{l'}(x) = \{x \leq x_{l'}\}$, where $x_1, \ldots, x_l$ is a grid of points and $\{\cdot\}$ denotes the indicator function. In this case, $(\boldsymbol{G})_{l',z} = F_z(x_{l'})$. Other approximating functions such as orthogonal polynomials are also possible. Whatever the choice of functions, the matrix

$\boldsymbol{G}$ is identified because the F_z are and the columns of $\boldsymbol{G}$ are linear functionals thereof. Use the edge weights $X_{1,2}, X_{1,3}, X_{3,4}$ from the path graph in Figure 1(c) to construct the $l \times l$ matrix $\boldsymbol{M}_\varphi$ as

$$(\boldsymbol{M}_\varphi)_{l_1,l_2} := \mathbb{E}(\alpha_{l_1}(X_{1,2})\,\varphi(X_{1,3})\,\alpha_{l_2}(X_{3,4})).$$

Noting that $X_{1,2}, X_{1,3}, X_{3,4}$ are independent conditional on the pair (Z_1, Z_3), we have

$$\boldsymbol{M}_\varphi = \boldsymbol{G}\,\boldsymbol{H}_\varphi \boldsymbol{G}',$$

where

$$(\boldsymbol{H}_\varphi)_{z_1,z_2} := p_{z_1} p_{z_2} \varphi_{z_1,z_2}.$$

Because $\boldsymbol{G}$ has maximal column rank $\boldsymbol{G}'\boldsymbol{G}$ is invertible and, thus, we obtain

$$\boldsymbol{H}_\varphi = (\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{M}_\varphi \boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G})^{-1}$$

With the p_z already shown to be identified this suffices to complete the proof. It is nonetheless useful to note that an application of the above argument to the constant function $\varphi(x) = 1$ gives $\boldsymbol{M}_1 = \boldsymbol{G}\,\boldsymbol{H}_1\boldsymbol{G}'$ where the matrix $\boldsymbol{H}_1$ has entries $(\boldsymbol{H}_1)_{z_1,z_2} := p_{z_1}p_{z_2}$. Hence,

$$\varphi_{z_1,z_2} = (\boldsymbol{H}_1)^{-1}_{z_1,z_2}(\boldsymbol{H}_\varphi)_{z_1,z_2},$$

which is a more convenient result for the purpose of estimation. The proof of Theorem 1 is complete.

Part (ii) of Theorem 1 can be applied to $\varphi(x) = \{x \leq x'\}$ for any chosen value x', leading to our first proposition.

PROPOSITION 1. *Suppose that Assumption 1 holds. Then the (weighted) stochastic blockmodel is nonparametrically identified from the distribution of edge weights involving* 4 *nodes.*

Proposition 1 is to be contrasted with the existing identification results to date. For the unweighted model with two communities Allman, Matias and Rhodes [2009, Theorem 7] showed identification from the complete subgraph involving 16 nodes. Assuming the number of communities is known, Allman,

Matias and Rhodes [2011, Theorems 14 and 15] obtained results for the general model, but they rely on the complete subgraph on 9 nodes to do so. Their results further require conditions on the support of the edge weights relative to the number of communities being sufficiently large, in addition to the demand that the F_{z_1,z_2} (for all $z_1 < z_2$) are linearly-independent functions.

Theorem 1 is useful beyond as an input to establish Proposition 1 as it can be used to directly show identification of various functionals of the conditional distributions. Its proof is also constructive and we will present estimators below.

4. Identification from larger subgraphs. Assumption 1 cannot be satisfied when the edge weights can take on strictly less values than there are latent communities. In such a case identification can be obtained by looking at subgraphs involving a larger set of nodes. This illustrates the interplay between the richness of the support of the edge weights, the number of latent communities, and the size of the subgraphs that are needed to achieve identification.

The argument is based on chaining nodes in a particular manner. We let

$$F_z^q(x_1,\ldots,x_q) := \mathbb{P}(X_{i,i_1} \leq x_1, X_{i_1,i_2} \leq x_2, \ldots, X_{i_1,i_q} \leq x_q | Z_i = z),$$

where the indices i and $i_1,\ldots,i_q$ are all distinct, and impose the following rank requirement.

ASSUMPTION 1'. *There exists a finite integer q such that the functions $F_1^q,\ldots,F_r^q$ are linearly independent.*

Let $q_{\min}$ be the smallest integer for which Assumption 1' holds.

PROPOSITION 1'. *If $q_{\min}$ is positive the (weighted) stochastic blockmodel is nonparametrically identified from the distribution of edge weights involving $3q_{\min}+1$ nodes.*

Proposition 1' reduces to Proposition 1 when $q_{\min} = 1$. It can be shown by applying a similar strategy as in the proof of Theorem 1. The proof is given in the Appendix.

5. Nonparametric estimation. An estimator of the number of latent communities, r, can be constructed along the lines of Kwon and Mbakop [2019]. Here we construct estimators of the distribution of the communities, $\boldsymbol{p}$, and the conditional distributions F_{z_1,z_2} and functionals thereof, building on the proof of Theorem 1. We consider a setting where we observe data from a single network involving n nodes.

Our proposal is to proceed in two steps. First, the matrix $\boldsymbol{G}$ is estimated by a modification of the diagonalization estimator of Bonhomme, Jochmans and Robin [2016b]. This estimator, $\hat{\boldsymbol{G}}$, is detailed below. We then appeal to the analogy principle to construct our second-step estimators of the different components of the stochastic blockmodel.

If we write $\boldsymbol{a} := (a_1, \ldots, a_l)'$ for $a_{l'} := \mathbb{E}(\alpha_{l'}(X_{i,j}))$ we have the univariate mixture representation

$$\boldsymbol{a} = \boldsymbol{G}\boldsymbol{p}.$$

Given $\hat{\boldsymbol{G}}$, a least-squares argument suggests estimating $\boldsymbol{p}$ by

$$\hat{\boldsymbol{p}} := (\hat{\boldsymbol{G}}'\hat{\boldsymbol{G}})^{-1}\hat{\boldsymbol{G}}'\hat{\boldsymbol{a}},$$

for $\hat{\boldsymbol{a}} = (\hat{a}_1, \ldots, \hat{a}_l)'$, with

$$\hat{a}_{l'} := \frac{2}{n(n-1)} \sum_{i<j} \alpha_{l'}(X_{i,j}).$$

This approach is inspired by Titterington [1983], where minimum-distance estimators of mixing proportions were considered.

Similarly,

$$\hat{\boldsymbol{H}}_{\varphi} := (\hat{\boldsymbol{G}}'\hat{\boldsymbol{G}})^{-1}\hat{\boldsymbol{G}}'\hat{\boldsymbol{M}}_{\varphi}\hat{\boldsymbol{G}}(\hat{\boldsymbol{G}}'\hat{\boldsymbol{G}})^{-1},$$

constructed with

$$(\hat{\boldsymbol{M}}_{\varphi})_{l_1,l_2} := \frac{1}{n(n-1)(n-2)(n-3)} \sum_{i_1 \neq i_2 \neq i_3 \neq i_4} \alpha_{l_1}(X_{i_1,i_2})\, \varphi(X_{i_2,i_3})\, \alpha_{l_2}(X_{i_3,i_4}),$$

yields the estimator

$$\hat{\varphi}_{z_1,z_2} := (\hat{\boldsymbol{H}}_1)^{-1}_{z_1,z_2} (\hat{\boldsymbol{H}}_\varphi)_{z_1,z_2}$$

of φ_{z_1,z_2}.

Below we will derive the sampling properties of these estimators under asymptotics where the number of nodes, n, grows large, assuming that the number of communities, r, is known. Derivations of the results to follow are collected in the appendix.

We remark that an alternative sampling scheme would be to sample n' independent networks, each of size n and generated from the same stochastic blockmodel. Under asymptotics where $n' \to \infty$ while n remains fixed, our estimators achieve the parametric rate of $\sqrt{n'}$ under the same regularity conditions as the ones introduced here. We omit further details for this case for brevity.

5.1. *Diagonalization step.* We construct the matrix $\hat{\boldsymbol{G}}$ by relying on an (approximate) simultaneous-diagonalization argument. We summarize the procedure here and refer to Bonhomme, Jochmans and Robin [2016a;b] for additional details on this approach in multivariate mixture and related latent-variable models. We begin by constructing the $l \times l$ matrix $\hat{\boldsymbol{A}}_0$, where

$$(\hat{\boldsymbol{A}}_0)_{l_1,l_2} := \frac{1}{n(n-1)(n-2)} \sum_{i_1 \neq i_2 \neq i_3} \alpha_{l_1}(X_{i_1,i_2})\, \alpha_{l_2}(X_{i_1,i_3}),$$

and perform an eigendecomposition on it to construct an $l \times r$ matrix $\hat{\boldsymbol{V}}$ for which $\hat{\boldsymbol{V}}\hat{\boldsymbol{A}}_0\hat{\boldsymbol{V}}' = \boldsymbol{I}_r$, the $r \times r$ identity matrix. We next form the $l \times l$ matrices $\hat{\boldsymbol{A}}_{l'}$, with

$$(\hat{\boldsymbol{A}}_{l'})_{l_1,l_2} := \frac{1}{n(n-1)(n-2)(n-3)} \sum_{i_1 \neq i_2 \neq i_3 \neq i_4} \alpha_{l_1}(X_{i_1,i_2})\, \alpha_{l'}(X_{i_1,i_3})\, \alpha_{l_2}(X_{i_1,i_4}),$$

where $l' = 1, \ldots, l$, and transform them using $\hat{\boldsymbol{V}}$ to obtain the $r \times r$ matrices

$$\hat{\boldsymbol{N}}_{l'} := \hat{\boldsymbol{V}}'\hat{\boldsymbol{A}}_{l'}\hat{\boldsymbol{V}}'.$$

We then find the matrix of joint (approximate) eigenvectors of these matrices as

$$\hat{\boldsymbol{Q}} := \arg\min_{\boldsymbol{Q}\in\boldsymbol{\mathcal{Q}}} \sum_{l'=1}^{l} \sum_{z_1\neq z_2} (\boldsymbol{Q}'\hat{\boldsymbol{N}}_{l'}\boldsymbol{Q})_{z_1,z_2}^2,$$

where we let $\boldsymbol{\mathcal{Q}}$ be the set of $r \times r$ orthonormal matrices. With this matrix at hand we construct $\hat{\boldsymbol{G}}$ as

$$(\hat{\boldsymbol{G}})_{l',z} := (\hat{\boldsymbol{Q}}'\hat{\boldsymbol{N}}_{l'}\hat{\boldsymbol{Q}})_{z,z}.$$

The minimization problem that defines $\hat{\boldsymbol{Q}}$ can be solved efficiently using the algorithm of Cardoso and Souloumiac [1993].

An alternative to the joint diagonalization approach would be to estimate the $(\boldsymbol{G})_{l',z} = \mathbb{E}(\alpha_{l'}(X_{i,j})|Z_i = z)$ using estimates of the F_z. These could be obtained by maximizing a parametric likelihood (using the EM algorithm, see McLachlan and Peel 2000), or by nonparametric procedures such as those given in Levine, Hunter and Chauveau [2011]. A practical advantage of our proposal is that it bypasses estimation of the complete mixture model. A theoretical advantage (relative to other nonparametric estimators) is that distribution theory for the matrix $\hat{\boldsymbol{G}}$ can be obtained by adapting the work of Bonhomme, Jochmans and Robin [2016b] to deal with the network structure of the data.

5.2. *Regularity conditions.* Three regularity conditions will be used. They are collected here. The first two of them impose conventional requirements on second moments.

ASSUMPTION 2. *The variables $\alpha_{l'}(X_{i,j})$ have finite variance.*

Note that this assumption can always be satisfied by working with bounded functions.

ASSUMPTION 3. *The variable $\varphi(X_{i,j})$ has finite variance.*

The third regularity condition concerns the $l \times l$ matrix $\boldsymbol{A}_0$, with

$$(\boldsymbol{A}_0)_{l_1,l_2} := \mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_3})).$$

Observe that $\boldsymbol{A}_0 = \boldsymbol{G}\,\mathrm{diag}(\boldsymbol{p})\,\boldsymbol{G}'$. Its rank is equal to r by Assumption 1 and so it has r non-zero eigenvalues. We represent its eigendecomposition as

$$\boldsymbol{A}_0 = \boldsymbol{U}\boldsymbol{L}\boldsymbol{U}',$$

with $\boldsymbol{L}$ the $r \times r$ diagonal matrix that collects the non-zero eigenvalues and $\boldsymbol{U}$ the $l \times r$ orthonormal matrix whose r columns contain the associated eigenvectors. Then

$$\boldsymbol{V} := \boldsymbol{L}^{-1/2}\boldsymbol{U}'$$

is the probability limit of $\hat{\boldsymbol{V}}$.

ASSUMPTION 4. *All non-zero eigenvalues of $\boldsymbol{A}_0$ are simple.*

This assumption implies continuity of $\boldsymbol{V}$ as a function of $\boldsymbol{A}_0$ and is helpful in deriving the properties of $\hat{\boldsymbol{V}}$.

We remark that, because $\hat{\boldsymbol{A}}_0$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{A}_0$, the rank condition on $\boldsymbol{G}$ can be tested by any of a number of standard procedures to test the rank of a matrix. This is related to, but different from, the approach put forth in Lei [2016].

5.3. *A linearization.* An important step in deriving the large-sample properties of our procedures lies in analyzing the first-step estimator, $\hat{\boldsymbol{G}}$. This estimator is a complicated function of the auxiliary estimators of the matrices $\boldsymbol{A}_0$ and $\boldsymbol{A} := (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_l)$, with the elements of the latter equal to

$$(\boldsymbol{A}_{l'})_{l_1,l_2} := \mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l'}(X_{i_1,i_3})\,\alpha_{l_2}(X_{i_1,i_4})).$$

Their respective influence functions are

$$\boldsymbol{\beta}_i(\boldsymbol{A}_0) := \mathrm{vec}\,(\boldsymbol{B}_i(\boldsymbol{A}_0) - \mathbb{E}(\boldsymbol{B}_i(\boldsymbol{A}_0))),$$

where

$$\begin{aligned}(\boldsymbol{B}_{i_1}(\boldsymbol{A}_0))_{l_1,l_2} :=& \mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_3})|Z_{i_1}) \\ &+\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_2,i_3})|Z_{i_1}) \\ &+\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_2})|Z_{i_1}),\end{aligned}$$

and

$$\boldsymbol{\beta}_i(\boldsymbol{A}) := \operatorname{vec}\left(\boldsymbol{B}_i(\boldsymbol{A}) - \mathbb{E}(\boldsymbol{B}_i(\boldsymbol{A}))\right),$$

where $\boldsymbol{B}_i(\boldsymbol{A}) := (\boldsymbol{B}_i(\boldsymbol{A}_1), \ldots, \boldsymbol{B}_i(\boldsymbol{A}_l))$ for

$$\begin{aligned}(\boldsymbol{B}_{i_1}(\boldsymbol{A}_{l'}))_{l_1,l_2} :=& \mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_3})\,\alpha_{l'}(X_{i_1,i_4})|Z_{i_1}) \\ &+\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_2,i_3})\,\alpha_{l'}(X_{i_2,i_4})|Z_{i_1}) \\ &+\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_2})\,\alpha_{l'}(X_{i_2,i_4})|Z_{i_1}) \\ &+\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_2,i_4})\,\alpha_{l'}(X_{i_1,i_2})|Z_{i_1}).\end{aligned}$$

From this a linearization of $\hat{\boldsymbol{G}}$ can be derived. Additional notation is needed in order to state the result.

The insight on which $\hat{\boldsymbol{G}}$ is based is that the $r \times r$ matrices $\boldsymbol{N}_{l'} := \boldsymbol{V}\boldsymbol{A}_{l'}\boldsymbol{V}'$ are diagonalizable in the same (orthonormal) basis. We write $\boldsymbol{Q}$ for the $r \times r$ matrix of their joint eigenvectors and let

$$\boldsymbol{D}_{l'} := \boldsymbol{Q}'\boldsymbol{N}_{l'}\boldsymbol{Q}$$

be the $r \times r$ diagonal matrices that contain their respective eigenvalues. We note that the main diagonal of $\boldsymbol{D}_{l'}$ corresponds to the l'-th row of matrix $\boldsymbol{G}$. If we let $\boldsymbol{D} := (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_l)'$, then

$$\operatorname{vec}(\boldsymbol{G}') = (\boldsymbol{I}_l \otimes \boldsymbol{S})\operatorname{vec}(\boldsymbol{D}'),$$

where $\boldsymbol{S} := (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r)'$ is the $r \times r^2$ selection matrix with $\boldsymbol{s}_z$ the vector whose $((z-1)(r+1)+1)$-th entry is equal to one and all other entries are equal to zero. The shorthand $\boldsymbol{R} := (\boldsymbol{D}_1 \ominus \boldsymbol{D}_1, \ldots, \boldsymbol{D}_l \ominus \boldsymbol{D}_l)'$, where $\boldsymbol{D}_{l'} \ominus \boldsymbol{D}_{l'} := (\boldsymbol{D}_{l'} \otimes \boldsymbol{I}_r) - (\boldsymbol{I}_r \otimes \boldsymbol{D}_{l'})$ is a Kronecker difference, will also be useful.

We have

$$\operatorname{vec}(\hat{\boldsymbol{G}}' - \boldsymbol{G}') = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\beta}_i(\boldsymbol{G}') + o_p(n^{-1/2}),$$

for

$$\boldsymbol{\beta}_i(\boldsymbol{G}') := (\boldsymbol{I}_l \otimes \boldsymbol{S})\left((\boldsymbol{I}_l \otimes \boldsymbol{I}_{r^2}) + \boldsymbol{P}_{\boldsymbol{R}}\right)\left((\boldsymbol{I}_l \otimes \boldsymbol{K}_{r^2})(\boldsymbol{T}_1 + \boldsymbol{T}_2)\boldsymbol{\beta}_i(\boldsymbol{A}_0) + \boldsymbol{T}_3\,\boldsymbol{\beta}_i(\boldsymbol{A})\right),$$

where $\boldsymbol{P}_{\boldsymbol{R}} := \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^*\boldsymbol{R}'$ with a $*$ superscript denoting the Moore-Penrose pseudo inverse of a matrix, $\boldsymbol{K}_{r^2} := \boldsymbol{I}_{r^2} + \boldsymbol{C}_{r^2}$ with $\boldsymbol{C}_{r^2}$ denoting the $r^2 \times r^2$ commutation matrix, and

$$\boldsymbol{T}_1 := -(\boldsymbol{D} \otimes \boldsymbol{I}_r)(\boldsymbol{Q}' \otimes \boldsymbol{Q}')(\boldsymbol{L} \ominus \boldsymbol{L})^*(\boldsymbol{L} \otimes \boldsymbol{I}_r)(\boldsymbol{V} \otimes \boldsymbol{V}),$$

and

$$\boldsymbol{T}_2 := -\frac{1}{2}(\boldsymbol{D} \otimes \boldsymbol{I}_r)(\boldsymbol{Q}' \overset{c}{\otimes} \boldsymbol{Q}')(\boldsymbol{V} \overset{r}{\otimes} \boldsymbol{V}), \qquad \boldsymbol{T}_3 := \boldsymbol{I}_l \otimes (\boldsymbol{Q}' \otimes \boldsymbol{Q}')(\boldsymbol{V} \otimes \boldsymbol{V}).$$

Here, we use $\overset{c}{\otimes}$ and $\overset{r}{\otimes}$ to denote the columnwise and row-wise Kronecker product, respectively.

5.4. *Limit behavior.* Given the large-sample behavior of the first-step estimator it is readily established that

$$\operatorname{vec}(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\beta}_i(\boldsymbol{G}^*) + o_p(n^{-1/2}),$$

for

$$\boldsymbol{\beta}_i(\boldsymbol{G}^*) := (((\boldsymbol{I}_l - \boldsymbol{G}\boldsymbol{G}^*) \otimes (\boldsymbol{G}'\boldsymbol{G})^{-1}) - (\boldsymbol{G}' \otimes \boldsymbol{G})^* \boldsymbol{C}_{rl})\, \boldsymbol{\beta}_i(\boldsymbol{G}').$$

We can now present the asymptotic behavior of our estimators of the main components of the stochastic blockmodel. We state these in the form of two theorems.

We first provide the limit distribution of the estimator $\hat{\boldsymbol{p}}$. The asymptotic variance of this estimator is equal to

$$\boldsymbol{V}_{\boldsymbol{p}} := \mathbb{E}(\boldsymbol{\theta}_i \boldsymbol{\theta}_i'),$$

where

$$\boldsymbol{\theta}_i := \boldsymbol{G}^* \boldsymbol{\beta}_i(\boldsymbol{a}) + (\boldsymbol{a}' \otimes \boldsymbol{I}_r)\, \boldsymbol{\beta}_i(\boldsymbol{G}^*),$$

with $\boldsymbol{\beta}_i(\boldsymbol{a}) := (\beta_i(a_1), \ldots, \beta_i(a_l))'$ for

$$\beta_i(a_{l'}) := 2\,(\mathbb{E}(\alpha_{l'}(X_{i,j})|Z_i) - a_{l'}).$$

Theorem 2 follows.

THEOREM 2. *Suppose that Assumptions 1, 2, and 4 hold. Then*

$$\sqrt{n}\,(\hat{\boldsymbol{p}} - \boldsymbol{p}) \rightsquigarrow N(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{p}}),$$

as $n \to \infty$.

We next state the limit distribution of our estimator of φ_{z_1,z_2}. We will need

$$\boldsymbol{\beta}_i(\boldsymbol{M}_\varphi) := \operatorname{vec}\left(\boldsymbol{B}_i(\boldsymbol{M}_\varphi) - \mathbb{E}(\boldsymbol{B}_i(\boldsymbol{M}_\varphi))\right),$$

with

$$\begin{aligned}(\boldsymbol{B}_{i_1}(\boldsymbol{M}_\varphi))_{l_1,l_2} :=& \mathbb{E}(\alpha_{l_1}(X_{i_1,i_3})\,\varphi(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_2,i_4})|Z_{i_1}) \\ &+\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\varphi(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_4})|Z_{i_1}) \\ &+\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\varphi(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_3,i_4})|Z_{i_1}) \\ &+\mathbb{E}(\alpha_{l_1}(X_{i_3,i_4})\,\varphi(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_2})|Z_{i_1}).\end{aligned}$$

We then have that

$$\boldsymbol{\beta}_i(\boldsymbol{H}_\varphi) := \boldsymbol{K}_{r^2}\,(\boldsymbol{G}^*\boldsymbol{M}_\varphi \otimes \boldsymbol{I}_r)\,\boldsymbol{\beta}_i(\boldsymbol{G}^*) + (\boldsymbol{G}^* \otimes \boldsymbol{G}^*)\,\boldsymbol{\beta}_i(\boldsymbol{M}_\varphi)$$

is the influence function of $\hat{\boldsymbol{H}}_\varphi$. From this, the asymptotic behavior of $\hat{\varphi}_{z_1,z_2}$ will follow after a linearization. Writing $\boldsymbol{I}_r = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_r)$, the influence function of $\hat{\varphi}_{z_1,z_2}$ is

$$\vartheta_i(z_1, z_2) := (p_{z_1}\,p_{z_2})^{-1}(\boldsymbol{e}'_{z_2} \otimes \boldsymbol{e}'_{z_1})\,(\boldsymbol{\beta}_i(\boldsymbol{H}_\varphi) - \varphi_{z_1,z_2}\,\boldsymbol{\beta}_i(\boldsymbol{H}_1)).$$

We let

$$v_\varphi(z_1, z_2) := \mathbb{E}(\vartheta_i(z_1, z_2)^2)$$

in the next theorem.

THEOREM 3. *Suppose that Assumptions 1, 2, 3, and 4 hold. Then,*

$$\sqrt{n}\,(\hat{\varphi}_{z_1,z_2} - \varphi_{z_1,z_2}) \rightsquigarrow N(0, v_\varphi(z_1, z_2)),$$

as $n \to \infty$.

This theorem covers distribution functions, probability mass functions when edge weights are discrete, and moments, for example. More generally, it can be used in combination with standard asymptotic theory to construct estimators of a parameter

$$\arg\max_{\boldsymbol{\delta}} \mathbb{E}(\varphi(X_{i,j};\boldsymbol{\delta})|Z_i = z_1, Z_j = z_2).$$

Under regularity conditions the implied estimator will be $\sqrt{n}$-consistent and asymptotically normal. This formulation is useful as it covers (linear and nonlinear) least-squares problems as well as generic applications of maximum likelihood, for example.

5.5. *Density estimation.* Consider the case where the edge weights are continuous and F_{z_1,z_2} admits a density function, f_{z_1,z_2}, say. Estimation of f_{z_1,z_2} may be of interest. Theorem 3 does not immediately cover this as such nonparametric estimators involve smoothing- or truncation bias. However, the dependence between edge weights reduces these issues to second-order problems. This is in line with the conclusion reached by Graham, Niu and Powell [2019], who considered estimation of the marginal density of the edge weights.

We consider a standard kernel estimator at a point x. The same type of conclusions may be established for locally-linear (or polynomial) versions of the kernel estimator, as well as for estimators based on series expansions. The kernel estimator can be cast into our generic formula for $\hat{\varphi}_{z_1,z_2}$, setting

$$\varphi(X_{i,j}) = \frac{1}{h_n} k\left(\frac{X_{i,j} - x}{h_n}\right),$$

where k is a kernel function and h_n is a non-negative bandwidth.

The following conditions are standard.

ASSUMPTION 3'. *The function k is symmetric, bounded, and integrates to one. The f_{z_1,z_2} are bounded and are twice differentiable with bounded derivatives.*

Assumption 3' replaces Assumption 3.

The dependence that the stochastic blockmodel induces between the edge weights means that the variance of the kernel estimator will be of the order

$$n^{-1} + (n^2 h_n)^{-1} + (n^3 h_n)^{-1} + (n^4 h_n)^{-1}.$$

These terms arise from the covariances between the (symmetrized) kernel of

$$\frac{1}{n(n-1)(n-2)(n-3)} \sum_{i_1 \neq i_2 \neq i_3 \neq i_4} \alpha_{l_1}(X_{i_1,i_2}) \, \frac{1}{h_n} k\left(\frac{X_{i_2,i_3} - x}{h_n}\right) \alpha_{l_2}(X_{i_3,i_4}),$$

evaluated at two different quadruples of nodes that have exactly one, two, three, or all four indices in common. Hence, if $nh_n \to \infty$, the variance is of order n^{-1}, and only terms involving quadruples of nodes that have one index in common contribute to the asymptotic variance. By the usual arguments for U-statistics on graphs [Janson and Nowicki 1991] their contribution is equal to the variance of the sample mean of the projection of the kernel onto the Z_i.

Further, exploiting the fact that the three terms in the kernel above are independent conditional on (Z_{i_2}, Z_{i_3}) it is readily confirmed that a standard argument, as validated by Assumption 3', implies that the expectation of the above statistic is $\sum_{z_1,z_2} (\boldsymbol{G})_{l_1,z_1} f_{z_1,z_2}(x) \, p_{z_1} p_{z_2} \, (\boldsymbol{G})_{l_2,z_2} + O(h_n^2)$. Hence, the smoothing bias of the kernel estimator is of the order h_n^2, which is the conventional result for such a procedure. In light of the variance being of the order n^{-1}, this means that asymptotic bias will be absent provided that $nh_n^4 \to 0$. Taking these observations together leads to the following conclusion.

PROPOSITION 2. *Suppose that Assumptions 1, 2, 3', and 4 hold. Then, if the bandwidth satisfies $nh_n \to \infty$ and $nh_n^4 \to 0$ as $n \to \infty$, Theorem 3 applies to the kernel density estimator.*

Note that, from above, the mean-squared error of the density estimator is of the order

$$h_n^4 + n^{-1} + (n^2 h_n)^{-1}.$$

Equating the rate of the first term to the rate of the second term gives us the optimal-rate requirement $h_n \propto n^{-2/5}$. This is compatible with the condition $nh_n^4 \to 0$ in Proposition 2. Hence, undersmoothing is not needed to prevent asymptotic bias.

6. Numerical experiments. We now provide simulation evidence for the standard blockmodel with binary outcomes and two latent communities. Here, the conditional distributions are fully characterized by their success probabilities and so we consider estimation of

$$\varphi_{z_1,z_2} = \mathbb{P}(X_{i,j} = 1 | Z_i = z_1, Z_j = z_2),$$

along with the relative sizes of the two latent communities, p_1 and p_2. We report results for several combinations of these probabilities. For each, we simulated $10,000$ networks of size $n = 100$, with $p_2 = .70$, and report the mean, median, standard deviation, and interquartile range across the Monte Carlo replications. To give a sense of numerical complexity, estimation of the model for a single replication takes just under $1/3$ of a second on my desktop computer.

Table 1 contains results for three designs that feature complementarity, i.e., success is more likely if agents are from the same community. The three designs vary in how much $\varphi_{1,1}$ is separated from $\varphi_{2,2}$. The specification is peculiar in that agents from different communities never generate successes. We do this to highlight that such a degeneracy does not cause problems for our procedure.

The table shows good performance of our procedure. The conditional distributions are accurately recovered. As $\varphi_{2,2}$ moves further away from $\varphi_{1,1}$ the standard deviation of the estimated success probabilities goes down, as expected. The estimator of the population shares of the communities equally does well across the designs. Its performance is essentially unaffected by the design changes.

TABLE 1
Simulation results

	$\varphi_{1,1}$	$\varphi_{1,2}$	$\varphi_{2,2}$	p_1	p_2
			Design 1		
true value	0.200	0.000	0.400	0.300	0.700
mean	0.220	0.000	0.392	0.285	0.715
median	0.219	0.000	0.392	0.284	0.716
std. dev.	0.029	0.006	0.014	0.044	0.044
iqr	0.039	0.008	0.018	0.059	0.059
			Design 2		
true value	0.200	0.000	0.600	0.300	0.700
mean	0.209	0.000	0.590	0.287	0.713
median	0.209	0.000	0.591	0.287	0.714
std. dev.	0.024	0.003	0.012	0.043	0.043
iqr	0.032	0.004	0.016	0.059	0.059
			Design 3		
true value	0.200	0.000	0.800	0.300	0.700
mean	0.202	0.000	0.789	0.287	0.713
median	0.202	0.000	0.789	0.287	0.713
std. dev.	0.023	0.002	0.009	0.044	0.044
iqr	0.030	0.002	0.012	0.058	0.058

Supplementary material. The proofs of all the technical results are available in the supplement to this paper [Jochmans 2021].

APPENDIX

PROOF OF PROPOSITION 1'. The proof is a modified version of the proof of Theorem 1. It suffices to consider the case where $q_{\min} = 2$, which is the smallest $q_{\min}$ for which the result is different from Proposition 1.

Part (i) of Theorem 1 goes through as a consequence of Kwon and Mbakop [2019, Proposition 2.5] with F_z replaced by F_z^2.

Part (ii) of Theorem 1 goes through by observing that the distributions of certain edge weights in subgraphs involving 7 nodes again have multivariate mixture representations. Indeed, first, the joint distribution of $X_{1,2}, X_{2,3}$ and $X_{1,4}, X_{4,5}$ and $X_{1,6}, X_{6,7}$ factors as

$$\sum_{z=1}^{r} p_z \, F_z^2 \otimes F_z^2 \otimes F_z^2,$$

from which identification of the F_z^2 and p_z follows in the same manner as before. Next, knowledge of the F_z^2 allows to select a set of $l \geq r$ bivariate

functions $\alpha_1, \ldots, \alpha_l$ with which to construct an $l \times r$ matrix $\boldsymbol{G}$ with entries

$$(\boldsymbol{G})_{l',z} := \mathbb{E}(\alpha_{l'}(X_{i,i_1}, X_{i_1,i_2})|Z_i = z)$$

that has full column rank. Then the observable $l \times l$ matrix $\boldsymbol{M}_\varphi$ with entries

$$(\boldsymbol{M}_\varphi)_{l_1,l_2} := \mathbb{E}(\alpha_{l_1}(X_{1,2}, X_{2,5})\, \varphi(X_{1,3})\, \alpha_{l_2}(X_{3,4}, X_{4,6}))$$

again factors as $\boldsymbol{G}\boldsymbol{H}_\varphi\boldsymbol{G}'$, from which we can obtain $\boldsymbol{H}_\varphi$ by re-arrangement as before. This yields identification of

$$\varphi_{z_1,z_2} = (\boldsymbol{H}_1)^{-1}_{z_1,z_2}(\boldsymbol{H}_\varphi)_{z_1,z_2}.$$

As this holds for any function φ we can set $\varphi(x) = \{x \leq x'\}$ for any value x', from which identification of F_{z_1,z_2} follows. □

REFERENCES

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

Allman, E. S., Matias, C. and Rhodes, J. A. (2011). Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference* **141** 1719–1736.

Bickel, P. J., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Annals of Statistics* **41** 1922–1943.

Bonhomme, S., Jochmans, K. and Robin, J. M. (2016a). Estimating multivariate latent-structure models. *Annals of Statistics* **44** 540–563.

Bonhomme, S., Jochmans, K. and Robin, J. M. (2016b). Nonparametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society, Series B* **78** 211–229.

Cardoso, J. F. and Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. *IEEE-Proceedings, F* **140** 362–370.

Celisse, A., Daudin, J. J. and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics* **6** 1847–1899.

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

Graham, B. S., Niu, F. and Powell, J. L. (2019). Kernel density estimation for undirected dyadic data. Mimeo.

Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098.

Holland, P., Laskey, K. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137.

Janson, S. and Nowicki, K. (1991). The asymptotic distributions of generalized U-statistics with applications to random graphs. *Probability Theory and Related Fields* **90** 341–375.

Jochmans, K. (2021). Supplement to Identification and estimation of stochastic blockmodels. Mimeo.

Kasahara, H. and Shimotsu, K. (2014). Nonparametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society, Series B* **76** 97–111.

Kwon, C. and Mbakop, E. (2019). Estimation of the number of components of nonparametric multivariate finite mixture models. Forthcoming in *Annals of Statistics*. Availabe at arXiv:1908.03656 [stat.ME].

Lei, J. (2016). A goodness-of-fit test for stochastic block models. *Annals of Statistics* **44** 401–424.

Levine, M., Hunter, D. R. and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98** 403–416.

Mariadassou, M., Robin, S. and Vacher, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics* **4** 715–742.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Blackwell.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96** 1077–1087.

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

Titterington, D. M. (1983). Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B* **45** 37–46.

Toulouse School of Economics
1 esplanade de l'Université
31080 Toulouse
France
E-mail: koen.jochmans@tse-fr.eu