A Details of estimation

In this section we supply details of estimation in support of Algorithm 1, beginning with the initialization of ρ. We then provide details of computing the expectations of $\ell_{\mathbf{y}}$ need for β maximization, and then details of computing the expectations of $\ell_{\mathbf{y}}$ need for ρ maximization. We close the section with the handling of missing data in the EMM algorithm.

A.1 Initialization of ρ *estimator*

An EM algorithm may take many iterations to converge, and selecting a starting point near the optima may significantly reduce the number of iterations required. We present a method of initializing $\widehat{\rho}^{(0)}$ using a mixture estimator. By examining the eigenvalues of Ω, it can be shown that ρ lies in the interval $[0, 1/2)$ when Ω is positive definite for arbitrary n (Marrs *et al.*, 2017). Thus $\widehat{\rho} = 0.25$ is a natural naive initialization point as it is the midpoint of the range of possible values. However, we also allow the data to influence the initialization point by taking a random subset $\mathscr{A}$ of Θ_2 of size $2n^2$, and estimating ρ using the data corresponding to relations in $\mathscr{A}$. Then, the final initialization point is defined as a mixture between the naive estimate $\widehat{\rho} = 0.25$ and the estimate based on the data. We weight the naive value as if it arose from $100n$ samples, such that the weights are even at $n = 50$, and for increasing n, the data estimate dominates:

$$\widehat{\rho}^{(0)} = \frac{100n}{4(100n + |\mathscr{A}|)} + \frac{|\mathscr{A}|}{(100n + |\mathscr{A}|)} \left(\frac{1}{|\mathscr{A}|} \sum_{jk,lm \in \mathscr{A}} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}] \right). \quad \text{(A 1)}$$

We compute the average $\frac{1}{|\mathscr{A}|} \sum_{jk,lm \in \mathscr{A}} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}]$ using the linearization approach described in Section A.3.

A.2 Implementation of β *expectation step*

Under general correlation structure, computation of the expectation $E[\varepsilon \,|\, \mathbf{y}]$ (step 1 in Algorithm 1, where we drop conditioning on $\rho^{(\nu)}$ and $\beta^{(\nu)}$ to lighten notation) for even small networks is prohibitive, since this expectation is an $\binom{n}{2}$-dimensional truncated multivariate normal integral. We exploit the structure of Ω to compute $E[\varepsilon \,|\, \mathbf{y}]$ using the law of total expectation and a Newton-Raphson algorithm.

First, we take a single relation jk and use the law of total expectation to write

$$E[\varepsilon_{jk} \,|\, \mathbf{y}] = E[E[\varepsilon_{jk} \,|\, \varepsilon_{-jk}, y_{jk}] \,|\, \mathbf{y}], \quad \text{(A 2)}$$

where ε_{-jk} is the vector of all entries in ε except relation jk. Beginning with the innermost conditional expectation, the distribution of ε_{jk} given ε_{-jk} and y_{jk} is truncated univariate normal, where the untruncated normal random variable has the mean and variance of ε_{jk} given ε_{-jk}. Based on the conditional multivarite normal distribution and the form of the inverse covariance matrix $\Omega^{-1} = \sum_{i=1}^{3} p_i \mathscr{S}_i$, we may write the untruncated distribution directly as

$$\begin{aligned} \varepsilon_{jk} \,|\, \varepsilon_{-jk} &\sim \mathrm{N}(\mu_{jk}, \sigma_n^2), \qquad \text{(A 3)} \\ \mu_{jk} &= -\sigma_n^2 \mathbf{1}_{jk}^T (p_2 \mathscr{S}_2 + p_3 \mathscr{S}_3) \widetilde{\varepsilon}_{-jk}, \\ \sigma_n^2 &= \frac{1}{p_1}, \end{aligned}$$

where $\mathbf{1}_{jk}$ is the vector of all zeros with a one in the position corresponding to relation jk and, for notational purposes, we define $\widetilde{\varepsilon}_{-jk}$ as the vector ε except with a zero in the location corresponding to relation jk. We note that the diagonal of the matrix $p_2\mathscr{S}_2 + p_3\mathscr{S}_3$ consists of all zeros so that μ_{jk} is free of ε_{jk}.

We now condition on y_{jk}. For general $z \sim \mathrm{N}(\mu, \sigma^2)$ and $y = \mathbb{1}[z > -\eta]$ we have that

$$E[z \,|\, y] = \mu + \sigma \frac{\phi(\widetilde{\eta})}{\Phi(\widetilde{\eta})(1 - \Phi(\widetilde{\eta}))} (y - \Phi(\widetilde{\eta})), \quad \text{(A 4)}$$

where $\widetilde{\eta} := (\eta + \mu)/\sigma$. Now, taking $z = (\varepsilon_{jk} \,|\, \varepsilon_{-jk})$, we have that

$$E[\varepsilon_{jk} \,|\, \varepsilon_{-jk}, y_{jk}] = \mu_{jk} + \sigma_n \left(\frac{\phi(\widetilde{\mu}_{jk}) \left(y_{jk} - \Phi(\widetilde{\mu}_{jk}) \right)}{\Phi(\widetilde{\mu}_{jk})(1 - \Phi(\widetilde{\mu}_{jk}))} \right), \quad \text{(A 5)}$$

where $\widetilde{\mu}_{jk} := (\mu_{jk} + \mathbf{x}_{jk}^T \beta)/\sigma_n$.

We now turn to the outermost conditional expectation in (A 2). Substituting the expression for μ_{jk} into (A 5), we have that

$$E[\varepsilon_{jk} \,|\, \mathbf{y}] = -\sigma_n^2 \mathbf{1}_{jk}^T (p_2 \mathscr{S}_2 + p_3 \mathscr{S}_3) E[\varepsilon \,|\, \mathbf{y}] + \sigma_n E \left[\frac{\phi(\widetilde{\mu}_{jk}) \left(y_{jk} - \Phi(\widetilde{\mu}_{jk}) \right)}{\Phi(\widetilde{\mu}_{jk})(1 - \Phi(\widetilde{\mu}_{jk}))} \,\Big|\, \mathbf{y} \right]. \quad \text{(A 6)}$$

This last conditional expectation is difficult to compute in general. Thus, in place of $\widetilde{\mu}_{lm}$, we substitute its conditional expectation $E[\widetilde{\mu}_{lm} \,|\, \mathbf{y}]$. Letting $w_{lm} := E[\varepsilon_{lm} \,|\, \mathbf{y}]$ and $\mathbf{w}$ be the vector of the expectations $\{w_{lm}\}_{lm}$, we define the following nonlinear equation for $\mathbf{w}$:

$$0 \approx g(\mathbf{w}) := (-\mathbf{I} + \mathbf{B})\mathbf{w} + \sigma_n \left(\frac{\phi(\widetilde{\mathbf{w}}) (\mathbf{y} - \Phi(\widetilde{\mathbf{w}}))}{\Phi(\widetilde{\mathbf{w}})(1 - \Phi(\widetilde{\mathbf{w}}))} \right), \quad \text{(A 7)}$$

where we define $\mathbf{B} := -\sigma_n^2 (p_2\mathscr{S}_2 + p_3\mathscr{S}_3)$, $\widetilde{\mathbf{w}} := (\mathbf{B}\mathbf{w} + \mathbf{X}\beta)/\sigma_n$, and the functions $\phi(.)$ and $\Phi(.)$ are applied element-wise. The approximation in (A 7) refers to the approximation made when replacing $\widetilde{\mu}_{jk}$ with its conditional expectation $E[\widetilde{\mu}_{jk}|\mathbf{y}]$. We use a Newton-Raphson algorithm to update $\mathbf{w}$ (Atkinson, 2008), initializing the algorithm using the expectation when $\rho = 0$,

$$\mathbf{w}_0 := \frac{\phi(\mathbf{X}\beta) (\mathbf{y} - \Phi(\mathbf{X}\beta))}{\Phi(\mathbf{X}\beta)(1 - \Phi(\mathbf{X}\beta))}. \quad \text{(A 8)}$$

The Newton-Raphson algorithm re-estimates $\mathbf{w}$ based on the estimate at iteration ν, $\widehat{\mathbf{w}}^{(\nu)}$, until convergence:

$$\widehat{\mathbf{w}}^{(\nu+1)} = \widehat{\mathbf{w}}^{(\nu)} - \left(\frac{\partial}{\partial \mathbf{w}^T} g(\widehat{\mathbf{w}}^{(\nu)}) \right)^{-1} g(\widehat{\mathbf{w}}^{(\nu)}). \quad \text{(A 9)}$$

The inverse in (A 9) is of a matrix that is not of the form $\sum_{i=1}^3 a_i \mathscr{S}_i$. To reduce the computational burden of the Netwon method updates, we numerically approximate the inverse in (A 9). First, we define $v(w_{jk}) = \sigma_n \frac{\phi(w_{jk})(y_{jk}-\Phi(w_{jk}))}{\Phi(w_{jk})(1-\Phi(w_{jk}))}$, where we define the vector $v(\mathbf{w}) = \{v(w_{jk})\}_{jk}$, and write the derivative

$$\frac{\partial}{\partial \mathbf{w}^T} g(\mathbf{w}) = \mathbf{B} - \mathbf{I} + \mathbf{DB}. \quad \text{(A 10)}$$

where we define

$$\mathbf{D} = \text{diag}\left\{ \frac{-w_{jk}\phi_{jk}(y_{jk}-\Phi_{jk})-\phi_{jk}^2-\phi_{jk}^2(y_{jk}-\Phi_{jk})(1-2\phi_{jk}\Phi_{jk})}{\Phi_{jk}(1-\Phi_{jk})} \right\}_{jk}.$$

where we let $\phi_{jk} = \phi(w_{jk})$ and $\Phi_{jk} = \Phi(w_{jk})$. The term $\mathbf{DB}$ arises from differentiating $v(\mathbf{w})$ with respect to $\mathbf{w}$. Using the expression in (A 10), we are then able to write the second term in (A 9) as

$$\begin{aligned}\left(\frac{\partial}{\partial \mathbf{w}^T} g(\widehat{\mathbf{w}})\right)^{-1} g(\widehat{\mathbf{w}}) &= (\mathbf{B} - \mathbf{I} + \mathbf{DB})^{-1}\left((\mathbf{B}-\mathbf{I})\mathbf{w} + v(\mathbf{w})\right), \\ &= \mathbf{B}^{-1}\left(\mathbf{I} + \mathbf{D} - \mathbf{B}^{-1}\right)^{-1}\left((\mathbf{B}-\mathbf{I})\mathbf{w} + v(\mathbf{w})\right). \quad \text{(A 11)}\end{aligned}$$

We notice that the matrix $\mathbf{I} + \mathbf{D}$ is diagonal, but not homogeneous (in which case we compute (A 11) directly, with limited computational burden, by exploiting the exchangeable structure). Instead, defining $\mathbf{Q} = (1+\delta)\mathbf{I} - \mathbf{B}^{-1}$ and $\mathbf{M} = \mathbf{D} - \delta\mathbf{I}$, which is diagonal, we make the approximation that

$$\left(\mathbf{I} + \mathbf{D} - \mathbf{B}^{-1}\right)^{-1} = (\mathbf{Q} + \mathbf{M})^{-1} \approx \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}^{-1},$$

which is based on a Neumann series of matrices and relies on the absolute eigenvalues of $\mathbf{M}$ being small (Petersen *et al.*, 2008). We choose δ to be the mean of the minimum and maximum value of $\mathbf{D}$. This choice of δ minimizes the maximum absolute eigenvalue of $\mathbf{M}$, and thus limits the approximation error. Since the inverse of $\mathbf{Q}$ may be computed using the exchangeable inversion formula discussed in Appendix B (in $O(1)$ time), the following approximation represents an improvement in computation from $O(n^3)$ to $O(n^2)$ time:

$$\left(\frac{\partial}{\partial \mathbf{w}^T} g(\widehat{\mathbf{w}})\right)^{-1} g(\widehat{\mathbf{w}}) \approx \mathbf{B}^{-1}\left(\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}^{-1}\right)\left((\mathbf{B}-\mathbf{I})\mathbf{w} + v(\mathbf{w})\right).$$

A.3 Approximation to ρ expectation step

The maximization of the expected likelihood with respect to ρ relies on the computation of $\gamma_i = E[\varepsilon^T \mathscr{S}_i \varepsilon \,|\, \mathbf{y}]/|\Theta_i|$, for $i \in \{1,2,3\}$ (step 2 in Algorithm 1). Under general correlation structure, computation of the expectation $\{\gamma_i\}_{i=1}^3$ for even small networks is prohibitive. To practically compute $\{\gamma_i\}_{i=1}^3$, we make two approximations, which we detail in the following subsections: (1) compute expectations conditioning only on the entries in $\mathbf{y}$ that correspond to the entries in ε being integrated, and (2) approximating these pairwise expectations as linear functions of ρ.

A.3.1 Pairwise expectation

Explicitly, the pairwise approximations to $\{\gamma_i\}_{i=1}^{3}$ we make are:

$$\gamma_1 = \frac{1}{|\Theta_1|}\sum_{jk} E[\varepsilon_{jk}^2 \,|\, \mathbf{y}] \approx \frac{1}{|\Theta_1|}\sum_{jk} E[\varepsilon_{jk}^2 \,|\, y_{jk}], \tag{A 12}$$

$$\gamma_2 = \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, \mathbf{y}] \approx \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}],$$

$$\gamma_3 = \frac{1}{|\Theta_3|}\sum_{jk,lm\in\Theta_3} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, \mathbf{y}] \approx \frac{1}{|\Theta_3|}\sum_{jk,lm\in\Theta_3} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}],$$

where Θ_i is the set of ordered pairs of relations (jk, lm) which correspond entries in $\mathscr{S}_i$ that are 1, for $i \in \{1,2,3\}$. These approximations are natural first-order approximations: recalling that $y_{jk} = \mathbb{1}[\varepsilon_{jk} > -\mathbf{x}_{jk}^T\beta]$, the approximations in (A 12) are based on the notion that knowing the domains of ε_{jk} and ε_{lm} is significantly more informative for $E[\varepsilon_{jk}\varepsilon_{lm} \,|\, \mathbf{y}]$ than knowing the domain of, for example, ε_{ab}.

The approximations in (A 12) are orders of magnitude faster to compute than the expectations when conditioning on all observations $E[\varepsilon_{jk}\varepsilon_{lm} \,|\, \mathbf{y}]$. In particular, when $i \in \{1,3\}$, the expectations are available in closed form:

$$E[\varepsilon_{jk}^2 \,|\, y_{jk}] = 1 - \eta_{jk}\frac{\phi(\eta_{jk})(y_{jk} - \Phi(\eta_{jk}))}{\Phi(\eta_{jk})(1-\Phi(\eta_{jk}))},$$

$$E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}] = \frac{\phi(\eta_{jk})\phi(\eta_{lm})(y_{jk} - \Phi(\eta_{jk}))(y_{lm} - \Phi(\eta_{lm}))}{\Phi(\eta_{jk})\Phi(\eta_{lm})(1-\Phi(\eta_{jk}))(1-\Phi(\eta_{lm}))},$$

where we define $\eta_{jk} = \mathbf{x}_{jk}^T\beta$ and the indices j,k,l and m are distinct. When $i = 2$, that is, $|\{j,k\} \cap \{l,m\}| = 1$, the expectation depends on a two dimensional normal probability integral:

$$\begin{aligned}
&E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}] = \\
&\rho\left(1 - \frac{\bar{\eta}_{jk}\phi(\eta_{jk})}{L_{jk,lm}}\Phi\left(\frac{\bar{\eta}_{lm} - \bar{\rho}\,\bar{\eta}_{jk}}{\sqrt{1-\rho^2}}\right) - \frac{\bar{\eta}_{lm}\phi(\eta_{lm})}{L_{jk,lm}}\Phi\left(\frac{\bar{\eta}_{jk} - \bar{\rho}\,\bar{\eta}_{lm}}{\sqrt{1-\rho^2}}\right)\right) \qquad \text{(A 13)} \\
&+ \frac{1}{L_{jk,lm}}\sqrt{\frac{1-\rho^2}{2\pi}}\,\phi\left(\sqrt{\frac{\eta_{jk}^2 + \eta_{lm}^2 - 2\rho\,\eta_{jk}\eta_{lm}}{1-\rho^2}}\right), \;\; |\{j,k\} \cap \{l,m\}| = 1, \\
L_{jk,lm} &= \mathbb{P}\left((2y_{jk}-1)\varepsilon_{jk} > -\eta_{jk} \cap (2y_{lm}-1)\varepsilon_{lm} > -\eta_{lm}\right),
\end{aligned}$$

where $\bar{\eta}_{jk} = (2y_{jk}-1)\eta_{jk}$, e.g., and $\bar{\rho} = (2y_{jk}-1)(2y_{lm}-1)\rho$.

A.3.2 Linearization

The computation of $E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}]$ in (A 13) requires the computation of $O(n^3)$ bivariate truncated normal integrals $L_{jk,lm}$, which are not generally available in closed form. We observe empirically, however, that the pairwise approximation to γ_2 described in Section A.3.1 above, $\gamma_2 \approx \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}]$, is approximately linear in ρ. This linearity is somewhat intuitive, as the sample mean $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}]$ has expectation equal to ρ, and is thus an asymptotically linear function of ρ. As the sample mean $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}]$ concentrates around its expectation, it concentrates around a linear function of ρ, and it is reasonable to approximate the sample mean $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}]$ as a linear function of ρ. To do so, we compute the approximate values of γ_2 at $\rho = 0$ and if $\rho = 1$. In particular,

$$\begin{aligned}
\gamma_2 &\approx a_2 + b_2\rho, \qquad\qquad \text{(A 14)}\\
a_2 &= \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\,|\,y_{jk}]E[\varepsilon_{lm}\,|\,y_{lm}],\\
&= \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} \frac{\phi(\eta_{jk})\phi(\eta_{lm})(y_{jk}-\Phi(\eta_{jk}))(y_{lm}-\Phi(\eta_{lm}))}{\Phi(\eta_{jk})\Phi(\eta_{lm})(1-\Phi(\eta_{jk}))(1-\Phi(\eta_{lm}))},\\
c_2 &= \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}]\Big|_{\rho=1},\\
b_2 &= c_2 - a_2.
\end{aligned}$$

To compute c_2, we must compute the value of $E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}]$ when $\rho = 1$. Computing $E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}]$ is simple when the values $y_{jk} = y_{lm}$, as in this case $E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}] = E[\varepsilon_{jk}^2\,|\,y_{jk} = y_{lm}]$ since, when $\rho = 1$, $\varepsilon_{jk} = \varepsilon_{lm}$. Approximations must be made in the cases when $y_{jk} \neq y_{lm}$. There are two such cases. In the first, there is overlap between the domains of ε_{jk} and ε_{lm} indicated by $y_{jk} = \mathbb{1}[\varepsilon_{jk} > -\eta_{jk}]$ and $y_{jk} = \mathbb{1}[\varepsilon_{lm} > -\eta_{lm}]$, respectively. We define the domain for ε_{jk} indicated by y_{jk} as $U_{jk} := \{u \in \mathbb{R} : u > (1-2y_{jk})\eta_{jk}\}$. As an example, there is overlap between U_{jk} and U_{lm} when $y_{jk} = 1, y_{lm} = 0$ and $\eta_{lm} < \eta_{jk}$. Then, the dersired expectation may be approximated $E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}] \approx E[\varepsilon_{jk}^2\,|\,\varepsilon_{jk} \in U_{jk} \cap U_{lm}]$. In the second case, when $y_{jk} \neq y_{lm}$ and $U_{jk} \cap U_{lm} =$, we make the approximation by integrating over the sets U_{jk} and U_{lm}. That is, by taking

$$\begin{aligned}
&E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}]\\
&\qquad\approx E[\varepsilon_{jk}^2\,|\,\varepsilon_{jk} \in U_{jk}]\,\mathbb{P}(\varepsilon_{jk} \in U_{jk}) + E[\varepsilon_{lm}^2\,|\,\varepsilon_{lm} \in U_{lm}]\,\mathbb{P}(\varepsilon_{lm} \in U_{lm}).
\end{aligned}$$

To summarize, we compute c_2 in (A 14) when $\rho = 1$ by using the following approximation to $E[\varepsilon_{jk}\varepsilon_{lm}\,|\,\mathbf{y}]\Big|_{\rho=1}$:

$$\begin{cases}
E[\varepsilon_{jk}^2\,|\,\varepsilon_{jk} > \max(-\eta_{jk},-\eta_{lm})], & y_{jk} = 1 \text{ and } y_{lm} = 1,\\
E[\varepsilon_{jk}^2\,|\,\varepsilon_{jk} < \min(-\eta_{jk},-\eta_{lm})], & y_{jk} = 0 \text{ and } y_{lm} = 0,\\
E[\varepsilon_{jk}^2\,|\,\varepsilon_{jk} \in U_{jk} \cap U_{lm}], & U_{jk} \cap U_{lm} \neq \emptyset,\\
E[\varepsilon_{jk}^2\,|\,\varepsilon_{jk} \in U_{jk}]\,\mathbb{P}(\varepsilon_{jk} \in U_{jk}) + E[\varepsilon_{lm}^2\,|\,\varepsilon_{lm} \in U_{lm}]\,\mathbb{P}(\varepsilon_{lm} \in U_{lm}) & U_{jk} \cap U_{lm} = \emptyset.
\end{cases}$$

A.4 Missing data

In this subsection, we describe estimation of the PX model in the presence of missing data. We present the maximization of $\ell_{\mathbf{y}}$ with respect to β first. Second, we discuss maximization of $\ell_{\mathbf{y}}$ with respect to ρ. Finally, we give a note on prediction from the PX model when data are missing.

Update β:

To maximize $\ell_{\mathbf{y}}$ with respect to β (Step 1 of Algorithm 1) in the presence of missing data, we impute the missing values of $\mathbf{X}$ and $\mathbf{y}$. We make the decision to impute missing values since much of the speed of estimation of the PX model relies on exploitation of the particular network structure, and, when data are missing, this structure is more difficult to leverage. We impute entries in **X** with the mean value of the covariates. For example, if $x_{jk}^{(1)}$ is missing, we replace it with the sample mean $\frac{1}{|\mathscr{M}^c|}\sum_{lm\in\mathscr{M}^c} x_{lm}^{(1)}$, where the superscript (1) refers to the first entry in $\mathbf{x}_{jk}$ and $\mathscr{M}$ is the set of relations for which data are missing. If y_{jk} is missing, we impute y_{jk} with $\mathbb{1}[w_{jk} > -\bar{\eta}]$, where $\bar{\eta} = \frac{1}{|\mathscr{M}^c|}\sum_{lm\in\mathscr{M}^c} \mathbf{x}_{lm}^T\widehat{\beta}$ and we compute $\mathbf{w} = E[\varepsilon \,|\, \mathbf{y}]$ using the procedure in Section A.2. We initialize this procedure at $\mathbf{w}^{(0)}$, where any missing entries $jk \in \mathscr{M}$ are initialized with $w_{jk}^{(0)} = 0$. Given the imputed **X** and **y**, the estimation routine may be accomplished as described in Algorithm 1.

Update ρ:
To maximize $\ell_{\mathbf{y}}$ with respect to ρ (Step 2 of Algorithm 1), we approximate $\{\gamma_i\}_{i=1}^3$ using only observed values. Using the pairwise expressions in (A 12), the expressions for the expectation step under missing data are

$$
\begin{aligned}
\gamma_1 &\approx \frac{1}{|\mathscr{M}^c|}\sum_{jk\in\mathscr{M}^c} E[\varepsilon_{jk}^2 \,|\, y_{jk}], &\text{(A 15)}\\
\gamma_2 &\approx \frac{1}{|\mathscr{A}^{(s)}|}\sum_{jk,lm\in\mathscr{A}^{(s)}} E[\varepsilon_{jk}\varepsilon_{lm} \,|\, y_{jk}, y_{lm}].\\
\gamma_3 &\approx \frac{\sum_{jk,lm\in\Theta_3} E[\varepsilon_{jk} \,|\, y_{jk}]E[\varepsilon_{lm} \,|\, y_{lm}]\mathbb{1}[jk\in\mathscr{M}^c]\mathbb{1}[lm\in\mathscr{M}^c]}{\sum_{jk,lm\in\Theta_3}\mathbb{1}[jk\in\mathscr{M}^c]\mathbb{1}[lm\in\mathscr{M}^c]},\\
&\approx \frac{1}{|\Theta_3|}\Bigg(\bigg(\frac{|\Theta_1|}{|\mathscr{M}^c|}\sum_{jk\in\mathscr{M}^c} E[\varepsilon_{jk} \,|\, y_{jk}]\bigg)^2 - \frac{|\Theta_1|}{|\mathscr{M}^c|}\sum_{jk\in\mathscr{M}^c} E[\varepsilon_{jk} \,|\, y_{jk}]^2\\
&\qquad - \frac{|\Theta_2|}{|\mathscr{A}^{(s)}|}\sum_{jk,lm\in\mathscr{A}^{(s)}} E[\varepsilon_{jk} \,|\, y_{jk}]E[\varepsilon_{lm} \,|\, y_{lm}]\Bigg),
\end{aligned}
$$

where we only subsample pairs of relations that are observed such that $\mathscr{A}^{(s)} \subset \Theta_2 \cap \mathscr{M}^c$. Then, given the values of $\{\gamma_i\}_{i=1}^3$ in (A 15), the maximization of $\ell_{\mathbf{y}}$ with respect to ρ (Step 2 in Algorithm 1) may proceed as usual.

Prediction:
Joint prediction in the presence of missing data is required for out-of-sample evaluation of the EMM estimator, for example, for cross validation studies in Section 8. In this setting, model estimation is accomplished by imputing values in **X** and **y** earlier in this section under the '**Update β**' subheading. Then, prediction may be performed by proceeding as described in Section 6 with the full observed **X** matrix and imputing the missing values in **y** (again as described above in this section under the '**Update β**' subheading).

B Parameters of undirected exchangeable network covariance matrices

In this section, we give a 3×3 matrix equation to invert Ω rapidly. This equation also gives a basis to compute the partial derivatives $\left\{\frac{\partial \phi_i}{\partial p_j}\right\}$, which we require for the EMM algorithm.

We define an *undirected exchangeable network covariance matrix* as those square, positive definite matrices of the form

$$\Omega(\phi) = \sum_{i=1}^{3} \phi_i \mathscr{S}_i.$$

We find empirically that the inverse matrix of any undirected exchangeable network covariance matrix has the same form, that is $\Omega^{-1} = \sum_{i=1}^{3} \mathbf{p}_i \mathscr{S}_i$. Using this fact and the particular forms of the binary matrices $\{\mathscr{S}_i\}_{i=1}^{3}$, one can see that there are only three possible row-column inner products in the matrix multiplication $\Omega\Omega^{-1}$, those pertaining to row-column pairs of the form (ij,ij), (ij,ik), and (ij,kl) for distinct indices i, j, k, and l. Examining the three products in terms of the parameters in ϕ and $\mathbf{p}$, and the fact that $\Omega\Omega^{-1} = \mathbf{I}$, we get the following matrix equation for the parameters $\mathbf{p}$ given ϕ

$$\mathbf{C}(\phi)\mathbf{p} = [1,0,0]^T, \quad \text{(B 1)}$$

where the matrix $\mathbf{C}(\phi)$ is given by

$$\begin{bmatrix} \phi_1 & 2(n-2)\phi_2 & \frac{1}{2}(n-2)(n-3)\phi_3 \\ \phi_2 & \phi_1 + (n-2)\phi_2 + (n-3)\phi_3 & (n-3)\phi_2 + \left(\frac{1}{2}(n-2)(n-3) - n + 3\right)\phi_3 \\ \phi_3 & 4\phi_2 + (2n-8)\phi_3 & \phi_1 + (2n-8)\phi_2 + \left(\frac{1}{2}(n-2)(n-3) - 2n + 7\right)\phi_3 \end{bmatrix}.$$

Then, we may invert Ω with a 3×3 inverse to find the parameters $\mathbf{p}$ of Ω^{-1}. Explicitly solving these linear equations, the expressions for $\mathbf{p}$ are given by

$$\begin{aligned} p_1 &= 1 - (2n-4)p_2, \qquad \text{(B 2)} \\ p_2 &= \frac{1 + (n-3)p_3}{(2n-4)\rho - n + 2 - 1/\rho}, \\ p_3 &= \frac{-4\rho^2}{(n-3)4\rho + (1 + (2n-8)\rho)((2n-4)\rho - n + 2 - 1/\rho)}. \end{aligned}$$

Taking only the largest terms in n, one may approximate the values in $\mathbf{p}$ as follows, which will be useful in following theoretical development:

$$\begin{aligned} p_1 &\approx \frac{1}{1-2\rho} + O(n^{-1}), \qquad \text{(B 3)} \\ p_2 &\approx \frac{-1}{n(1-2\rho)} + O(n^{-2}), \\ p_3 &\approx \frac{2}{n^2(1-2\rho)} + O(n^{-3}). \end{aligned}$$

The equation in (B 1) allows one to compute the partial derivatives $\left\{\frac{\partial \phi_i}{\partial p_j}\right\}$. First, based on (B 1), we can write $\mathbf{C}(\mathbf{p})\phi = [1,0,0]^T$. Then, we note that the matrix function $\mathbf{C}(\phi)$ in (B 1) is linear in the terms ϕ, and thus, we may write $\mathbf{C}(\mathbf{p}) = \sum_{j=1}^{3} p_j \mathbf{A}_j^{(n)}$ for some matrices $\left\{\mathbf{A}_j^{(n)}\right\}_{j=1}^{3}$ that depend on n. Differentiating both sides of $\mathbf{C}(\mathbf{p})\phi = [1,0,0]^T$ with respect to p_j and solving gives

$$\frac{\partial \phi}{\partial p_j} = -\mathbf{C}(\mathbf{p})^{-1}\mathbf{A}_j^{(n)}\mathbf{C}(\mathbf{p})^{-1}[1,0,0]^T,$$

which holds for all $j \in \{1,2,3\}$.

C Theoretical support

In this section, we outline proofs suggesting that the estimators resulting from the EMM algorithm are consistent.

C.1 Consistency of $\widehat{\beta}_{EMM}$

The estimator of β resulting from the EMM algorithm, $\widehat{\beta}_{EMM}$, depends on the estimated value of ρ, $\widehat{\rho}_{EMM}$, through the covariance matrix Ω. Explicitly, given Ω, the EMM estimator

$$\widehat{\beta}_{EMM} = (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Omega^{-1}\widehat{E[\mathbf{z} \mid \mathbf{y}]}, \quad \text{(C 1)}$$

where $\widehat{E[\mathbf{z} \mid \mathbf{y}]}$ represents the estimation and approximation of $E[\mathbf{z} \mid \mathbf{y}]$ described in the EMM algorithm. This estimator is difficult to analyze in general, because, in principle, $\widehat{E[z_{jk} \mid \mathbf{y}]}$ depends on every entry in $\mathbf{y}$, and the effects of the approximations are difficult to evaluate. Instead of direct analysis, to evaluate consistency of $\widehat{\beta}_{EMM}$, we define a bounding estimator that is easier to analyze,

$$\widehat{\beta}_{bound} = (\mathbf{X}^T\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Omega^{-1}\mathbf{u}, \quad u_{jk} = E[z_{jk} \mid y_{jk}]. \quad \text{(C 2)}$$

It is immediately clear that $\widehat{\beta}_{bound}$ is unbiased, since $E[u_{jk}] = \mathbf{x}_{jk}^T\beta$. Further, the approximations made in the EMM algorithm are meant to bound $||\widehat{\beta}_{EMM} - \beta^*_{MLE}||_2^2 \leq ||\widehat{\beta}_{bound} - \beta^*_{MLE}||_2^2$, where β^*_{MLE} is the true maximum likelihood estimator. That is, the expectation estimator we compute $\widehat{E[\mathbf{z} \mid \mathbf{y}]}$ takes into account correlation information through Ω, and is thus closer to the true expectation, $E[\mathbf{z} \mid \mathbf{y}]$, than $\mathbf{u}$. Then, we also have that $\widehat{\beta}_{EMM}$ is closer to β^*_{MLE} than $\widehat{\beta}_{bound}$. Then, consistency of $\widehat{\beta}_{bound}$ implies consistency of $\widehat{\beta}_{EMM}$, since we assume that the true MLE is consistent.

We now establish consistency of $\widehat{\beta}_{bound}$. We make the following assumptions:

1. The true model follows a latent variable model,

$$\begin{aligned} \mathbb{P}(y_{ij} = 1) &= \mathbb{P}\left(\mathbf{x}_{ij}^T\beta + \varepsilon_{ij} > 0\right), \quad \text{(C 3)} \\ E[\varepsilon_{jk}] &= 0. \end{aligned}$$

   where ε is not necessarily normally distributed.
2. The design matrix $\mathbf{X}$ is such that the expressions $n^{-(1+i)}\mathbf{X}^T\mathscr{S}_i\mathbf{X}$, for $i \in \{1,2,3\}$, converge in probability to constant matrices.
3. The fourth moments of $\mathbf{X}$ and ε are bounded, $||\mathbf{x}_{jk}||_4 \leq C_1 < \infty$ and $E[\varepsilon_{jk}^4] \leq C_2 < \infty$.
4. The estimator of ρ is such that $\Omega(\widehat{\rho})$ converges in probability to some positive definite matrix.
5. The independence assumption for relations that do not share an actor holds, such that ε_{jk} is independent ε_{lm} whenever actors j, k, l, and m are distinct.

The first assumption defines the meaning of the true coefficient β. The second assumption is a standard condition required for most regression problems; a similar condition is required for consistency of any estimator which accounts for correlation in generalized linear model. We evaluate the second assumption in the following section, when we analyze $\widehat{\rho}_{EMM}$. The fourth assumption defines the minimal independence structure.

We start by noticing that $\mathbf{u} = \mathbf{X}\beta + \varepsilon$, such that

$$\widehat{\beta}_{bound} = \beta + \left(n^{-2} \sum_{i=1}^{3} p_i \mathbf{X}^T \mathscr{S}_i \mathbf{X} \right)^{-1} \left(n^{-2} \sum_{i=1}^{3} p_i \mathbf{X}^T \mathscr{S}_i \mathbf{v} \right), \quad v_{jk} = E[\varepsilon_{jk} \mid y_{jk}]. \tag{C 4}$$

Then, as noted in the previous paragraph, the bounding estimator is unbiased, $E[\widehat{\beta}_{bound}] = \beta$. It remains to establish sufficient conditions for which $\widehat{\beta}_{bound}$ converges to its expectation in probability. Noting the orders of $\{p_i\}_i$ in (B 3), we immediately have that $n^{-2}\mathbf{X}^T\Omega^{-1}\mathbf{X}$ converges in probability to a constant. A sufficient condition to establish that $\left(n^{-2}\sum_{i=1}^{3} p_i \mathbf{X}^T \mathscr{S}_i \mathbf{v}\right)$ converges in probability to its expectation (zero) is that its variance tends to zero. Expanding this variance expression,

$$\begin{aligned} \operatorname{var}\left(n^{-2} \sum_{i=1}^{3} p_i \mathbf{X}^T \mathscr{S}_i \mathbf{v} \right) &= n^{-4} \sum_{i=1}^{3} \sum_{j=1}^{3} p_i p_j \mathbf{X}^T \mathscr{S}_i E[\mathbf{v}\mathbf{v}^T] \mathscr{S}_j \mathbf{X}, \qquad \text{(C 5)} \\ &= n^{-4} \sum_{i=1}^{3} \sum_{j=1}^{3} p_i p_j \sum_{jk,lm \in \Theta_i} \sum_{rs,tu \in \Theta_j} \mathbf{x}_{jk} \mathbf{x}_{rs}^T E[v_{lm} v_{tu}]. \end{aligned}$$

By assumption, every term in the sum expression in (C 5) is bounded. Also by assumption, the expectation $E[v_{lm}v_{tu}]$ is zero whenever the relations lm and tu do not share an actor. Using the expressions in (B 3) ($p_i \propto n^2|\Theta_i|^{-1}$) and counting terms,

$$\operatorname{var}\left(n^{-2} \sum_{i=1}^{3} p_i \mathbf{X}^T \mathscr{S}_i \mathbf{v} \right) \propto n^{-4} \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{n^2}{|\Theta_i|} \frac{n^2}{|\Theta_j|} \frac{|\Theta_i||\Theta_j|}{n} = O(n^{-1}).$$

Thus, the variance of $\widehat{\beta}_{bound}$ converges to zero, so that $\widehat{\beta}_{bound}$ converges in probability to the true β, as does $\widehat{\beta}_{EMM}$.

C.2 Consistency of $\widehat{\rho}_{EMM}$

Using the expressions in (B 3) and differentiating the expected log-likelihood with respect to ρ, the maximum likelihood estimator is

$$\widehat{\rho}_{MLE} = \frac{1}{2} + \frac{1}{n^3} E[\varepsilon^T \mathscr{S}_2 \varepsilon \mid \mathbf{y}] - \frac{1}{n^2} E[\varepsilon^T \varepsilon \mid \mathbf{y}] - \frac{2}{n^4} E[\varepsilon^T \mathscr{S}_3 \varepsilon \mid \mathbf{y}] + O(n^{-1}). \tag{C 6}$$

In the EMM algorithm, we approximate the expectations in (C 6) using pairwise conditioning. Then, we have that

$$\begin{aligned} \widehat{\rho}_{EMM} = \frac{1}{2} + \frac{1}{n^3} \sum_{jk,lm \in \Theta_2} E[\varepsilon_{jk}\varepsilon_{lm} \mid y_{jk}, y_{lm}] - \frac{1}{n^2} \sum_{jk} E[\varepsilon_{jk}^2 \mid y_{jk}] \ldots \text{(C 7)} \\ \ldots - \frac{2}{n^4} \sum_{jk,lm \in \Theta_3} E[\varepsilon_{jk} \mid y_{jk}] E[\varepsilon_{lm} \mid y_{lm}] + O(n^{-1}). \end{aligned}$$

According to the exchangeability assumption of the errors, the pairwise expectations are known, and the EMM estimator of ρ is unbiased, $E[\widehat{\rho}_{EMM}] = E[\varepsilon_{jk}\varepsilon_{lm}] = \rho$. The EMM estimator $\widehat{\rho}_{EMM}$, converges to its expectation when the sums of conditional expectations in (C 7) converge to their expectations. This occurs when the variances of these sums tend to zero. This fact can be established by similar counting arguments as in the previous subsection. For example,

$$\operatorname{var}\left(\frac{1}{n^3}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm} \mid y_{jk}, y_{lm}]\right) = n^{-6}\sum_{jk,lm\in\Theta_2}\sum_{jk,lm\in\Theta_2}\left(E[E[\varepsilon_{jk}\varepsilon_{lm} \mid y_{jk}, y_{lm}]E[\varepsilon_{rs}\varepsilon_{tu} \mid y_{rs}, y_{tu}]] - \rho^2\right),$$

$$= n^{-6}\frac{|\Theta_2||\Theta_2|}{n} = O(n^{-1}),$$

since $E[\varepsilon_{jk}\varepsilon_{lm} \mid y_{jk}, y_{lm}]$ is independent $E[\varepsilon_{rs}\varepsilon_{tu} \mid y_{rs}, y_{tu}]$ whenever all the indices $\{j,k,l,m,r,s,t,u\}$ are distinct. Thus, each of the sums of expectations in (C 7) has variance that tends to zero, so that they converge to their marginal expectations, and $\widehat{\rho}_{EMM}$ is consistent.

C.3 Consistency under misspecification

In the discussion of consistency of the EMM estimator, we did not require the assumption of latent normality, nor of exchangeability of the latent errors (we do require a small assumption that the sequence of constants $n^{-3}E[\varepsilon^T \mathscr{S}_2 \varepsilon_{lm}]$ converges to some constant on $[0,1/2)$). Hence, when the data generating mechanism is non-Gaussian and non-exchangeable, we expect $\widehat{\rho}_{EMM}$ to converge to the pseudo-true ρ. The pseudo-true ρ is the value which minimizes the Kullback-Leibler divergence from the modeled (Gaussian, exchangeable) distribution to the true distribution (Huber, 1967; Dhaene, 1997). In the discussion of consistency of $\widehat{\beta}_{EMM}$, we only require that $\widehat{\rho}_{EMM}$ converges to a fixed value on the interval $[0,1/2)$, such that $\Omega(\rho)$ is positive definite. Again, when the data generating mechanism is non-Gaussian and non-exchangeable, we expect $\widehat{\beta}_{EMM}$ to converge to the pseudo-true β. When the true data generating mechanism is Gaussian (but not necessarily exchangeable), the limiting pseudo-true value for $\widehat{\beta}_{EMM}$ should be the true value.

D Simulation studies

In this section we present details pertaining to the second simulation study in Section 7.

D.1 Evaluation of estimation of β

See Section 7.2 for a description of the simulation study to evaluate performance in estimating β. We provide further details in the rest of this paragraph. We generated each $\{x_{1i}\}_{i=1}^n$ as iid Bernoulli$(1/2)$ random variables, such that the second covariate is an indicator of both $x_{1i} = x_{1j} = 1$. Each of $\{x_{2i}\}_{i=1}^n$ and $\{x_{3ij}\}_{ij}$ were generated from iid standard normal random variables. We fixed $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]^T = [-1, 1/2, 1/2, 1/2]^T$ throughout the simulation study. When generating from the latent eigenmodel in (5), we set $\Lambda = \mathbf{I}$, $\sigma_a^2 = 1/6$, $\sigma_u^2 = 1/\sqrt{6}$, and $\sigma_\xi^2 = 1/3$.

To further investigate the source of poor performance of the `amen` estimators of the social relations and latent eigenmodels, we computed the bias and the variance of estimators when generating from the PX model and the latent eigenmodel in Figures D 1 and D 2, respectively. Figures D 1 and D 2 show that the variances of the `amen` estimators of the social relations and latent eigenmodels are similar to the PX model, however, that the bias of the `amen` estimators are substantially larger.

Both the EMM estimator of the PX model and `amen` estimator of the social relations model provide estimates of ρ. We computed the RMSE for each estimator, for each $\mathbf{X}$

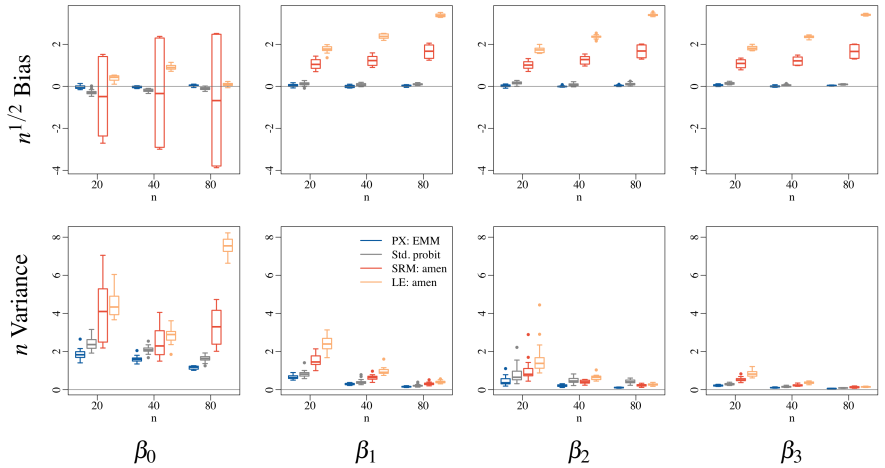


Fig. D 1. **PX model:** Scaled bias and variance of estimators of β for a given $\mathbf{X}$ when generating from the PX model. Variability captured by the boxplots reflects variation with $\mathbf{X}$.

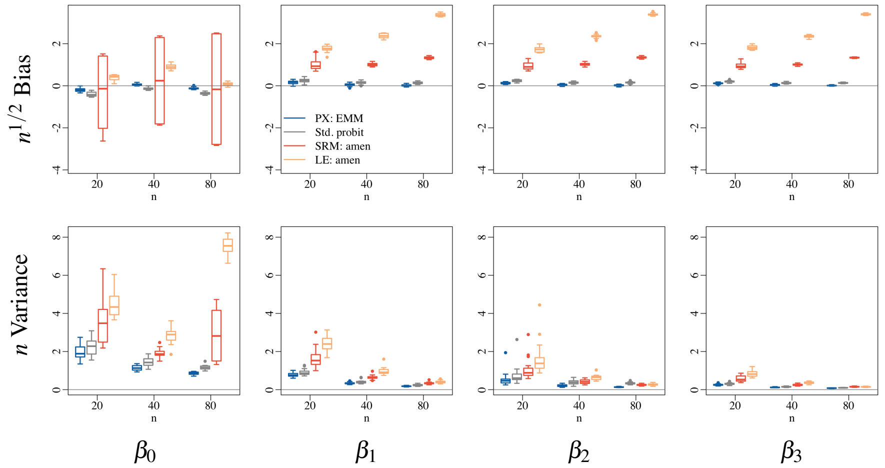


Fig. D 2. **LE model:** Scaled bias and variance of estimators of β for a given $\mathbf{X}$ when generating from the latent eigenmodel. Variability captured by the boxplots reflects variation with $\mathbf{X}$.

realization, when generating from the PX model. In Figure D 3, the RMSE plot for $\widehat{\rho}$ shows that the MSE, and the spread of the MSE, decreases with n for the EMM estimator, suggesting that the EMM estimator of ρ is consistent. As with the β parameters, the `amen` estimator displays substantially larger RMSE than the EMM estimator of ρ.

D.2 Remaining coefficients in t simulation

We simulated from the PX model, modified to have heavier-tailed t_5 error distribution. The scaled RMSE when estimating all entries in β is given in Figure D 4. All coefficient estimators, for both PX: EMM and standard probit regression, appear consistent, but the PX: EMM has lower RSME.

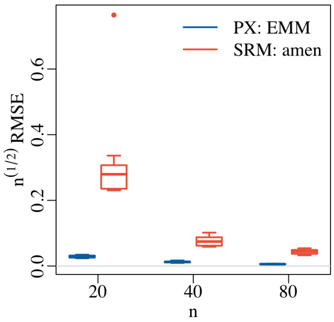
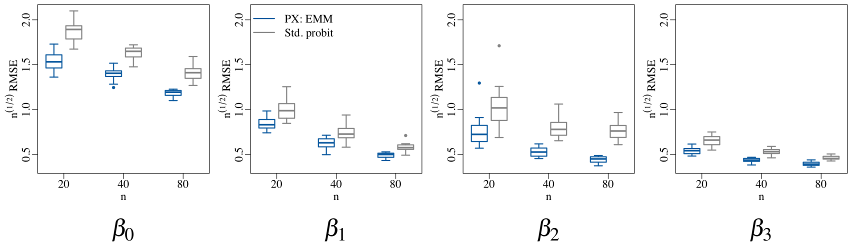


Fig. D 3. RMSE, scaled by $n^{1/2}$, of the EMM estimator and `amen` estimator of the social relations model of ρ when generating from the PX model. Variability captured by the boxplots reflects variation in $n^{1/2}$RMSE with $\mathbf{X}$.



β_0 β_1 β_2 β_3

Fig. D 4. *t* **model:** Scaled RMSE, for PX: EMM and standard probit regression, when generating from the PX model modified to have latent errors with heavier-tailed distribution.

E Analysis of political books network

In this section, we present additional predictive results and verify the efficacy of an approximation made by the EMM algorithm when analyzing the political books network data set.

E.1 Prediction performance using ROC AUC

In Section 8, we use area under the precision-recall curve to evaluation predictive performance on the political books network data set. Figure E 1 shows the results of the cross validation study, described in Section 8, as measured by area under the receiver operating characteristic (ROC AUC). The conclusions are the same as those given in Section 8: the PX model appears to account for the inherent correlation in the data with estimation runtimes that are orders of magnitude faster than existing approaches.

E.2 Linear approximation in ρ in EMM algorithm

In Section 5.2, we discuss a series of approximations to the E-step of an EM algorithm to maximize $\ell_{\mathbf{y}}$ with respect to ρ. One approximation is a linearization of the sample average $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk}, y_{lm}]$ with respect to ρ. In Figure E 2, we confirm that this approximation is reasonable for the political books network data set. Figure E 2 shows that the linear approximation to $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk}, y_{lm}]$ (dashed blue line), as described in detail in Section A.3, agrees well with the true average of the pairwise expectations (solid orange line).

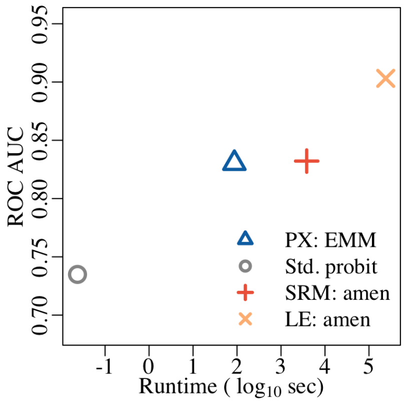


Fig. E 1. Out-of-sample performance in 10-fold cross validation, as measured by area under the precision-recall curve (ROC AUC), plotted against mean runtime in the cross validation for Krebs' political books network. The estimators are standard probit assuming independent observations (Std. probit), the proposed PX estimator as estimated by EMM (PX: EMM), the social relations model as estimated by `amen` (SRM: amen), and the latent eigenmodel as estimated by `amen` (LE: amen).

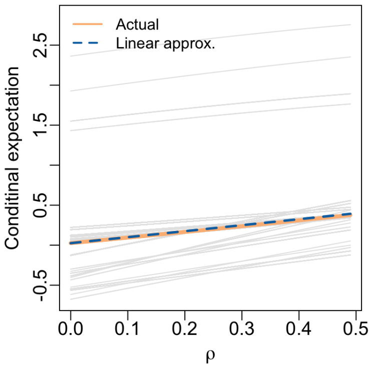


Fig. E 2. The average of all pairwise expectations $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}]$ is shown in orange, and the linear approximation to this average, described in Section 5, is shown in dashed blue. In addition, pairwise conditional expectations $E[\varepsilon_{jk}\varepsilon_{lm}\,|\,y_{jk},y_{lm}]$ are shown in light gray, for a random subset of 500 relation pairs $(jk,lm)\in\Theta_2$.