Supplementary Material: Inference for network regression models with community structure

A Proof of Theorem 5.1

We first restate Theorem 5.1, then provide a complete proof.

Assume (a) the error vector satisfies the block-exchangeability assumption, with two blocks of sizes n_1 and n_2, (b) $\boldsymbol{X}$ is a full rank $(n(n-1) \times 2)$ matrix, (c) covariates $\{X_{ij}\}$ are independent and identically distributed, (d) the fourth moment of the errors and covariates are bounded, (e) errors Ξ and $\boldsymbol{X}$ are independent, and (f) the number of blocks B is $\mathcal{O}(1)$. As $n_1 \to \infty, n_2 \to \infty$, and $n_1/n_2 \to \alpha$, where α is a constant such that $0 < \alpha < \infty$,

$$n\left(\hat{V}_B(\hat{\boldsymbol{\beta}}) - \hat{V}_E(\hat{\boldsymbol{\beta}})\right) \xrightarrow{p} c(\boldsymbol{X}). \tag{1}$$

where $c(\boldsymbol{X})$ is a weighted linear combination of the differences between the true block exchangeable parameters and corresponding exchangeable parameters when the block exchangeable parameters are appropriately averaged within configuration type and convergence is pointwise. Furthermore, when X_{ij} is independent of g_i and g_j, $c(\boldsymbol{X}) = \boldsymbol{0}$ and thus the estimators are asymptotically equivalent.

We now proceed with the proof. We begin by defining $c(\boldsymbol{X})$:

$$\begin{aligned} c(\boldsymbol{X}) &= \sum_{M,q \in Q_M} f_{M,q}(M_q - M) \\ &= \sum_{u,v\in\{1,2\}} f_{\sigma^2,(u,v)}(\sigma^2_{(u,v)} - \sigma^2) + \sum_{u,v\in\{1,2\}} f_{\phi_A,\{u,v\}}(\phi_{A,\{u,v\}}) - \phi_A) + ... \end{aligned} \tag{2}$$

where $f_{M,q}$ are functions of $\boldsymbol{X}$. More specifically, given M and q, $f_{M,q}$ is a function of elements in the set $\{[X_{ij}, X_{kl}] | [(i,j),(k,l)] \in \Phi_{M,q}\}$. The parameter

$$\sigma^2 = \frac{n_1(n_1-1)\sigma^2_{(1,1)} + n_2(n_2-1)\sigma^2_{(2,2)} + n_1 n_2(\sigma^2_{(1,2)} + \sigma^2_{(2,1)})}{n(n-1)} \tag{3}$$

We now present a proof of Theorem 5.1.

$$\begin{aligned}
&n\left(\hat{V}_B(\hat{\boldsymbol{\beta}}) - \hat{V}_E(\hat{\boldsymbol{\beta}})\right)\\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T(\widehat{\Omega}_B - \widehat{\Omega}_E)X(\boldsymbol{X}^T\boldsymbol{X})^{-1}\\
&= \frac{n}{n^2(n-1)^2}\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n(n-1)}\right)^{-1}\left(\sum_{M\in\mathcal{M}}\sum_{q\in Q_M}\frac{\sum\limits_{(j,k),(m,n)\in\Phi_{M,q}}\boldsymbol{X}_{jk}\boldsymbol{X}_{mn}^T\left(\widehat{M_q}-\widehat{M}\right)|\Phi_{M,q}|}{|\Phi_{M,q}|}\right)\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n(n-1)}\right)^{-1}\\
&= \sum_{M\in\mathcal{M}}\sum_{q\in Q_M}\frac{|\Phi_{M,q}|}{n(n-1)^2}\left(\widehat{M_q}-\widehat{M}\right)\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n(n-1)}\right)^{-1}\left(\frac{\sum\limits_{(j,k),(m,n)\in\Phi_{M,q}}\boldsymbol{X}_{jk}\boldsymbol{X}_{mn}^T}{|\Phi_{M,q}|}\right)\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n(n-1)}\right)^{-1}\\
&= \sum_{M\in\mathcal{M}}\sum_{q\in Q_M}\frac{c_{M,q}\cdot|\Phi_M|}{n(n-1)^2}\left(\widehat{M_q}-\widehat{M}\right)h_{M,q}(\boldsymbol{X})\\
&= \sum_{M\in\mathcal{M}}\sum_{q\in Q_M}c'_M c_{M,q}\left(\widehat{M_q}-\widehat{M}\right)h_{M,q}(\boldsymbol{X}) \qquad (4)
\end{aligned}$$

where $c'_M = \frac{|\Phi_M|}{n(n-1)^2}$, $c_{M,q}$ is the proportion of dyad pairs with configuration M and block specification q over all dyad pairs with configuration M, and $h_{M,q}$ contains the remaining terms which are functions of $\boldsymbol{X}$. Because we assume B is $\mathcal{O}(1)$, each $|\Phi_M|$ is at most $\mathcal{O}(n^3)$, so each $c'_M \to d_M$ for some constant d_M. Marrs et al. (2017) (Eq.27) show that
$h_{M,q}(\boldsymbol{X}) \xrightarrow{p} h'_{M,q}(\boldsymbol{X}) =$
$$\begin{cases}\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T]^{-1}\,\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T|(j,k)\in\Phi_{\sigma^2,q}]\,\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T]^{-1}, & \text{for } M=\sigma^2\\ \mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T]^{-1}\,\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{mn}^T|(j,k),(m,n)\in\Phi_{M,q}]\,\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T]^{-1}, & \text{for } M\in\mathcal{M}\setminus\sigma^2\end{cases}$$
We have shown $c_{M,q}$ and $h_{M,q}$ both converge in probability to constants. So the only part left in Equation 4 is $\left(\widehat{M_q}-\widehat{M}\right)$. Previous work (Marrs et al., 2017) has shown that

$$\widehat{M_q} \xrightarrow{p} M_q \text{ and } \widehat{M} \xrightarrow{p} M, \qquad (5)$$

where

$$M = \frac{\sum\limits_{q\in Q_M} M_q\cdot|\Phi_{M,q}|}{\sum\limits_{q\in Q_M}|\Phi_{M,q}|} = \sum_{q\in Q_M} M_q\cdot c_{M,q} \qquad (6)$$

Thus, by Slutsky's theorem,

$$n\left(\hat{V}_B(\hat{\boldsymbol{\beta}}) - \hat{V}_E(\hat{\boldsymbol{\beta}})\right) \xrightarrow{p} \sum_{M\in\mathcal{M}}\sum_{q\in Q_M}\left(M_q - M\right)f_{M,q}(\boldsymbol{X}), \qquad (7)$$

where $f_{M,q} = c_{M,q}\cdot d_M\cdot h'_{M,q}(\boldsymbol{X})$ is a constant when distribution of $\boldsymbol{X}$ is known, and M_q is the true parameter in Ω_B. When the distribution of $\boldsymbol{X}$ is independent of block membership, we have $f_{M,q}(\boldsymbol{X}) = f_M(\boldsymbol{X}) \ \forall q$. In addition, $\sum\limits_{q \in Q_M} c_{M,q} = 1 \ \forall M$. Therefore,

$$\begin{aligned} n\left(\hat{V}_B(\hat{\boldsymbol{\beta}}) - \hat{V}_E(\hat{\boldsymbol{\beta}})\right) &\xrightarrow{p} \sum_{M \in \mathcal{M}} d_M f_M(\boldsymbol{X}) \sum_{q \in Q_M} c_{M,q}\left(M_q - M\right) \\ &= \sum_{M \in \mathcal{M}} d_M f_M(\boldsymbol{X}) \left(\sum_{q \in Q_M} c_{M,q} M_q - \sum_{q \in Q_M} c_{M,q} M\right) \\ &= \sum_{M \in \mathcal{M}} d_M f_M(\boldsymbol{X}) \left(M - M\right) = 0 \end{aligned} \tag{8}$$

Therefore, we have shown that when $\boldsymbol{X}$ is independent of g, $n\left(\hat{V}_B(\hat{\boldsymbol{\beta}}) - \hat{V}_E(\hat{\boldsymbol{\beta}})\right) \xrightarrow{p} 0$.

In the case of two blocks,

$$\begin{aligned} n\left(\hat{V}_B(\hat{\boldsymbol{\beta}}) - \hat{V}_E(\hat{\boldsymbol{\beta}})\right) &= \sum_{u,v \in \{1,2\}} \left(\sigma^2_{(u,v)} - \sigma^2\right) f_{\sigma^2,(u,v)}(\boldsymbol{X}) \\ &+ \sum_{u,v \in \{1,2\}} \left(\phi_{A,\{u,v\}} - \phi_A\right) f_{\phi_A,(u,v)}(\boldsymbol{X}) + \sum_{u,v,w \in \{1,2\}} \left(\phi_{B,(u,\{v,w\})} - \phi_B\right) f_{\phi_B,(u,\{v,w\})}(\boldsymbol{X}) \\ &+ \sum_{u,v,w \in \{1,2\}} \left(\phi_{C,(u,\{v,w\})} - \phi_C\right) f_{\phi_C,(u,\{v,w\})}(\boldsymbol{X}) + \sum_{u,v,w \in \{1,2\}} \left(\phi_{D,(uv,w)} - \phi_D\right) f_{\phi_D,(u,v,w)}(\boldsymbol{X}), \end{aligned}$$

where

- $\sigma^2 = \dfrac{n_1(n_1-1)\sigma^2_{(1,1)} + n_2(n_2-1)\sigma^2_{(2,2)} + n_1 n_2(\sigma^2_{(1,2)} + \sigma^2_{(2,1)})}{n(n-1)}$
- $\phi_A = \dfrac{n_1(n_1-1)\phi_{A,\{1,1\}} + n_2(n_2-1)\phi_{A,\{2,2\}} + 2n_1 n_2 \phi_{A,\{1,2\}}}{n(n-1)}$
- $$\begin{aligned} \phi_B &= \frac{n_1(n_1-1)(n_1-2)\phi_{B(1,\{1,1\})} + 2n_1(n_1-1)n_2\phi_{B(1,\{1,2\})} + n_1 n_2(n_2-1)\phi_{B(1,\{2,2\})}}{n(n-1)(n-2)} \\ &+ \frac{+n_2(n_2-1)(n_2-2)\phi_{B(2,\{2,2\})} + 2n_2 n_1(n_2-1)\phi_{B(2,\{1,2\})} + n_2 n_1(n_1-1)\phi_{B(2,\{1,1\})}}{n(n-1)(n-2)} \end{aligned}$$
- $$\begin{aligned} \phi_C &= \frac{n_1(n_1-1)(n_1-2)\phi_{C(1,\{1,1\})} + 2n_1(n_1-1)n_2\phi_{C(1,\{1,2\})} + n_1 n_2(n_2-1)\phi_{C(1,\{2,2\})}}{n(n-1)(n-2)} \\ &+ \frac{+n_2(n_2-1)(n_2-2)\phi_{C(2,\{2,2\})} + 2n_2 n_1(n_2-1)\phi_{C(2,\{1,2\})} + n_2 n_1(n_1-1)\phi_{C(2,\{1,1\})}}{n(n-1)(n-2)} \end{aligned}$$
- $$\begin{aligned} \phi_D &= \frac{n_1(n_1-1)(n_1-2)\phi_{D(1,1,1)} + n_1(n_1-1)n_2\phi_{D(1,1,2)} + n_1(n_1-1)n_2\phi_{D(1,2,1)}}{n(n-1)(n-2)} \\ &+ \frac{n_1 n_2(n_2-1)\phi_{D(1,2,2)} + n_2(n_2-1)(n_2-2)\phi_{D(2,2,2)} + n_2 n_1(n_2-1)\phi_{D(2,1,2)}}{n(n-1)(n-2)} \\ &+ \frac{n_2 n_1(n_2-1)\phi_{D(2,2,1)} + n_2 n_1(n_1-1)\phi_{D(2,1,1)}}{n(n-1)(n-2)}. \end{aligned}$$

B Additional simulation details

In this section, we provide additional details about the simulation presented in Section 6 of the manuscript. To begin, take the generative model as:

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + \xi_{ij},\ \xi_{ij} = a_i + b_j + z_i^T z_j + \gamma_{(ij)} + \epsilon_{ij},\ (a_i, b_i)|g_i \sim N_2(0, \Sigma_{ab,g_i});$$
$$\Sigma_{ab,g_i} = \begin{pmatrix} \sigma^2_{a,g_i} & \rho_{ab}\sigma_{a,g_i}\sigma_{b,g_i} \\ \rho_{ab}\sigma_{a,g_i}\sigma_{b,g_i} & \sigma^2_{b,g_i} \end{pmatrix};$$
$$z_i|g_i \sim N_d\left(0, \sigma^2_{z,g_i} I_d\right);\ \epsilon_{ij} \sim N(0, \sigma^2_\epsilon);$$
$$\gamma_{(ij)} = \gamma_{(ji)}|g_i, g_j \sim (0, \sigma^2_{\gamma,\{g_i,g_j\}}).$$

Under the generative model, the variance and covariances are:

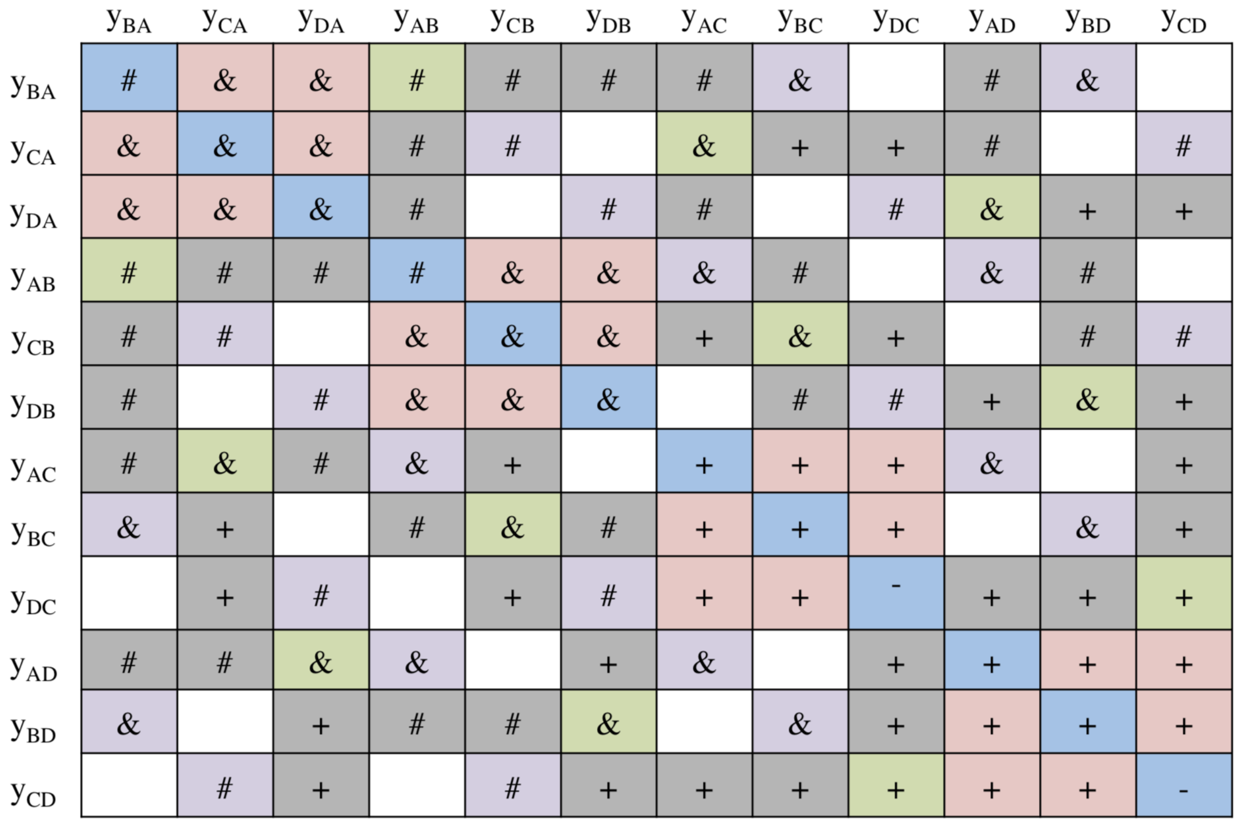
- $\text{Var}(\xi_{ij}) = \sigma^2_{a,g_i} + \sigma^2_{b,g_j} + d\sigma^2_{z,g_i}\sigma^2_{z,g_j} + \sigma^2_{\gamma,\{g_i,g_j\}} + \sigma^2_{\epsilon;}$
- $\text{Cov}(\xi_{ij}, \xi_{ji}) = \rho_{ab}\sigma_{a,g_i}\sigma_{b,g_i} + \rho_{ab}\sigma_{a,g_j}\sigma_{b,g_j} + d\sigma^2_{z,g_i}\sigma^2_{z,g_j} + \sigma^2_{\gamma,\{g_i,g_j\};}$
- $\text{Cov}(\xi_{ij}, \xi_{il}) = \sigma^2_{a,g_i;}$
- $\text{Cov}(\xi_{ij}, \xi_{kj}) = \sigma^2_{b,g_j;}$
- $\text{Cov}(\xi_{ij}, \xi_{ki}) = \rho_{ab}\sigma_{a,g_i}\sigma_{b,g_i.}$

We recognize that the error vector satisfies the block-exchangeability by making the observation that $\text{Cov}(\xi_{ij}, \xi_{kl}) = \text{Cov}(\xi_{\pi(i)\pi(j)}, \xi_{\pi(k)\pi(l)})$ with $g_i = g_{\pi(i)}, g_j = g_{\pi(j)}, g_k = g_{\pi(k)},$ and $g_l = g_{\pi(l)}$. However, this does not correspond to the most general form of the covariance matrix Ω_B that satisfy block-exchangeability. For example, under the error generating model, $\text{Cov}(\xi_{ij}, \xi_{il})$ takes B parameters, compared to $B^2(B+1)/2$ in the most general form in Table 1 in the main document.

Figure B.1 shows a visualization of the covariance matrix Ω_B under the error generating model. Entries shaded with the same color and symbol share the same covariance value. Compared to Figure 1 in the main text, the error generative model does not correspond to the most general formulation of block-exchangeability covariance structure. For example, $cov(\xi_{ij}, \xi_{ik})$ can take B values under the error generating model, but on the order of B^3 with the most general formulation.

We generate three types of covariates, each having three sub-cases regarding the correlation between the covariate and block membership:

1. $X_{ij,1} = \mathbb{1}_{X_i = X_j}$, where $X_i \sim \text{Bernoulli}(p_{g_i})$ and
   (a) p_{g_i} is uncorrelated with g_i, i.e., p_{g_i} is a fixed number
   (b) $p_{g_i}|g_i = 2 > p_{g_j}|g_j = 1 > 0.5$, which suggests that high $\text{Var}(X_{ij,1})$ is associated with high $\text{Var}(\xi_{ij})$
   (c) $p_{g_i}|g_i = 1 > p_{g_j}|g_j = 2 > 0.5$, which suggests that high $\text{Var}(X_{ij,1})$ is associated with low $\text{Var}(\xi_{ij})$
2. $X_{ij,2} = |X_i - X_j|$, where $X_i \sim \text{N}(0, \sigma_{g_i})$ and



	y_{BA}	y_{CA}	y_{DA}	y_{AB}	y_{CB}	y_{DB}	y_{AC}	y_{BC}	y_{DC}	y_{AD}	y_{BD}	y_{CD}
y_{BA}	#	&	&	#	#	#	#	&		#	&	
y_{CA}	&	&	&	#	#		&	+	+	#		#
y_{DA}	&	&	&	#		#	#		#	&	+	+
y_{AB}	#	#	#	#	&	&	&	#		&	#	
y_{CB}	#	#		&	&	&	+	&	+		#	#
y_{DB}	#		#	&	&	&		#	#	+	&	+
y_{AC}	#	&	#	&	+		+	+	+	&		+
y_{BC}	&	+		#	&	#	+	+	+		&	+
y_{DC}		+	#		+	#	+	+	-	+	+	+
y_{AD}	#	#	&	&		+	&		+	+	+	+
y_{BD}	&		+	#	#	&		&	+	+	+	+
y_{CD}		#	+		#	+	+	+	+	+	+	-

Figure B.1: Visualization of covariance matrix Ω under the error generating model used in simulation. Entries shaded with the same color and symbol share the same parameter value, and a white box indicates a covariance of zero.

(a) σ_{g_i} is uncorrelated with g_i, i.e., σ_{g_i} is a fixed number

(b) $\sigma_{g_i}|g_i = 1 > \sigma_{g_i}|g_i = 2$, which suggests that high $\text{Var}(X_{ij,2})$ is associated with high $\text{Var}(\xi_{ij})$.

(c) $\sigma_{g_i}|g_i = 1 < \sigma_{g_i}|g_i = 2$, which suggests that high $\text{Var}(X_{ij,2})$ is associated with low $\text{Var}(\xi_{ij})$.

3. $X_{ij,3} \sim N(0, \sigma^2_{g_i,g_j})$ and

(a) σ_{g_i,g_j} is uncorrelated with g_i, g_j, i.e., σ_{g_i,g_j} is a fixed number

(b) $\sigma_{g_i,g_j}|g_i = 1, g_j = 1 > \sigma_{g_i,g_j}|g_i = 2, g_j = 2$, which suggests that high $\text{Var}(X_{ij,3})$ is associated with high $\text{Var}(\xi_{ij,3})$

(c) $\sigma_{g_i,g_j}|g_i = 1, g_j = 1 < \sigma_{g_i,g_j}|g_i = 2, g_j = 2$, which suggests that high $\text{Var}(X_{ij,3})$ is associated with low $\text{Var}(\xi_{ij,3})$.

We set the parameters for generating covariates such that the noise to signal ratio, which is defined as the ratio of sum of squared errors over total sum of squares, is consistent across all three scenarios. Let NTS denote the noise-to-signal ratio, then

$$NTS_{ij} = \frac{\text{Var}(\xi_{ij})}{\text{Var}(Y_{ij})}, \text{ where } \text{Var}(\xi_{ij}) = \sigma^2_{(g_i,g_j)} \text{ and}$$

$$\text{Var}(Y_{ij}) = E(\text{Var}(Y_{ij}|X_{ij})) + \text{Var}(E(Y_{ij}|X_{ij})) = \sigma^2_{(g_i,g_j)} + \beta_1^2 \text{Var}(X_{ij}).$$

Therefore, for all three types of covariates:

1. $X_{ij,1} = \mathbb{1}_{X_i = X_j}$, where $X_i \sim \text{Bernoulli}(p_{g_i})$.

   $NTS_{ij} \mid g_i, g_j = \frac{\sigma^2_{g_i,g_j}}{\sigma^2_{g_i,g_j} + \beta_1^2 p_{ij}(1-p_{ij})}$, where $p_{ij} = p_i p_j + (1-p_i)(1-p_j)$

2. $X_{ij,2} = |X_i - X_j|$, where $X_i \sim \text{N}(0, a^2_{g_i})$.

   $NTS_{ij} \mid g_i, g_j = \frac{\sigma^2_{g_i,g_j}}{\sigma^2_{g_i,g_j} + \beta_1^2 (a^2_{g_i} + a^2_{g_j})(1-2/\pi)}$

3. $X_{ij,3} \sim N(0, a^2_{g_i,g_j})$.

   $NTS_{ij} \mid g_i, g_j = \frac{\sigma^2_{g_i,g_j}}{\sigma^2_{g_i,g_j} + \beta_1^2 a^2_{g_i,g_j}}$

With two blocks and equal block size, we set the equations $(\sum_{(u,v)\in\{(1,1),(1,2),(2,1),(2,2)\}} NTS_{ij} \mid g_i = u, g_j = v)/4 = 0.45$ and solve for the parameters.

C Additional simulations: Evaluating Block Membership Estimation

This section aims to show how well we recover block labels (Step 2-4 in Algorithm) as well as graphical proof of concept for why we construct the similarity metric between a pair of nodes as in Step 2 of the Algorithm. We consider a simple linear regression model with two blocks:

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + \xi_{ij},$$

where $X_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ and $g_i \in \{1, 2\}$. We vary the strength of block structure in errors and show how the algorithm recovers block membership.

Based on the error generating model in Section 6 of the main text, we set parameters as follows:

- $[\sigma_{a,1} \;\; \sigma_{a,2}] = [\sqrt{2}\alpha_1 \;\; \sqrt{2}r\alpha_1]$
- $[\sigma_{b,1} \;\; \sigma_{b,2}] = [\alpha_1 \;\; r\alpha_1]$
- $[\sigma_{z,1} \;\; \sigma_{z,2}] = [\alpha_1 \;\; r\alpha_1]$
- $[\sigma_{\gamma,\{1,1\}} \;\; \sigma_{\gamma,\{1,2\}} \;\; \sigma_{\gamma,\{2,2\}}] = [\alpha_1 \;\; \sqrt{r}\alpha_1 \;\; r\alpha_1]$
- $\sigma_\epsilon = \alpha_1$, $\rho = 0.5$, and $d = 2$.

We immediately see that r quantifies the strength of block structure in errors. A trivial $r = 1$ suggests that there is no block structure, while an r value far away from one suggests a strong block structure. As functions of r and α_1, the variance and covariances are:

$$\text{Var}(\xi_{ij}) = \begin{cases} 5\alpha_1^2 + 2\alpha_1^4 & \text{if } g_i = 1, g_j = 1 \\ (r^2 + r + 3)\alpha_1^2 + 2r^2\alpha_1^4 & \text{if } g_i = 1, g_j = 2 \\ (2r^2 + r + 2)\alpha_1^2 + 2r^2\alpha_1^4 & \text{if } g_i = 2, g_j = 1 \\ (4r^2 + 1)\alpha_1^2 + 2r^4\alpha_1^4 & \text{if } g_i = 2, g_j = 2 \end{cases}$$

$$\text{Cov}(\xi_{ij}, \xi_{ji}) = \begin{cases} (\sqrt{2} + 1)\alpha_1^2 + 2\alpha_1^4 & \text{if } g_i = 1, g_j = 1 \\ (1/\sqrt{2} + r + 1/\sqrt{2}r^2)\alpha_1^2 + 2r^2\alpha_1^4 & \text{if } g_i = 1, g_j = 2 \\ (1/\sqrt{2} + r + 1/\sqrt{2}r^2)\alpha_1^2 + 2r^2\alpha_1^4 & \text{if } g_i = 2, g_j = 1 \\ (\sqrt{2} + 1)r^2\alpha_1^2 + 2r^4\alpha_1^4 & \text{if } g_i = 2, g_j = 2 \end{cases}$$

$$\text{Cov}(\xi_{ij}, \xi_{il}) = \begin{cases} 2\alpha_1^2 & \text{if } g_i = 1 \\ 2r^2\alpha_1^2 & \text{if } g_i = 2 \end{cases}$$

$$\text{Cov}(\xi_{ij}, \xi_{kj}) = \begin{cases} \alpha_1^2 & \text{if } g_j = 1 \\ r^2\alpha_1^2 & \text{if } g_j = 2 \end{cases}$$

$$\text{Cov}(\xi_{ij}, \xi_{ki}) = \begin{cases} 1/\sqrt{2}\alpha_1^2 & \text{if } g_i = 1 \\ 1/\sqrt{2}r^2\alpha_1^2 & \text{if } g_i = 2 \end{cases}$$

We perform simulation study on three values of r: $r = 1/4, r = 1/2$, and $r = 3/4$. Again we see that $r = 1/4$ has the strongest block structure in errors, as the differences in variance and covariances between different blocks are largest. For example, $\text{Cov}(\xi_{ij}, \xi_{il}|g_i = 1) - \text{Cov}(\xi_{ij}, \xi_{il}|g_i = 2) = 2(1 - r^2)\alpha_1^2$, and $(1 - r^2)$ is a decreasing function in $r \in (0, 1]$. Because all three values of r are between 0 and 1, We also observe that:

- $\text{Var}(\xi_{ij})|g_i = 1, g_j = 1 > \text{Var}(\xi_{ij})|g_i = 1, g_j = 2 > var(\xi_{ij})|g_i = 2, g_j = 1 > \text{Var}(\xi_{ij})|g_i = 2, g_j = 2$

- $\text{Cov}(\xi_{ij}, \xi_{ji})|g_i = 1, g_j = 1 > \text{Cov}(\xi_{ij}, \xi_{ji})|g_i = 1, g_j = 2 = \text{Cov}(\xi_{ij}, \xi_{ji})|g_i = 2, g_j = 1 > \text{Cov}(\xi_{ij}, \xi_{ji})|g_i = 2, g_j = 2$

- $\text{Cov}(\xi_{ij}, \xi_{il})|g_i = 1 > \text{Cov}(\xi_{ij}, \xi_{il})|g_i = 2$

- $\text{Cov}(\xi_{ij}, \xi_{kj})|g_j = 1 > \text{Cov}(\xi_{ij}, \xi_{kj})|g_j = 2$

- $\text{Cov}(\xi_{ij}, \xi_{ki})|g_i = 1 > \text{Cov}(\xi_{ij}, \xi_{ki})|g_i = 2.$

C.1 Simulation Results

In this section, we provide simulation evidence for Step 2 and 3 in Algorithm 2, as well as how well we recover the block membership. Step 2 calculates the set of residual products for a specific actor and dyad configuration, and step 3 calculates the Kolmogorov-Smirnov statistic of the residual products between a pair of actors. Using simulated data, we show that the distributions of residual products for actors i and i' ($g_i \neq g_{i'}$) are more similar as block strength decreases, which is evidence why using the KS statistic between them is a reasonable way to construct a similarity matrix.

Figure C.1 shows the distribution of residual products calculated in Algorithm 2 Step 2 on each of the five cases at different values of r. Each column represents one of the five cases $M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, and each row represents a given r value. The red and blue curves represent the distribution in Block 1 and Block 2, respectively. The densities are constructed on all actors from 10 simulations of a network of size 80. The KS statistic on each plot is calculated between the distribution of residual products. At $r = 1/4$, all five plots show that the red curve is more spread out. This is because we set the simulation parameters such that variance and covariances involving actors in Block 1 is always larger than those involving Block 2. Since residual products are estimators of variance and covariances, we observe that $\forall M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, the distribution of $\boldsymbol{R}_{M,i}|g_i = 1$ is more spread out. As r decreases, the strength of block in errors decreases, so we observe a smaller difference between the two densities on all five cases. At $r = 3/4$, the two densities coincide on $M \in \{\phi_C, \phi_D\}$. This shows that as we have stronger block structure in errors, we have a larger difference between the distribution of residual products.

Figure C.2 shows the distribution of KS statistic $KS_{i,j,M}$ calculated in Algorithm 2 Step 3 on each of the five cases at different values of r. Each column represents one of the five cases $M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, and each row represents a given r value. The red curves represent the distribution where the two actors share the same block membership ($g_i = g_j$), while the blue curves represent the distribution where the two actors are in different blocks ($g_i \neq g_j$). The densities are constructed on all actors from 10 simulations of a network of size 80. The KS statistic on each plot is calculated between the distribution of KS statistics. At $r = 1/4$, we observe that the blue curve is more spread out. This is expected because the difference in distributions of residual products involving actors i and that involving actor j is larger when $g_i \neq g_j$, which leads to larger KS statistic between the two distributions. We also observe that when $M = \sigma^2$, the KS statistic between two distributions of KS statistic is largest, which is evidence that the distribution of $\boldsymbol{R}_{\sigma^2,i}$ is most effective in identifying whether two actors belong to the same block. At $r = 3/4$, we observe that the two curves are similar. Since the block structure is not strong in errors, the distribution of $\boldsymbol{R}_{M,i}$ and $\boldsymbol{R}_{M,j}$ are not too different even when $g_i \neq g_j$.

Figure C.3 shows the number of misclustered nodes at different values of r. The number of misclustered nodes is defined as $\min(\Pi_{g_i} \sum_{i=1}^{n} |g_i - \hat{g}_i|)$, which is the minimum number of nodes in the wrong block under permutation of the block labels. In the network of size n, the number of misclustered nodes ranges from 0 to $n/2$. The boxplots in Figure C.3 shows the distribution of the proportion of misclustered nodes, which is defined as the number of misclustered nodes over n, where the red, blue, yellow color represent network size $n = 20, 40, 80, 160$ respectively. The line in the box is the median proportion, the boundaries of the box is 10 and 90 percentile, and the whiskers are 2.5 and 97.5 percentile. We observe that the proportion decreases with increasing n and increases with increasing r, which shows that we recover block membership well at large network size and strong block structure in errors.

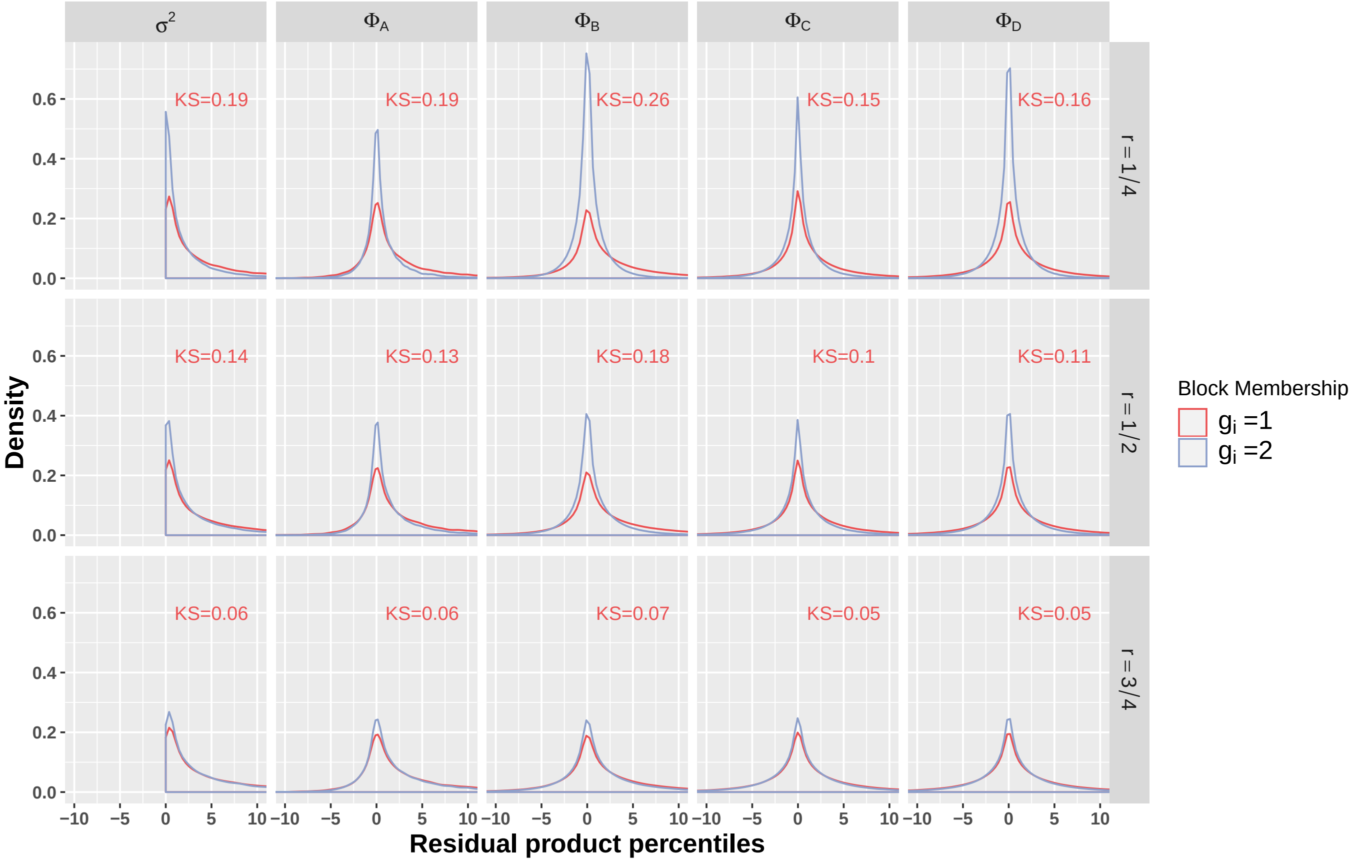


Figure C.1: Residual products of five dyads from 10 simulation of $n = 80$. Each column represents one of the five cases $M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, and each row represents a given r value. The red and blue curves represent the distribution in Block 1 and Block 2, respectively. The KS statistic on each plot is calculated between the distribution of residual products.

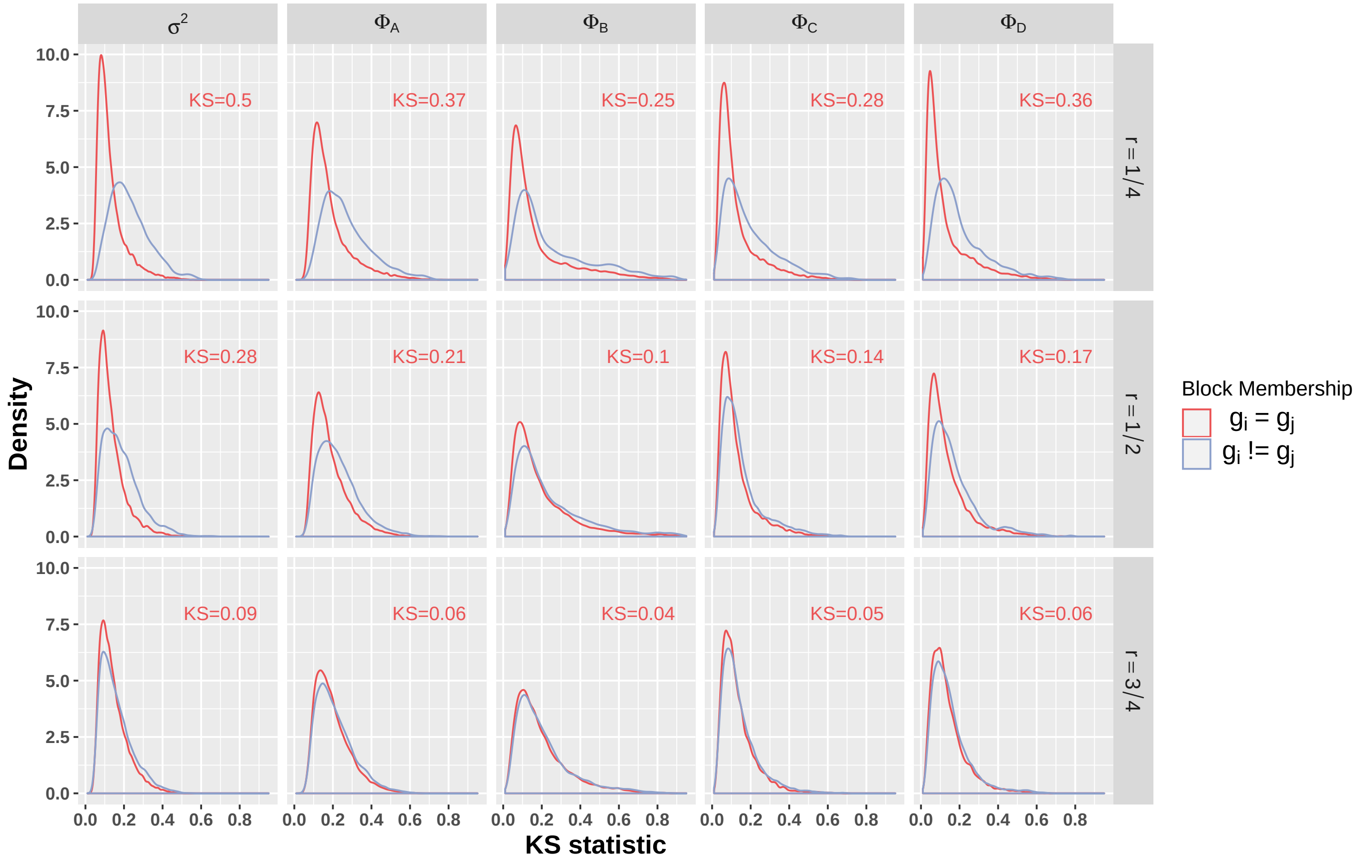


Figure C.2: Distribution of KS statistic between residual products of five dyads from 10 simulation of $n = 80$. Each column represents one of the five cases $M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, and each row represents a given r value. The red curves represent the distribution where the two actors share the same block membership $(g_i = g_j)$, while the blue curves represent the distribution where the two actors are in different blocks $(g_i \neq g_j)$. The KS statistic on each plot is calculated between the distribution of KS statistics.

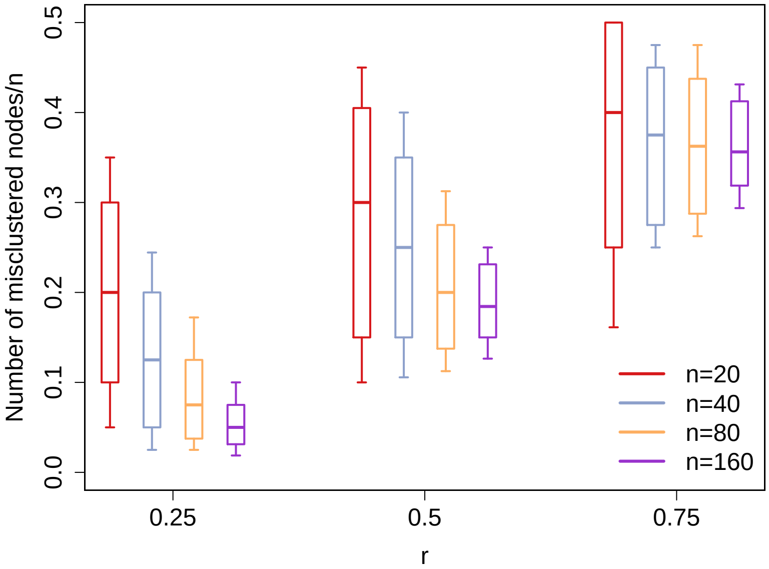


Figure C.3: Number of misclustered nodes over n at different r.

D Additional details on the air traffic data

In this section, we provide more details about fitting the proposed model in the context of our illustrative data example. A challenge posed by these data is the large number of zeros that arise when there are no passenger seats from one airport to another. Figure D.1 shows the distribution of passenger seats and the log number of seats between a destination for cases where the number is greater than zero. We develop a pseudolikelihood approach to address the structure of these data. Besag (1975) introduces pseudo-likelihood methodology using an objective function that maximizes a product of conditional densities instead of the joint likelihood, and Arnold & Strauss (1991) shows that when using pseudo-likelihood as the objective, the parameter estimates are asymptotically normal with mean as the true parameter and covariance matrix as the sandwich estimator. In the field of network analysis, pseudo-likelihood approach has been widely used to make inference for exponential family random graph models (ERGM) (Strauss & Ikeda (1990)), due to the fact that computation of conditional densities are easier. Assume we have n independent, identically distributed observed vectors $Y^{(i)}$, researchers have also used pseudo-likelihood approach as maximizing a sum of pairwise marginal log likelihoods:

$$l(\boldsymbol{\theta}; Y^{(1)}, ..., Y^{(n)}) = \sum_i l(\boldsymbol{\theta}; Y^{(i)}) \text{ ,where } \quad l(\boldsymbol{\theta}; Y^{(i)}) = \sum_{s>t} \log L(Y_s^{(i)}, Y_t^{(i)}; \boldsymbol{\theta}), \qquad (9)$$

where $L(Y_s^{(i)}, Y_t^{(i)}; \boldsymbol{\theta})$ is the likelihood of observing a pair of values $Y_s^{(i)}$ and $Y_t^{(i)}$ given parameter $\boldsymbol{\theta}$. Cox & Reid (2004) presents conditions for obtaining consistent estimates when using such approach. Fieuws & Verbeke (2006) apply this method to the case of longitudinal observations, where individual random effects lead to non-zero covariance between multiple observations on the same individual. Solomon & Weissfeld (2017) extend this application to a case where observations are left-censored. Other applications of pairwise likelihood approach include Kuk & Nott (2000) and Renard et al. (2004).

We first present the likelihood for a pair of observations (y_{ij}, y_{kl}) when one or both observations mat be censored. We consider a setting of relational observations left-censored at zero for the regression model below:

$$y_{ij} = \begin{cases} y^*_{ij}, & \text{if } y^*_{ij} \geq 0 \\ 0, & \text{if } y^*_{ij} < 0 \end{cases}$$

where $y^*_{ij} = \boldsymbol{\beta}^T \boldsymbol{X}_{ij} + \xi_{ij}, \quad i, j \in \{1, ..., n\}, i \neq j$. Let $\boldsymbol{\theta}$ denote the parameter vector containing $\boldsymbol{\beta}$ and covariance terms. Let $\rho_{(\epsilon_{ij}, \epsilon_{kl})} = \text{Cov}(\epsilon_{ij}, \epsilon_{kl}) / \sqrt{\text{Var}(\epsilon_{ij})\text{Var}(\epsilon_{kl})}$ denote the correlation coefficient between ϵ_{ij} and ϵ_{kl}. The likelihood of a pair of relational observations $L(y_{ij}, y_{kl}; \boldsymbol{\theta})$ takes one of the four following values.

- If $y_{ij} > 0$ and $y_{kl} > 0$, then $L(y_{ij}, y_{kl}; \boldsymbol{\theta}) = f_{Y_{ij}, Y_{kl}}(y_{ij}, y_{kl})$,

  where $\begin{pmatrix} Y_{ij} \\ Y_{kl} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\beta}^T \boldsymbol{X}_{ij} \\ \boldsymbol{\beta}^T \boldsymbol{X}_{kl} \end{bmatrix}, \begin{bmatrix} \text{Var}(\epsilon_{ij}) & \text{Cov}(\epsilon_{ij}, \epsilon_{kl}) \\ \text{Cov}(\epsilon_{ij}, \epsilon_{kl}) & \text{Var}(\epsilon_{kl}) \end{bmatrix} \right)$.

- If $y_{ij} > 0$ and $y_{kl} = 0$, then $L(y_{ij}, y_{kl}; \boldsymbol{\theta}) = f_{Y_{ij}}(y_{ij}) \cdot F_{Y_{kl}|Y_{ij}}(0)$,

  where $Y_{ij} \sim \mathcal{N}(\boldsymbol{\beta}^T \boldsymbol{X}_{ij}, \text{Var}(\epsilon_{ij}))$
  and $Y_{kl} \mid Y_{ij} \sim \mathcal{N}(\boldsymbol{\beta}^T \boldsymbol{X}_{kl} + \text{Cov}(\epsilon_{ij}, \epsilon_{kl}) / \text{Var}(\epsilon_{ij}) \cdot (y_{ij} - \boldsymbol{\beta}^T \boldsymbol{X}_{ij}), (1 - \rho^2_{(\epsilon_{ij}, \epsilon_{kl})}) \cdot \text{Var}(\epsilon_{kl}))$

- if $y_{ij} = 0$ and $y_{kl} > 0$, then $L(y_{ij}, y_{kl}; \boldsymbol{\theta}) = f_{Y_{kl}}(y_{kl}) \cdot F_{Y_{ij}|Y_{kl}}(0)$,

  where $Y_{kl} \sim \mathcal{N}(\boldsymbol{\beta}^T \boldsymbol{X}_{kl}, \text{Var}(\epsilon_{kl}))$
  and $Y_{ij} \mid Y_{kl} \sim \mathcal{N}(\boldsymbol{\beta}^T \boldsymbol{X}_{ij} + \text{Cov}(\epsilon_{ij}, \epsilon_{kl}) / \text{Var}(\epsilon_{kl}) \cdot (y_{kl} - \boldsymbol{\beta}^T \boldsymbol{X}_{kl}), (1 - \rho^2_{(\epsilon_{ij}, \epsilon_{kl})}) \cdot \text{Var}(\epsilon_{ij}))$

- if $y_{ij} = 0$ and $y_{kl} = 0$, then $L(y_{ij}, y_{kl}; \boldsymbol{\theta}) = F_{Y_{ij}, Y_{kl}}(0, 0)$

  where $\begin{pmatrix} Y_{ij} \\ Y_{kl} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\beta}^T \boldsymbol{X}_{ij} \\ \boldsymbol{\beta}^T \boldsymbol{X}_{kl} \end{bmatrix}, \begin{bmatrix} \text{Var}(\epsilon_{ij}) & \text{Cov}(\epsilon_{ij}, \epsilon_{kl}) \\ \text{Cov}(\epsilon_{ij}, \epsilon_{kl}) & \text{Var}(\epsilon_{kl}) \end{bmatrix} \right)$.

The likelihood we present above applies to one pair of observations. To calculate the pseudo log likelihood of all pairs of observations, we have

$$l(\boldsymbol{\theta}; Y) = \sum_{i,j,k,l \in [n], i \neq j, k \neq l} \log L(y_{ij}, y_{kl}; \boldsymbol{\theta}). \tag{10}$$

In Equation 10, $\boldsymbol{\theta}$ contains $\boldsymbol{\beta}$ and all variance and covariance terms. With B blocks and $p - 1$ covariates, the total number of parameters is on the order of p or B^3, depending on which one is larger. Estimating all parameters at the same time is too difficult for state of art optimization algorithms. The covariance between y_{ij} and y_{kl} depends on the dyad configuration of $[(i, j), (k, l)]$ and their block memberships. Therefore, in order to decrease the number of parameters in each numerical optimization, we decompose the likelihood into a sum of sub-likelihoods involving pairs of observations that share the same covariance. The parameter vector in each sub-likelihood contains $\boldsymbol{\beta}$, $\text{Var}(\epsilon_{ij})$, $\text{Var}(\epsilon_{ij})$, and $\text{Cov}(\epsilon_{ij}, \epsilon_{kl})$. The number of parameters in each sub-likelihood is on the order p, which greatly reduces the difficulties for numerical optimization.

The likelihood formula is

$$\begin{aligned} l(\boldsymbol{\theta}; Y) &= \sum_{M,q \in Q_M} \sum_{[(i,j),(k,l)] \in \Phi_{M,q}} \log L(y_{ij}, y_{kl}; \boldsymbol{\theta}_{M,q}) \\ &= \sum_{M,q \in Q_M} l(\boldsymbol{\theta}_{M,q}, Y) \end{aligned}$$

where $l(\boldsymbol{\theta}_{M,q}, Y) = \sum_{[(i,j),(k,l)] \in \Phi_{M,q}} \log L(y_{ij}, y_{kl}; \boldsymbol{\theta}_{M,q})$. Instead of finding the set of parameters that maximize $l(\boldsymbol{\theta}; Y)$, we now find the set of parameters that maximize $l(\boldsymbol{\theta}_{M,q}, Y)$.

Let s denote the index of optimization. For example, $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{\phi_A,\{1,1\}}$, $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{\phi_A,\{1,2\}}$,$\boldsymbol{\theta}_3 = \boldsymbol{\theta}_{\phi_A,\{2,2\}}$, $\boldsymbol{\theta}_4 = \boldsymbol{\theta}_{\phi_B,(1,\{1,1\})\ldots}$. Let $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4, ...]$. The asymptotic distribution of $\hat{\boldsymbol{\Theta}}$ is

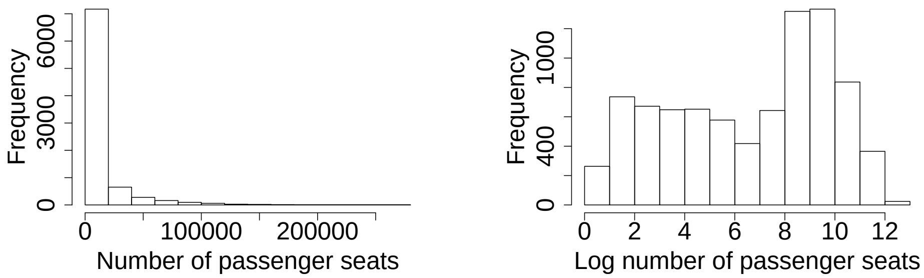
$$\sqrt{n}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \xrightarrow{d} MVN(0, A(\boldsymbol{\Theta})^{-1} B(\boldsymbol{\Theta}) A(\boldsymbol{\Theta})^{-1}), \qquad (11)$$

where $A(\boldsymbol{\Theta}) = E\left[-\frac{\partial^2 \sum_s l(\boldsymbol{\theta}_s, Y)}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}'}\right]$ and $B(\boldsymbol{\Theta}) = E\left[\frac{\partial \sum_s l(\boldsymbol{\theta}_s, Y)}{\partial \boldsymbol{\Theta}} \left(\frac{\partial \sum_s l(\boldsymbol{\theta}_s, Y)}{\partial \boldsymbol{\Theta}}\right)'\right]$.

Because $l_s(\boldsymbol{\theta}_s, Y)$ only involves $\boldsymbol{\theta}_s$, $A(\boldsymbol{\Theta})$ is a block-diagonal matrix with blocks

$$A(\boldsymbol{\Theta})_{ss} = E\left[-\frac{\partial^2 l_s(\boldsymbol{\theta}_s, Y)}{\partial \boldsymbol{\theta}_s \partial \boldsymbol{\theta}_s'}\right],$$

and $B(\boldsymbol{\Theta})$ is a symmetric matrix where $B(\boldsymbol{\Theta})_{st} = E\left[\frac{\partial l_s(\boldsymbol{\theta}_s, Y)}{\partial \boldsymbol{\theta}_s} \left(\frac{\partial l_t(\boldsymbol{\theta}_t, Y)}{\partial \boldsymbol{\theta}_t}\right)'\right]$.



(a) Raw passenger seats distribution (b) Log passenger seats distribution

Figure D.1: Distribution of number of passenger seats from departure airport to arrival airport. The right figure shows the log of the number of seats when the number exceeds zero.

In Fieuws & Verbeke (2006) and Solomon & Weissfeld (2017), independent observations are drawn from a multivariate distribution and longitudinal observations on the same individual are correlated. Since we deal with network data, we can not simply calculate the empirical version of $B(\boldsymbol{\Theta})$ by taking averages with $\hat{\boldsymbol{\Theta}}$ substituted. Therefore, we make the modification that observations used in maximizing $l_s(\boldsymbol{\theta}_s, Y)$ and $l_t(\boldsymbol{\theta}_t, Y)$ are distinct. Then $\hat{B}(\boldsymbol{\Theta})_{st} = \mathbf{0}$ and we can get $\hat{A}(\boldsymbol{\Theta})_{ss}$ and $\hat{B}(\boldsymbol{\Theta})_{ss}$ by taking averages with $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_t$ substituted.

The last step in getting $\hat{\boldsymbol{\beta}}$ and $SE(\hat{\boldsymbol{\beta}})$ is to take weighted averages of $\boldsymbol{\theta}_s \ \forall s$. Let $\hat{\boldsymbol{\theta}} = \boldsymbol{A}\hat{\boldsymbol{\Theta}}$, where $\boldsymbol{A}$ is the matrix that calculates the weighted averages, with weights proportional to the sample size used in each optimization. Then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} MVN(0, \boldsymbol{A}\Sigma(\hat{\boldsymbol{\Theta}})\boldsymbol{A}'), \tag{12}$$

where $\Sigma(\hat{\boldsymbol{\Theta}})$ is the covariance matrix for $\hat{\boldsymbol{\Theta}}$ obtained by using Equation (11).

E Definitions and notation

In this section, we formally define the notation defined conceptually in the paper. Q_M, which is the the set of block pairs/triplets for dyad configuration M given $[B]$ (Section 4.1) is defined as:

- $Q_{\sigma^2} = \{(u, v) : u, v \in [B]\}$
- $Q_{\phi_A} = \{\{u, v\} : u, v \in [B]\}$
- $Q_{\phi_B} = \{(u, \{v, w\}) : u, v, w \in [B]\}$
- $Q_{\phi_C} = \{(u, \{v, w\}) : u, v, w \in [B]\}$
- $Q_{\phi_D} = \{(u, v, w) : u, v, w \in [B]\}$

$\Phi_{M,q}$ is defined as:

- $\Phi_{\sigma^2,(u,v)} = \{[(i, j), (i, j)] : i, j \in [n], i \neq j, g_i = u, g_j = v\}$
- $\Phi_{\phi_A,\{u,v\}} = \{[(i, j), (j, i)] : i, j \in [n], i \neq j, g_i = u, g_j = v\}$
- $\Phi_{\phi_B,(u,\{v,w\})} = \{[(i, j), (i, k)] : i, j, k \in [n], i \neq j \neq k, g_i = u, g_j = v, g_k = w\}$
- $\Phi_{\phi_C,(u,\{v,w\})} = \{[(j, i), (k, i)] : i, j, k \in [n], i \neq j \neq k, g_i = u, g_j = v, g_k = w\}$
- $\Phi_{\phi_D,(u,v,w)} = \{[(i, j), (k, i)] : i, j, k \in [n], i \neq j \neq k, g_i = u, g_j = v, g_k = w\}$

$\Phi_{M,i}$ is defined as:

- $\Phi_{\sigma^2,i} = \{[(i, j), (i, j)] : j \in [n], i \neq j\} \cup \{[(j, i), (j, i)] : j \in [n], i \neq j\}$
- $\Phi_{\phi_A,i} = \{[(i, j), (j, i)] : j \in [n], i \neq j\}$
- $\Phi_{\phi_B,i} = \{[(i, j), (i, k)] : j \in [n], k \in [n], i \neq j \neq k\}$
- $\Phi_{\phi_C,i} = \{[(j, i), (k, i)] : j \in [n], k \in [n], i \neq j \neq k\}$
- $\Phi_{\phi_D,i} = \{[(i, j), (k, i)] : j \in [n], k \in [n], i \neq j \neq k\}$

References

Arnold, B. C. and Strauss, D. Pseudolikelihood estimation: Some examples. *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 233–243, 1991.

Besag, J. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

Cox, D. R. and Reid, N. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

Fieuws, S. and Verbeke, G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431, 2006.

Kuk, A. Y. and Nott, D. J. A pairwise likelihood approach to analyzing correlated binary data. *Statistics & Probability Letters*, 47(4):329–335, 2000.

Marrs, F. W., Fosdick, B. K., and McCormick, T. H. Standard errors for regression on relational data with exchangeable errors. *arXiv preprint arXiv:1701.05530*, 2017.

Renard, D., Molenberghs, G., and Geys, H. A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, 44(4):649–667, 2004.

Solomon, G. and Weissfeld, L. Pseudo maximum likelihood approach for the analysis of multivariate left-censored longitudinal data. *Statistics in Medicine*, 36(1):81–91, 2017.

Strauss, D. and Ikeda, M. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.