APPENDIX TO

ESTIMATION AND INFERENCE FOR STOCHASTIC BLOCK MODELS

Koen Jochmans*

Toulouse School of Economics

This version: January 31, 2023

1 Lemmata

This section collects intermediate results that are used in the derivation of the asymptotic behavior of our estimators, along with their proofs. We will routinely exploit the fact that the stochastic block model implies that the collection of edge weights can be represented as

$$X_{i,j} = X(Z_i, Z_j, Y_{i,j}) \tag{1}$$

for a function X that is symmetric in (Z_i, Z_j) and random variables $Y_{i,j}$ that are independent of (Z_i, Z_j) and independent and identically distributed across pairs (i, j).

Lemma 1. *Suppose that Assumption 2 holds. Then*

$$\hat{\boldsymbol{a}} - \boldsymbol{a} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\beta}_i(\boldsymbol{a}) + o_p(n^{-1/2})$$

as $n \to \infty$.

Proof of Lemma 1. Recalling that $X_{i,j} = X_{j,i}$ and using (1), gives

$$\hat{a}_{l'} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i<j} \alpha_{l'}(X_{i,j}) = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{i<j} \alpha_{l'}(X(Z_i, Z_j, Y_{i,j})).$$

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

The final representation reveals that $\hat{a}_{l'}$ is a generalized U-statistic in the sense of Janson and Nowicki (1991). An application of their Lemma 3 yields

$$\hat{a}_{l'} - a_{l'} = \frac{1}{n}\sum_{i=1}^{n} 2\left(\mathbb{E}(\alpha_{l'}(X_{i,j})|Z_i) - a_{l'}\right) + o_p(n^{-1/2}).$$

The summand is equal to the l'-th entry of $\boldsymbol{\beta}_i(\boldsymbol{a})$. The result follows from stacking these equations over l'. □

Lemma 2. *Suppose that Assumption 2 holds. Then*

$$\operatorname{vec}(\hat{\boldsymbol{A}}_0 - \boldsymbol{A}_0) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\beta}_i(\boldsymbol{A}_0) + o_p(n^{-1/2})$$

as $n \to \infty$.

Proof of Lemma 2. We start by symmetrizing the kernel of

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

as

$$\begin{aligned}\alpha_{l_1,l_2}(X_{i_1,i_2}, X_{i_1,i_3}, X_{i_2,i_3}) :=& \frac{\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_3})}{6} + \frac{\alpha_{l_1}(X_{i_1,i_3})\,\alpha_{l_2}(X_{i_1,i_2})}{6} \\ &+ \frac{\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_2,i_3})}{6} + \frac{\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_2})}{6} \\ &+ \frac{\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_3})}{6} + \frac{\alpha_{l_1}(X_{i_1,i_3})\,\alpha_{l_2}(X_{i_2,i_3})}{6}\end{aligned}$$

to write

$$(\hat{\boldsymbol{A}}_0)_{l_1,l_2} = \binom{n}{3}^{-1} \sum_{i_1 < i_2 < i_3} \alpha_{l_1,l_2}(X_{i_1,i_2}, X_{i_1,i_3}, X_{i_2,i_3}).$$

This is a generalized U-statistic. By (1) its kernel is a symmetric function in the variables $Z_{i_1}, Z_{i_2}, Z_{i_3}$ and $Y_{i_1,i_2}, Y_{i_1,i_3}, Y_{i_2,i_3}$, and it is easy to see that its variance is finite. Janson and Nowicki's [1991] Lemma 3 can be applied to deduce that $(\hat{\boldsymbol{A}}_0)_{l_1,l_2} - (\boldsymbol{A}_0)_{l_1,l_2}$ equals

$$\frac{3}{n}\sum_{i=1}^{n}(\mathbb{E}(\alpha_{l_1,l_2}(X_{i_1,i_2}, X_{i_1,i_3}, X_{i_2,i_3})|Z_{i_1}) - (\boldsymbol{A}_0)_{l_1,l_2}) + o_p(n^{-1/2}).$$

It is readily verified that

$$\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_3})|Z_{i_1}) = \mathbb{E}(\alpha_{l_1}(X_{i_1,i_3})\,\alpha_{l_2}(X_{i_1,i_2})|Z_{i_1}),$$
$$\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_2,i_3})|Z_{i_1}) = \mathbb{E}(\alpha_{l_1}(X_{i_1,i_3})\,\alpha_{l_2}(X_{i_2,i_3})|Z_{i_1}),$$
$$\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_2})|Z_{i_1}) = \mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_3})|Z_{i_1}),$$

as each of the expectations on the left-hand side is invariant to a permutation of the indices (i_2, i_3). Hence,

$$\mathbb{E}(\alpha_{l_1,l_2}(X_{i_1,i_2}, X_{i_1,i_3}, X_{i_2,i_3})|Z_{i_1})$$

is equal to

$$\frac{\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\alpha_{l_2}(X_{i_1,i_3}) + \alpha_{l_1}(X_{i_1,i_2})\alpha_{l_2}(X_{i_2,i_3}) + \alpha_{l_1}(X_{i_2,i_3})\alpha_{l_2}(X_{i_1,i_2})|Z_{i_1})}{3}.$$

This is equal to the (l_1, l_2)-th entry of $\boldsymbol{B}_{i_1}(\boldsymbol{A}_0)$, up to the factor $\frac{1}{3}$. Hence, collecting terms over (l_1, l_2) and noting that $\mathbb{E}(\boldsymbol{B}_i(\boldsymbol{A}_0)) = 3\boldsymbol{A}_0$ we arrive at

$$\operatorname{vec}(\hat{\boldsymbol{A}}_0 - \boldsymbol{A}_0) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\beta}_i(\boldsymbol{A}_0) + o_p(n^{-1/2}),$$

as claimed. □

Lemma 3. *Suppose that Assumption 2 holds. Then*

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

as $n \to \infty$.

Proof of Lemma 3. The proof follows the same steps as the proof of Lemma 2. Start with the estimator $\hat{\boldsymbol{A}}_{l'}$ for a given l', which is equal to

$$\frac{1}{n(n-1)(n-2)(n-3)} \sum_{i_1 \neq i_2 \neq i_3 \neq i_4} \alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_3})\,\alpha_{l'}(X_{i_1,i_4}).$$

Its kernel can be symmetrized by considering all possible rearrangements of the indices (i_1, i_2, i_3, i_4). This yields $4! = 24$ terms. It is easy to see, however, that the projection of many of these terms onto Z_{i_1} co-incide. First, there are six terms where the edge weights all have the index i_1 in common. Their projections are all equal to

$$\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_3})\,\alpha_{l'}(X_{i_1,i_4})|Z_{i_1}).$$

Next, their are six terms where the edge weights all have the index i_2 in common. The index i_1 enters the kernel either through α_{l_1}, α_{l_2}, or $\alpha_{l'}$; each of these configurations occurs twice. The corresponding projections thus are

$$\begin{aligned}
&\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_2,i_3})\,\alpha_{l'}(X_{i_2,i_4})|Z_{i_1}),\\
&\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_2})\,\alpha_{l'}(X_{i_2,i_4})|Z_{i_1}),\\
&\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_2,i_4})\,\alpha_{l'}(X_{i_1,i_2})|Z_{i_1}),
\end{aligned}$$

respectively. There are an additional 12 terms in the symmetrized kernel, those where the common index is either i_3 or i_4. Their projections are equal to those where the common index is i_2, however. Consequently, if we collect terms as

$$\begin{aligned}
(\boldsymbol{B}_{i_1}(\boldsymbol{A}_{l'}))_{l_1,l_2} = {} & \frac{\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_1,i_3})\,\alpha_{l'}(X_{i_1,i_4})|Z_{i_1})}{4}\\
&+\frac{\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\,\alpha_{l_2}(X_{i_2,i_3})\,\alpha_{l'}(X_{i_2,i_4})|Z_{i_1})}{4}\\
&+\frac{\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_1,i_2})\,\alpha_{l'}(X_{i_2,i_4})|Z_{i_1})}{4}\\
&+\frac{\mathbb{E}(\alpha_{l_1}(X_{i_2,i_3})\,\alpha_{l_2}(X_{i_2,i_4})\,\alpha_{l'}(X_{i_1,i_2})|Z_{i_1})}{4},
\end{aligned}$$

we have

$$\operatorname{vec}(\hat{\boldsymbol{A}}_{l'}-\boldsymbol{A}_{l'}) = \frac{4}{n}\sum_{i=1}^{n}\operatorname{vec}(\boldsymbol{B}_{i_1}(\boldsymbol{A}_{l'})-\boldsymbol{A}_{l'}) + o_p(n^{-1/2}).$$

Stacking over l' and noting that $\boldsymbol{B}_i(\boldsymbol{A}) = 4\,(\boldsymbol{B}_1(\boldsymbol{A}_1),\ldots,\boldsymbol{B}_1(\boldsymbol{A}_l))$ we have

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

where, recall, $\boldsymbol{A} = (\boldsymbol{A}_1,\ldots,\boldsymbol{A}_l)$ and $\boldsymbol{\beta}_i(\boldsymbol{A}) = \operatorname{vec}(\boldsymbol{B}_i(\boldsymbol{A}) - \mathbb{E}(\boldsymbol{B}_i(\boldsymbol{A})))$. □

Lemma 4. *Suppose that Assumptions 2 and 3 hold. Then*

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

as $n\to\infty$.

Proof of Lemma 4. The proof follows the same steps as the proof of Lemma 3, but with $\alpha_{l'}$ set to φ and the order of the indices altered accordingly. $\square$

Lemma 5. *Let* $\boldsymbol{N} := (\boldsymbol{N}_1, \ldots, \boldsymbol{N}_l)$ *and* $\hat{\boldsymbol{N}} := (\hat{\boldsymbol{N}}_1, \ldots, \hat{\boldsymbol{N}}_l)$*. Suppose that Assumptions 2 and 4 hold. Then, as* $n \to \infty$,

$$\operatorname{vec}(\hat{\boldsymbol{N}} - \boldsymbol{N}) = \frac{1}{n}\sum_{i=1}^{n}((\boldsymbol{I}_l \otimes \boldsymbol{K}_{r^2})\boldsymbol{W}\,\boldsymbol{\beta}_i(\boldsymbol{A}_0) + (\boldsymbol{I}_l \otimes \boldsymbol{V} \otimes \boldsymbol{V})\,\boldsymbol{\beta}_i(\boldsymbol{A})) + o_p(n^{-1/2}),$$

where $\boldsymbol{W}_{l'} := -(\boldsymbol{N}_{l'} \otimes \boldsymbol{I}_r)\,(\boldsymbol{L} \ominus \boldsymbol{L})^*(\boldsymbol{L}\boldsymbol{V} \otimes \boldsymbol{V}) - \frac{1}{2}(\boldsymbol{N}_{l'} \overset{c}{\otimes} \boldsymbol{I}_r)\,(\boldsymbol{V} \overset{r}{\otimes} \boldsymbol{V})$ *and* $\boldsymbol{W} := (\boldsymbol{W}_1', \ldots, \boldsymbol{W}_l')'$.

Proof of Lemma 5. Recall that

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

We have $\hat{\boldsymbol{N}}_{l'} - \boldsymbol{N}_{l'} = O_p(n^{-1/2})$ because $\hat{\boldsymbol{A}}_{l'} - \boldsymbol{A}_{l'} = O_p(n^{-1/2})$ by Lemma 3 below, and $\hat{\boldsymbol{V}} - \boldsymbol{V} = O_p(n^{-1/2})$ by Lemma 2 combined with Lemma S.2 of Bonhomme, Jochmans and Robin (2016). Moreover, a linearization gives

$$\hat{\boldsymbol{N}}_{l'} - \boldsymbol{N}_{l'} = \boldsymbol{V}(\hat{\boldsymbol{A}}_{l'} - \boldsymbol{A}_{l'})\boldsymbol{V}' + (\hat{\boldsymbol{V}} - \boldsymbol{V})\boldsymbol{A}_{l'}\boldsymbol{V}' + \boldsymbol{V}\boldsymbol{A}_{l'}(\hat{\boldsymbol{V}} - \boldsymbol{V})' + o_p(n^{-1/2}).$$

By elementary properties of the Kronecker product and the commutation matrix $\operatorname{vec}(\hat{\boldsymbol{N}}_{l'} - \boldsymbol{N}_{l'})$ equals

$$(\boldsymbol{V} \otimes \boldsymbol{V})\operatorname{vec}(\hat{\boldsymbol{A}}_{l'} - \boldsymbol{A}_{l'}) + \boldsymbol{K}_{r^2}(\boldsymbol{V}\boldsymbol{A}_{l'} \otimes \boldsymbol{I}_r)\operatorname{vec}(\hat{\boldsymbol{V}} - \boldsymbol{V}) + o_p(n^{-1/2}).$$

Using Lemma S.2 of Bonhomme, Jochmans and Robin (2016) in tandem with Lemma 2 we have that

$$(\boldsymbol{V}\boldsymbol{A}_{l'} \otimes \boldsymbol{I}_r)\operatorname{vec}(\hat{\boldsymbol{V}} - \boldsymbol{V}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_{l'}\,\boldsymbol{\beta}_i(\boldsymbol{A}_0) + o_p(n^{-1/2}),$$

while

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

follows from Lemma 3. Collecting terms yields

$$\begin{aligned}\operatorname{vec}(\hat{\boldsymbol{N}} - \boldsymbol{N}) = \frac{1}{n}\sum_{i=1}^{n}&((\boldsymbol{I}_l \otimes \boldsymbol{K}_{r^2})\boldsymbol{W}\boldsymbol{\beta}_i(\boldsymbol{A}_0) \\ &+(\boldsymbol{I}_l \otimes \boldsymbol{V}' \otimes \boldsymbol{V}')\boldsymbol{\beta}_i(\boldsymbol{A})) + o_p(n^{-1/2}),\end{aligned}$$

for $\boldsymbol{W} := (\boldsymbol{W}_1', \ldots, \boldsymbol{W}_l')'$, as claimed. $\square$

Lemma 6. *Suppose that Assumptions 2 and 4 hold. Then*

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

as $n \to \infty$.

Proof of Lemma 6. Let

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

and recall that $\boldsymbol{D}_{l'} = \boldsymbol{Q}' \boldsymbol{N}_{l'} \boldsymbol{Q}$. A linearization gives

$$\hat{\boldsymbol{D}}_{l'} - \boldsymbol{D}_{l'} = \boldsymbol{Q}'(\hat{\boldsymbol{N}}_{l'} - \boldsymbol{N}_{l'})\boldsymbol{Q} + (\hat{\boldsymbol{Q}} - \boldsymbol{Q})'\boldsymbol{N}_{l'}\boldsymbol{Q} + \boldsymbol{Q}'\boldsymbol{N}_{l'}(\hat{\boldsymbol{Q}} - \boldsymbol{Q}) + o_p(n^{-1/2}),$$

with the order of the remainder term following from Lemma 3 and from Theorem 5 of Bonhomme and Robin (2009), which is applicable in the present setting by our Lemmas 2 and 3. By an application of the delta method,

$$\operatorname{vec}(\hat{\boldsymbol{Q}}' - \boldsymbol{Q}') = -(\boldsymbol{Q} \otimes \boldsymbol{Q}') \operatorname{vec}(\hat{\boldsymbol{Q}} - \boldsymbol{Q}) + o_p(n^{-1/2}),$$

which, after some re-arrangement, allows us to write

$$\begin{aligned}\operatorname{vec}(\hat{\boldsymbol{D}}_{l'} - \boldsymbol{D}_{l'}) = {} & -(\boldsymbol{D}_{l'} \ominus \boldsymbol{D}_{l'})(\boldsymbol{I}_r \otimes \boldsymbol{Q}') \operatorname{vec}(\hat{\boldsymbol{Q}} - \boldsymbol{Q}) \\ & + (\boldsymbol{Q}' \otimes \boldsymbol{Q}')\operatorname{vec}(\hat{\boldsymbol{N}}_{l'} - \boldsymbol{N}_{l'}) + o_p(n^{-1/2}).\end{aligned}$$

Recalling that $\boldsymbol{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_l)'$ and that $\boldsymbol{R} = (\boldsymbol{D}_1 \ominus \boldsymbol{D}_1, \ldots, \boldsymbol{D}_l \ominus \boldsymbol{D}_l)'$, this gives

$$\operatorname{vec}(\hat{\boldsymbol{D}}' - \boldsymbol{D}') = -\boldsymbol{R}(\boldsymbol{I}_r \otimes \boldsymbol{Q}')\operatorname{vec}(\hat{\boldsymbol{Q}} - \boldsymbol{Q}) + (\boldsymbol{I}_l \otimes \boldsymbol{Q}' \otimes \boldsymbol{Q}')\operatorname{vec}(\hat{\boldsymbol{N}} - \boldsymbol{N}) + o_p(n^{-1/2}).$$

For the first of these right-hand side terms we can rely on Theorem 5 of Bonhomme and Robin (2009), which implies that

$$\operatorname{vec}(\hat{\boldsymbol{Q}} - \boldsymbol{Q}) = -(\boldsymbol{I}_r \otimes \boldsymbol{Q})\,(\boldsymbol{R}'\boldsymbol{R})^{*}\boldsymbol{R}'(\boldsymbol{I}_l \otimes \boldsymbol{Q}' \otimes \boldsymbol{Q}')\operatorname{vec}(\hat{\boldsymbol{N}} - \boldsymbol{N}) + o_p(n^{-1/2}).$$

Plugging this expression into the equation for $\operatorname{vec}(\hat{\boldsymbol{D}}' - \boldsymbol{D}')$ and using the shorthand $\boldsymbol{P}_{\boldsymbol{R}} = \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{*}\boldsymbol{R}'$ gives

$$\operatorname{vec}(\hat{\boldsymbol{D}}' - \boldsymbol{D}') = (\boldsymbol{I}_l \otimes \boldsymbol{I}_{r^2} + \boldsymbol{P}_{\boldsymbol{R}})\,(\boldsymbol{I}_l \otimes \boldsymbol{Q}' \otimes \boldsymbol{Q}')\operatorname{vec}(\hat{\boldsymbol{N}} - \boldsymbol{N}) + o_p(n^{-1/2}).$$

The result then follows from Lemma 5, where the influence function of $\mathrm{vec}(\hat{\boldsymbol{N}} - \boldsymbol{N})$ is given, together with

$$\mathrm{vec}(\hat{\boldsymbol{G}}^{'} - \boldsymbol{G}^{'}) = (\boldsymbol{I}_l \otimes \boldsymbol{S})\,\mathrm{vec}(\hat{\boldsymbol{D}}^{'} - \boldsymbol{D}^{'}).$$

To see this observe that, because

$$(\boldsymbol{Q}' \otimes \boldsymbol{Q}')\,(\boldsymbol{N}_{l'} \otimes \boldsymbol{I}_r) = (\boldsymbol{D}_{l'} \otimes \boldsymbol{I}_r)\,(\boldsymbol{Q}' \otimes \boldsymbol{Q}')$$

and

$$(\boldsymbol{Q}' \otimes \boldsymbol{Q}')\,(\boldsymbol{N}_{l'} \overset{c}{\otimes} \boldsymbol{I}_r) = (\boldsymbol{D}_{l'} \otimes \boldsymbol{I}_r)\,(\boldsymbol{Q}' \overset{c}{\otimes} \boldsymbol{Q}'),$$

we have $(\boldsymbol{I}_l \otimes \boldsymbol{Q}' \otimes \boldsymbol{Q}')\,\boldsymbol{W} = \boldsymbol{T}_1 + \boldsymbol{T}_2$, where $\boldsymbol{W}$ is defined in Lemma 5. □

Lemma 7. *Suppose that Assumptions 2 and 4 hold and that $\boldsymbol{G}$ has maximal column rank. Then*

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

as $n \to \infty$.

Proof of Lemma 7. We start with

$$\hat{\boldsymbol{G}}^{*} - \boldsymbol{G}^{*} = (\boldsymbol{G}'\boldsymbol{G})^{-1}(\hat{\boldsymbol{G}}^{'} - \boldsymbol{G}^{'}) + ((\hat{\boldsymbol{G}}^{'}\hat{\boldsymbol{G}})^{-1} - (\boldsymbol{G}'\boldsymbol{G})^{-1})\,\boldsymbol{G}' + o_p(n^{-1/2}),$$

where the order of the remainder term follows from Lemma 6. The first term on the right-hand side equals

$$\mathrm{vec}((\boldsymbol{G}'\boldsymbol{G})^{-1}(\hat{\boldsymbol{G}}^{'} - \boldsymbol{G}^{'})) = (\boldsymbol{I}_l \otimes (\boldsymbol{G}'\boldsymbol{G})^{-1})\,\mathrm{vec}(\hat{\boldsymbol{G}}^{'} - \boldsymbol{G}^{'}) + o_p(n^{-1/2})$$

by the same lemma. For the second right-hand side term we similarly obtain

$$\begin{aligned}\mathrm{vec}(((\hat{\boldsymbol{G}}^{'}\hat{\boldsymbol{G}})^{-1} - (\boldsymbol{G}'\boldsymbol{G})^{-1})\,\boldsymbol{G}') = &- (\boldsymbol{G}\boldsymbol{G}^{*} \otimes (\boldsymbol{G}'\boldsymbol{G})^{-1}))\mathrm{vec}(\hat{\boldsymbol{G}}^{'} - \boldsymbol{G}^{'}) \\ &- (\boldsymbol{G}' \otimes \boldsymbol{G})^{*}\mathrm{vec}(\hat{\boldsymbol{G}} - \boldsymbol{G}) + o_p(n^{-1/2}).\end{aligned}$$

This expression follows from an application of the delta method to get

$$\begin{aligned}\mathrm{vec}((\hat{\boldsymbol{G}}^{'}\hat{\boldsymbol{G}})^{-1} - (\boldsymbol{G}'\boldsymbol{G})^{-1}) = &- ((\boldsymbol{G}'\boldsymbol{G})^{-1} \otimes (\boldsymbol{G}'\boldsymbol{G})^{-1}) \\ &\times \mathrm{vec}((\hat{\boldsymbol{G}}^{'}\hat{\boldsymbol{G}}) - (\boldsymbol{G}'\boldsymbol{G})) + o_p(n^{-1/2}),\end{aligned}$$

together with the linearization

$$\begin{aligned}\operatorname{vec}((\hat{\boldsymbol{G}}'\hat{\boldsymbol{G}}) - (\boldsymbol{G}'\boldsymbol{G})) = (\boldsymbol{G}' \otimes \boldsymbol{I}_r)\operatorname{vec}(\hat{\boldsymbol{G}}' - \boldsymbol{G}') \\ + (\boldsymbol{I}_r \otimes \boldsymbol{G}')\operatorname{vec}(\hat{\boldsymbol{G}} - \boldsymbol{G}) + o_p(n^{-1/2}).\end{aligned}$$

Collecting terms yields

$$\begin{aligned}\operatorname{vec}(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*) = ((\boldsymbol{I}_l - \boldsymbol{G}\boldsymbol{G}^*) \otimes (\boldsymbol{G}'\boldsymbol{G})^{-1})\operatorname{vec}(\hat{\boldsymbol{G}}' - \boldsymbol{G}') \\ - (\boldsymbol{G}' \otimes \boldsymbol{G})^*\operatorname{vec}(\hat{\boldsymbol{G}} - \boldsymbol{G}) + o_p(n^{-1/2}).\end{aligned}$$

The final conclusion follows from $\operatorname{vec}(\hat{\boldsymbol{G}} - \boldsymbol{G}) = \boldsymbol{C}_{lr}\operatorname{vec}(\hat{\boldsymbol{G}}' - \boldsymbol{G}')$. □

2 Omitted proofs

This section collects the proofs of the main results in the paper.

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 continues to go through as a consequence of Kwon and Mbakop (2021, Proposition 2.5), only now 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. First, the joint distribution of the three pairs $X_{1,2}, X_{2,3}$ 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 $\alpha_1, \ldots, \alpha_l$ that may be used 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. □

Proof of Theorem 2. By a linearization,

$$\hat{\boldsymbol{p}} - \boldsymbol{p} = \hat{\boldsymbol{G}}^* \hat{\boldsymbol{a}} - \boldsymbol{G}^* \boldsymbol{a} = \boldsymbol{G}^*(\hat{\boldsymbol{a}} - \boldsymbol{a}) + (\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*)\, \boldsymbol{a} + o_p(n^{-1/2}),$$

where the order of the remainder term follows from Lemmas 1 and 7. By the same lemmas,

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

and

$$(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*)\, \boldsymbol{a} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{a}' \otimes \boldsymbol{I}_r)\, \boldsymbol{\beta}_i(\boldsymbol{G}^*) + o_p(n^{-1/2}),$$

because $(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*)\, \boldsymbol{a} = (\boldsymbol{a}' \otimes \boldsymbol{I}_r)\, \mathrm{vec}(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*)$ by an elementary relationship between the Kronecker product and the vec operator, which we rely on frequently throughout the derivations in this appendix. On collecting terms,

$$\hat{\boldsymbol{p}} - \boldsymbol{p} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta}_i + o_p(n^{-1/2})$$

follows, as asserted in the theorem. □

Proof of Theorem 3. First we linearize

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

as

$$(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*)\boldsymbol{M}_\varphi(\boldsymbol{G}^*)' + \boldsymbol{G}^*(\hat{\boldsymbol{M}}_\varphi - \boldsymbol{M}_\varphi)(\boldsymbol{G}^*)' + \boldsymbol{G}^*\boldsymbol{M}_\varphi(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*)' + o_p(n^{-1/2}).$$

The order of the remainder term, as well as the linearizations stated next, follow from Lemmas 4 and 7. The linearizations in question take the form

$$(\boldsymbol{G}^*\boldsymbol{M}_\varphi \otimes \boldsymbol{I}_r)\,\mathrm{vec}(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*) = (\boldsymbol{G}^*\boldsymbol{M}_\varphi \otimes \boldsymbol{I}_r)\,\frac{1}{n}\sum_{i=1}^n \boldsymbol{\beta}_i(\boldsymbol{G}^*) + o_p(n^{-1/2}),$$

and

$$(\boldsymbol{G}^* \otimes \boldsymbol{G}^*)\,\mathrm{vec}(\hat{\boldsymbol{M}}_\varphi - \boldsymbol{M}_\varphi) = (\boldsymbol{G}^* \otimes \boldsymbol{G}^*)\,\frac{1}{n}\sum_{i=1}^n \boldsymbol{\beta}_i(\boldsymbol{M}_\varphi) + o_p(n^{-1/2}).$$

Also, $(\boldsymbol{I}_r \otimes \boldsymbol{G}^*\boldsymbol{M}_\varphi)\,\mathrm{vec}((\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*)') = (\boldsymbol{I}_r \otimes \boldsymbol{G}^*\boldsymbol{M}_\varphi)\,\boldsymbol{C}_{r^2}\,\mathrm{vec}(\hat{\boldsymbol{G}}^* - \boldsymbol{G}^*)$ by the defining property of the commutation matrix. Therefore, on collecting terms

$$\mathrm{vec}(\hat{\boldsymbol{H}}_\varphi - \boldsymbol{H}_\varphi) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\beta}_i(\boldsymbol{H}_\varphi) + o_p(n^{-1/2})$$

with

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

As $\boldsymbol{e}_{z_1}'(\hat{\boldsymbol{H}}_\varphi - \boldsymbol{H}_\varphi)\,\boldsymbol{e}_{z_2} = (\boldsymbol{e}_{z_2}' \otimes \boldsymbol{e}_{z_1}')\,\mathrm{vec}(\hat{\boldsymbol{H}}_\varphi - \boldsymbol{H}_\varphi)$ we have therefore obtained

$$\boldsymbol{e}_{z_1}'(\hat{\boldsymbol{H}}_\varphi - \boldsymbol{H}_\varphi)\boldsymbol{e}_{z_2} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{e}_{z_2}' \otimes \boldsymbol{e}_{z_1}')\,\boldsymbol{\beta}_i(\boldsymbol{H}_\varphi) + o_p(n^{-1/2}).$$

This holds for any function φ, and thus also for $\varphi(x) = 1$. A linearization of

$$\hat{\varphi}_{z_1,z_2} = \frac{\boldsymbol{e}_{z_1}'\hat{\boldsymbol{H}}_\varphi\boldsymbol{e}_{z_2}}{\boldsymbol{e}_{z_1}'\hat{\boldsymbol{H}}_1\boldsymbol{e}_{z_2}}$$

then gives

$$\hat{\varphi}_{z_1,z_2} - \varphi_{z_1,z_2} = \frac{\boldsymbol{e}_{z_1}'(\hat{\boldsymbol{H}}_\varphi - \boldsymbol{H}_\varphi)\boldsymbol{e}_{z_2}}{\boldsymbol{e}_{z_1}'\boldsymbol{H}_1\boldsymbol{e}_{z_2}} - \varphi_{z_1,z_2}\,\frac{\boldsymbol{e}_{z_1}'(\hat{\boldsymbol{H}}_1 - \boldsymbol{H}_1)\boldsymbol{e}_{z_2}}{\boldsymbol{e}_{z_1}'\boldsymbol{H}_1\boldsymbol{e}_{z_2}} + o_p(n^{-1/2}).$$

The result then follows on using that $\boldsymbol{e}_{z_1}'\boldsymbol{H}_1\boldsymbol{e}_{z_2} = p_{z_1}\,p_{z_2}$. $\square$

3 Computational notes

The matrices that serve as inputs to our estimator of the stochastic block model are all U-statistics of order two up to four. Brute force calculation of these statistics (and of their influence functions) is time consuming and can be prohibitive for large n. Fortunately, such calculations can be avoided.

To illustrate, consider

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

for fixed (l_1, l_2). The complexity of direct evaluation of this expression is n^3. Further, as

$$((\hat{\boldsymbol{A}}_0)_{l_1,l_2} - (\boldsymbol{A}_0)_{l_1,l_2}) = \frac{1}{n} \sum_{i=1}^{n} ((\boldsymbol{B}_{i_1}(\boldsymbol{A}_0))_{l_1,l_2} - 3(\boldsymbol{A}_0)_{l_1,l_2})) + o_p(n^{-1/2}),$$

where, recall, $(\boldsymbol{B}_{i_1}(\boldsymbol{A}_0))_{l_1,l_2}$ takes the form

$$\mathbb{E}(\alpha_{l_1}(X_{i_1,i_2})\, \alpha_{l_2}(X_{i_1,i_3}) + \alpha_{l_1}(X_{i_1,i_2})\, \alpha_{l_2}(X_{i_2,i_3}) + \alpha_{l_1}(X_{i_2,i_3})\, \alpha_{l_2}(X_{i_1,i_2}) | Z_{i_1})$$

(see Lemma 2), inference also requires to estimate this projection for each node. A natural estimator is

$$(\hat{\boldsymbol{B}}_{i_1}(\boldsymbol{A}_0))_{l_1,l_2} := a_{i_1} + b_{i_1} + c_{i_1},$$

for

$$a_{i_1} := \frac{1}{(n-1)(n-2)} \sum_{i_2 \neq i_1} \sum_{i_3 \neq i_1, i_2} \alpha_{l_1}(X_{i_1,i_2})\, \alpha_{l_2}(X_{i_1,i_3}),$$
$$b_{i_1} := \frac{1}{(n-1)(n-2)} \sum_{i_2 \neq i_1} \sum_{i_3 \neq i_1, i_2} \alpha_{l_1}(X_{i_1,i_2})\, \alpha_{l_2}(X_{i_2,i_3}),$$
$$c_{i_1} := \frac{1}{(n-1)(n-2)} \sum_{i_2 \neq i_1} \sum_{i_3 \neq i_1, i_2} \alpha_{l_1}(X_{i_2,i_3})\, \alpha_{l_2}(X_{i_1,i_2}).$$

Direct calculation for all n again involves $O(n^3)$ operations.

Observe that

$$
\begin{aligned}
a_{i_1} &= \frac{1}{(n-1)} \sum_{i_2 \neq i_1} \alpha_{l_1}(X_{i_1,i_2}) \, \frac{1}{(n-2)} \sum_{i_3 \neq i_1, i_2} \alpha_{l_2}(X_{i_1,i_3}) \\
&= \frac{1}{(n-1)} \sum_{i_2 \neq i_1} \alpha_{l_1}(X_{i_1,i_2}) \, \frac{1}{(n-2)} \left(\sum_{i_3 \neq i_1} \alpha_{l_2}(X_{i_1,i_3}) - \alpha_{l_2}(X_{i_1,i_2}) \right) \\
&= \frac{1}{(n-1)(n-2)} \left(\sum_{i_2 \neq i_1} \alpha_{l_1}(X_{i_1,i_2}) \right) \left(\sum_{i_3 \neq i_1} \alpha_{l_2}(X_{i_1,i_3}) \right) \\
&\quad - \frac{1}{(n-1)(n-2)} \left(\sum_{i_2 \neq i_1} \alpha_{l_1}(X_{i_1,i_2}) \, \alpha_{l_2}(X_{i_1,i_2}) \right).
\end{aligned}
$$

The last expression reveals that we only need to calculate the sums

$$
\sum_{i_2 \neq i_1} \alpha_{l_1}(X_{i_1,i_2}), \quad \sum_{i_2 \neq i_1} \alpha_{l_2}(X_{i_1,i_2}), \quad \text{and} \quad \sum_{i_2 \neq i_1} \alpha_{l_1}(X_{i_1,i_2}) \, \alpha_{l_2}(X_{i_1,i_2}),
$$

for each node to obtain the $(\hat{\boldsymbol{B}}_{i_1}(\boldsymbol{A}_0))_{l_1,l_2}$. This is immediate in any matrix-based language. Further, with $(\hat{\boldsymbol{B}}_{i_1}(\boldsymbol{A}_0))_{l_1,l_2}$ at hand we can compute

$$
(\hat{\boldsymbol{A}}_0)_{l_1,l_2} = \frac{1}{3} \frac{1}{n} \sum_{i=1}^{n} (\hat{\boldsymbol{B}}_{i_1}(\boldsymbol{A}_0))_{l_1,l_2}.
$$

Consequently, we both the estimator and its influence function can be obtained in the same calculation. In this way the computational cost is linear in n, making our estimator fast to compute, even for large n.

References

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

Bonhomme, S. and J.-M. Robin (2009). Consistent noisy independent component analysis. *Journal of Econometrics 149*, 12–25.

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

Kwon, C. and E. Mbakop (2021). Estimation of the number of components of non-parametric multivariate finite mixture models. *Annals of Statistics 49*, 2178–2205.