Network Science **11** (3): 502–535, 2023
doi:10.1017/nws.2023.12

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Regression of binary network data with exchangeable latent errors

Frank W. Marrs[1] and Bailey K. Fosdick[2]

[1]Los Alamos National Laboratory, Los Alamos, NM, USA and [2]Department of Biostatistics & Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
Corresponding author: Frank W. Marrs; Email: fmarrs3@lanl.gov

Action Editor: Paolo Pin

Abstract

Undirected, binary network data consist of indicators of symmetric relations between pairs of actors. Regression models of such data allow for the estimation of effects of exogenous covariates on the network and for prediction of unobserved data. Ideally, estimators of the regression parameters should account for the inherent dependencies among relations in the network that involve the same actor. To account for such dependencies, researchers have developed a host of latent variable network models; however, estimation of many latent variable network models is computationally onerous and which model is best to base inference upon may not be clear. We propose the probit exchangeable (PX) model for undirected binary network data that is based on an assumption of exchangeability, which is common to many of the latent variable network models in the literature. The PX model can represent the first two moments of any exchangeable network model. We leverage the EM algorithm to obtain an approximate maximum likelihood estimator of the PX model that is extremely computationally efficient. Using simulation studies, we demonstrate the improvement in estimation of regression coefficients of the proposed model over existing latent variable network models. In an analysis of purchases of politically aligned books, we demonstrate political polarization in purchase behavior and show that the proposed estimator significantly reduces runtime relative to estimators of latent variable network models, while maintaining predictive performance.

Keywords: expectation-maximization; latent variable models; probit regression; exogenous regression; political networks

1. Introduction

Undirected binary network data measure the presence or absence of a relationship between pairs of actors and have recently become extremely common in the social and biological sciences. Some examples of data that are naturally represented as undirected binary networks are international relations among countries (Fagiolo et al., 2008), gene co-expression (Zhang & Horvath, 2005), and interactions among students (Han et al., 2016). We focus on an example of politically aligned books, where a relation exists between two books if they were frequently purchased by the same person on Amazon.com. Our motivations are estimation of the effects of exogenous covariates, such as the effect of alignment of political ideologies of pairs of books on the propensity for books to be purchased by the same consumer, and the related problem of predicting unobserved relations using book ideological information. For example, predictions of relations between new books and old books could be used to recommend new books to potential purchasers.

A binary, undirected network $\{y_{ij} \in \{0,1\}: i,j \in \{1,\ldots,n\}, i<j\}$, which we abbreviate $\{y_{ij}\}_{ij}$, may be represented as an $n \times n$ symmetric adjacency matrix which describes the presence or absence of relationships between unordered pairs of n actors. The diagonal elements of the matrix

$\{y_{ii} : i \in \{1, \ldots, n\}\}$ are assumed to be undefined, as we do not consider actor relations with him/herself. We use $\mathbf{y}$ to refer to the $\binom{n}{2}$ vector of network relations formed by a columnwise vectorization of the upper triangle of the matrix corresponding to $\{y_{ij}\}_{ij}$.

A regression model for the probability of observing a binary outcome is the probit model, which can be expressed

$$\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij} > 0\right), \tag{1}$$

where ϵ_{ij} is a mean-zero normal random error, $\mathbf{x}_{ij}$ is a fixed vector of covariates corresponding to relation ij, and $\boldsymbol{\beta}$ is a vector of coefficients to be estimated. When each entry in the error network $\{\epsilon_{ij}\}_{ij}$ is independent of the others, estimation of the probit regression model in (1) is straightforward and proceeds via standard gradient methods for maximum likelihood estimation of generalized linear models (Greene, 2003). The assumption of independence of $\{\epsilon_{ij}\}_{ij}$ may be appropriate when the mean $\left\{\mathbf{x}_{ij}^T \boldsymbol{\beta}\right\}_{ij}$ represents nearly all of the dependence in the network $\{y_{ij}\}_{ij}$. However, network data naturally contain excess dependence beyond the mean: the errors ϵ_{ij} and ϵ_{ik} both concern actor i (see Faust & Wasserman, 1994, e.g., for further discussion of dependencies in network data). In the context of the political books data set, the propensity of "Who's Looking Out For You?" by Bill O'Reilly to be purchased by the same reader as "Deliver Us from Evil" by Sean Hannity may be similar to the propensity of "Who's Looking Out For You?" and "My Life" by Bill Clinton to be co-purchased simply because "Who's Looking Out For You?" is a popular book. Or, in a student friendship network, the friendship that Julie makes with Steven may be related to the friendship that Julie makes with Asa due to Julie's gregariousness. Unlike the case of typical linear regression, the estimator that maximizes the likelihood of the generalized linear regression model in (1), when assuming independence of each entry in the error network $\{\epsilon_{ij}\}_{ij}$, is not unbiased for $\boldsymbol{\beta}$. Ignoring the excess dependence in $\{\epsilon_{ij}\}_{ij}$ can thus be expected to result in poor estimation of $\boldsymbol{\beta}$ and poor out-of-sample predictive performance. We observe this phenomenon in the simulation studies and analysis of the political books network (see Sections 7 and 8, respectively). Thus, estimators of $\boldsymbol{\beta}$ and $\mathbb{P}(y_{ij} = 1)$ in (1) for the network $\{y_{ij}\}_{ij}$ should ideally account for the excess dependence of network data. A host of regression models exist in the literature that do just this; we briefly review these here.

A method used to account for excess dependence in regression of binary network data is the estimation of generalized linear mixed models, which were first introduced for repeated measures studies (Stiratelli et al., 1984; Breslow & Clayton, 1993). In these models, a random effect, that is, latent variable, is estimated for each individual in the study, to account for possible individual variation. Warner et al. (1979) used latent variables to account for excess network dependence when analyzing data with continuous measurements of relationships between actors, and Holland & Leinhardt (1981) extended their approach to networks consisting of binary observations. Hoff et al. (2002) further extended this approach to include nonlinear functions of latent variables, and since then, many variations have been proposed (Handcock et al., 2007; Hoff, 2008; Sewell & Chen, 2015). We refer to parametric network models wherein the observations are independent conditional on random latent variables as "latent variable network models," which we discuss in detail in Section 2. Separate latent variable approaches may lead to vastly different estimates of $\boldsymbol{\beta}$, and it may not be clear which model's estimate of $\boldsymbol{\beta}$, or prediction, to choose. Goodness-of-fit checks are the primary method of assessing latent variable network model fit (Hunter et al., 2008b); however, selecting informative statistics is a well-known challenge. Finally, latent variable network models are typically computationally burdensome to estimate, often relying on Markov chain Monte Carlo methods.

Another approach to estimating covariate effects on network outcomes is the estimation of exponential random graph models, known as ERGMs. ERGMs represent the probability of relation formation using a generalized exponential family distribution, $\mathbb{P}(y_{ij} = 1) \propto \exp(\mathbf{t}(\mathbf{y}_{ij}, \mathbf{x}_{ij})^T \theta)$, where θ is a vector of parameters to be estimated. In this flexible formulation, the effects of the

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

exogenous covariates are included in the network statistics $\mathbf{t}(\mathbf{y}_{ij}, \mathbf{x}_{ij})$. ERGMs also account for excess network dependence using the network statistics $\mathbf{t}(\mathbf{y}_{ij}, \mathbf{x}_{ij})$, such as counts of the number of observed triangles or the number of "2-stars"—pairs of indicated relations that share an actor. ERGMs were developed by Frank & Strauss (1986) and Snijders et al. (2006) and are typically estimated using Markov chain Monte Carlo (MCMC) approximations to posterior distributions (Snijders, 2002; Handcock et al., 2019; Hunter et al., 2008a). ERGMs have been shown to be prone to place unrealistic quantities of probability mass on networks consisting of all "1"s or all "0"s (Handcock et al., 2003; Schweinberger, 2011), and the estimation procedures may be slow to complete (Caimo & Friel, 2011). Further, parameter estimates typically cannot be generalized to populations outside the observed network (Shalizi & Rinaldo, 2013).

A final approach to account for excess network dependence is to explicitly model the correlation among network observations. This is the approach we take in this paper. In this approach, an unobserved normal random variable, z_{ij}, is proposed to underlie each data point, such that $y_{ij} = \mathbf{1}[z_{ij} > 0]$ for $\mathbf{z} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}(\boldsymbol{\theta}))$. In this formulation, excess dependence due to the network is accounted for in $\boldsymbol{\Omega}$. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ of the distribution of the unobserved normal random variables $\{z_{ij}\}_{ij}$ may be estimated using likelihood methods. For example, Ashford & Sowden (1970) propose likelihood ratio hypothesis tests and Ochi & Prentice (1984) give closed-form parameter estimators for studies of repeated observations on the same individual, such that $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is block diagonal. In more general scenarios, such as unrestricted correlation structures, methods such as semi-parametrics (Connolly & Liang, 1988), pseudo-likelihoods (Le Cessie & Van Houwelingen, 1994), and MCMC approximations to EM algorithms (Chib & Greenberg, 1998; Li & Schafer, 2008) are employed for estimation.

In this paper, we propose the probit exchangeable (PX) model, a parsimonious regression model for undirected binary network data based on an assumption of exchangeability of the unobserved normal random variables $\{z_{ij}\}_{ij}$. The assumption of exchangeability is pervasive in random network models and, in fact, underlies many of the latent variable network models (see Section 3 for a detailed discussion of exchangeability).[1] We show that, under exchangeability, the excess network dependence in $\{z_{ij}\}_{ij}$ may be quantified using a single parameter ρ such that $\boldsymbol{\Omega}(\boldsymbol{\theta}) = \boldsymbol{\Omega}(\rho)$. This fact remains regardless of the particular exchangeable generating model, and thus, our approach can be seen as subsuming exchangeable latent network variable models, at least up to the second moment of their latent distributions. The proposed model may be rapidly estimated using an expectation-maximization (EM) algorithm to attain a numerical approximation to the maximum likelihood estimator, where we make approximations in the expectation step for runtime considerations. The estimation scheme we employ is similar to those used to estimate generalized linear mixed models in the literature (Littell et al., 2006; Gelman & Hill, 2006).

This paper is organized as follows. As latent variable network models are strongly related to our work, we review them in detail in Section 2. We provide supporting theory for exchangeable random network models and their connections to latent variable network models in Section 3. In Section 4, we define the PX model and then the estimation thereof in Section 5. In Section 6, we give a method for making predictions on unobserved relations. We provide simulation studies demonstrating consistency of the proposed estimation algorithm and demonstrating the improvement with the proposed model over latent variable network models in estimating $\boldsymbol{\beta}$ in Section 7. We analyze a network of political books in Section 8, demonstrating the reduction in runtime when PX model, and compare its out-of-sample performance to existing latent variable network models. A discussion with an eye toward future work is provided in Section 9.

2. Latent variable network models

In this section, we briefly summarize a number of latent variable network models in the literature that are used to capture excess dependence in network observations. All latent variable network models we consider here may be written in the common form

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

$$\mathbb{P}(y_{ij}=1)=\mathbb{P}\left(\mu_{ij}+f_{\boldsymbol{\theta}}(\mathbf{v}_i,\mathbf{v}_j)+\mathbf{x}\boldsymbol{i}_{ij}>0\right), \tag{2}$$
$$\mathbf{v}_i \overset{iid}{\sim} (\mathbf{0},\boldsymbol{\Sigma}_v), \qquad \mathbf{x}\boldsymbol{i}_{ij} \overset{iid}{\sim} \mathrm{N}(0,\sigma^2),$$

where $\mathbf{v}_i \in \mathbb{R}^K$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_v$, and μ_{ij} is fixed. We avoid specifying a distribution for the latent vectors $\{\mathbf{v}_i\}_{i=1}^n$, although they are often taken to be multivariate Gaussian. We set the total variance of the latent variable representation to be $1=\sigma^2+\mathrm{var}[f_{\boldsymbol{\theta}}(\mathbf{v}_i,\mathbf{v}_j)]$, since it is not identifiable. The function of the latent variables $f_{\boldsymbol{\theta}}:\mathbb{R}^K\times\mathbb{R}^K\to\mathbb{R}$, parametrized by $\boldsymbol{\theta}$, serves to distinguish the latent variable network models discussed below. Regression latent variable network models are formed when the latent mean is represented as a linear function of exogenous covariates $\mathbf{x}_{ij}\in\mathbb{R}^p$, such that $\mu_{ij}=\mathbf{x}_{ij}^T\boldsymbol{\beta}$. The latent nodal random vectors $\{\mathbf{v}_i\}_{i=1}^n$ represent excess network dependence—beyond the mean μ_{ij}. Since relations y_{ij} and y_{ik} share latent vector $\mathbf{v}_i$ corresponding to shared actor i, and thus, y_{ij} and y_{ik} have related distributions through the latent function $f_{\boldsymbol{\theta}}(\mathbf{v}_i,\mathbf{v}_j)$. Many popular models for network data may be represented as in (2), such as the social relations model, the latent position model, and the latent eigenmodel.

2.1 Social relations model

The social relations model was first developed for continuous, directed network data (Warner et al., 1979; Wong, 1982; Snijders & Kenny, 1999). In the social relations model for binary network data (Hoff, 2005), $f_{\boldsymbol{\theta}}(\mathbf{v}_i,\mathbf{v}_j)=\mathbf{v}_i+\mathbf{v}_j$ and $\mathbf{v}_i=a_i\in\mathbb{R}$ for each actor i, such that

$$\mathbb{P}(y_{ij}=1)=\mathbb{P}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}+a_i+a_j+\mathbf{x}\boldsymbol{i}_{ij}>0\right), \tag{3}$$
$$a_i \overset{iid}{\sim}\left(0,\sigma_a^2\right), \qquad \mathbf{x}\boldsymbol{i}_{ij} \overset{iid}{\sim} \mathrm{N}(0,\sigma^2).$$

Each actor's latent variable $\{a_i\}_{i=1}^n$ may be thought of as the actor's sociability: large values of a_i correspond to actors with a higher propensity to form relations in the network. The random $\{a_i\}_{i=1}^n$ in (3) also accounts for the excess correlation in network data; any two relations that share an actor, for example, y_{ij} and y_{ik}, are marginally correlated.

2.2 Latent position model

A more complex model for representing excess dependence in social network data is the latent position model (Hoff et al., 2002). The latent position model extends the idea of the social relations model by giving each actor i a latent position $\mathbf{u}_i$ in a Euclidean latent space, for example $\mathbb{R}^K$. Then, actors whose latent positions are closer together in Euclidean distance are more likely to share a relation:

$$\mathbb{P}(y_{ij}=1)=\mathbb{P}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}+a_i+a_j-||\mathbf{u}_i-\mathbf{u}_j||_2+\mathbf{x}\boldsymbol{i}_{ij}>0\right), \tag{4}$$
$$a_i \overset{iid}{\sim}\left(0,\sigma_a^2\right), \quad \mathbf{u}_i \overset{iid}{\sim} (0,\boldsymbol{\Sigma}_u), \quad \mathbf{x}\boldsymbol{i}_{ij} \overset{iid}{\sim} \mathrm{N}(0,\sigma^2).$$

In the form of (2), the latent position model contains latent random vector $\mathbf{v}_i=[a_i,\mathbf{u}_i]^T\in\mathbb{R}^{K+1}$, and $f_{\boldsymbol{\theta}}(\mathbf{v}_i,\mathbf{v}_j)=a_i+a_j-||\mathbf{u}_i-\mathbf{u}_j||_2$. Hoff et al. (2002) show that the latent position model is capable of representing transitivity, that is, when $y_{ij}=1$ and $y_{jk}=1$, it is more likely that $y_{ik}=1$. Models that are transitive often display a pattern observed in social network data: a friend of my friend is also my friend (Wasserman & Faust, 1994).

2.3 Latent eigenmodel

The latent eigenmodel also associates each actor with a latent position $\mathbf{u}_i$ in a latent Euclidean space; however, the inner product between latent positions (weighted by symmetric parameter

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

matrix Λ) measures the propensity of actors i and j to form a relation, rather than the distance between positions (Hoff, 2008):

$$\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + a_i + a_j + \mathbf{u}_i^T\Lambda\mathbf{u}_j + \mathbf{x}i_{ij} > 0\right), \tag{5}$$
$$a_i \overset{iid}{\sim}\left(0, \sigma_a^2\right), \quad \mathbf{u}_i \overset{iid}{\sim} (0, \boldsymbol{\Sigma}_u), \quad \mathbf{x}i_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2).$$

In the context of (2), the function $f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j) = a_i + a_j + \mathbf{u}_i^T\Lambda\mathbf{u}_j$ for the latent eigenmodel, where the parameters $\boldsymbol{\theta}$ are the entries in Λ and $\mathbf{v}_i = [a_i, \mathbf{u}_i]^T \in \mathbb{R}^{K+1}$. Hoff (2008) shows that the latent eigenmodel is capable of representing transitivity and that the latent eigenmodel generalizes the latent position model given sufficiently large dimension of the latent vectors K.

In addition to transitivity, a second phenomenon observed in social networks is structural equivalence, wherein different groups of actors in the network form relations in a similar manner to others in their group. One form of structural equivalence is associative community structure, where the social network may be divided into groups of nodes that share many relations within group, but relatively few relations across groups. Such behavior is common when cliques are formed in high school social networks or around subgroups in online social networks. A form of structural equivalence is when actors in a given group are more likely to form relations with actors in other groups than with actors in their own group, for example, in networks of high-functioning brain regions when performing cognitively demanding tasks (Betzel et al., 2018). Two models that are aimed at identifying subgroups of nodes that are structurally equivalent are the latent class model of Nowicki & Snijders (2001) and the mixed membership stochastic blockmodel (Airoldi et al., 2008). Hoff (2008) shows that the latent eigenmodel is capable of representing stochastic equivalence in addition to transitivity and that the latent eigenmodel generalizes latent class models given sufficiently large dimension of the latent vectors K. For this reason, we focus on the latent eigenmodel, and the simpler social relations model, as reference models in this paper.

2.4 Drawbacks

The latent variable network models discussed in this section were developed based on the patterns often observed in real-world social networks. Latent variable network models contain different terms to represent the social phenomena underlying these patterns, and thus, different models may lead to substantially different estimates of $\boldsymbol{\beta}$. It may not be clear which model's estimate of $\boldsymbol{\beta}$, or which model's prediction of $\{y_{ij}\}_{ij}$, is best. Generally, latent variable network models are evaluated using goodness-of-fit checks (Hunter et al., 2008b), rather than rigorous tests, and it is well-known that selecting informative statistics for the goodness-of-fit checks is challenging. The latent variable network models described in this section are typically estimated using a Bayesian Markov chain Monte Carlo (MCMC) approach, which may be slow, especially for large data sets. Some recent advances do directly attempt to maximize the likelihood of network models with latent spaces (Ma et al., 2020; Zhang et al., 2022); however, public software implementations of these methods do not appear available, and they require certain covariate types (relation-level and actor-level, respectively) and certain latent space structures, such as the Euclidean distance latent space.

3. Exchangeable network models

To motivate the formulation of the proposed model, we briefly discuss the theory of exchangeable random network models and their relationship to latent variable network models. A random network model for $\{\epsilon_{ij}\}_{ij}$ is *exchangeable* if the distribution of $\{\epsilon_{ij}\}_{ij}$ is invariant to permutations of

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

the actor labels, that is, if

$$\mathbb{P}\big(\{\epsilon_{ij}\}_{ij}\big) = \mathbb{P}\big(\{\epsilon_{\pi(i)\pi(j)}\}_{ij}\big)\,, \tag{6}$$

for any permutation $\pi(.)$. There is a rich theory of exchangeable network models, dating back to work on exchangeable random matrices (Hoover, 1979; Aldous, 1981), upon which we draw in this section.

All the latent variable network models discussed in Section 2 have latent error networks $\{\epsilon_{ij}\}_{ij}$ that are exchangeable, where we define $\epsilon_{ij} = f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j) + \mathbf{x}\boldsymbol{i}_{ij}$ from (2), the random portion of a general latent variable network model. Further, under constant mean $\mu_{ij} = \mu$, all the latent variable network models for the observed network $\{y_{ij}\}_{ij}$ in Section 2 are exchangeable. In fact, any exchangeable network model may be represented by a latent variable network model. Specifically, the theory of exchangeable network models states that every exchangeable random network model may be represented in the following form (see, for example, Lovász & Szegedy, 2006; Kallenberg, 2006):

$$\begin{aligned} &\mathbb{P}(\epsilon_{ij} = 1) = \mathbb{P}\big(\mu + h(u_i, u_j) + \mathbf{x}\boldsymbol{i}_{ij} > 0\big)\,, \\ &u_i \overset{iid}{\sim} \text{Uniform}(0,1), \quad \mathbf{x}\boldsymbol{i}_{ij} \overset{iid}{\sim} \text{N}(0, \sigma^2), \end{aligned} \tag{7}$$

where the function $h : [0,1] \times [0,1] \to \mathbb{R}$ has finite integral $\int_{[0,1]\times[0,1]} h(u,v)dudv < \infty$ and serves to distinguish the various exchangeable network models. It can be shown that (7) is equivalent to the graphon representation of exchangeable random network models, where the graphon is the canonical probabilistic object of exchangeable random network models (Lovász & Szegedy, 2006; Borgs et al., 2014). Noting that we may always map the random scalar u_i to some random vector $\mathbf{v}_i$, the expression in (7) illustrates how every exchangeable random network model may be represented by a latent variable network model in the sense of (2).

3.1 Covariance matrices of exchangeable network models

The expression in (7) shows that any exchangeable network model for binary network data must correspond to a latent random network $\{\epsilon_{ij}\}_{ij}$ that is continuous and exchangeable. The covariance matrix of *any* undirected exchangeable network model has the same form and contains at most two unique nonzero values. (Marrs et al. (2017) show that directed exchangeable network models with continuous values all have covariance matrices of the same form with at most five unique nonzero terms). This fact can be seen by simply considering the ways that any pair of relations can share an actor. In addition to a variance, the remaining covariances are between relations that do and do not share an actor:

$$\text{var}[\epsilon_{ij}] = \sigma_\epsilon^2, \quad \text{cov}[\epsilon_{ij}, \epsilon_{ik}] := \rho, \quad \text{cov}[\epsilon_{ij}, \epsilon_{kl}] = 0, \tag{8}$$

where the indices i, j, k, and l are unique. It is easy to see the second equality holds for any pair of relations that share an actor by the exchangeability property, that is, by permuting the actor labels. The third equality results from the fact that the only random elements in (7) are the actor random variables u_i, u_j, and the random error $\mathbf{x}\boldsymbol{i}_{ij}$. When the random variables corresponding to two relations ϵ_{ij} and ϵ_{kl} share no actor, the pair of relations are independent by the generating process. Finally, we note that exchangeable network models have relations that are marginally identically distributed, and thus, relations therein have the same expectation and variance. That said, in the generalized linear regression case of (2), the means $\mu_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta}$ are non-constant, and thus, the observations $\{y_{ij}\}_{ij}$ are not exchangeable; only the latent error network $\{\epsilon_{ij}\}_{ij}$ is exchangeable in the generalized linear regression case. In the proposed model, rather than put forth a particular parametric model for the latent network $\{\epsilon_{ij}\}_{ij}$, we simply model the covariance structure outlined in (8), which is sufficient to represent the covariance structure of *any* exchangeable network model for the errors.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

4. The probit exchangeable (PX) model

In this section, we propose the probit exchangeable network regression model, which we abbreviate as the "PX" model. In the PX model, the vectorized mean of the network is characterized by a linear combination of covariates, $\mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a p-length vector of coefficients that are the subject of inference and $\mathbf{X}$ is a $\binom{n}{2} \times p$ matrix of covariates. The excess network dependence beyond that captured in $\mathbf{X}\boldsymbol{\beta}$ is represented by an unobservable mean-zero error vector $\boldsymbol{\epsilon}$, a vectorization of $\{\epsilon_{ij}\}_{ij}$, that is exchangeable in the sense of (6). The PX model is

$$\begin{aligned} &\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij} > 0\right), \\ &\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}), \end{aligned} \tag{9}$$

where we note that the variance of ϵ_{ij} is not identifiable, and thus, we choose $\mathrm{var}[\epsilon_{ij}] = 1$ without loss of generality. We focus on normally distributed unobserved errors $\boldsymbol{\epsilon}$ in this paper; however, other common distributions, such as the logistic distribution, could be used. We note that the normal distribution assumption implies that (9) is a typical probit regression model, but with correlation among the observations due to network structure.

As discussed in Section 3, under the exchangeability assumption, the covariance matrix of the latent error network $\mathrm{var}[\boldsymbol{\epsilon}] = \boldsymbol{\Omega}$ has at most two unique nonzero parameters. Taking $\mathrm{var}[\epsilon_{ij}] = 1$, the covariance matrix of $\boldsymbol{\epsilon}$ has a single parameter $\rho = \mathrm{cov}[\epsilon_{ij}, \epsilon_{ik}]$. We may thus write

$$\boldsymbol{\Omega}(\rho) = \mathscr{S}_1 + \rho\mathscr{S}_2, \tag{10}$$

where we define the binary matrices $\{\mathscr{S}_i\}_{i=1}^3$ indicating unique entries in $\boldsymbol{\Omega}$. The matrix $\S_1$ is a diagonal matrix indicating the locations of the variance in $\boldsymbol{\Omega}$, and $\S_2$ and $\S_3$ indicate the locations in $\boldsymbol{\Omega}$ corresponding to the covariances $\mathrm{cov}[\epsilon_{ij}, \epsilon_{ik}]$, and $\mathrm{cov}[\epsilon_{ij}, \mathrm{cov}[\epsilon_{ij}, \epsilon_{kl}]$, respectively, where the indices i, j, k, and l are unique.

The PX model unifies many of the latent variable network models discussed in Sections 2 and 3. Similar to (7), the PX model may be seen as representing the covariance structure of the latent variables $\{f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j) + \mathbf{x}i_{ij}\}_{ij}$ with $\{\epsilon_{ij}\}_{ij}$, the unobservable error network of the PX model in (9). As both networks $\{f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j) + \mathbf{x}i_{ij}\}_{ij}$ and $\{\epsilon_{ij}\}_{ij}$ are exchangeable, they have covariance matrices of the same form (see discussion in Section 3). As every exchangeable random network model may be represented by a latent variable network model, the PX model may represent the latent correlation structure of *any* exchangeable network model, yet without specifying a particular exchangeable model. Further, we now show that the PX model is equivalent to the social relations model under certain conditions.

Proposition 4.1. *Suppose that the random effects* $\{a_i\}_{i=1}^n$ *for the social relations model in* (3) *are normally distributed. Then, there exists* $\rho \in [0, 1/2]$ *such that* $\{y_{ij}\}_{ij}$ *in the PX model in* (9) *is equal in distribution to* $\{y_{ij}\}_{ij}$ *as specified by the social relations model in* (3).

Proof. As the PX and social relations models are probit regression models with the same mean structure, given by $\mathbf{X}\boldsymbol{\beta}$, it is sufficient to show that their latent covariance matrices are equivalent, that is, that $\mathrm{var}[\{a_i + a_j + \mathbf{x}i_{ij}\}_{ij}] = \mathrm{var}[\{\epsilon_{ij}\}_{ij}]$. By exchangeability, the latent covariance matrices of the PX and social relations models have the same form and by assumption have variance 1. It is easy to see that, given $\sigma_a^2 \leq 1$ (a necessary condition for $\mathrm{var}[\epsilon_{ij}] = 1$), we may take $\rho = \sigma_a^2/2$ for the PX model, which establishes equality in the model distributions. □

Exact distributional equivalence between the PX model and latent variable models other than the social relations model will typically not hold. For example, the latent eigenmodel in (5) includes non-Gaussian random variables, so that exact distributional equivalence is impossible. Similarly, it appears likely that the general latent variable model in (2) may generate non-Gaussian

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

random variables through the function $f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j)$. Importantly however, there does exist ρ such that the covariance of the latent errors of every pair of relations, $\text{cov}[\epsilon_{ij}, \epsilon_{kl}]$, is equal to the covariance of the latent errors in *any* exchangeable latent variable model, $\text{cov}[f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j) + \mathbf{x}\boldsymbol{i}_{ij}, f_{\boldsymbol{\theta}}(\mathbf{v}_k, \mathbf{v}_l) + \mathbf{x}\boldsymbol{i}_{kl}]$. Hence, the PX model may be seen as a generalized exchangeable latent variable model that focuses all modeling effort on the first two moments of the data.

Proposition 4.1 states that the PX model and social relations model are equivalent under normality of their latent error networks. In principle, the social relations model is simply a generalized linear mixed model; however, existing software packages, such as `lme4` in `R` (Bates et al., 2015), do not appear to accommodate the random effects specification of the social relations model in (3) since the indices i and j pertain to random effects a_i and a_j from the same set (as opposed to a_i and b_j in a random crossed design). Nevertheless, the estimation scheme proposed in Section 5 employs the same strategies as those commonly used to estimate generalized linear mixed models (Littell et al., 2006; Gelman & Hill, 2006). In the estimation algorithm in `lme4`, the marginal likelihood of the data is approximated and then maximized using numerical approximations with respect to $\boldsymbol{\beta}$ and random effects variance, for example σ_a^2 in the social relations model. Rather than an approximate likelihood, we propose maximizing the true likelihood with respect to $\boldsymbol{\beta}$ and ρ, yet also use numerical approximations to accomplish this maximization.

It is important to note that, although the latent errors $\{\epsilon_{ij}\}_{ij}$ in the PX model form an exchangeable random network, the random network y_{ij} represented by the PX model is almost certainly not exchangeable. For example, each y_{ij} may have a different marginal expectation $\Phi\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)$. Then, the relations in the network are not marginally identically distributed, which is a necessary condition for exchangeability. Further, the covariances between pairs of relations, say y_{ij} and y_{ik}, depend on the marginal expectations:

$$\text{cov}\left[y_{ij}, y_{ik}\right] = E\left[y_{ij}y_{ik}\right] - E\left[y_{ij}\right]E\left[y_{ik}\right] = \int_{-\mathbf{x}_{ij}^T\boldsymbol{\beta}}^{\infty}\int_{-\mathbf{x}_{ik}^T\boldsymbol{\beta}}^{\infty} dF_\rho - \Phi\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}\right)\Phi\left(\mathbf{x}_{ik}^T\boldsymbol{\beta}\right).$$

Here, dF_ρ is the bivariate standard normal distribution with correlation ρ. Since the covariance $\text{cov}[y_{ij}, y_{ik}]$ depends on the latent means $\mathbf{x}_{ij}^T\boldsymbol{\beta}$ and $\mathbf{x}_{ik}^T\boldsymbol{\beta}$, $\text{cov}[y_{ij}, y_{ik}]$ is only equal to $\text{cov}[y_{ab}, y_{ac}]$ when the latent means are equal. As a result, although the covariance matrix of the unobserved errors $\boldsymbol{\Omega}$ is of a simple form with entries $\{1, \rho, 0\}$, the covariances between elements of the vector of observed relations $\mathbf{y}$ are heterogeneous (in general) and depend on ρ in a generally more complicated way.

5. Estimation

In this section, we propose an estimator of $\{\boldsymbol{\beta}, \rho\}$ in the PX model that approximates the maximum likelihood estimator (MLE). The algorithm we propose is based on the EM algorithm (Dempster et al., 1977). Although the covariance matrix for the PX model is highly structured, as in (10), a closed-form expression for the MLE does not appear available. While we explored pseduo-likelihood pairwise approximations (also called "composite likelihoods" in some literature) to the complete PX likelihood (Heagerty & Lele, 1998), we found no substantial advantage—neither in performance nor runtime—over the proposed estimation scheme in this paper.

The proposed estimation algorithm consists of alternating computation of the expected complete likelihood with maximization with respect to ρ and $\boldsymbol{\beta}$, iterating until convergence. Since the algorithm iterates expectation and two maximization steps, we term it the EMM algorithm. To improve algorithm efficiency, we initialize $\boldsymbol{\beta}$ at the ordinary probit regression estimator (assuming independence of the latent errors), and initialize ρ with a mixture estimator based on possible values of ρ such that $\boldsymbol{\Omega}$ is positive definite, as detailed in Section A.1. The complete EMM algorithm is presented in Algorithm 1. In the following text, we detail the EMM algorithm, beginning

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

Algorithm 1. EMM estimation of the PX model

0. **Initialization:**
Initialize $\widehat{\boldsymbol{\beta}}^{(0)}$ using probit regression assuming independence and initialize $\widehat{\rho}^{(0)}$ as described in Section A.1. Set positive convergence threshold τ, scaling $\delta \in [0, 1]$ and set iteration $\nu = 0$.
1. **Expectation step:**
Given $\widehat{\rho}^{(\nu)}$ and $\widehat{\boldsymbol{\beta}}^{(\nu)}$, compute $E\left[\boldsymbol{\epsilon} \mid \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}\right]$ using the procedure described in Section A.2, and approximate $\{\gamma_i\}_{i=1}^3$ as described in Section A.3.
2. **Maximization with respect to**
ρ: Given $s = 0$ and $\widehat{\rho}^{(\nu,s)} = \widehat{\rho}^{(\nu)}$, $\widehat{\boldsymbol{\beta}}^{(\nu)}$, and $\{\gamma_i\}_{i=1}^3$, compute $\widehat{\rho}^{(\nu,s+1)}$ by alternating (12) and (13) until ρ changes by less than $\delta\tau$. Set $\rho^{(\nu+1)}$ equal to the final ρ value.
3. **Maximization with respect to $\boldsymbol{\beta}$**:
Compute the updated estimate
$$\widehat{\boldsymbol{\beta}}^{(\nu+1)} = \widehat{\boldsymbol{\beta}}^{(\nu)} + \left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}E\left[\boldsymbol{\epsilon} \mid \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}\right].$$
4. If $\max\left\{\left|\widehat{\boldsymbol{\beta}}^{(\nu+1)} - \widehat{\boldsymbol{\beta}}^{(\nu)}\right| / \widehat{\boldsymbol{\beta}}^{(\nu)}, \left|\widehat{\rho}^{(\nu+1)} - \widehat{\rho}^{(\nu)}\right| / \widehat{\rho}^{(\nu)}\right\} > \tau$, then increment ν by 1 and return to Step 1. Otherwise, end.

with maximization with respect to ρ, and then proceeding to maximization with respect to $\boldsymbol{\beta}$. We define $\gamma_i = E\left[\boldsymbol{\epsilon}^T\mathscr{S}_i\boldsymbol{\epsilon} \mid \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}\right] / |\Theta_i|$, where Θ_i is the set of relation pairs indicated by binary matrices $\mathscr{S}_i$. By default, we typically set $\tau = 10^{-2}$ and $\delta = 10^{-1}$.

5.1 Expectation

Consider the log-likelihood, $\ell_{\mathbf{z}}$, of the latent continuous random vector $\mathbf{z}$. Taking the expectation of $\ell_{\mathbf{z}}$ conditional on $\mathbf{y}$, the expectation step for a given iteration ν of the EM algorithm is

$$\begin{aligned} &E\left[\ell_{\mathbf{z}} \mid \mathbf{y}, \rho = \widehat{\rho}^{(\nu)}, \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(\nu)}\right] \\ &\quad = -\frac{1}{2}\log 2\pi|\boldsymbol{\Omega}| - \frac{1}{2}E\left[(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) \mid \mathbf{y}, \rho = \widehat{\rho}^{(\nu)}, \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(\nu)}\right], \end{aligned} \tag{11}$$

where $\widehat{\rho}^{(\nu)}$ and $\widehat{\boldsymbol{\beta}}^{(\nu)}$ are the estimators of ρ and $\boldsymbol{\beta}$ at iteration ν. In discussing the maximization step for ρ, we will show that that the ρ update depends on the data through the expectations denoted by γ_i for $i \in \{1, 2, 3\}$. In discussing the maximization step for β, we will show that that the β update depends on the data only through the expectation $E\left[\boldsymbol{\epsilon} \mid \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}\right]$.

5.1.1 Approximations

The computation of $E\left[\boldsymbol{\epsilon} \mid \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}\right]$ in (14) is nontrivial, as it is a $\binom{n}{2}$-dimensional truncated multivariate normal integral. We exploit the structure of $\boldsymbol{\Omega}$ to compute $E\left[\boldsymbol{\epsilon} \mid \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}\right]$ using the law of total expectation. A Newton-Raphson algorithm, along with an approximate matrix inverse, is employed to compute an approximation of $E\left[\boldsymbol{\epsilon} \mid \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}\right]$. Details of the implementation of the approximations are given in Section A.2.

The expectations $\{\gamma_i\}_{i=1}^3$ require the computation of $\binom{n}{2}$-dimensional truncated multivariate normal integrals, which are onerous for even small networks. Thus, we make two approximations to $\{\gamma_i\}_{i=1}^3$ to reduce the runtime of the EMM algorithm. First, we compute the expectations conditioning only on the entries in $\mathbf{y}$ that correspond to the entries in $\boldsymbol{\epsilon}$ being integrated, for example,

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

instead of computing $E[\epsilon_{jk}\epsilon_{lm} \mid \mathbf{y}]$, we compute $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$. This first approximation is most appropriate when ρ is small, since y_{lm} is maximally informative for ϵ_{jk} when ρ is large (for l, m, j, and k distinct). Second, we find empirically that $\gamma_2 = E\left[\boldsymbol{\epsilon}^T \mathscr{S}_2 \boldsymbol{\epsilon} \mid \mathbf{y}\right] / |\Theta_2|$ is approximately linear in ρ, since this sample mean of conditional expectations concentrates around a linear function of ρ. Thus, we compute γ_2 for $\rho = 0$ and $\rho = 1$ and use a line connecting these two values to compute γ_2 for arbitrary values of ρ (see evidence of linearity of γ_2 for the political books network in Appendix E). The details of the approximations to $\{\gamma_i\}_{i=1}^3$ are given in Section A.3.

5.2 Maximization with respect to ρ

To derive the maximization step for ρ, we use the method of Lagrange multipliers, since differentiating (11) directly with respect to ρ gives complex nonlinear equations that are not easily solvable. We first define the set of parameters $\{\phi_i\}_{i=1}^3$, representing the variance and two possible covariances in $\boldsymbol{\Omega}$,

$$\operatorname{var}\left[\epsilon_{ij}\right] = \phi_1, \quad \operatorname{cov}\left[\epsilon_{ij}, \epsilon_{ik}\right] = \phi_2 = \rho, \quad \operatorname{cov}\left[\epsilon_{ij}, \epsilon_{kl}\right] = \phi_3,$$

where the indices i, j, k, and l are distinct. In addition, we let $\mathbf{p} = [p_1, p_2, p_3]$ parametrize the precision matrix $\boldsymbol{\Omega}^{-1} = \sum_{i=1}^3 p_i \mathscr{S}_i$, which has the same form as the covariance matrix $\boldsymbol{\Omega}$ (see Marrs et al. (2017) for a similar result when $\{\epsilon_{ij}\}_{ij}$ forms a directed network). The objective function, incorporating the restrictions that $\phi_1 = 1$ and $\phi_3 = 0$, is

$$Q_{\mathbf{y}}(\boldsymbol{\phi}) := E[\ell_{\mathbf{z}} \mid \mathbf{y}] + \frac{1}{2}\lambda_1(\phi_1 - 1) + \frac{1}{2}\lambda_3\phi_3,$$

where $\boldsymbol{\phi} = [\phi_1, \phi_2, \phi_3]$ and the "$\frac{1}{2}$" factors are included to simplify algebra. Then, differentiating $Q_{\mathbf{y}}$ with respect to $\mathbf{p}$, λ_1, and λ_3, the estimators for ρ, $\{\lambda_1, \lambda_3\}$ are

$$\widehat{\rho} = \gamma_2 - \frac{1}{|\Theta_2|}\begin{bmatrix} \frac{\partial\phi_1}{\partial p_2} & \frac{\partial\phi_3}{\partial p_2} \end{bmatrix}^T \begin{bmatrix} \lambda_1 \\ \lambda_3 \end{bmatrix} \tag{12}$$

$$\begin{bmatrix} \widehat{\lambda}_1 \\ \widehat{\lambda}_3 \end{bmatrix} = \begin{bmatrix} \frac{\partial\phi_1}{\partial p_1} & \frac{\partial\phi_3}{\partial p_1} \\ \frac{\partial\phi_1}{\partial p_3} & \frac{\partial\phi_3}{\partial p_3} \end{bmatrix}^{-1} \begin{bmatrix} |\Theta_1| & 0 \\ 0 & |\Theta_3| \end{bmatrix} \begin{bmatrix} \gamma_1 - 1 \\ \gamma_3 \end{bmatrix}, \tag{13}$$

where again Θ_i is the set of pairs of relations (jk, lm) that share an actor in the i^{th} manner, for $i \in \{1, 2, 3\}$. For instance, Θ_2 consists of pairs of relations of the form (jk, jl), where j, k, and l are distinct indices. In (12) and (13), the partial derivatives $\left\{\partial\phi_i/\partial p_j\right\}$ are available in closed form and are easily computable in $O(1)$ time using the forms of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^{-1}$. See Appendix B for details.

Alternation of the estimators for ρ and $\{\lambda_1, \lambda_3\}$ in (12) and (13) constitutes a block coordinate descent for $\rho = \phi_2$ subject to the constraints $\phi_1 = 1$ and $\phi_3 = 0$. This block coordinate descent makes up the maximization step of the EMM algorithm for ρ.

5.3 Maximization with respect to $\boldsymbol{\beta}$

The maximization step with respect to $\boldsymbol{\beta}$ in the EMM algorithm can be obtained directly. Setting the derivative of (11) with respect to $\boldsymbol{\beta}$ equal to zero, the maximization step for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}^{(\nu+1)} = \widehat{\boldsymbol{\beta}}^{(\nu)} + \left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}E\left[\boldsymbol{\epsilon} \mid \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}\right], \tag{14}$$

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

where we use the identity $\boldsymbol{\epsilon} = \mathbf{z} - \mathbf{X}\boldsymbol{\beta}$. In Appendix B, we show that the leading terms of the unique entries in Ω^{-1}, $\mathbf{p}$, depend only on ρ through a multiplicative factor,

$$\mathbf{p} \approx f(\rho)[g_1(n),\ g_2(n),\ g_3(n)]^T.$$

Thus, we may factor $f(\rho)$ out of (14), and the $\boldsymbol{\beta}$ maximization is asymptotically ρ-free (except for the expectation term). Similarly, the maximization with respect to ρ in Section 5.2 is free from $\boldsymbol{\beta}$ except for the expectation term. Hence, only a single maximization step with respect to each $\boldsymbol{\beta}$ and ρ is required for each expectation.

5.4 Consistency of the EMM estimator

The complete multivariate normal likelihood for $\mathbf{z}$ is a non-curved, identifiable likelihood. Then, it is known that each expectation and maximization step in an EM algorithm increases the current likelihood value (Wu, 1983). Whenever there is a unique, single local maximum, the EM algorithm yields consistent, and efficient, estimators. We make a series of approximations to the expectations in the EMM algorithm to reduce computational demands, so that the theory of EM estimator convergence may not be directly applicable. Yet, we find that the EMM estimators, $\widehat{\beta}_{\text{EMM}}, \widehat{\rho}_{\text{EMM}}\}$, maintain consistency. Taking the leading terms of $\widehat{\rho}_{\text{EMM}}$,

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

which has expectation $E[\widehat{\rho}_{\text{EMM}}] = \rho + O(n^{-1})$. Then, consistency can be established by showing that the variance of $\widehat{\rho}_{\text{EMM}}$ tends to zero. We provide details in Appendix C. The estimator $\widehat{\beta}_{\text{EMM}}$ is particularly difficult to analyze, as $E[\epsilon_{jk} \mid \mathbf{y}]$ depends on every entry in $\mathbf{y}$, and because we approximate this expectation. We provide a sketch for a proof of consistency in Appendix C by bounding the distance between the EMM estimator for β and the true MLE, $\left\|\widehat{\beta}_{\text{EMM}} - \widehat{\beta}_{\text{MLE}}\right\|_2^2$, using an easier-to-analyze estimator for β which replaces $E[\epsilon_{jk} \mid \mathbf{y}]$ with $E[\epsilon_{jk} \mid y_{jk}]$. As in the argument for consistency of $\widehat{\rho}_{\text{EMM}}$, we establish consistency of the bounding estimator by showing the expectation is asymptotically equal to the true value of β and that the variance of the bounding estimator tends to zero. We also discuss performance of $\widehat{\beta}_{\text{EMM}}$ under model misspecification, showing that it maintains consistency even under violation of the normality and exchangeability assumptions.

6. Prediction

In this section, we describe how to use the PX model, and the approximations in service of Algorithm 1, to make predictions for an unobserved network relation without undue computational cost. The predicted value we seek is the probability of observing $y_{jk} = 1$ given all the other values $\mathbf{y}_{-jk}$, where $\mathbf{y}_{-jk}$ is the vector of observations $\mathbf{y}$ excluding the single relation jk. As in estimation, the desired probability is again equal to a $\binom{n}{2}$-dimensional multivariate truncated normal integral, which is computationally burdensome. Thus, we approximate the desired prediction probability

$$\mathbb{P}(y_{jk} = 1 \mid \mathbf{y}_{-jk}) = E\left[E\left[\mathbf{1}\left[\epsilon_{jk} > -\mathbf{x}_{jk}^T\boldsymbol{\beta}\right] \mid \boldsymbol{\epsilon}_{-jk}\right] \mid \mathbf{y}_{-jk}\right], \tag{15}$$
$$\approx \Phi\left(\frac{E[\epsilon_{jk} \mid \mathbf{y}] + \mathbf{x}_{jk}^T\boldsymbol{\beta}}{\sigma_n}\right).$$

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

The approximation in (15) is based on the fact that $[\epsilon_{jk} \mid \boldsymbol{\epsilon}_{-jk}]$ is normally distributed:

$$\begin{aligned} \epsilon_{jk} \mid \boldsymbol{\epsilon}_{-jk} &\sim \mathrm{N}\left(m_{jk}, \sigma_n^2\right), \\ m_{jk} &= -\sigma_n^2 \mathbf{1}_{jk}^T \left(p_2 \mathscr{S}_2 + p_3 \mathscr{S}_3\right) \widetilde{\boldsymbol{\epsilon}}_{-jk}, \quad \sigma_n^2 = \frac{1}{p_1}, \end{aligned} \tag{16}$$

where $\mathbf{1}_{jk}$ is the vector of all zeros with a one in the position corresponding to relation jk and, for notational simplicity, we define $\widetilde{\boldsymbol{\epsilon}}_{-jk}$ is the vector $\boldsymbol{\epsilon}$ with a zero in the entry 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 m_{jk} is free of ϵ_{jk}. Then, the inner expectation in (15) is

$$E\Big[\mathbf{1}[\epsilon_{jk} > -\mathbf{x}_{jk}^T \boldsymbol{\beta}] \mid \boldsymbol{\epsilon}_{-jk}\Big] = \Phi\left(\frac{m_{jk} + \mathbf{x}_{jk}^T \boldsymbol{\beta}}{\sigma_n}\right). \tag{17}$$

Of course, m_{jk} depends on $\boldsymbol{\epsilon}_{-jk}$ which is unknown, and thus, we replace m_{jk} with its conditional expectation $E[m_{jk} \mid \mathbf{y}_{-jk}] = E[\epsilon_{jk} \mid \mathbf{y}_{-jk}]$.

Computing $E[\epsilon_{jk} \mid \mathbf{y}_{-jk}]$ is extremely difficult; however, computing $E[\epsilon_{jk} \mid \mathbf{y}]$ proves feasible if we exploit the structure of $\boldsymbol{\Omega}$. Thus, we approximate the desired expectation by imputing y_{jk} with the mode of the observed data:

$$E\Big[\epsilon_{jk} \mid \mathbf{y}_{-jk}\Big] \approx E\Big[\epsilon_{jk} \mid \mathbf{y}_{-jk}, y_{jk} = y^*\Big] = E\big[\epsilon_{jk} \mid \mathbf{y}\big], \tag{18}$$

where y^* is the mode of $\mathbf{y}_{-jk}$. The error due to this approximation is small and shrinks as n grows. Substituting (18) for m_{jk} in (17) gives the final expression in (15).

7. Simulation studies

In this section, we describe three simulation studies. The first verifies that the performance of the EMM estimator in Algorithm 1 provides improvement over standard probit regression. The second simulation study verifies consistency of the EMM estimators of $\boldsymbol{\beta}$, and compares the performance of these estimators to the estimators of $\boldsymbol{\beta}$ from the social relations model and the latent eigenmodel. The third simulation study evaluates the robustness of the PX model, and EMM algorithm, to the assumption that the latent random variables are normally distributed.

For both simulation studies, we generated data with mean consisting of three covariates and an intercept:

$$y_{ij} = \mathbf{1}\Big[\beta_0 + \beta_1 \mathbf{1}[x_{1i} \in C]\mathbf{1}[x_{1j} \in C] + \beta_2 |x_{2i} - x_{2j}| + \beta_3 x_{3ij} + \epsilon_{ij} > 0\Big]. \tag{19}$$

In the model in (19), β_0 is an intercept; β_1 is a coefficient on a binary indicator of whether individuals i and j both belong to a pre-specified class C; β_2 is a coefficient on the absolute difference of a continuous, actor-specific covariate x_{2i}; and β_3 is that for a pair-specific continuous covariate x_{3ij}. We fixed $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]^T$ at a single set of values. Since the accuracy of estimators of $\boldsymbol{\beta}$ may depend on $\mathbf{X}$, we generated 20 random design matrices $\mathbf{X}$ for each sample size of $n \in \{20, 40, 80\}$ actors. We emphasize that, although these may appear to be only moderately sized networks, each consists of $\binom{n}{2} \in \{190, 780, 3160\}$ observations. For each design matrix, we simulated 100 error realizations of $\{\epsilon_{ij}\}_{ij}$, with distribution that depended on the generating model. When generating from the PX model, half of the total variance in ϵ_{ij} was due to correlation $\rho = 1/4$, and the remaining half was due to the unit variance of ϵ_{ij}. When generating from the latent eigenmodel in (5), one-third of the variance in ϵ_{ij} was due to each term $a_i + a_j$, $\mathbf{u}_i^T \Lambda \mathbf{u}_j$, and $\mathbf{x}i_{ij}$, respectively. For additional details of the simulation study procedures, see Section D.1.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

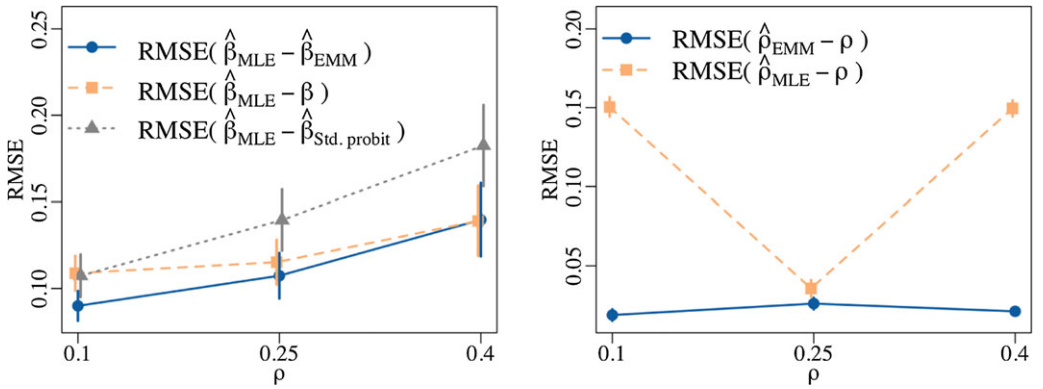

Figure 1. The left panel depicts performance in estimating β: RMSE between the EMM estimator and the MLE (RMSE($\widehat{\beta}_{\text{MLE}} - \widehat{\beta}_{\text{EMM}}$)), between the MLE and the truth (RMSE($\widehat{\beta}_{\text{MLE}} - \beta$)), and between the MLE and the standard probit estimator (RMSE($\widehat{\beta}_{\text{MLE}} - \widehat{\beta}_{\text{Std. probit}}$)). The right panel depicts performance in estimating ρ: RMSE between the MLE and the EMM estimator (RMSE($\widehat{\rho}_{\text{MLE}} - \widehat{\rho}_{\text{EMM}}$)) and between the MLE and the truth (RMSE($\widehat{\rho}_{\text{MLE}} - \rho$)). The RMSEs are plotted as a function of the true values of ρ, and solid vertical lines denote Monte Carlo error bars. Some points obscure their Monte Carlo error bars.

7.1 Evaluation of approximations in Algorithm 1

To evaluate the efficacy of the approximations described in the estimation procedure in Algorithm 1, we simulated from (19) for a single $\mathbf{X}$ with $n = 40$ (larger n caused multivariate normal integral failures in `R`). We simulated 100 networks from the PX model in (9) using this $\mathbf{X}$, for each value of $\rho \in \{0.1, 0.25, 0.4\}$ (we note that we require $\rho < 1/2$ for the error covariance matrix Ω to be positive definite). For each realization, we estimated β in the PX model using EMM in Algorithm 1. To estimate β in the standard probit model, we used the function `glm` in `R`. To compute the MLE, we numerically optimized the data log-likelihood using the Broyden-Fletcher-Goldfarb-Shanno algorithm as implemented in the `optim` function in `R`, initializing at the true values of $\{\beta, \rho\}$.

In the left panel of Figure 1, we evaluate the performance of the EMM estimator by comparing the root mean square error (RMSE) between the EMM coefficient estimate, $\widehat{\beta}_{\text{EMM}}$, and the MLE obtained by the optimization procedure $\widehat{\beta}_{\text{MLE}}$. As a baseline, we compute the RMSE between $\widehat{\beta}_{\text{MLE}}$ and the true value β. If the approximations in the EMM algorithm are small, we expect the RMSE between $\widehat{\beta}_{\text{EMM}}$ and $\widehat{\beta}_{\text{MLE}}$ to be much smaller than the RMSE between $\widehat{\beta}_{\text{MLE}}$ and β. Generally, the RMSE between $\widehat{\beta}_{\text{EMM}}$ and $\widehat{\beta}_{\text{MLE}}$ is smaller than the RMSE between $\widehat{\beta}_{\text{MLE}}$ and β. However, the discrepancy between the two RMSEs decreases as the true ρ grows. As a reference, the MSE between $\widehat{\beta}_{\text{Std. probit}}$ and $\widehat{\beta}_{\text{MLE}}$ is also shown in the left panel of Figure 1; the EMM estimator is closer to $\widehat{\beta}_{\text{MLE}}$ than the standard probit estimator is to $\widehat{\beta}_{\text{MLE}}$ for all values of ρ. Raw RMSE values between the estimators and the truth, shown in Figure 2, confirm that the EMM algorithm does perform better than standard probit in RMSE with respect to estimation of β. The results of this simulation study suggest that the EMM algorithm improves estimation of β over the standard probit estimator for $\rho > 0$ and that the EMM estimator is reasonably close to the MLE, signifying the approximations in the EMM algorithm are reasonable.

In the right panel of Figure 1, the EMM estimator of ρ is closer to the MLE, $\widehat{\rho}_{\text{MLE}}$, than the MLE is close to the true value of ρ for all values of ρ examined. This fact suggests that the approximation error in estimating ρ in the EMM algorithm is small. Further, the raw RMSE values shown in Figure 2 illustrate that $\widehat{\rho}_{\text{EMM}}$ may be as good an estimator of ρ as is $\widehat{\rho}_{\text{MLE}}$. The approximations in $\widehat{\rho}_{\text{EMM}}$ appear to be stable over the range of ρ values examined. Overall, since the degradation in performance of the EMM algorithm is most pronounced in estimation of β, we postulate that the degradation may be due to the approximations in computing $E[\epsilon_{jk} \mid \mathbf{y}]$ (see Section A.2).

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

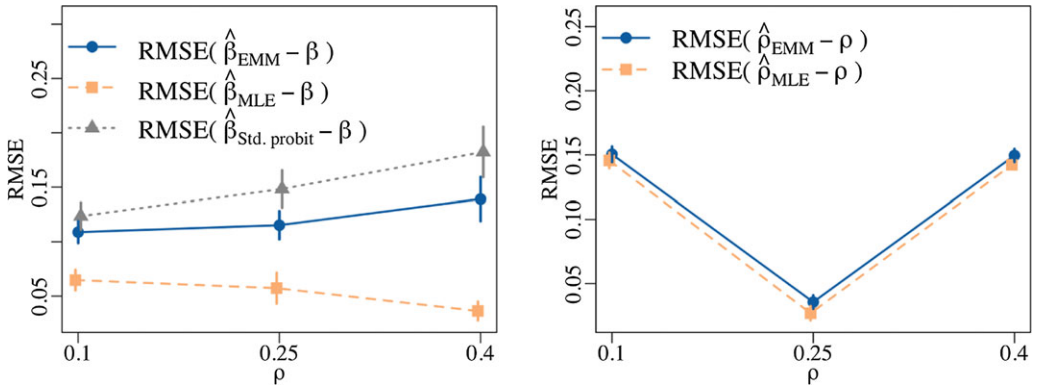

Figure 2. The left panel depicts the RMSE in estimating β using the EMM algorithm, MLE, and standard probit regression. The right panel depicts the same for ρ. The MSEs are plotted as a function of the true values of ρ, and solid vertical lines denote Monte Carlo error bars.

7.2 Performance in estimation of β

To evaluate the performance of the PX estimator in estimating linear coefficients $\boldsymbol{\beta}$, we compared estimates of $\boldsymbol{\beta}$ by the EMM algorithm to estimators of the social relations and latent eigenmodels on data generated from the PX model and data generated from the latent eigenmodel. We used the `amen` package in `R` to estimate the social relations model and latent eigenmodel (Hoff et al., 2017). We again compared these estimators to the standard probit regression model assuming independence as a baseline, which we estimated using the function `glm` in `R`. We focused on the value of $\rho = 0.25$, in the center of the range of possible ρ values.

In Figure 3, we plot the RMSE (scaled by $n^{1/2}$) of the β coefficients estimated for the PX model, standard probit model, social relations model, and latent eigenmodels. We see that the EMM estimator for the PX model has a downward trend in $n^{1/2}$RMSE with n, and a reducing spread of $n^{1/2}$RMSE with n, for both the PX and latent eigenmodel generating models. These facts suggest that the PX estimator is consistent for $\boldsymbol{\beta}$, at a rate $n^{1/2}$ or better, for both the PX and latent eigenmodel generating models, confirming the claims in Section 5.4. Further, the EMM estimator has the lowest median $n^{1/2}$RMSE of any of the estimators for all entries in $\boldsymbol{\beta}$, where $n^{1/2}$RMSE is evaluated for each $\mathbf{X}$ realization (across the error realizations), and the median is computed across the 20 $\mathbf{X}$ realizations. We observe similar patterns for the correlation parameter ρ; see Section D.1. Interestingly, the superiority of the PX estimator holds whether we generate from the PX or latent eigenmodel, which suggests that any benefit in correctly specifying the latent eigenmodel is lost in the estimating routine. The larger $n^{1/2}$RMSEs of the `amen` estimator of the social relations and latent eigenmodels are a result of bias; see Section D.1 for bias-variance decomposition of the MSEs.

7.3 Runtimes

We evaluated the average runtimes of the algorithms used to estimate the simulated data. The average runtimes are plotted in Figure 4. The improvement in runtime offered by the EMM estimation scheme over SRM and LE MCMC estimation is several orders of magnitude. Interestingly, the runtime cost of EMM appears to grow faster than the MCMC routines and faster than standard probit regression. A contributing factor is the sum over $O(n^3)$ terms in the maximization of ρ in the EMM algorithm. We have experimented with using only a random subset of $O(n^2)$ relation pairs in the maximization step, which results in gains in runtime with small cost in estimation performance. Such a tradeoff may become attractive for networks of sufficient size n.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

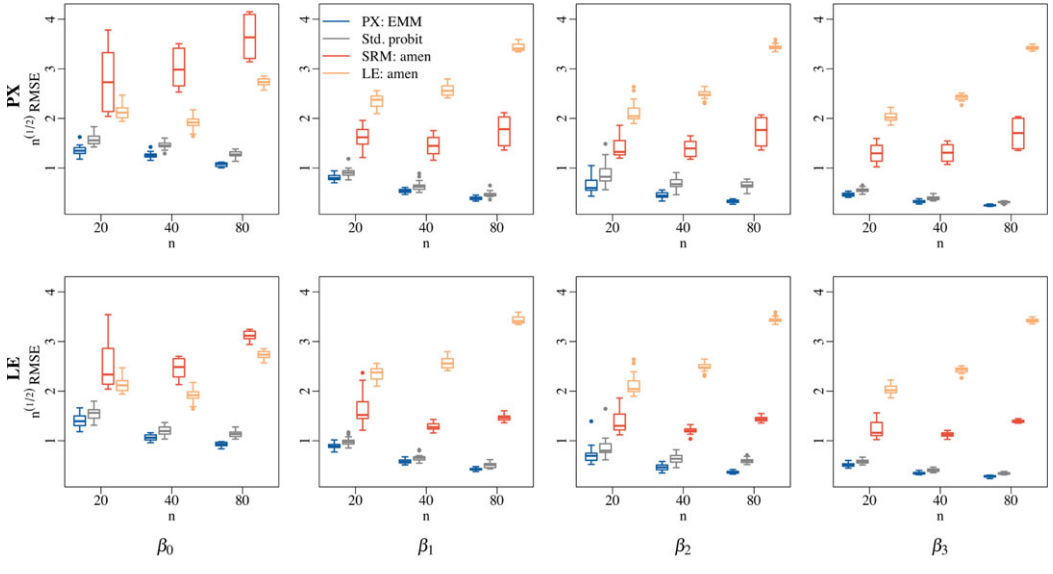

Figure 3. Performance ($n^{1/2}$ RMSE) of estimators of β, for a given **X**, when generating from the PX model (top row) and the latent eigenmodel (LE; bottom row). Variability captured by the boxplots reflects variation in RMSE with **X**.

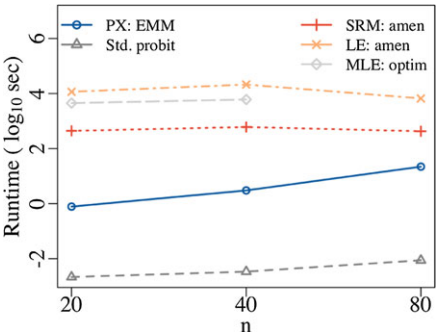

Figure 4. Average runtimes of various algorithms used on simulated data.

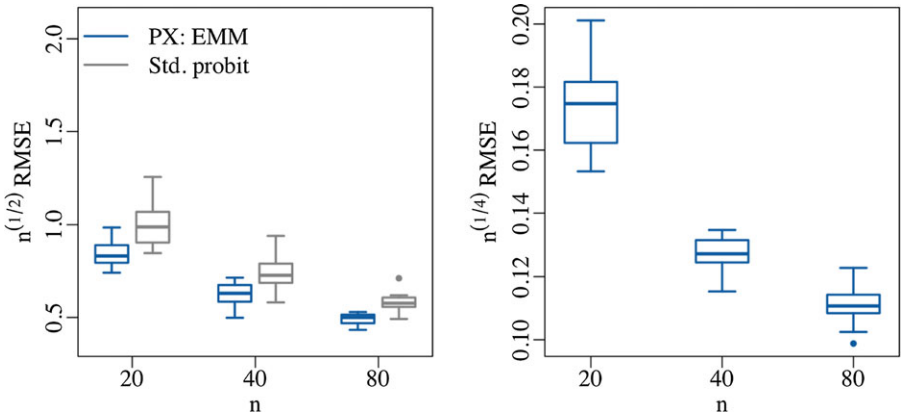

Figure 5. The left panel depicts $n^{1/2}$RMSE in estimating β_1, using the EMM algorithm and standard probit regression, under t distribution of the errors. The right panel depicts $n^{1/4}$RMSE in estimating ρ using the EMM algorithm in the same simulation. Variability captured by the boxplots reflects variation in RMSE with **X**.

7.4 Evaluation of latent normality assumption

To evaluate the performance of the PX model under violation of the normality assumption on the latent errors $\{\epsilon_{ij}\}_{ij}$, we repeated the simulation study with t-distributed latent random variables. Specifically, we simulated from (19), but replaced the latent error vector $\boldsymbol{\epsilon}$ with $\sigma^{-1}\,\Omega^{1/2}\mathbf{u}$, where $\mathbf{u}$ consists of independently and identically distributed t random variables with 5 degrees of freedom. The scaling factor $\sigma = \sqrt{5/3}$ ensures that $\mathbf{u}$ has unit population variance, for consistency with the Gaussian case, and $\Omega^{1/2}$ is the matrix square root of Ω, with $\rho = 0.25$. This model thus has the same latent mean and covariance matrix as in the Gaussian case, but the latent errors have substantially heavier tails.

The left panel of Figure 5 shows the performance of the EMM algorithm in estimating $\boldsymbol{\beta}_1$, compared to the standard probit regression estimates. As in the Gaussian case, the EMM algorithm produces estimates with $n^{1/2}$RMSE tending to zero as n grows. Also as in the Gaussian case, EMM estimation of the PX model improves estimation of $\boldsymbol{\beta}$ over standard probit regression. We observed the same results in estimation of the remaining coefficients (see Section D.2). Unlike the Gaussian case, $n^{1/2}$RMSE in estimating ρ did not appear to tend towards zero. However, in the right panel of Figure 5, the error in estimating ρ scaled by $n^{1/4}$, $n^{1/4}$ RMSE, does tend towards zero. This study confirms the claim in Section 5.4 that the EMM algorithm produces consistent estimators $\widehat{\boldsymbol{\beta}}, \widehat{\rho}\}$, even under violation of the normality assumption of the PX model.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

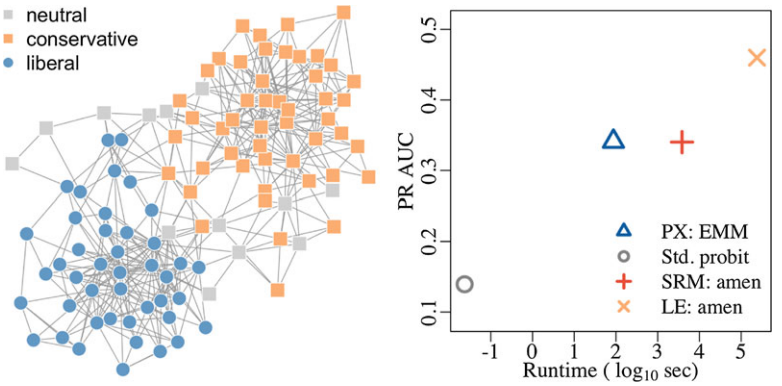

Figure 6. Krebs' political books network (left) and out-of-sample performance in 10-fold cross-validation, as measured by area under the precision-recall curve (PRAUC, right), plotted against mean runtime in the cross-validation. The estimators are standard probit assuming independent observations (Std. probit), the PX model as estimated by the EMM algorithm (PX), the social relations model estimator (SRM), and the latent eigenmodel estimator (LE).

8. Analysis of a network of political books

We live in a time of political polarization. We investigate this phenomenon by analyzing a network of $n = 105$ books on American politics published around the time of the 2004 presidential election.[2] These data were compiled by Dr. Valdis Krebs using the "customers who bought this book also bought these books" list on Amazon.com. At the time, when browsing a particular book, Amazon listed the books that were bought by individuals who also bought the book in question. Thus, a relation between two books in the network indicates that they were frequently purchased by the same buyer on Amazon. Political books on the best-seller list of The New York Times were used as actors in the network. Finally, the books were labeled as conservative, liberal, or neutral based on each book's description (Figure 6). Work by Dr. Krebs on a similar network was in a 2004 *New York Times* article (Eakin, 2004), where it was shown that there were many relations between books with similar ideologies yet relatively few across ideologies. The work by Dr. Krebs has inspired similar analyses of book purchasing networks in the fields of nanotechnology (Schummer, 2005) and climate science (Shi et al., 2017).

To confirm previous work by Dr. Krebs, we develop a model that assigns a different probability of edge formation between books i and j depending on whether the books are ideologically aligned. By examining the network in Figure 6, we observe that neutral books appear to have ties than books that are labeled conservative or liberal. Thus, we add a nodal effect indicating whether either book in a relation is labeled neutral. The regression model specified is

$$\begin{aligned} \mathbb{P}(y_{ij} = 1) = \mathbb{P}(&\beta_0 + \beta_1 \mathbf{1}[c(i) = c(j)] \\ &+ \beta_2 \mathbf{1}\left[\{c(i) = \text{neutral}\} \cup \left\{c(j) = \text{neutral}\right\}\right] + \boldsymbol{\epsilon}_{ij} > 0), \\ \boldsymbol{\epsilon} \sim (&\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{20}$$

where $c(i)$ represents the class of book i (neutral, conservative, or liberal), and the distribution and covariance matrix of $\boldsymbol{\epsilon}$ are determined by the particular model being estimated. In this section, we estimate the PX model (PX), the equivalent social relations model (SRM), the latent eigenmodel (LE), and, as a baseline, the standard probit regression model assuming independence of observations (which we label "std. probit").

We used a 10-fold cross-validation to compare the out-of-sample predictive performance of the estimators and the runtimes of the algorithms for the models in question. We used the proposed EMM algorithm to estimate the PX model, the `amen` package in `R` to estimate the social relations model and latent eigenmodel (Hoff et al., 2017), and the `glm(.)` command in the `R` package `stats`

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

Table 1. Results of fitting the Krebs political books data using the EMM estimator for the PX model and the `amen` estimator for the social relations and latent eigenmodels (SRM and LE, respectively). Point estimates for the coefficients are given to the left of the vertical bar, and runtimes (in seconds) and minimum effective sample sizes across the coefficient estimates are given to the right

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	Runtime (s)	Min(*ESS*)
PX: EMM	−1.87	1.21	1.12	68	–
SRM: amen	−2.70	0.98	1.55	7984	195
LE: amen	−3.90	1.63	2.06	62565	26

to estimate the standard probit model. We randomly divided the $\binom{105}{2}$ relations into 10 disjoint sets, termed "folds," of roughly the same size. Then, for each fold, we estimated the models on the remaining nine folds and made predictions for the data in the fold that was not used for estimation (for details of estimation of the PX model with missing data, see Section A.4). Repeating this operation for each complete data set of out-of-sample predictions for each estimating model. The procedure to make marginal predictions from the PX model is described in Section 6. To compare with the PX model, we make marginal predictions from the social relations model and the latent eigenmodel, that is, by integrating over the random effect space. The predictions from the social relations model and the latent eigenmodel are automatically output from `amen` in the presence of missing data. The predictions from the standard probit model are marginal by default as there is no correlation structure.

We use area under the precision-recall curve (PRAUC) to measure performance of the predictions relative to the observed data, although using area under the receiver operating characteristic (ROC) yields the same conclusions (see Appendix E). In Figure 6, the proposed EMM estimator produces an improvement in PRAUC over standard probit prediction that is roughly equivalent to the improvement of the social relations model over standard probit, yet with an average runtime that is 45 times faster (about a minute compared with an hour). The latent eigenmodel produces an improvement in PRAUC over the proposed EMM algorithm and the social relations model, however, at the expense of significant increase in average runtime, that of about 3,000 times slower than EMM and taking almost three days to complete. Note that we selected the number of MCMC iterations for the social relations and latent eigenmodels that resulted in sets of samples from the posterior distributions (after burn-in) that had a effective sample sizes roughly equal to 100 independent samples of the $\boldsymbol{\beta}$ parameters. Increasing the number of iterations, which may be desirable, would result in even longer runtimes for the estimators of the social relations and latent eigenmodels. Taken together, the results of the cross-validation study suggest that the PX model accounts for a large portion of the correlation in network data with estimation runtime that, depending upon stopping criterion, is orders of magnitude faster the runtime than existing approaches.

To estimate the complete data set under the mean model in (20), we used the EMM algorithm for the PX model and the `amen` package for the social relations model (SRM) and latent eigenmodel (LE), which we ran for 1×10^6 iterations after a burn-in of 5×10^4 iterations (with runtimes of roughly two hours for SRM and 17 hours for LE). The coefficient estimates in Table 1 suggest that books that share the same ideology are more likely to be frequently purchased together, as all $\widehat{\beta}_1 > 0$. This positive coefficient estimate demonstrates political polarization in the network: conservative books are more likely to be purchased with other conservative books rather than with liberal books. The second coefficient estimate, $\widehat{\beta}_2 > 0$, suggests that, relative to a random pair of ideologically misaligned books, pairs of books where at least one of the books is neutral are more likely to be purchased together. Neutral books are thus generally more likely to be purchased with books of disparate ideologies and have a unifying effect in the book network.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

Returning briefly to Table 1, the runtimes highlight that EMM reduces computational burden by order(s) of magnitude over existing approaches.

9. Discussion

In this paper, we present the PX model, a probit regression model for undirected, binary networks. The PX model adds a single parameter—latent correlation ρ—to the ordinary probit regression model that assumes independence of observations. Our focus in this paper is estimation of the effects of exogenous covariates on the observed network, $\boldsymbol{\beta}$, and prediction of unobserved network relations. Thus, we do not present uncertainty estimators for $\widehat{\boldsymbol{\beta}}$ or $\widehat{\rho}$. However, practitioners estimating the PX model may require uncertainty estimators to perform inference. Development and evaluation of estimators of the uncertainty in estimators of network data is nontrivial; indeed, entire papers are dedicated to this task for the simpler linear regression case (see, for example, Aronow et al., 2015; Marrs et al., 2017). Future development of uncertainty estimators for the PX model may draw upon existing literature for uncertainty in EM estimators (Louis, 1982) and the numerical approximations in this paper.

A popular notion in the analysis of network data is the presence of higher-order dependencies, meaning beyond second order (Hoff, 2005). The representation of triadic closure, a form of transitivity—the friend of my friend is likely to also be my friend— is one motivation for the latent eigenmodel (Hoff, 2008). The PX model does represent triadic closure to a degree. One can show that, given two edges of a triangle relation exist, $y_{ij} = y_{jk} = 1$, the probability that the third edge exists, $\mathbb{P}(y_{ik} = 1)$, increases as ρ increases. However, the increase in probability describing triadic closure under the PX model is fixed based on the estimated value of ρ, which is informed only by the first two moments of the data when using the EMM estimator. It may be desirable to develop a test for whether the PX model sufficiently represents the level of triadic closure as suggested by the data. One such test might compute the empirical probability that $\mathbb{P}(y_{ik} = 1 \mid y_{ij} = y_{jk} = 1)$ and compare this statistic to its distribution under the null that the PX model is the true model with correlation parameter $\rho = \widehat{\rho}$. Future work consists of theoretical development of the distributions of the test statistic(s) of choice under the null. Statistics of interest will likely be related to various clustering coefficients in the network literature (Wasserman & Faust, 1994; Watts & Strogatz, 1998).

We focus on the probit model in this paper. However, we find that this choice may limit the degree of covariance in the observed network $\{y_{ij}\}_{ij}$ that the PX model can represent. For constant mean $\mathbf{x}_{ij}^T\boldsymbol{\beta} = \mu$, the maximum covariance the PX model can represent is bounded by

$$\text{cov}[y_{ij}, y_{ik}] \leq \lim_{\rho \to 1/2} \int_{-\mu}^{\infty} \int_{-\mu}^{\infty} dF_\rho - \Phi(\mu)^2, \tag{21}$$

where dF_ρ is the bivariate standard normal distribution with correlation ρ. The use of different latent distributions for $\boldsymbol{\epsilon}$ other than normal may allow a model analogous to the PX model to represent a larger range of observed covariances $\text{cov}[y_{ij}, y_{ik}]$. Future work may consider a logistic distribution for $\boldsymbol{\epsilon}$, as some researchers prefer to make inference with logistic regression models for binary data due to the ease of interpretation.

Acknowledgements. This work utilized the RMACC Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The RMACC Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University. This work was also partially supported by the National Science Foundation under Grant no. 1856229.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

Competing interests. None.

Supplementary materials. For supplementary material for this article, please visit http://doi.org/10.1017/nws.2023.12.

Notes

1 We consider infinite exchangeability such that the exchangeable generating process is valid for arbitrarily large numbers of actors n, as in Hoover (1979) and Aldous (1981).
2 These unpublished data were compiled by Dr. Valdis Krebs for his website http://www.orgnet.com/ and are hosted, with permission, by Dr. Mark Newman at http://www-personal.umich.edu/mejn/netdata/polbooks.zip

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, *9*, 1981–2014.
Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, *11*(4), 581–598.
Aronow, P. M., Samii, C., & Assenova, V. A. (2015). Cluster–robust variance estimation for dyadic data. *Political Analysis*, *23*(4), 564–577.
Ashford, J. R., & Sowden, R. R. (1970). Multi-variate probit analysis. In *Biometrics* (pp. 535–546).
Atkinson, K. E. (2008). *An introduction to numerical analysis*. Hoboken, NJ: John Wiley & Sons.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.
Betzel, R. F., Bertolero, M. A., & Bassett, D. S. (2018). Non-assortative community structure in resting and task-evoked functional brain networks. *bioRxiv*, 355016.
Borgs, C., Chayes, J. T., Cohn, H., & Zhao, Y. (2014). An Lp theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *arXiv: 1401.2906*.
Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.
Caimo, A., & Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, *33*(1), 41–55.
Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361.
Connolly, M. A., & Liang, K.-Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika*, *75*(3), 501–506.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *39*(1), 1–22.
Dhaene, G. (1997). Pseudo-true values. In *Encompassing: Formulation, properties and testing* (pp. 7–30).
Eakin, E. (2004). Study finds a nation of polarized readers. *The New York Times*, *9*, 9.
Fagiolo, G., Reyes, J., & Schiavo, S. (2008). On the topological properties of the world trade web: A weighted network analysis. *Physica A*, *387*(15), 3868–3873.
Faust, K., & Wasserman, S. (1994). *Social network analysis: Methods and applications* (Vol. *249*). Cambridge: Cambridge University Press.
Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, *81*(395), 832–842.
Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
Greene, W. H. (2003). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.
Han, G., McCubbins, O. P., & Paulsen, T. H. (2016). Using social network analysis to measure student collaboration in an undergraduate capstone course. *Nacta Journal*, *60*(2), 176.
Handcock, M. S., Robins, G., Snijders, T., Moody, J., & Besag, J. (2003). *Assessing degeneracy in statistical models of social networks*. Tech. rept. Citeseer.
Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(2), 301–354.
Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., & Morris, M. (2019). *ergm: Fit, simulate and diagnose exponential-family models for networks*. The Statnet Project (https://statnet.org). R package version 3.10.4.
Heagerty, P. J., & Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, *93*(443), 1099–1111.
Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances neural information processing systems* (pp. 657–664).

Hoff, P., Fosdick, B., Volfovsky, A., & He, Y. (2017). *amen: Additive and multiplicative effects models for networks and relational data*. R package version 1.3.
Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, *100*(469), 286–295.
Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098.
Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, *76*(373), 33–50.
Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. Princeton, NJ: Institute for Advanced Study. Preprint.
Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability: Weather modification* (p. 221). Berkeley, CA: University of California Press.
Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008a). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, *24*(3), 1–29.
Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008b). Goodness of fit of social network models. *Journal of the American Statistical Association*, *103*(481), 248–258.
Kallenberg, O. (2006). *Probabilistic symmetries and invariance principles*. New York: Springer Science & Business Media.
Le Cessie, S., & Van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Journal of the Royal Statistical Society*, *43*(1), 95–108.
Li, Y., & Schafer, D. W. (2008). Likelihood analysis of the multivariate ordinal probit regression model for repeated ordinal responses. *Computional Statistics and Data Analysis*, *52*(7), 3474–3492.
Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Oliver, S. (2006). *SAS for mixed models*. Cary, NC: SAS Publishing,
Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *44*(2), 226–233.
Lovász, L., & Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial. Theory, Series B*, *96*(6), 933–957.
Ma, Z., Ma, Z., & Yuan, H. (2020). Universal latent space model fitting for large networks with edge covariates. *The Journal of Machine Learning Research*, *21*(1), 86–152.
Marrs, F. W., Fosdick, B. K., & McCormick, T. H. (2017). Standard errors for regression on relational data with exchangeable errors. *arxiv: 1701.05530*.
Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*(455), 1077–1087.
Ochi, Y., & Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, *71*(3), 531–543.
Petersen, K. B., & Pedersen, M. S. (2008). The matrix cookbook. *Technical University of Denmark*, *7*(15), 510.
Schummer, J. (2005). Reading nano: The public interest in nanotechnology as reflected in purchase patterns of books. *Public Understanding of Science*, *14*(2), 163–183.
Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, *106*(496), 1361–1370.
Sewell, D. K., & Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, *110*(512), 1646–1657.
Shalizi, C. R., & Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *Annals of Statistics*, *41*(2), 508.
Shi, F., Shi, Y., Dokshin, F. A., Evans, J. A., & Macy, M. W. (2017). Millions of online book co-purchases reveal partisan differences in the consumption of science. *Nature Human Behavior*, *1*(4), 0079.
Snijders, T. A. B. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, *3*(2), 1–40.
Snijders, T. A. B., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, *6*(4), 471–486.
Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, *36*(1), 99–153.
Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, *40*(4), 961–971.
Warner, R. M., Kenny, D. A., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, *37*(10), 1742–1757.
Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. *8*). Cambridge: Cambridge University Press.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *Nature*, *393*(6684), 440–442.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

Wong, G. Y. (1982). Round robin analysis of variance via maximum likelihood. *Journal of the American Statistical Association*, *77*(380), 714–724.
Wu, C. F. J. (1983). On the convergence properties of the em algorithm. In *The annals of statistics* (pp. 95–103).
Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), Article17.
Zhang, X., Xu, G., & Zhu, J. (2022). Joint latent space models for network data with high-dimensional node variables. *Biometrika*, *109*(3), 707–720.

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 $\mathbf{\Omega}$, it can be shown that ρ lies in the interval $[0, 1/2)$ when $\mathbf{\Omega}$ 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:

$$\widehat{\rho}^{(0)} = \frac{100n}{4(100n + |\mathscr{A}|)} + \frac{|\mathscr{A}|}{(100n + |\mathscr{A}|)} \left(\frac{1}{|\mathscr{A}|} \sum_{jk,lm \in \mathscr{A}} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}] \right). \tag{A1}$$

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

A.2 Implementation of β expectation step

Under general correlation structure, computation of the expectation $E[\boldsymbol{\epsilon} \mid \mathbf{y}]$ (step 1 in Algorithm 1, where we drop conditioning on $\rho^{(\nu)}$ and $\boldsymbol{\beta}^{(\nu)}$ 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 $\mathbf{\Omega}$ to compute $E[\boldsymbol{\epsilon} \mid \mathbf{y}]$ 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

$$E[\epsilon_{jk} \mid \mathbf{y}] = E[E[\epsilon_{jk} \mid \boldsymbol{\epsilon}_{-jk}, y_{jk}] \mid \mathbf{y}], \tag{A2}$$

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

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

$$\begin{aligned}\epsilon_{jk} \mid \boldsymbol{\epsilon}_{-jk} &\sim \mathrm{N}\left(\mu_{jk}, \sigma_n^2\right), \\ \mu_{jk} &= -\sigma_n^2 \mathbf{1}_{jk}^T \left(p_2 \mathscr{S}_2 + p_3 \mathscr{S}_3\right) \widetilde{\boldsymbol{\epsilon}}_{-jk}, \\ \sigma_n^2 &= \frac{1}{p_1},\end{aligned} \tag{A3}$$

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 $\widetilde{\boldsymbol{\epsilon}}_{-jk}$ as the vector $\boldsymbol{\epsilon}$ 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

$$E[z \mid y] = \mu + \sigma \frac{\phi(\widetilde{\eta})}{\Phi(\widetilde{\eta})(1 - \Phi(\widetilde{\eta}))}(y - \Phi(\widetilde{\eta})), \tag{A4}$$

where $\widetilde{\eta} := (\eta + \mu)/\sigma$. Now, taking $z = (\epsilon_{jk} \mid \boldsymbol{\epsilon}_{-jk})$, we have that

$$E[\epsilon_{jk} \mid \boldsymbol{\epsilon}_{-jk}, y_{jk}] = \mu_{jk} + \sigma_n \left(\frac{\phi(\widetilde{\mu}_{jk})\left(y_{jk} - \Phi(\widetilde{\mu}_{jk})\right)}{\Phi(\widetilde{\mu}_{jk})(1 - \Phi(\widetilde{\mu}_{jk}))} \right), \tag{A5}$$

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

We now turn to the outermost conditional expectation in (A2). Substituting the expression for μ_{jk} into (A5), we have that

$$E[\epsilon_{jk} \mid \mathbf{y}] = -\sigma_n^2 \mathbf{1}_{jk}^T \left(p_2 \mathscr{S}_2 + p_3 \mathscr{S}_3\right) E[\boldsymbol{\epsilon} \mid \mathbf{y}] + \sigma_n E\left[\frac{\phi(\widetilde{\mu}_{jk})\left(y_{jk} - \Phi(\widetilde{\mu}_{jk})\right)}{\Phi(\widetilde{\mu}_{jk})(1 - \Phi(\widetilde{\mu}_{jk}))} \,\middle|\, \mathbf{y} \right]. \tag{A6}$$

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

$$0 \approx g(\mathbf{w}) := (-\mathbf{I} + \mathbf{B})\mathbf{w} + \sigma_n \left(\frac{\phi(\widetilde{\mathbf{w}})\left(\mathbf{y} - \Phi(\widetilde{\mathbf{w}})\right)}{\Phi(\widetilde{\mathbf{w}})(1 - \Phi(\widetilde{\mathbf{w}}))} \right), \tag{A7}$$

where we define $\mathbf{B} := -\sigma_n^2\left(p_2\mathscr{S}_2 + p_3\mathscr{S}_3\right)$, $\widetilde{\mathbf{w}} := (\mathbf{B}\mathbf{w} + \mathbf{X}\boldsymbol{\beta})/\sigma_n$, and the functions $\phi(.)$ and $\Phi(.)$ are applied element-wise. The approximation in (A7) 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 $\rho = 0$,

$$\mathbf{w}_0 := \frac{\phi(\mathbf{X}\boldsymbol{\beta})\left(\mathbf{y} - \Phi(\mathbf{X}\boldsymbol{\beta})\right)}{\Phi(\mathbf{X}\boldsymbol{\beta})(1 - \Phi(\mathbf{X}\boldsymbol{\beta}))}. \tag{A8}$$

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

$$\widehat{\mathbf{w}}^{(\nu+1)} = \widehat{\mathbf{w}}^{(\nu)} - \left(\frac{\partial}{\partial \mathbf{w}^T} g\left(\widehat{\mathbf{w}}^{(\nu)}\right) \right)^{-1} g\left(\widehat{\mathbf{w}}^{(\nu)}\right). \tag{A9}$$

The inverse in (A9) 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 Newton method updates, we numerically approximate the inverse in (A9). First, we define $\nu(w_{jk}) = \sigma_n \frac{\phi(w_{jk})(y_{jk} - \Phi(w_{jk}))}{\Phi(w_{jk})(1 - \Phi(w_{jk}))}$, where we define the vector $\nu(\mathbf{w}) = \{\nu(w_{jk})\}_{jk}$, and write the derivative

$$\frac{\partial}{\partial \mathbf{w}^T} g(\mathbf{w}) = \mathbf{B} - \mathbf{I} + \mathbf{D}\mathbf{B}. \tag{A10}$$

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

where we define

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

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 (A10), we are then able to write the second term in (A9) as

$$\begin{aligned}\left(\frac{\partial}{\partial \mathbf{w}^T} g(\widehat{\mathbf{w}})\right)^{-1} g(\widehat{\mathbf{w}}) &= (\mathbf{B} - \mathbf{I} + \mathbf{DB})^{-1}((\mathbf{B} - \mathbf{I})\mathbf{w} + v(\mathbf{w}))\,, \\ &= \mathbf{B}^{-1}\left(\mathbf{I} + \mathbf{D} - \mathbf{B}^{-1}\right)^{-1}((\mathbf{B} - \mathbf{I})\mathbf{w} + v(\mathbf{w}))\,. \end{aligned} \tag{A11}$$

We notice that the matrix $\mathbf{I} + \mathbf{D}$ is diagonal, but not homogeneous (in which case we compute (A11) 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:

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

A.3 Approximation to ρ expectation step

The maximization of the expected likelihood with respect to ρ relies on the computation of $\gamma_i = E[\boldsymbol{\epsilon}^T \mathscr{S}_i \boldsymbol{\epsilon} \mid \mathbf{y}]/|\Theta_i|$, 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 $\boldsymbol{\epsilon}$ 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 as follows:

$$\begin{aligned}\gamma_1 &= \frac{1}{|\Theta_1|}\sum_{jk} E\left[\epsilon_{jk}^2 \mid \mathbf{y}\right] \approx \frac{1}{|\Theta_1|}\sum_{jk} E\left[\epsilon_{jk}^2 \mid y_{jk}\right], \\ \gamma_2 &= \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E\left[\epsilon_{jk}\epsilon_{lm} \mid \mathbf{y}\right] \approx \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E\left[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}\right], \\ \gamma_3 &= \frac{1}{|\Theta_3|}\sum_{jk,lm\in\Theta_3} E\left[\epsilon_{jk}\epsilon_{lm} \mid \mathbf{y}\right] \approx \frac{1}{|\Theta_3|}\sum_{jk,lm\in\Theta_3} E\left[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}\right], \end{aligned} \tag{A12}$$

where Θ_i is the set of ordered pairs of relations (jk, lm) which correspond entries in $\mathscr{S}_i$ that is 1, for $i \in \{1, 2, 3\}$. These approximations are natural first-order approximations: recalling that $y_{jk} =$

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

$\mathbb{1}\left[\epsilon_{jk} > -\mathbf{x}_{jk}^T\boldsymbol{\beta}\right]$, the approximations in (A12) are based on the notion that knowing the domains of ϵ_{jk} and ϵ_{lm} is significantly more informative for $E[\epsilon_{jk}\epsilon_{lm} \mid \mathbf{y}]$ than knowing the domain of, for example, ϵ_{ab}.

The approximations in (A12) are orders of magnitude faster to compute than the expectations when conditioning on all observations $E[\epsilon_{jk}\epsilon_{lm} \mid \mathbf{y}]$. In particular, when $i \in \{1, 3\}$, the expectations are available in closed form:

$$E\left[\epsilon_{jk}^2 \mid y_{jk}\right] = 1 - \eta_{jk}\frac{\phi(\eta_{jk})(y_{jk} - \Phi(\eta_{jk}))}{\Phi(\eta_{jk})(1 - \Phi(\eta_{jk}))},$$

$$E\left[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}\right] = \frac{\phi(\eta_{jk})\phi(\eta_{lm})(y_{jk} - \Phi(\eta_{jk}))(y_{lm} - \Phi(\eta_{lm}))}{\Phi(\eta_{jk})\Phi(\eta_{lm})(1 - \Phi(\eta_{jk}))(1 - \Phi(\eta_{lm}))},$$

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:

$$\begin{aligned}
&E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}] \\
&\quad = \rho\left(1 - \frac{\bar{\eta}_{jk}\phi(\eta_{jk})}{L_{jk,lm}}\Phi\left(\frac{\bar{\eta}_{lm} - \bar{\rho}\,\bar{\eta}_{jk}}{\sqrt{1-\rho^2}}\right) - \frac{\bar{\eta}_{lm}\phi(\eta_{lm})}{L_{jk,lm}}\Phi\left(\frac{\bar{\eta}_{jk} - \bar{\rho}\,\bar{\eta}_{lm}}{\sqrt{1-\rho^2}}\right)\right) \\
&\qquad + \frac{1}{L_{jk,lm}}\sqrt{\frac{1-\rho^2}{2\pi}}\phi\left(\sqrt{\frac{\eta_{jk}^2 + \eta_{lm}^2 - 2\rho\,\eta_{jk}\eta_{lm}}{1-\rho^2}}\right), \quad |\{j, k\} \cap \{l, m\}| = 1, \\
&L_{jk,lm} = \mathbb{P}\left((2y_{jk} - 1)\epsilon_{jk} > -\eta_{jk} \cap (2y_{lm} - 1)\epsilon_{lm} > -\eta_{lm}\right),
\end{aligned} \tag{A13}$$

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

A.3.2 Linearization

The computation of $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ in (A13) 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, $\gamma_2 \approx \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$, is approximately linear in ρ. This linearity is somewhat intuitive, as the sample mean $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ has expectation equal to ρ, and is thus an asymptotically linear function of ρ. As the sample mean $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ concentrates around its expectation, it concentrates around a linear function of ρ, and it is reasonable to approximate the sample mean $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ 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,

$$\begin{aligned}
\gamma_2 &\approx a_2 + b_2\rho, \\
a_2 &= \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E\left[\epsilon_{jk} \mid y_{jk}\right] E[\epsilon_{lm} \mid y_{lm}], \\
&= \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} \frac{\phi(\eta_{jk})\phi(\eta_{lm})(y_{jk} - \Phi(\eta_{jk}))(y_{lm} - \Phi(\eta_{lm}))}{\Phi(\eta_{jk})\Phi(\eta_{lm})(1 - \Phi(\eta_{jk}))(1 - \Phi(\eta_{lm}))}, \\
c_2 &= \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E\left[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}\right]\Big|_{\rho=1}, \\
b_2 &= c_2 - a_2.
\end{aligned} \tag{A14}$$

To compute c_2, we must compute the value of $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ when $\rho = 1$. Computing $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ is simple when the values $y_{jk} = y_{lm}$, as in this case $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}] =$

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

$E\left[\epsilon_{jk}^2 \mid y_{jk} = y_{lm}\right]$ since, when $\rho = 1$, $\epsilon_{jk} = \epsilon_{lm}$. 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 $y_{jk} = \mathbb{1}[\epsilon_{jk} > -\eta_{jk}]$ and $y_{jk} = \mathbb{1}[\epsilon_{lm} > -\eta_{lm}]$, 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 desired expectation may be approximated $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}] \approx E\left[\epsilon_{jk}^2 \mid \epsilon_{jk} \in U_{jk} \cap U_{lm}\right]$. In the second case, when $y_{jk} \neq y_{lm}$ and $U_{jk} \cap U_{lm} = \varnothing$, we make the approximation by integrating over the sets U_{jk} and U_{lm}. That is, by taking

$$\begin{aligned} &E\left[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}\right] \\ &\quad \approx E\left[\epsilon_{jk}^2 \mid \epsilon_{jk} \in U_{jk}\right] \mathbb{P}\left(\epsilon_{jk} \in U_{jk}\right) + E\left[\epsilon_{lm}^2 \mid \epsilon_{lm} \in U_{lm}\right] \mathbb{P}(\epsilon_{lm} \in U_{lm})\,. \end{aligned}$$

To summarize, we compute c_2 in (A14) when $\rho = 1$ by using the following approximation to $E[\epsilon_{jk}\epsilon_{lm} \mid \mathbf{y}]\Big|_{\rho=1}$:

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

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 $\boldsymbol{\beta}$ 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.

A.4.1 Update $\boldsymbol{\beta}$

To maximize $\ell_{\mathbf{y}}$ with respect to $\boldsymbol{\beta}$ (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 $\mathbf{X}$ 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 $\bar{\eta} = \frac{1}{|\mathscr{M}^c|}\sum_{lm \in \mathscr{M}^c} \mathbf{x}_{lm}^T \widehat{\boldsymbol{\beta}}$ and we compute $\mathbf{w} = E[\boldsymbol{\epsilon} \mid \mathbf{y}]$ 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 $\mathbf{X}$ and $\mathbf{y}$, the estimation routine may be accomplished as described in Algorithm 1.

A.4.2 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 (A12), the expressions for the expectation step under missing data are

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

$$\gamma_1 \approx \frac{1}{|\mathscr{M}^c|} \sum_{jk \in \mathscr{M}^c} E\left[\epsilon_{jk}^2 \mid y_{jk}\right], \tag{A15}$$

$$\gamma_2 \approx \frac{1}{|\mathscr{A}^{(s)}|} \sum_{jk,lm \in \mathscr{A}^{(s)}} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}].$$

$$\gamma_3 \approx \frac{\sum_{jk,lm \in \Theta_3} E[\epsilon_{jk} \mid y_{jk}] E[\epsilon_{lm} \mid y_{lm}] \mathbb{1}[jk \in \mathscr{M}^c] \mathbb{1}[lm \in \mathscr{M}^c]}{\sum_{jk,lm \in \Theta_3} \mathbb{1}[jk \in \mathscr{M}^c] \mathbb{1}[lm \in \mathscr{M}^c]},$$

$$\approx \frac{1}{|\Theta_3|} \left(\left(\frac{|\Theta_1|}{|\mathscr{M}^c|} \sum_{jk \in \mathscr{M}^c} E[\epsilon_{jk} \mid y_{jk}] \right)^2 - \frac{|\Theta_1|}{|\mathscr{M}^c|} \sum_{jk \in \mathscr{M}^c} E[\epsilon_{jk} \mid y_{jk}]^2 - \frac{|\Theta_2|}{|\mathscr{A}^{(s)}|} \sum_{jk,lm \in \mathscr{A}^{(s)}} E[\epsilon_{jk} \mid y_{jk}] E[\epsilon_{lm} \mid y_{lm}] \right),$$

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 (A15), the maximization of $\ell_{\mathbf{y}}$ with respect to ρ (Step 2 in Algorithm 1) may proceed as usual.

A.4.3 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 $\mathbf{X}$ and $\mathbf{y}$ earlier in this section under the "**Update $\boldsymbol{\beta}$**" subheading. Then, prediction may be performed by proceeding as described in Section 6 with the full observed $\mathbf{X}$ matrix and imputing the missing values in $\mathbf{y}$ (again as described above in this section under the "**Update $\boldsymbol{\beta}$"** subheading).

B. Parameters of undirected exchangeable network covariance matrices

In this section, we give a 3×3 matrix equation to invert $\boldsymbol{\Omega}$ 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

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

We find empirically that the inverse matrix of any undirected exchangeable network covariance matrix has the same form, that is $\boldsymbol{\Omega}^{-1} = \sum_{i=1}^{3} \mathbf{p}_i \mathscr{S}_i$. 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 $\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}$, 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 $\boldsymbol{\phi}$ and $\mathbf{p}$, and the fact that $\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1} = \mathbf{I}$, we get the following matrix equation for the parameters $\mathbf{p}$ given $\boldsymbol{\phi}$

$$\mathbf{C}(\boldsymbol{\phi})\mathbf{p} = [1, 0, 0]^T, \tag{B1}$$

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

where the matrix $\mathbf{C}(\boldsymbol{\phi})$ 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 $\boldsymbol{\Omega}$ with a 3×3 inverse to find the parameters $\mathbf{p}$ of $\boldsymbol{\Omega}^{-1}$. Explicitly solving these linear equations, the expressions for $\mathbf{p}$ are given by

$$\begin{aligned} p_1 &= 1 - (2n-4)p_2, \\ p_2 &= \frac{1 + (n-3)p_3}{(2n-4)\rho - n + 2 - 1/\rho}, \\ p_3 &= \frac{-4\rho^2}{(n-3)4\rho + (1 + (2n-8)\rho)((2n-4)\rho - n + 2 - 1/\rho)}. \end{aligned} \tag{B2}$$

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:

$$\begin{aligned} p_1 &\approx \frac{1}{1-2\rho} + O(n^{-1}), \\ p_2 &\approx \frac{-1}{n(1-2\rho)} + O(n^{-2}), \\ p_3 &\approx \frac{2}{n^2(1-2\rho)} + O(n^{-3}). \end{aligned} \tag{B3}$$

The equation (B1) allows one to compute the partial derivatives $\left\{\frac{\partial \phi_i}{\partial p_j}\right\}$. First, based on (B1), we can write $\mathbf{C}(\mathbf{p})\boldsymbol{\phi} = [1, 0, 0]^T$. Then, we note that the matrix function $\mathbf{C}(\boldsymbol{\phi})$ in (B1) is linear in the terms $\boldsymbol{\phi}$, 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 $\mathbf{C}(\mathbf{p})\boldsymbol{\phi} = [1, 0, 0]^T$ with respect to p_j and solving gives

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

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, $\widehat{\beta}_{\text{EMM}}$, depends on the estimated value of ρ, $\widehat{\rho}_{\text{EMM}}$, through the covariance matrix $\boldsymbol{\Omega}$. Explicitly, given $\boldsymbol{\Omega}$, the EMM estimator

$$\widehat{\beta}_{\text{EMM}} = (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\widehat{E[\mathbf{z} \mid \mathbf{y}]}, \tag{C1}$$

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

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

direct analysis, to evaluate consistency of $\widehat{\beta}_{\text{EMM}}$, we define a bounding estimator that is easier to analyze,

$$\widehat{\beta}_{\text{bound}} = (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{u}, \quad u_{jk} = E[z_{jk} \mid y_{jk}]. \tag{C2}$$

It is immediately clear that $\widehat{\beta}_{\text{bound}}$ is unbiased, since $E[u_{jk}] = \mathbf{x}_{jk}^T\boldsymbol{\beta}$. Further, the approximations made in the EMM algorithm are meant to bound $\left|\left|\widehat{\beta}_{\text{EMM}} - \beta^*_{\text{MLE}}\right|\right|_2^2 \leq \left|\left|\widehat{\beta}_{\text{bound}} - \beta^*_{\text{MLE}}\right|\right|_2^2$, 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 $\widehat{\beta}_{\text{EMM}}$ is closer to β^*_{MLE} than $\widehat{\beta}_{\text{bound}}$. Then, consistency of $\widehat{\beta}_{\text{bound}}$ implies consistency of $\widehat{\beta}_{\text{EMM}}$, since we assume that the true MLE is consistent.

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

1. The true model follows a latent variable model,
$$\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij} > 0\right), \tag{C3}$$
$$E[\epsilon_{jk}] = 0.$$
where $\boldsymbol{\epsilon}$ is not necessarily normally distributed.
2. The design matrix $\mathbf{X}$ is such that the expressions $n^{-(1+i)}\mathbf{X}^T\mathcal{S}_i\mathbf{X}$, for $i \in \{1, 2, 3\}$, converge in probability to constant matrices.
3. The fourth moments of $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are bounded, $||\mathbf{x}_{jk}||_4 \leq C_1 < \infty$ and $E\left[\epsilon_{jk}^4\right] \leq C_2 < \infty$.
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 $\widehat{\rho}_{\text{EMM}}$. The fourth assumption defines the minimal independence structure.

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

$$\widehat{\beta}_{\text{bound}} = \beta + \left(n^{-2}\sum_{i=1}^{3} p_i\mathbf{X}^T\mathcal{S}_i\mathbf{X}\right)^{-1}\left(n^{-2}\sum_{i=1}^{3} p_i\mathbf{X}^T\mathcal{S}_i\mathbf{v}\right), \quad v_{jk} = E[\epsilon_{jk} \mid y_{jk}]. \tag{C4}$$

Then, as noted in the previous paragraph, the bounding estimator is unbiased, $E\left[\widehat{\beta}_{\text{bound}}\right] = \beta$. It remains to establish sufficient conditions for which $\widehat{\beta}_{\text{bound}}$ converges to its expectation in probability. Noting the orders of $\{p_i\}_i$ in (B3), 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 $\left(n^{-2}\sum_{i=1}^{3} p_i\mathbf{X}^T\mathcal{S}_i\mathbf{v}\right)$ converges in probability to its expectation (zero) is that its variance tends to zero. Expanding this variance expression,

$$\begin{aligned}\text{var}\left(n^{-2}\sum_{i=1}^{3} p_i\mathbf{X}^T\mathcal{S}_i\mathbf{v}\right) &= n^{-4}\sum_{i=1}^{3}\sum_{j=1}^{3} p_ip_j\mathbf{X}^T\mathcal{S}_iE[\mathbf{v}\mathbf{v}^T]\mathcal{S}_j\mathbf{X}, \\ &= n^{-4}\sum_{i=1}^{3}\sum_{j=1}^{3} p_ip_j\sum_{jk,lm\in\Theta_i}\sum_{rs,tu\in\Theta_j}\mathbf{x}_{jk}\mathbf{x}_{rs}^T E[v_{lm}v_{tu}].\end{aligned} \tag{C5}$$

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

By assumption, every term in the sum expression in (C5) is bounded. Also by assumption, the expectation $E[\nu_{lm}\nu_{tu}]$ is zero whenever the relations lm and tu do not share an actor. Using the expressions in (B3) ($p_i \propto n^2|\Theta_i|^{-1}$) 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 $\widehat{\beta}_{\text{bound}}$ converges to zero, so that $\widehat{\beta}_{\text{bound}}$ converges in probability to the true $\boldsymbol{\beta}$, as does $\widehat{\beta}_{\text{EMM}}$.

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

Using the expressions in (B3) and differentiating the expected log-likelihood with respect to ρ, the maximum likelihood estimator is

$$\widehat{\rho}_{\text{MLE}} = \frac{1}{2} + \frac{1}{n^3}E\left[\boldsymbol{\epsilon}^T \mathscr{S}_2 \boldsymbol{\epsilon} \mid \mathbf{y}\right] - \frac{1}{n^2}E\left[\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \mid \mathbf{y}\right] - \frac{2}{n^4}E\left[\boldsymbol{\epsilon}^T \mathscr{S}_3 \boldsymbol{\epsilon} \mid \mathbf{y}\right] + O(n^{-1}). \tag{C6}$$

In the EMM algorithm, we approximate the expectations in (C6) using pairwise conditioning. Then, we have that

$$\widehat{\rho}_{\text{EMM}} = \frac{1}{2} + \frac{1}{n^3}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}] - \frac{1}{n^2}\sum_{jk} E\left[\epsilon_{jk}^2 \mid y_{jk}\right] \ldots \tag{C7}$$

$$\ldots - \frac{2}{n^4}\sum_{jk,lm\in\Theta_3} E[\epsilon_{jk} \mid y_{jk}]E[\epsilon_{lm} \mid y_{lm}] + O(n^{-1}).$$

According to the exchangeability assumption of the errors, the pairwise expectations are known, and the EMM estimator of ρ is unbiased, $E[\widehat{\rho}_{\text{EMM}}] = E[\epsilon_{jk}\epsilon_{lm}] = \rho$. The EMM estimator $\widehat{\rho}_{\text{EMM}}$ converges to its expectation when the sums of conditional expectations in (C7) 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,

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

since $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ is independent $E[\epsilon_{rs}\epsilon_{tu} \mid y_{rs}, y_{tu}]$ whenever all the indices $\{j, k, l, m, r, s, t, u\}$ are distinct. Thus, each of the sums of expectations in (C7) has variance that tends to zero, so that they converge to their marginal expectations, and $\widehat{\rho}_{\text{EMM}}$ 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 $n^{-3}E[\epsilon^T \mathscr{S}_2 \epsilon_{lm}]$ converges to some constant on $[0, 1/2)$). Hence, when the data-generating mechanism is non-Gaussian and non-exchangeable, we expect $\widehat{\rho}_{\text{EMM}}$ 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,

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

1967; Dhaene, 1997). In the discussion of consistency of $\widehat{\beta}_{\text{EMM}}$, we only require that $\widehat{\rho}_{\text{EMM}}$ 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 $\widehat{\beta}_{\text{EMM}}$ 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 $\widehat{\beta}_{\text{EMM}}$ 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 $\boldsymbol{\beta}$. 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 $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]^T = [-1, 1/2, 1/2, 1/2]^T$ 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 $\sigma_{\mathbf{x}}^2 i = 1/3$.

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 D1 and D2, respectively. Figures D1 and D2 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 is 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 $\mathbf{X}$ realization, when

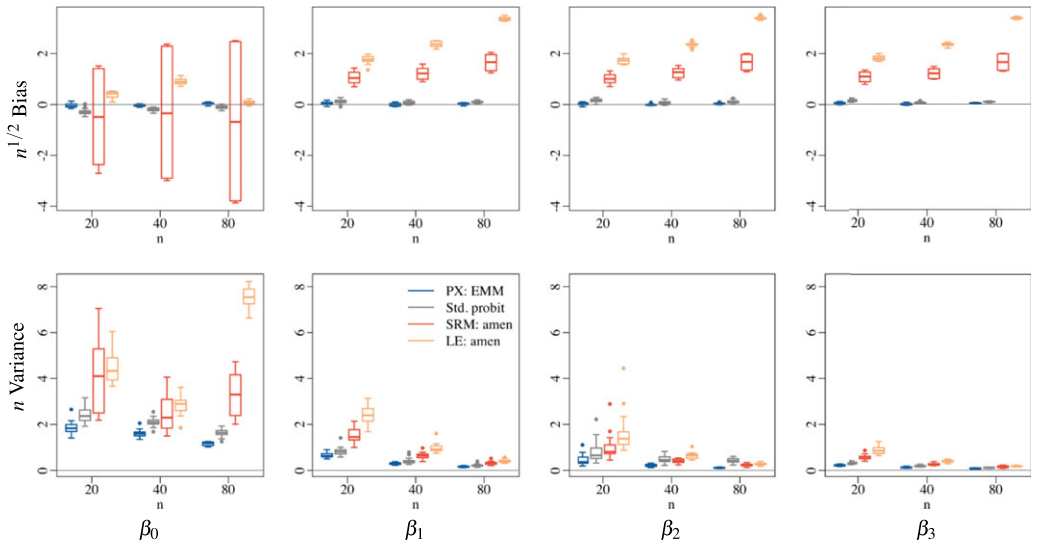

Figure D1. PX model: Scaled bias and variance of estimators of $\boldsymbol{\beta}$ for a given $\mathbf{X}$ when generating from the PX model. Variability captured by the boxplots reflects variation with $\mathbf{X}$.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

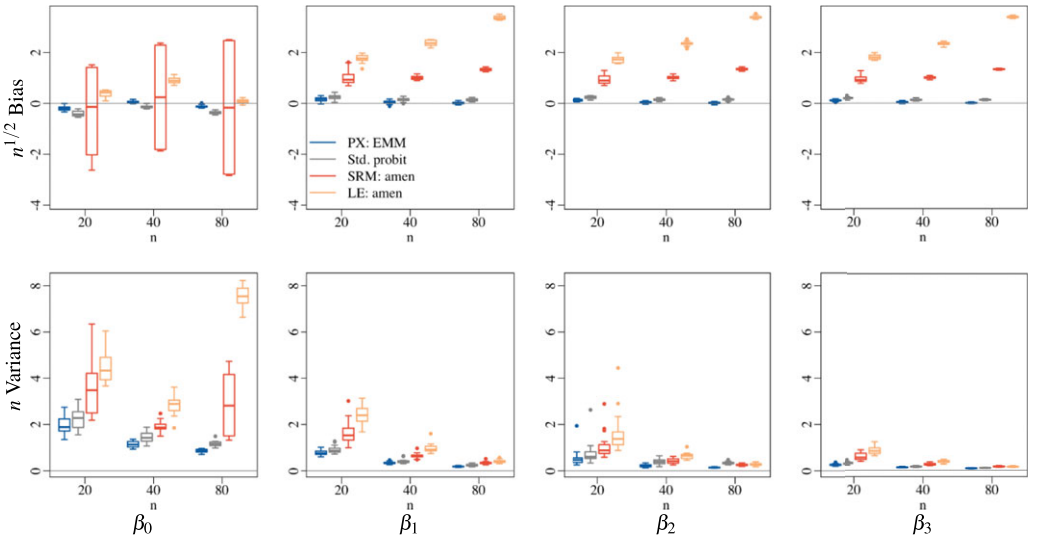

Figure D2. LE model: Scaled bias and variance of estimators of $\boldsymbol{\beta}$ for a given **X** when generating from the latent eigenmodel. Variability captured by the boxplots reflects variation with **X**.

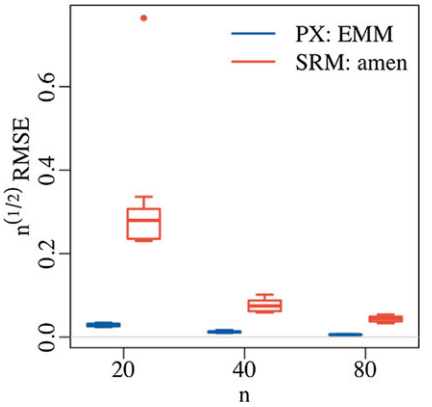

Figure D3. RMSE, scaled by $n^{1/2}$, of the EMM estimator and `amen` estimator of the social relations model of ρ when generating from the PX model. Variability captured by the boxplots reflects variation in $n^{1/2}$RMSE with **X**.

generating from the PX model. In Figure D3, 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 $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ is given in Figure D4. All coefficient estimators, for both PX: EMM and standard probit regression, appear consistent, but the PX: EMM has lower RSME.

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

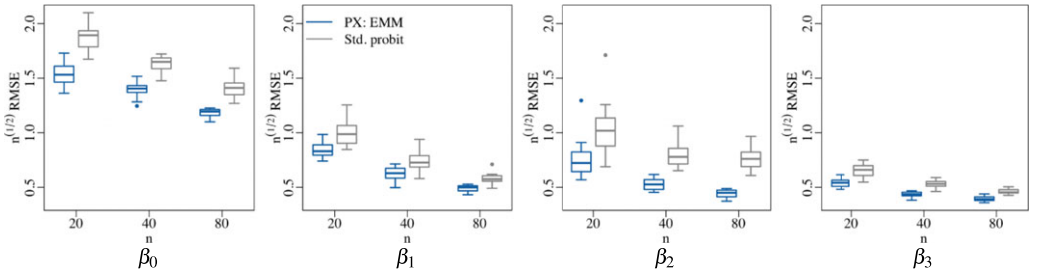

Figure D4. *t***model:** Scaled RMSE, for PX: EMM and standard probit regression, when generating from the PX model modified to have latent errors with heavier-tailed distribution.

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 E1 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 $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ with respect to ρ. In Figure E2, we confirm that this

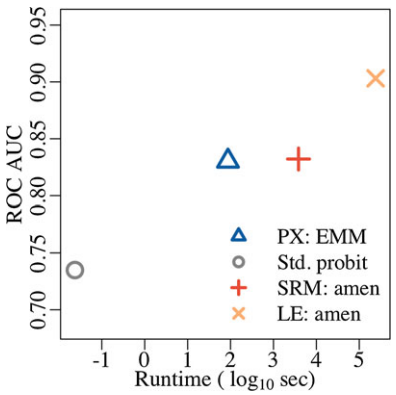

Figure E1. Out-of-sample performance in 10-fold cross-validation, as measured by area under the precision-recall curve (ROC AUC), plotted against mean runtime in the cross-validation for Krebs' political books network. The estimators are standard probit assuming independent observations (Std. probit), 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).

https://doi.org/10.1017/nws.2023.12 Published online by Cambridge University Press

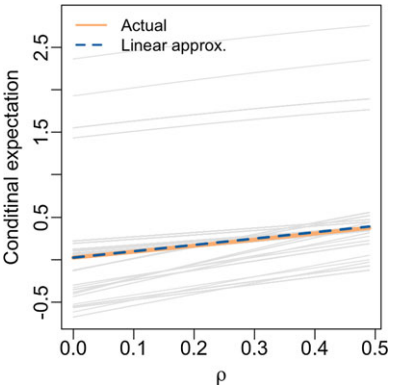

Figure E2. The average of all pairwise expectations $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ is shown in orange, and the linear approximation to this average, described in Section 5, is shown in dashed blue. In addition, pairwise conditional expectations $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ are shown in light gray, for a random subset of 500 relation pairs $(jk, lm) \in \Theta_2$.

approximation is reasonable for the political books network data set. Figure E2 shows that the linear approximation to $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ (dashed blue line), as described in detail in Section A.3, agrees well with the true average of the pairwise expectations (solid orange line).

Cite this article: Marrs F. W. and Fosdick B. K. (2023). Regression of binary network data with exchangeable latent errors. *Network Science* **11**, 502–535. https://doi.org/10.1017/nws.2023.12