DISSERTATION

REGRESSION OF NETWORK DATA: DEALING WITH DEPENDENCE

Submitted by

Frank W. Marrs

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2019

Doctoral Committee:

Advisor: Bailey K. Fosdick

F. Jay Breidt
Wen Zhou
James B. Wilson

ABSTRACT

REGRESSION OF NETWORK DATA: DEALING WITH DEPENDENCE

Network data, which consist of measured relations between pairs of actors, characterize some of the most pressing problems of our time, from environmental treaty legislation to human migration flows. A canonical problem in analyzing network data is to estimate the effects of exogenous covariates on a response that forms a network. Unlike typical regression scenarios, network data often naturally engender excess statistical dependence – beyond that represented by covariates – due to relations that share an actor. For analyzing bipartite network data observed over time, we propose a new model that accounts for excess network dependence directly, as this dependence is of scientific interest. In an example of international state interactions, we are able to infer the networks of influence among the states, such as which states' military actions are likely to incite other states' military actions. In the remainder of the dissertation, we focus on situations where inference on effects of exogenous covariates on the network is the primary goal of the analysis, and thus, the excess network dependence is a nuisance effect. In this setting, we leverage an exchangeability assumption to propose novel parsimonious estimators of regression coefficients for both binary and continuous network data, and new estimators for coefficient standard errors for continuous network data. The exchangeability assumption we rely upon is pervasive in network and array models in the statistics literature, but not previously considered when adjusting for dependence in a regression of network data.Although the estimators we propose are aligned with many network models in the literature, our estimators are derived from the assumption of exchangeability rather than proposing a particular parametric model for representing excess network dependence in the data.

ACKNOWLEDGEMENTS

I am supremely grateful for the support, feedback, and guidance of many people during my graduate experience. First, I would like to thank my parents, Frank and Ellen Marrs, and my brother Matt Marrs, for their continual belief in me. My success is only a reflection of their dedication to me. I would also like to thank my partner Zoe Iasilli, without whose love and compassionate support I would not be writing this dissertation.

I owe a large debt of gratitude to my advisor, Dr. Bailey Fosdick. Her belief in me from the start of our research together, and her patience with my many failures, were instrumental in completion of this work. Her kind advice, from technical writing to career decisions, have made me a much better researcher.

I have received terrific input from many CSU faculty, within and without the Statistics Department. I especially thank the entire Statistics Department, whose welcoming atmosphere meant that I never lacked someone to talk to. Specifically, I would like to thank my committee, Dr. Jay Breidt, Dr. Wen Zhou, and Dr. James Wilson, for their feedback on this dissertation. I appreciate the thoughtful conversations and career advice from Dr. Jennifer Hoeting, Dr. Colleen Webb, Dr. Mevin Hooten, and Dr. Josh Keller. In addition to career advice, Dr. Julia Sharp provided excellent guidance in the Statistical Collaboration Lab.

In addition to faculty at CSU, I am grateful for the guidance from and collaborative experiences with faculty outside the university. Thank you to Dr. Skyler Cranmer, Dr. Tobias Bohmelt, Dr. Benjamin Campbell, and Dr. Tyler McCormick for being such excellent collaborators. I appreciate input from Dr. Dan Larremore and Dr. James Wilson on future career and research directions. I am sincerely grateful for Dr. Karim Lounici for getting me started in Statistics and encouraging me to pursue my PhD.

I would like to thank the CSU Statistics faculty and department for their flexibility in PhD course requirements, which has allowed me to complete this dissertation (slightly) before the five

year standard. I appreciate the help in navigating all administrative red tape from Katy Jackson, Kristin Stephens, and Karena Alons in the CSU Statistics office.

I have made many friends in the Statistics Department, and I am truly grateful for their support as I pursued my degree. I especially want to mention good friend Dr. Zach Weller for listening to my complaints throughout the program. Dr. Henry Scharf and Dr. Josh Hewitt were excellent sounding boards for some of my wilder ideas. Others who were instrumental in my success are David Brown, Connor Gibbs, Ryan Haunfelder, Ryan Hicks, Dr. Soo-Young Kim, Robert McAndrew, and Dr. Ben Zheng.

I would like to thank the friends outside the Statistics Department as well. I especially would like to thank Alison Zak and the Zak family, who gave much-needed emotional support. Others who helped me through this process, and helped me take my mind off it, are Travis Engle, Mike Callahan, and Chris Large. The belief of my friends from back in Georgia – Ed Bolian, Brian French, Michael O'Leary, and Evan Wimpey – was a constant inspiration for me. I would like to thank Amanda Marrs for encouraging me to pursue my PhD.

DEDICATION

I would like to dedicate this dissertation to my father, Frank Wayne Marrs, Jr.

TABLE OF CONTENTS

Chapter 1

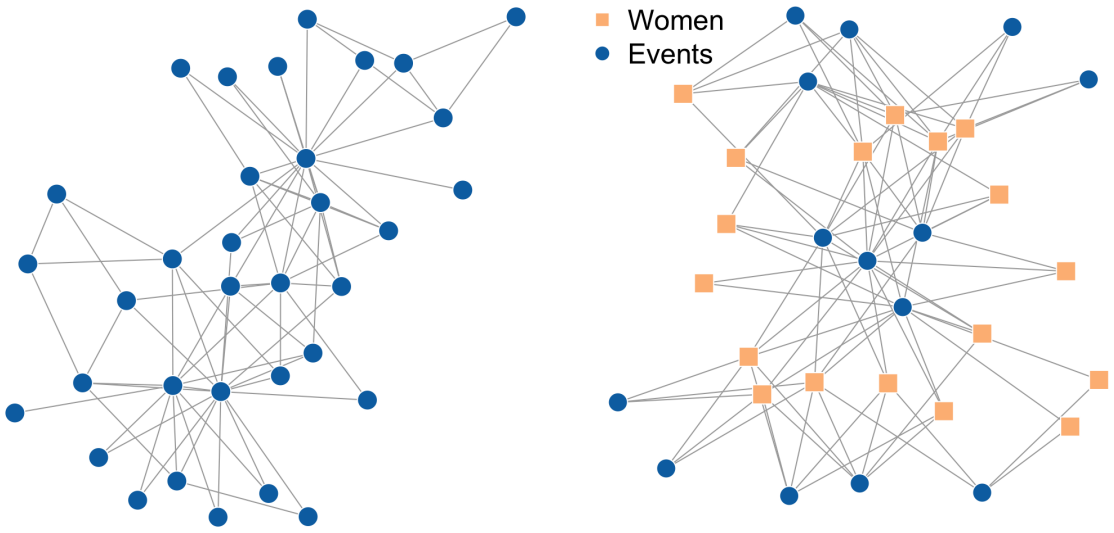
Introduction

1.1 Network data

The most pressing problems of our time, from environmental treaty legislation (Campbell et al., 2019) to human migration flows (Aleskerov et al., 2017), may be characterized by network data. Yet, the study of network data is not new. Analysis of network data dates back to sociometry studies relating social structures and sociological well-being (Moreno, 1934). Up until recently, only small and simple networks were available for analysis. With the advent of modern data-gathering techniques, the demand for network analysis tools that can answer more complex questions and be used on larger and larger networks has exploded. As evidence, numerous journals are now dedicated to the analysis of network data, such as *Network Science* and *The Journal of Complex Networks*.

Network data consist of measurements of relations between pairs of actors. A classic example of a network is a binary social network, where each relation is a binary indicator of whether two actors share a friendship. Figure 1.1 shows an example binary social network from Zachary (1977), where actors are depicted as points and lines connect pairs of actors if the two actors participated in activities together outside of a karate club, that is, a line between two actors indicates a relation between them. We may represent the Zachary karate club network of n actors as an $n \times n$ symmetric binary matrix $\{y_{ij} \in \{0, 1\} : i, j \in \{1, ..., n\}\}$, which we abbreviate $\{y_{ij}\}_{ij}$, where the relation y_{ij} is '1' if actors i and j participated in activities together outside of the karate club, and '0' otherwise. The diagonal of the matrix $\{y_{ij}\}_{ij}$ is undefined, as an actor does not participate in activities outside the karate club with him/herself.

We note that the karate club network is *undirected*, meaning that relations are symmetric between actors, and thus, the matrix $\{y_{ij}\}_{ij}$ is symmetric. We call networks where the relations are asymmetric *directed*, such as a network $\{y_{ij}\}_{ij}$ where $y_{ij} \in \{0, 1\}$ indicates whether actor i named



Zachary karate club Davis southern women

Figure 1.1: Examples of binary networks, where a line between points indicate edges that are '1', and edges that are '0' suppressed. An edge is '1' between individuals in the Zachary karate club (left, from Zachary (1977)) if both individuals shared activities outside of karate. The Davis southern women network (right, from Davis et al. (1941)) is an example of a bipartite network. Edges indicate that women attended particular events (note that edges only exist between women and events).

actor j as a close friend in a survey. We note that actor i may name actor j a close friend when actor j does not name i, and thus, $y_{ij} \neq y_{ji}$ in general in a directed network. An example of a directed network where the relations are continuous valued is a network where relations quantify the value of economic goods transported from one nation to another in a given year. Naturally, the value of goods that the US sends to China may not be equal to the value of goods that China sends to the US in a given year. We analyze both directed and undirected network data, and continuous valued and binary valued network data in this dissertation, and some of these network data consist of multiple observations over time. We do not consider networks wherein an actor has a relation with him/herself; for example, the US does not trade with the US.

A canonical problem in analyzing network data is to estimate the effects of exogenous covariates on a response that forms a network. In the karate club example, for instance, researchers might have interest in inferring whether pairs of actors with the same gender are more likely to interact outside the karate club than those with disparate gender. In this example, the covariate

x_{ij} is an indicator of whether actors i and j are the same gender. Unlike typical regression scenarios, network data naturally engender excess statistical dependence – beyond that represented by covariates – since every relation shares an actor with many other relations. For instance, club member Jason may have a higher likelihood of associating with Andrew than that explained by Jason's and Andrew's shared gender, simply due to Andrew's gregariousness. Regression models that are accurate in prediction and estimation of regression coefficients should explicitly recognize this excess network dependence. This dissertation is focused on developing methods to capture the excess dependence in regression models of network data.

A generalization of the matrix $\{y_{ij}\}_{ij}$ which we consider in this dissertation is one where the rows and columns of the matrix refer to different actor sets. An example of a network with two different actor sets is an academic citation network, where a relation exists between a researcher and a given paper whenever the researcher was listed as an author on the paper. We call these networks with two actor types *bipartite* networks. Figure 1.1 depicts another example of a bipartite network, where squares denote southern women and points denote events, and lines connect women that attended particular events (Davis et al., 1941). We may mathematically represent the bipartite network with a binary $m \times n$ matrix, $\{y_{ij} \in \{0,1\} : i \in \{1,...,m\}, j \in \{1,...,n\}\}$, such that y_{ij} is '1' if woman i attended event j, and '0' otherwise. In this Introduction, we summarize the proposed methods for accounting for excess network dependence – beyond that represented by covariates – in bipartite network data in Section 1.2.

With the advent of social networking sites, such as Facebook, and other modern sources of big network data sets, methods are needed for analyzing larger and larger networks. Such networks consist of hundreds or even thousands of actors. Note that networks of hundreds of nodes contains tens of thousands of pairwise relation observations. In this dissertation, we present novel parsimonious methods for accounting for excess network dependence in regression models for continuous and binary valued, unipartite network data, which we introduce in Section 1.3. These methods are based on an assumption of joint exchangeability, common to many random network models, which we introduce in more detail in Section 1.3.1.

1.2 Influence networks in longitudinal bipartite network data

Longitudinal bipartite relational data characterize the evolution of relations between pairs of actors of disparate types, and can be represented as a sequence of networks, one for each time point. A common goal is to understand the temporal dependencies in these data, specifically which actor relations incite later actor relations. For example, consider a longitudinal bipartite network data set of countries ratifying environmental treaties (Campbell et al., 2019). A proxy for influence of country i on country j is the ratification of a given treaty by country j the year after the same treaty is ratified by country i. Similarly, one environmental treaty may be likely to be ratified in the year after another treaty is ratified by the same country. Further, these dependencies may not be able to be fully explained by covariates about the countries or treaties, such that a standard regression model is insufficient. The year-over-year dependence between countries that ratify the same treaty and between treaties that are ratified by the same country are the excess network dependencies which we seek to model explicitly, and upon which we wish to make inference.

There are two primary existing approaches to modeling excess dependence in longitudinal bipartite network data. The first projects the bipartite data in each time period to a unipartite network and then uses methods for uniparite networks to analyze the projected networks (Newman, 2001; Barabási et al., 2002; Wu et al., 2014). Unfortunately, information is lost in calculating the projection and generative models for networks obtained through this process are scarce. The second approach explicitly models the excess network dependence using two unipartite *influence networks*, corresponding to the two actor types. In the example of countries ratifying treaties, this approach infers a network of influence among the countries and a network of influence among the treaties. The existing regression model that takes this approach is bilinear in the influence networks, creating challenges in computation and interpretation (Hoff, 2015).

In Chapter 2, we propose the Bipartite Longitudinal Influence Network (BLIN) model that permits estimation of continuous valued, directed influence networks and does not suffer from the shortcomings of the existing model. The proposed model is linear in the influence networks, permitting inference using off-the-shelf software tools. We prove our estimators of the influence

networks are consistent under cases of model misspecification and nearly asymptotically equivalent to the existing estimators in the existing bilinear model. We demonstrate the performance of the proposed model and estimators in simulation studies and an analysis of weekly international state interactions.

1.3 Regression models for network data

A primary task in many data analyses is estimation of generalized linear regression models, and this fact remains true when the data form a network. The main challenge in estimating generalized linear models when the outcomes form a network is accounting for the network dependence, beyond that ccounted for by covariates, without undue computational burden.

We focus on continuous valued network data in Chapter 3. We propose and evaluate a new class of standard error estimators for coefficients in linear regression models where the outcomes form a network. Existing estimators of parameter standard errors that recognize network dependence rely on estimating extremely complex, heterogeneous structure across actors (Fafchamps and Gubert, 2007; Aronow et al., 2015). Leveraging an exchangeability assumption on the model errors, we derive parsimonious standard error estimators for the regression coefficients that pool information across actors and are substantially more accurate than existing estimators in a variety of settings. This exchangeability assumption is pervasive in network and array models in the statistics literature, but not previously considered when adjusting for dependence in a regression setting with relational data. We briefly introduce exchangeability in Section 1.3.1. We show that our estimator outperforms the current state-of-the-art estimator in mean-square error and demonstrate improvements in inference through simulation and a data set involving international trade.

After addressing continuous valued network data, we turn to regression models for binary valued network data in Chapter 4. The binary setting is more complicated, as the mean and variance of the network generative process are no longer separable. Additionally, as binary data can be viewed as thresholded continuous data, there is inherently less information in binary network data than in continuous network data. As in the continuous case, the challenge with estimating network

regression models for binary network data is accounting for network dependence – beyond that accounted for by covariates – due to relations that share the same actor. The literature has developed a host of latent variable models to account for the inherent dependence of network data (Hoff et al., 2002; Hoff, 2008), however, estimation of these latent variable models is computationally onerous, and which model produces the best predictions or estimations may not be clear. We propose the Probit Exchangeable (PX) model for network data based on a key exchangeability assumption. The PX model can represent the second moments of any exchangeable network model, yet is specifies no particular parametric model. We propose an approximate maximum likelihood estimator for the PX model that allows for rapid estimation. Using simulation studies, we demonstrate the improvement in estimation of regression coefficients of the proposed model over existing latent variable models. In an analysis of purchases of politically-aligned books, we demonstrate that the proposed model significantly reduces runtime relative to latent variable models while maintaining predictive performance, and demonstrate political polarization in the network of book purchases.

1.3.1 Joint exchangeability of network models

The contributions of Chapters 3 and 4 rely heavily on joint exchangeability, a property that many network models share. A network model that is jointly exchangeable is one where the node labelling is uninformative to the distribution of the relational data, or equivalently, the distribution of matrix $\{y_{ij}\}_{ij}$ is invariant under simultaneous permutations of its rows and columns (Hoover, 1979; Aldous, 1981; Lovász and Szegedy, 2006). For instance, consider a binary undirected network. In the case of joint exchangeability, the probability of observing any "shape" of network is the same, regardless of the node labelling. In Figure 1.2, we depict two binary undirected networks of four actors, each of which forms a line. If a network model is jointly exchangeable, the probability of observing these networks is the same, even though the node labellings of the two networks are different.

A key result of our work is that relations in jointly exchangeable network models all have covariance matrices of the same form, which is due to the symmetries resulting from the assumption

Figure 1.2: Examples of two four actor networks of the same shape, a line, but different node labelling. Under a jointly exchangeable network model, the probability of observing either network is the same.

of joint exchangeability. If the jointly exchangeable network model is undirected, then the covariance matrix has at most three unique terms, and if the network model is directed, then the covariance matrix has at most six unique terms. We use these covariance matrices to account for the excess dependence in network data, whether the network data are continuous or binary valued, and directed or undirected.

1.4 Outline

This dissertation is organized as follows. First, we discuss a novel model for longitudinal bipartite relational data, the BLIN model, in Chapter 2. Then, we introduce a parsimonious standard error estimator for regression of continuous network data in Chapter 3. In Chapter 4, we provide an approximate maximum likelihood estimator of the proposed PX regression model for binary network data. Finally, we summarize and provide other potential directions for research in Chapter 5.

Chapter 2

Inferring influence networks from longitudinal bipartite relational data

2.1 Introduction

Longitudinal bipartite relational data are being collected at unprecedented rates to study complex phenomena in both the social and biological sciences. These data characterize the evolution of relations between pairs of actors, where each actor is one of two distinct types, and relations exist only between disparate actor types. Studies involving such data have focused on, e.g., films (Watts and Strogatz, 1998), international relations (Boulet et al., 2016; Campbell et al., 2019), metabolic interactions (Jeong et al., 2000), recommender systems (Linden et al., 2003), or transportation systems (Zhang et al., 2006). For example, in international affairs, researchers might study countries' financial contributions to international organizations over the past few decades. Here, the countries and international organizations are the actors and the relations of interest are yearly financial contributions.

In many studies of longitudinal bipartite relational data, the relevant scientific questions surround relations among actors of a single type; the set of which can be represented in a *unipartite* network. For example, researchers may be interested in the (unobserved) relationships among countries that affect the amount of financial contributions to different international organizations. The financial contribution of China to the UN, for instance, may be influenced by the US's recent announcement to cut its budget obligations for the next years. The degree of change in China's contribution based on the US could be viewed as a measure of US influence on China. Such influences may exist between international organizations as well: the contribution of the US to the United Nations (UN) can be related to its financial obligations to the Worl Trade Organization (WTO). In the examples given, the influences are occurring over time and are allowed to be asymmetric

such that, for example, China may be influenced by the US to a high degree while the US is not influenced by China. A goal then in studying longitudinal bipartite data is to infer the relationships among actors of each type. We term these sets of unipartite relations *influence networks*. For example, in the US-China illustration, the country influence network is denoted $\mathbf{A} = \{a_{ij}\}_{i,j=1}^{S}$, where S is the number of countries and a_{ij} represents the amount of influence country i has on country j. Similarly, the organization influence network is denoted $\mathbf{B} = \{b_{ij}\}_{i,j=1}^{L}$, where L is the number of international organizations and b_{ij} represents the influence of organization i on organization j. Inferring these latent influences are of substantive interest in many studies and, in the case of the states and their contributions to international organizations, have the potential to inform international policy-making in effective and previously unknown ways.

The idea of summarizing bipartite data in terms of unipartite influence networks is not new. Newman (2001) analyzes the relationships among academic authors by estimating the unipartite author-author network from data on academics and the papers they authored over five years. Newman (2001) ignores the temporal component of the data and defines the relationship a_{ij} between author i and author j as the the number of papers i and j co-authored during the five-year period. The resulting influence network is often referred to as a (one-mode) projection. If the binary data matrix of authorship is denoted by a rectangular matrix $\mathbf{Y} = \{y_{ik}\}$, where y_{ik} is an indicator of whether academic i authored paper k, then the author influence network can be expressed as $\mathbf{A} = \mathbf{Y}\mathbf{Y}^T$. Notice that $\mathbf{A}$ is symmetric by construction and represents behavioral co-occurrence (in this case, co-authorship), rather than influences over time as described earlier. Investigating temporal patterns in publications, Barabási et al. (2002) estimated yearly influence networks among academic authors using one-mode projections and analyzed the evolution of summary statistics of the yearly projections. Extensions of one-mode projections exist for longitudinal bipartite networks (Wu et al., 2014), weighted bipartite networks (Newman, 2004; Liu et al., 2009), and for creating directed influence networks (Zhou et al., 2007). These various extensions involve different weightings of the original bipartite relations (e.g., weight each paper by the number of co-authors).

Due to the plethora of tools available for unipartite networks, bipartite data are often cast into one-mode projections that can be subsequently analyzed using standard network modeling techniques (Zhou et al., 2007). Although various weighting schemes have been investigated for one-mode projections (Wu et al., 2014), a key disadvantage of this approach is that information in the original bipartite data is inherently lost in the projection, regardless of the selected weighting scheme. In addition, since the data naturally arise in a bipartite format, specifying a generative model for the projection on which to base inference is fundamentally challenging.

There exists some previous work that directly models the observed bipartite network as well. Skvoretz and Faust (1999) and Wang et al. (2009), for example, propose generative Exponential Random Graph Models (ERGMs) for bipartite networks, however this work aims to infer the effect of certain network motifs (such as triangles) on the strength of relations rather than infer the latent influence networks. Another thread of research seeks to explain network formation of heterogeneous information networks, those that consist of disparate node types *and* links, of which bipartite networks are a subset (Sun et al., 2011; Sun and Han, 2012; Shi et al., 2017a). A particularly similar line of work to ours explains network formation as a function of influence networks among node types, although these models have no temporal component and thus model simultaneous influence rather than the particular sequential influence we consider (Liu et al., 2010, 2012).

In this chapter, we propose a novel bipartite longitudinal influence network (BLIN) model, which permits inference on the influence networks for each set of actors. This work builds upon recent developments on statistical models for longitudinal unipartite relational data (Banks and Carley, 1996; Snijders, 2005; Krackhardt and Handcock, 2007; Sewell and Chen, 2015; Carnegie et al., 2015). Specifically, Almquist and Butts (2013, 2014) propose an autoregressive model for unipartite networks that may be expressed as a generalized linear model. In a similar vein, we propose an autoregressive, generalized linear model for bipartite networks, wherein the influence networks are autoregressive parameters. Although the proposed model is conceptually similar to an existing diffusion model (Desmarais et al., 2015) and a bilinear regression model (Hoff,

2015), our model has key advantages over these existing methods with regard to estimability and interpretability.

The rest of the article is organized as follows. We introduce the BLIN model in Section 2.2. We discuss approaches to modeling longitudinal bipartite data and then explore various extensions to our model. We describe maximum likelihood estimation procedures for the BLIN model in Section 2.3 and give properties of the resulting estimators in Section 2.4, including performance under misspecification. In Section 2.5, we compare the performance of our model to existing approaches in simulation studies. In Section 2.6, we demonstrate our methodology using a data set of material and verbal interactions between international states, where the disparate actor types are the source countries and the target countries of these actions (e.g. humanitarian aid, boycotting, or intent to negotiate). Finally, we discuss future work in Section 2.7.

2.2 BLIN model

Let the matrix $\mathbf{Y}_t = \{y_{ij}^t\} \in \mathbb{R}^{S \times L}$ denote the t^{th} observation of the bipartite relations among S actors of one type (e.g., countries) and L actors of a second type (e.g., international organizations), where the time index $t \in \{1, 2, \ldots, T\}$. For example, in the illustration introduced above, the (i, j) entry in $\mathbf{Y}_t$, y_{ij}^t, is the financial contribution by country i to international organization j in year t.

The BLIN model expresses the relations at time t, $\mathbf{Y}_t$, as a function of the p previous relations $\{\mathbf{Y}_k : k \in \{t-1, t-2, \ldots, t-p\}\}$, and the influence matrices $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{Y}_t = \mathbf{A}^T \sum_{k=1}^{p_A} \mathbf{Y}_{t-k} + \sum_{k=1}^{p_B} \mathbf{Y}_{t-k}\mathbf{B} + \mathbf{E}_t, \tag{2.1}$$

where $\mathbf{E}_t$ is an $S \times L$ matrix of mean zero, independent and identically distributed errors and $p = \max(p_A, p_B)$. The constants p_A and p_B represent the number of previous time periods which influence $\mathbf{Y}_t$ through the networks $\mathbf{A}$ and $\mathbf{B}$, respectively.

The BLIN model can alternatively be expressed as:

$$y_{ij}^{t} = \sum_{s=1}^{S} a_{si} \left(\sum_{k=1}^{p_A} y_{sj}^{t-k} \right) + \sum_{\ell=1}^{L} b_{\ell j} \left(\sum_{k=1}^{p_B} y_{i\ell}^{t-k} \right) + e_{ij}^{t}. \tag{2.2}$$

From this representation, it is more easily seen that y_{ij}^{t} is exclusively a function of those $y_{k\ell}^{t-k}$ when either $k = i$ or $\ell = j$, or both. Consider $p_A = p_B = 1$. In the context of the international affairs example, this means that China's financial contribution to the UN in 2015 depends on the contributions of all other countries to the UN in 2014 through entries in $\mathbf{A}$, and on China's financial contributions to all other international organizations in 2014 through the entries in $\mathbf{B}$. The interpretation of the individual $\mathbf{A}$ and $\mathbf{B}$ parameters follow from the linearity of the BLIN model. For example, a_{si} is the expected increase in financial contributions of country i to a given international institution when country s has raised its contribution by one unit to the same organization in the previous year. Similarly, the coefficient $b_{\ell j}$ is the expected increase in budget obligations to international organization j by a given country when international organization ℓ has received one unit of financial contributions from the same country in the previous year. Figure 2.1 depicts these influences using the international affairs example. A positive value of $a_{US,CHI}$ in $\mathbf{A}$, corresponding to the influence of the US on China, means that if the US increased its contributions to the UN in 2014, then China is expected to increase its expenditure to the UN in 2015. Similarly, a positive value of $b_{UN,WTO}$ in $\mathbf{B}$, corresponding to influence of financial obligations to the UN on contributions to the WTO, implies that if the US spent more money on the UN in 2014, then it is expected to increase its WTO expenditure in 2015. Since the influence matrices are time invariant, the entries in $\mathbf{A}$ and $\mathbf{B}$ represent the average influence over all time periods under consideration.

Figure 2.1 depicts types of direct influence patterns the BLIN model captures (i.e. those between y_{ij}^{t} and $y_{k\ell}^{t-1}$ where $i = k$ and/or $j = \ell$). Note, however, that secondary influences (such as that between y_{ij}^{t} and $y_{k\ell}^{t-s}$ where $i \neq k$, and $j \neq \ell$) may propagate through the BLIN model over multiple time periods, i.e. for $s > 1$. For example, although US contributions to the UN in 2014 may not affect China's contribution to the WTO in 2015, it may do so in 2016. This may occur if,

say, US financial transfers to the UN in 2014 affect China's UN expenditure in 2015 via a nonzero value of $a_{US,\,CHI}$. Then, China's contribution to the UN in 2015 impact its own contribution to the WTO in 2016 through a nonzero value $b_{UN,\,WTO}$. In this way, the BLIN model allows both direct and secondary influences through different mechanisms.

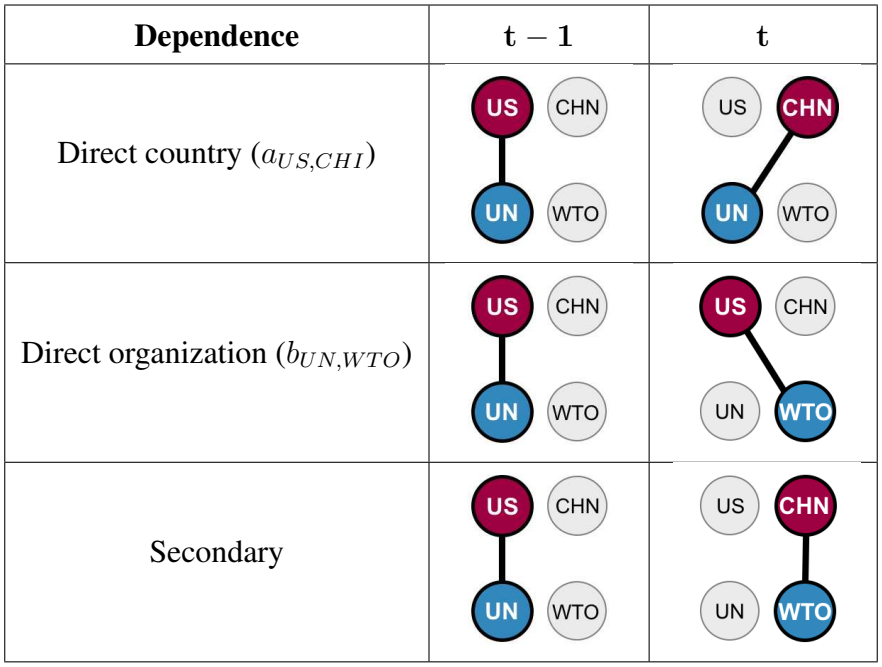


Figure 2.1: Influence types in longitudinal bipartite relational data in the parlance of the country/international organization example for two countries and two institutions when $lag = 1$. Dark red nodes represent countries (US and China) and light blue nodes represent international organizations (UN and WTO).

A key flexibility of the BLIN model is that it allows for $p_A \neq p_B$; that is, $\mathbf{A}$ and $\mathbf{B}$ may represent influences over differing time scales. This is natural. For example, in the international affairs example, country contributions may be influenced only by other countries' contributions in the past year ($p_A = 1$). However, the US may have long-standing pattern of contributions to the UN and WTO such that the dependence through $\mathbf{B}$ is much longer, say $p_B = 5$.

A key property of the BLIN model is that it may be written as a linear model. Letting $\mathbf{y}_t$ and $\mathbf{e}_t$ denote the column-wise vectorization of matrices $\mathbf{Y}_t$ and $\mathbf{E}_t$, respectively, the model in (2.2) can alternatively be expressed

$$\mathbf{y}_t = \left(\sum_{k=1}^{p_A} \mathbf{Y}_{t-k}^T \otimes \mathbf{I}_S \right) \mathrm{vec}(\mathbf{A}^T) + \left(\mathbf{I}_L \otimes \sum_{k=1}^{p_B} \mathbf{Y}_{t-k} \right) \mathrm{vec}(\mathbf{B}) + \mathbf{e}_t, \tag{2.3}$$

$$:= \mathbb{X}_B^{(t)} \boldsymbol{\theta} + \mathbf{e}_t, \tag{2.4}$$

where '$\otimes$' is the Kronecker product and $\mathrm{vec}(\mathbf{B})$ denotes the column-wise vectorization of matrix $\mathbf{B}$. In the second line, $\mathbb{X}_B^{(t)}$ is the $SL \times (S^2 + L^2)$ matrix $[\sum_{k=1}^{p_A} \mathbf{Y}_{t-k}^T \otimes \mathbf{I}_S,\ \mathbf{I}_L \otimes \sum_{k=1}^{p_B} \mathbf{Y}_{t-k}]$ and $\boldsymbol{\theta}^T = [\mathrm{vec}(\mathbf{A}^T)^T,\ \mathrm{vec}(\mathbf{B})^T]$ is the vector of parameters. Since the BLIN model may be written as a linear model, numerous off-the-shelf tools exist for estimation (including regularization) of the BLIN model, making inference on the influence networks straightforward. In what follows, we focus on the model without covariates, although we may easily incorporate covariate information by adding the term $\mathbf{W}_t\boldsymbol{\beta}$ to the right hand side of (2.3), for $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\mathbf{W}_t \in \mathbb{R}^{SL \times p}$. Thus, we assume throughout the chapter that $\mathbf{Y}_t$ is mean zero for all $t \in \{1, 2, \ldots, T\}$ without loss of generality.

Another useful representation of the BLIN model is as a vector autoregressive (VAR) model, a generalization of the univariate autoregressive model (Sims, 1980). Letting $\boldsymbol{\Theta}_1 = \mathbf{I}_L \otimes \mathbf{A}^T + \mathbf{B}^T \otimes \mathbf{I}_S$ and $\boldsymbol{\Theta}_2 = \mathbb{1}_{[p_A > p_B]} \left(\mathbf{I}_L \otimes \mathbf{A}^T \right) + \mathbb{1}_{[p_B > p_A]} \left(\mathbf{B}^T \otimes \mathbf{I}_S \right)$, the BLIN model in (2.1) may be rewritten

$$\mathbf{y}_t = \boldsymbol{\Theta}_1 \left(\sum_{k=1}^{q} \mathbf{y}_{t-k} \right) + \boldsymbol{\Theta}_2 \left(\sum_{k=q+1}^{p} \mathbf{y}_{t-k} \right) + \mathbf{e}_t, \tag{2.5}$$

where $q = \min(p_A, p_B)$. We note that, when $p_A = p_B = p = 1$, the VAR representation of the BLIN model reduces to $\mathbf{y}_t = \boldsymbol{\Theta}_1 \mathbf{y}_{t-1} + \mathbf{e}_t$, the standard lag-1 VAR model. An unstructured coefficient matrix $\boldsymbol{\Theta}_1$ has S^2L^2 unknown parameters, while the BLIN $\boldsymbol{\Theta}_1$ has only $S^2 + L^2$ unknowns. In this light, the BLIN model may be viewed as reducing the number of unknown parameters in the coefficient matrix $\boldsymbol{\Theta}_1$ by imposing interpretable bipartite structure on $\boldsymbol{\Theta}_1$.

The BLIN model in (2.1) is not fully identifiable such that for any $c \in \mathbb{R}$, the transformation $\{\mathbf{A}, \mathbf{B}\} \rightarrow \{\mathbf{A} + c\mathbf{I}_S,\ \mathbf{B} - c\mathbf{I}_L\}$ results in the *exact* same model for the data $\mathbf{Y}_t$. This non-identifiability means that we are unable to determine a_{ii} and b_{jj} separately, but that the sum $a_{ii} + b_{jj}$

is identifiable. Specifically, we may estimate the effect of y_{ij}^{t-1} on y_{ij}^{t}, but we cannot decompose this effect into the contribution of country i and organization j. Nevertheless, we may compare the marginal country effect among countries and among international organizations, respectively. For example, although the absolute values of $a_{US,US}$ and $a_{CHI,CHI}$ are not identifiable, their difference $a_{US,US} - a_{CHI,CHI}$ is identifiable. If this difference is positive, we may conclude, for example, that a US increase in financial contributions given a unit expenditure in the previous year is higher than China's increase in contributions.

2.2.1 Comparison to existing approaches

Diffusion models (e.g., Berry and Berry, 1990) are popular for studying the interdependencies of institutions in political science (Desmarais et al., 2015). In these models, an outside institution puts transmission pressure on for a particular policy on the focal institution, making the latter more likely to adopt that policy. The network of these transmissions forms a directed tree, where there is at most one path from one institution to another. The diffusion model is distinct from the approach we propose in several ways. In the parlance of the international affairs example, the former supposes that each country's financial contribution to a specific international organization is influenced by at most a single other country. In addition, a binary network is inferred, rather than a weighted network which can encode both positive and negative influences. Furthermore, methods for quantifying uncertainty in the estimated network and incorporating covariates are unavailable.

Hoff (2015) proposes a generative model for bipartite longitudinal data termed the bilinear model, which can be expressed

$$\mathbf{Y}_t = \mathbf{A}^T \sum_{k=1}^{p} \mathbf{Y}_{t-k}\mathbf{B} + \mathbf{E}_t. \tag{2.6}$$

Hoff (2015) presents an estimator that proceeds by alternating estimates of $\mathbf{A}$ and $\mathbf{B}$, however this estimator is guaranteed to converge only to a local optimum. Thus, global optimality of the existing estimator is not guaranteed. We illustrate this issue in simulation studies (Section 2.5).

Also, unlike the BLIN model, since the bilinear model is nonlinear, many standard off-the-shelf tools for regularization and uncertainty quantification are not applicable.

The matrices $\mathbf{A}$ and $\mathbf{B}$ in the bilinear model, as in the BLIN model, measure actor influences. However, the bilinear model combines the direct and secondary dependencies in Figure 2.1 into the same mechanism. This results in a different interpretation of the parameters. To illustrate the interpretation, we rewrite the bilinear model in (2.6) as

$$y_{ij}^t = \sum_{s=1}^{S} \sum_{\ell=1}^{L} a_{si} b_{\ell j} \left(\sum_{k=1}^{p} y_{s\ell}^{t-k} \right) + e_{ij}^t. \tag{2.7}$$

Here we see $\mathbf{Y}_t$ depends on *every* entry in $\mathbf{Y}_{t-1}$, as y_{ij}^t may be affected by both y_{sj}^{t-1} (direct) and $y_{s\ell}^{t-1}$ (secondary) through a_{si}. A consequence of the multiplicative nature of the model is that the influence parameters must be interpreted in conjunction with one another. For example, $a_{US,\,CHI}$ represents the expected increase in Chinese contributions to international institution k *for each* b_{jk} *unit* of expenditure of the US to organization j in the previous year. Thus, the interpretations of the country influences in $\mathbf{A}$ and the international-organization influences in $\mathbf{B}$ are intertwined. While there may be instances where $y_{s\ell}^t$ is influenced by y_{kj}^{t-1}, we argue that secondary dependencies are often likely of a smaller magnitude than the direct dependencies. In these cases, it would be undesirable to use, for example, a single parameter a_{ik} to simultaneously capture direct and secondary influences. The BLIN model assumes secondary dependencies are zero and focuses estimation on the direct dependencies.

The influence matrices $\mathbf{A}$ and $\mathbf{B}$ in the bilinear model are identifiable up to a multiplicative constant. For any $c \in \mathbb{R}$, the transformation $\{\mathbf{A}, \mathbf{B}\} \rightarrow \{c\mathbf{A}, \mathbf{B}/c\}$ leaves the model for $\mathbf{Y}_t$ invariant. This implies that the relative scales of the networks represented by $\mathbf{A}$ and $\mathbf{B}$ and the signs of the elements are not estimable. However, the ratio of elements within each influence network is identifiable, e.g. $a_{US,CHN}/a_{US,UK}$.

2.2.2 Extensions of the BLIN model

In this section, we discuss various extensions of the BLIN model. First, in the definition of the BLIN model in (2.1), the first observation of the bipartite relations $\mathbf{Y}_t$ is a function of the p past observations $\{\mathbf{Y}_{t-k}\}_{k=1}^{p}$. For simplicity, each past observations affects the current observation equally. Obviously, more complex dependence functions may be considered, such as an exponentially decaying influence of the past time periods on the current time period. In general, $\sum_{k=1}^{p_A} \mathbf{Y}_{t-k}$ and $\sum_{k=1}^{p_B} \mathbf{Y}_{t-k}$ in (2.1) may be replaced by any $S \times L$ matrix function of the past observations, $f(\mathbf{Y}_{t-})$, where $\mathbf{Y}_{t-} := \{\mathbf{Y}_{t-k}\}_{k=1}^{p}$ and again $p = \max(p_A, p_B)$. The selection and estimation of such functions is a current area of research: see Krackhardt and Handcock (2007); Krivitsky (2009); Hanneke et al. (2010); Almquist and Butts (2014) for a discussion of general unipartite temporal network models and autoregressive models for unipartite temporal networks.

The linear nature of the BLIN model simplifies its extension to other types of outcomes, e.g. binary or count observations. Let y_{ij}^t be a general measure of the relation between actors i and j at time t. Then, a general BLIN model may be expressed

$$g(E[\mathbf{Y}_t|\mathbf{Y}_{t-}]) = \mathbf{A}^T \sum_{k=1}^{p_A} \mathbf{Y}_{t-k} + \sum_{k=1}^{p_B} \mathbf{Y}_{t-k}\mathbf{B}, \tag{2.8}$$

where $g(.)$ is an appropriate link function based on the form of $\mathbf{Y}_t$ (McCullagh and Nelder, 1989). Off-the-shelf tools are again available for estimation of the model in (2.8) if g is a standard link function. When g is the canonical link function for logistic regression, for example, the BLIN model in (2.8) may be viewed as a conditional logistic discrete choice model (McFadden, 1973); models of this type have recently been employed in network representations (Overgoor et al., 2018).

2.3 Estimation of the BLIN model

In the following, we discuss several estimation procedures for the BLIN model. First, we propose an estimator that results from minimizing a least squares criterion. We then consider more

parsimonious estimators using sparsity-inducing penalties and reduced-rank approaches. For ease of notation, we define the regressor matrices in (2.1) as $\mathbf{X}_t := \sum_{k=1}^{p_A} \mathbf{Y}_{t-k}$ and $\mathbf{Z}_t := \sum_{k=1}^{p_B} \mathbf{Y}_{t-k}$, such that the BLIN model can be expressed

$$\mathbf{Y}_t = \mathbf{A}^T\mathbf{X}_t + \mathbf{Z}_t\mathbf{B} + \mathbf{E}_t. \tag{2.9}$$

In the theory that follows in this section and the next, we treat $\mathbf{X}_t$ and $\mathbf{Z}_t$ generally as, in principle, they may be any sequence of matrices of appropriate size.

2.3.1 Least squares estimator

Based on the vector representation of the BLIN model in (2.3), we propose minimizing the following least squares criterion to construct an estimator for the $\mathbf{A}$ and $\mathbf{B}$ matrices:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\mathbf{y} - \mathbb{X}_B\boldsymbol{\theta})^T(\mathbf{y} - \mathbb{X}_B\boldsymbol{\theta}), \tag{2.10}$$

$$= \underset{\{\mathbf{A},\mathbf{B}\}}{\operatorname{argmin}} \sum_t ||\mathbf{Y}_t - \mathbf{A}^T\mathbf{X}_t - \mathbf{Z}_t\mathbf{B}||_F^2 \tag{2.11}$$

where $\mathbf{y}^T := [\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_T^T]$ such that $\mathbf{y} \in \mathbb{R}^{SLT}$ and $\mathbb{X}_B \in \mathbb{R}^{SLT\times(S^2+L^2)}$ is the column-wise stacking of the design matrices $\{\mathbb{X}_B^{(t)}\}_{t=1}^T$ in the vector representation of the BLIN model in (2.4). The explicit solution to (2.10) is

$$\begin{bmatrix} \operatorname{vec}\left(\widehat{\mathbf{A}}^T\right) \\ \operatorname{vec}\left(\widehat{\mathbf{B}}\right) \end{bmatrix} = \begin{bmatrix} \left(\sum_{t=1}^T \mathbf{X}_t\mathbf{X}_t^T\right) \otimes \mathbf{I}_S & \sum_{t=1}^T \mathbf{X}_t \otimes \mathbf{Z}_t \\ \sum_{t=1}^T \mathbf{X}_t^T \otimes \mathbf{Z}_t^T & \mathbf{I}_L \otimes \left(\sum_{t=1}^T \mathbf{Z}_t^T\mathbf{Z}_t\right) \end{bmatrix}^- \begin{bmatrix} \operatorname{vec}\left(\sum_{t=1}^T \mathbf{Y}_t\mathbf{X}_t^T\right) \\ \operatorname{vec}\left(\sum_{t=1}^T \mathbf{Z}_t^T\mathbf{Y}_t\right) \end{bmatrix}, \tag{2.12}$$

where $\mathbf{H}^-$ denotes the generalized inverse of square matrix $\mathbf{H}$.

Computing the solution in (2.12) requires inversion of a matrix of dimension $S^2 + L^2$. Using an iterative algorithm to solve (2.10), this computation can be replaced by repeated inversions of square matrices of dimensions S and L. Specifically, Algorithm 1 details a block coordinate descent procedure, which alternates between solving for $\mathbf{A}$ and $\mathbf{B}$. Particularly, the iteration scheme reduces memory demand and reduces the complexity of computation from $O((S^2+L^2)^3)$ in (2.12)

to $O(T \cdot \max(S, L)^3)$. In the data analysis in Section 2.6 (with $S = L = 25$ and $T = 543$), we find the iteration scheme advantageous over building the design matrix $\mathbb{X}_B$. The estimator in Algorithm 1 converges to a unique minimum when $\left(\sum_t \mathbf{Z}_t^T \mathbf{Z}_t\right)$ and $\left(\sum_t \mathbf{X}_t \mathbf{X}_t^T\right)$ are full rank.

Algorithm 1 Block coordinate descent estimation of BLIN model

0. Set threshold for convergence η. Set number of iterations $\nu = 1$. Initialize $\widehat{\mathbf{A}}^{(0)} = \mathbf{I}_S$, $\widehat{\mathbf{B}}^{(0)} = \mathbf{I}_L$ and $Q_0 = \sum_t ||\mathbf{Y}_t||_F^2$.
1. Compute $\widehat{\mathbf{B}}^{(\nu)} = \left(\sum_t \mathbf{Z}_t^T \mathbf{Z}_t\right)^{-1} \left(\sum_t \mathbf{Z}_t^T \widetilde{\mathbf{Y}}_t^{(A)}\right)$, where $\widetilde{\mathbf{Y}}_t^{(A)} := \mathbf{Y}_t - (\widehat{\mathbf{A}}^{(\nu-1)})^T \mathbf{X}_t$ for all t.
2. Compute $(\widehat{\mathbf{A}}^{(\nu)})^T = \left(\sum_t (\widetilde{\mathbf{Y}}_t^{(B)})^T \mathbf{X}_t\right) \left(\sum_t \mathbf{X}_t \mathbf{X}_t^T\right)^{-1}$, where $\widetilde{\mathbf{Y}}_t^{(B)} := \mathbf{Y}_t - \mathbf{Z}_t \widehat{\mathbf{B}}^{(\nu)}$ for all t.
3. Compute the least squares criterion $Q_\nu = \sum_t ||\mathbf{Y}_t - (\widehat{\mathbf{A}}^{(\nu)})^T \mathbf{X}_t - \mathbf{Z}_t \widehat{\mathbf{B}}^{(\nu)}||_F^2$. If $|Q_\nu - Q_{\nu-1}| > \eta$, increment ν and return to 1.

2.3.2 Sparse coefficients

Although influence may be multifactorial, it is easy to imagine scenarios where many entries in $\mathbf{A}$ and $\mathbf{B}$ are small or zero. In the international affairs example, democracies may only influence other democracies, or organizations dealing with issues at a global level may only be influenced by other global institutions. To leverage this fact in parameter estimation, we propose augmenting the least-squares criterion of the of the BLIN model in (2.10) with a sparsity-inducing penalty. Here, we consider the Lasso penalty (Tibshirani, 1996), which uses an L^1 norm on $\boldsymbol{\theta}$ to simultaneously perform variable selection and regularization. We term this model the *sparse BLIN* model as elements of $\widehat{\boldsymbol{\theta}}$ are forced to the zero. The estimation objective function is

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\mathbf{y} - \mathbb{X}_B \boldsymbol{\theta})^T (\mathbf{y} - \mathbb{X}_B \boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1, \tag{2.13}$$

where λ is a tuning parameter, and larger values of λ correspond to more regularization. Since the BLIN model is linear, the vector of parameters $\boldsymbol{\theta}$ may be regularized with any of a host of existing penalty terms (Hoerl and Kennard, 1970; Friedman et al., 2001).

We note that, as presented in (2.13), the sparse BLIN estimator in loses some of the scalability of the full BLIN estimator presented in Algorithm 1. However, it may be possible to implement an estimator of the sparse BLIN model in the spirit of Algorithm 1, that is, with alternating updates of $\mathbf{A}$ and $\mathbf{B}$. As we find no issue estimating (2.13) in the data analysis in Section 2.6 with $S = L = 25$ and $T = 543$, we leave this implementation for future work.

2.3.3 Reduced-rank coefficients

Thus far we discussed sparsity-inducing penalties for the vector of regression model coefficients $\boldsymbol{\theta}$. However, several other penalties on the singular value decomposition (SVD) of coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ have been proposed. These penalties result in coefficient estimates of $\mathbf{A}$ and $\mathbf{B}$ with reduced-rank, or approximately reduced-rank. Yuan et al. (2007) propose a nuclear norm penalty, which is an L^1 penalty of the singular values of $\mathbf{A}$. Similarly, Bunea et al. (2011) recommend a rank selection criteria penalty that is proportional to the rank of $\mathbf{A}$, i.e. a L^0 penalty on the singular values $\mathbf{A}$. This second approach provides simultaneous shrinkage on $\mathbf{A}$ and consistent estimation of its (reduced) rank.

Here we consider estimation of reduced-rank $\mathbf{A}$ and $\mathbf{B}$, i.e., $\text{rank}(A) = k < S$ and $\text{rank}(B) = m < L$. This assumption may be appropriate when there is lower-dimensional structure inherent in $\mathbf{A}$ and $\mathbf{B}$: for example, the influences among countries may be grouped by region. This approach is employed in reduced-rank regression, first developed by Anderson (1951), and has connections to principal component analysis, as shown in Izenman (1975). For any ranks $k < S$ and $m < L$ of $\mathbf{A}$ and $\mathbf{B}$, respectively, we may define a *reduced-rank BLIN model* by writing $\mathbf{A}^T = \mathbf{U}\mathbf{V}^T$ for $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{S\times k}$ and $\mathbf{B} = \mathbf{R}\mathbf{S}^T$ for $\mathbf{R}, \mathbf{S} \in \mathbb{R}^{L\times m}$. These decompositions are not identifiable up to a full-rank transformation of the decompositions, e.g. $\{\mathbf{U}, \mathbf{V}\}$ and $\{\mathbf{U}\mathbf{G}, \mathbf{G}^{-1}\mathbf{V}\}$ result in the

same influence matrix $\mathbf{A}$ for any invertible matrix $\mathbf{G}$. However, the estimands $\mathbf{A}$ and $\mathbf{B}$ remain identifiable up to an additive constant along the diagonal, as discussed in Section 2.2.

We estimate a reduced-rank BLIN model by minimizing the least-squares criteria in (2.11) with $\mathbf{A}$ replaced by $\mathbf{U}\mathbf{V}^T$, $\mathbf{B}$ replaced by $\mathbf{R}\mathbf{S}^T$, and minimizing over $\{\mathbf{U}, \mathbf{V}, \mathbf{R}, \mathbf{S}\}$. This optimization problem is easily solved using a block coordinate descent algorithm similar to Algorithm 1 for the full BLIN model and with similar computational complexity (see Appendix A.1 for details). In what follows, we refer to the BLIN model with no constraints on the parameters $\mathbf{A}$ and $\mathbf{B}$ as the *full BLIN* model, in order to distinguish it from the sparse and reduced-rank versions.

2.4 Estimator properties

This section examines properties of the least squares estimators of the full and reduced-rank BLIN models (all proofs are provided in Appendix A.2). We show that the least squares estimators are unique for a relatively small number of observations T and give some sufficient conditions for their asymptotic normality and efficiency. We then examine the properties of the least squares estimators under misspecification, providing sufficient conditions for their consistency. As in the previous section, we use the representation of the BLIN model in (2.9).

2.4.1 Uniqueness and efficiency of least squares estimators

When $\mathbf{X}_t$ and $\mathbf{Z}_t$ in (2.9) represent past observations of $\mathbf{Y}_t$, the BLIN model is a VAR model as discussed following (2.5). For the estimators in Section 2.2.2 to be consistent, stationarity of the time series is required (Brockwell et al., 1991). When $p = 1$, a sufficient condition for stationarity is that the eigenvalues of $\mathbf{\Theta}_1$ in (2.5) all have modulus less than one. To define a sufficient condition for stationarity in the general case of $p > 1$ and $p_A \neq p_B$, we rewrite the VAR version of the BLIN model in (2.5) in its companion form (Zivot and Wang, 2006) by augmenting the vector of observations at t with the p previous observations, $\boldsymbol{\xi}_t^T = [\mathbf{y}_t^T, \mathbf{y}_{t-1}^T, \ldots, \mathbf{y}_{y-p}^T]$, and augmenting the error vector at time t with an appropriate number of zeros, $\mathbf{v}_t^T = [\mathbf{e}_t^T, \mathbf{0}^T, \ldots, \mathbf{0}^T]$. Then, the BLIN model can be expressed

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t, \qquad \mathbf{F} := \begin{bmatrix} \mathbf{1}_q^T \otimes \boldsymbol{\Theta}_1 \ ; \ \mathbf{1}_{p-q}^T \otimes \boldsymbol{\Theta}_2 \\ \mathbf{I}_{SL(p-1)} \ ; \ \mathbf{0} \end{bmatrix}, \tag{2.14}$$

where $q = \min(p_A, p_B)$ and $\mathbf{1}_n$ is the vector of n ones. Then, a sufficient condition for stationarity of the BLIN model is that moduli of the eigenvalues of the companion matrix $\mathbf{F}$ are all less than one (Zivot and Wang, 2006).

The optimization problem described in (2.10) is convex, and under stationarity, by the Gauss-Markov theorem, it has a unique solution whenever $\mathbb{X}_B$ is full rank (see, e.g., Graybill, 1976). Due to the non-identifiability of the diagonal entries of $\mathbf{A}$ and $\mathbf{B}$, $\mathbb{X}_B$ is never full rank. However, the projection of $\mathbf{y}$ onto the column space of $\mathbb{X}_B$, i.e., $\widehat{\mathbf{y}} = \mathbb{X}_B\widehat{\boldsymbol{\theta}}$, is always unique. Thus, when the column space of $\mathbb{X}_B$ spans the space of possible $\mathbf{A}$ and $\mathbf{B}$ matrices up to their non-identifiability, the BLIN estimator in (2.12) is unique (up to the non-identifiability properties). This is true when the rank of $\mathbb{X}_B$ is maximized, that is one less than the number of columns: $\text{rank}(\mathbb{X}_B) = S^2+L^2-1$. The '$-1$' results from the fact that if a single diagonal entry in $\{a_{ii}\}_{i=1}^S$ or $\{b_{jj}\}_{j=1}^L$ is known, then the rest of the diagonal entries are identifiable. We now provide a proposition that states conditions under which $\mathbb{X}_B$ is maximal rank; these conditions are satisfied with probability one when, e.g., $\{\mathbf{X}_t\}_{t=1}^T$ and $\{\mathbf{Z}_t\}_{t=1}^T$ are distributed array normal.

Proposition 1. *Without loss of generality, take $S \leq L$. Assume that the $TS \times L$ matrices formed by the column-wise concatenation $[\mathbf{X}_1; \mathbf{X}_2; \ldots; \mathbf{X}_t]$ and $[\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_t]$ are full rank. Then, the design matrix has* $\text{rank}(\mathbb{X}_B) = \min(TSL, S^2 + L^2 - 1)$.

A consequence of Prop. 1 is that the full BLIN model has a unique solution when $TSL \geq S^2 + L^2 - 1$. A key implication of this is that a unique solution exists for relatively small T. For instance, if $S = L$, then the BLIN model has a unique solution (modulo the non-identifiability) when $T = 2$. Figure 2.2 plots values of S, L, and T for which the full BLIN model has a unique solution. As T grows, the space of values for which the BLIN model has a unique solution rapidly spans all values of S and L, except when their values are extremely disparate.

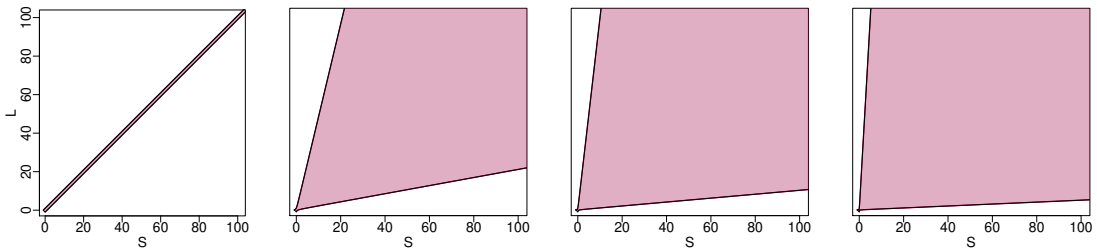


Figure 2.2: Dimensions of square matrices $\mathbf{A}$ and $\mathbf{B}$, S and L respectively, when the BLIN model has a unique solution. Shaded areas denote unique solutions.

Under the BLIN model in (2.1) when $\mathbf{X}_t$ is independent $\mathbf{E}_t$ and the true coefficients have companion matrix $\mathbf{F}$ in (2.14) with eigenvalues in the unit circle, the Gauss-Markov theorem (Graybill, 1976) states that the BLIN estimator $u^T\widehat{\boldsymbol{\theta}}$, where $\widehat{\boldsymbol{\theta}}$ is any solution to (2.10) and $u^T\boldsymbol{\theta}$ defines any identifiable linear combination of $\boldsymbol{\theta}$, is the best linear unbiased estimator (BLUE) of $u^T\boldsymbol{\theta}$. Additionally, these estimates are asymptotically normal at rate $\sqrt{T}$. Finally, we note that the estimator in (2.10) is the maximum likelihood estimator when $\mathbf{y}$ is normally distributed with homogeneous variance. Thus, under regularity conditions (Lehmann and Casella, 2006), the limiting normal random variable for $u^T\widehat{\boldsymbol{\theta}}$, has minimum asymptotic variance.

Recall that the least squares estimates of the reduced-rank BLIN model can be obtained using an iterative block-coordinate descent algorithm (see Appendix A.1 for details). When every matrix inverse in the update equations is unique, the estimates $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ from this algorithm converge to a local minimum, which is unique up to non-identifiabilities in $\mathbf{A}$ and $\mathbf{B}$ provided that $\mathrm{rank}(\mathbb{X}_B) = S^2 + L^2 - 1$. The reduced-rank estimators are the maximum likelihood estimators for $\mathbf{A}$ and $\mathbf{B}$ with ranks k and m, respectively, under the assumptions of normally distributed independent errors $\mathbf{E}_t$ with homogeneous variance. Therefore, under these conditions, $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ resulting from the iterative block-coordinate descent algorithm are consistent and asymptotically normal with minimum asymptotic variance.

2.4.2 Least squares estimator properties under misspecification

Thus far, we have discussed the attractive properties of the least squares estimators of the BLIN model, $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$, under the assumption that the model is correctly specified. It is useful to determine the limiting values of the estimators when the model is misspecified. The limiting values of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$, which we denote $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$, respectively, are referred to as pseudo-true parameters in the model misspecification literature, first investigated by Huber (1967). The pseudo-true parameters, by definition, are

$$\{\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\} = \underset{\{\mathbf{A},\mathbf{B}\}}{\operatorname{argmin}} \; E\left[\sum_{t=1}^{T} ||\mu(\mathbf{X}_t, \mathbf{Z}_t) - \mathbf{A}^T\mathbf{X}_t - \mathbf{Z}_t\mathbf{B}||_F^2\right], \tag{2.15}$$

where $\mathbf{Y}_t$ has expectation $E[\mathbf{Y}_t|\mathbf{X}_t, \mathbf{Z}_t] = \mu(\mathbf{X}_t, \mathbf{Z}_t)$, some general function of $\mathbf{X}_t$ and $\mathbf{Z}_t$, and the expectation in (2.15) is over $\mathbf{X}_t$ and $\mathbf{Z}_t$. Note that this expression holds regardless of the distribution of $\mathbf{Y}_t$, $\mathbf{X}_t$, and $\mathbf{Z}_t$ and the form of $\mu(\mathbf{X}_t, \mathbf{Z}_t)$. We note that the limiting values $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ minimize the Kullback-Leibler divergence from the true distribution of $\mathbf{Y}_t$ to the distribution of $\mathbf{Y}_t$ under the BLIN model with Gaussian errors.

In this section, we assume that the variance-covariance matrices of $\mathbf{x}_t := \text{vec}(\mathbf{X}_t)$ and $\mathbf{z}_t := \text{vec}(\mathbf{Z}_t)$ are Kronecker-structured, that is $E[\mathbf{x}_t\mathbf{x}_t^T] = \mathbf{\Omega}_X \otimes \mathbf{\Psi}_X$ and $E[\mathbf{z}_t\mathbf{z}_t^T] = \mathbf{\Omega}_Z \otimes \mathbf{\Psi}_Z$ for $\mathbf{X}_t$ and $\mathbf{Z}_t$ mean zero. This is the case if, for example, $\mathbf{X}_t$ and $\mathbf{Z}_t$ are distributed matrix normal (Gupta and Nagar, 2000). This assumption is consistent with the theoretical treatment of the bilinear model in Hoff (2015). We also assume that the errors are additive. Below, we formalize the conditions for the theory to follow.

Conditions 1.

1. *Each $\{\mathbf{X}_t\}_{t=1}^T$ is identically distributed with mean zero and $E[\mathbf{x}_t\mathbf{x}_t^T] = \mathbf{\Omega}_X \otimes \mathbf{\Psi}_X$, and each $\{\mathbf{Z}_t\}_{t=1}^T$ is identically distributed with mean zero and $E[\mathbf{z}_t\mathbf{z}_t^T] = \mathbf{\Omega}_Z \otimes \mathbf{\Psi}_Z$, for positive-definite matrices $\mathbf{\Omega}_X, \mathbf{\Omega}_Z \in \mathbb{R}^{L\times L}$ and $\mathbf{\Psi}_X, \mathbf{\Psi}_Z \in \mathbb{R}^{S\times S}$ with finite entries.*

2. *For all* $t \in \{1, 2, \ldots, T\}$, $\mathbf{Y}_t = \mu(\mathbf{X}_t, \mathbf{Z}_t) + \mathbf{E}_t$, *where* $\mu(\mathbf{X}_t, \mathbf{Z}_t)$ *is a general function of* $\mathbf{X}_t$ *and* $\mathbf{Z}_t$, *and every entry in* $\mathbf{E}_t$ *is independent and identically distributed with homogeneous, finite variance.*

3. *The sequence of* $\{\mathbf{Y}_t\}_{t=1}^T$ *is weakly stationary.*

In the remainder of this section, we examine some properties of the pseudo-true parameters (all proofs are provided in Appendix A.2). The (i, j) entry in $\widetilde{\mathbf{A}}$, denoted $\widetilde{a}_{ij}$, estimates the linear relationship between row i of $\mathbf{X}_t$ and row j of $\mathbf{Y}_t$ across all time. Similarly, the (i, j) entry in $\widetilde{\mathbf{B}}$, denoted $\widetilde{b}_{ij}$, estimates the linear relationship between column i of $\mathbf{Z}_t$ and column j of $\mathbf{Y}_t$ across all time. The first proposition states that when there is no linear relationship, the appropriate pseudo-true parameter is zero. We denote row i of $\mathbf{X}_t$ as $\mathbf{x}_{i\cdot t}$ and column j of $\mathbf{X}_t$ as $\mathbf{x}_{\cdot jt}$, and do the same for $\mathbf{Y}_t$ and $\mathbf{Z}_t$.

Proposition 2. *Under Conditions 1, if* $\mathbf{\Omega}_X$ *is diagonal and* $E[\mathbf{y}_{j\cdot t}^T \mathbf{x}_{i\cdot t}] = 0$ *for all* t, *then* $\widetilde{a}_{ij} = 0$. *Alternatively, if* $\mathbf{\Psi}_Z$ *is diagonal and* $E[\mathbf{z}_{\cdot it}^T \mathbf{y}_{\cdot jt}] = 0$ *for all* t, *then* $\widetilde{b}_{ij} = 0$.

In the setting of the international policy example, Proposition 2 states that if country i's expenditure is uncorrelated with country j's contribution to the same organization in the following year, then the least squares estimator of a_{ij} in the BLIN model will converge to $\widetilde{a}_{ij} = 0$. The following proposition provides alternative conditions under which the pseudo-true parameters are equal to zero. It allows for more general covariance structure at the cost of assuming that the conditional mean of $\mathbf{y}_t$ is linear in $\mathbf{x}_t$.

Proposition 3. *Assume that there exists a linear relationship* $E\left[\mathbf{y}_t | \mathbf{x}_t, \mathbf{z}_t\right] = \mathbf{\Theta}_X \mathbf{x}_t + \mathbf{\Theta}_Z \mathbf{z}_t$ *for all* t *and Conditions 1 hold. If all entries in* $\mathbf{\Theta}_X$ *relating row* i *in* $\mathbf{X}_t$ *to row* j *of* $\mathbf{Y}_t$ *are zero and* $i \neq j$, *then* $\widetilde{a}_{ij} = 0$. *If all entries in* $\mathbf{\Theta}_Z$ *relating column* i *in* $\mathbf{Z}_t$ *to column* j *of* $\mathbf{Y}_t$ *are zero and* $i \neq j$, *then* $\widetilde{b}_{ij} = 0$.

2.4.3 Comparison of BLIN and bilinear least squares estimators

We now compare least squares estimates of the BLIN model to those of the bilinear model. As the bilinear model accommodates only a single lag for $\mathbf{A}$ and $\mathbf{B}$, in all that follows we fix $p_A = p_B = p$ so that $\mathbf{Z}_t = \mathbf{X}_t$, $\mathbf{\Omega}_X = \mathbf{\Omega}_Z = \mathbf{\Omega}$, and $\mathbf{\Psi}_X = \mathbf{\Psi}_Z = \mathbf{\Psi}$. Both the BLIN and bilinear models aim to quantify the influences of the rows (columns) of $\mathbf{X}_t$ on the rows (columns) of $\mathbf{Y}_t$, but with different emphasis on the type of influence quantified (recall the discussion in the Introduction and Section 2.2.1). We provide a theorem and proposition stating that, when data are generated from the bilinear model, the least squares BLIN estimators of the off-diagonal entries in $\mathbf{A}$ and $\mathbf{B}$ converge to the corresponding $\mathbf{A}$ and $\mathbf{B}$ values used in the bilinear generative model, up to numerical constants. The analogous result holds when switching the roles of the bilinear and BLIN models. See Appendix A.2 for proofs.

Theorem 4. *Under Conditions 1,*

1. *If $\{\mathbf{Y}_t\}_{t=1}^T$ are generated from the bilinear model in* (2.6)*, then for all $i \neq j$ and $k \neq \ell$*

$$\widetilde{a}_{ij} = \frac{\mathrm{tr}(\mathbf{\Omega B})}{\mathrm{tr}(\mathbf{\Omega})} a_{ij}, \quad \textit{and} \quad \widetilde{b}_{k\ell} = \frac{\mathrm{tr}(\mathbf{\Psi A})}{\mathrm{tr}(\mathbf{\Psi})} b_{k\ell}.$$

2. *If $\{\mathbf{Y}_t\}_{t=1}^T$ are generated from the BLIN model in* (2.1)*, then for all $i \neq j$ and $k \neq \ell$*

$$\bar{a}_{ij} = \frac{\mathrm{tr}(\mathbf{\Omega}\bar{\mathbf{B}})}{\mathrm{tr}(\mathbf{\Omega}\bar{\mathbf{B}}\bar{\mathbf{B}}^T)} a_{ij} \quad \textit{and} \quad \bar{b}_{k\ell} = \frac{\mathrm{tr}(\mathbf{\Psi}\bar{\mathbf{A}})}{\mathrm{tr}(\mathbf{\Psi}\bar{\mathbf{A}}\bar{\mathbf{A}}^T)} b_{k\ell},$$

   where $\bar{\mathbf{A}} = \{\bar{a}_{ij}\}$ and $\bar{\mathbf{B}} = \{\bar{b}_{k\ell}\}$ are the pseudo-true parameters of least-squares estimation of the bilinear model.

We now address the diagonals of the $\mathbf{A}$ and $\mathbf{B}$ matrices. We provide conditions under which the diagonals (up to their non-identifiabilities) are asymptotically equivalent. To be clear, we compare the estimated influence of x_{ij}^t on y_{ij}^t from the two models. Under the BLIN model, this influence is $a_{ii} + b_{jj}$, whereas under the bilinear model the influence is $a_{ii} b_{jj}$.

Proposition 5. *Suppose the true* $\mathbf{A}$ *and* $\mathbf{B}$ *matrices are constant along the diagonal with nonzero values* α *and* β*, respectively, and* $\alpha + \beta \neq 0$*. Then, under Conditions 1,*

1. *If* $\{\mathbf{Y}_t\}_{t=1}^T$ *are generated from the bilinear model in* (2.6)*, then the pseudo-true parameter* $\widetilde{a}_{ii} + \widetilde{b}_{jj}$ *of least-squares estimation of the BLIN model is equal to the true diagonal specification* $\alpha\beta$*, and*

2. *If* $\{\mathbf{Y}_t\}_{t=1}^T$ *are generated from the BLIN model in* (2.1)*, then the pseudo-true parameter* $\bar{a}_{ii}\bar{b}_{jj}$ *of least-squares estimation of the bilinear model does <u>not</u> equal the true diagonal specification* $\alpha + \beta$ *in general.*

The conditions for equivalence of the diagonals are more restrictive than those for the equivalence of the off-diagonal elements of $\mathbf{A}$ and $\mathbf{B}$. Furthermore, we see the BLIN model diagonals are consistent in misspecification situations when the bilinear diagonals are not consistent. In Appendix A.3.3, we evaluate the convergence of the bilinear and full BLIN estimators in support of Theorem 4 and Proposition 5.

Although much of this section describes the similarities between the BLIN and bilinear models, we emphasize that these similarities lie in estimation of the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$. The estimated mean function of the BLIN and bilinear models are very different (even asymptotically), which has implications for model fit and prediction. To illustrate the difference in mean functions between the BLIN and bilinear models, we show that equivalence of $\mathbf{A}$ and $\mathbf{B}$ between the BLIN and bilinear models implies equivalence in the estimated mean $\widehat{\mathbf{Y}}_t$ only when $\mathbf{A}$ and $\mathbf{B}$ are diagonal.

Proposition 6. *Let the estimators from the BLIN and bilinear models have diagonals that are equal up to a constant, as in Theorem 4. Additionally, let the estimators of the diagonals be constant as in Proposition 5, part 1. Then, under the conditions of Proposition 5 with* $\mathbf{\Omega}$ *and* $\mathbf{\Psi}$ *diagonal, equivalence of* $\widehat{\mathbf{Y}}_t$ *for both models implies that the BLIN and bilinear estimators,* $\{\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\}$ *and* $\{\bar{\mathbf{A}}, \bar{\mathbf{B}}\}$*, respectively, are all diagonal.*

2.5 Simulation study

As discussed in the previous section, although the BLIN and bilinear estimators of $\mathbf{A}$ and $\mathbf{B}$ are asymptotically equivalent under certain conditions, their mean functions are only equivalent under strict conditions (see Proposition 6). Thus, the ability of these models to represent the variation in a bipartite relational data set will heavily depend on whether the true mean is of BLIN or bilinear form. To compare the ability of each model to represent mean structure under model misspecification, we conducted a simulation study.

We generated from a lag-1 vector-autoregressive model

$$\mathbf{y}_t = \mathbf{\Theta}\mathbf{y}_{t-1} + \mathbf{e}_t, \qquad t \in \{1, 2, \ldots, T\}, \tag{2.16}$$

where $\mathbf{y}_t = \text{vec}(\mathbf{Y}_t)$ is the columnwise vectorization of the 10×10 matrix of bipartite relations $\mathbf{Y}_t$, $\mathbf{e}_t$ consists of i.i.d. standard normal entries, and the number of time periods $T \in \{10, 20, 50\}$. Recall that both the BLIN and bilinear models may be cast as VAR models of the form of (2.16), where $\mathbf{\Theta} = \mathbf{A}^T \otimes \mathbf{I}_L + \mathbf{I}_S \otimes \mathbf{B}^T$ for the BLIN model and $\mathbf{\Theta} = \mathbf{B}^T \otimes \mathbf{A}^T$ for the bilinear model. We created weighted, directed $\mathbf{A}$ and $\mathbf{B}$ matrices and used these to generate data from both models.

The process for specifying $\mathbf{A}$ and $\mathbf{B}$ was motivated by a desire to construct matrices with network structure that we might expect in an influence network. Initially, $\mathbf{A}$ and $\mathbf{B}$ were randomly generated as matrices of rank 1, which may be viewed as latent factor models of rank 1 (Hoff, 2008; Li et al., 2011). The diagonal entries and smallest $q = 0.9$ fraction of off-diagonal entries were set to zero, such that the matrices were approximately low rank and sparse. Finally, we scaled the $\mathbf{\Theta}$ matrix of each generating model to control the signal-to-noise ratio, such that the true model had an R^2 of approximately 0.75 for T approaching infinity. For further details on the simulation study, including investigation of $q = 0.5$ and $q = 0.0$, please see Appendix A.3.

We generated 100 data sets from both the BLIN and bilinear model for $T = 50$, and evaluated the out-of-sample predictions from the models in a 10-fold cross validation study. To compare performance of the estimators on data sets containing varying amounts of time periods, we evaluated

performance on the complete data sets with $T = 50$, as well as the performance when the data were trimmed to include only the last 10 time periods and only the last 20 time periods. In each of these three scenarios, i.e. $T \in \{10, 20, 50\}$, we performed a 10-fold cross validation. The time periods were randomly partitioned into 10 sets (the partitions were the same for all data sets of a given size T for the sake of equal comparison). Models were fit to the data in nine of the ten sets, and then predictions for the values in the left-out time period set were obtained. This fitting procedure was repeated 10 times: once for each of the partitions. Models were evaluated based on the R^2 value between the ten sets of predicted values and the true values, for each data set and each T. R^2 is a natural measure of model fit as the data are normally distributed and hence high R^2 values corresponds to large likelihood values. For each of the two generative models (BLIN and bilinear) and each of the three data set sizes ($T \in \{10, 20, 50\}$), we compare the performance of four models: the bilinear model and the full, reduced rank (with rank set to 1), and sparse BLIN models.

We plot the resulting R^2 values in Figure 2.3 for sparsity of the generating coefficients $q = 0.9$. A dotted horizontal line is drawn at $R^2 = 0$, which denotes the expected performance of fitting no model at all, that is, predicting $\widehat{\mathbf{Y}}_t = \mathbf{0}$ for all t. An additional dotted horizontal line is drawn at $R^2 = 0.75$, the expected large-sample R^2 value when the true model is known. When generating from the BLIN model (left panel of Figure 2.3), the estimated full and sparse BLIN models perform well for all values of $T \in \{10, 20, 50\}$. This is as expected, since the true matrices $\mathbf{A}$ and $\mathbf{B}$ are sparse. The reduced rank BLIN model is not able to represent this sparse structure, and thus its out-of-sample performance falls short of the full and sparse BLIN models. When generating from the BLIN model, the bilinear model results in extremely poor predictions when $T = 10$ and marginal performance when $T = 20$, yet it is on par with the BLIN models for $T = 50$.

When generating from the bilinear model (right panel of Figure 2.3), as expected, the bilinear model performs best for the largest number of replications $T = 50$. Surprisingly, only for this value does the bilinear model consistently outperform predicting simply $\widehat{\mathbf{Y}}_t = \mathbf{0}$, even though the data are generated from the bilinear model. Overall, the BLIN models perform poorly when

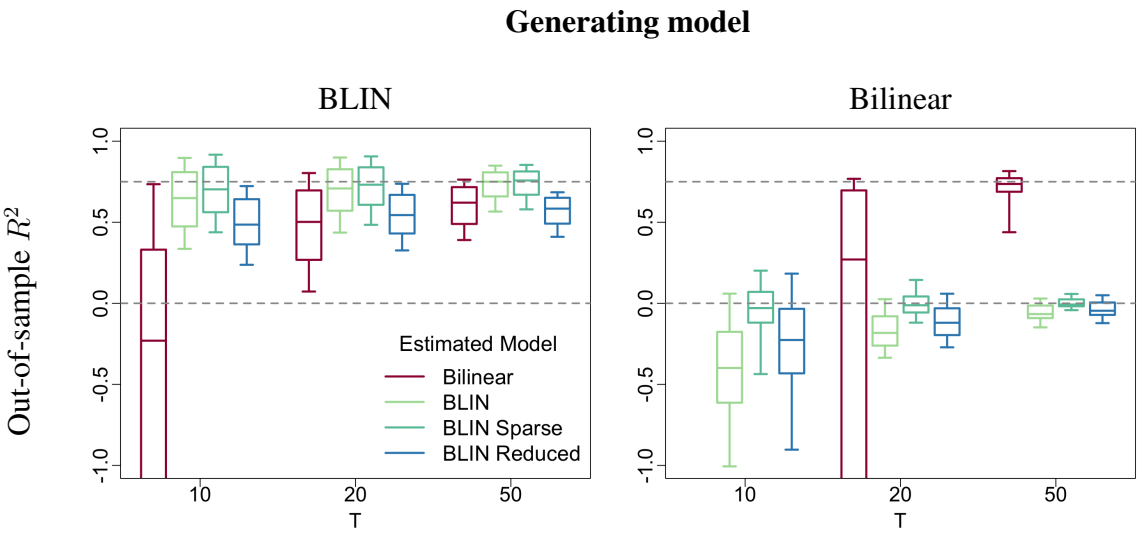


Figure 2.3: Out-of-sample R^2 values for each estimation procedure applied to 100 data realizations generated from the BLIN model (left panel) and generated from the bilinear model (right panel). The centers of the boxplots represent the median R^2 value, the boxes represent the middle 80% of R^2 values, and the whiskers correspond to the maximum and minimum R^2 values across 100 simulated data sets. Plots are truncated such that R^2 values less than -1 are not shown.

generating from the bilinear model. However, we note that the sparse BLIN model guards against poor performance (the typical R^2 is always about 0), whereas the full and reduced rank BLIN models do not.

We investigated the source of errors when the data were generated from the bilinear model by examining the likelihood surface of the bilinear model near the estimated and true values of the influence matrices (Appendix A.3). These investigations showed that, when $T \in \{10, 20\}$, the likelihood surface of the BLIN model is multimodal and that the highest mode may be "far" from the mode that is nearest the true parameter values. This means that estimating the model on one portion of data may not generalize well to other portions of the data, which is exactly what we observe in the negative R^2 values of the bilinear estimator when the data are generated from the bilinear model for $T \in \{10, 20\}$ (Figure 2.3). A similar analysis for the BLIN model confirmed the unimodality implied by Propostion 1. The results of the simulation study suggest that the bilinear model may be difficult to estimate unless a large number of replications T are observed. In addition, although the coefficients estimated by the BLIN and bilinear models may be similar

for large T (as suggested by theory and confirmed by simulation in Appendix A.3), neither model may be a good predictive substitute for the other in small samples.

2.6 Temporal state interaction data analysis

Power, defined as the ability of an actor to influence another to do something they may not otherwise do, is perhaps the most essential concept in the study of politics broadly, and interstate relations specifically (Morgenthau, 1978; Dahl, 1957; Waltz, 1979; Lukes, 2004). Thus, both network researchers and political scientists are frequently interested in inferring patterns of influence in interactions between states (Minhas et al., 2017). Examining country behaviors, we may use the proposed approach to directly infer the most powerful and influential actors in modern international relations and also gain insights that may yield valuable predictions of international policy changes, such as ratification of environmental treaties (Campbell et al., 2019). Inferring these influences resolves a significant shortcoming in the international relations literature, which has been forced to measure power in terms of the relative military or economic strength because of the empirical difficulties associated with measuring power as influence (Singer et al., 1972; Hart, 1976; Johnson, 2017).

To infer country influences, we analyzed country interactions at weekly intervals from 2004 to mid-2014, giving $T = 543$ weeks of data. The relations were obtained from the Integrated Crisis Early Warnings System (ICEWS) (Boschee et al., 2015), previously analyzed in Hoff (2015), which automatically identifies and encodes interaction intensities (between -10 and 10) from news stories. Each relation is one of four interaction types from a source state to a target state: material negative actions (*mn*), material positive actions (*mp*), verbal negative actions (*vn*), and verbal positive actions (*vp*). We analyzed only the 25 most active states. An example of each interaction type is boycotting for leadership change (*mn*), providing humanitarian aid (*mp*), denying accusations by a target country (i.e. charges of genocide or other human rights violations) (*vn*), and expressing intent to negotiate (*vp*).

We denote the intensity of relation in week t as y_{ijk}^t, where i is the source state, j is the target state, k is the relation type, and t is the week of the observation. Each time series $\{y_{ijk}^t\}_{t=1}^T$ is centered and standardized. Here we analyze the *change* in relations using the differences $d_{ijk}^t := y_{ijk}^t - y_{ijk}^{t-1}$. The BLIN model defined in Section 2.2 is for longitudinal bipartite data, however the ICEWS data set contains tripartite data, such that observations are indexed by (source, target, type) triples. For this reason, here we introduce an extension of the BLIN model for tripartite data, which we abbreviate the TLIN (tripartite longitudinal influence network) model, and point interested readers to Appendix A.4 for details on an extension of the BLIN model specification for arbitrary multipartite data. In the TLIN model, the change in relation intensity in each week t, d_{ijk}^t, depends on the previous actions of states that influence source state i through source influence matrix $\mathbf{A}$, the previous actions of states that influence target state j through target influence matrix $\mathbf{B}$, and interaction types that influence interaction type k through the interaction type influence matrix $\mathbf{C}$. For the ICEWS data the TLIN model can be expressed

$$d_{ijk}^t = \sum_{s=1}^{25} a_{si} \left(\sum_{r=1}^{p_A} d_{sjk}^{t-r} \right) + \sum_{\ell=1}^{25} b_{\ell j} \left(\sum_{r=1}^{p_B} d_{i\ell k}^{t-r} \right) + \sum_{u=1}^{4} c_{uk} \left(\sum_{r=1}^{p_C} d_{iju}^{t-r} \right) + e_{ijk}^t, \tag{2.17}$$

where each e_{ijk}^t is an independent, mean zero random error. The TLIN model in (2.17) states that, if there is a positive source influence from the Russia to France in $\mathbf{A}$, then an observed increase in, say, boycott intensity from Russia to China implies that we should expect France to increase its boycotting of China in the following weeks. For $\mathbf{B}$, the model states that if there exists a negative target influence from Lebanon to Pakistan, then an increase in US humanitarian aid sent to Lebanon would indicate a decrease in aid sent from US to Pakistan in the following weeks. Finally, a positive influence of *vn* on *mn* in $\mathbf{C}$ suggests that verbal negative interactions precede material negative interactions; that is, if the US threatens to increase boycotts on North Korea, then we might expect the US to increase tariffs on North Korea in the following weeks. Finally, we contend that imposing a boycott is fundamentally different than having a boycott imposed upon

one's state, and thus, although the country sets of sources and targets are the same, we treat these nodes as separate types and infer separate influence networks among source and target countries.

To choose the lag values $\{p_A, p_B, p_C\}$, corresponding to influences among source countries, target countries, and interaction types, respectively, we fit a sparse TLIN model to the ICEWS data for a range of lags using an ℓ^1 penalty on the entries in the influence networks. We then chose the lags that gave the best balance of model fit and model parsimony based on a comparison of likelihood values of the estimated models, penalized by twice the number of nonzero estimated parameters, in the vein of Akaike's Information Criteria (Akaike, 1998). This procedure resulted in lag values $\{p_A = 5, p_B = 3, p_C = 1\}$. We note that the choice to difference the responses is a departure from the analysis in Hoff (2015). For more details of the data analysis, see Appendix A.5.

The estimated source country ($\mathbf{A}$), target country ($\mathbf{B}$), and interaction type ($\mathbf{C}$) influence networks are depicted in Figure 2.4, where we focus on the entries that constitute the largest 5% of magnitudes across all networks (see Figure A.6 in Appendix A.5). Across all networks, there are generally larger and more positive entries than negative entries (Figure 2.4). This fact suggests that positive influence is more consequential than negative influence: e.g., increases in aid generally lead to other increases in aid, rather than decreases. For example, in the target influence network $\mathbf{B}$, there is a positive influence from North Korea (PRK) to China (CHN) and a smaller, negative influence from North Korea to Afghanistan (AFG). This suggests that an increase in boycotts by Great Britain on North Korea leads one to expect an increase in boycotts by Great Britain on China in the following three weeks, and a smaller decrease in boycotts by Great Britain on Afghanistan.

One of the largest positive influences in the $\mathbf{A}$ network is that of the US on Germany (DEU). This means that if the US increases humanitarian aid sent to Syria, for example, then we expect Germany to increase the amount of humanitarian aid it sends to Syria in the following $p_A = 5$ weeks. This result matches conventional wisdom as the US and Germany have been among the closest North Atlantic Treaty Organization (NATO) allies following the Cold War, also sharing many common geopolitical interests outside this particular alliance. Within $\mathbf{A}$, we note that the US and Russia have the strongest ties with other countries and the most ties across continents.

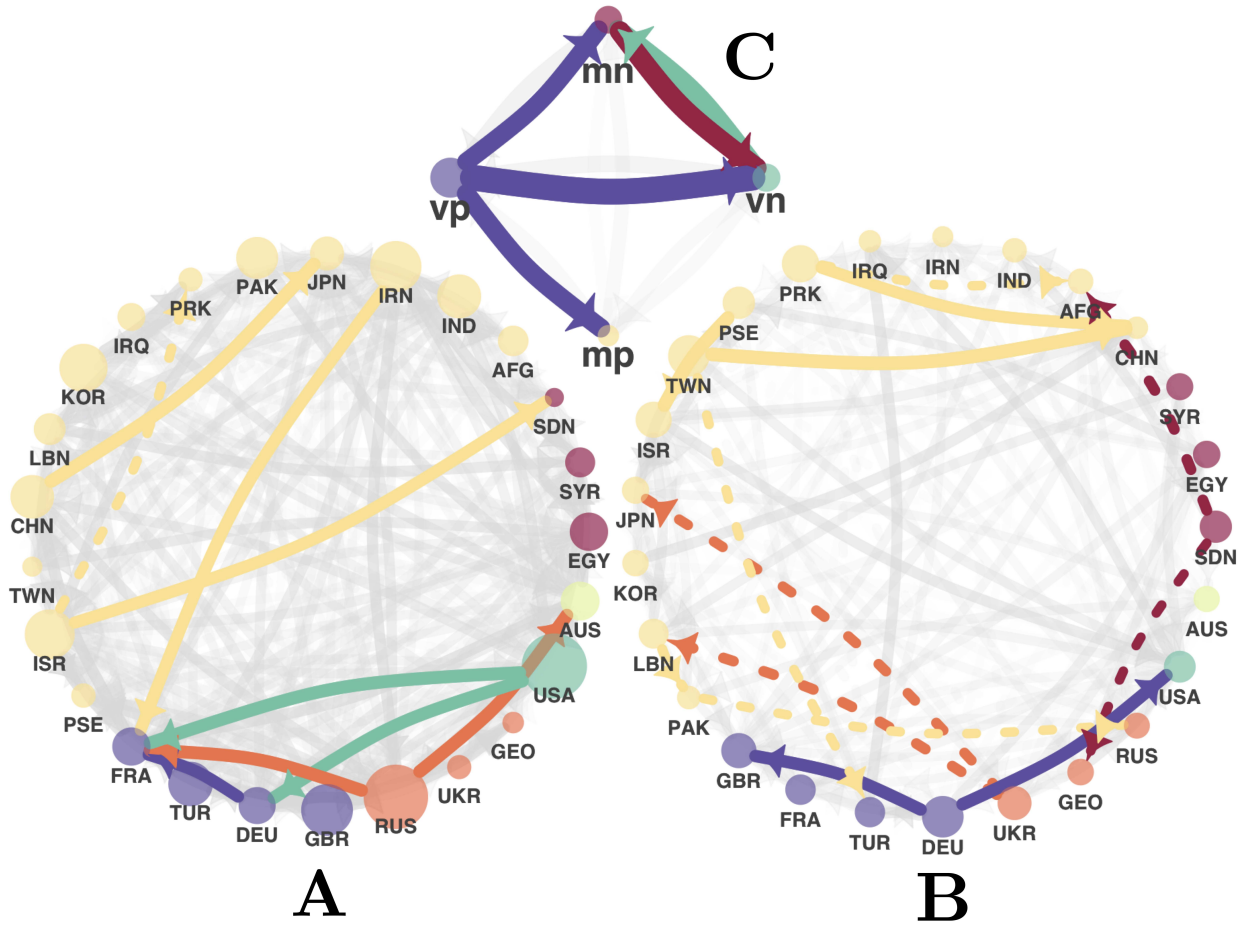


Figure 2.4: The largest 5% of edges (in magnitude) are shown in color; the smaller relations are shown in grayscale proportional to their magnitude. The source country network is in the bottom-left panel (**A**), the target country network is in the bottom-right panel (**B**), and the interaction type network is in the top-center panel (**C**). Nodes are sized proportional to the sum of the magnitudes of outgoing relations (comparable across networks) and nodes in **A** and **B** are colored according to continent. Edges are sized proportional to absolute edge weight (comparable across networks) and colored according to the originating node. Solid lines denote positive relations and dotted lines denote negative relations.

This also may have been expected since great powers should possess the most influence relationships and the most substantively important influence relationships (Morgenthau, 1978; Waltz, 1979; Mearsheimer, 2001).

In the target country **B** influence network, we observe that the most influential country is Germany (DEU), suggesting again that Germany is very central in the international community. We see that the **B** network has many more large negative entries than **A** and **C**. This fact is natural, as, for some relations, target states are competing for resources and hence, influences

may be limited. The material positive and material negative interactions are zero-sum, that is, a dollar sent to Sudan cannot also be sent to Afghanistan. For example, there is a large negative influence in $\mathbf{B}$ from the Sudan (SDN) to Afghanistan (AFG). This influence indicates that, when Sudan receives an increase in, say, humanitarian aid from the US, Afghanistan is expected to have reduced aid from the US in the coming weeks. Hence, these countries are essentially competing for the resources of source countries, of which resources there are finite amounts. We do not observe this phenomenon in the $\mathbf{A}$ network as the Germany following the US in aiding Sudan employs both the resources of the German and US people.

From $\mathbf{C}$ we glean understanding of the relationships between interaction types. For example, we observe positive relations between material negative and verbal negative interaction types, which means, for example, that increases (or decreases) in verbal negative interactions from the US to China may signal increases (decreases) in material negative interactions from the US to China. We also observe large positive influences from *vp* to the other relation types. This fact indicates that changes in verbal positive interactions are often followed by changes in other interaction types, but that a change in verbal positive relations is relatively uninformative to what type of interaction may change afterwards. Finally, since $\mathbf{C}$ is entirely positive, this indicates positive feedback loops among the interaction types and a lack of negative feedback loops.

With the development of the BLIN model and multipartite extensions, international relations scholars no longer must rely upon the widely used proxy measures of power and influence, such as national capabilities or economic development. Instead, they are now permitted to measure power directly, as the influence exercised by one state over another. This approach has already been adopted and shown to be powerful in modeling the influence that countries exercise over one another in the ratification of environmental treaties (Campbell et al., 2019).

2.7 Discussion

In this chapter, we present the Bipartite Longitudinal Influence Network model, a novel generative model for the evolution of bipartite relational data over time. The BLIN model allows for

the estimation of the weighted and directed influence networks among each of the two actor types in bipartite data, each with their own separate time scale of influence. The BLIN model can be expressed as a generalized linear model, lending itself to use with a litany of off-the-shelf tools for estimation and to straightforward parameter interpretation.

In the BLIN model, the entries in the influence networks $\mathbf{A}$ and $\mathbf{B}$ may be interpreted as "average" influences over the entire data set. There are scenarios where the influences among countries may evolve over time. For example, major international trade agreements or wars might change the structures of the influence networks. To determine whether $\mathbf{A}$ and $\mathbf{B}$ evolve over time, we might consider testing for changepoints in the influence networks. Of course, this requires the development of a more flexible model with time-varying influence networks, such as modeling $\mathbf{A}$ and $\mathbf{B}$ as linear functions of known covariates as in Minhas et al. (2017).

Chapter 3

Regression of relational data with exchangeable errors

3.1 Introduction

Relational arrays have recently become extremely common in the social and biological sciences. Entries in relational arrays quantify pairwise interactions between actors that may be of multiple types or may be observed over time. Examples include annual flows of migrants between countries (Aleskerov et al. (2017)) and interactions among students over the course of a semester (Han et al. (2016)). In economics, relational arrays can be used to describe monetary transfers between individuals as part of informal insurance markets (see, for example, Bardham (1984); Fafchamps (2006); Foster and Rosenzweig (2001); Attanasio et al. (2012); Banerjee et al. (2013)). Other examples of data that are naturally represented as relational arrays include gene expression data (Zhang and Horvath (2005)) and international relations among countries (Fagiolo et al. (2008)).

A common task in analyzing relational arrays is to infer the effects of exogenous covariates on the values in the relational array. Taking again informal insurance markets, Fafchamps and Gubert (2007) examine how covariates such as geographical proximity and kinship relate to risk sharing relations after economic shocks. Recent extensions of this work (e.g. Aker (2010); Blumenstock et al. (2011); Jack and Suri (2014)) explore the strength of the association between physical proximity and financial transactions among individuals with access to mobile phones. Thus, our motivation is analyses where the primary goal is inference for the effects of exogenous covariates on the values in the relational array.

A relational array $Y = \left\{ y_{ij}^{(r)} : i, j \in \{1, ..., n\}, i \neq j, r \in \{1, ..., R\} \right\}$ is composed of a series of R $\left(n \times n\right)$ matrices, each of which describes the directed pairwise relationships among n actors

of type r, e.g. time period r or relation context r. The diagonal elements of each matrix $\{y_{ii}^{(r)} : i \in \{1, ..., n\}\}$ are assumed to be undefined, as we do not consider actor relations with his/herself. For much of this chapter, we focus on data sets represented as a single matrix of relations, i.e. $R = 1$, where the array Y is often simply referred to as a weighted network.

We consider regression models that express the entries in a relational array as a linear function of observable covariates:

$$y_{ij}^{(r)} = \boldsymbol{\beta}^T \mathbf{x}_{ij}^{(r)} + \xi_{ij}^{(r)}, \qquad i, j \in \{1, ..., n\},\ i \neq j,\ r \in \{1, ..., R\}, \tag{3.1}$$

where inference for $\boldsymbol{\beta}$ is the primary goal. In (3.1), $y_{ij}^{(r)}$ is a (continuous) directed measure of the rth relation from actor i to actor j and $\mathbf{x}_{ij}^{(r)}$ is a $\big(p \times 1\big)$ vector of covariates, which are unrelated (i.e. exogenous) to the mean-zero error $\xi_{ij}^{(r)}$. The data sets we consider are in contrast to those where the response is a vector of actors which are connected in a possibly unknown network, for example, see Zhou and Song (2016). In a study on international trade, $y_{ij}^{(r)}$ may denote the value of trade exported from country i to country j in year r and the covariates may include country-specific attributes such as GDP and population, as well as country pair characteristics such as geographic distance. Throughout the chapter, we assume that relations are directed such that the relationship from actor i to actor j may differ than that from j to i, however, the methods we propose extend to the undirected/symmetric relation case in a straightforward manner. We discuss the extension to undirected arrays in Appendix B.1.

A core statistical challenge in modeling relational arrays arises from the innate dependencies among relations involving the same actor. For example, dependence often exists between trade relations involving the same country and between economic transfers originating from the same individual. This dependence may arise, for example, from differences in production levels between nations or, in the informal insurance markets, individual differences in risk aversion. Substantial dependence in the errors precludes the use of standard regression techniques, as these techniques may lead to faulty inference. While unbiased estimation for the $\boldsymbol{\beta}$ coefficients in (3.1) is possible

via ordinary least squares (OLS), accurate uncertainty quantification for β, i.e. standard errors, requires consideration and estimation of any auxiliary dependence. Approaches for addressing this challenge have appeared in the statistics, biostatistics, and econometrics literatures and can be characterized into two broad classes.

The first set of approaches impose a parametric model on the errors. Specifically, they either use latent variables to model the array measurements as conditionally independent given the latent structure (e.g. see Holland et al. (1983); Wang and Wong (1987); Hoff et al. (2002); Li and Loken (2002); Hoff (2005)) or model the error covariance structure directly subject to a set of simplifying assumptions (e.g. see Hoff (2011); Fosdick and Hoff (2014); Hoff (2015)). While these methods provide parsimonious representations of the underlying error structure, the accuracy of inference on β depends on the extent to which the true error structure is consistent with the specified parametric model. In addition, many of these models are estimated in a Bayesian paradigm using Markov chain Monte Carlo approaches so they are expensive to estimate for networks of even a few hundred nodes.

The second approach to accounting for error dependence relies heavily on empirical estimates of the error structure based on the residuals in an estimating equation/moment condition framework, first proposed by Fafchamps and Gubert (2007) and based on the spatial dependence work of Conley (1999). In contrast to the first approach, this framework makes as few assumptions as possible about the data generating process and utilizes a sandwich covariance estimator for the standard errors of the regression coefficients. Sandwich estimators employ the regression residuals to "adjust" the standard error estimate in case the moment conditions are misspecified or there is dependence structure within the errors. As a result, the sandwich estimator is commonly known as a *robust* estimator of the standard error. The quality of this correction depends on the accuracy of the error covariance estimate based on the residuals. In finite samples, current error covariance estimators for relational regression are hindered by the need to estimate a large number of covariance parameters with limited observations. These practical limitations have been recognized in other contexts (see King and Roberts (2015) for a discussion) and is the reason why Wakefield (2013)

suggests such estimators be labeled "empirical" rather than "robust." We examine this estimator in more detail in the following section. Other empirical approaches to the problem of accounting for error dependence in relational data bear mentioning, such as permutation testing Dekker et al. (2007) and developing work in bootstrapping Menzel (2017); Green and Shalizi (2017). Although we do not address these methods further, the vast majority are based on the exchangeability assumption examined in this chapter.

In this chapter, we extend the estimating equation/moment condition framework by incorporating an exchangeability assumption. This assumption is implicit in many of the model-based approaches discussed previously and is a hallmark of Bayesian hierarchical models, including models for data outside the relational context (Orbanz and Roy (2015)). Let $Y^{(r)}$ denote the r^{th} $\big(n \times n\big)$ matrix slice in the array containing all relations of type r. We propose leveraging the assumption of exchangeability assumption within, and potentially across, each matrix $Y^{(r)}$ to derive a parsimonious estimator of the relational dependence. Our approach produces a dramatically simplified estimator that results in superior performance in inference, which we demonstrate both theoretically and empirically in simulation studies. The two key features of our approach are that it is model agnostic, not assuming any specific underlying parametric model, and is significantly easier to compute compared to existing Bayesian model-based approaches.

this chapter is organized as follows. The remainder of this section provides background on the estimating equation framework in the context of relational data. Section 3.2 describes current inference approaches arising from the network econometrics literature and literature on moment condition estimators with cross-sectional dependence, focusing specifically on data sets where $R = 1$ (e.g. Conley (1999); Hansen (2015)). We discuss what it means for relational data to be exchangeable in Section 3.3 and present our proposed covariance matrix estimator based on an exchangeability assumption. Section 3.4 describes the improvements in mean-square error of our proposed method compared to the current state of practice, as supported by extensive theoretical results and simulation evidence. We discuss extensions of our method for use with arrays with

$R > 1$ in Section 3.5 and demonstrate our methodology using a data set of international trade flow in Section 3.6. We conclude with a discussion in Section 3.7.

3.1.1 Accounting for correlated errors in relational regression

A key statistical challenge in relational regression is accounting for the correlation structure present in the $n \times n \times R$ array of error terms $\Xi = \left\{\xi_{ij}^{(r)} : i, j \in \{1, ..., n\}, i \neq j, r \in \{1, ..., R\}\right\}$. First, consider a single matrix of relations $Y^{(r)} = \left\{y_{ij}^{(r)} : i, j \in \{1, ..., n\},\ i \neq j\right\}$ and error matrix $\Xi^{(r)}$ corresponding to relation type r. There are two primary types of correlation we might expect among the errors. The first type is between relations within the same row or within the same column of the matrix $\Xi^{(r)}$. Revisiting the international trade example, this dependence corresponds to correlation among a country's exports (i.e. within a rows of $\Xi^{(r)}$) and correlation among a country's imports (i.e. within a column of $\Xi^{(r)}$). These dependence patterns are often seen in array data in general, even when each dimension of the array is distinct Hoff (2011); Fosdick and Hoff (2014). The second type of correlation we expect, which is specific to relational data, stems from the fact that the row and column index sets represent the same entities. Again, in the context of the trade data, we might expect France's exports to Germany to depend upon the amount Spain exports to France. This corresponds to dependence between errors, say, $\xi_{ij}^{(r)}$ and $\xi_{ki}^{(r)}$.

We use an estimating equations/moment conditions framework to perform inference on $\boldsymbol{\beta}$ (Wakefield, 2013; Hansen, 2015). In relational regression, estimating equations g are defined such that for all (i, j, r), $\mathbb{E}\left[g(y_{ij}^{(r)}, \boldsymbol{\beta})\right] = \mathbf{0}_p$, where $\mathbf{0}_p$ is the p-dimensional vector of zeros. The estimator $\widehat{\boldsymbol{\beta}}$ is then defined as that which satisfies

$$G(Y, \widehat{\boldsymbol{\beta}}) := \sum_{i,j,r} g(y_{ij}^{(r)}, \widehat{\boldsymbol{\beta}}) = \mathbf{0}_p. \tag{3.2}$$

The estimating equations g characterize specific features of the population distribution (e.g. the first moment), but critically, this approach does not fully specify the population distribution.

Consider the relational regression model as defined in (3.1). There are many g functions one could specify which would provide reasonable $\boldsymbol{\beta}$ estimates. One common specification is (see, for

example, Chapter 11 in Hansen (2015) or Chapter 5 in Wakefield (2013))

$$g(y_{ij}^{(r)}, \boldsymbol{\beta}) = \mathbf{x}_{ij}^{(r)} \left(y_{ij}^{(r)} - \boldsymbol{\beta}^T \mathbf{x}_{ij}^{(r)} \right). \tag{3.3}$$

This corresponds to the score function of the multivariate normal likelihood assuming homoskedastic, independent errors and gives rise to the familiar ordinary least squares estimate of $\boldsymbol{\beta}$: $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y_v}$, where X is an $\big(n(n-1)R \times p\big)$ matrix of covariate vectors $\{\mathbf{x}_{ij}^{(r)}\}$ and $\boldsymbol{Y_v}$ is a vectorized representation of Y. Under regularity conditions Van der Vaart (2000); Cameron et al. (2011), the estimator satisfying (3.2) is consistent ($\widehat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$) and moreover asymptotically normal:

$$\sqrt{n} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \to_d \mathrm{N}\Big(\mathbf{0}_p,\ A^{-1} B (A^T)^{-1} \Big), \tag{3.4}$$

where $A = \mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{\beta}^T} G(Y, \boldsymbol{\beta}) \Big]$ and $B = \mathbb{E}\Big[G(Y, \boldsymbol{\beta}) G(Y, \boldsymbol{\beta})^T \Big]$, such that $G(Y, \boldsymbol{\beta})$ is as defined in (3.2). Estimating the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ then amounts to estimating A and B. Asymptotic covariance estimators of the form $\widehat{A}^{-1} \widehat{B} (\widehat{A}^T)^{-1}$ are commonly referred to as "sandwich" estimators Huber (1967); White (1980). Assuming independence across observations, the elements of the covariance can be estimated as

$$\widehat{A} = \frac{1}{n(n-1)R} \sum_{i,j,r} \frac{\partial}{\partial \boldsymbol{\beta}^T} g\big(y_{ij}^{(r)}, \widehat{\boldsymbol{\beta}}\big) \quad \text{and} \quad \widehat{B} = \frac{1}{n(n-1)R} \sum_{i,j,r} g\big(y_{ij}^{(r)}, \widehat{\boldsymbol{\beta}}\big) g\big(y_{ij}^{(r)}, \widehat{\boldsymbol{\beta}}\big)^T.$$

When g is defined as in (3.3) for relational data, $A = X^T X$ and $B = X^T \Omega X$, where $\Omega = \mathrm{V}[\boldsymbol{Y_v} | X]$ is the covariance matrix of the relations, equivalently the errors. Ω appears in the form of the variance for most g functions commonly used to estimate $\boldsymbol{\beta}$. When observations are independent and homoskedastic, Ω is proportional to the identity matrix and the form of the variance simplifies to that from standard linear regression. However, independence among the errors is often violated in relational data as we expect relations involving the same actor(s) will be dependent. More complex covariance structures have been considered that assume only subsets of the observations be independent. These independent subsets are often specified based on distance metrics derived

from observable features of the data White and Domowitz (1984); Liang and Zeger (1986); Conley (1999). In the next section, we discuss in detail the estimators proposed for relational data.

3.2 Dyadic clustering estimator

To facilitate presentation, we first describe the current state-of-the-art sandwich covariance estimation framework with a single relation $Y^{(1)}$, then move to arrays with $R > 1$. For notational simplicity, we presently drop the superscript (1) indexing the relation type and reintroduce it in Section 3.5 when needed. Thus, $y_{ij} = y_{ij}^{(1)}$, $\mathbf{x}_{ij} = \mathbf{x}_{ij}^{(1)}$, $\boldsymbol{Y_v}$ is an $\big(n(n-1) \times 1\big)$ vector of relational observations in $Y^{(1)}$, and X is the $\big(n(n-1) \times p\big)$ matrix of covariates for these relations.

Consider a ordered pair (i, j) and define Θ_{ij} as the set consisting of all ordered pairs that contain an overlapping member with the pair (i, j). In other words, $\Theta_{ij} = \big\{(k, l) : \{i, j\} \cap \{k, l\} \neq \varnothing\big\}$. Generalizing the standard estimating equation framework, Fafchamps and Gubert (2007), Cameron et al. (2011), Aronow et al. (2015), and Tabord-Meehan (2018) propose and describe the properties of a flexible standard error estimator for relational regression which makes the sole assumption that two relations (i, j) and (k, l) are independent if (i, j) and (k, l) do not share an actor (i.e. $(k, l) \notin \Theta_{ij}$). This implies that $\text{Cov}(y_{ij}, y_{kl}|X) = \text{Cov}(\xi_{ij}, \xi_{kl}) = 0$ for non-overlapping pairs, but places no restrictions on the covariance elements for pairs that involve the same actor. Let Ω_{DC} denote the covariance matrix $\text{V}[\boldsymbol{Y_v}|X]$ subject to this non-overlapping pair independence assumption. Fafchamps and Gubert (2007) propose estimating each nonzero entry of Ω_{DC} with a product of residuals, e.g. $\widehat{\text{Cov}}(\xi_{ij}, \xi_{ik}) = e_{ij}e_{ik}$. This may be expressed in matrix form as

$$\widehat{\Omega}_{DC} = \mathbf{e}\mathbf{e}^T \circ \mathbf{1}_{[\{i,j\}\cap\{k,l\}\neq\varnothing]}, \tag{3.5}$$

where $\mathbf{e}$ is the vector of residuals $\{e_{ij} = y_{ij} - \widehat{\boldsymbol{\beta}}^T\mathbf{x}_{ij}\}$ for all relations, $\mathbf{1}_{[\{i,j\}\cap\{k,l\}\neq\varnothing]}$ is an $\big(n(n-1) \times n(n-1)\big)$ matrix of indicators denoting which relation pairs share an actor, and '$\circ$' denotes the matrix Hadamard (entry-wise) product. The estimator $\widehat{\Omega}_{DC}$ can be seen as that which takes the empirical covariance of the residuals defined by $\mathbf{e}\mathbf{e}^T$ and systematically introduces zeros to

enforce the non-overlapping pair independence assumption. We refer to the covariance estimator $\widehat{\Omega}_{DC}$ as the **dyadic clustering (DC) covariance estimator** as it owes its derivation to the extensive literature on "cluster-robust" standard error estimates. Restricting the covariances in $\widehat{\Omega}_{DC}$ between non-overlapping relations to be zero makes this estimator similar to that resulting from a two-way clustering approach which clusters on each relation sender (i.e. the rows of Ξ) and also clusters on each relation receiver (i.e. the columns of Ξ).

When the $\boldsymbol{\beta}$ estimator is that based on ordinary least squares (i.e. that associated with (3.3)), Fafchamps and Gubert (2007) propose a sandwich variance estimator for $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ based on the DC covariance estimator, which is equal to

$$\widehat{V}_{DC} = (X^T X)^{-1} X^T \widehat{\Omega}_{DC} X (X^T X)^{-1}. \tag{3.6}$$

We will refer to this as the DC estimator of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$. Aronow et al. (2015) show that $\widehat{V}_{DC}$ is consistent by showing that as the number of actors n grows, the number independent pairs of actors grows with n^4 whereas the number of dependent pairs grows with n^3. Tabord-Meehan (2018) gives a general theoretical treatment of $\widehat{V}_{DC}$ under the exchangeability assumption examined in this chapter.

The DC estimator of the variance in (3.6) is widely used and has the attractive properties that it is asymptotically consistent and theoretically robust to a wide range of error dependence structures (making minimal assumptions). However, we contend its utility is limited in practice for several reasons. First, the DC approach estimates each of the $\mathcal{O}(n^3)$ nonzero covariance elements separately with a *unique residual product*: e.g. $\widehat{\mathrm{Cov}}(\xi_{ij}, \xi_{ik}) = e_{ij} e_{ik}$. Thus, $\widehat{V}_{DC}$ is inherently more variable than more parsimonious estimators, such as the one proposed in this chapter. Secondly, only when there is extreme heterogeneity in the true covariance structure is the DC method ideal, and this method will suffer a loss of efficiency otherwise. Lastly, methods are not currently available to adjust estimates of $\boldsymbol{\beta}$ using $\widehat{\Omega}_{DC}$ as in, for example, feasible generalized least squares.

3.3 Standard errors under exchangeability

In this section, we propose a novel estimator for $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ that leverages an exchangeability assumption in the estimation of Ω. In short, this assumption induces structure among portions of the covariance matrix Ω corresponding to subsets of the relations with a similar arrangement, and pools information within these subsets. For this section, we continue discussion in terms of data sets containing a single relation $Y = Y^{(1)}$, and discuss extensions of our proposed methodology to arrays in Section 3.5.

3.3.1 Exchangeability in relational models

A common modeling assumption for relational and array structured errors is exchangeability. Defined by de Finetti for a univariate sequence of random variables, exchangeability was generalized to array data and relational data by Hoover (1979) and Aldous (1981). The errors in a relational data model are jointly exchangeable if the probability distribution of the error array, Ξ, is invariant under any permutation of the rows and columns. Mathematically, this means

$$\mathrm{P}(\Xi) = \mathrm{P}(\Pi(\Xi)),$$

where $\Pi(\Xi) = \{\xi_{\pi(i)\pi(j)}\}$ is the error array with its rows and columns reordered according to permutation operator π. Intuitively, exchangeability in the context of linear regression on an array simply means the observed covariates are sufficiently informative such that the ordering of the row and column labeling in the error array is uninformative. Each of the conditionally independent parametric network models discussed in the introduction have this joint exchangeability property (see Hoff (2008) and Bickel and Chen (2009) for further discussion).

3.3.2 Impact of exchangeability on covariance structure

Under exchangeability, the covariance matrix Ω has at most six unique elements. To see this result intuitively, note that any relation has five distinguishable types of covariance configurations involving another relation, plus one variance term associated with the relation itself. Figure 3.1

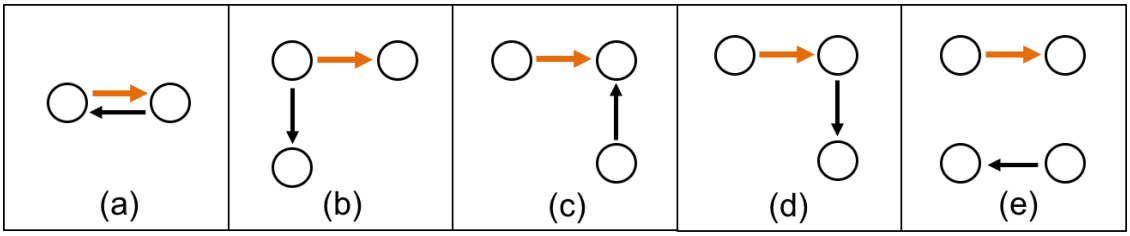


Figure 3.1: Five distinguishable configurations of relation pairs involving the bold orange relation in an exchangeable relational model: (a) reciprocal relations; (b) relations share common sender; (c) relations share common receiver; (d) shared actor is the sender of one relation and receiver of the other; (e) no shared actors among the two relations.

shows the five distinguishable configurations of relation pairs that comprise the covariance structure. If a probability model for Ξ is jointly exchangeable, then all entries ξ_{ij} are marginally identically distributed under the model. This implies that each of the covariances corresponding to a particular configuration in Figure 3.1 (plus the variance term) should have the same value across all possible actor labels. Li and Loken (2002) noted three specific random effects models have this covariance structure, however, we formalize and extend this observation showing that this result holds for *any* jointly exchangeable distribution over Ξ.

Proposition 7. *If a probability model for a directed relational matrix Ξ is jointly exchangeable and has finite second moments, then the covariance matrix of Ξ contains at most six unique values.*

Proof: Consider a probability model for a directed relational matrix Ξ that satisfies the joint exchangeability and second moment criteria defined above. For any four, possibly non-unique, actors $\{i, j, k, l\}$, observe that the covariance between the errors ξ_{ij} and ξ_{kl} takes one of the following six values, depending on the relationships between the actor indices:

- $\text{Var}(\xi_{ij})$ if $i = k$ and $j = l$;
- $\text{Cov}(\xi_{ij}, \xi_{ji})$ if $i = l$ and $j = k$;
- $\text{Cov}(\xi_{ij}, \xi_{il})$ if $i = k$ and $j \neq l$;
- $\text{Cov}(\xi_{ij}, \xi_{kj})$ if $i \neq k$ and $j = l$;
- $\text{Cov}(\xi_{ij}, \xi_{ki})$ if $i = l$ and $j \neq k$;
- $\text{Cov}(\xi_{ij}, \xi_{kl})$ if $i \neq k$ and $j \neq l$.

Now consider an arbitrary permutation operation $\pi(\cdot)$ of the entire actor set $\{1, ..., n\}$. Note that exchangeability implies the bivariate distribution of the pair (ξ_{ij}, ξ_{kl}) must be the same as distribution of $(\xi_{\pi(i)\pi(j)}, \xi_{\pi(k)\pi(l)})$. Thus, the covariance of $\xi_{\pi(i)\pi(j)}$ and $\xi_{\pi(k)\pi(l)}$ must equal that of the original pair:

$$\text{Cov}(\xi_{ij}, \xi_{kl}) = \text{Cov}(\xi_{\pi(i)\pi(j)}, \xi_{\pi(k)\pi(l)}) \quad \text{for any } i, j, k, l.$$

By exchangeability this is true for all permutations $\pi(\cdot)$, establishing the result. □

To illustrate the correspondence between joint exchangeability and the covariance entries, consider the bilinear mixed effects network regression model proposed in Hoff (2005). This model uses an inner product measure to model the error structure in relations and can be expressed as follows:

$$y_{ij} = \boldsymbol{\beta}^T \mathbf{x}_{ij} + \xi_{ij}; \qquad \xi_{ij} = a_i + b_j + z_i^T z_j + \gamma_{(ij)} + \epsilon_{ij}; \tag{3.7}$$

$$(a_i, b_i) \sim \mathbf{N}_2(0, \Sigma_{ab}); \qquad \Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \rho_{ab}\sigma_a\sigma_b \\ \rho_{ab}\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix};$$

$$z_i, z_j \sim \mathbf{N}_d(0, \sigma_z^2 I_d); \qquad \gamma_{(ij)} = \gamma_{(ji)} \sim \mathbf{N}(0, \sigma_\gamma^2); \qquad \epsilon_{ij} \sim \mathbf{N}(0, \sigma_\epsilon^2).$$

where a_i, b_j, z_i, z_j, and ϵ_{ij} are independent. Note that $\mathbb{E}[\xi_{ij}] = 0$.
As presented in Hoff (2005), the elements of $\text{V}[\boldsymbol{\Xi}_{\boldsymbol{v}}] = \text{V}[\boldsymbol{Y}_{\boldsymbol{v}}|X]$ are

- $\text{Var}(\xi_{ij}) = \sigma_a^2 + \sigma_b^2 + d\sigma_z^4 + \sigma_\gamma^2 + \sigma_\epsilon^2$,
- $\text{Cov}(\xi_{ij}, \xi_{kj}) = \sigma_b^2$,
- $\text{Cov}(\xi_{ij}, \xi_{ji}) = 2\rho_{ab}\sigma_a\sigma_b + d\sigma_z^4 + \sigma_\gamma^2$,
- $\text{Cov}(\xi_{ij}, \xi_{ki}) = \text{Cov}(\xi_{ij}, \xi_{jk}) = \rho_{ab}\sigma_a\sigma_b$,
- $\text{Cov}(\xi_{ij}, \xi_{il}) = \sigma_a^2$,
- $\text{Cov}(\xi_{ij}, \xi_{kl}) = 0$.

Note that there are six unique terms, corresponding to the five relation pair configurations shown in Figure 3.1 and a variance term. Moreover, these terms depend only on the population-level parameters of the data generating process and not on individual-level latent variables.

Like the results in Hoff (2005), our work draws on a much deeper, general literature on variance decompositions for structured and symmetric models. In regard to symmetry, a related notion to exchangeability, Dawid (1988) states that "the specification of the relevant symmetry represents a pre-modelling phase from which many important consequences flow." Our work leverages these symmetries, assuming only, again quoting Dawid (1988), that there is "no reason to consider the observations in any one order rather than any other." Work by Li and Loken (2002), Li et al. (2002), and Li (2006) generalize the social relations model (SRM) of Warner et al. (1979) to describe the family of symmetric probability distributions for dyadic data. Though these approaches confirm our findings on the gains of assuming exchangeability, their approach to modeling the covariance structure is quite different. These approaches draw inspiration from the variance decomposition literature in statistics. This motivation leads to developing hypothesis tests that explore restrictions on the symmetries (i.e. invariance to transformations) as a null hypothesis, but impose a parametric form on the error terms (e.g. involving sender, receiver, and pairwise effects in the Warner et al. (1979) social relations model) and in some cases assume a Gaussian likelihood. In contrast, our motivation comes from econometric methods for nonparametric standard error estimation. As a result, we leverage the exchangeability assumption only to simplify the existing estimating equation uncertainty estimates, rather than attempt to fully specify a probability distribution for the data.

3.3.3 Covariance matrices of exchangeable relational arrays

Proposition 7 implies that at most six parameters are required to describe the dependence structure arising from jointly exchangeable relational models. Thus, we introduce a new class of covariance matrices Ω_E which contain five unique nonzero entries: one variance parameter σ^2 along the diagonal of Ω_E and four covariance parameters $\{\phi_a, \phi_b, \phi_c, \phi_d\}$ associated with (a-d) in Figure 3.1. Similar to the DC covariance model, we assume non-overlapping directed pairs of relations are independent, such that $\text{Cov}(\xi_{ij}, \xi_{kl}) = 0$, corresponding to (e) in Figure 3.1 and implying $\phi_e = 0$. Though there may be association between non-overlapping pairs of relations, we expect this dependence to be small compared to dependence between pairs

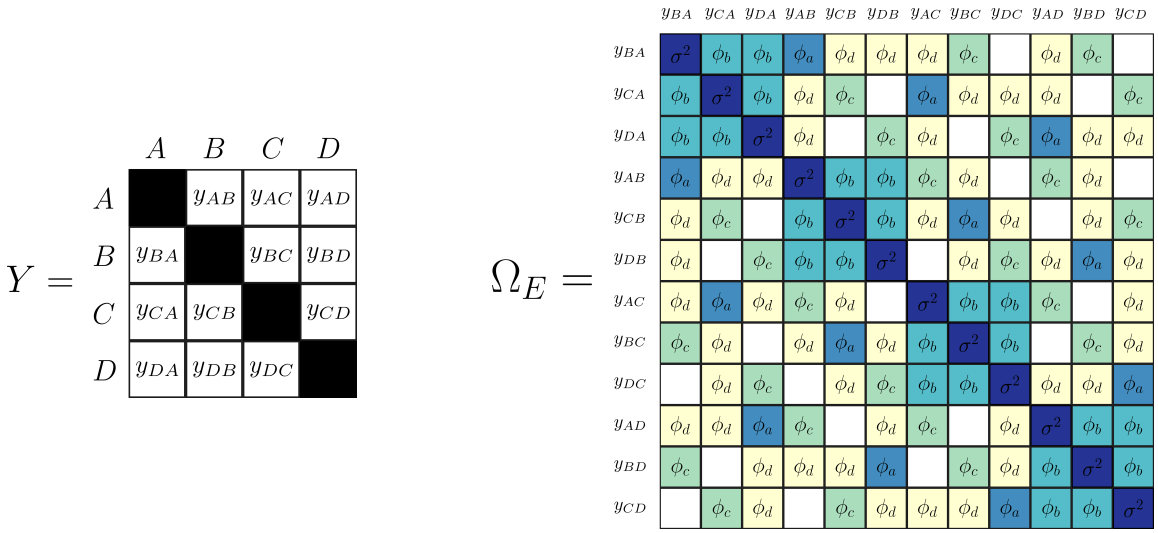


Figure 3.2: Consider a matrix Y containing the relations among four actors $\{A, B, C, D\}$ shown on the left. Since the relation between an actor and itself is undefined, the diagonal entries (blacked out in the picture) are not regarded as part of Y. Assuming joint exchangeability of the actors and that relations involving non-overlapping sets of actors are independent, the covariance matrix Ω_E contains five unique values.

of relations that share a member such that our assumption is reasonable. Figure 3.2 shows the structure of Ω_E for a relational matrix with four actors $\{A, B, C, D\}$. We formally define the class Ω_E below.

Definition 8. *An exchangeable covariance matrix is defined as* $\Omega_E = \mathbb{E}[\boldsymbol{\Xi}_{\boldsymbol{v}} \boldsymbol{\Xi}_{\boldsymbol{v}}^T]$ *arising from mean-zero random vector* $\boldsymbol{\Xi}_{\boldsymbol{v}} = \text{vec}(\Xi)$*, where* Ξ *is a jointly exchangeable random matrix with* ξ_{ij} *independent* ξ_{kl} *whenever* $\{i, j\} \cap \{k, l\} \neq \varnothing$*.* Ω_E *has five unique terms consisting of a variance and four covariances:* $\{\sigma^2, \phi_a, \phi_b, \phi_c, \phi_d\}$.

We now present a theorem that states that, for a linear model with error covariance matrix in the class of Ω_E, the OLS estimate of the coefficients $\boldsymbol{\beta}$ is asymptotically normal. This theorem is similar to that in Tabord-Meehan (2018), although under different conditions. Our asymptotic regime is the addition of actors to the relational data set, leading to asymptotics in n, where we treat the matrix X as a random variable. To examine the asymptotic behavior of $\widehat{\boldsymbol{\beta}}$, we must make some assumptions about the distribution of X. As each covariate pertains to entries in a relational array, it is natural to assume that there may be dependence among the rows in X. In the context of trade between countries, for example, if scalar $x_{jk}^{(1)}$ measures the difference in GDP between France (j) and Germany (k), we expect the difference in GDP between Spain (l) and France, $x_{lj}^{(1)}$, to be correlated with the former $x_{jk}^{(1)}$. Thus, we assume the *rows* of the matrix X are jointly exchangeable, meaning that, any permutation $\pi(.)$ of the rows $\{\mathbf{x}_{jk}\}_{j,k=1}^n$ to $\{\mathbf{x}_{\pi(j)\pi(k)}\}_{j,k=1}^n$ leaves

the distribution of matrix X invariant. As with the dependence in the errors, we assume that two *rows* of X that correspond to relations which do not share an actor are independent, that is row $\mathbf{x}_{ij}^T$ is independent row $\mathbf{x}_{kl}^T$ whenever $\{i,j\} \cap \{k,l\} \neq \varnothing$. This dependence in the rows of matrix X (along with some assumptions on the finiteness of its moments) implies the following:

$$\sum_{(jk,lm)\in\Theta_i} \frac{\mathbf{x}_{jk}\mathbf{x}_{lm}^T}{|\Theta_i|} \xrightarrow{\mathbb{P}_X} M_i, \quad i \in \{0,a,b,c,d\}, \tag{3.8}$$

where Θ_i is the set of *pairs* of relations (jk,lm) that share a member in the i^{th} manner and '0' refers to self-relation (i.e. variance).

Theorem 9. *Define the following data generating process:*

(A1) The true data generating model is $\mathbf{Y}_{\boldsymbol{v}} = X\boldsymbol{\beta} + \boldsymbol{\Xi}_{\boldsymbol{v}}$, where the errors $\boldsymbol{\Xi}_{\boldsymbol{v}}$ are mean-zero with exchangeable covariance matrix as defined in Definition 8.

(A2) At least one of $\{\phi_b, \phi_c, \phi_d\}$ is nonzero.

In addition, consider the following regularity conditions:

(B1) The covariate matrix X has rows that are jointly exchangeable with at least one of $\{M_i\}_{i\in\{b,c,d\}}$ in (3.8) *nonzero, and where row $\mathbf{x}_{ij}^T$ is independent row $\mathbf{x}_{kl}^T$ whenever $\{i,j\} \cap \{k,l\} \neq \varnothing$.*

(B2) The fourth moments of each the errors and the covariates are bounded: $\mathbb{E}[|\xi_{jk}|^4] < C < \infty$ and $\max_{l\in\{1,2,\ldots,p\}} \mathbb{E}[|x_{jk}^{(l)}|^4] < C' < \infty$ where $\mathbf{x}_{jk} = [x_{jk}^{(1)}, \ldots, x_{jk}^{(p)}]^T$.

(B3) The errors Ξ and covariates X are independent.

(B4) X is full rank.

Given $(A1) - (A2)$ and $(B1) - (B4)$, the ordinary least squares estimate $\widehat{\boldsymbol{\beta}}$ is asymptotically normal:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, M_0^{-1}\big(\phi_b M_b + \phi_c M_c + 2\phi_d M_d\big)M_0^{-1}),$$

where $\{M_i\}_{i\in\{0,b,c,d\}}$ are as in (3.8) *and '$\to_d$' denotes element-wise convergence in distribution.*

The proof of Theorem 9 is given in Appendix B.2. Note that only the covariances $\{\phi_b, \phi_c, \phi_d\}$ appear in asymptotic variance of $\widehat{\boldsymbol{\beta}}$. This results from the fact that there are an order of magnitude more of terms $\{\phi_b, \phi_c, \phi_d\}$ in the covariance matrix Ω_E than there are of the terms $\{\phi_a, \sigma^2\}$. In particular, in Ω_E there are $n(n-1)(n-2)$ pairs of relations (ξ_{ij}, ξ_{kl}) of each of type (b) and type (c), $2n(n-1)(n-2)$ pairs of type (d), and $n(n-1)$ pairs of each of type (a) and σ^2. We make the assumption that at least one of the covariances $\{\phi_b, \phi_c, \phi_d\}$ is nonzero. Should the assumption be violated, then all $\binom{n}{2}$ dyadic pairs of the form (ξ_{ij}, ξ_{ji}) are independent of one another, and the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ follows from the usual independent data arguments. In this case, the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ is of rate n with asymptotic variance $M_0^{-1}\left(\sigma^2 M_0 + \phi_a M_a\right) M_0^{-1}$. The canonical case of independent and identically distributed errors is recovered when $\phi_a = \phi_b = \phi_c = \phi_d = 0$. In the canonical case, $\widehat{\boldsymbol{\beta}}$ has asymptotic variance $\sigma^2 M_0^{-1}$ occurring at rate n. Note that both of these final cases are rate n (and not $\sqrt{n}$) since there are $n(n-1)$ entries in $\boldsymbol{Y_v}$.

3.3.4 Exchangeable covariance estimator

As emphasized above, the DC estimators in (3.5) and (3.6) estimate each nonzero element in Ω_{DC} using a single product of residuals. Here we introduce novel estimators inspired by the covariance structure Ω_E associated with relations are jointly exchangeable. Specifically we consider estimates of Ω in the class of exchangeable covariance matrices, as in Definition 8. Our new **exchangeable (EXCH) covariance estimator** $\widehat{\Omega}_E$, and corresponding estimator of $\mathrm{V}[\widehat{\boldsymbol{\beta}}]$ can be written, respectively, as

$$\widehat{\Omega}_E = \widehat{\sigma}^2 I_{n(n-1)} + \sum_{s=a}^{d} \widehat{\phi}_s \mathcal{S}_s, \quad \text{and} \quad \widehat{V}_E = (X^T X)^{-1} X^T \widehat{\Omega}_E X (X^T X)^{-1}, \tag{3.9}$$

where $\mathcal{S}_s$ denotes the $(n(n-1) \times n(n-1))$ binary matrix with 1s in the entries corresponding to relation pairs of type $s \in \{a, b, c, d\}$ as defined in Figure 3.1.

We propose estimating the five parameters in Ω_E by averaging the residual products across pairs having the same index configurations corresponding to (a)-(d) in Figure 3.1. These empirical mean estimates can be expressed

- $\widehat{\sigma}^2 = \widehat{\mathrm{Var}}(\xi_{ij}) = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} e_{ij}^2$,
- $\widehat{\phi}_a = \widehat{\mathrm{Cov}}(\xi_{ij}, \xi_{ji}) = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} e_{ij} e_{ji}$,

- $\widehat{\phi}_b = \widehat{\text{Cov}}(\xi_{ij}, \xi_{kj}) = \frac{1}{n(n-1)(n-2)} \sum_i \sum_{j \neq i} e_{ij} \Big(\sum_{k \neq j} e_{kj} - e_{ij} \Big)$,
- $\widehat{\phi}_c = \widehat{\text{Cov}}(\xi_{ij}, \xi_{ik}) = \frac{1}{n(n-1)(n-2)} \sum_i \sum_{j \neq i} e_{ij} \Big(\sum_{k \neq i} e_{ik} - e_{ij} \Big)$,
- $\widehat{\phi}_d = \widehat{\text{Cov}}(\xi_{ij}, \xi_{ki}) = \widehat{\text{Cov}}(\xi_{ij}, \xi_{jl}) = \frac{1}{2n(n-1)(n-2)} \sum_i \sum_{j \neq i} e_{ij} \Big(\sum_{k \neq i} e_{ki} + \sum_{k \neq j} e_{jk} - 2e_{ji} \Big)$.

We implement the above estimator, along with many methods to follow in this chapter, in the `R` software package `netregR` Marrs et al. (2018).

Even if the underlying data generating model is not jointly exchangeable, the proposed estimator will work well if the variability in the covariances among relations of the same type (i.e. (a)-(d) in Figure 3.1) is small. In this case, it is likely that the reduction in estimation variance that arises from pooling will outweigh the small bias introduced in the estimation of each covariance entry.

3.4 Evaluating the exchangeable estimator

In this section, we theoretically and empirically evaluate the properties of our estimators in (3.9) for data with a single matrix of relations (i.e. $R = 1$). We first prove that our variance estimator $\widehat{V}_E$ is consistent for the true variance of $\widehat{\boldsymbol{\beta}}$ when the data generating process is jointly exchangeable. Then, in the spirit of, for example, Kauermann and Carroll (2001), we show that the mean-square error (MSE) of $\widehat{V}_E$ is lower than that of $\widehat{V}_{DC}$ with high probability. We then present simulation evidence of improved inference for $\widehat{\boldsymbol{\beta}}$ by simulating data from both exchangeable and non-exchangeable generative models.

3.4.1 Consistency of the exchangeable estimator

We begin theoretical justification of the exchangeable estimator $\widehat{V}_E$ by stating that exchangeable covariance estimator is consistent for the true variance of the coefficients, as the number of actors increases.

Theorem 10. *Under the conditions of Theorem 9, the exchangeable covariance estimator is consistent in the sense that*

$$n\widehat{V}_E - nV[\widehat{\boldsymbol{\beta}}] \to_p 0 \quad as \;\; n \to \infty, \tag{3.10}$$

where '$\to_p$' denotes element-wise convergence in probability.

The proof of Theorem 10 is given in Appendix B.3.

3.4.2 MSE of DC and exchangeable estimators

We continue our theoretical justification of the exchangeable estimator $\widehat{V}_E$ by showing that the MSE of the exchangeable covariance estimator V_E is lower than that of the dyadic clustering covariance estimator V_{DC} with high probability. We evaluate the MSE conditional on X, and then evaluate the probability that the difference is positive for random X. We show that the probability that the MSE of the exchangeable estimator is less than that of the dyadic clustering estimator tends to 1 as n tends to infinity.

Theorem 11. *Under the assumptions of Theorem 9 and for covariate matrix X with bounded eighth moments, that is $\max_{l\in\{1,2,\ldots,p\}} \mathbb{E}\left(|x_{jk}^{(l)}|^8\right) < C' < \infty$, the MSE of the exchangeable estimator is less than that of the dyadic clustering estimator with probability approaching 1, that is*

$$\mathbb{P}_X\left(MSE_\xi\left(\widehat{V}_{DC}|X\right) \geq MSE_\xi\left(\widehat{V}_E|X\right)\right) \to 1.$$

The proof of Theorem 11 is provided in Appendix B.4.

3.4.3 Simulation evidence

We performed a simulation study to compare the performance of our estimator to the DC estimator under various scenarios. We consider three different data generating models for the errors $\Xi = \{\xi_{ij}\}$: (i) independent and identically distributed errors, (ii) errors generated from the (exchangeable) bilinear mixed effects model of Hoff (2005) shown in (3.7) and (iii) errors generated from a non-exchangeable model. We note that the exchangeable error model is a generalization of the "additive common shocks" error structure used in simulation studies to justify the DC estimator (Cameron and Miller (2014); Aronow et al. (2015); Tabord-Meehan (2018)).

The non-exchangeable model included systematic noise in the upper-left quadrant of the relational error matrix Ξ. Since noise was added to actor relations in the same position in Ξ in each simulation run, the distribution of the relations was not exchangeable: the distribution of the errors would be different for a reordering of the rows and columns. Note that the non-exchangeable model violates the assumption of both exchangeable and DC estimators that ξ_{jk} is independent of ξ_{lm} when $\{i,j\} \cap \{k,l\} \neq \varnothing$.

For each simulation setting, we employed the following three-covariate regression model:

$$y_{ij} = \beta_1 + \beta_2 \mathbf{1}_{x_{2i} \in C} \mathbf{1}_{x_{2j} \in C} + \beta_3 |x_{3i} - x_{3j}| + \beta_4 x_{4ij} + \xi_{ij}. \tag{3.11}$$

In this model, β_1 is an intercept; β_2 is a coefficient on a binary indicator of whether individuals i and j both belong to a pre-specified class C; β_3 is a coefficient on the absolute difference of a continuous, actor-specific covariate x_{3i}; and β_4 is that for a pair-specific continuous covariate x_{4ij}. We note here that the matrix X satisfies the jointly exchangeable assumption (B1) in Theorem 9. For the entirety of the study, we fixed $\boldsymbol{\beta}$ at a single set of values. Since the variance of $\widehat{\boldsymbol{\beta}}$ explicitly depends on X, we generated 500 random design matrices X for each sample size of actors, and for each design matrix simulated 1,000 error matrices under each of the three models to assess the variability of the standard error estimates and accuracy of the subsequent confidence intervals for $\boldsymbol{\beta}$. For additional details on the simulation study procedure, see Appendix B.5.

Figures 3.3 - 3.5 display the coverage probabilities for 95% confidence intervals for each β for the three error settings. Along with the dyadic clustering (DC) and exchangeable (EXCH) estimators, we also include the standard heteroskedasticity consistent (HC) estimator as a baseline, as in Aronow et al. (2015).

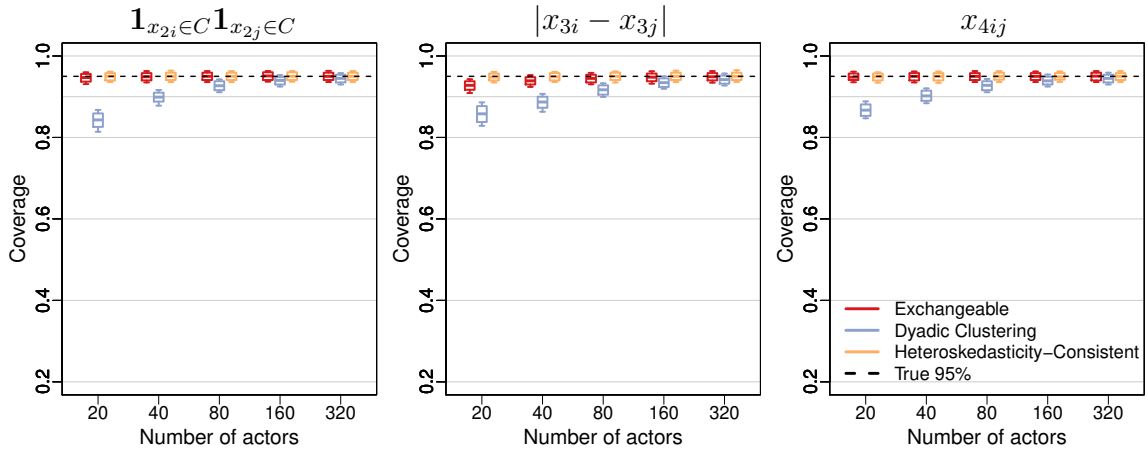


Figure 3.3: (IID Errors) Probability β is in 95% confidence interval across 500 random X draws when the errors are independent and identically distributed. Lines in the boxplots denote the median coverage, the box denotes the middle 80% of coverages, and the whiskers denote the middle 95% of coverages across the set of design matrices.

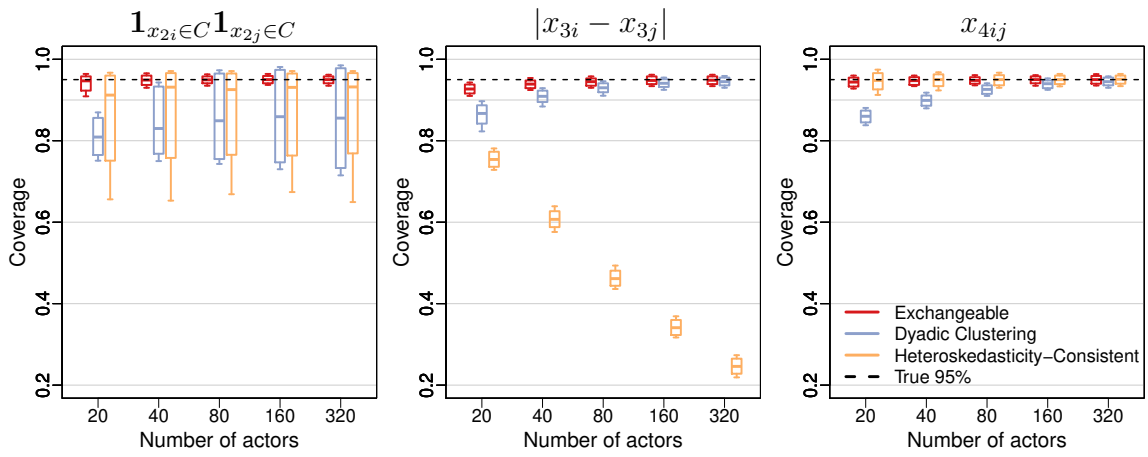


Figure 3.4: (Exchangeable Errors) Probability β is in 95% confidence interval across 500s random X draws when the errors are generated according to the exchangeable bilinear effects model. Lines in the boxplots denote the median coverage, the box denotes the middle 80% of coverages, and the whiskers denote the middle 95% of coverages across the set of design matrices.

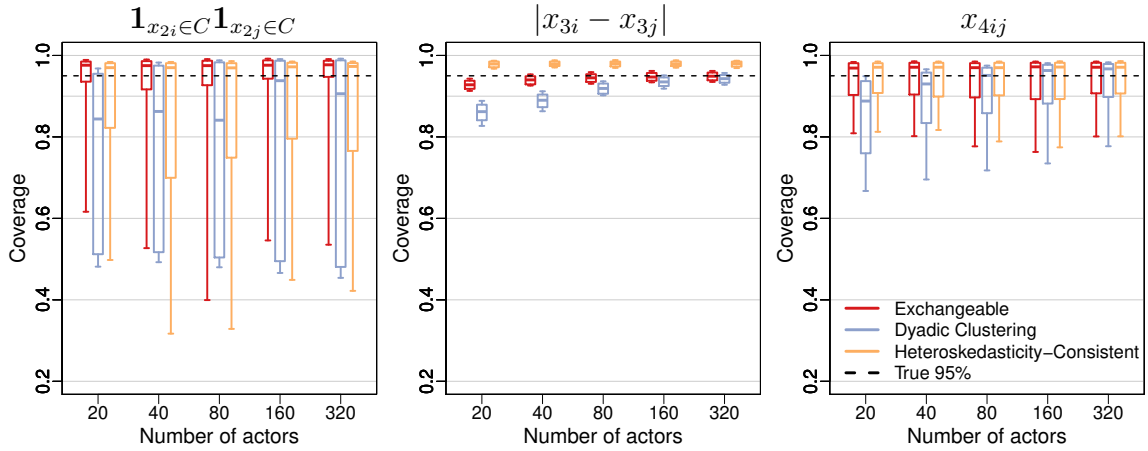


Figure 3.5: (Non-exchangeable Errors) Probability β is in 95% confidence interval across 500 random X draws when the errors are generated from a non-exchangeable network model. Lines in the boxplots denote the median coverage, the box denotes the middle 80% of coverages, and the whiskers denote the middle 95% of coverages across the set of design matrices.

We draw two key conclusions from our simulations. First, our proposed approach performs extremely well compared to the DC and HC alternatives, even when the assumption of exchangeability of the errors is violated. Specifically, we see that the EXCH estimator produces confidence intervals with nominal, or near nominal, coverage for a variety of data generating processes. In addition, we see the variability in coverage across different X realizations for the EXCH estimator is substantially smaller than that for the other estimators. Intuitively the observed reduction in variability is a result of the averaging inherent in the

EXCH estimator. In particular, the EXCH estimator replaces DC's $\mathcal{O}(n^3)$ unique residual products with five averages over subsets of these products. In Appendix B.5.1, we plot the standard deviation of the EXCH and DC standard error estimates, where we clearly see the reduction in variability of the EXCH estimator relative to the DC estimator. We also plot the expected error (given X) of the DC and EXCH standard error estimates relative to the true standard errors. We find that both estimators generally underestimate the true standard errors (and thus confidence interval width), although the EXCH estimator underestimates to a significantly lesser degree. Thus, the price of the robustness of the DC estimator is a loss of efficiency that results in anti-conservative confidence intervals, i.e. undercoverage.

Returning to the coverage plots, it is interesting that even when the exchangeable assumption is incorrect, as in Figure 3.5, we see better performance from the EXCH estimator than the others. This is despite the fact that, in this heterogeneous case, we expect the empirical DC estimator to perform best. The performance of the EXCH estimator under non-exchangeable errors suggests that the reduction in the variability of the covariance entry estimates in the exchangeable estimator can outweigh the covariance model misspecification.

The second key observation we glean from the study is that the type of covariate (e.g. continuous actor-level characteristic versus product of binary indicators) affects the performance of all standard error estimators. For example, Figures 3.4 and 3.5 show that when there is structure in the errors, the variability in the confidence interval coverage across design matrices is far greater for the binary covariate than for either of the continuous covariates. Focusing specifically on the boxes representing the middle 80% of coverage levels across the 500 simulations associated with the binary coefficient (left-most plots in Figures 3.3 through 3.5), we see the EXCH estimator coverage varies from about 93-98%, whereas the DC estimator varies between 50-95% with no improvement as the sample size n increases.

3.5 Regressions involving relational arrays

In this section we extend our discussion of exchangeable estimators to the case when $R > 1$. We introduce three notions of exchangeability for relational array data and discuss models consistent with these assumptions. We separately consider the cases when the underlying model for the error array is exchangeable along the third dimension and that when it is not. Figure 3.6 illustrates the former case and two variations

of the latter. Before dissecting the spectrum of possible exchangeability assumptions, we first revisit the treatment of error arrays with $R > 1$ by Aronow et al. (2015).

3.5.1 Dyadic clustering

Aronow et al. (2015) examine relational regression standard errors when the third dimension, indexed by r in $Y = \{y_{ij}^{(r)}\}$, denotes time. Data consistent with this structure is, for example, country to country trade over time. Aronow et al. (2015)'s treatment is a direct extension of dyadic clustering in two dimensions: two errors $\xi_{ij}^{(r)}$ and $\xi_{k\ell}^{(s)}$ are assumed to be independent if the associated dyads do not share a member (i.e. $\{i,j\} \cap \{k,\ell\} = \varnothing$), regardless of the third dimension indices r and s. As in the $R = 1$ case, each nonzero covariance entry is estimated by the corresponding residual product, i.e. $\widehat{\text{Cov}}\left(\xi_{ij}^{(r)}, \xi_{k\ell}^{(s)}\right) = e_{ij}^{(r)} e_{k\ell}^{(s)}$. Note that this specification makes no assumptions about the dependence structure along the third dimension.

3.5.2 Exchangeability in the third dimension

Here we consider relational data that are fully exchangeable in the third dimension. Intuitively, the numbering of the row, column, and depth indices of an array with this property are uninformative. For example, consider the case where the relational array Y represents the quantity of trade between pairs of countries, decomposed by various categories of goods traded (e.g. intangible vs. tangible). Without reason to believe some pairs of good types are more dependent than others, we might be willing to assume the dependence structure along the third dimension is exchangeable.

Define a permutation of the third dimension indices $\nu(.)$ in addition to the row and column permutation $\pi(.)$ defined previously. An array probability model that is jointly exchangeable, as well as exchangeable in the third dimension, has the property that

$$\text{Cov}\left(\xi_{ij}^{(r)}, \xi_{k\ell}^{(s)}\right) = \text{Cov}\left(\xi_{\pi(i)\pi(j)}^{(\nu(r))}, \xi_{\pi(k)\pi(\ell)}^{(\nu(s))}\right). \tag{3.12}$$

It follows that the covariance matrix, denoted $\Omega_{Ea} = V[\boldsymbol{\Xi}_{\boldsymbol{v}}]$, consists of 10 distinct nonzero parameters, corresponding to two submatrices Ω_1 and Ω_2, which each have exchangeable structure as in Definition 8. Along the diagonal of Ω_{Ea} there are R instances of the $n(n-1) \times n(n-1)$ matrix Ω_1 represents covariance between observations that share the same third index, i.e. $r = s$. The off-diagonal blocks of Ω_{Ea} are

populated with a second $n(n-1) \times n(n-1)$ exchangeable error matrix Ω_2 for errors that do not share the same third index, i.e. $r \neq s$. This structure is depicted in Figure 3.6(a). As previously, we propose estimating each of the 10 unique values with the average of the corresponding residual products.

Jointly exchangeable models that model the slices of the error array $\left\{\Xi^{(1)}, \Xi^{(2)}, ..., \Xi^{(R)}\right\}$ as independent constitute a subclass of the models with full exchangeability. Specifically, they make the additional assumption that $\text{Cov}\left(\xi_{ij}^{(r)}, \xi_{k\ell}^{(s)}\right) = 0$ for $r \neq s$. In Figure 3.6(a), an assumption of independence along the third dimension corresponds to $\Omega_2 = 0$.

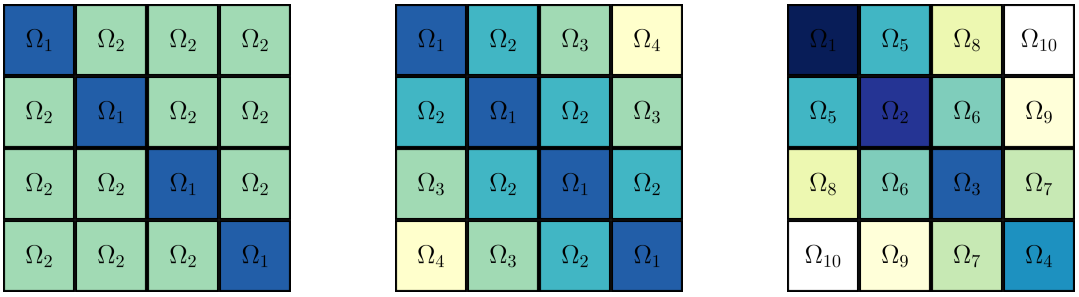


Figure 3.6: Covariance matrices $\Omega = \text{V}[\boldsymbol{\Xi_v}]$ for exchangeable arrays with depth $R = 4$ where $\boldsymbol{\Xi}_{\boldsymbol{v}}^T = \left((\boldsymbol{\Xi}_{\boldsymbol{v}}^{(\mathbf{1})})^T, (\boldsymbol{\Xi}_{\boldsymbol{v}}^{(\mathbf{2})})^T, (\boldsymbol{\Xi}_{\boldsymbol{v}}^{(\mathbf{3})})^T, (\boldsymbol{\Xi}_{\boldsymbol{v}}^{(\mathbf{4})})^T\right)$. All matrices are symmetric, where the (i, j) block denotes $\text{Cov}(\boldsymbol{\Xi}_{\boldsymbol{v}}^{(\boldsymbol{i})}, \boldsymbol{\Xi}_{\boldsymbol{v}}^{(\boldsymbol{j})})$. Subfigure (a) corresponds to full exchangeability yielding two unique blocks, (b) corresponds to no exchangeability in the third dimension with stationarity assumption yielding $R = 4$ unique blocks, and (c) corresponds to no exchangeability in the third dimension yielding $\binom{R}{2} + R = 10$ unique blocks. Each block contains five unique nonzero terms as in Ω_E in Figure 3.2.

3.5.3 Partial exchangeability or no exchangeability in the third dimension

The assumption of exchangeability along the third dimension can be unnatural and inappropriate for certain data sets, so here we consider relaxing the fully exchangeable assumption introduced Section 3.5.2. Consider again the quantity of trade between countries i and j as the relational response, except where trade decomposed by time period rather than by good type. We would expect the temporal index in the third dimension to be non-exchangeable, as we might expect errors associated with nearby time periods will be more dependent than those far apart.

Here we consider arrays which are jointly exchangeable along the rows and columns only, such that the ordering of the array in the third (depth) dimension must remain the same for the probability distribution

to remain invariant. Intuitively, this property corresponds to one where the labeling of rows and columns is inconsequential, but the labeling of the third dimension is material. This exchangeability assumption implies

$$\text{Cov}\left(\xi_{ij}^{(r)}, \xi_{k\ell}^{(s)}\right) = \text{Cov}\left(\xi_{\pi(i)\pi(j)}^{(r)}, \xi_{\pi(k)\pi(\ell)}^{(s)}\right). \tag{3.13}$$

The full covariance matrix, denoted $\Omega_{Ec} = \text{V}[\boldsymbol{\Xi}_{\boldsymbol{v}}]$, contains a separate $n(n-1)\times n(n-1)$ exchangeable covariance matrix for each of the $\binom{R}{2}$ unique third index pairings and each of the R diagonal variance matrices (see Figure 3.6(c)). Covariance matrices of this form contain $5\left(\binom{R}{2} + R\right)$ unique parameters.

This type of exchangeability assumption is extremely unrestrictive. Specifically, it places no constraints on the evolution of the dependence along the third dimension. However, a more restrictive assumption specifying the relationships among the covariances in the third dimension may be appropriate when we expect the behavior in this dimension to vary in a particular manner. For example, if the third dimension corresponds to different time periods, it may be reasonable to assume stationarity along the third dimension, whereby the covariance across time periods only depends on the absolute difference in the time indices. In this case, there are five unique nonzero covariances for each difference in time $|r-s|$, yielding $5R$ unique nonzero values in the covariance matrix. We denote a covariance matrix with this structure by Ω_{Eb} (see Figure 3.6(b)).

3.6 Patterns in international trade

In this section we demonstrate our exchangeable standard error estimator using data on international trade flow over multiple decades. We fit the model using Generalized Estimating Equations (GEE), which weights the estimating equations, g, in (3.3) by an estimate of the inverse of the "working" covariance matrix of the observations (see, for example, Chapter 8 in Wakefield (2013) for a review of GEE). When the assumed covariance structure is correct, this approach yields an estimator $\widehat{\boldsymbol{\beta}}$ which has improved efficiency over that based on unweighted equations in (3.3). In the remainder of this section, we outline how the exchangeable estimator can be used in a method of moments (weighted least squares) approach to estimation and present results from the international trade data.

3.6.1 Inference via GEE

Inference via GEE proceeds by first specifying a "working" covariance matrix for the errors, which serves as a weight for estimating equations. The choice of the working covariance matrix represents a trade-off between robustness and efficiency. If the working matrix resembles the true underlying covariance structure, then the efficiency of $\widehat{\boldsymbol{\beta}}$ improves over that resulting from the estimating equations in (3.3). Even if the working covariance is misspecified, the standard error estimates for $\widehat{\boldsymbol{\beta}}$ can be 'corrected' using the sandwich standard error estimators with an appropriate estimator $\widehat{\Omega}$. However, these standard error estimates can be unstable if the assumed working structure differs greatly from the truth, which, of course, is unknown in practice (see discussion in Chapter 8 of Wakefield (2013), for example).

The GEE algorithm proceeds as follows. Let W^{-1} be the working covariance matrix, then the estimate of $\boldsymbol{\beta}$ is the solution to the GEE estimating equation

$$XW(\boldsymbol{Y_v} - \boldsymbol{\beta}^T X) = 0,$$

and the corresponding variance estimator of the coefficients is

$$V[\widehat{\boldsymbol{\beta}}] = (X^T W X)^{-1} X^T W \Omega W X (X^T W X)^{-1}.$$

Our estimation algorithm is composed of a two-step iteration procedure. Given initial estimates $\widehat{\boldsymbol{\beta}}^{(0)}$ and corresponding residuals $\widehat{\Xi}^{(0)}$, we iterate between two steps, such that for iteration $\tau + 1$:

1. Solving the estimating equations $X\widehat{W}^{(\tau)} \left(\boldsymbol{Y_v} - X\widehat{\boldsymbol{\beta}}^{(\tau+1)} \right) = 0$, set $\widehat{\boldsymbol{\beta}}^{(\tau+1)} = \left(X^T \widehat{W}^{(\tau)} X \right)^{-1} X^T \widehat{W}^{(\tau)} \boldsymbol{Y_v}$.

2. Use $\widehat{\boldsymbol{\beta}}^{(\tau+1)}$ to calculate $\widehat{\Xi}^{(\tau+1)}$, and obtain estimates $\widehat{\Omega}^{(\tau+1)}$ and $\widehat{W}^{(\tau+1)}$.

These steps are repeated until convergence.

3.6.2 International trade models

We demonstrate the implications of using our exchangeable standard error estimator in a study of international trade between 58 countries. These data were previously analyzed and made available by Westveld

and Hoff (2011)[1]. For each pair of countries, we observe yearly total volume of trade between the two countries for a period from 1981-2000. Following Westveld and Hoff (2011) and Tinbergen (1962), we model (log) trade in a given year using a modified gravity mean model. The gravity model, proposed by Tinbergen (1962), posits that the total trade between countries is proportional to overall economic activity of the countries weighted by the inverse of the distance between them (raised to a power). Following Ward and Hoff (2007) we also add an indicator for whether the nations' militaries cooperated in the given year and a measure of democracy, i.e. polity, which ranges from 0 (highly authoritarian) to 20 (highly democratic).

The complete model has the form:

$$\begin{aligned} \ln \text{Trade}_{ijt} &= \beta_{0t} + \beta_{1t} \ln \text{GDP}_{it} + \beta_{2t} \ln \text{GDP}_{jt} + \beta_{3t} \ln \text{D}_{ijt} \\ &\quad + \beta_{4t} \text{Pol}_{it} + \beta_{5t} \text{Pol}_{jt} + \beta_{6t} \text{CC}_{ijt} + \beta_{7t} (\text{Pol}_{it} \times \text{Pol}_{jt}) + \epsilon_{ijt}, \end{aligned}$$

where $\ln \text{Trade}_{ijt}$ is the (log) volume of trade between countries i and j at time t; $\ln \text{GDP}_{it}$ and $\ln \text{GDP}_{jt}$ are the (log) Gross Domestic Product of nations i and j, respectively; $\ln \text{D}_{ijt}$ is the (log) geographic distance between nations; CC_{ijt} is the measure of cooperation in conflict (coded as +1 if nations were on the same side of a dispute and -1 if they were on opposing sides); and Pol_{it} and Pol_{jt} are the polity measures for i and j, respectively.

We fit the regression model above using the GEE approach, where both the working covariance matrix W^{-1} and the population covariance matrix Ω have the covariance structure Ω_{Ea}, as in panel (a) of Figure 3.6. The estimator of $\boldsymbol{\beta}$ is then based the assumption that the error covariance structure is fully exchangeable. We place no further restrictions on the covariance structure beyond this exchangeability.

The exchangeability assumption underlying our approach differs substantially from assumptions frequently made in analyses of temporal relational data. For example, Westveld and Hoff (2011) propose a model – which we refer to as the hierarchical, longitudinal mixed effects model (HLMEM) – which explicitly decomposes the error term ϵ_{ijt} for each time period and pair of actors into time-dependent sender and receiver effects. The error structure in Figure 3.6(a), in contrast, imposes a temporal structure that implies the covariance is the same between overlapping dyads in different time points.

[1] See https://doi.org/10.1214/10-AOAS403SUPP for data.

We compare our results to those from Westveld and Hoff (2011), as well as to the DC estimator of Aronow et al. (2015) described in Section 3.5.1. Although the DC estimator makes even fewer assumptions about the structure of the error dependence than our method, it cannot be directly used for GEE because the covariance matrix estimator $\widehat{\Omega}_{DC}$ is always singular. Of course, this is by design, as the DC method is aimed at estimating $V[\widehat{\boldsymbol{\beta}}]$ rather than $\widehat{\Omega}_{DC}$ and is inherently post-hoc. Thus, in the following comparison we use ordinary least squares to estimate $\boldsymbol{\beta}$ as in (3.3) and estimate confidence intervals using $\widehat{V}_{DC}$. In the Appendix B.6, we provide a proof that $\widehat{\Omega}_{DC}$ is always singular and in Appendix B.7, we provide a method for efficiently inverting $\widehat{\Omega}_E$.

3.6.3 International trade results

The estimated coefficients and corresponding 95% confidence intervals and posterior credible intervals are shown in Figure 3.7. Coefficient estimates for the HLMEM are posterior medians and 95% credible intervals based on Bayesian Markov chain Monte Carlo. Interpreting Figure 3.7 requires care as (i) there is no ground truth and (ii) we are comparing three different inference paradigms. Nonetheless, focusing on two aspects of Figure 3.7 reveals important insights for practitioners determining which paradigm to use. First, consider the overall trends in estimated $\widehat{\boldsymbol{\beta}}$ across time for the three methods. Fitting using GEE produces coefficients much closer to those estimated for the HLMEM than OLS. In particular, the intercept estimated via OLS is approximately three times larger in magnitude than either the HLMEM or our GEE estimates. We see a similar result with the other coefficient that is relatively constant over time, log GDP of exporter. Both the GEE and the OLS estimates roughly match the temporal trends of the HLMEM estimates, with the exception of the cooperation in conflict variable. For this case, the HLMEM and GEE estimators are both nearly zero from 1990 onward. The OLS estimator, however, demonstrates substantial fluctuations that are not present in the other methods.

The second aspect to notice is confidence interval width. The widths of the confidence intervals for the exchangeable GEE approach are generally comparable to those corresponding to the HLMEM, while the DC interval widths are noticeably larger. The exchangeable GEE approach incorporates information about the covariance structure of the errors when estimating the regression coefficients. The OLS estimate of $\boldsymbol{\beta}$, however, is identical to a GEE estimate when the working covariance $W^{-1} = I$. If the working covariance estimate is close to the (unknown) true covariance we expect efficiency gains in the GEE estimate of $\boldsymbol{\beta}$ over

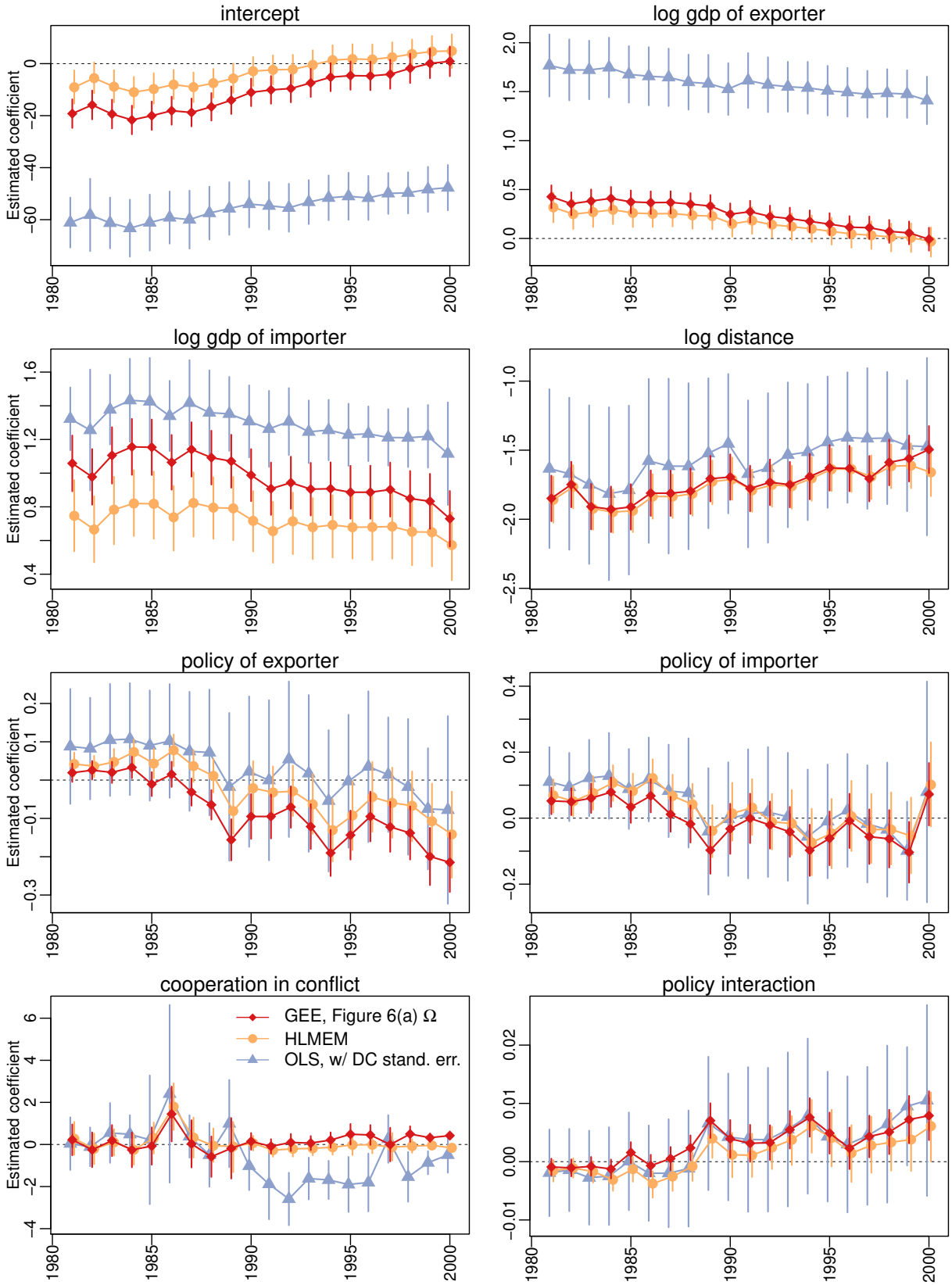


Figure 3.7: Estimated coefficients for at time periods using three different estimation techniques.

Table 3.1: MSPE at $T = 11$ and $T = 20$ of the estimating procedures.

	MSPE	
Estimator	$T = 11$	$T = 20$
Exchangeable GEE	9.85	9.51
HLMEM	12.78	9.77
OLS	18.2	12.85

the OLS estimate. The widths of the HLMEM and GEE intervals also tend to be more consistent across time periods than those from OLS/DC. For the cooperation in conflict variable, for example, the OLS/DC confidence intervals become markedly wider during the upward spikes, one of which, in the late 1980s, is only present in OLS/DC estimate.

The proposed exchangeable GEE approach required significantly less computational resources than those required to estimate the HLMEM. The GEE procedure completed in less than ten minutes, while the Markov chain Monte Carlo proposed by Westveld and Hoff (2011) to estimate HLMEM required almost two days (both procedures were run on a dedicated compute node at 2.50GHz and with 4GB RAM). The procedure for HLMEM generated 55,000 iterations from the Markov chain, resulting in 2,250 samples after burn-in and thinning.

To ensure that the underlying model of the exchangeable GEE procedure is a reasonable representation of the data, we compared the GEE and HLMEM models in a brief out-of-sample performance study. In this study, we estimated both models on the first 10 and 19 years of trade data and used the estimates to predict the trade values in the the 11^{th} and 20^{th} years, respectively. As a baseline, we did the same with ordinary least squares. Table 3.1 provides the mean-square prediction error (MSPE) for the three procedures and two time points. There is a large reduction in MSPE from that based on the OLS estimates to the HLMEM results, and further reduction in MSPE for the proposed exchangeable GEE approach. This suggests the exchangeable model proposed here represents the data at least as well as the HLMEM and significantly better than one which assumes all errors are independent. Further details about the prediction study are provided in Appendix B.9.

3.7 Discussion

This chapter develops a new set of uncertainty estimators for regression models on relational arrays. The proposed estimators strike a balance between making additional assumptions to decrease variability and remaining robust to dependence heterogeneity and model misspecification. We show that the proposed estimators are consistent and, when the error structure is exchangeable, that the mean-square error is less than that of the state of the art estimator with probability approaching one. In simulation studies, our proposed estimators achieve better coverage than currently available methods, even when the underlying generative model violates the assumptions. Lastly, we demonstrate that our covariance matrix estimator can be used to weight coefficient estimates in GEE in an analysis of data on international trade flows.

The proposed estimator is not appropriate when the dependence among relations, i.e. the covariance structure, is extremely heterogeneous. This can happen in two ways: (i) heterogeneity in the covariance structure is endogenous with an observed covariate that is not included in the regression (a variant of omitted variable bias) and (ii) there is heterogeneity in the error variances even after accounting for all observables in the "true" generating model. In both cases, we could consider an extension of our approach that would further compromise between the unstructured covariance structure of the DC estimator and our exchangeable covariance structure. One could, for example, use a two-stage approach that first estimates actor clusters using the residuals, then assumes exchangeability within but not across clusters. While these methods are in development, practitioners should consider using the exchangeable working covariance in GEE procedures and possibly "correcting" the population covariance using the DC estimator $\widehat{\Omega}_{DC}$. When the standard errors resulting from the population covariance estimators $\widehat{\Omega}_{DC}$ and $\widehat{\Omega}_{E}$ differ, a conservative approach of selecting the maximum standard error estimate may be appropriate, as recommended by Angrist and Pischke (2008).

Many relational data sets contain binary or count measures, such as the presence or absence of relations between actors or number of interactions. Estimating equation and GEE procedures are often used with non-continuous data whereby the g equations in (3.2) involve a link function connecting the observed relation to the covariates, mirroring generalized linear regression procedures. While it is possible to impose an exchangeability assumption on the covariance matrix of the observations with non-continuous data, it is unclear how the assumption translates to an assumption about the data generating process. For example, consider the logit and probit regression models for binary data. Both models possess latent variable constructions which involve thresholding a latent continuous outcome composed of the linear regression

function plus a random error. Exchangeability of these errors does not imply the relations themselves are exchangeable (conditional on the covariates) as the covariates impact the dependence structure among the relational measures. For this reason, the methods proposed here cannot be trivially applied to non-continuous relational data. We address regression of binary network data in the following chapter.

Chapter 4

Regression of binary network data with exchangeable latent errors

4.1 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 and Horvath, 2005), and interactions among students over the course of a semester (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 alignment of political ideologies of pairs of books, on the 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 $\{y_{ij} \in \{0,1\} : i,j \in \{1,...,n\}, i < j\}$, 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 $\{y_{ii} : i \in \{1,...,n\}\}$ 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

$$\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij} > 0\right), \tag{4.1}$$

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 the error network $\{\epsilon_{ij}\}_{ij}$ are independent, estimation of the probit regression model in (4.1) 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 $\{\mathbf{x}_{ij}^T\boldsymbol{\beta}\}_{ij}$ 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 and 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. Ignoring the excess dependence in $\{\epsilon_{ij}\}_{ij}$ will often result in poor estimation of $\boldsymbol{\beta}$ and poor out-of-sample predictive performance; we observe this in the simulation studies and analysis of the political books network (see Sections 4.7 and 4.8, respectively). Thus, estimators of $\boldsymbol{\beta}$ and $\mathbb{P}(y_{ij} = 1)$ in (4.1) for the network $\{y_{ij}\}_{ij}$ would 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 and Clayton, 1993). In these models, a random effect, i.e. 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 and 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 and 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 4.2. 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., 2008a), 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 $\mathbb{P}(y_{ij} = 1)$ as a function of exogenous covariates and statistics of the network itself, 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 and 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., 2008b). 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 and Friel, 2011). Further, parameter estimates typically cannot be generalized to populations outside the observed network (Shalizi and 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 this chapter. In this approach, an unobserved normal random variable, z_{ij}, is proposed to underlie each data point, such that $y_{ij} = \mathbb{1}[z_{ij} > 0]$ for $\mathbf{z} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}(\boldsymbol{\theta}))$. In this formulation, excess dependence 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 and Sowden (1970) propose likelihood ratio hypothesis tests and Ochi and 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 and Liang, 1988), pseudo-likelihoods (Le Cessie and Van Houwelingen, 1994), and MCMC approximations to EM algorithms (Chib and Greenberg, 1998; Li and Schafer, 2008) are employed for estimation.

In this chapter, 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 4.3 for a detailed discussion of exchangeability)[2]. 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 a block coordinate descent to attain a numerical approximation to the maximum likelihood estimator. The estimation scheme we employ is similar to those used to estimate generalized linear mixed models in the literature (Littell et al., 2006; Gelman and Hill, 2006).

This chapter is organized as follows. As latent variable network models are strongly related to our work, we review them in detail in Section 4.2. We provide supporting theory for exchangeable random network models and their connections to latent variable network models in Section 4.3. In Section 4.4, we define the PX model and then the estimation thereof in Section 4.5. 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 4.7. We analyze a network of political books in Section 4.8, demonstrating the reduction in runtime when using the 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 4.9.

4.2 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

$$\begin{aligned} &\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mu_{ij} + f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j) + \xi_{ij} > 0\right), \\ &\mathbf{v}_i \overset{iid}{\sim} (\mathbf{0}, \boldsymbol{\Sigma}_v), \qquad \xi_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), \end{aligned} \tag{4.2}$$

[2]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).

where $\mathbf{v}_i \in \mathbb{R}^K$ and μ_{ij} is fixed. We set the total variance of the latent variable representation to be $1 = \sigma^2 + var[f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j)]$, 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 similar distributions through latent function $f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j)$. Many popular models for network data may be represented as in (4.2), such as the social relations model, the latent position model, and the latent eigenmodel.

4.2.1 Social relations model

The social relations model was first developed for continuous, directed network data (Warner et al., 1979; Wong, 1982; Snijders and 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

$$\begin{aligned} &\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mathbf{x}_{ij}^T \boldsymbol{\beta} + a_i + a_j + \xi_{ij} > 0\right), \\ &a_i \overset{iid}{\sim} (0, \sigma_a^2), \qquad \xi_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2). \end{aligned} \tag{4.3}$$

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 (4.3) also account for the excess correlation in network data; any two relations that share an actor, e.g. y_{ij} and y_{ik} are marginally correlated.

4.2.2 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:

$$\mathbb{P}(y_{ij}=1) = \mathbb{P}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + a_i + a_j - ||\mathbf{u}_i - \mathbf{u}_j||_2 + \xi_{ij} > 0\right), \tag{4.4}$$
$$a_i \overset{iid}{\sim} (0, \sigma_a^2), \qquad \mathbf{u}_i \overset{iid}{\sim} (0, \boldsymbol{\Sigma}_u), \qquad \xi_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2).$$

In the form of (4.2), 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) shows 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 and Faust, 1994).

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

$$\mathbb{P}(y_{ij}=1) = \mathbb{P}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + a_i + a_j + \mathbf{u}_i^T\Lambda\mathbf{u}_j + \xi_{ij} > 0\right), \tag{4.5}$$
$$a_i \overset{iid}{\sim} (0, \sigma_a^2), \qquad \mathbf{u}_i \overset{iid}{\sim} (0, \boldsymbol{\Sigma}_u), \qquad \xi_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2).$$

In the context of (4.2), 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 clustering, 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 and 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 chapter.

4.2.4 Drawbacks

The latent variable network models discussed in this section were developed based on the types of patterns often seen 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., 2008a), rather than rigorous tests, and it is well-known that selecting informative statistics for the goodness-of-fit checks is challenging. Finally, 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.

4.3 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 the actor labels, that is, if

$$\mathbb{P}\left(\{\epsilon_{ij}\}_{ij}\right) = \mathbb{P}\left(\{\epsilon_{\pi(i)\pi(j)}\}_{ij}\right), \tag{4.6}$$

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

All the latent variable network models discussed in Section 4.2 have latent error networks $\{\epsilon_{ij}\}_{ij}$ that are exchangeable, where we define $\epsilon_{ij} = f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j) + \xi_{ij}$ from (4.2), 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 4.2 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 and Szegedy (2006); Kallenberg (2006)):

$$\begin{aligned} &\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mu + h(u_i, u_j) + \xi_{ij} > 0\right), \\ &u_i \overset{iid}{\sim} \text{Uniform}(0,1), \qquad \xi_{ij} \overset{iid}{\sim} \text{N}(0, \sigma^2), \end{aligned} \tag{4.7}$$

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 (4.7) is equivalent to the graphon representation of exchangeable random network models, where the graphon is the canonical probabilistic object of exchangeable random network models and the ξ_{ij} are traditionally uniformly distributed (Lovász and 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 (4.7) shows that every exchangeable random network model may be represented by a latent variable network model in the sense of Section 4.2.

4.3.1 Covariance matrices of exchangeable network models

The expression in (4.7) 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. Marrs et al. (2017) shows that directed exchangeable network models with continuous values all have covariance matrices of the same form with at most five unique nonzero terms. Similarly, the covariance matrix of *any* undirected exchangeable network model has the same form and contains at most two unique nonzero values. 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:

$$var[\epsilon_{ij}] = \sigma_\epsilon^2, \qquad cov[\epsilon_{ij}, \epsilon_{ik}] := \rho, \qquad cov[\epsilon_{ij}, \epsilon_{kl}] = 0, \tag{4.8}$$

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, i.e. by permuting the actor labels. The third equality

results from the fact that the only random elements in (4.7) are the actor random variables u_i, u_j, and the random error ξ_{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, $E[y_{ij}] = \Phi(\mu)$ and $var[y_{ij}] = \Phi(\mu)(1 - \Phi(\mu))$ for all relations ij, where $\Phi(a)$ is the standard normal cumulative distribution function evaluated at a That said, in the linear regression case, 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 network $\{\epsilon_{ij}\}_{ij}$ remains exchangeable in the 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 (4.8).

4.4 The Probit Exchangeable (PX) model

In this section, we propose the probit exchangeable network regression model, which we abbreviate 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 Section 4.3. The PX model is

$$\begin{aligned} &\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij} > 0\right), \\ &\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}), \end{aligned} \tag{4.9}$$

where we note that the variance of ϵ_{ij} is not identifiable, and thus we choose $var[\epsilon_{ij}] = 1$ without loss of generality. We focus on normally-distributed unobserved errors $\boldsymbol{\epsilon}$ in this chapter, however, other common distributions, such as the logistic distribution, could be used. We note that the normal distribution assumption implies that (4.9) is a probit regression model with correlation among the observations.

As discussed in Section 4.3, under the exchangeability assumption, the covariance matrix of the latent error network $var[\boldsymbol{\epsilon}] = \boldsymbol{\Omega}$ has at most two unique nonzero parameters. Taking $var[\epsilon_{ij}] = 1$, the covariance matrix of $\boldsymbol{\epsilon}$ has a single parameter $\rho = cov[\epsilon_{ij}, \epsilon_{ik}]$. We may thus write

$$\boldsymbol{\Omega}(\rho) = \mathbf{S}_1 + \rho \mathcal{S}_2, \tag{4.10}$$

where we define the binary matrices $\{\mathbf{S}_i\}_{i=1}^3$ indicating unique entries in $\boldsymbol{\Omega}$. The matrix $\mathbf{S}_1$ is a diagonal matrix indicating the locations of the variance in $\boldsymbol{\Omega}$, and $\mathbf{S}_2$ and $\mathbf{S}_3$ indicate the locations in $\boldsymbol{\Omega}$ corresponding to the covariances $cov[\epsilon_{ij}, \epsilon_{ik}]$, and $cov[\epsilon_{ij}, \epsilon_{kl}]$, 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 4.2 and 4.3. Similar to (4.7), the PX model may be seen representing the covariance structure of the latent variables $\{f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j)+\xi_{ij}\}_{ij}$ with $\{\epsilon_{ij}\}_{ij}$, the unobservable error network of the PX model in (4.9). As both networks $\{f_{\boldsymbol{\theta}}(\mathbf{v}_i, \mathbf{v}_j) + \xi_{ij}\}_{ij}$ and $\{\epsilon_{ij}\}_{ij}$ are exchangeable, they have covariance matrices of the same form (see discussion in Section 4.3). 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 12. *Suppose that the random effects* $\{a_i\}_{i=1}^n$ *for the social relations model in* (4.3) *are normally distributed. Then, there exists* $\rho \geq 0$ *such that* $\{y_{ij}\}_{ij}$ *in the PX model in* (4.9) *is equal in distribution to* $\{y_{ij}\}_{ij}$ *as specified by the social relations model in* (4.3).

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 $var[\{a_i+ a_j+\xi_{ij}\}_{ij}] = var[\{\epsilon_{ij}\}_{ij}]$. 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 $var[\epsilon_{ij}] = 1$), we may take $\rho = \sigma_a^2/2$ for the PX model, which establishes equality in the model distributions. □

Proposition 12 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 (4.3) 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 4.5 employs the same strategies as those

commonly used to estimate generalized linear mixed models (Littell et al., 2006; Gelman and 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 $\Phi(\mathbf{x}_{ij}^T\boldsymbol{\beta})$. 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:

$$cov[y_{ij}, y_{ik}] = E\left[y_{ij}y_{ik}\right] - E\left[y_{ij}\right]E\left[y_{ik}\right] = \int_{-\mathbf{x}_{ij}^T\boldsymbol{\beta}}^{\infty}\int_{-\mathbf{x}_{ik}^T\boldsymbol{\beta}}^{\infty} dF_\rho - \Phi(\mathbf{x}_{ij}^T\boldsymbol{\beta})\Phi(\mathbf{x}_{ik}^T\boldsymbol{\beta}). \tag{4.11}$$

Here, dF_ρ is the bivariate standard normal distribution with correlation ρ. Since the covariance $cov[y_{ij}, y_{ik}]$ depends on the latent means $\mathbf{x}_{ij}^T\boldsymbol{\beta}$ and $\mathbf{x}_{ik}^T\boldsymbol{\beta}$, $cov[y_{ij}, y_{ik}]$ is only equal to $cov[y_{ab}, y_{ac}]$ 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.

4.5 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 a block coordinate descent with steps based on expectation-maximization (EM) algorithms (Dempster et al., 1977). Generally, the MLE of the parameters in a correlated probit regression model may be written only for particular covariance structures of the unobserved errors. Although the covariance matrix for the PX model is highly structured, as in (4.10), a closed-form expression for the MLE does not appear available.

The proposed estimation algorithm consists of alternating maximization of the data log-likelihood, $\ell_{\mathbf{y}}$, with respect to $\boldsymbol{\beta}$ and ρ, respectively. For each $\boldsymbol{\beta}$ and ρ maximization, we use an EM algorithm to maximize $\ell_{\mathbf{y}}$ with respect to each parameter by writing the likelihood as a function of latent relation ij values

$z_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}$, such that $y_{ij} = \mathbb{1}[z_{ij} > 0]$. Since the proposed algorithm is an embedding of EM algorithms within a block coordinate descent, we term it the BC-EM algorithm. To improve algorithm efficiency, we initialize $\boldsymbol{\beta}$ at the ordinary probit regression estimator, assuming independence, and initialize ρ with a mixture estimator based on possible values of ρ such that $\boldsymbol{\Omega}$ is positive definite, as detailed in Appendix C.1.1. The complete BC-EM algorithm is presented in Algorithm 2. In what follows, we detail the BC-EM algorithm, beginning with maximization with respect to $\boldsymbol{\beta}$, and then proceeding to maximization with respect to ρ.

Algorithm 2 BC-EM 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 Appendix C.1.1. Set positive convergence thresholds τ, τ_β, τ_ρ, and set iteration $\nu = 0$.

1. $\boldsymbol{\beta}$ **block:**
Set $s = 0$ and $\widehat{\boldsymbol{\beta}}^{(\nu,\,s)} = \widehat{\boldsymbol{\beta}}^{(\nu)}$.

   1.1. **Expectation:** Given $\widehat{\rho}^{(\nu)}$ and $\widehat{\boldsymbol{\beta}}^{(\nu,\,s)}$, compute $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu,s)}]$ using the procedure described in Appendix C.1.2.

   1.2. **Maximization:** Compute the updated estimate

$$\widehat{\boldsymbol{\beta}}^{(\nu,\,s+1)} = \widehat{\boldsymbol{\beta}}^{(\nu,\,s)} + (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu,s)}].$$

   1.3. If $||\widehat{\boldsymbol{\beta}}^{(\nu,\,s+1)} - \widehat{\boldsymbol{\beta}}^{(\nu,\,s)}||_1 < \tau_\beta$, then set $\widehat{\boldsymbol{\beta}}^{(\nu+1)} = \widehat{\boldsymbol{\beta}}^{(\nu,\,s+1)}$. Otherwise, increment s by 1 and return to step 1.1.

2. ρ **block:**
Set $s = 0$ and $\widehat{\rho}^{(\nu,\,s)} = \widehat{\rho}^{(\nu)}$.

   2.1. **Subsample:** Randomly sample $10n(n-1)$ pairs of relations that share an actor, and define this sample $\mathcal{A}^{(s)} \subset \Theta_2$.

   2.2. **Expectation:** Given $\widehat{\rho}^{(\nu,\,s)}$ and $\widehat{\boldsymbol{\beta}}^{(\nu+1)}$, approximate $\{\gamma_i\}_{i=1}^3$ as described in Appendix C.1.3.

   2.3. **Maximization:** Given $\widehat{\rho}^{(\nu,\,s)}$, $\widehat{\boldsymbol{\beta}}^{(\nu+1)}$, and $\{\gamma_i\}_{i=1}^3$, compute $\widehat{\rho}^{(\nu,\,s+1)}$ by alternating (4.15) and (4.16) until ρ changes by less than τ_ρ. The final value of ρ is $\widehat{\rho}^{(\nu,\,s+1)}$.

   2.4. If $|\widehat{\rho}^{(\nu,\,s+1)} - \widehat{\rho}^{(\nu,\,s)}| < \tau_\rho$, then set $\widehat{\rho}^{(\nu+1)} = \widehat{\rho}^{(\nu,\,s+1)}$. Otherwise, increment s by 1 and return to step 2.1.

3. If $\max\{||\widehat{\boldsymbol{\beta}}^{(\nu+1)} - \widehat{\boldsymbol{\beta}}^{(\nu)}||_1, |\widehat{\rho}^{(\nu+1)} - \widehat{\rho}^{(\nu)}|\} > \tau$, then increment ν by 1 and return to Step 1. Otherwise, end.

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

To maximize the data log-likelihood, $\ell_{\mathbf{y}}$, with respect to $\boldsymbol{\beta}$, we utilize an EM algorithm. We begin by discussing expectation of the likelihood $\ell_{\mathbf{z}}$ with respect to $\mathbf{z}$ (E-step), follow by discussing maximization of the resulting expression as a function of $\boldsymbol{\beta}$ (M-step), and then discuss approximations to make the estimation computationally feasible.

E-step:

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

$$\begin{aligned} E[\ell_{\mathbf{z}} \,|\, \mathbf{y}, \rho = \widehat{\rho}^{(\nu)}, \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(\nu)}] = \\ -\frac{1}{2}\log 2\pi|\boldsymbol{\Omega}| - \frac{1}{2}E\left[(\mathbf{z}-\mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}) \,|\, \mathbf{y}, \rho = \widehat{\rho}^{(\nu)}, \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(\nu)}\right], \end{aligned} \tag{4.12}$$

where $\widehat{\rho}^{(\nu)}$ and $\widehat{\boldsymbol{\beta}}^{(\nu)}$ are the estimators of ρ and $\boldsymbol{\beta}$ at iteration ν. In discussing the M-step for β, we will show that that the β update depends on the data only through the expectation $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}]$.

M-step:

Setting the derivative of (4.12) with respect to $\boldsymbol{\beta}$ equal to zero, the maximization step for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}^{(\nu+1)} = \widehat{\boldsymbol{\beta}}^{(\nu)} + \left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}], \tag{4.13}$$

where we use the identity $\boldsymbol{\epsilon} = \mathbf{z} - \mathbf{X}\boldsymbol{\beta}$. Noting that (4.13) only requires the expectation $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}]$, an EM algorithm to maximize $\ell_{\mathbf{y}}$ with respect to $\boldsymbol{\beta}$ consists of alternating computation of $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}]$ in the E-step with computing the next $\boldsymbol{\beta}$ estimate given by (4.13) in the M-step.

Approximations:

The computation of $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}]$ in (4.13) is nontrivial, as it is a $\binom{n}{2}$-dimensional truncated multivariate normal integral. We exploit the structure of $\boldsymbol{\Omega}$ to compute $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}]$ using the law of total expectation. A Newton-Raphson algorithm, along with an approximate matrix inverse, are employed to

compute an approximation of of $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu)}]$. Details of the implementation of the EM algorithm for $\boldsymbol{\beta}$ are given in Appendix C.1.2.

4.5.2 Maximization with respect to ρ

To maximize $\ell_{\mathbf{y}}$ with respect to ρ, we again utilize an EM algorithm. We begin by discussing expectation of the likelihood $\ell_{\mathbf{z}}$ with respect to $\mathbf{z}$ (E-step), follow by discussing maximization of the resulting expression as a function of ρ (M-step), and then discuss approximations to make the estimation computationally feasible.

E-step:

The expectation step is the same as in (4.12), although evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(\nu+1)}$. In discussing the M-step for ρ, we will show that that the ρ update depends on the data through the expectations $E[\boldsymbol{\epsilon}^T \mathcal{S}_i \boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu+1)}]$ for $i \in \{1, 2, 3\}$.

M-step:

To derive the maximization step for ρ, we use the method of Lagrange multipliers, since differentiating (4.12) 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 $\boldsymbol{\Omega}$,

$$var[\epsilon_{ij}] = \phi_1, \qquad cov[\epsilon_{ij}, \epsilon_{ik}] = \phi_2 = \rho, \qquad cov[\epsilon_{ij}, \epsilon_{kl}] = \phi_3, \tag{4.14}$$

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 $\boldsymbol{\Omega}^{-1} = \sum_{i=1}^3 p_i \mathbf{S}_i$, 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

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

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

$$\widehat{\rho} = \gamma_2 - \frac{1}{|\Theta_2|}\begin{bmatrix} \frac{\partial \phi_1}{\partial p_2} & \frac{\partial \phi_3}{\partial p_2} \end{bmatrix}^T \begin{bmatrix} \lambda_1 \\ \lambda_3 \end{bmatrix} \tag{4.15}$$

$$\begin{bmatrix} \widehat{\lambda}_1 \\ \widehat{\lambda}_3 \end{bmatrix} = \begin{bmatrix} \frac{\partial \phi_1}{\partial p_1} & \frac{\partial \phi_3}{\partial p_1} \\ \frac{\partial \phi_1}{\partial p_3} & \frac{\partial \phi_3}{\partial p_3} \end{bmatrix}^{-1} \begin{bmatrix} |\Theta_1| & 0 \\ 0 & |\Theta_3| \end{bmatrix} \begin{bmatrix} \gamma_1 - 1 \\ \gamma_3 \end{bmatrix}, \tag{4.16}$$

where $\gamma_i := E[\boldsymbol{\epsilon}^T \mathcal{S}_i \boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu+1)}]/|\Theta_i|$ and 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 (4.15) and (4.16), the partial derivatives $\left\{\frac{\partial \phi_i}{\partial p_j}\right\}$ 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 C.2 for details.

Alternation of the estimators for ρ and $\{\lambda_1, \lambda_3\}$ in (4.15) and (4.16) 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 M-step of the EM algorithm for ρ. The M-step depends on the data through $\{\gamma_i\}_{i=1}^3$, the computation of which constitutes the E-step of the EM algorithm. We describe the approximation to the E-step, that is, computation of approximate values of $\{\gamma_i\}_{i=1}^3$ below.

Approximations:

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 three approximations to $\{\gamma_i\}_{i=1}^3$ to reduce runtime of the BC-EM 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, instead of computing $E[\epsilon_{jk}\epsilon_{lm} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu+1)}]$, we compute $E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu+1)}]$. Second, we find empirically that $\gamma_2 = E[\boldsymbol{\epsilon}^T \mathbf{S}_2 \boldsymbol{\epsilon} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu+1)}]/|\Theta_2|$ is approximately linear in ρ, and 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 C.4). Third, it can be shown that $\gamma_2 = \rho + o_p(n^{-1/2})$, yet, γ_2 is an average of $O(n^3)$ terms. We take a random subset of $O(n^2)$ of these terms at each iteration to reduce the computational burden (note that γ_1 and γ_3 may be computed with $O(n^2)$ operations given

the pairwise approximation that $E[\epsilon_{jk}\epsilon_{lm} \,|\, \mathbf{y}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu+1)}] \approx E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}, \widehat{\rho}^{(\nu)}, \widehat{\boldsymbol{\beta}}^{(\nu+1)}]$). Additional details of the approximations to $\{\gamma_i\}_{i=1}^3$ are given in Appendix C.1.3.

4.6 Prediction

In this section, we describe how to use the PX model to make predictions for an unobserved network relation. 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 for pair jk. This 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

$$\begin{aligned} \mathbb{P}(y_{jk} = 1 \,|\, \mathbf{y}_{-jk}) &= E\left[E\left[\mathbb{1}[\epsilon_{jk} > -\mathbf{x}_{jk}^T\boldsymbol{\beta}] \,|\, \boldsymbol{\epsilon}_{-jk}\right] \,|\, \mathbf{y}_{-jk}\right], \\ &\approx \Phi\left(\frac{E[\epsilon_{jk} \,|\, \mathbf{y}] + \mathbf{x}_{jk}^T\boldsymbol{\beta}}{\sigma_n}\right). \end{aligned} \tag{4.17}$$

The approximation in (4.17) is based on the fact that $[\epsilon_{jk} \,|\, \boldsymbol{\epsilon}_{-jk}]$ is normally distributed:

$$\begin{aligned} \epsilon_{jk} \,|\, \boldsymbol{\epsilon}_{-jk} &\sim \mathrm{N}(\mu_{jk}, \sigma_n^2), \\ \mu_{jk} = -\sigma_n^2 \mathbf{1}_{jk}^T \left(p_2\mathcal{S}_2 + p_3\mathcal{S}_3\right) \widetilde{\boldsymbol{\epsilon}}_{-jk}, &\qquad \sigma_n^2 = \frac{1}{p_1}, \end{aligned} \tag{4.18}$$

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 $p_2\mathcal{S}_2 + p_3\mathcal{S}_3$ consists of all zeros so that μ_{jk} is free of ϵ_{jk}. Then, the inner expectation in (4.17) is

$$E\left[\mathbb{1}[\epsilon_{jk} > -\mathbf{x}_{jk}^T\boldsymbol{\beta}] \,|\, \boldsymbol{\epsilon}_{-jk}\right] = \Phi\left(\frac{\mu_{jk} + \mathbf{x}_{jk}^T\boldsymbol{\beta}}{\sigma_n}\right). \tag{4.19}$$

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

Computing $E[\epsilon_{jk} \,|\, \mathbf{y}_{-jk}]$ is extremely difficult, however computing $E[\epsilon_{jk} \,|\, \mathbf{y}]$ 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:

$$E\left[\epsilon_{jk} \,|\, \mathbf{y}_{-jk}\right] \approx E\left[\epsilon_{jk} \,|\, \mathbf{y}_{-jk}, y_{jk} = \mathbf{y}^*\right] = E\left[\epsilon_{jk} \,|\, \mathbf{y}\right], \tag{4.20}$$

where $\mathbf{y}^*$ is the mode of $\mathbf{y}_{-jk}$. The error due to this approximation is small and shrinks as n grows. Substituting (4.20) for μ_{jk} in (4.19) gives the final expression in (4.17).

4.7 Simulation studies

In this section, we describe two simulation studies. The first verifies that the performance of the BC-EM estimator in Algorithm 2 is not substantially worse than the MLE. The second simulation study verifies consistency of the BC-EM 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.

4.7.1 Evaluation of approximations in Algorithm 2

To evaluate the efficacy of the approximations described in the estimation procedure in Algorithm 2, we conducted a simulation study comparing BC-EM, and MLE estimators of β. We estimated β in the standard probit model assuming independence between observations (which we abbreviate "std. probit") as a baseline. We fixed $\mathbf{X}$ for the study to contain a single covariate (column) of Bernoulli$(p = 0.25)$ random variables. We simulated 100 networks from the PX model in (4.9) using this $\mathbf{X}$, for each combination of parameters $\beta = 1$ and $\rho \in \{0.1, 0.2, 0.3, 0.4\}$. For each realization, we estimated β in the PX model using BC-EM in Algorithm 2. To estimate β in the standard probit model, we used the the function `glm` in `R`. We numerically optimized the data log-likelihood using the `optim` function in `R` to compute the MLE. Since this numerical optimization is computationally onerous, we simulated networks of size $n = 10$ for this study.

In the left panel of Figure 4.1, we evaluate the performance of the BC-EM estimator by comparing the mean square error (MSE) between its coefficient estimate, $\widehat{\beta}_{BC-EM}$, and MLE obtained by the optimization procedure $\widehat{\beta}_{MLE}$. As a baseline, we compute the MSE between $\widehat{\beta}_{MLE}$ and the true value β. If the approximations in the BC-EM algorithm are small, we expect the MSE between $\widehat{\beta}_{BC-EM}$ and $\widehat{\beta}_{MLE}$ to be much smaller than the MSE between $\widehat{\beta}_{MLE}$ and β. Generally, the MSE between $\widehat{\beta}_{BC-EM}$ and $\widehat{\beta}_{MLE}$ is

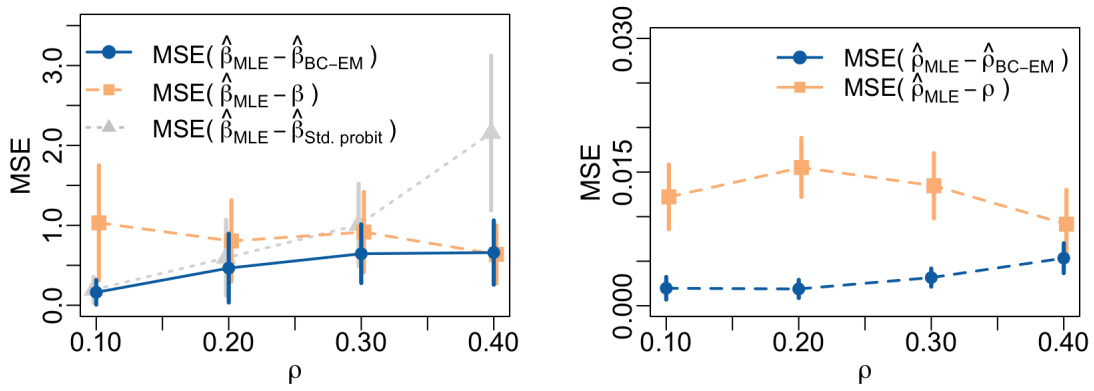


Figure 4.1: The left panel depicts performance in estimating β: MSE between the BC-EM estimator and the MLE (MSE($\widehat{\beta}_{MLE} - \widehat{\beta}_{BC-EM}$)), between the MLE and the truth (MSE($\widehat{\beta}_{MLE} - \beta$)), and between the MLE and the standard probit estimator (MSE($\widehat{\beta}_{MLE} - \widehat{\beta}_{Std.probit}$)). The right panel depicts performance in estimating ρ: MSE between the MLE and the BC-EM estimator (MSE($\widehat{\rho}_{MLE} - \widehat{\rho}_{BC-EM}$)) and between the MLE and the truth (MSE($\widehat{\rho}_{MLE} - \rho$). The MSEs are plotted as a function of the true values of ρ, and solid vertical lines denote Monte Carlo error bars.

smaller than the MSE between $\widehat{\beta}_{MLE}$ and β, however, the discrepancy between the two MSEs decreases as the true ρ grows. As a reference, the MSE between $\widehat{\beta}_{Std.\,probit}$ and $\widehat{\beta}_{MLE}$ is also shown in the left panel of Figure 4.1. For true $\rho > 0.2$, the BC-EM estimator is substantially closer to $\widehat{\beta}_{MLE}$ than the standard probit estimator is to $\widehat{\beta}_{MLE}$. Raw MSE values between the estimators and the truth, shown in Appendix C.3.1, confirm that the BC-EM algorithm does perform better than standard probit in MSE with respect to estimation of β. The results of this simulation study suggest that the BC-EM algorithm improves estimation of β over the standard probit estimator for $\rho > 0$, and that the BC-EM estimator is relatively close to the MLE, signifying the approximations in the BC-EM algorithm are not unreasonable. It is worth noting that the approximations used in the BC-EM algorithm are best for large n so we would expect better and better results as n increases.

In the right panel Figure 4.1, we see that the the BC-EM estimator of ρ is substantially closer to the MLE, $\widehat{\rho}_{MLE}$, than the MLE is close to the true value of ρ. This suggests that the approximation error in estimating ρ in the BC-EM algorithm is small. Further, the raw MSE values shown in Appendix C.3.1 show that $\widehat{\rho}_{BC-EM}$ is actually closer to the true ρ than is $\widehat{\rho}_{MLE}$. The difference in the MSE between $\widehat{\rho}_{MLE}$ and $\widehat{\rho}_{BC-EM}$ and the MSE between $\widehat{\rho}_{MLE}$ and ρ again decreases as the true value of ρ grows. This trend and the similar trend in the left panel of Figure 4.1 suggest that the approximations in the BC-EM algorithm degrade as the true value of ρ grows, at least for $n = 10$.

4.7.2 Performance in estimation of $\boldsymbol{\beta}$

To evaluate the performance of the PX estimator in estimating linear coefficients $\boldsymbol{\beta}$, we compared estimates of $\boldsymbol{\beta}$ by the BC-EM 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`.

To conduct the desired simulation study, we generated data with mean consisting of three covariates and an intercept:

$$y_{ij} = \mathbb{1}\Big[\beta_0 + \beta_1 \mathbb{1}[x_{1i} \in C]\mathbb{1}[x_{1j} \in C] + \beta_2 |x_{2i} - x_{2j}| + \beta_3 x_{3ij} + \epsilon_{ij} > 0\Big]. \tag{4.21}$$

In the model in (4.21), 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. 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 $\rho = 1/4$ (the remaining half was due to noise ξ_{ij}). When generating from the latent eigenmodel in (4.5), one third the variance in ϵ_{ij} was due to each term $a_i + a_j$, $\mathbf{u}_i^T \Lambda \mathbf{u}_j$, and ξ_{ij}, respectively. For additional details of the simulation study procedures, see Appendix C.3.2.

In Figure 4.2, we see that the BC-EM estimator for the PX model has a downward trend in MSE with n, and a reducing spread of the MSE with n, for both the PX and latent eigenmodel generating models. These facts suggest that the PX estimator is consistent for $\boldsymbol{\beta}$ for both the PX and latent eigenmodel generating models. Further, the BC-EM estimator has the lowest median MSE of any of the estimators for all entries in $\boldsymbol{\beta}$, where the MSE 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 Appendix C.3.2. 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 MSEs of the `amen` estimator of the social relations and latent eigenmodels are a result of bias; see Appendix C.3.2 for bias-variance decomposition of the MSEs.

4.8 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[3]. 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 labelled as conservative, liberal, or neutral based on each book's description (Figure 4.3). Work by Dr. Krebs on a similar network was described 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., 2017b).

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 4.3, we observe that neutral books appear to have fewer ties to other books than books that are labelled conservative or liberal. Thus, we add a nodal effect indicating whether either book in a relation is labelled neutral. The regression model specified is

$$\mathbb{P}(y_{ij} = 1) = \mathbb{P}\left(\beta_0 + \beta_1 \mathbb{1}[c(i) = c(j)] + \beta_2 \mathbb{1}\left[\{c(i) = \text{neutral}\} \cup \{c(j) = \text{neutral}\}\right] + \boldsymbol{\epsilon}_{ij} > 0\right), \quad (4.22)$$

$$\boldsymbol{\epsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}),$$

[3]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

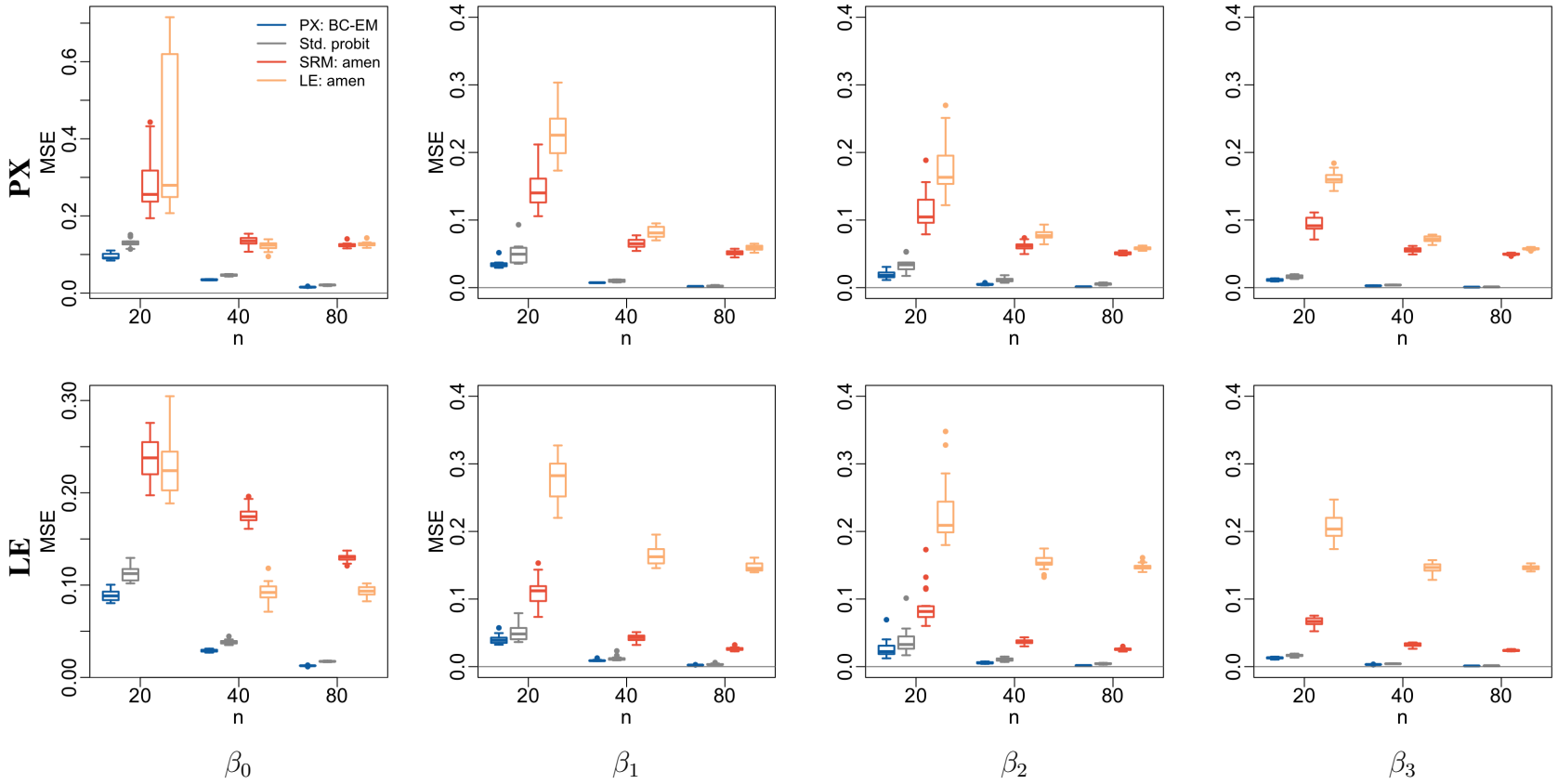


Figure 4.2: MSE of estimators of β_1 for a given $\mathbf{X}$ when generating from the PX model (top row) and the latent eigenmodel (LE; bottom row). Variability captured by the boxplots reflects variation in MSE with $\mathbf{X}$. Note that the intercept, β_0, has MSEs on different scales than the remaining coefficients.

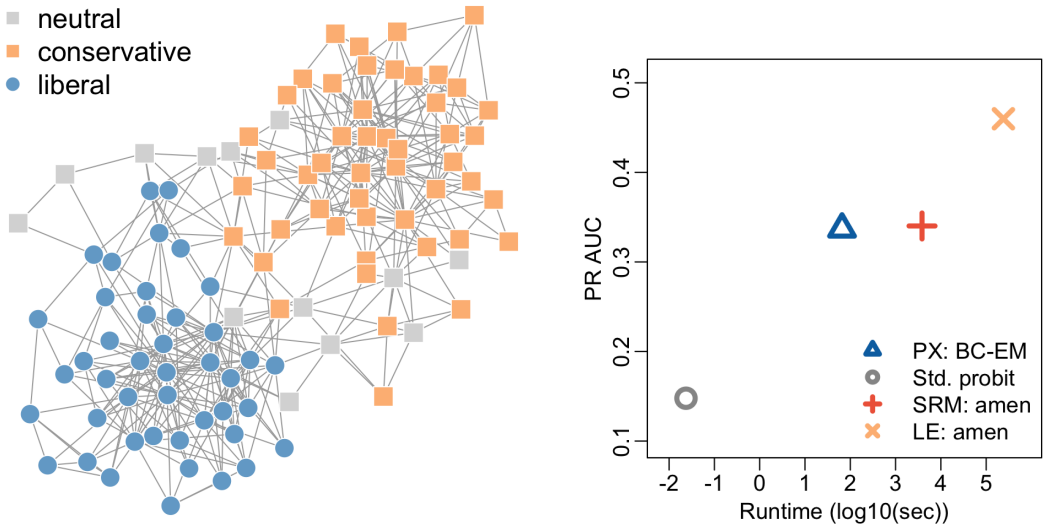


Figure 4.3: Krebs' political books network (left) and out-of-sample performance in 10-fold cross validation, as measured by area under the precision-recall curve (PRAUC, right), plotted against mean runtime in the cross validation. The estimators are standard probit assuming independnet observations (Std. probit), the PX model as estimated by the BC-EM algorithm (PX), the social relations model estimator (SRM), and the latent eigenmodel estimator (LE).

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 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 BC-EM 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` 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 Appendix C.1.4). Repeating this operation for each fold gave a 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 4.6. 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 C.4). In Figure 4.3, the proposed BC-EM 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 50 times faster (about a minute compared with an hour). The latent eigenmodel produces an improvement in PRAUC over the proposed BC-EM algorithm and the social relations model, however, at the expense of significant increase in average runtime, that of about 3,600 times slower than BC-EM 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 a set of samples from the posterior distribution (after burn-in) that has an effective sample size equal to roughly 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, may be orders of magnitude faster the runtime than existing approaches.

To estimate the complete data set under the mean model in (4.22), we used the BC-EM 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 4.1 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.

Returning briefly to Table 4.1, the runtimes highlight that BC-EM reduces computational burden by order(s) of magnitude over existing approaches.

Table 4.1: Results of fitting the Krebs political books data using the BC-EM 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: BC-EM	-1.61	0.93	0.97	54	–
SRM: amen	-2.70	1.55	0.98	7984	195
LE: amen	-3.90	2.06	1.63	62565	26

4.9 Discussion

In this chapter we present the PX model, a probit regression model for binary, undirected 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 chapter 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 network data estimators is non-trivial, indeed, entire papers are dedicated to this task (see, for example, Aronow et al. (2015); Marrs et al. (2017)). Thus, we leave the development of uncertainty estimators for the PX model to future work.

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, $\mathbb{P}(y_{ik} = 1)$, is increased. 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 BC-EM estimator. It is 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 $\mathbb{P}(y_{ik} = 1 \,|\, y_{ij} = y_{jk} = 1)$ 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 in 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 networks literature (Wasserman and Faust, 1994; Watts and Strogatz, 1998).

We focus on the probit model in this chapter. 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

$$cov[y_{ij}, y_{ik}] \;\leq\; \lim_{\rho \to 1/2} \int_{-\mu}^{\infty} \int_{-\mu}^{\infty} dF_\rho - \Phi(\mu)^2, \tag{4.23}$$

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 $cov[y_{ij}, y_{ik}]$. 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.

Chapter 5

Conclusion

In this dissertation, we have presented methods for accounting for excess dependence, beyond that represented by the covariates, in regression of network data. In the analysis of a bipartite network (Chapter 2), we explicitly modeled and inferred the excess network dependence. In contrast, in Chapter 3, we presented a parsimonious estimator for making inference on regression coefficient estimators when the response forms a continuous valued network, avoiding the need to explicitly model the form of the excess dependence. We presented similar methods for accounting for excess network dependence in regression of binary network data in Chapter 4. At the end of each of Chapters 2-4, we discussed potential extensions of the work therein. We now discuss additional possible research directions for modeling of bipartite network data and estimating regressions of continuous and binary network data in Sections 5.1 through 5.3. We conclude with a perspective on the broader impacts of this dissertation in Section 5.4.

5.1 Modeling of bipartite network data

A shortcoming of the Bipartite Longitudinal Influence Network (BLIN) model in Chapter 2 is that it represents a dependence structure that is strictly for discrete, equally-spaced event data. However, many phenomena – including international relations – occur in time regimes that are continuous and irregular. For example, in an analysis of a longitudinal bipartite network of environmental treaties, Campbell et al. (2019) observe ratifications of environmental treaties at the resolution of days, yet, the authors analyze the data at the resolution of years. Recall that, in the BLIN model, the influence of one ratifictation on the potential of future ratifications to occur is constant for a set time interval, such as three days. Thus, the BLIN model is ill-suited to the higher resolution at the level of days, since it is unlikely that these ratifications occur on consecutive days, and it is likely that the dependence in the data is non-constant in time. Instead, it may be best to estimate a level of influence of one ratification on the potential for another to occur that is a function of the time t between them, such that ratifications that are closer together in time are more highly correlated than those that are farther apart in time. One possible function to control the magnitude of influence is one which decays exponentially over time, for example, $f(t) = exp(-\kappa t)$, where κ is a single additional

parameter to be estimated that controls the rate of decay of all influences (we note that $f(t) = 1$ in the BLIN model proposed in Chapter 2). An approach that estimates κ as well as $\mathbf{A}$ and $\mathbf{B}$ matrices should verify that estimating a $\kappa \neq 0$ provides an improvement in model performance over $\kappa = 0$, that is, that there is evidence in the data that such a decay model is warranted.

The BLIN model in Chapter 2 may be written as a vector autoregressive model with coefficient matrix $\boldsymbol{\Theta}$. As we discuss in Chapter 2, the BLIN model may be seen as one way to impose a particular structure on $\boldsymbol{\Theta}$ such that $\boldsymbol{\Theta}$ is estimable. Further, we employ shrinkage in the form of lasso regression to estimate nonzero elements in $\boldsymbol{\Theta}$. When the true $\boldsymbol{\Theta}$ is outside the space of coefficient matrices spanned by those of the BLIN model, this regression will fail (even under lasso regularization). Instead, it may be desirable to shrink the fully populated $\boldsymbol{\Theta}$ matrix while still encouraging the BLIN structure. Griffin and Hoff (2019) develop shrinkage methods for regression coefficients in generalized linear models that simultaneously encourage sparsity and structure, for example, when regression coefficients are vectorized matrices. It may be possible to build off this approach to shrink $\boldsymbol{\Theta}$ while simultaneously encouraging the structure of the BLIN model.

5.2 Testing assumptions of regression of network data

The methods in both Chapters 3 and 4 rely on assumptions of joint exchangeability. This assumption is common in the network literature and is natural whenever the covariates contain all the heterogeneity in the probability distribution of the network. However, it is reasonable to suspect that in some cases, there may be omitted covariates that render the residual structure heterogeneous. Thus, it is desirable to test whether the assumption of joint exchangeability is appropriate for a particular data set.

One form of heterogeneity in the residual structure is heteroskedasticity. Indeed, homogeneity of variances of the errors is a requirement for the joint exchangeability assumed in Chapters 3 and 4. Thus, existing tests for heteroskedasticity may be sufficient to reject the null hypothesis of joint exchangeability of the errors (Breusch and Pagan, 1979; Jarque and Bera, 1980; White, 1980). However, we must consider that the existing tests are developed for vector regression problems, and thus, may be underpowered for the network regression problems considered in this dissertation, as the networks have corresponding matrix representations. Thus, one direction for research is to develop tests for joint exchangeability that build upon existing tests. Following White (1980), a test statistic that compares estimators of $V[\widehat{\boldsymbol{\beta}}]$ that require joint exchangeability for consistency against those estimators of $V[\widehat{\boldsymbol{\beta}}]$ that are consistent under heteroskedasticity,

respectively, may be a statistic that could be used to test for joint exchangeability. It remains to investigate the consistency of such a test statistic and demonstrate its power under possible non-exchangeable alternative distributions for the errors.

5.3 Nonparametric models for network data

We have shown in Section 3.3 that all exchangeable network models have a covariance matrix of the same form. This powerful result suggests a representation theorem may exist for jointly exchangeable network models that is made up of a covariance matrix and a marginal distribution of the unobserved errors. Development of theory in support of such a representation would be a contribution. Further, if any jointly exchangeable random network model may be represented by a covariance matrix and a marginal distribution of the unobserved errors, then this representation provides a setting for new estimators of these models. It may be efficacious, for example, to estimate the covariance parameter ρ for the undirected binary data in Chapter 4 while simultaneously estimating a non-parametric distribution of the unobserved errors (and possibly also estimating the effects of exogenous covariates). This approach is likely similar to non-parametric estimators of jointly exchangeable network models, see Bickel and Chen (2009) for example, however the Bickel and Chen (2009) approach is not directly applicable to regression problems.

As an example, consider Chapter 4, where we select a normal marginal distribution for the unobserved (jointly exchangeable) errors in the PX model. This assumption of normality may be unrealistic and/or inconvenient; for example, see Section 4.9 where we discuss logistic regression as an alternative to the proposed probit regression. The distribution of the errors may be modeled by *any* marginal distribution. Thus, estimating ρ and allowing the data to inform the marginal distribution of the errors may improve estimation of $\boldsymbol{\beta}$ and prediction of $\{y_{ij}\}_{ij}$ beyond that of the BC-EM estimator of the PX model presented in Chapter 4.

5.4 Broader impacts

In this dissertation, we have presented methods for accounting for the excess dependence – beyond the dependence captured by covariates – that may be expected when regressing network data. However, network data are only one form of discrete data; the methods and approaches proposed in this dissertation may be

applicable to other types of data structures. Other discrete data types with similar dependence are time series and spatially resolved data. The methods in Chapter 4 may be useful in analyzing binary time series, a current area of research (Fokianos and Moysiadis, 2017; Gao et al., 2018). Analysis of spatially resolved data, in ecological data e.g. (Ver Hoef et al., 2018), often use autoregressive models to capture dependence; the regression techniques in Chapter 3 may be useful for modelling dependence in the field of spatial statistics as well. Finally, spatial-temporal regression models must represent dependence in both space and time, which can be difficult using available covariates, as in applications in ecology and epidemiology (de Espindolaa et al., 2011; Noth et al., 2011). The bipartite methods in Chapter 2 are well-suited to modelling discrete dependence in two dimensions.

Bibliography

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008) Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, **9**, 1981–2014.

Akaike, H. (1998) Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, 199–213. Springer.

Aker, J. C. (2010) Information from markets near and far: Mobile phones and agricultural markets in Niger. *American Economic Joural*, **2**, 46–59.

Aldous, D. J. (1981) Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, **11**, 581–598.

Aleskerov, F., Meshcheryakova, N., Rezyapova, A. and Shvydun, S. (2017) Network analysis of international migration. In *Models, Algorithms, and Technologies for Network Analysis* (eds. V. A. Kalyagin, A. I. Nikolaev, P. M. Pardalos and O. A. Prokopyev), 177–185. Cham: Springer International Publishing.

Almquist, Z. W. and Butts, C. T. (2013) Dynamic network logistic regression: A logistic choice analysis of inter-and intra-group blog citation dynamics in the 2004 us presidential election. *Political Analysis*, **21**, 430–448.

— (2014) Logistic network regression for scalable analysis of networks with joint edge/vertex dynamics. *Sociological Methodology*, **44**, 273–321.

Anderson, T. W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 327–351.

Angrist, J. D. and Pischke, J.-S. (2008) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Aronow, P. M., Samii, C. and Assenova, V. A. (2015) Cluster–robust variance estimation for dyadic data. *Political Analysis*, **23**, 564–577.

Ashford, J. and Sowden, R. (1970) Multi-variate probit analysis. *Biometrics*, 535–546.

Atkinson, K. E. (2008) *An Introduction to Numerical Analysis*. John Wiley & Sons.

Attanasio, O., Barr, A., Cardenas, J. C., Genicot, G. and Meghir, C. (2012) Risk pooling, risk preferences, and social networks. *American Economic Journal*, **4**, 134–167.

Banerjee, A., Chandrasekhar, A. G., Duflo, E. and Jackson, M. O. (2013) The diffusion of microfinance. *Science*, **341**, 363–373.

Banks, D. L. and Carley, K. M. (1996) Models for network evolution. *The Journal of Mathematical Sociology*, **21**, 173–196. URL: https://doi.org/10.1080/0022250X.1996.9990179.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. and Vicsek, T. (2002) Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, **311**, 590–614.

Bardham, P. (1984) *Land, Labor and Rural Poverty*. Oxford University Press.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Berry, F. S. and Berry, W. D. (1990) State lottery adoptions as policy innovations: An event history analysis. *American Political Science Review*, **84**, 395–415.

Betzel, R. F., Bertolero, M. A. and Bassett, D. S. (2018) Non-assortative community structure in resting and task-evoked functional brain networks. *bioRxiv*, 355016.

Bickel, P. J. and Chen, A. (2009) A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, **106**, 21068–21073.

Blumenstock, J. E., Fafchamps, M. and Eagle, N. (2011) Risk and reciprocity over the mobile phone network: Evidence from Rwanda. *Available at SSRN 1958042*.

Bolthausen, E. (1982) On the central limit theorem for stationary mixing random fields. *The Annals of Probability*, **10**, 1047–1050.

Borgs, C., Chayes, J. T., Cohn, H. and Zhao, Y. (2014) An Lp theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *arXiv preprint arXiv:1401.2906*.

Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J. and Ward, M. (2015) ICEWS Coded Event Data. URL: https://doi.org/10.7910/DVN/28075.

Boulet, R., Barros-Platiau, A. F. and Mazzega, P. (2016) 35 years of multilateral environmental agreements ratifications: A network analysis. *Artificial Intelligence and Law*, **24**, 133–148.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Breusch, T. S. and Pagan, A. R. (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 1287–1294.

Brockwell, P. J., Davis, R. A. and Fienberg, S. E. (1991) *Time Series: Theory and Methods*. Springer Science & Business Media.

Bunea, F., She, Y. and Wegkamp, M. H. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 1282–1309.

Caimo, A. and Friel, N. (2011) Bayesian inference for exponential random graph models. *Social Networks*, **33**, 41–55.

Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2011) Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, **29**.

Cameron, A. C. and Miller, D. L. (2014) Robust inference for dyadic data. *Working paper, University of California-Davis*.

Campbell, B. W., Marrs, F. W., Böhmelt, T., Fosdick, B. K. and Cranmer, S. J. (2019) Latent influence networks in global environmental politics. *PLOS ONE*, **14**, e0213284.

Carnegie, N. B., Krivitsky, P. N., Hunter, D. R. and Goodreau, S. M. (2015) An approximation method for improving dynamic network model fitting. *Journal of Computational and Graphical Statistics*, **24**, 502–519.

Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

Conley, T. G. (1999) GMM estimation with cross sectional dependence. *Journal of Econometrics*, **92**, 1–45.

Connolly, M. A. and Liang, K.-Y. (1988) Conditional logistic regression models for correlated binary data. *Biometrika*, **75**, 501–506.

Cramér, H. and Wold, H. (1936) Some theorems on distribution functions. *Journal of the London Mathematical Society*, **1**, 290–294.

Dahl, R. A. (1957) The concept of power. *Behavioral Science*, **2**, 201–215.

Davis, A., Gardner, B. and Gardner, M. (1941) *Deep South*. Chicago Press.

Dawid, A. P. (1988) Symmetry models and hypotheses for structured data layouts. *Journal of the Royal Statistical Society: Series B (Methodological).*, **50**, 1–34.

De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000) A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, **21**, 1253–1278.

Dekker, D., Krackhardt, D. and Snijders, T. A. (2007) Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika*, **72**, 563–581.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**, 1–22.

Desmarais, B. A., Harden, J. J. and Boehmke, F. J. (2015) Persistent policy pathways: Inferring diffusion networks in the American states. *American Political Science Review*, **109**, 392–406.

Eakin, E. (2004) Study finds a nation of polarized readers. *The New York Times*, 9–9.

de Espindolaa, G. M., Pebesmab, E. and Câmaraa, G. (2011) Spatio-temporal regression models for deforestation in the Brazilian Amazon. In *The International Symposium on Spatial-Temporal Analysis and Data Mining, University College London*.

Fafchamps, M. (2006) Development and social capital. *The Journal of Development Studies*, **42**, 1180–1198.

Fafchamps, M. and Gubert, F. (2007) The formation of risk sharing networks. *Journal of Development Economics*, **83**, 326–350.

Fagiolo, G., Reyes, J. and Schiavo, S. (2008) On the topological properties of the world trade web: A weighted network analysis. *Physica A: Statistical Mechanics and its Applications*, **387**, 3868–3873.

Faust, K. and Wasserman, S. (1994) *Social Network Analysis: Methods and Applications*, vol. 249. Cambridge: Cambridge University Press.

Fokianos, K. and Moysiadis, T. (2017) Binary time series models driven by a latent process. *Econometrics and Statistics*, **2**, 117–130.

Fosdick, B. K. and Hoff, P. D. (2014) Separable factor analysis with applications to mortality data. *The Annals of Applied Statistics*, **8**, 120–147.

Foster, A. D. and Rosenzweig, M. R. (2001) Imperfect commitment, altruism, and the family: Evidence from transfer behavior in low-income rural areas. *Review of Economics and Statistics*, **83**, 389–407.

Frank, O. and Strauss, D. (1986) Markov graphs. *Journal of the American Statistical Association*, **81**, 832–842.

Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The Elements of Statistical Learning*. Berlin: Springer Series in Statistics.

Gao, X., Shahbaba, B. and Ombao, H. (2018) Modeling binary time series using Gaussian processes with application to predicting sleep states. *Journal of Classification*, **35**, 549–579.

Gelman, A. and Hill, J. (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Graybill, F. A. (1976) *Theory and Application of the Linear Model*. North Scituate, Massachusetts: Duxbury Press.

Green, A. and Shalizi, C. R. (2017) Bootstrapping exchangeable random graphs. *arXiv:1711.00813*.

Greene, W. H. (2003) *Econometric Analysis*. Prentice Hall.

Griffin, M. and Hoff, P. D. (2019) Structured shrinkage priors. *arXiv preprint arXiv:1902.05106*.

Gupta, A. K. and Nagar, D. (2000) *Matrix variate distributions*. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics ; 104. Boca Raton, FL: Chapman & Hall.

Guyon, X. (1995) *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer Science & Business Media.

Han, G., McCubbins, O. and Paulsen, T. (2016) Using social network analysis to measure student collaboration in an undergraduate capstone course. *NACTA Journal*, **60**, 176.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N. and Morris, M. (2019) *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (https://statnet.org). URL: https://CRAN.R-project.org/package=ergm. R package version 3.10.4.

Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007) Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170**, 301–354.

Handcock, M. S., Robins, G., Snijders, T., Moody, J. and Besag, J. (2003) Assessing degeneracy in statistical models of social networks. *Tech. Rep. 39*, Center for Statistics and the Social Sciences: University of Washington.

Hanneke, S., Fu, W., Xing, E. P. et al. (2010) Discrete temporal models of social networks. *Electronic Journal of Statistics*, **4**, 585–605.

Hansen, B. E. (2015) *Econometrics*. Online at http://www.ssc.wisc.edu/~bhansen/econometrics/.

Hart, J. (1976) Three approaches to the measurement of power in international relations. *International Organization*, **30**, 289–305.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoff, P., Fosdick, B., Volfovsky, A. and He, Y. (2017) *amen: Additive and Multiplicative Effects Models for Networks and Relational Data*. URL: https://CRAN.R-project.org/package=amen. R package version 1.3.

Hoff, P. D. (2005) Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, **100**, 286–295.

— (2008) Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, 657–664.

— (2011) Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis*, **6**, 179–196.

— (2015) Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, **9**, 1169–1193.

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

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

Holland, P. W. and Leinhardt, S. (1981) An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, **76**, 33–50.

Hoover, D. N. (1979) Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, **2**.

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*, vol. 1, 221–233.

Hunter, D. R., Goodreau, S. M. and Handcock, M. S. (2008a) Goodness of fit of social network models. *Journal of the American Statistical Association*, **103**, 248–258.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M. (2008b) ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, **24**, 1–29.

Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, **5**, 248–264.

Jack, W. and Suri, T. (2014) Risk sharing and transactions costs: Evidence from Kenya's mobile money revolution. *The American Economic Review*, **104**, 183–223.

Jarque, C. M. and Bera, A. K. (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, **6**, 255–259.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A.-L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651.

Johnson, J. C. (2017) External threat and alliance formation. *International Studies Quarterly*, **61**, 736–745.

Kallenberg, O. (2006) *Probabilistic Symmetries and Invariance Principles*. Springer Science & Business Media.

Kauermann, G. and Carroll, R. J. (2001) A note on the efficiency of sandwich covariance matrix estimation. *Journal of the Americana Statistical Association*, **96**, 1387–1396.

King, G. and Roberts, M. E. (2015) How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, **23**, 159–179.

Kolda, T. G. and Bader, B. W. (2009) Tensor decompositions and applications. *SIAM Review*, **51**, 455–500.

Krackhardt, D. and Handcock, M. S. (2007) Heider vs Simmel: Emergent features in dynamic structures. In *Statistical Network Analysis: Models, Issues, and New Directions*, 14–27. Springer.

Krivitsky, P. N. (2009) *Statistical models for social network data and processes*. Ph.D. thesis, University of Washington.

Le Cessie, S. and Van Houwelingen, J. (1994) Logistic regression for correlated binary data. *Applied Statistics*, 95–108.

Lehmann, E. L. and Casella, G. (2006) *Theory of Point Estimation*. Springer Science & Business Media.

Li, H. (2006) The covariance structure and likelihood function for multivariate dyadic data. *Journal of Multivariate Analysis*, **97**, 1263–1271.

Li, H. and Loken, E. (2002) A unified theory of statistical analysis and inference for variance component models for dyadic data. *Statistica Sinica*, 519–535.

Li, H. et al. (2002) Modeling through group invariance: An interesting example with potential applications. *The Annals of Statistics*, **30**, 1069–1080.

Li, W.-J., Yeung, D.-Y. and Zhang, Z. (2011) Generalized latent factor models for social network analysis. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Li, Y. and Schafer, D. W. (2008) Likelihood analysis of the multivariate ordinal probit regression model for repeated ordinal responses. *Computational Statistics & Data Analysis*, **52**, 3474–3492.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 13–22.

Linden, G., Smith, B. and York, J. (2003) Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, **7**, 76–80.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Oliver, S. (2006) *SAS for Mixed Models*. SAS publishing.

Liu, J., Shang, M. and Chen, D. (2009) Personal recommendation based on weighted bipartite networks. In *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 5, 134–137. IEEE.

Liu, L., Tang, J., Han, J., Jiang, M. and Yang, S. (2010) Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 199–208. ACM.

Liu, L., Tang, J., Han, J. and Yang, S. (2012) Learning influence from heterogeneous social networks. *Data Mining and Knowledge Discovery*, **25**, 511–544.

Lovász, L. and Szegedy, B. (2006) Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, **96**, 933–957.

Lukes, S. (2004) *Power: A radical view*. Macmillan International Higher Education.

Lumley, T. and Hamblett, N. M. (2003) Asymptotics for marginal generalized linear models with sparse correlations. *UW Biostatistics Working Paper Series*.

Marrs, F. W., Fosdick, B. K. and McCormick, T. H. (2018) *netregR: Regression of network responses*. URL: https://CRAN.R-project.org/package=netregR. R package version 0.3.0.

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

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, vol. 37. CRC press.

McFadden, D. (1973) Conditional logit analysis of qualitative choice behavior.

Mearsheimer, J. J. (2001) *The Tragedy of Great Power Politics*. WW Norton & Company.

Menzel, K. (2017) Bootstrap with clustering in two or more dimensions. *arXiv:1703.03043*.

Minhas, S., Hoff, P. D. and Ward, M. D. (2017) Influence networks in international relations. *arXiv preprint arXiv:1706.09072*.

Moreno, J. L. (1934) Who shall survive?: A new approach to the problem of human interrelations.

Morgenthau, H. (1978) *Politics Among Nations: The Struggle for Power and Peace*. New York: Alfred A. Knopf Inc.

Newman, M. E. (2001) The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, **98**, 404–409.

— (2004) Analysis of weighted networks. *Physical Review E*, **70**, 056131.

Noth, E. M., Hammond, S. K., Biging, G. S. and Tager, I. B. (2011) A spatial-temporal regression model to predict daily outdoor residential PAH concentrations in an epidemiologic study in Fresno, CA. *Atmospheric Environment*, **45**, 2394–2403.

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

Ochi, Y. and Prentice, R. L. (1984) Likelihood inference in a correlated probit regression model. *Biometrika*, **71**, 531–543.

Orbanz, P. and Roy, D. M. (2015) Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 437–461.

Overgoor, J., Benson, A. R. and Ugander, J. (2018) Choosing to grow a graph: Modeling network formation as discrete choice. *arXiv preprint arXiv:1811.05008*.

Petersen, K. B., Pedersen, M. S. et al. (2008) The matrix cookbook. *Technical University of Denmark*, **7**, 510.

Schummer, J. (2005) Reading nano: The public interest in nanotechnology as reflected in purchase patterns of books. *Public Understanding of Science*, **14**, 163–183.

Schweinberger, M. (2011) Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, **106**, 1361–1370.

Sewell, D. K. and Chen, Y. (2015) Latent space models for dynamic networks. *Journal of the American Statistical Association*, **110**, 1646–1657.

Shalizi, C. R. and Rinaldo, A. (2013) Consistency under sampling of exponential random graph models. *The Annals of Statistics*, **41**, 508.

Shi, C., Li, Y., Zhang, J., Sun, Y. and Philip, S. Y. (2017a) A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, **29**, 17–37.

Shi, F., Shi, Y., Dokshin, F. A., Evans, J. A. and Macy, M. W. (2017b) Millions of online book co-purchases reveal partisan differences in the consumption of science. *Nature Human Behaviour*, **1**, 0079.

Sims, C. A. (1980) Macroeconomics and reality. *Econometrica*, **48**, 1–48. URL: http://www.jstor.org/stable/1912017.

Singer, J. D., Bremer, S. and Stuckey, J. (1972) Capability distribution, uncertainty, and major power war, 1820-1965. *Peace, War, and Numbers*, **19**, 48.

Skvoretz, J. and Faust, K. (1999) Logit models for affiliation networks. *Sociological Methodology*, **29**, 253–280.

Snijders, T. A. (2002) Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, **3**, 1–40.

— (2005) Models for longitudinal network data. *Models and Methods in Social Network Analysis*, **1**, 215–247.

Snijders, T. A. and Kenny, D. A. (1999) The social relations model for family data: A multilevel approach. *Personal Relationships*, **6**, 471–486.

Snijders, T. A., Pattison, P. E., Robins, G. L. and Handcock, M. S. (2006) New specifications for exponential random graph models. *Sociological Methodology*, **36**, 99–153.

Stiratelli, R., Laird, N. and Ware, J. H. (1984) Random-effects models for serial observations with binary response. *Biometrics*, 961–971.

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C. and Han, J. (2011) Co-author relationship prediction in heterogeneous bibliographic networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, 121–128. IEEE.

Sun, Y. and Han, J. (2012) Mining heterogeneous information networks: Principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, **3**, 1–159.

Tabord-Meehan, M. (2018) Inference with dyadic data: Asymptotic behavior of the dyadic-robust t-statistic. *Journal of Business & Economic Statistics*, 1–10.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tinbergen, J. (1962) *Shaping the World Economy: Suggestions for an International Economic Policy*. Twentieth Century Fund, New York.

Tucker, L. R. (1964) The extension of factor analysis to three-dimensional matrices. *Contributions to Mathematical Psychology*, **110119**.

Van der Vaart, A. W. (2000) *Asymptotic Statistics*, vol. 3. Cambridge University Press.

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M. and Fortin, M.-J. (2018) Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, **88**, 36–59.

Wakefield, J. (2013) *Bayesian and Frequentist Regression Methods*. Springer Science & Business Media.

Waltz, K. N. (1979) *Theory of International Politics*. Waveland Press.

Wang, P., Sharpe, K., Robins, G. L. and Pattison, P. E. (2009) Exponential random graph (p*) models for affiliation networks. *Social Networks*, **31**, 12–25.

Wang, Y. J. and Wong, G. Y. (1987) Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**, 8–19.

Ward, M. D. and Hoff, P. D. (2007) Persistent patterns of international commerce. *Journal of Peace Research*, **44**, 157–175.

Warner, R. M., Kenny, D. A. and Stoto, M. (1979) A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, **37**, 1742.

Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications*, vol. 8. Cambridge University Press.

Watts, D. J. and Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440.

Westveld, A. H. and Hoff, P. D. (2011) A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *The Annals of Applied Statistics*, 843–872.

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 817–838.

White, H. and Domowitz, I. (1984) Nonlinear regression with dependent observations. *Econometrica*, 143–161.

Wong, G. Y. (1982) Round robin analysis of variance via maximum likelihood. *Journal of the American Statistical Association*, **77**, 714–724.

Wu, T., Yu, S.-H., Liao, W. and Chang, C.-S. (2014) Temporal bipartite projection and link prediction for online social networks. In *Big Data, 2014 IEEE International Conference on*, 52–59. IEEE.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007) Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 329–346.

Zachary, W. W. (1977) An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**, 452–473.

Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**.

Zhang, P.-P., Chen, K., He, Y., Zhou, T., Su, B.-B., Jin, Y., Chang, H., Zhou, Y.-P., Sun, L.-C., Wang, B.-H. et al. (2006) Model and empirical study on some collaboration networks. *Physica A: Statistical Mechanics and its Applications*, **360**, 599–616.

Zhou, T., Ren, J., Medo, M. and Zhang, Y.-C. (2007) Bipartite network projection and personal recommendation. *Physical Review E*, **76**, 046115.

Zhou, Y. and Song, P. X.-K. (2016) Regression analysis of networked data. *Biometrika*, **103**, 287–301.

Zivot, E. and Wang, J. (2006) Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-Plus®*, 385–429.

Appendix A

Influence networks for longitudinal bipartite network data

A.1 Least squares estimation of reduced-rank BLIN model

In this section, we provide a procedure for obtaining the maximum likelihood estimator of the BLIN model assuming reduced rank coefficient matrices $\mathbf{A}$ and $\mathbf{B}$, where the respective ranks are known. The log-likelihood of the data $\{\mathbf{Y}_t\}_{t=1}^T$ is simply the sum of the log-likelihood at each time period since we assume the errors $\mathbf{E}_t$ are independent of each other. The log-likelihood, in terms of the unknown matrices $\{\mathbf{U}, \mathbf{V}, \mathbf{R}, \mathbf{S}\}$, is proportional to

$$\ell \propto \frac{-1}{2} \sum_{t=1}^{T} ||\mathbf{Y}_t - \mathbf{U}\mathbf{V}^T\mathbf{X}_t - \mathbf{Z}_t\mathbf{R}\mathbf{S}^T||_F^2, \tag{A.1}$$

where $\mathbf{U}\mathbf{V}^T$ is a rank k decomposition of $\mathbf{A}$, and $\mathbf{R}\mathbf{S}^T$ is a rank m decomposition of $\mathbf{B}$.

We introduce a block coordinate descent algorithm in Algorithm 3 to obtain the maximum likelihood estimator of the unknown matrices $\{\mathbf{U}, \mathbf{V}, \mathbf{R}, \mathbf{S}\}$. Algorithm 3 estimates each unknown matrix $\{\mathbf{U}, \mathbf{V}, \mathbf{R}, \mathbf{S}\}$ in turn until convergence, in an analogous procedure to the estimation of the full BLIN model in Algorithm 1. The update equation for each unknown matrix is derived by differentiating (A.1) with respect to the unknown matrix, e.g. $\mathbf{U}$, setting this derivative to zero, and solving for the unknown matrix.

Algorithm 3 Block coordinate descent LS estimation of reduced-rank BLIN model

0. Set threshold for convergence η. Set number of iterations $\nu = 1$. Initialize $\{\widehat{\mathbf{U}}^{(0)}, \widehat{\mathbf{V}}^{(0)}, \widehat{\mathbf{R}}^{(0)}, \widehat{\mathbf{S}}^{(0)}\}$ with independent standard normal entries and $Q_0 = \sum_t ||\mathbf{Y}_t||_F^2$.

1. Compute

$$(\widehat{\mathbf{U}}^{(\nu)})^T = \left((\widehat{\mathbf{V}}^{(\nu-1)})^T \sum_{t=1}^{T} \left(\mathbf{X}_t \mathbf{X}_t^T \right) \widehat{\mathbf{V}}^{(\nu-1)} \right)^{-1} (\widehat{\mathbf{V}}^{(\nu-1)})^T \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{Y}_t^T - \sum_{t=1}^{T} \mathbf{X}_t \widehat{\mathbf{S}}^{(\nu-1)} (\widehat{\mathbf{R}}^{(\nu-1)})^T \mathbf{Z}_t^T \right)$$

2. Compute

$$\widehat{\mathbf{V}}^{(\nu)} = \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t^T \right)^{-1} \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{Y}_t^T - \sum_{t=1}^{T} \mathbf{X}_t \mathbf{S}^{(\nu-1)} (\mathbf{R}^{(\nu-1)})^T \mathbf{Z}_t^T \right) \widehat{\mathbf{U}}^{(\nu)} \left((\widehat{\mathbf{U}}^{(\nu)})^T \widehat{\mathbf{U}}^{(\nu)} \right)^{-1}$$

3. Compute

$$\widehat{\mathbf{R}}^{(\nu)} = \left(\sum_{t=1}^{T} \mathbf{Z}_t^T \mathbf{Z}_t \right)^{-1} \left[\sum_{t=1}^{T} \mathbf{Z}_t^T \mathbf{Y}_t - \sum_{t=1}^{T} \mathbf{Z}_t^T \mathbf{U}^{(\nu)} (\mathbf{V}^{(\nu)})^T \mathbf{X}_t \right] \mathbf{S} \left(\mathbf{S}^T \mathbf{S} \right)^{-1}$$

4. Compute

$$(\widehat{\mathbf{S}}^{(\nu)})^T = \left((\mathbf{R}^{(\nu)})^T \left(\sum_{t=1}^{T} \mathbf{Z}_t^T \mathbf{Z}_t \right) \mathbf{R}^{(\nu)} \right)^{-1} (\mathbf{R}^{(\nu)})^T \left[\sum_{t=1}^{T} \mathbf{Z}_t^T \mathbf{Y}_t - \sum_{t=1}^{T} \mathbf{Z}_t^T \mathbf{U}^{(\nu)} (\mathbf{V}^{(\nu)})^T \mathbf{X}_t \right]$$

5. Compute the least squares criterion

$$Q_\nu = \sum_t ||\mathbf{Y}_t - \mathbf{U}^{(\nu)} (\mathbf{V}^{(\nu)})^T \mathbf{X}_t - \mathbf{Z}_t \mathbf{R}^{(\nu)} (\mathbf{S}^{(\nu)})^T||_F^2.$$

If $|Q_\nu - Q_{\nu-1}| > \eta$, increment ν and return to 1.

A.2 Proofs of theoretical results

We begin with the proof of Proposition 1. We then prove Theorem 4, as expressions in this proof support the proofs of the remaining Propositions 2, 3, 5, and 6. We take $\mathbf{X}_t$ and $\mathbf{Y}_t$ mean zero without loss of generality.

Proof of Proposition 1. Noting that $\mathbb{X}_B$ is of dimension $TSL \times S^2 + L^2 - 1$, it is sufficient to show that $\mathbb{X}_B$ is full rank under the assumptions. We treat two cases: (1) $TSL \leq S^2 + L^2 - 1$ and (2) $TSL > S^2 + L^2 - 1$.

Case (1):

We first show that $TSL \leq (S^2 + L^2 - 1)$ implies $TS \leq L$. Assume towards a contradiction that $TS > L$ and let $S = \alpha L$ for $\alpha \in (0, 1]$. Then,

$$\begin{aligned} TSL &\leq S^2 + L^2 - 1, \\ T\alpha L^2 &\leq (1 + \alpha^2)L^2 - 1, \\ T &\leq 1/\alpha + \alpha - 1/L^2. \end{aligned} \tag{A.2}$$

If $TS > L$, then $T > 1/\alpha$. As there is no integer T that satisfies (A.2) and $T > 1/\alpha$, we have that $TS \leq L$.

Now, as $TSL \leq S^2 + L^2 - 1$, we have that $\text{rank}(\mathbb{X}_B) \leq TSL$. Assume towards a contradiction that $\text{rank}(\mathbb{X}_B) < TSL$. Then, for some nonzero $u \in \mathbb{R}^{TSL}$, the assumption implies that $u^T \mathbb{X}_B = 0$. Consider the columns $S^2 + 1$ through $S^2 + 1 + L$ of $\mathbb{X}_B$. The assumption implies that, for some nonzero $v \in \mathbb{R}^{TS}$, that $v^T[\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_t]$. For $TS \leq L$, this is a contradiction of the assumption that $[\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_t]$ is full rank. Thus, we have that $\mathbb{X}_B$ is full rank in case (1).

Case (2):

Now we take $TSL > S^2 + L^2 - 1$, such that $\text{rank}(\mathbb{X}_B) \leq S^2 + L^2 - 1$. Assume towards a contradiction that $\text{rank}(\mathbb{X}_B) < S^2 + L^2 - 1$. Then, there exists some $u_1 \in \mathbb{R}^{S^2}$ and $u_2 \in \mathbb{R}^{L^2}$ such that $\mathbb{X}_B u = 0$, for $u^T = [u_1, u_2]$, nonzero and not utilizing the nonidentifiability exclusively. By the single nonidentifiability of the BLIN model, for $\mathbb{X}_B u = 0$, either u utilizes the nonidentifiability or both u_1 and u_2 are in the null spaces of $[\mathbf{X}_1^T \otimes \mathbf{I}_L; \mathbf{X}_2^T \otimes \mathbf{I}_L; \ldots; \mathbf{X}_t^T \otimes \mathbf{I}_L]$ and $[\mathbf{I}_S \otimes \mathbf{Z}_1; \mathbf{I}_S \otimes \mathbf{Z}_2; \ldots; \mathbf{I}_S \otimes \mathbf{Z}_t]$, respectively. However, by the discussion under case (1), we have that $TS \geq L$ such that, by assumption, $[\mathbf{X}_1; \mathbf{X}_2; \ldots; \mathbf{X}_t]$ and $[\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_t]$ are of rank L. This also implies that the $TL \times S$ matrix $[\mathbf{X}_1^T; \mathbf{X}_2^T; \ldots; \mathbf{X}_t^T]$ is full rank,

which is rank S as $TL \geq S$. Then, again using the discussion in case (1), the two matrices $[\mathbf{I}_S \otimes \mathbf{Z}_1; \mathbf{I}_S \otimes \mathbf{Z}_2; \ldots; \mathbf{I}_S \otimes \mathbf{Z}_t]$ and $[\mathbf{X}_1^T \otimes \mathbf{I}_L; \mathbf{X}_2^T \otimes \mathbf{I}_L; \ldots; \mathbf{X}_t^T \otimes \mathbf{I}_L]$ are full rank, which are S^2 and L^2, respectively. So, the only way that $\mathbb{X}_B u = 0$ and u is nonzero is utilizing the nonidentifiability. This is a contradiction, as this implies that $\text{rank}(\mathbb{X}_B) = S^2 + L^2 - 1$, and $\mathbb{X}_B$ is full rank under case (2). □

Proof of Theorem 4. Recall that we work in the setting of $p = p_A = p_B$, as the bilinear model cannot accommodate different lags for the different influence types. Thus, we have $\mathbf{Z}_t = \mathbf{X}_t$ for all $t \in \{1, 2, \ldots, T\}$ and the covariance matrices $\mathbf{\Omega}_X = \mathbf{\Omega}_Z = \mathbf{\Omega}$ and $\mathbf{\Psi}_X = \mathbf{\Psi}_Z = \mathbf{\Psi}$.

For either generative model, we may write the useful form

$$E[\mathbf{y}_t | \mathbf{x}_t] = \mathbf{\Theta}\mathbf{x}_t, \tag{A.3}$$

such that $\mathbf{\Theta} = [\mathbf{I}_L \otimes \mathbf{A}; \mathbf{B}^T \otimes \mathbf{I}_S]$ for the BLIN model and $\mathbf{\Theta} = \mathbf{B}^T \otimes \mathbf{A}$ for the bilinear model. We examine the impacts of specifying either the BLIN or bilinear model as the generating model in the proof. First, however, we write the pseudo-true parameters for least squares estimators of the BLIN and bilinear models, respectively, under the general generative structure in (A.3).

For least squares estimation of the BLIN model, we may write the pseudo-true parameters for $\mathbf{A}$ as

$$\widetilde{\mathbf{A}}^T = E\left[\mathbf{Y}_t \mathbf{X}_t^T - \mathbf{X}_t \widetilde{\mathbf{B}} \mathbf{X}_t^T\right] E\left[\mathbf{X}_t \mathbf{X}_t^T\right]^{-1}, \tag{A.4}$$

$$= E\left[\sum_{j=1}^{L} E[\mathbf{y}_{\cdot jt} \mid \mathbf{x}_t] \mathbf{x}_{\cdot jt}^T - \sum_{i=1}^{L}\sum_{j=1}^{L} \widetilde{b}_{ij} \mathbf{x}_{\cdot it} \mathbf{x}_{\cdot jt}^T\right] E\left[\mathbf{X}_t \mathbf{X}_t^T\right]^{-1}, \tag{A.5}$$

where we obtain (A.4) by maximizing the expression for the BLIN pseudo-true parameters in (2.15) with respect to $\mathbf{A}$. Then, we may exploit the assumption that $E[\mathbf{y}_{\cdot jt} \mid \mathbf{x}_t]$ is a linear function of $\mathbf{x}_t$. Letting $\mathcal{C}_{k\ell}$ be the $S \times S$ partition of $\mathbf{\Theta}$ relating column ℓ of $\mathbf{X}_t$ to column k of $\mathbf{Y}_t$, substituting into (A.5), we obtain

$$\widetilde{\mathbf{A}}^T = \left(\sum_{k=1}^{L}\sum_{\ell=1}^{L} \mathcal{C}_{k\ell}\omega_{k\ell} - \text{tr}(\mathbf{\Omega}\widetilde{\mathbf{B}})\mathbf{I}_S\right) \Big/ \text{tr}(\mathbf{\Omega}), \tag{A.6}$$

where and $\omega_{k\ell}$ is the (k, ℓ) entry in $\mathbf{\Omega}$ (and we use the symmetric property of $\mathbf{\Omega}$) and the terms concerning $\mathbf{\Psi}$ cancel. We note that the BLIN pseudo-true parameters $\widetilde{\mathbf{A}}$ exist by applying Conditions 1 to the explicit

expression for $\widehat{\mathbf{A}}$ in (2.12), using the law of large numbers. Hoff (2015) writes the pseudo-true parameters for the bilinear estimator of $\mathbf{A}$, which we denote $\bar{\mathbf{A}}$, under least squares estimation as:

$$\bar{\mathbf{A}}^T = E\left[\mathbf{Y}_t\bar{\mathbf{B}}\mathbf{X}_t\right] E\left[\mathbf{X}_t\bar{\mathbf{B}}\bar{\mathbf{B}}^T\mathbf{X}_t^T\right]^{-1}, \tag{A.7}$$

$$= \mathrm{tr}(\mathbf{\Omega}\bar{\mathbf{B}}\bar{\mathbf{B}}^T)^{-1}\sum_{j=1}^{L}\sum_{k=1}^{L}\left(\sum_{\ell=1}^{L}\mathbf{\Omega}_{k\ell}\bar{b}_{j\ell}\right)\mathcal{C}_{jk}, \tag{A.8}$$

where $\bar{b}_{j\ell}$ is the (j,ℓ) entry in $\bar{\mathbf{B}}$, the pseudo-true parameters for the bilinear estimator of $\mathbf{B}$. The bilinear pseudo-true parameters $\bar{\mathbf{A}}$ exist as Conditions 1 satisfy those given in Hoff (2015).

We have derived the pseudo-true parameters for the least squares estimators of the BLIN and bilinear models in (A.6) and (A.8), respectively, whenever $E[\mathbf{y}_t|\mathbf{x}_t] = \mathbf{\Theta}\mathbf{x}_t$. Now, we address the specific BLIN and bilinear generating models. That is, we specify the $S \times S$ matrices $\{\mathcal{C}_{k\ell}\}_{k,\ell}$ that partition $\mathbf{\Theta}$ under each generating model and examine the resulting pseudo-true parameters.

When the data $\{\mathbf{Y}_t\}_{t=1}^T$ are generated by the BLIN model, the matrix $\mathcal{C}_{jk} = \mathbf{A}^T + b_{jk}\mathbf{I}_s$ when $k = \ell$ and $\mathcal{C}_{jk}$ is diagonal otherwise. Substituting into (A.8), we see:

$$\bar{\mathbf{A}}^T = \mathrm{tr}(\mathbf{\Omega}\bar{\mathbf{B}}\bar{\mathbf{B}}^T)^{-1}\sum_{j=1}^{L}\left(\sum_{\ell=1}^{L}\omega_{j\ell}\bar{b}_{j\ell}\right)\mathbf{A}^T + c_1\mathbf{I}_s, \tag{A.9}$$

$$= \frac{\mathrm{tr}(\mathbf{\Omega}\bar{\mathbf{B}})}{\mathrm{tr}(\mathbf{\Omega}\bar{\mathbf{B}}\bar{\mathbf{B}}^T)}\mathbf{A}^T + \frac{\mathrm{tr}(\mathbf{\Omega}\bar{\mathbf{B}}\mathbf{B}^T)}{\mathrm{tr}(\mathbf{\Omega}\bar{\mathbf{B}}\bar{\mathbf{B}}^T)}\mathbf{I}_s. \tag{A.10}$$

Thus, the off-diagonal pseudo-true parameters $\bar{\mathbf{A}}$ are equal to the off-diagonal entries in $\mathbf{A}$, up to a multiplicative constant, when the data are generated by the BLIN model.

When the data $\{\mathbf{Y}_t\}_{t=1}^T$ are generated by the bilinear model, the matrix $\mathcal{C}_{jk} = b_{jk}\mathbf{A}^T$ for all (j,k). Substituting into (A.6), we see

$$\widetilde{\mathbf{A}}^T = \left(\sum_{k=1}^{L}\sum_{\ell=1}^{L} b_{k\ell}\omega_{\ell k}\mathbf{A}^T - \mathrm{tr}(\mathbf{\Omega}\widetilde{\mathbf{B}})\mathbf{I}_S\right)\Big/\mathrm{tr}(\mathbf{\Omega}), \tag{A.11}$$

$$= \frac{\mathrm{tr}(\mathbf{\Omega}\mathbf{B})}{\mathrm{tr}(\mathbf{\Omega})}\mathbf{A}^T - \frac{\mathrm{tr}(\mathbf{\Omega}\widetilde{\mathbf{B}})}{\mathrm{tr}(\mathbf{\Omega})}\mathbf{I}_s, \tag{A.12}$$

and $\mathrm{tr}(\mathbf{\Omega}) \neq 0$ by assumption of $\mathbf{\Omega}$ positive definite. Thus, the off-diagonal pseudo-true parameters for $\widetilde{\mathbf{A}}$ are equal to the off-diagonal entries in $\mathbf{A}$, up to a multiplicative constant, when the data are generated by the bilinear model.

These results hold for the off-diagonal pseudo-true parameters as estimated for the BLIN and bilinear models, $\widetilde{\mathbf{B}}$ and $\bar{\mathbf{B}}$, respectively. This fact can be seen by transposing the BLIN and bilinear models and swapping the roles $\mathbf{B}$ for $\mathbf{A}$ in the above proof, i.e. writing the BLIN model as $E[\mathbf{Y}_t^T|\mathbf{X}_t] = \mathbf{B}^T\mathbf{X}_t^T + \mathbf{X}_t^T\mathbf{A}$ and applying the above arguments for $\mathbf{A}$ to $\mathbf{B}^T$. □

Proof of Proposition 2. The pseudo-true parameter under least squares estimation of the BLIN model, $\widetilde{\mathbf{A}}$, is given in (A.4). Under the assumptions in Proposition 2, the matrix $E\left[\mathbf{X}_t\mathbf{X}_t^T\right]^{-1}$ is diagonal with entries $1/E[\mathbf{x}_{i\cdot t}^T\mathbf{x}_{i\cdot t}]$. Then, the off-diagonal (i,j) entry of $\widetilde{A}$ is

$$\widetilde{a}_{ij} = \frac{E[\mathbf{x}_{i\cdot t}^T\mathbf{y}_{j\cdot t}] - \mathrm{tr}(E[\mathbf{x}_{i\cdot t}\mathbf{z}_{j\cdot t}^T]\widetilde{\mathbf{B}}^T)}{E[\mathbf{x}_{i\cdot t}^T\mathbf{x}_{i\cdot t}]}, \quad i \neq j. \tag{A.13}$$

By the assumptions given in Proposition 2, both terms in the numerator are zero as is the coefficient.

The result for $\widetilde{b}_{ij}$ when $i \neq j$ follows from considering the transpose of the BLIN model as in the proof of Theorem 4, that is, swapping the roles of $\mathbf{A}$ and $\mathbf{B}$ and the roles of $\mathbf{X}_t$ and $\mathbf{Z}_t$ in (A.13). □

Proof of Proposition 3. Refer to the expression for the pseudo-true parameter under least squares estimation of the BLIN model, $\widetilde{\mathbf{A}}$, in (A.6). The entries relating row i of $\mathbf{X}_t$ to row j of $\mathbf{Y}_t$ are the (j,i) entries in $\mathcal{C}_{k\ell}$ for all (k,ℓ). These are zero by assumption when $i \neq j$. Thus, the assumptions imply $\widetilde{a}_{ij} = 0$ when $i \neq j$. There is no issue when $\mathbf{Z}_t \neq \mathbf{X}_t$, as this change only enters (A.6) through $\mathbf{\Omega}$, which value is immaterial to the argument.

Again, the result for $\widetilde{b}_{ij}$ when $i \neq j$ follows from considering the transpose of the BLIN model as in the proof of Theorem 4. □

Proof of Proposition 5. As in the proof of Theorem 4, we work in the setting of $p = p_A = p_B$, Thus, we have $\mathbf{Z}_t = \mathbf{X}_t$ for all $t \in \{1, 2, \ldots, T\}$ and the covariance matrices $\mathbf{\Omega}_X = \mathbf{\Omega}_Z = \mathbf{\Omega}$ and $\mathbf{\Psi}_X = \mathbf{\Psi}_Z = \mathbf{\Psi}$.

Under least squares estimation of the BLIN model, using (A.12), we may write the pseudo-true diagonal specification as

$$\widetilde{a}_{ii} + \widetilde{b}_{jj} = \frac{\mathrm{tr}(\mathbf{B})}{L} a_{ii} + \frac{\mathrm{tr}(\mathbf{A})}{S} b_{jj} - \frac{\mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})}{SL}, \tag{A.14}$$

where we use the condition of matrices $\mathbf{\Omega}$ and $\mathbf{\Psi}$ proportional to the identities of appropriate size. Then, the traces of $\mathbf{A}$ and $\mathbf{B}$ are simply αS and βL, respectively. Substituting, we find

$$\widetilde{a}_{ii} + \widetilde{b}_{jj} = \alpha\beta + \alpha\beta - \frac{SL\alpha\beta}{SL} = \alpha\beta \; \forall \, i, j, \tag{A.15}$$

which is the specification of the diagonal entries under the bilinear model, $\alpha\beta = a_{ii} b_{jj} \; \forall \, i, j$.

We now turn to least squares estimation of the bilinear model. It is sufficient to provide a counterexample to prove the claim. Set $\mathbf{A} = \alpha \mathbf{I}_S + \mathbf{A}_0$, where $\mathbf{A}_0$ has diagonal of all zeros, and the same for $\mathbf{B} = \beta \mathbf{I}_L + \mathbf{B}_0$, for some $\alpha + \beta \neq 0$. Using the expression for the bilinear pseudo-true parameters in (A.10), we may write pseudo-true diagonal specification under least squares estimation of the bilinear model as

$$\bar{a}_{ii} \bar{b}_{jj} = \frac{\mathrm{tr}(\bar{\mathbf{A}})\mathrm{tr}(\bar{\mathbf{B}})}{\mathrm{tr}(\bar{\mathbf{A}}^T \bar{\mathbf{A}})\mathrm{tr}(\bar{\mathbf{B}}^T \bar{\mathbf{B}})} \left(\alpha + \beta + \frac{\mathrm{tr}(\bar{\mathbf{A}} \mathbf{A}_0^T)}{\mathrm{tr}(\bar{\mathbf{A}})} \right) \left(\alpha + \beta + \frac{\mathrm{tr}(\bar{\mathbf{B}} \mathbf{B}_0^T)}{\mathrm{tr}(\bar{\mathbf{B}})} \right), \tag{A.16}$$

for any pair (i, j). Again using the expression for the bilinear pseudo-true parameters in (A.10), we have that

$$\frac{\mathrm{tr}(\bar{\mathbf{A}})}{\mathrm{tr}(\bar{\mathbf{A}}^T \bar{\mathbf{A}})} \frac{\mathrm{tr}(\bar{\mathbf{B}})}{\mathrm{tr}(\bar{\mathbf{B}}^T \bar{\mathbf{B}})} = \alpha + \beta + \frac{\mathrm{tr}(\bar{\mathbf{A}} \mathbf{A}_0^T)}{\mathrm{tr}(\bar{\mathbf{A}})} + \frac{\mathrm{tr}(\bar{\mathbf{B}} \mathbf{B}_0^T)}{\mathrm{tr}(\bar{\mathbf{B}})}. \tag{A.17}$$

Substituting (A.17) into (A.16), we see that

$$\begin{aligned} \bar{a}_{ii} \bar{b}_{jj} &= \alpha + \beta + \frac{c_A c_B}{\alpha + \beta + c_A + c_B}, \\ \text{where } \; c_A &:= \frac{\mathrm{tr}(\bar{\mathbf{A}} \mathbf{A}_0^T)}{\mathrm{tr}(\bar{\mathbf{A}})}, \\ c_B &:= \frac{\mathrm{tr}(\bar{\mathbf{B}} \mathbf{B}_0^T)}{\mathrm{tr}(\bar{\mathbf{B}})}. \end{aligned} \tag{A.18}$$

This expression is equal to the true diagonal specification $\alpha + \beta$ if and only at least one of c_A or c_B is equal zero. For example, take all entries in $\mathbf{A}_0$ and $\mathbf{B}_0$ to be zero except for a_{12} and b_{12}. Then,

$$c_A c_B = \frac{a_{12}^2 b_{12}^2}{SL\left(\alpha + \beta + c_A\right)\left(\alpha + \beta + c_B\right)} \neq 0. \tag{A.19}$$

This case establishes a counterexample for $\bar{a}_{ii}\bar{b}_{jj} \neq \alpha + \beta$. □

Proof of Proposition 6. By assumption, the diagonals of the BLIN estimators $\{\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}\}$ are constant; we let these values be α and β, respectively. Thus, we have

$$\widetilde{\mathbf{A}} = \alpha \mathbf{I}_S + \mathbf{A}_0, \qquad \widetilde{\mathbf{B}} = \beta \mathbf{I}_L + \mathbf{B}_0, \tag{A.20}$$

where $\mathbf{A}_0$ and $\mathbf{B}_0$ are matrices with zeros along the diagonal.

Now, the assumption of equivalences of the estimators states that the bilinear estimator, for example $\bar{\mathbf{A}}$, has off-diagonal entries that are equivalent to $\mathbf{A}_0$ multiplied by the diagonal entry β, by Theorem 4. The same is true for the off-diagonal entries of $\bar{\mathbf{B}}$, which are equivalent to $\alpha\mathbf{B}_0$. The diagonal entries of the bilinear estimators $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are not unique, but $\bar{a}_{ii} = \text{sign}(\alpha + \beta)\sqrt{|\alpha + \beta|}$ and $\bar{b}_{jj} = \sqrt{|\alpha + \beta|}$ satisfies the equivalence relation $\bar{a}_{ii}\bar{b}_{jj} = \widetilde{a}_{ii} + \widetilde{b}_{jj}$ for all $i \in \{1, 2, \ldots, S\}$ and $j \in \{1, 2, \ldots, L\}$. Thus, we have the blinear estimators

$$\bar{\mathbf{A}} = \text{sign}(\alpha + \beta)\sqrt{|\alpha + \beta|}\mathbf{I}_S + \beta \mathbf{A}_0, \qquad \bar{\mathbf{B}} = \sqrt{|\alpha + \beta|}\mathbf{I}_L + \alpha \mathbf{B}_0, \tag{A.21}$$

Now assume towards a contradiction that the estimated means of the BLIN and bilinear models are equal, that is,

$$\widetilde{\mathbf{A}}^T \mathbf{X}_t + \mathbf{X}_t \widetilde{\mathbf{B}} = \bar{\mathbf{A}}^T \mathbf{X}_t \bar{\mathbf{B}}, \quad t \in \{1, 2, \ldots, T\}, \tag{A.22}$$

Then, substituting the matrices in (A.20) and (A.21), (A.22) implies the following:

$$\beta \mathbf{A}_0{}^T \mathbf{X}_t + \alpha \mathbf{X}_t \mathbf{B}_0 = \beta \mathbf{A}_0{}^T \mathbf{X}_t + \alpha \mathbf{X}_t \mathbf{B}_0 + \mathbf{A}_0{}^T \mathbf{X}_t \mathbf{B}_0, \tag{A.23}$$

$$\mathbf{0} = \mathbf{A}_0{}^T \mathbf{X}_t \mathbf{B}_0, \quad t \in \{1, 2, \ldots, T\}, \tag{A.24}$$

which is true for general $\mathbf{X}_t$ with probability zero unless both $\mathbf{A}_0$ and $\mathbf{B}_0$ are zero. □

A.3 Details of simulation studies

We provide details of the simulation studies performed to compare the bilinear and BLIN models. We first detail the cross-validation study. We then discuss a convergence study verifying Theorem 4 and Proposition 5.

A.3.1 Cross-validation study

We began by generating weighted networks of rank 1, that is,

$$\mathbf{A}_0 = \mathbf{u}\mathbf{v}^T, \qquad \mathbf{B}_0 = \mathbf{r}\mathbf{s}^T,$$

with each vector $\mathbf{u}$, $\mathbf{v}$, $\mathbf{r}$, and $\mathbf{s}$ of length 10 and consisting of independent and identically distributed standard normal random variables. Then, to remove self-loops in the networks, we set the diagonal entries of $\mathbf{A}_0$ and $\mathbf{B}_0$ to zero, thus arriving at true influence networks $\mathbf{A}$ and $\mathbf{B}$. To generate values of $\mathbf{A}$ and $\mathbf{B}$ with fractions of zeros $q = 0.5$ and $q = 0.9$, we simply set the absolute smallest 50% and 90% of entries in $\mathbf{A}$ and $\mathbf{B}$ to zero, respectively.

To control the signal-to-noise ratio of the models, we set the large-sample R^2 of each generative model to be about 0.75. To do so, we scaled the true $\mathbf{A}$ and $\mathbf{B}$ by a multiplicative constant based on the generating model and sparsity level. We now derive the formulas for these constants for the general lag 1 autoregressive model in (2.16). By definition, the in-sample R^2 of any estimator of the mean of $\mathbf{y}_t$, denoted here $\{\widehat{\mathbf{y}}_t\}_{t=1}^T$, is

$$R^2 = \frac{\frac{1}{T}\sum_{t=1}^T ||\mathbf{y}_t||_2^2 - \frac{1}{T}\sum_{t=1}^T ||\widehat{\mathbf{y}}_t - \mathbf{y}_t||_2^2}{\frac{1}{T}\sum_{t=1}^T ||\mathbf{y}_t||_2^2}. \tag{A.25}$$

Now, we assume that $\boldsymbol{\Theta}$ is known, such that $\widehat{\mathbf{y}}_t = \boldsymbol{\Theta}\mathbf{y}_{t-1}$ for all t. Then, for large samples and independent errors, the Law of Large Numbers implies that the sample averages in (A.25) converge (in probability) to their expectations. Thus, we have that, for large samples,

$$R^2 \approx \frac{\text{tr}\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T E[\mathbf{y}_{t-1}\mathbf{y}_{t-1}^T]\right)}{\text{tr}\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T E[\mathbf{y}_{t-1}\mathbf{y}_{t-1}^T]\right) + E[\mathbf{e}_t^T\mathbf{e}_t]} = \frac{\text{tr}\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T E[\mathbf{y}_{t-1}\mathbf{y}_{t-1}^T]\right)}{\text{tr}\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T E[\mathbf{y}_{t-1}\mathbf{y}_{t-1}^T]\right) + SL}, \tag{A.26}$$

with $S = L = 10$ in this simulation. Further, it can be shown that for any stationary mean-zero VAR model of the form of (2.16) that

$$\text{vec}\left(E[\mathbf{y}_{t-1}\mathbf{y}_{t-1}^T]\right) = \left(\mathbf{I}_{S^2L^2} - \boldsymbol{\Theta}\otimes\boldsymbol{\Theta}\right)^{-1}\text{vec}\left(\mathbf{I}_{SL}\right). \tag{A.27}$$

For the BLIN model, we scaled both $\mathbf{A}$ and $\mathbf{B}$ by a constant k_{BLIN} and defined $\boldsymbol{\Theta}_{k,q} = k_{BLIN}[\mathbf{I}_L \otimes \mathbf{A}_q; \mathbf{B}_q^T \otimes \mathbf{I}_S]$, where the subscript '$k$' emphasizes that $\boldsymbol{\Theta}_{k,q}$ depends on k_{BLIN} and the subscript 'q' signifies the $\mathbf{A}_q$ and $\mathbf{B}_q$ influence networks correspond to one of the sparsity levels $q \in \{0.0, 0.5, 0.9\}$. We then defined

$$g(k_{BLIN}, q) = \text{tr}\left(\boldsymbol{\Theta}_{k,q}\boldsymbol{\Theta}_{k,q}^T E[\mathbf{y}_{t-1}\mathbf{y}_{t-1}^T]\right) = \text{tr}\left(\boldsymbol{\Theta}_{k,q}\boldsymbol{\Theta}_{k,q}^T E[\mathbf{y}_{t-1}\mathbf{y}_{t-1}^T]\right) \tag{A.28}$$

$$= \text{vec}\left(\boldsymbol{\Theta}_{k,q}\boldsymbol{\Theta}_{k,q}^T\right)\left(\mathbf{I}_{S^2L^2} - \boldsymbol{\Theta}_{k,q}\otimes\boldsymbol{\Theta}_{k,q}\right)^{-1}\text{vec}\left(\mathbf{I}_{SL}\right). \tag{A.29}$$

For each $q \in \{0.0, 0.5, 0.9\}$, we selected k_{BLIN} such that $0.75 \approx g(k_{BLIN}, q)/(g(k_{BLIN}, q) + SL)$, where the '$\approx$' simply indicates that k_{BLIN} was selected by discrete search. We repeated the same procedure to select $k_{bilinear}$, although the $\boldsymbol{\Theta}_{k,q}$ in this case is $\boldsymbol{\Theta}_{k,q} = k_{bilinear}^2\mathbf{B}_q^T \otimes \mathbf{A}_q$. Every generating model selected was stationary.

For each combination of $q \in \{0.0, 0.5, 0.9\}$ and generating model, e.g. BLIN or bilinear, ($3 \times 2 = 6$ total simulation settings), we generated 100 data realizations. To ensure stationarity was reached, we simulated 100 time periods of each realization and used the final T observations, for $T \in \{10, 20, 50\}$, as simulated data realizations. To compute out-of-sample R^2 for each simulated data set, we estimated the sparse, (reduced) rank 1, full BLIN models, and the bilinear model on each of 10 training data sets within a 10-fold cross-validation. This procedure is described in detail in Section 2.5. We focused on $q = 0.9$ in the text in Section 2.5, however, the out-of-sample results for $q \in \{0.0, 0.5\}$ are given in Figure A.1.

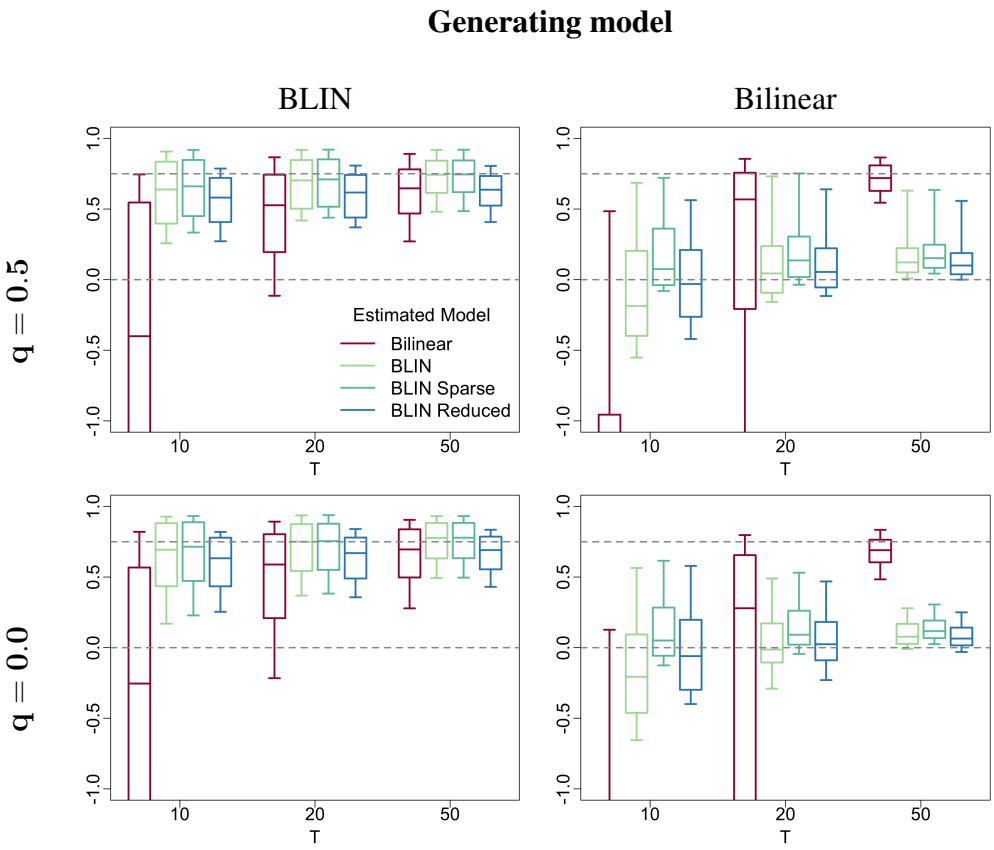


Figure A.1: Out-of-sample R^2 values for each estimation procedure applied to 100 data realizations, with true coefficient matrices of sparsity $q \in \{0.0, 0.5\}$. The left panel is results for data generated from the BLIN model and the right panel shows that for data generated from the bilinear model. The centers of the boxplots represent the median R^2 value, the boxes represent the middle 80% of R^2 values, and the whiskers correspond to the maximum and minimum R^2 values across 100 simulated data sets. Plots are truncated such that R^2 values less than -1 are not shown.

For completeness, we examined in-sample R^2 values for the generated data as well. In this case, we performed no cross-validation, but simply estimated each model on each complete data realization, i.e. for a given generating model, q, and T. The in-sample R^2 values are shown in Figure A.2.

A.3.2 Likelihood of bilinear model

In this section, we conduct an investigation into the source of poor performance of the bilinear estimator when data is generated from this same model. To do so, we examine in-sample and out-of-sample R^2 values at various coefficient values. These coefficient values are weighted averages of true and estimated coefficients. Recall that, for normally-distributed data, large R^2 values correspond to large likelihood values and vice versa, so we use the terms, e.g., "high R^2" and "large likelihood" interchangeably for this discussion.

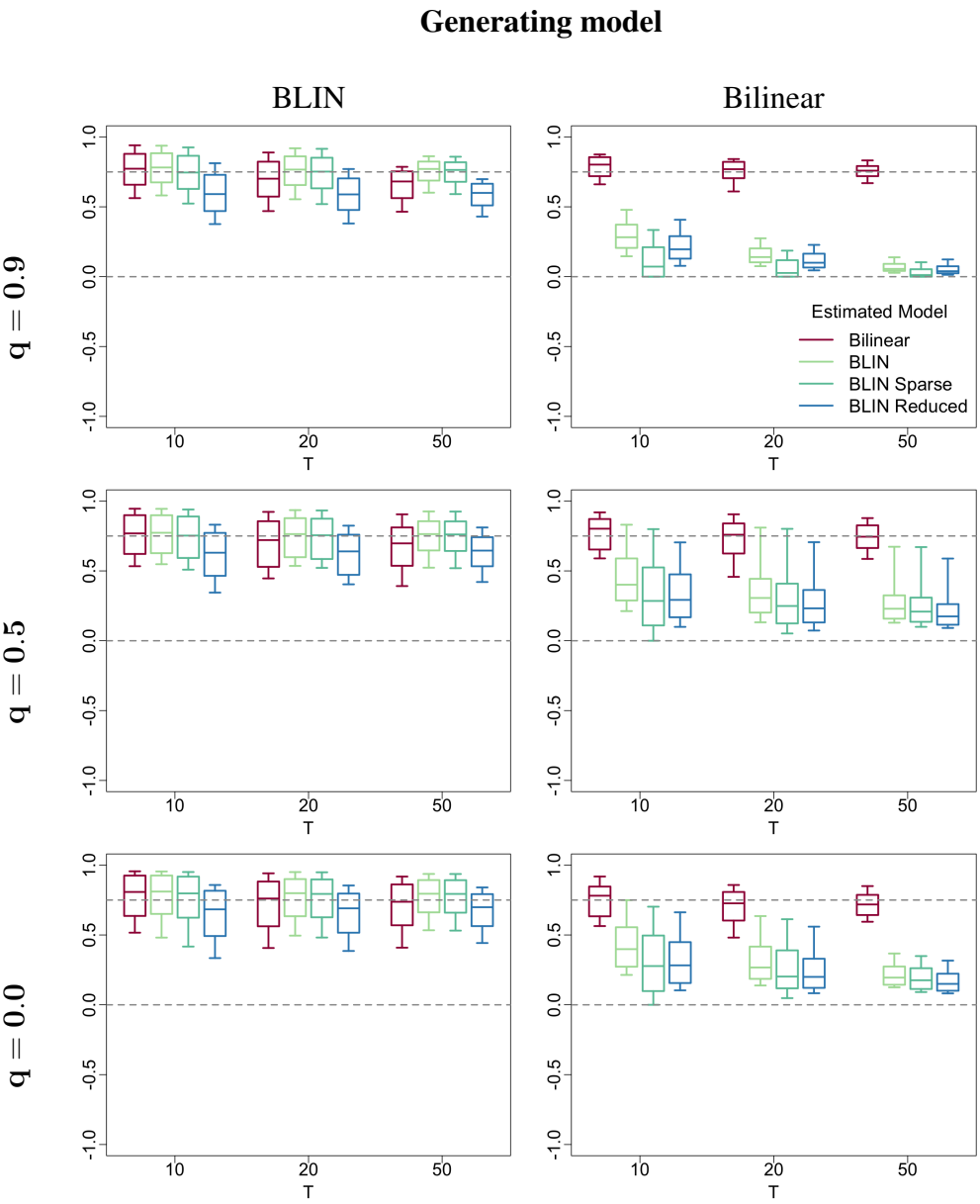


Figure A.2: In-sample R^2 values for each estimation procedure applied to 100 data realizations, with true coefficient matrices of sparsity $q \in \{0.0, 0.5, 0.9\}$. The left panel is results for data generated from the BLIN model and the right panel shows that for data generated from the bilinear model. The centers of the boxplots represent the median R^2 value, the boxes represent the middle 80% of R^2 values, and the whiskers correspond to the maximum and minimum R^2 values across 100 simulated data sets.

Recall that the bilinear model was estimated in a 10-fold cross-validation (see Section 2.5), and thus, the bilinear estimator was computed on a subset of the data containing approximately 90% of the full simulated data set. For a particular subset of a particular simulated data set, we computed in-sample and out-of-sample R^2 values at a mixture of the true coefficients and the coefficients estimated on the particular subset. That

is, we computed R^2 as a function of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector of the entries in $\mathbf{A}$ and $\mathbf{B}$ matrices defined by the mixture

$$\boldsymbol{\theta}^T = (1-\xi)[\text{vec}(\mathbf{A}), \text{vec}(\mathbf{B})] + \xi[\text{vec}(\widehat{\mathbf{A}}), \text{vec}(\widehat{\mathbf{B}})], \tag{A.30}$$

where $\{\mathbf{A}, \mathbf{B}\}$ are the true coefficients and $\{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}\}$ are the coefficients estimated on the 90% data subset. Thus, we examined the likelihood surface along a line in the coefficient space connecting the true and estimated coefficients. This likelihood surface may show multiple local optima, should they exist.

In Figure A.3, we show the R^2 values when generating from the bilinear model and estimating the bilinear model. We see that, when $\xi = 1$, the value of in-sample R^2 is high, as it must be, since $\boldsymbol{\theta}$ is the MLE when $\xi = 1$. However, we also see that in-sample R^2 is high near $\xi = 0$, i.e. near the true coefficients. Between $\xi = 0$ and $\xi = 1$, the in-sample R^2 drops substantially. This is evidence of multiple modes, or multiple local optima, in the likelihood of the bilinear model. Of course, we observe that the out-of-sample R^2 near $\xi = 0$ is also high, that is, selecting coefficient values near the true coefficients generalize well in prediction out-of-sample. However, near $\xi = 1$, even though the in-sample R^2 mode is higher than the one near $\xi = 0$, the out-of-sample R^2 value is low. Thus, choosing the mode with highest in-sample R^2, as is done in maximum-likelihood estimation procedures, does not necessarily generalize well out-of-sample, as shown in this case. This fact, taken with the poor representation of $\mathbf{A}$ by the $\widehat{\mathbf{A}}$ matrix in Figure A.3 (we observe similar issues for $\mathbf{B}$ and $\widehat{\mathbf{B}}$), suggests that $\xi = 1$ is far from $\xi = 0$ in coefficient space. That is, the likelihood of the bilinear model (at least for $T = 10$ and $q = 0.50$) has multiple modes that are not near one another. Finding all of the modes is an extremely difficult – if not impossible – task, and selecting an incorrect mode may result in poor estimator performance. We find similar patterns to those in Figure A.3 for $T = 20$ and $q \in \{0.0, 0.9\}$ when generating from the bilinear model and estimating the bilinear model.

A similar analysis for the BLIN model confirms the unimodality implied by Proposition 1 (see Figure A.4). Since the in-sample R^2 value does not dip between $\xi = 0$ and $\xi = 1$, the estimator (at $\xi = 1$) resides in a mode that contains the true coefficient values. The fact that the out-of-sample R^2 is flat between $\xi = 0$ and $\xi = 1$ suggests that any value of ξ in this range will generalize well out-of-sample, that is, any of these estimators will perform well. As $\widehat{\mathbf{A}}$ looks remarkably similar to $\mathbf{A}$ in Figure A.4 ($\widehat{\mathbf{B}}$ also looks similar

to $\mathbf{B}$), taken with the flat out-of-sample likelihood, we can be confident the estimator $\widehat{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}$ are close in the coefficient space.

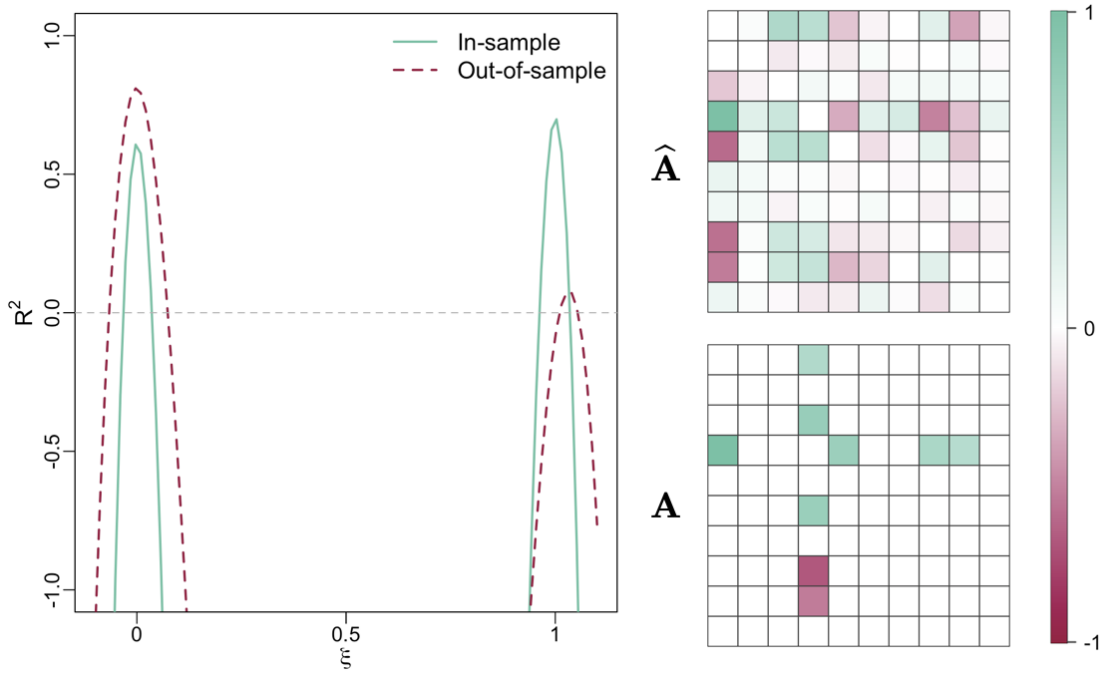


Figure A.3: R^2 values for a mixture of true coefficients and coefficients estimated for the bilinear model (see (A.30)) when generating from the bilinear model with $q = 0.9$ and $T = 10$, for a single fold of the cross-validation. The estimated $\mathbf{A}$ matrix (top right) is scaled by a constant c (and $\mathbf{B}$ by c^{-1}) such that the estimated $\mathbf{A}$ and $\mathbf{B}$ are as close as possible in L_2 norm to the true $\mathbf{A}$ and $\mathbf{B}$ matrices. The pair of matrices $\mathbf{A}$ and $\widehat{\mathbf{A}}$ are normalized to lie in $[-1, 1]$ for sake of plotting.

A.3.3 Convergence study

We conduct a simulation study (to larger dimension T up to 10^5) to examine the theoretical results of Section 2.4.2. In line with Section 2.4.2, we generate from and estimated the bilinear and full BLIN models, using the BLIN representation in (2.9) with $\mathbf{X}_t = \mathbf{Z}_t$ for all $t \in \{1, 2, \ldots, T\}$. For simplicity, we generate every element in $\mathbf{X}_t$ and $\mathbf{E}_t$ with independent and identically distributed (iid) standard normal random variables. We fix $\mathbf{A}$ and $\mathbf{B}$ throughout the study, with $\mathbf{A}^T = \mathbf{U}\mathbf{V}^T$ and $\mathbf{B} = \mathbf{R}\mathbf{S}^T$, where $\{\mathbf{U}, \mathbf{V}\}$ are $S \times S$ matrices and $\{\mathbf{R}, \mathbf{S}\}$ are $L \times L$ matrices, all with iid standard normal entries. Although these $\mathbf{A}$ and $\mathbf{B}$ matrices resemble those in the reduced rank BLIN model, by construction, they are full rank. The

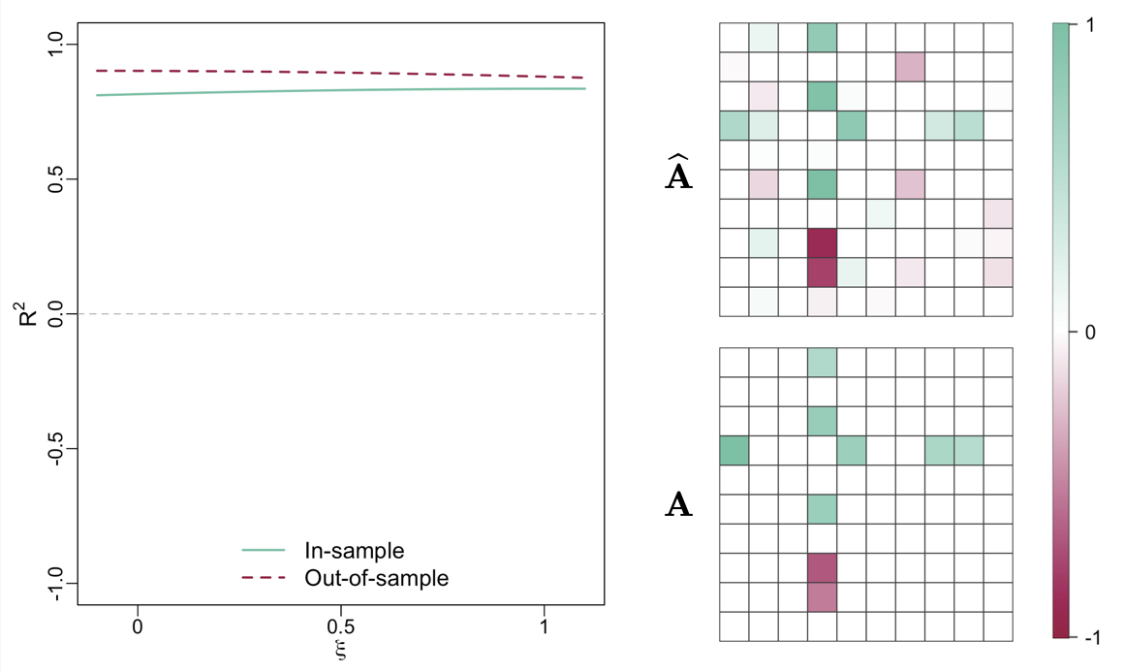


Figure A.4: R^2 values for a mixture of true coefficients and coefficients estimated for the BLIN model (see (A.30)) when generating from the BLIN model with $q = 0.9$ and $T = 10$, for a single fold of the cross-validation. The estimated $\mathbf{A}$ matrix (top right) is scaled by a constant c (and $\mathbf{B}$ by c^{-1}) such that the estimated $\mathbf{A}$ and $\mathbf{B}$ are as close as possible in L_2 norm to the true $\mathbf{A}$ and $\mathbf{B}$ matrices. The pair of matrices $\mathbf{A}$ and $\widehat{\mathbf{A}}$ are normalized to lie in $[-1, 1]$ for sake of plotting.

number of time periods T grows from 10^2 to 10^5 in $10^{1/2}$ multiplicative increments. We choose $S = 10$, $L = 9$, and repeat for 1,000 replications at each combination of T and generating model.

We compare the estimated coefficients to the truth by normalizing the off-diagonal entries such that, if the off-diagonals of the estimate $\widehat{\mathbf{A}} = c\mathbf{A}$ for any $c \in \mathbb{R}$ from any estimation procedure, then we say that the off-diagonal estimates of $\mathbf{A}$ have no error. To do this, we compute the mean squared error between the off-diagonal entries in $\widehat{\mathbf{A}}$ divided by their sum and the off-diagonal entries in $\mathbf{A}$ divided by their sum. We do the same for $\mathbf{B}$. Finally, we calculate the mean squared error of the autoregressive parameters as appropriate for the generating and estimating model. For instance, when generating from the bilinear model but estimating the BLIN model, the (i, j) contribution to the mean squared error is $(\widehat{a}_{ii} + \widehat{b}_{jj} - a_{ii}b_{jj})^2$, and the mean square error of the diagonals is the sum over all i and j. Lastly, we estimate the convergence rate of each estimator to the truth by fitting a simple linear regression model to $\log_{10}(MSE)$ against $\log_{10}(T)$ for each coefficient and combination of generating and estimating models.

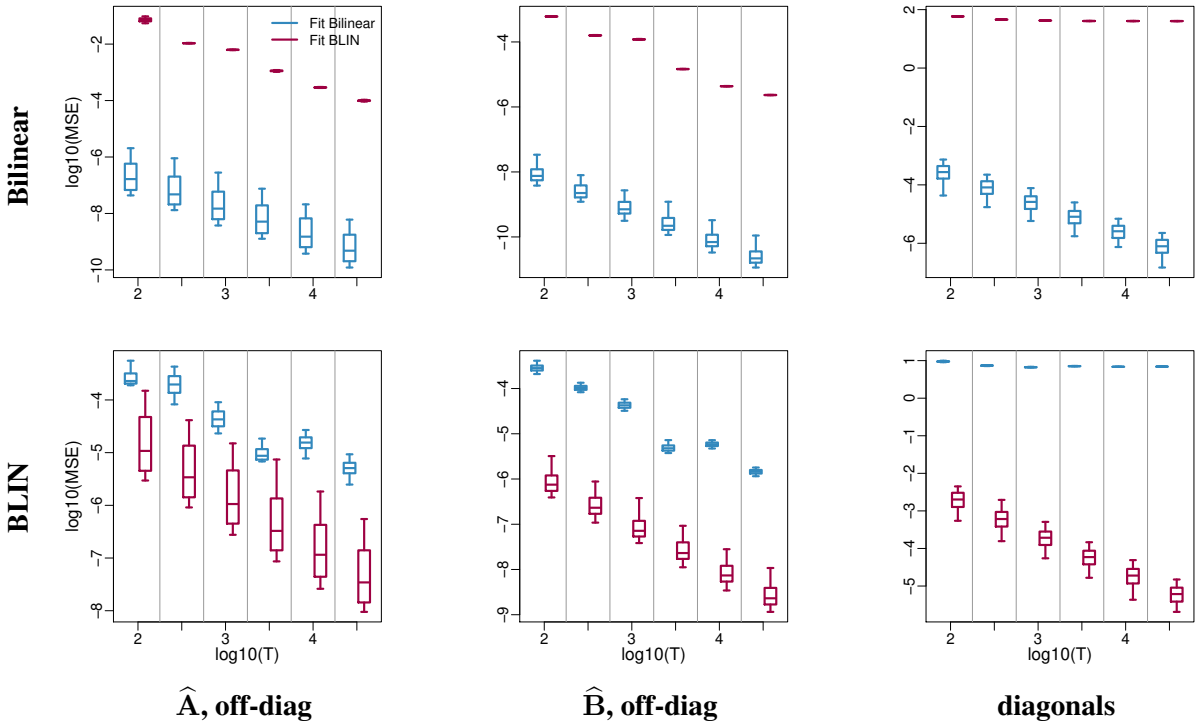


Figure A.5: Mean square error between coefficient estimates and true coefficients used to generate data from the bilinear model (top row) and BLIN model (bottom row). The estimators are the least squares estimators of the BLIN and bilinear models.

In Figure A.5, both the BLIN and bilinear estimators of the off-diagonal estimator of $\mathbf{A}$ and $\mathbf{B}$ have MSEs that tend to zero regardless of the generating model. This suggests that both estimators are consistent under both generating models, confirming Theorem 4. In addition, we observe that neither diagonal estimator appears to converge when the estimation is not of the true model, e.g. the BLIN diagonals do not converge to the true bilinear diagonals. In this simulation, we have that $\mathbf{\Psi} \propto \mathbf{I}_S$ and $\mathbf{\Omega} \propto \mathbf{I}_L$, however, the diagonals of $\mathbf{A}$ and $\mathbf{B}$ are non-constant, verifying that this latter condition is critical in Proposition 5.

In Table A.1, we observe $\sqrt{T}$ convergence for all off-diagonal elements estimated by the BLIN model (when data is generated from either the BLIN or bilinear models), which agrees with the discussion in Section 2.4.2. Although the bilinear model attains the same $\sqrt{T}$ convergence rate when correctly specified, the convergence rate is significantly slower when the true generating model is the BLIN model. This suggests that the least squares estimator of the BLIN model is more efficient when misspecified than that of the bilinear model.

Table A.1: Estimates of decrease in (base 10 logarithm of) mean squared error of least squares estimates of BLIN and bilinear models, under simulated data from each model, when T increases by a factor of 10. Bolded and starred slopes are those which are significantly *not* -1 ($p < 0.05$), where the p-value is computed from a simple linear regression of $\log_{10} MSE$ on $\log_{10} T$.

Estimator	Bilinear	BLIN
BLIN **A**	-1.13	-1.00
BLIN **B**	-1.10	-1.00
BLIN diag.	**-0.05***	-1.00
bilinear **A**	-1.00	**-0.71***
bilinear **B**	-1.01	**-0.92***
bilinear diag.	-1.01	**-0.04***

A.4 Multipartite relational data

In this paper, we primarily discuss the setting where $\mathbf{Y}_t$ is a matrix, equivalently a 2-mode array, with replications over time, such that $\{\mathbf{Y}_t\}_{t=1}^T$ may be considered a tensor, or a 3-mode array. The $S \times L \times T$ array $\mathbf{Y}$ may be constructed by concatenating the T matrices $\{\mathbf{Y}_t\}_{t=1}^T$, each of dimension $S \times L$. That is, $\mathbf{Y} = [\mathbf{Y}_1; \mathbf{Y}_2; \ldots; \mathbf{Y}_T]$, where we let ';' denote this concatenation operation. The BLIN model is easily extendable to model K-mode arrays. A K-mode BLIN model is appropriate when there are more than two actor types in the relational dataset, which may be termed *multipartite* as opposed to bipartite. In the ICEWS data example, we consider interaction type, in addition to source country and target country, as the third mode. In this example, each relation is one of four interaction types from a source state to a target state. Then, each $\mathbf{Y}_t$ is a 3-mode array with entries y_{ijk}^t measuring relation intensity, where i is the source state, j is the target state, k is the relation type, and t is the week of the observation. Please see Section 2.6 for more detail.

We are interested in making inference on influence networks $\{\mathbf{B}_1, ..., \mathbf{B}_K\}$, where each influence network is associated with a specific mode of the array. For the ICEWS data example, our motivation is inference on influence networks of source countries ($\mathbf{B}_1$), target countries ($\mathbf{B}_2$), and interaction types ($\mathbf{B}_3$). As discussed in Section 2.2, the (i, j) component of $\mathbf{B}_k$ reflects the influence of the i^{th} "slice" of $\mathbf{X}_t$ on the j^{th} slice of $\mathbf{Y}_t$, where the slice is along mode k. For example, the (i, j) entry of $\mathbf{B}_3$ estimates the influence of $\mathbf{x}_{\cdot\cdot i}$ on $\mathbf{y}_{\cdot\cdot j}$, when controlling for row and column dependencies between $\mathbf{Y}_t$ and $\mathbf{X}_t$. In the example above, this entry characterizes the influence of interaction type i on interaction type j when controlling for source country and target country influences.

We lean on the Tucker product (Tucker (1964)) and related results to write the array BLIN model for general K-mode arrays (see also De Lathauwer et al., 2000; Kolda and Bader, 2009; Hoff, 2011). First, we rewrite the expectation of the BLIN model of (2.9) using the Tucker product notation as

$$E[\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{X} \times \{\mathbf{A}^T, \mathbf{I}_L, \mathbf{I}_T\} + \mathbf{Z} \times \{\mathbf{I}_S, \mathbf{B}^T, \mathbf{I}_T\}, \tag{A.31}$$

where the $S \times L \times T$ arrays $\mathbf{X}$ and $\mathbf{Z}$ are the results of concatenating $\{\mathbf{X}_t\}_{t=1}^T$ and $\{\mathbf{Z}_t\}_{t=1}^T$, respectively, as $\{\mathbf{Y}_t\}_{t=1}^T$ is concatenated to form $\mathbf{Y}$.

We now extend (A.31) to general K-mode arrays, allowing a different predictor for each mode (i.e. $\mathbf{X}$ and $\mathbf{Z}$ for the 2-mode approach in (A.31)). Let $\mathbf{Y}_t$ be an $m_1 \times \ldots \times m_K$ array observed over $t \in \{1, 2, \ldots, T\}$ time periods, with K corresponding predictor arrays $\{\mathbf{X}_t^{(k)}\}_{k=1}^K$ each of dimension $m_1 \times \ldots \times m_K$ as well. To form the full model, we concatenate the arrays, as above in (A.31), such that $\mathbf{Y}$ is of dimension $m_1 \times m_2 \times \ldots \times m_K \times T$, as is every one of the K predictor arrays $\{\mathbf{X}^{(k)}\}_{k=1}^K$. Then, the expectation of the multipartite extension of the BLIN model is as follows:

$$E\left[\mathbf{Y} \mid \{\mathbf{X}^{(k)}\}_{k=1}^K\right] = \sum_{k=1}^K \mathbf{X}^{(k)} \times \{\mathbf{I}_{m_1}, \mathbf{I}_{m_2}, \ldots, \mathbf{I}_{m_{k-1}}, \mathbf{B}_k^T, \mathbf{I}_{m_{k+1}}, \ldots, \mathbf{I}_{m_K}, \mathbf{I}_R\} \tag{A.32}$$

$$E\left[\mathbf{y} \mid \{\mathbf{x}^{(k)}\}_{k=1}^K\right] = \sum_{k=1}^K \left(\mathbf{I}_{m_{k+1}m_{k+2}\ldots m_K T} \otimes \mathbf{B}_k \otimes \mathbf{I}_{m_1 m_2 \ldots m_{k-1}}\right)\mathbf{x}^{(k)}, \tag{A.33}$$

where each $\mathbf{B}_k$ is an $m_k \times m_k$ matrix of coefficients and $\mathbf{y}$ and $\mathbf{x}^{(k)}$ are the vectorizations of $\mathbf{Y}$ and $\mathbf{X}^{(k)}$, respectively. As the array form the BLIN model in (A.32) is of course linear, an appropriate design matrix as in (2.4) may be constructed to estimate the sparse array BLIN model.

We now provide array manipulations such that a block coordinate descent, similar to that in Algorithm 1, may be derived to iteratively estimate $\{\mathbf{B}_k\}_{k=1}^K$ in the array BLIN model. First, let $\mathbf{M}$ be any array of dimensions $m_1 \times m_2 \times m_3$. The mode-1 matricization of the array $\mathbf{M}_{(1)}$ is defined as an $m_1 \times m_2 m_3$ matrix and if $\mathbf{M} = \mathbf{X} \times \{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3\}$, then $\mathbf{M}_{(1)} = \mathbf{C}_1 \mathbf{X}_{(1)} (\mathbf{C}_3 \otimes \mathbf{C}_2)^T$. The following mode-1 and mode-2 matricizations of the expectation of the matrix BLIN model are then

$$E[\mathbf{Y}_{(1)} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{A}^T \mathbf{X}_{(1)} + \mathbf{Z}_{(1)}\big(\mathbf{I}_T \otimes \mathbf{B}\big), \tag{A.34}$$

$$E[\mathbf{Y}_{(2)} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{X}_{(2)}\big(\mathbf{I}_T \otimes \mathbf{A}\big) + \mathbf{B}^T \mathbf{Z}_{(2)}. \tag{A.35}$$

The mode-i matricization of the expectation of the K-mode BLIN model in (A.32) may be written as

$$E\left[\mathbf{Y}_{(i)} \mid \{\mathbf{X}^{(k)}\}_{k=1}^{K}\right] = \mathbf{B}_i^T \mathbf{X}_{(i)}^{(i)} + \sum_{k \neq i} \mathbf{X}_{(i)}^{(k)} \left(\mathbf{I}_{n_1^k} \otimes \mathbf{B}_k \otimes \mathbf{I}_{n_2^k}\right), \tag{A.36}$$

$$\text{where } n_1^k = T \prod_{\substack{j \geq k+1 \\ j \neq i}} m_j \text{ and } n_2^k = \prod_{\substack{j \leq k-1 \\ j \neq i}} m_j.$$

Representation the model in this form enables straightforward estimation of $\mathbf{B}_i$ given $\{\mathbf{B}_k\}_{k \neq i}$.

A.5 Details of data analysis

In this section, we provide supporting materials for performing the analysis of the ICEWS data set. In the notation of Appendix A.4, the BLIN model for this analysis may be written:

$$\begin{aligned} \mathbf{Y}_t = &\left(\sum_{k=1}^{p_A} \mathbf{Y}_{t-k}\right) \times \{\mathbf{A}^T, \mathbf{I}_S, \mathbf{I}_R\} + \left(\sum_{k=1}^{p_B} \mathbf{Y}_{t-k}\right) \times \{\mathbf{I}_S, \mathbf{B}^T, \mathbf{I}_R\} \\ &+ \left(\sum_{k=1}^{p_C} \mathbf{Y}_{t-k}\right) \times \{\mathbf{I}_S, \mathbf{I}_S, \mathbf{C}^T\} + \mathbf{E}_t, \end{aligned} \tag{A.37}$$

where $\mathbf{Y}_t$ is a $25 \times 25 \times 4$ array representing senders, receivers, and interaction types. We also examined replacing $\mathbf{Y}_t$ with $\mathbf{D}_t := \mathbf{Y}_t - \mathbf{Y}_{t-1}$, the week-over-week increase in interaction value. To evaluate the performance of the model to represent the undifferenced and differenced data, we estimated the sparse BLIN model on each of differenced and undifferenced data for all combinations of lags $\{p_A, p_B, p_C\}$ between 1 and 5 (inclusive).

We found that the BLIN model represented the differenced data significantly better, that is, with in-sample R^2 values above 0.30 for differenced data rather than those below 0.10 for undifferenced data. Thus, we proceeded with the differenced data only. To select the model that best balanced model fit and model parsimony, we used an analog of Akaike's Information Criterion (Akaike, 1998):

$$\widehat{AIC} = 2\left(||\widehat{\mathbf{A}}||_1 + ||\widehat{\mathbf{B}}||_1 + ||\widehat{\mathbf{C}}||_1\right) + S^2 RT \log \sum_{t=1}^{T} ||\widehat{\mathbf{D}}_t - \mathbf{D}_t||_2^2, \tag{A.38}$$

where $||\mathbf{H}||_1$ is the number of nonzero entries in $\mathbf{H}$, e.g., and $\widehat{\mathbf{D}}_t$ is the estimated mean corresponding to the estimated networks $\{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}\}$. In (A.38), the first term is a penalty for the number of parameters, i.e.

model complexity, and the second term quantifies the goodness of fit. Smaller values of $\widehat{AIC}$ are better; the five smallest values are given in Table A.2.

Table A.2: The five smallest information criteria ($\widehat{AIC}$) for selected lags of sparse BLIN estimates of differenced data.

p_A	p_B	p_C	$\widehat{AIC}$	R^2	Number of parameters
5	3	1	58307	0.308	563
3	5	1	58460	0.308	613
5	1	3	58738	0.307	628
1	5	3	58860	0.307	463
3	1	5	59074	0.307	617

To select the network entries to be highlighted in Figure 2.4, we plotted the ordered nonzero coefficients (Figure A.6). Then, we visually cut off the coefficient values at natural breaks in the nonzero entries. The visual cutoffs are near the 2.5% and 97.5% quantiles of the nonzero coefficient values.

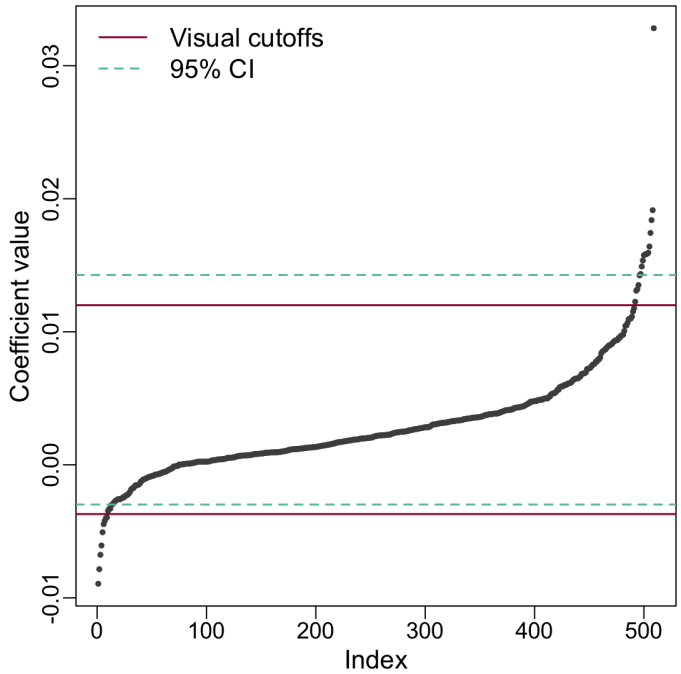


Figure A.6: Estimated influence network values and cutoffs for relations highlighted in Figure 2.4.

Appendix B

Regression of relational data with exchangeable errors

B.1 Undirected arrays

This section specializes the results presented in the manuscript to undirected relational data. Consider the case when $R = 1$ and suppose the relational data contains the relations among n actors. The covariance of the errors Ω contains three unique elements

$$\text{Cov}(\xi_{ij}, \xi_{ij}) := \theta, \quad \text{Cov}(\xi_{ij}, \xi_{ki}) := \phi, \quad \text{Cov}(\xi_{ij}, \xi_{kl}) := 0.$$

As in the directed case, we assume the last covariance, corresponding to two relations which share no common member, is zero. We again estimate the two remaining terms using the residual matrix $\mathbf{E} = \{e_{ij}\} \in \mathbb{R}^{n \times n}$. Note that the residual matrix we consider is for the entire $n \times n$ matrix of relations and thus contains duplicate off-diagonal entries corresponding to pairs $\{(i,j),(j,i)\}$. We set the diagonal of $\mathbf{E}$ to zero as the relation between an actor and itself is undefined.

The estimate of θ is the empirical mean of each squared residual and can be expressed

$$\widehat{\theta} = \frac{tr(\mathbf{EE})}{n(n-1)}$$

where $tr(\cdot)$ denotes the matrix trace operator.

Similarly, the estimate of ϕ is

$$\begin{aligned}\widehat{\phi} &= \frac{1}{2m} \sum_i \sum_{j \neq i} e_{ij} \left(\sum_{k \neq i} e_{ik} + \sum_{k \neq j} e_{kj} - 2e_{ij} \right) \\ &= \frac{1}{m} \mathbf{1}^T (\mathbf{EE}) \, \mathbf{1} - tr(\mathbf{EE}) \qquad \text{where } \; m = n(n-1)(n-2).\end{aligned}$$

B.2 Proof of asymptotic normality of OLS

For this proof, and throughout Appendix B, we adopt slightly different notation to simplify the representation of the exchangeable covariance estimator. Recall that the exchangeable covariance estimator for the OLS estimating equations is

$$\widehat{V}_E = (X^T X)^{-1} X^T \widehat{\Omega}_E X (X^T X)^{-1},$$

where $\widehat{\Omega}_E$ is the exchangeable estimate of the error covariance matrix, consisting of five averages of residual products. Here we express $\widehat{\Omega}_E$ as

$$\widehat{\Omega}_E = \sum_{i=1}^{5} \widehat{\phi}_i \mathcal{S}_i, \quad \text{where } \widehat{\phi}_i = \frac{\sum_{(jk,\ell m)\in\Theta_i} e_{jk} e_{\ell m}}{|\Theta_i|} \quad \text{for } i \in \{1,2,3,4,5\}. \tag{B.1}$$

This amounts to mapping $\sigma^2 \mapsto \phi_1$, $\phi_a \mapsto \phi_2$, ..., $\phi_d \mapsto \phi_5$, and re-indexing the $\mathcal{S}$ and M matrices accordingly. Additionally, when we consider sequences of jointly exchangeable random variables $\{W_{ij}\}_{i,j=1}^n$, it is understood that the sequence arises from a relational array such that entries with $i = j$ are undefined. Thus, sums over the sequence are of $n(n-1)$ terms and we define $\sum_{ij} W_{ij} = \sum_{i\neq j} W_{ij}$.

We work in the asymptotic regime where actors are added incrementally to the relational data set, i.e. n is continually increasing. To establish asymptotic normality of $\widehat{\boldsymbol{\beta}}$, we wish to show

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, M_1^{-1}\big(\phi_3 M_3 + \phi_4 M_4 + 2\phi_5 M_5\big) M_1^{-1}\big), \tag{B.2}$$

where $\{M_i\}_{i\in\{1,3,4,5\}}$ are as in (3.8) and '$\to_d$' denotes element-wise convergence in distribution.

The motivation for the proof argument follows from the expression

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{\sum_{jk} \mathbf{x}_{jk}\mathbf{x}_{jk}^T}{n(n-1)}\right)^{-1} \frac{\sqrt{n}\sum_{jk} \mathbf{x}_{jk}\xi_{jk}}{n(n-1)}. \tag{B.3}$$

We note that $\left(\frac{\sum_{jk} \mathbf{x}_{jk}\mathbf{x}_{jk}}{n(n-1)}\right)^{-1}$ converges in probability to M_1^{-1} and then show asymptotic normality of the second multiplicative term in (B.3).

To analyze $\{\mathbf{x}_{ij}\xi_{ij}\}_{i,j=1}^n$, we note that, by condition (B1), the joint exchangeability and independence of non-overlapping pairs of the sequence $\{\xi_{ij}\}_{i,j=1}^n$ extends to the component sequences in the vectors

$\{\mathbf{x}_{ij}\xi_{ij}\}_{i,j=1}^{n}$. Thus, to prove asymptotic normality of $\widehat{\boldsymbol{\beta}}$, we first prove a theorem stating that the average of a mean-zero sequence of jointly exchangeable random variables is asymptotically normal. Specifically, for $\{W_{ij}\}_{i,j=1}^{n}$ mean zero and jointly exchangeable, we show

$$k_n \frac{\sum_{ij} W_{ij}}{\sigma} \to_d \mathrm{N}(0,1) \tag{B.4}$$

for some normalizing constant σ and fixed sequence $k_n \to 0$ as $n \to \infty$.

To prove (B.4), we rely on a result from Bolthausen (1982), as well as a supporting lemma which we present here. Below we outline the significance of these results in the proof.

- **Lemma 13** (Bolthausen (1982)): Provides a sufficient condition for asymptotic normality of a sequence of measures based on the standard normal characteristic function.
- **Lemma 14:** Provides a bound for a variance that surfaces in the proof of asymptotic normality in (B.4).

From (B.4), we immediately have the marginal asymptotic normality of the sample mean of each of the vector components in the sequence $\{\mathbf{x}_{ij}\xi_{ij}\}_{i,j=1}^{n}$. To establish joint asymptotic normality, we employ the Cramér-Wold device Cramér and Wold (1936), where asymptotic normality of $\{\mathbf{v}^T\mathbf{x}_{ij}\xi_{ij}\}_{i,j=1}^{n}$, for all $\mathbf{v} \in \mathbb{R}^p$ with $||\mathbf{v}|| = 1$, establishes joint normality. To achieve the asymptotic normality of this inner product, we simply recognize that this inner product is itself the mean of an exchangeable sequence of random variables. Joint asymptotic normality of the mean of the sequence of vectors $\{\mathbf{x}_{ij}\xi_{ij}\}_{i,j=1}^{n}$ establishes joint asymptotic normality of $\widehat{\boldsymbol{\beta}}$ via (B.3).

B.2.1 Lemmas and theorem in support of Theorem 9

The following is Lemma 2 in Bolthausen (1982) and provides a sufficient condition for asymptotic normality. We abuse notation slightly, letting i be the imaginary unit where appropriate.

Lemma 13 (Bolthausen (1982))**.** *Let ν_n be a sequence of probability distributions over $\mathbb{R}$ which satisfies*

1. $\sup_n \int x^2 d\nu_n(x) < \infty$*, and*
2. *for all* $\lambda \in \mathbb{R}$, $\lim_n \int (i\lambda - x)\mathrm{e}^{i\lambda x} d\nu_n(x) = 0$.

Then, $\nu_n \to_d \mathrm{N}(0,1)$.

To provide intuition for Lemma 13, the integral in condition (2) is identically zero when ν_n is the standard normal distribution.

The next lemma provides a sufficient condition on the dependence structure in $\{W_{ij}\}_{i,j=1}^n$ necessary for the proof of asymptotic normality in (B.4). Again we emphasize that terms in $\{W_{ij}\}_{i,j=1}^n$ with $i = j$ are undefined.

Lemma 14. *Let* $\{W_{ij}\}_{i,j=1}^n$ *be a sequence of jointly exchangeable random variables as in Definition 8 with* $||W_{ij}||_4 < L < \infty$*, where* $||W_{ij}||_p := \mathbb{E}\left[|W_{ij}|^p\right]^{1/p}$ *for* $p > 0$*. Then,*

$$\frac{1}{n^6} V\left[\sum_{ij}\sum_{kl\in\Theta_{ij}} W_{ij}W_{kl}\right] < \frac{CL^4}{n} \to 0 \text{ as } n \to \infty, \tag{B.5}$$

for some $C < \infty$*, where* Θ_{ij} *is the set of ordered pairs* (k,l) *that share an index with* (i,j)*.*

Proof. By definition we write

$$\frac{1}{n^6} V\left[\sum_{ij}\sum_{kl\in\Theta_{ij}} W_{ij}W_{kl}\right] = \frac{1}{n^6}\sum_{ij}\sum_{kl\in\Theta_{ij}}\sum_{rs}\sum_{tu\in\Theta_{rs}} \mathrm{Cov}(W_{ij}W_{kl}, W_{rs}W_{tu}). \tag{B.6}$$

Each covariance of (B.6) is bounded by L^4. To bound the variance, we will show the number of nonzero entries in the sum is $\mathcal{O}(n^5)$. For $\mathrm{Cov}(W_{ij}W_{kl}, W_{rs}W_{tu}) \neq 0$, there must be overlap between the index sets $\{i,j,k,l\}$ and $\{r,s,t,u\}$. Further, the sum in (B.6) is taken over index sets that themselves contain overlap, i.e. $\{i,j\} \cap \{k,l\} \neq \varnothing$ and $\{r,s\} \cap \{t,u\} \neq \varnothing$. For example, the index sets $\{i,j,i,l\}$ and $\{i,s,i,u\}$ have nonzero covariance in (B.6). Since there are 5 unique indices in the union of the sets $\{i,j,i,l\}$ and $\{i,s,i,u\}$, there are $\mathcal{O}(n^5)$ such index set pairs of this form in total. There are 96 pairs of index sets that result in nonzero covariance $\mathrm{Cov}(W_{ij}W_{kl}, W_{rs}W_{tu})$. For example, another such pair of index sets is $\{i,j,i,l\}$ and $\{i,j,i,j\}$. Each of these 96 pairs of index sets is $\mathcal{O}(n^5)$. Thus, the sum of covariances in (B.6) is over $\mathcal{O}(n^5)$ bounded elements. □

It is worth noting that we repeat the counting argument in the proof of Lemma 14 in many of the following proofs, including those in later sections. Now that we have Lemma 13 and 14, we prove that a general sequence of mean-zero exchangeable random variables is asymptotically normal.

Theorem 15. *Let $\{W_{ij}\}_{i,j=1}^{n}$ be a mean-zero sequence of jointly exchangeable random variables with at least one of $\{\phi_3, \phi_4, \phi_5\}$ nonzero. If $||W_{ij}||_4 < L < \infty$, then*

$$\frac{\sqrt{n}\sum_{ij} W_{ij}}{n(n-1)} \rightarrow_d \mathrm{N}(0, \phi_3 + \phi_4 + 2\phi_5) \text{ as } n \rightarrow \infty. \tag{B.7}$$

Proof. We first show that $\phi_3 + \phi_4 + 2\phi_5$ is the correct limiting variance. Writing the variance of the expression on the left hand side of (B.7) explicitly and recalling that entries such that $i = j$ are undefined, we see

$$\begin{aligned} V\left[\frac{\sqrt{n}}{n(n-1)}\sum_{ij} W_{ij}\right] &= \frac{n}{n^2(n-1)^2}\sum_{ij}\sum_{kl\in\Theta_{ij}} \mathrm{Cov}\left(W_{ij}, W_{kl}\right) \\ &= \frac{n^2(n-1)\left(\phi_1+\phi_2\right) + n^2(n-1)(n-2)\left(\phi_3+\phi_4+2\phi_5\right)}{n^2(n-1)^2} \\ &\rightarrow \phi_3 + \phi_4 + 2\phi_5 \text{ as } n \rightarrow \infty, \end{aligned} \tag{B.8}$$

by the properties of joint exchangeability of $\{W_{ij}\}_{i,j=1}^{n}$ as described in Section 3.3.3. This variance is finite and nonzero by assumption. To prove (B.7), it is sufficient to show

$$\bar{S}_n := \frac{\sum_{ij} W_{ij}}{n^{3/2}\sqrt{\phi_3 + \phi_4 + 2\phi_5}} \rightarrow_d N(0,1). \tag{B.9}$$

Define the limiting variance as $\sigma_n^2 = n^3(\phi_3 + \phi_4 + 2\phi_5)$ and the sum $S_n = \sum_{ij} W_{ij}$.

To establish (B.9), we employ Lemma 13, where ν_n is the probability measure corresponding to $\bar{S}_n$ for all n. The first condition of Lemma 13 is satisfied since

$$\mathbb{E}[(\bar{S}_n)^2] = \frac{V\left[\sum_{ij} W_{ij}\right]}{n^3(\phi_3 + \phi_4 + 2\phi_5)} < CL^2 \tag{B.10}$$

for $C < \infty$ and all n. Thus, to prove (B.9), it is sufficient to show the second condition of Lemma 13: for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[(i\lambda - \bar{S}_n)\mathrm{e}^{i\lambda\bar{S}_n}\right] \rightarrow 0 \text{ as } n \rightarrow \infty. \tag{B.11}$$

We decompose the term in the expectation as in Guyon (1995) and Lumley and Hamblett (2003):

$$(i\lambda - \bar{S}_n)\mathrm{e}^{i\lambda\bar{S}_n} = A_1 - A_2 - A_3, \tag{B.12}$$

$$\begin{aligned}
\text{where } A_1 &= i\lambda\mathrm{e}^{i\lambda\bar{S}_n}\left(1 - \frac{1}{\sigma_n^2}\sum_{ij} W_{ij}S_{ij,n}\right), \\
A_2 &= \frac{\mathrm{e}^{i\lambda\bar{S}_n}}{\sigma_n}\sum_{ij} W_{ij}\left(1 - i\lambda\bar{S}_{ij,n} - \mathrm{e}^{-i\lambda\bar{S}_{ij,n}}\right), \\
A_3 &= \frac{1}{\sigma_n}\sum_{ij} W_{ij}\mathrm{e}^{i\lambda(\bar{S}_n - \bar{S}_{ij,n})}, \\
S_{ij,n} &= \sum_{kl\in\Theta_{ij}} W_{kl}, \text{ and } \bar{S}_{ij,n} = S_{ij,n}/\sigma_n.
\end{aligned}$$

To satisfy (B.11) it remains to be shown that $\lim_{n\to\infty}\mathbb{E}[A_m] = 0$ for each $m \in \{1,2,3\}$.

$\mathbf{A_1}$: First notice that $|\mathrm{e}^{i\lambda\bar{S}_n}| \leq 1$. Using this fact and Lemma 14,

$$0 \leq \mathbb{E}[|A_1|]^2 \leq \mathbb{E}[|A_1^2|] \leq \lambda^2\mathbb{E}\left[\left|1 - \frac{1}{\sigma_n^2}\sum_{ij} W_{ij}S_{ij,n}\right|^2\right] \tag{B.13}$$

$$= \frac{\lambda^2}{\sigma_n^4}V\left[\sum_{ij}\sum_{kl\in\Theta_{ij}} W_{ij}W_{kl}\right] + \lambda^2\left(1 - \frac{V\left[\sum_{ij} W_{ij}\right]}{\sigma_n^2}\right)^2 \tag{B.14}$$

$$\leq \lambda^2\frac{CL^4}{n} + \lambda^2\left(1 - \frac{\sigma_n^2 + \mathcal{O}(n^{-1})}{\sigma_n^2}\right)^2 \tag{B.15}$$

$$= \lambda^2\left(\frac{CL^4}{n} + \frac{\mathcal{O}\left(n^{-2}\right)}{\sigma_n^2}\right) \to 0 \tag{B.16}$$

for all real λ. $\mathbb{E}[|A_1|]^2$ limiting to zero implies $\mathbb{E}[|A_1|]$ limits to zero, and hence $\mathbb{E}[A_1]$ limits to zero.

$\mathbf{A_2}$: By Taylor expansion of $\mathrm{e}^{-i\lambda\bar{S}_{ij,n}}$, we can write

$$\left|1 - i\lambda\bar{S}_{ij,n} - \mathrm{e}^{-i\lambda\bar{S}_{ij,n}}\right| \leq c\lambda^2\left(\bar{S}_{ij,n}\right)^2, \tag{B.17}$$

for some $0 < c < \infty$ and all n, λ. Using this bound and the fact that $|\Theta_{ij}| = 4n - 6$, we evaluate $\mathbb{E}[|A_2|]$ directly below:

$$\mathbb{E}[|A_2|] \leq \frac{1}{\sigma_n}\mathbb{E}\left[\sum_{ij}|W_{ij}|\left|1-i\lambda\bar{S}_{ij,n}-\mathrm{e}^{-i\lambda\bar{S}_{ij,n}}\right|\right], \tag{B.18}$$

$$\leq \frac{c\lambda^2}{\sigma_n^3}\sum_{ij}\mathbb{E}\left[|W_{ij}|\left(S_{ij,n}\right)^2\right], \tag{B.19}$$

$$\leq \frac{c\lambda^2}{\sigma_n^3}n(n-1)(4n-6)^2L^3 \to 0, \tag{B.20}$$

for all real λ. As $\mathbb{E}[|A_2|]$ limits to zero, so does $\mathbb{E}[A_2]$.

$\mathbf{A_3}$: Note that $S_{ij,n}$ sums all terms in the sequence $\{W_{ij}\}_{i,j=1}^n$ that depend upon W_{ij}, including W_{ij} itself. Thus, W_{ij} and $\bar{S}_n - \bar{S}_{ij,n}$ are independent. It follows immediately that

$$\mathbb{E}\left[\frac{1}{\sigma_n}\sum_{ij}W_{ij}\mathrm{e}^{i\lambda(\bar{S}_n-\bar{S}_{ij,n})}\right] = \frac{1}{\sigma_n}\sum_{ij}\mathbb{E}\big[W_{ij}\big]\mathbb{E}\big[\mathrm{e}^{i\lambda(\bar{S}_n-\bar{S}_{ij,n})}\big] = 0, \tag{B.21}$$

since $\mathbb{E}\left[W_{ij}\right] = 0$ for all ordered pairs (i,j).

Hence, $\lim_{n\to\infty}\mathbb{E}[A_m] = 0$ for each $m \in \{1,2,3\}$ and we have the convergence in (B.11), implying $\bar{S}_n \to_d \mathrm{N}(0,1)$ by Lemma 13, which gives the desired result in (B.7). □

B.2.2 Proof of Theorem 9

We begin by writing

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \left(\frac{\sum_{jk}\mathbf{x}_{jk}\mathbf{x}_{jk}^T}{n(n-1)}\right)^{-1}\frac{\sqrt{n}\sum_{jk}\mathbf{x}_{jk}\xi_{jk}}{n(n-1)}, \tag{B.22}$$

again emphasizing that entries in the sum with $j = k$ are undefined and omitted. Addressing the first multiplicative term in (B.22), we recall that the inverse map is continuous. Then, by (3.8) and the continuous mapping theorem, we have

$$\left(\frac{\sum_{jk}\mathbf{x}_{jk}\mathbf{x}_{jk}^T}{n(n-1)}\right)^{-1} \to_p M_1^{-1}. \tag{B.23}$$

We now analyze the second multiplicative term in (B.22). Showing asymptotic normality of this second multiplicative term is sufficient to show asymptotic normality of the expression on the left hand side of (B.22). We wish to show that the sum of vectors

$$\mathbf{U}_n := \frac{\sqrt{n}}{n(n-1)} \sum_{jk} \mathbf{x}_{jk}\xi_{jk} \to_d \mathrm{N}(0, \Sigma), \tag{B.24}$$

for some limiting variance Σ, where we recall the definition $\mathbf{x}_{jk}^T = [x_{jk}^{(1)}, x_{jk}^{(2)}, \ldots, x_{jk}^{(p)}]$. By the Cramér-Wold device (Cramér and Wold (1936)), $\mathbf{U}_n$ is asymptotically normal with asymptotic variance Σ if and only if $\mathbf{v}^T\mathbf{U}_n$ is asymptotically normal with asymptotic variance $\mathbf{v}^T\Sigma\mathbf{v}$ for every vector $\mathbf{v} \in \mathbb{R}^p$ such that $||\mathbf{v}|| = 1$. Clearly,

$$\mathbf{v}^T\mathbf{U}_n := \frac{\sqrt{n}}{n(n-1)} \sum_{jk} \widetilde{x}_{jk}\xi_{jk}, \tag{B.25}$$

where we define $\widetilde{x}_{jk} = \mathbf{v}^T\mathbf{x}_{jk}$. We wish to apply Theorem 15 to the sequence $\{\widetilde{x}_{jk}\xi_{jk}\}_{j,k=1}^n$. First, the condition of finite moments in (B2) of Theorem 9 and $||\mathbf{v}|| = 1$ implies that $||\widetilde{x}_{jk}\xi_{jk}||_4 < L$ for some finite $L < \infty$. Secondly, by the independence of X and Ξ in (B3) of Theorem 9, the sequence $\{\widetilde{x}_{jk}\xi_{jk}\}_{j,k=1}^n$ is a mean-zero exchangeable sequence of scalar random variables. Taking the variance directly, we have that the variance $V\left[\sum_{jk} \widetilde{x}_{jk}\xi_{jk}\right]$ is

$$n^3\mathbf{v}^T(\phi_1 M_1 \mathcal{O}(n^{-1}) + \phi_2 M_2 \mathcal{O}(n^{-1}) + \phi_3 M_3 + \phi_4 M_4 + 2\phi_5 M_5)\mathbf{v}. \tag{B.26}$$

Then, we apply Theorem 15 with $\sigma_n^2 = V[\sum_{jk} \widetilde{x}_{jk}\xi_{jk}]$ from (B.26), which gives that

$$\mathbf{v}^T\mathbf{U}_n \to_d \mathrm{N}(0, \mathbf{v}^T(\phi_3 M_3 + \phi_4 M_4 + 2\phi_5 M_5)\mathbf{v}). \tag{B.27}$$

Thus, by the Cramér-Wold device, we get the desired joint asymptotic normality

$$\frac{\sqrt{n}\sum_{jk} \mathbf{x}_{jk}\xi_{jk}}{n(n-1)} \to_d \mathrm{N}\big(0, \phi_3 M_3 + \phi_4 M_4 + 2\phi_5 M_5\big). \tag{B.28}$$

Combining the convergence in probability in (B.23) and the asymptotic normality of (B.28), we obtain the desired result. □

B.3 Proof of consistency of the exchangeable estimator

For the proof of the consistency of the exchangeable estimator $\widehat{V}_E$, we adopt the same change in notation as in Appendix B.2, defined in (B.1). We deviate slightly in that we denote Θ_i to denote dyadic pairs (j,k) and (l,m) that share a member in the i^{th} manner. For example, for $i=3$ we must have $j=l$ and $m \neq k$. We use the same assumptions as in Theorem 9.

This proof is outlined as follows. We initially prove that the exchangeable estimator $\widehat{V}_E$ is consistent if the exchangeable parameter estimates $\{\widehat{\phi}_i : i = 1,\ldots,5\}$ are consistent for the true parameters. We then prove consistency of $\{\widehat{\phi}_i\}$ in two steps: (a) we show parameter estimates $\{\widetilde{\phi}_i\}$ based on the unobserved true errors Ξ are consistent and then (b) we show that the parameter estimates $\{\widehat{\phi}_i\}$ are asymptotically equivalent to $\{\widetilde{\phi}_i\}$. We require the consistency of $\widehat{\boldsymbol{\beta}}$ result (implied by Theorem 9) for this last step.

B.3.1 Proof of Theorem 10

We first note that from Theorem 9 the order of convergence of $\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}$. Thus, we choose the rate n as our asymptotic regime for consistency of $\widehat{V}_E$. We wish to show that

$$n\widehat{V}_E - nV[\widehat{\boldsymbol{\beta}}] \to_p 0. \tag{B.29}$$

1. **Sufficient to show consistency of** $\{\widehat{\phi}_i\}$

Here we show that to prove consistency of $\widehat{V}_E$, it is sufficient to prove the consistency of the parameter estimates $\{\widehat{\phi}_i\}$ for the true parameters. We begin by writing the difference of variances $n\widehat{V}_E - nV[\widehat{\boldsymbol{\beta}}]$ in (B.29), as

$$\begin{aligned}
&n(X^TX)^{-1}X^T\big(\widehat{\Omega}_E - \Omega_E\big)X(X^TX)^{-1} \\
&= \frac{n}{n^2(n-1)^2}\left(\frac{X^TX}{n(n-1)}\right)^{-1}\left(\frac{X^T\sum_{i=1}^5|\Theta_i|\left(\widehat{\phi}_i-\phi_i\right)\mathcal{S}_iX}{|\Theta_i|}\right)\left(\frac{X^TX}{n(n-1)}\right)^{-1} \\
&= \sum_{i=1}^5\frac{|\Theta_i|}{n(n-1)^2}\left(\widehat{\phi}_i-\phi_i\right)\left(\frac{X^TX}{n(n-1)}\right)^{-1}\left(\frac{\sum_{(jk,\ell m)\in\Theta_i}\mathbf{x}_{jk}\mathbf{x}_{\ell m}^T}{|\Theta_i|}\right)\left(\frac{X^TX}{n(n-1)}\right)^{-1} \\
&:= \sum_{i=1}^5 c_{i,n}\left(\widehat{\phi}_i-\phi_i\right)h_{i,n}(X),
\end{aligned} \tag{B.30}$$

where $c_{i,n} = |\Theta_i|/n(n-1)^2$ and $h_{i,n}(X)$ contains the remaining terms which are functions of X. By the counting argument used to show Lemma 14, each $|\Theta_i|$ is at most $\mathcal{O}(n^3)$, so each $c_{i,n} \to d_i$ for some finite constant d_i, namely $d_i = 0$ for $i \in \{1, 2\}$, $d_i = 1$ when $i \in \{3, 4\}$, and $d_i = 2$ for $i = 5$. To obtain the result in (B.29), it is sufficient then to show $\widehat{\phi}_i - \phi_i \to_p 0$ and $h_{i,n}(X)$ converges in probability to some constant for all i. The latter comes easily, that is, by assumption and Slutsky's theorem,

$$h_{i,n}(X) \to_p M_1^{-1} M_i M_1^{-1}, \;\; i \in \{1, \ldots, 5\}. \tag{B.31}$$

The continuous mapping theorem allows us to take the probability limit of $\frac{X^T X}{n(n-1)}$ before inversion, as the inversion map is continuous.

We now consider consistency of the parameter estimates $\widehat{\phi}_i$. First, define error averages $\{\widetilde{\phi}_i : i = 1, 2, 3, 4, 5\}$ analogous to the parameter estimates, such that for each i

$$\widetilde{\phi}_i = \frac{1}{|\Theta_i|} \sum_{(jk,\ell m) \in \Theta_i} \xi_{jk} \xi_{\ell m}. \tag{B.32}$$

We will show $\widetilde{\phi}_i - \phi_i$ converges in probability to zero, and then do the same for $\widehat{\phi}_i - \widetilde{\phi}_i$. This is sufficient for showing $\widehat{\phi}_i - \phi_i \to_p 0$ as $\widehat{\phi}_i - \phi_i = (\widehat{\phi}_i - \widetilde{\phi}_i) + (\widetilde{\phi}_i - \phi_i)$.

2. **Consistency of $\widetilde{\phi}_i$ for ϕ_i**

To show convergence in probability of $\widetilde{\phi}_i - \phi_i$ to zero, we use the argument that the bias and variance both tend to zero. By assumption (A1), $\mathbb{E}[\xi_{jk}\xi_{\ell m}] = \phi_i$ for every relation pair $(jk, \ell m) \in \Theta_i$. Thus, $\mathbb{E}[\widetilde{\phi}_i - \phi_i] = 0$ for all n and $i \in \{1, \ldots, 5\}$. We now turn to the variance:

$$V\left[\widetilde{\phi}_i\right] = \frac{1}{|\Theta_i|^2} \sum_{(jk,\ell m) \in \Theta_i} \sum_{(rs,tu) \in \Theta_i} \mathrm{Cov}\big(\xi_{jk}\xi_{\ell m}, \xi_{rs}\xi_{tu}\big). \tag{B.33}$$

We again make a counting argument similar to that in Lemma (14). By condition (B2), each of the $|\Theta_i|^2$ covariances in the sum above are bounded. The covariance between $\xi_{jk}\xi_{\ell m}$ and $\xi_{rs}\xi_{tu}$ is nonzero only if there is overlap between their two index sets. This reduces the number of nonzero covariances from the maximum possible $|\Theta_i|^2$ by a factor of at least n. Again, consider the case of $i = 3$ where $|\Theta_3| = \mathcal{O}(n^3)$.

Each pair of relations in Θ_3 must be of the form (jk, jm), and thus the second set of indices must be of the form (js, ju), for example, for the covariance to be nonzero. The set of indices $\{j, k, j, m, j, s, j, u\}$ is of order $\mathcal{O}(n^5) = |\Theta_3|^2 n^{-1}$. There are other forms of relation pairs in the second sum that give rise to nonzero covariance, such as (ks, ku) and so on. However, there are nine such forms, each of which is $\mathcal{O}(n^5)$. Thus, the number of nonzero covariances is $\mathcal{O}(n^5)$, and hence, we have

$$V\left[\widetilde{\phi}_i\right] = \frac{|\Theta_i|^2 \mathcal{O}(n^{-1})}{|\Theta_i|^2} \to 0. \tag{B.34}$$

This same argument holds for all i, and thus, we have the desired consistency: $\widetilde{\phi}_i - \phi_i \to_p 0$ for $i = 1, \ldots, 5$.

3. **Asymptotic equivalence of $\widehat{\phi}_i$ and $\widetilde{\phi}_i$**

We now show that $\widehat{\phi}_i - \widetilde{\phi}_i$ converges in probability to zero. We first write the expression in terms of the estimated coefficients $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned}\widehat{\phi}_i \; - \; \widetilde{\phi}_i &= \frac{\sum_{(jk,\ell m)\in\Theta_i} e_{jk}e_{\ell m} - \xi_{jk}\xi_{\ell m}}{|\Theta_i|} \\ &= \frac{1}{|\Theta_i|} \sum_{(jk,\ell m)\in\Theta_i} \left((\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T (\mathbf{x}_{jk}\mathbf{x}_{\ell m}^T)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) - (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T (\xi_{jk}\mathbf{x}_{\ell m} + \xi_{\ell m}\mathbf{x}_{jk}) \right). \end{aligned} \tag{B.35}$$

By Theorem 9, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ converges to zero in probability. By Slutsky's theorem, if the terms in (B.35) involving elements of X and $\boldsymbol{\Xi}_{\boldsymbol{v}}$ converge in probability to any constant, then $\widehat{\phi}_i - \widetilde{\phi}_i$ converges in probability to zero. By (B1) and (3.8) we have the convergence in probability of the term involving $\mathbf{x}_{jk}\mathbf{x}_{\ell m}^T$. Furthermore, by condition (B3), we have that $\mathbb{E}[\xi_{jk}\mathbf{x}_{\ell m}] = \mathbb{E}[\xi_{\ell m}\mathbf{x}_{jk}] = 0$. It remains to be shown that the variance of the error-covariate averages tend to zero. Consider the variance of the first error-covariate averages:

$$V\left[\frac{1}{|\Theta_i|} \sum_{(jk,\ell m)\in\Theta_i} \xi_{jk}\mathbf{x}_{\ell m}\right] = \frac{1}{|\Theta_i|^2} \sum_{(jk,\ell m)\in\Theta_i} \sum_{(rs,tu)\in\Theta_i} \operatorname{Cov}\left(\xi_{jk}\mathbf{x}_{\ell m}, \xi_{rs}\mathbf{x}_{tu}\right), \tag{B.36}$$

$$= \frac{1}{|\Theta_i|^2} \sum_{(jk,\ell m)\in\Theta_i} \sum_{(rs,tu)\in\Theta_i} \mathbb{E}\left[\mathbf{x}_{\ell m}\mathbf{x}_{tu}^T\right] \operatorname{Cov}\left(\xi_{jk}, \xi_{rs}\right). \tag{B.37}$$

In writing (B.37), we use condition (B3) and simplify by conditioning on X and using the law of total variance. By the same counting arguments used to establish (B.34), there are $|\Theta_i|^2\mathcal{O}(n^{-1})$ nonzero bounded

covariances in (B.37). Thus, we have

$$V\left[\frac{1}{|\Theta_i|}\sum_{(jk,\ell m)\in\Theta_i}\xi_{jk}\mathbf{x}_{\ell m}\right] = \frac{|\Theta_i|^2\mathcal{O}(n^{-1})}{|\Theta_i|^2} \to 0. \tag{B.38}$$

Since the expectation and variance both tend to zero, we have

$$\frac{1}{|\Theta_i|}\sum_{(jk,\ell m)\in\Theta_i}\xi_{jk}\mathbf{x}_{\ell m} \to_p 0. \tag{B.39}$$

The same argument applies to the second error-covariate term in (B.35). Thus, we have shown that consistency of $\widehat{\boldsymbol{\beta}}$ implies

$$\widehat{\phi}_i - \widetilde{\phi}_i \to_p 0. \tag{B.40}$$

□

B.4 Proof of MSE result

In this section, we prove that the MSE of $\widehat{V}[\widehat{\boldsymbol{\beta}}]$, conditional on X, is lower when using the exchangeable estimator than that when using the dyadic clustering estimator with high probability in X, assuming that the error structure is exchangeable. Before proving the theorem, we provide a lemma that states that the MSE of the each estimator is asymptotically equivalent to the MSE of each estimator based on the true errors, which vastly simplifies the proof of the MSE theorem. Even so, we must consider higher order moments of ξ than the covariances $\mathrm{Cov}\,(\xi_{jk},\xi_{lm})$. So, we also provide a lemma in which we define the covariance of any pair of product of error relations $\mathrm{Cov}\,(\xi_{jk}\xi_{lm},\xi_{rs}\xi_{tu})$ and define the limiting values of the covariance of the error averages, $n\mathrm{Cov}\left(\widetilde{\phi}_v,\widetilde{\phi}_w\right)$, for every pair $(v,w)\in\{1,2,\ldots,5\}\times\{1,2,\ldots,5\}$.

In this Section, we use the notation $\mathcal{O}(n^a)$ and $\boldsymbol{\Theta}(n^a)$, for some $a\in\mathbb{R}$, to denote the convergence a sequence of numbers to a constant (possibly zero) and a nonzero constant, respectively, as n grows to infinity. In other words, $X_n=\mathcal{O}(n^a)$ means that the sequence $n^{-a}X_n$ converges to a constant that may be zero. The notation $X_n=\boldsymbol{\Theta}(n^a)$ means that the sequence $n^{-a}X_n$ converges to a nonzero constant. Lastly, it follows that $X_n=\mathcal{O}(n^{a-\epsilon})$ means that the sequence $n^{-a}X_n$ converges to zero.

We use similar notation for convergence of sequences of random variables. The notation $X_n = \mathcal{O}_p(n^a)$ for some $a \in \mathbb{R}$ means that the sequence $n^{-a}X_n$ converges in distribution to a random variable (possibly a constant). The notation $X_n = o_p(n^a)$ for some $a \in \mathbb{R}$ means that the sequence $n^{-a}X_n$ converges in probability to zero. Finally, we define $X_n = \mathbf{\Theta}_p(n^a)$ to mean that $n^{-a}X_n$ converges in distribution to a random variable with distribution that is not a point mass at zero, and thus possibly a nonzero constant (as will always be the case in this Section).

B.4.1 Lemmas in support of Theorem 11

The first lemma describes the covariances of parameter estimates based on the errors, which arise in the proof of the MSE theorem. Of interest are the covariances $\text{Cov}\left(\widetilde{\phi}_v, \widetilde{\phi}_w\right)$ for $(v, w) \in \{3, 4, 5\} \times \{3, 4, 5\}$, as there are $\mathbf{\Theta}(n)$ times as many of these covariances in $\widehat{V}_E[\widehat{\boldsymbol{\beta}}]$ as those covariances where at least one of v or w is in $\{1, 2\}$. However, we provide limiting values of all covariances for completeness. The proof of this lemma follows from recognizing that $\widetilde{\phi}_v$ is a sample average and from defining all possible covariances that make up $\text{Cov}\left(\widetilde{\phi}_v, \widetilde{\phi}_w\right)$ and their multiplicities.

Lemma 16. *If $\boldsymbol{\xi}$ is a mean zero random vector with positive definite covariance matrix in the exchangeable class, $\Omega = \sum_{i=1}^{5} \phi_i \mathcal{S}_i$, and $\mathbb{E}[\xi_{jk}^4] < L < \infty$, then the covariance $n\text{Cov}\left(\widetilde{\phi}_v, \widetilde{\phi}_w\right)$ for $(v, w) \in \{1, 2, \ldots, 5\} \times \{1, 2, \ldots, 5\}$ converges to*

$$n\text{Cov}\left(\widetilde{\phi}_v, \widetilde{\phi}_w\right) \to \begin{cases} \sum_{i=1}^{4} \alpha_i \beta_v \beta_w C\left(v, w\right)_i & (v, w) \in \{3, 4, 5\} \times \{3, 4, 5\} \\ \sum_{j=1}^{3} \gamma_j F(v, w)_j & (v, w) \in \{1, 2\} \times \{1, 2\}, \\ \sum_{k=1}^{4} \gamma_k D(v, w)_k & o.w., \end{cases} \tag{B.41}$$

$$\textit{where } \alpha_i := 1 + \mathbb{1}[i > 1] + \mathbb{1}[i = 4], \;\; i \in \{1, 2, 3, 4\},$$

$$\beta_i := 1 + \mathbb{1}[i = 5], \;\; i \in \{1, 2, 3, 4, 5\},$$

$$\gamma_i := 1 + \mathbb{1}[i > 2], \;\; i \in \{1, 2, 3, 4\},$$

and $C\left(v, w\right)_i$, $D\left(v, w\right)_i$, and $F\left(v, w\right)_i$ are unknown finite constants equal to $\text{Cov}\left(\xi_{jk}\xi_{lm}, \xi_{rs}\xi_{tu}\right)$ for various configurations of the sets $\{j, k, l, m\}$ and $\{r, s, t, u\}$.

Proof. By definition, the covariance $n\text{Cov}\left(\widetilde{\phi}_v, \widetilde{\phi}_w\right)$ is

$$n|\Theta_v|^{-1}|\Theta_w|^{-1} \sum_{(jk,lm)\in\Theta_v} \sum_{(rs,tu)\in\Theta_w} \mathrm{Cov}\left(\xi_{jk}\xi_{lm}, \xi_{rs}\xi_{tu}\right). \tag{B.42}$$

The sum is over $|\Theta_v||\Theta_w|$ terms. Whenever $\{j,k,l,m\} \cap \{r,s,t,u\} = \varnothing$, the covariance is zero. This removes a power of n from the sum in (B.42), such that the sum is over $\mathcal{O}(|\Theta_v||\Theta_w|n^{-1})$ possibly nonzero covariances. The scaled sum in (B.42) converges – provided that the number of values that $\mathrm{Cov}\left(\xi_{jk}\xi_{lm}, \xi_{rs}\xi_{tu}\right)$ can take is finite – as each covariance is finite by assumption and the sequence of covariances is homogeneous as n grows by exchangeability. In the remainder of the proof, we enumerate and define the covariances $\mathrm{Cov}\left(\xi_{jk}\xi_{lm}, \xi_{rs}\xi_{tu}\right)$ in (B.42) for particular pairs $(v,w) \in \{1,2,\ldots,5\} \times \{1,2,\ldots,5\}$, showing that the number of values that $\mathrm{Cov}\left(\xi_{jk}\xi_{lm}, \xi_{rs}\xi_{tu}\right)$ can take is finite. This is sufficient to establish convergence.

Both v and w in $\{3,4,5\}$:

We begin by analyzing the case of interest, that is when both v and w are members of $\{3,4,5\}$. As an example, we focus on $v = 3$ and $w = 4$, where the first product of error relations corresponds to the same-sender covariance (b) in Figure 3.1 and the second corresponds to the same-receiver covariance (c) in Figure 3.1. In this case, both $|\Theta_3| = |\Theta_4| = \boldsymbol{\Theta}(n^3)$. When $v = 3$ and $w = 4$, the covariance in (B.42) becomes

$$n\mathrm{Cov}\left(\widetilde{\phi}_3, \widetilde{\phi}_4\right) =_a n^{-5} \sum_{jk} \sum_{l\notin\{j,k\}} \sum_{rs} \sum_{t\notin\{r,s\}} \mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{sr}\xi_{tr}\right), \tag{B.43}$$

where '$=_a$' denotes equality in the limit as n grows to infinity.

Only pairs of relation products that share a single actor will remain in the limit, as there are an order of n fewer covariances resulting from pairs of relation products that share two actors. One such pair of relation products that share a single actor correspond to the case when $s = j$, i.e. $\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{jr}\xi_{tr}\right)$, of which there are $\boldsymbol{\Theta}(n^5)$ in the sum in (B.43). There are $\boldsymbol{\Theta}(n^4)$ covariances corresponding to the case when $s = j$ and $r = k$, i.e. $\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{jk}\xi_{tk}\right)$. The values of all covariances in (B.43) are finite by assumption and not equal in general. However, by exchangeability, covariances resulting from pairs of relations that share an actor in the same way *are* equal. For example, the covariance corresponding to $s = j$

is the same regardless of the node labeling, that is $\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{jr}\xi_{tr}\right) = \mathrm{Cov}\left(\xi_{ab}\xi_{ac}, \xi_{ad}\xi_{ed}\right)$ for any set $\{a, b, c, d, e\} \subset \{1, 2, \ldots, n\}$ with $|\{a, b, c, d, e\}| = 5$.

There are nine ways that we may have $|\{j, k, l\} \cap \{r, s, t\}| = 1$ in (B.43), i.e. there are nine ways that exactly one of $\{j, k, l\}$ equals exactly one of $\{r, s, t\}$. However, these reduce into four unique covariance values for each pair $(v, w) \in \{3, 4, 5\} \times \{3, 4, 5\}$. As an example, when $t = j$ the covariance is the same as that when $s = j$, that is $\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{jr}\xi_{tr}\right) = \mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{sr}\xi_{jr}\right)$. Now we define these four covariance values and their multiplicities out of the nine possible ways that exactly one of $\{j, k, l\}$ equals exactly one of $\{r, s, t\}$:

- When $r = j$, we define the covariance $C(3,4)_1 := \mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{sj}\xi_{tj}\right)$, of which there is one out of nine possible;
- When $s = j$, the covariance is the same as when $t = j$ (multiplicity two), and we define this covariance $C(3,4)_2 := \mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{jr}\xi_{tr}\right)$;
- When $r = k$, the covariance is the same as when $r = l$ (multiplicity two), and we define this covariance $C(3,4)_3 := \mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{sk}\xi_{tk}\right)$;
- We define the covariance when $s = k$ to be $C(3,4)_4 := \mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{kr}\xi_{tr}\right)$, of which there are four, the remaining terms of which correspond to $t = k$, $s = l$, and $t = l$.

Now, noting that there are $n^5 + \mathbf{\Theta}(n^4)$ covariances in the sum (B.43) corresponding to each of the nine possible ways that exactly one of $\{j, k, l\}$ equals exactly one of $\{r, s, t\}$, we see that

$$\begin{aligned} & n\mathrm{Cov}\left(\widetilde{\phi}_3, \widetilde{\phi}_4\right) \rightarrow \\ & \rightarrow \mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{sj}\xi_{tj}\right) + 2\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{jr}\xi_{tr}\right) \\ & \qquad + 2\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{sk}\xi_{tk}\right) + 4\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{kr}\xi_{tr}\right), \\ & := C(3,4)_1 + 2C(3,4)_2 + 2C(3,4)_3 + 4C(3,4)_4, \end{aligned} \tag{B.44}$$

where '$\rightarrow$' denotes convergence in the limit as n goes to infinity. Under appropriate definition of $C(v, w)_i$ for $i \in \{1, 2, 3, 4\}$, the same argument applies when both v and w are one of $\{3, 4\}$. When $w = 5$ (relation products of the form $\{\xi_{jk}\xi_{kl}\}$ and $\{\xi_{jk}\xi_{lj}\}$) and $v = 3$, however, we then must consider covariances

$\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{rs}\xi_{tr}\right)$ *and* $\mathrm{Cov}\left(\xi_{jk}\xi_{jl}, \xi_{sr}\xi_{rt}\right)$ from $w = 5$, which doubles the coefficients in (B.44). This accounts for $\beta_w = 2$ when $w = 5$ and $\beta_w = 1$ otherwise in (B.41). The same argument applies when $v = 5$.

Both v and w in $\{1, 2\}$:

We now analyze both v and w in $\{1, 2\}$, corresponding to variance and the reciprocal covariance (a) in Figure 3.1. In this case, both $|\Theta_v| = |\Theta_w| = n(n-1)$. Taking $v = 1$ and $w = 1$ as an example, in the limit, the covariance in (B.42) is

$$n\mathrm{Cov}\left(\widetilde{\phi}_1, \widetilde{\phi}_2\right) =_a n^{-3} \sum_{jk} \sum_{rs} \mathrm{Cov}\left(\xi_{jk}^2, \xi_{rs}^2\right). \tag{B.45}$$

Again, we only consider covariances corresponding to pairs of relation products that share a single actor as only these covariances survive in the limit. There are four possible ways that $\{j, k\}$ shares exactly one actor with $\{r, s\}$. We define the three unique covariances and their multiplicities corresponding to the four ways that $\{j, k\}$ shares exactly one actor with $\{r, s\}$ as follows:

- When $r = j$, we define the covariance $F(1,1)_1 := \mathrm{Cov}\left(\xi_{jk}^2, \xi_{js}^2\right)$, of which there is one out of the four possibilities;
- When $s = k$, we define the covariance $F(1,1)_2 := \mathrm{Cov}\left(\xi_{jk}^2, \xi_{rk}^2\right)$, of which there is one;
- When $s = j$, we define the covariance $F(1,2)_3 := \mathrm{Cov}\left(\xi_{jk}^2, \xi_{rj}^2\right)$, which is the same as when $r = k$, accounting for the remaining two possibilities.

Now, the fact that there are $n^3 + \mathbf{\Theta}(n^2)$ covariances in the sum (B.45) corresponding to each of the four possible ways that $\{j, k\}$ shares exactly one actor with $\{r, s\}$ gives that

$$\begin{aligned} n\mathrm{Cov}\left(\widetilde{\phi}_1, \widetilde{\phi}_1\right) &\to \mathrm{Cov}\left(\xi_{jk}^2, \xi_{js}^2\right) + \mathrm{Cov}\left(\xi_{jk}^2, \xi_{rk}^2\right) + 2\mathrm{Cov}\left(\xi_{jk}^2, \xi_{rj}^2\right) \\ &:= F(1,1)_1 + F(1,1)_2 + 2F(1,1)_3. \end{aligned} \tag{B.46}$$

Of course, the same argument applies to any v and w both in $\{1, 2\}$. In the case where $v = 1$ and $w = 2$, by symmetry, $F(1,2)_1 = F(1,2)_2$. Similarly, for $v = w = 2$, all $F(2,2)_1 = F(2,2)_2 = F(2,2)_3$.

One of v and w in $\{1,2\}$ and the other in $\{3,4,5\}$:

Similar counting arguments to those in the previous paragraphs apply when one of v, w is in $\{3,4,5\}$ and the other is in $\{1,2\}$. As an example, consider $v = 1$ and $w = 3$. Once again, only pairs of relations that share a single actor will remain in the limit. Then, in the limit, the covariance in (B.42) becomes

$$n\text{Cov}\left(\widetilde{\phi}_2, \widetilde{\phi}_3\right) =_a n^{-4} \sum_{jk} \sum_{rs} \sum_{t \notin \{r,s\}} \text{Cov}\left(\xi_{jk}^2, \xi_{rs}\xi_{rt}\right). \tag{B.47}$$

Now, there are six ways in which the first pair of relations share an actor with the second pair, i.e. all sets with exactly one actor from $\{j,k\}$ equal to exactly one other from $\{r,s,t\}$. We define the covariances corresponding to the six possibilities below:

- When $r = j$, we define the covariance $D(1,3)_1 := \text{Cov}\left(\xi_{jk}^2, \xi_{js}\xi_{jt}\right)$, of which there is one out of the six possibilities;
- When $r = k$, we define the covariance $D(1,3)_2 := \text{Cov}\left(\xi_{jk}^2, \xi_{ks}\xi_{kt}\right)$, of which there is one;
- The overlaps where $s = j$ and $t = j$ result in the same covariance (multiplicity two), which we define $D(1,3)_3 := \text{Cov}\left(\xi_{jk}^2, \xi_{rj}\xi_{rt}\right)$;
- The overlaps where $s = k$ and $t = k$ result in the same covariance (multiplicity two), which we define $D(1,3)_4 := \text{Cov}\left(\xi_{jk}^2, \xi_{rk}\xi_{rt}\right)$.

Then, noting that there are $n^4 + \mathbf{\Theta}(n^3)$ covariances in the sum (B.47) corresponding to each of the six possible ways that exactly one actor from $\{j,k\}$ is equal to exactly one other from $\{r,s,t\}$, we have that $n\text{Cov}\left(\widetilde{\phi}_2, \widetilde{\phi}_3\right)$ converges to

$$\begin{aligned} &\to \text{Cov}\left(\xi_{jk}^2, \xi_{js}\xi_{jt}\right) + \text{Cov}\left(\xi_{jk}^2, \xi_{ks}\xi_{kt}\right) + 2\text{Cov}\left(\xi_{jk}^2, \xi_{rj}\xi_{rt}\right) + 2\text{Cov}\left(\xi_{jk}^2, \xi_{rk}\xi_{rt}\right), \\ &:= D(1,3)_1 + D(1,3)_2 + 2D(1,3)_3 + 2D(1,3)_4. \end{aligned} \tag{B.48}$$

When $D(v,w)_k$ for $k \in \{1,2,3,4\}$ is appropriately defined, the same argument applies for all settings where one of v, w is in $\{3,4\}$ and the other is in $\{1,2\}$. When $w = 5$ (relation products of the form $\{\xi_{jk}\xi_{kl}\}$ and $\{\xi_{jk}\xi_{lj}\}$), however, we then must consider covariances in (B.47) $\text{Cov}\left(\xi_{jk}^2, \xi_{rs}\xi_{tr}\right)$ *and* $\text{Cov}\left(\xi_{jk}^2, \xi_{sr}\xi_{rs}\right)$, which doubles the coefficients in (B.48). This accounts for $\beta_w = 2$ when $w = 5$ and

$\beta_w = 1$ otherwise in (B.41). We note that when $v = 2$, for example, we have the simplification that $D(2,3)_1 = D(2,3)_2$ and $D(2,3)_3 = D(2,3)_4$. □

The expressions for the estimators based on the errors are simpler to analyze than those based on the residuals. For example, when comparing the MSEs of the exchangeable and dyadic clustering estimators, it is desirable to analyze $MSE_\xi\left(\widetilde{V}_E|X\right)$ instead of $MSE_\xi\left(\widehat{V}_E|X\right)$. The following lemma allows us to do just this. This lemma states that MSEs of the estimators based on the errors are asymptotically equivalent to the MSEs of those based on the residuals. The proof consists of first evaluating the MSE conditional on X. We then show that $n^3 MSE_\xi\left(\widetilde{V}_E|X\right)$ converges in X-probability to a nonzero constant in general and that $n^3 MSE_\xi\left(\widetilde{V}_E|X\right) - n^3 MSE_\xi\left(\widehat{V}_E|X\right)$ converges in X-probability to zero, implying that the difference between $MSE_\xi\left(\widetilde{V}_E|X\right)$ and $MSE_\xi\left(\widehat{V}_E|X\right)$ is asymptotically negligible. We repeat the procedure for $MSE_\xi\left(\widetilde{V}_{DC}|X\right)$ and $MSE_\xi\left(\widehat{V}_{DC}|X\right)$.

Lemma 17. *Assuming $\mathbb{E}\left(|x_{jk}^{(l)}|^8\right) < L^8 < \infty$ for all $l \in \{1, 2, \ldots, p\}$ and under the assumptions of Theorem 9, the MSE for both the exchangeable and dyadic clustering estimators based on the residuals is asymptotically equivalent to the MSE of each respective estimator based on the errors. That is,*

$$n^3 MSE_\xi\left(\widehat{V}_E|X\right) = n^3 MSE_\xi\left(\widetilde{V}_E|X\right) + \mathcal{O}_p(n^{-1/2}) = \mathcal{O}_p(1), \tag{B.49}$$

and analogously for dyadic clustering.

Proof. We will focus on the exchangeable estimator first, and then the dyadic clustering estimator. Throughout, we drop the conditioning on X in the MSE as it is understood, for example $MSE_\xi\left(\widehat{V}_E\right) := MSE_\xi\left(\widehat{V}_E|X\right)$.

Exchangeable estimator:

By definition, the MSE of the exchangeable estimator is

$$MSE_\xi\left(\widehat{V}_E\right) = \mathbb{E}\left[\left(\widehat{V}_E - V^*\right)^2 \,\middle|\, X\right], \tag{B.50}$$

$$= MSE_\xi\left(\widetilde{V}_E\right) + \mathbb{E}\left[\left(\widehat{V}_E - \widetilde{V}_E\right)^2 \,\middle|\, X\right] + 2\mathbb{E}\left[\left(\widehat{V}_E - \widetilde{V}_E\right)\left(\widetilde{V}_E - V^*\right) \mid X\right], \tag{B.51}$$

where $V^* := V[\widehat{\beta}]$, the true variance of the OLS coefficient estimate. By the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\left(\widehat{V}_E - \widetilde{V}_E\right)\left(\widetilde{V}_E - V^*\right) \mid X\right] \leq \sqrt{MSE_\xi\left(\widetilde{V}_E\right)\mathbb{E}\left[\left(\widehat{V}_E - \widetilde{V}_E\right)^2 \mid X\right]}. \tag{B.52}$$

If we show that $n^3 MSE_\xi\left(\widetilde{V}_E\right)$ converges in X-probability to a constant, i.e. $MSE_\xi\left(\widetilde{V}_E\right) = \mathcal{O}_p(n^{-3})$, and that $\mathbb{E}\left[\left(\widehat{V}_E - \widetilde{V}_E\right)^2 \mid X\right] = \mathcal{O}_p(n^{-4})$, then (B.52) implies that the third additive term of (B.51) is $\mathcal{O}_p(n^{-7/2})$. This is sufficient to establish (B.49). We begin with showing $n^3 MSE_\xi\left(\widetilde{V}_E\right) = \mathcal{O}_p(1)$. By definition, the scaled MSE is $n^3 MSE_\xi\left(\widetilde{V}_E\right) = n^3\mathbb{E}\left[\mathrm{tr}\left(\widetilde{V}_E^2\right) \mid X\right]$, which is equal to

$$\sum_{v=1}^{5}\sum_{w=1}^{5} n\mathrm{Cov}\left(\widetilde{\phi}_v, \widetilde{\phi}_w\right)\mathrm{tr}\left(A_n^{-1}\left(\frac{X^T\mathcal{S}_v X}{n^3}\right)A_n^{-2}\left(\frac{X^T\mathcal{S}_w X}{n^3}\right)A_n^{-1}\right), \tag{B.53}$$

where $A_n := \frac{X^T X}{n^2}$. By Lemma 16, $n\mathrm{Cov}\left(\widetilde{\phi}_v, \widetilde{\phi}_w\right)$ converges to a finite constant for every $(v, w) \in \{1, 2, \ldots, 5\}\times\{1, 2, \ldots, 5\}$. The convergence in probability of each multiplicative term in (B.53) containing X is defined by assumption (B1); only those with both v and w in $\{3, 4, 5\}$ survive in the limit as these have $|\Theta_v| = \mathbf{\Theta}(n^3)$ whereas $|\Theta_v| = \mathbf{\Theta}(n^2)$ for $v \in \{1, 2\}$. Thus, we have that

$$n^3 MSE_\xi\left(\widetilde{V}_E\right) \xrightarrow{\mathbb{P}_X} \sum_{v=3}^{5}\sum_{w=3}^{5}\sum_{i=1}^{4}\alpha_i\beta_v\beta_w C(v, w)_i \mathrm{tr}\left(M_1^{-1}M_v M_1^{-2}M_w M_1^{-1}\right), \tag{B.54}$$

which is finite.

It remains to show that $\mathbb{E}\left[\left(\widehat{V}_E - \widetilde{V}_E\right)^2 \mid X\right] = \mathcal{O}_p(n^{-4})$. To establish this fact, it is sufficient to show that $\widehat{V}_E - \widetilde{V}_E = \mathcal{O}_p(n^{-2})$, and then, by the continuous mapping theorem, $\left(\widehat{V}_E - \widetilde{V}_E\right)^2 = \mathcal{O}_p(n^{-4})$, which implies the desired result. Writing directly,

$$\widehat{V}_E - \widetilde{V}_E = \frac{1}{n}\sum_{v=1}^{5}\left(\widehat{\phi}_v - \widetilde{\phi}_v\right)\left(\frac{X^T X}{n^2}\right)^{-1}\left(\frac{X^T\mathcal{S}_v X}{n^3}\right)\left(\frac{X^T X}{n^2}\right)^{-1}. \tag{B.55}$$

By assumption (B1), the multiplicative terms involving X converge in probability to constants. To establish $\widehat{V}_E - \widetilde{V}_E = \mathcal{O}_p(n^{-2})$, it is sufficient show that $\widehat{\phi}_v - \widetilde{\phi}_v = \mathcal{O}_p(n^{-1})$ for all $v \in \{1, 2, \ldots, 5\}$. Writing this expression directly, the difference $\widehat{\phi}_v - \widetilde{\phi}_v$ is

$$-(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\left(\sum_{(jk,lm)\in\Theta_v}\frac{\mathbf{x}_{jk}\xi_{lm}+\mathbf{x}_{lm}\xi_{jk}}{|\Theta_v|}\right)+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\left(\sum_{(jk,lm)\in\Theta_v}\frac{\mathbf{x}_{jk}\mathbf{x}_{lm}^T}{|\Theta_v|}\right)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}). \tag{B.56}$$

By Theorem 9, $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\mathcal{O}_p(n^{-1/2})$. Also, by assumption (B1), the sum involving X in the second term converges in probability to a constant; thus, the second additive term in (B.56) is $\mathcal{O}_p(n^{-1})$. Turning to the first additive term, we notice its expectation is zero, that is $\mathbb{E}[\mathbf{x}_{jk}\xi_{lm}]=0$ for all relations jk and lm. The variance is

$$V\left[\sum_{(jk,lm)\in\Theta_v}\frac{\mathbf{x}_{jk}\xi_{lm}}{|\Theta_v|}\right]=\frac{1}{|\Theta_v|^2}\sum_{(jk,lm)\in\Theta_v}\sum_{(rs,tu)\in\Theta_v}\mathbb{E}\left[\mathbf{x}_{jk}\mathbf{x}_{rs}^T\xi_{lm}\xi_{tu}\right]=\mathcal{O}(n^{-1}), \tag{B.57}$$

where we use the fact that $\mathbb{E}\left[\mathbf{x}_{jk}\mathbf{x}_{rs}^T\xi_{lm}\xi_{tu}\right]$ is only nonzero when relation lm shares an actor with relation tu since $\mathbb{E}\left[\xi_{lm}\xi_{tu}\right]=0$ whenever $\{j,k\}\cap\{l,m\}=\varnothing$ and X is independent $\boldsymbol{\xi}$ by assumption (B3). This fact removes at least a factor of n from the sum. Thus, we have that $\sum_{(jk,lm)\in\Theta_v}\frac{\mathbf{x}_{jk}\xi_{lm}+\mathbf{x}_{lm}\xi_{jk}}{|\Theta_v|}=\mathcal{O}_p(n^{-1/2})$, which gives that $\widehat{\phi}_v-\widetilde{\phi}_v=\mathcal{O}_p(n^{-1})$ and

$$\mathbb{E}\left[\left(\widehat{V}_E-\widetilde{V}_E\right)^2\mid X\right]=\mathcal{O}_p(n^{-4}), \tag{B.58}$$

which establishes (B.49) for the exchangeable estimator.

Dyadic clustering estimator:

The same argument following (B.51) applies to the dyadic clustering estimator. To establish (B.49) for the dyadic clustering estimator, it is thus sufficient to show that $n^3MSE_\xi\left(\widetilde{V}_{DC}\right)$ converges in X-probability to a constant and that $\mathbb{E}\left[\left(\widehat{V}_{DC}-\widetilde{V}_{DC}\right)^2\mid X\right]=\mathcal{O}_p(n^{-4})$. We begin with the former. By definition, $n^3MSE_\xi\left(\widetilde{V}_{DC}\right)$ is

$$\begin{aligned}&\frac{1}{n^5}\sum_{(jk,lm)\in\Theta_0}\sum_{(rs,tu)\in\Theta_0}\mathrm{Cov}\left(\xi_{jk}\xi_{lm},\xi_{rs}\xi_{tu}\right)\times\ldots\\&\quad\ldots\times\mathrm{tr}\left(\left(\frac{X^TX}{n^2}\right)^{-1}\mathbf{x}_{jk}\mathbf{x}_{lm}^T\left(\frac{X^TX}{n^2}\right)^{-2}\mathbf{x}_{rs}\mathbf{x}_{tu}^T\left(\frac{X^TX}{n^2}\right)^{-1}\right),\end{aligned} \tag{B.59}$$

where Θ_0 is the set of relation pairs that share an actor in *any* manner. Then, substituting the asymptotic values for $\text{Cov}\left(\xi_{jk}\xi_{lm}, \xi_{rs}\xi_{tu}\right)$ from Lemma 16 and separating the sum by the five ways that two relations may share an actor,

$$n^3 MSE_\xi\left(\widetilde{V}_{DC}\right) =_a \frac{1}{n^5}\sum_{v=3}^{5}\sum_{w=3}^{5}\sum_{i=1}^{4}\sum_{T(v,w)_i} C(v,w)_i\ \mathbf{x}_{lm}^T M_1^{-2}\mathbf{x}_{rs}\mathbf{x}_{tu}^T M_1^{-2}\mathbf{x}_{jk}, \tag{B.60}$$

where '$=_a$' denotes equality in the limit and $T(v,w)_i$ is the set of relations (jk, lm, rs, tu) such that $(jk, lm) \in \Theta_v$ and $(rs, tu) \in \Theta_w$ and such that the pairs of relations (jk, lm) and (rs, tu) share a single actor as appropriate for $C(v,w)_i$ in Lemma 16. In (B.60), we substitute the limit of $\left(\frac{X^TX}{n^2}\right)^{-2}$ from assumption (B1). Also in (B.60), only terms with v and w both in $\{3,4,5\}$ survive in the limit as the set $|T(v,w)_i| = \mathbf{\Theta}(n^5)$ (as detailed in Lemma 16), while the order is less for either v or w in $\{1,2\}$, so these terms vanish in the limit. Evaluating the vector products, the expression on the right hand side of (B.60) equal to

$$\sum_{v=3}^{5}\sum_{w=3}^{5}\sum_{i=1}^{4}\sum_{a=1}^{p}\sum_{b=1}^{p}\sum_{c=1}^{p}\sum_{d=1}^{p} C(v,w)_i\ \left(m_1^{-2}\right)_{ab}\left(m_1^{-2}\right)_{cd}\left(\frac{1}{n^5}\sum_{T(v,w)_i} x_{lm}^{(a)}x_{rs}^{(b)}x_{tu}^{(c)}x_{jk}^{(d)}\right), \tag{B.61}$$

where $\left(m_1^{-2}\right)_{ab}$ is the (a,b) entry in M_1^{-2}, e.g., and $x_{jk}^{(a)}$ is the entry in X pertaining to column a and relation jk. Further, the variance
$V\left[\frac{1}{n^5}\sum_{T(v,w)_i} x_{lm}^{(a)}x_{rs}^{(b)}x_{tu}^{(c)}x_{jk}^{(d)}\right]$ is

$$\frac{1}{n^{10}}\sum_{T(v,w)_i}\sum_{U(v,w)_i}\text{Cov}\left(x_{lm}^{(a)}x_{rs}^{(b)}x_{tu}^{(c)}x_{jk}^{(d)}, x_{ef}^{(a)}x_{gh}^{(b)}x_{np}^{(c)}x_{yz}^{(d)}\right), \tag{B.62}$$

$$= \frac{\mathbf{\Theta}(n^9)}{n^{10}} \to 0, \tag{B.63}$$

where $U(v,w)_i = T(v,w)_i$ and (lm, rs, tu, jk) indexes the first sum and (ef, gh, np, yz) indexes the second sum. The convergence is the result of the independence portion of assumption in (B1) and the bounded moment assumption on X. The variance in (B.62) converges to zero for every set of covariates $\{a,b,c,d\}$, every relation type v and w both in $\{3,4,5\}$, and every covariance type $i \in \{1,2,3,4\}$. Thus, provided that the expectation of $n^{-5}\sum_{T(v,w)_i} x_{lm}^{(a)}x_{rs}^{(b)}x_{tu}^{(c)}x_{jk}^{(d)}$ converges to a constant, this expression converges in probability to that same constant. This expectation is

$$\mathbb{E}\left[x_{lm}^{(a)}x_{rs}^{(b)}x_{tu}^{(c)}x_{jk}^{(d)}\right] = \mathrm{Cov}\left(x_{jk}^{(d)}x_{lm}^{(a)}, x_{rs}^{(b)}x_{tu}^{(c)}\right) + (m_v)_{ad}\,(m_w)_{bc}\,, \tag{B.64}$$

where $(m_v)_{ij}$ is the (i,j) entry in M_v and we use the symmetry of M_v for all $v \in \{3,4,5\}$. Unlike $\boldsymbol{\xi}$, for a given $i \in \{1,2,3,4\}$, the covariances in (B.64) may not be the same for two relation sets in $T(v,w)_i$. However, by assumption (B1) and taking $i=1$, $v=3$, and $w=4$ for example, we still have that

$$\mathrm{Cov}\left(x_{jl}^{(d)}x_{sj}^{(a)}, x_{tj}^{(b)}x_{jk}^{(c)}\right) = \mathrm{Cov}\left(x_{ef}^{(d)}x_{ge}^{(a)}, x_{he}^{(b)}x_{ep}^{(c)}\right), \tag{B.65}$$

for $|\{j,k,l,s,t,e,f,g,h,p\}| = 10$. That is, covariances that share an actor in the same way are still equal. So, for fixed $i \in \{1,2,3,4\}$ and pair of v and w both in $\{3,4,5\}$, we may collect the α_i possible covariances and average them to attain the convergent value. We thus define the limit

$$\frac{1}{n^5}\sum_{T(v,w)_i}\mathrm{Cov}\left(x_{jk}^{(d)}x_{lm}^{(a)}, x_{rs}^{(b)}x_{tu}^{(c)}\right) \to \alpha_i\beta_v\beta_w\frac{1}{\alpha_i}\sum_{W(v,w)_i}\mathrm{Cov}\left(x_{jk}^{(d)}x_{lm}^{(a)}, x_{rs}^{(b)}x_{tu}^{(c)}\right), \tag{B.66}$$

$$:= \alpha_i\beta_v\beta_w C_X^{(d,a,b,c)}(v,w)_i, \tag{B.67}$$

where $W(v,w)_i$ is the set of α_i *ways* that (jk,lm,rs,tu) correspond to $T(v,w)_i$. For example, when $i=4$, $v=3$, and $w=4$, $W(v,w)_i$ contains four index sets corresponding to the four multiplicities of $C(v,w)_4$ as defined in Lemma (3). The convergence of (B.66) results from (B.65). As the average over is $W(v,w)_i$ is over a finite number of terms, i.e. each α_i is bounded, there is no possibility of divergence. We note that the covariances in (B.66) are finite by assumption (B2) and $\beta_v\beta_w$ arises from the asymptotic limit of $n^{-5}|T(v,w)_i|$. Taking (B.66) together with (B.64), we have the convergence of the expectation of $n^{-5}\sum_{T(v,w)_i}x_{lm}^{(a)}x_{rs}^{(b)}x_{tu}^{(c)}x_{jk}^{(d)}$. Along with (B.62), convergence of the expectation of $n^{-5}\sum_{T(v,w)_i}x_{lm}^{(a)}x_{rs}^{(b)}x_{tu}^{(c)}x_{jk}^{(d)}$ establishes the convergence of $n^{-5}\sum_{T(v,w)_i}x_{lm}^{(a)}x_{rs}^{(b)}x_{tu}^{(c)}x_{jk}^{(d)}$ to the same limit.

Now, for a particular v and w both in $\{3,4,5\}$ and $i \in \{1,2,3,4\}$, we collect the set of $\{C_X^{(d,a,b,c)}(v,w)_i\}$ for every a,b,c, and d in $\{1,2,\ldots,p\}$ from (B.67) into a $p^2 \times p^2$ matrix defined $D_X(v,w)_i$. Substituting this definition into (B.61) while noting each $\{M_v\}_{v=1}^5$ is symmetric, the convergent value for the dyadic clustering estimator is

$$n^3 MSE_\xi\left(\widetilde{V}_{DC}\right) \xrightarrow{\mathbb{P}_X} \sum_{v=3}^{5}\sum_{w=3}^{5}\sum_{i=1}^{4}\alpha_i\beta_v\beta_w C(v,w)_i \times \ldots \tag{B.68}$$
$$\ldots \times \left(\text{vec}\left(M_1^{-2}\right)^T D_X(v,w)_i \text{vec}\left(M_1^{-2}\right) + \text{tr}\left(M_1^{-2}M_v M_1^{-2} M_w\right)\right).$$

Noting that the convergent value in (B.68) is a finite constant, it remains to show that $\mathbb{E}\left[\left(\widehat{V}_{DC} - \widetilde{V}_{DC}\right)^2 \mid X\right] = \mathcal{O}_p(n^{-4})$. As with the exchangeable estimator, it is sufficient to show that $\widehat{V}_{DC} - \widetilde{V}_{DC} = \mathcal{O}_p(n^{-2})$. Using the residual definition $e_{jk} = \xi_{jk} + \mathbf{x}_{jk}^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})$, the expression for $\widehat{V}_{DC} - \widetilde{V}_{DC}$ is

$$\frac{1}{n}\sum_{v=1}^{5}\sum_{(jk,lm)\in\Theta_v}\left(\frac{X^TX}{n^2}\right)^{-1}\left(\frac{\mathbf{x}_{jk}\mathbf{x}_{lm}^T}{n^3}\left(e_{jk}e_{lm} - \xi_{jk}\xi_{lm}\right)\right)\left(\frac{X^TX}{n^2}\right)^{-1} \tag{B.69}$$
$$=_a \frac{1}{n}\sum_{v=1}^{5}\sum_{(jk,lm)\in\Theta_v} M_1^{-1}\left(\frac{\mathbf{x}_{jk}\mathbf{x}_{lm}^T}{n^3}\left(\mathbf{x}_{jk}^T\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\mathbf{x}_{lm}\right) - 2\mathbf{x}_{jk}^T\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\xi_{lm}\right)M_1^{-1},$$

where we substitute the convergence of $\left(\frac{X^TX}{n^2}\right)^{-1}$ to M_1^{-1} and have used the exchangeability property to get the factor of two on the second additive term in the center of (B.69). Analyzing the first additive term in the center of (B.69),

$$\sum_{v=1}^{5}\sum_{(jk,lm)\in\Theta_v}\frac{\mathbf{x}_{jk}\mathbf{x}_{lm}^T}{n^3}\left(\mathbf{x}_{jk}^T\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\mathbf{x}_{lm}\right)$$
$$= \sum_{jk}\sum_{lm\in\Theta_{jk}}\left(\frac{\mathbf{x}_{jk}\mathbf{x}_{jk}^T}{n^2}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\left(\frac{\mathbf{x}_{lm}\mathbf{x}_{lm}^T}{n}\right), \tag{B.70}$$
$$= \boldsymbol{\Theta}_p(1)\mathcal{O}_p(n^{-1/2})\mathcal{O}_p(n^{-1/2})\boldsymbol{\Theta}_p(1) = \mathcal{O}_p(n^{-1}), \tag{B.71}$$

recalling the notation that Θ_{jk} is the set of all relations that share an actor with relation jk and that $|\Theta_{jk}| = \boldsymbol{\Theta}(n)$. We attain the convergence rate by noting that the X-terms in (B.70) converge in probability to constants by assumption (B1) and $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathcal{O}_p(n^{-1/2})$ by Theorem 9. The convergences in (B.70) are for $p \times p$ matrices; these convergences are element-wise.

We now analyze the convergence rate of the second additive term in the center of (B.69),

$$\sum_{v=1}^{5} \sum_{(jk,lm)\in\Theta_v} \frac{\mathbf{x}_{jk}\mathbf{x}_{lm}^T}{n^3} \left(\mathbf{x}_{jk}^T \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xi_{lm}\right) = \sum_{lm} \sum_{jk\in\Theta_{lm}} \left(\frac{\mathbf{x}_{jk}\mathbf{x}_{jk}^T}{n}\right) \left(\frac{\xi_{lm}\mathbf{x}_{lm}^T}{n^2}\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right), \tag{B.72}$$

$$= \boldsymbol{\Theta}_p(1)\mathcal{O}_p(n^{-1/2})\mathcal{O}_p(n^{-1/2}) = \mathcal{O}_p(n^{-1}). \tag{B.73}$$

Again, the convergence of the first multiplicative term is a result of assumption (B1) and $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathcal{O}_p(n^{-1/2})$ by Theorem 9. The mean $n^{-2}\sum_{lm} \mathbf{x}_{lm}\xi_{lm}$ is expectation zero and $\mathcal{O}_p(n^{-1/2})$ by previous arguments, for example in (B.24). Thus, we have $\widehat{V}_{DC} - \widetilde{V}_{DC} = \mathcal{O}_p(n^{-2})$, and the dyadic clustering estimator satisfies the relation in (B.49). □

B.4.2 Proof of Theorem 11

We now establish that the MSE of the exchangeable estimator is less than that of the dyadic clustering estimator with high probability. To do so, we show that the value to which the difference in MSEs converges is nonnegative. Throughout the proof, we drop the conditioning on X in the MSE as it is understood, for example $MSE_\xi\left(\widehat{V}_E\right) := MSE_\xi\left(\widehat{V}_E|X\right)$.

The asymptotic difference in MSEs is as follows, where we substitute the expressions for the estimators based on the errors in (B.54) and (B.68), as justified by Lemma 17:

$$n^3\left(MSE_\xi\left(\widehat{V}_{DC}\right) - MSE_\xi\left(\widehat{V}_E\right)\right) \xrightarrow{\mathbb{P}_X}$$

$$\xrightarrow{\mathbb{P}_X} \sum_{v=3}^{5}\sum_{w=3}^{5}\sum_{i=1}^{4} \alpha_i\beta_v\beta_w C(v,w)_i \text{vec}\left(M_1^{-2}\right)^T D_X(v,w)_i \text{vec}\left(M_1^{-2}\right). \tag{B.74}$$

It remains to show that this is a nonnegative constant. To do so, we show that the matrix in the quadratic form in (B.74) is the limit of a variance matrix, and thus positive semi-definite. We will show that the scaled variance

$$\frac{1}{n^5} V\left[\sum_{v=1}^{5} \sum_{jk,lm\in\Theta_v} \xi_{jk}\xi_{lm} \text{vec}\left(\mathbf{x}_{jk}\mathbf{x}_{lm}^T\right)\right]$$

$$= \frac{1}{n^5}\sum_{v=1}^{5}\sum_{w=1}^{5} \sum_{jk,lm\in\Theta_v} \sum_{rs,tu\in\Theta_w} \text{Cov}\left(\xi_{jk}\xi_{lm}\text{vec}\left(\mathbf{x}_{jk}\mathbf{x}_{lm}^T\right),\ \xi_{rs}\xi_{tu}\text{vec}\left(\mathbf{x}_{rs}\mathbf{x}_{tu}^T\right)\right), \tag{B.75}$$

converges to the desired matrix. First, we note that the sum in (B.75) is $\boldsymbol{\Theta}(n^5)$ as the relations jk and lm must share at least one actor with the relations rs and tu for the covariance to be nonzero. Then, by the arguments in Lemma 16, only pairs of relations that share a single actor survive in the limit. Finally, by assumption (B3), X is independent $\boldsymbol{\xi}$ and the variance in (B.75) is asymptotically equivalent to

$$\frac{1}{n^5}\sum_{v=3}^{5}\sum_{w=3}^{5}\sum_{jk,lm\in\Theta_v}\sum_{rs,tu\in\Theta_w}\mathbb{E}\left[\left(\xi_{jk}\xi_{lm}-\phi_v\right)\left(\xi_{rs}\xi_{tu}-\phi_w\right)\right]\times\ldots$$
$$\ldots\times\mathbb{E}\left[\left(\text{vec}\left(\mathbf{x}_{jk}\mathbf{x}_{lm}^T\right)-\text{vec}\left(M_v\right)\right)\left(\text{vec}\left(\mathbf{x}_{rs}\mathbf{x}_{tu}^T\right)-\text{vec}\left(M_w\right)\right)^T\right], \tag{B.76}$$

where only terms with both v and w in $\{3,4,5\}$ survive in the limit. Then, by assumptions (A1) and (B1) and applying Lemma 16, the variance converges to

$$\frac{1}{n^5}V\left[\sum_{v=1}^{5}\sum_{jk,lm\in\Theta_v}\xi_{jk}\xi_{lm}\text{vec}\left(\mathbf{x}_{jk}\mathbf{x}_{lm}^T\right)\right]\rightarrow\sum_{v=3}^{5}\sum_{w=3}^{5}\sum_{i=1}^{4}\alpha_i\beta_v\beta_w C(v,w)_i D_X(v,w)_i, \tag{B.77}$$

where we substitute the definition of $D_X(v,w)_i$ following (B.67). Thus, the matrix in (B.77) is positive semi-definite. Now, (B.74) becomes

$$n^3\left(MSE_\xi\left(\widehat{V}_{DC}\right)-MSE_\xi\left(\widehat{V}_E\right)\right)\xrightarrow{\mathbb{P}_X}$$
$$\xrightarrow{\mathbb{P}_X}\text{vec}\left(M_1^{-2}\right)^T\left(\sum_{v=3}^{5}\sum_{w=3}^{5}\sum_{i=1}^{4}\alpha_i\beta_v\beta_w C(v,w)_i D_X(v,w)_i\right)\text{vec}\left(M_1^{-2}\right)\geq 0. \tag{B.78}$$

□

B.5 Simulation study details

As noted in Section 3.4, 500 random realizations of covariates were generated for each sample size of actors $n\in\{20,40,80,160,320\}$. For each covariate realization, 1,000 random error realizations were generated for each of the three error settings: IID, exchangeable, and non-exchangeable. Using (3.11), a simulated data set was created from each covariate realization and error realization pair. The regression model was fit using ordinary least squares to each data set, and standard errors were estimated using the exchangeable, dyadic clustering, and heterskedasticity-consistent sandwich variance estimators. Confidence interval

Table B.1: Approximate standard deviations for exchangeable error setting.

σ_ϵ	σ_a	σ_b	σ_γ	σ_z
0.866	0.957	0.677	0.677	0.677

coverage was estimated *for each covariate realization* by counting the fraction of confidence intervals that contain the true coefficient.

For all simulations, we fixed true coefficients $\boldsymbol{\beta} = [1, 1, 1, 1]^T$. We drew each x_{2i} from a Bernoulli$(1/2)$ distribution independently. In the rare event that $x_{2i} = x_{2j}$ for all (i, j) pairs, one realization x_{2k} was randomly flipped to a 1 or 0. All x_{3i} and x_{4ij} were drawn independently from a standard normal distribution.

Each error setting was specified to have the same total variance:

$$\sum_{ij} \mathrm{Var}(\xi_{ij}) = 3n(n-1).$$

This variance was chosen so that the variance of the error would be similar to that of the regression mean model $\boldsymbol{\beta}^T \mathbf{x}_{ij}$. In the IID errors setting, $\xi_{ij} \sim_{iid} \mathrm{N}(0, 3)$ for all (i, j). To generate the non-exchangeable errors, a mean-zero random effect was added to the upper left quadrant of $\mathrm{V}[\boldsymbol{\Xi}_v]$. The errors for the non-exchangeable error setting may be written

$$\xi_{ij} = \tau \mathbf{1}_{i \leq \lfloor n/2 \rfloor} \mathbf{1}_{j \leq \lfloor n/2 \rfloor} + \epsilon_{ij}, \quad \tau \sim \mathrm{N}\left(0, \frac{9n}{4\lfloor n/2 \rfloor}\right), \quad \epsilon_{ij} \sim_{iid} \mathrm{N}(0, 3/4).$$

Finally, the distribution of the exchangeable (bilinear mixed effects model) error setting is defined in (3.7). We selected the dimension of the latent space to be $d = 2$, the correlation between sender and receiver effects as $\rho_{ab} = 1/2$, and the sender variance to be twice that of receiver variance: $\sigma_a^2 = 2\sigma_b^2$. We further specified $\sigma_z = \sigma_\gamma = \sigma_b$. Finally, we selected $\sigma_\epsilon^2 = 3/4$. With the aforementioned choices, the restriction $\sum_{ij} \mathrm{Var}(\xi_{ij}) = 3n(n-1)$ generated a quadratic equation in σ_b. The standard deviations that resulted from solving this quadratic equation are shown in Table B.1.

B.5.1 Confidence interval widths

To examine the relative confidence interval widths between the exchangeable and dyadic clustering sandwich variance estimators, it is sufficient to examine the values of the standard error estimates. In all

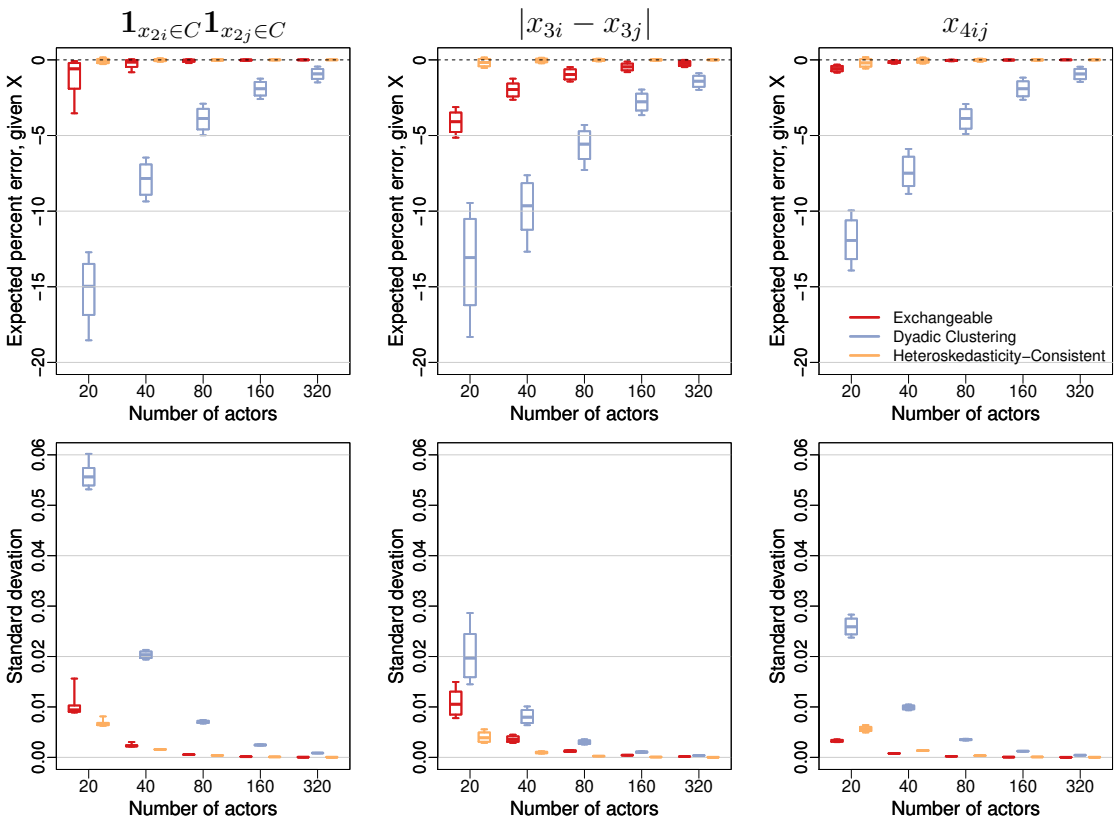


Figure B.1: (IID Errors) The top row of plots are the average differences in standard errors across random realizations of X, where the average is taken over $1,000$ error realizations. The bottom row of plots show the standard deviations of the standard error estimates across random X. Lines in the boxplots denote the median, the box denotes the middle 80% of values, and the whiskers denote the middle 95% of values.

simulations we generate 95% confidence intervals by using the typical normal approximation of plus or minus 1.96 times the standard error. We plot the empirical expected standard error given X relative to the true standard error given X in Figures B.1-B.3. We estimate the expectation by averaging the standard error estimates across the 1,000 error realizations, for each X realization. We also compute the standard deviation of the standard error estimates given X.

We observe that, for IID and exchangeable error structures in Figures B.1 and B.2, the standard errors resulting from the exchangeable estimator are much closer to the true standard errors than those resulting from the dyadic clustering estimator. This fact suggests that the dyadic clustering estimator fails to account for a portion of the dependency in the error structure. We note that both procedures generally produce underestimates of the true standard errors, however, the dyadic clustering estimator trades some efficiency for robustness. We observe that the standard deviation of the standard error estimates when us-

ing the exchangeable estimator are typically lower than those when using dyadic clustering under IID and exchangeable errors. Intuitively, the lower variability of the exchangeable estimator relative to the dyadic clustering estimator the result of the averaging present in the exchangeable estimator. Finally, the trends of larger expectation and smaller standard deviation are present for most of the realizations of X under non-exchangeable errors (Figure B.3). This is true despite the fact that we might expect the dyadic clustering estimator to account for the heterogenous, non-exchangeable error structure more effectively than the exchangeable estimator since the dyadic clustering estimator is claimed to be robust.

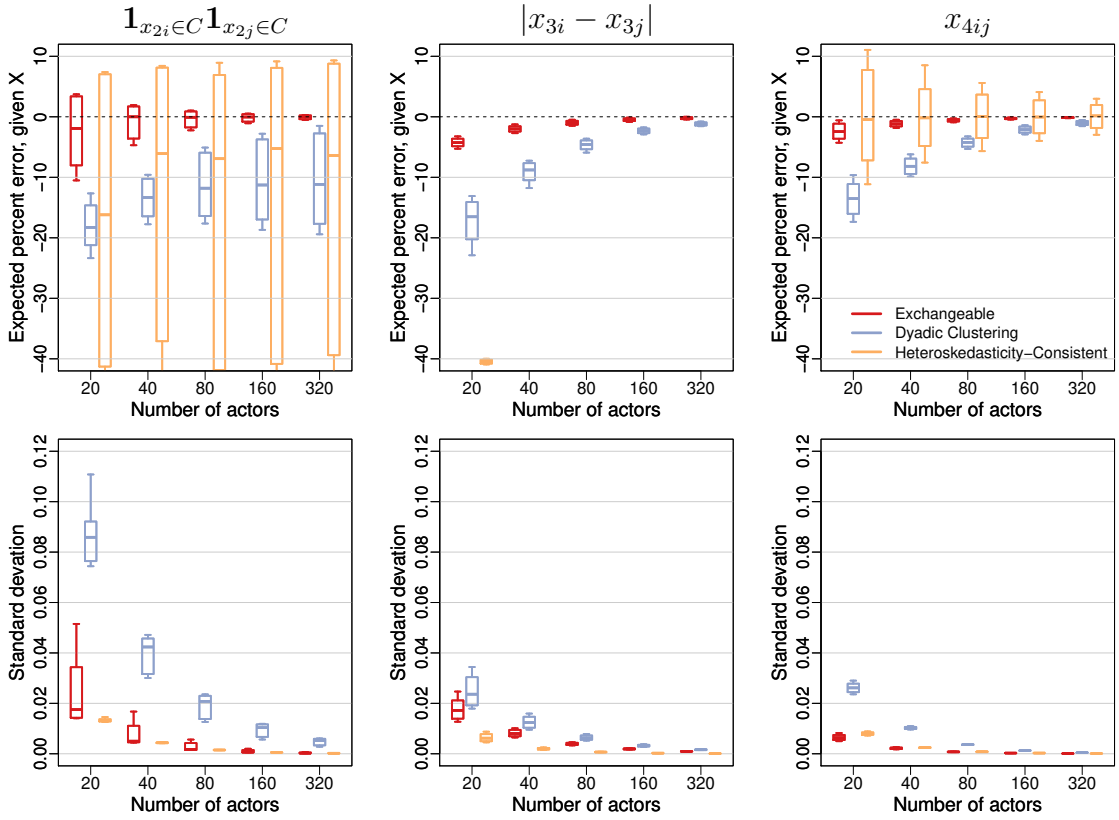


Figure B.2: (Exchangeable Errors) The top row of plots are the average differences in standard errors across random realizations of X, where the average is taken over $1,000$ error realizations. The bottom row of plots show the standard deviations of the standard error estimates across random X. Lines in the boxplots denote the median, the box denotes the middle 80% of values, and the whiskers denote the middle 95% of values. The ordinate axis is truncated where appropriate to show the estimators of interest.

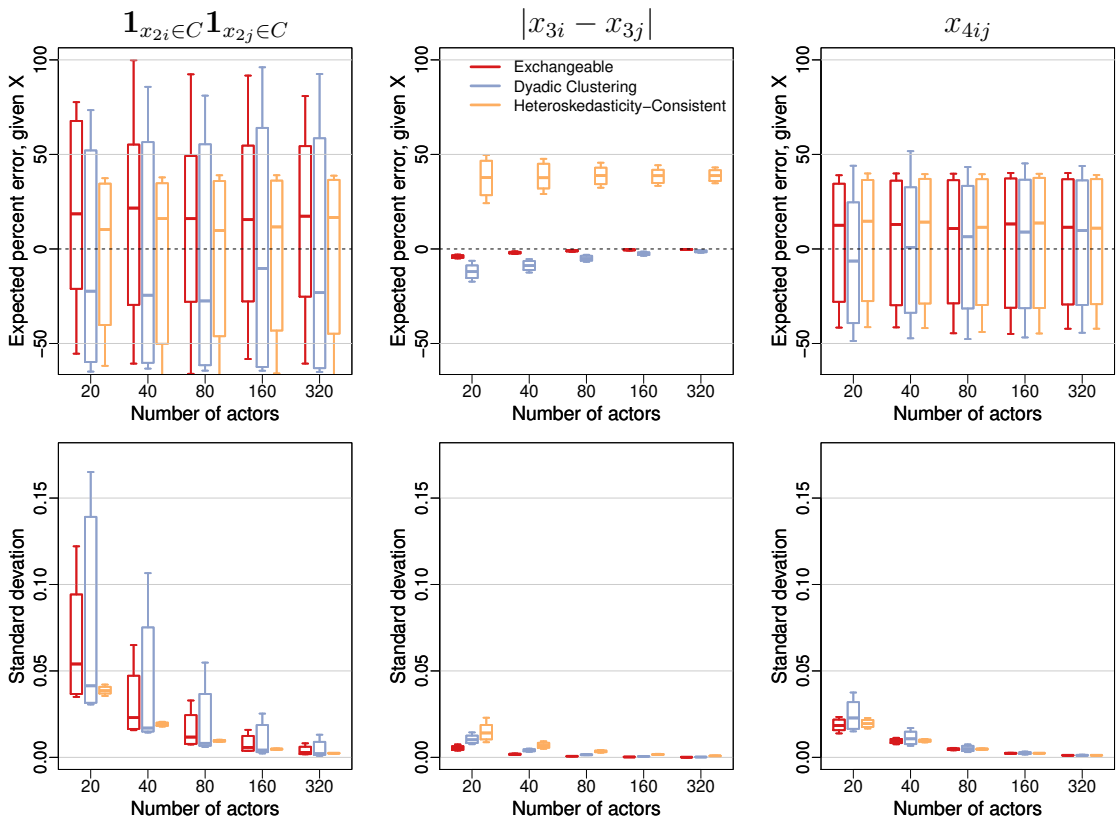


Figure B.3: (Non-exchangeable Errors) The top row of plots are the average differences in standard errors across random realizations of X, where the average is taken over $1,000$ error realizations. The bottom row of plots show the standard deviations of the standard error estimates across random X. Lines in the boxplots denote the median, the box denotes the middle 80% of values, and the whiskers denote the middle 95% of values. The ordinate axis is truncated where appropriate to show the estimators of interest.

B.6 DC covariance matrix invertibility

Ideally, for a covariance matrix estimate $\widehat{\Omega}$ to be of utmost utility, it must be invertible. For example, if we wish to reweight the estimating equations, as in GEE, and solve iteratively for both the variance matrix and regression coefficients simultaneously, the estimate of the the covariance matrix must be nonsingular. However, in many cases the DC estimator is singular and hence cannot be used as a reweighting matrix. In cases when the DC estimator is singular, it can still be used in the 'meat' (B matrix) in the coefficient sandwich estimator covariance matrix.

Theorem 18. *The dyadic clustering estimate of the error variance, $\widehat{\Omega}_{DC}$, is singular for directed data.*

Proof. The DC estimator can be written as the Hadamard product between the outer product of the residuals and a matrix of indicators of whether the dyad indices share a member.

$$\widehat{\Omega}_{DC} = ee^T \circ \mathbf{1}_{[\{i,j\}\cap\{k,l\}\neq\varnothing]} \tag{B.79}$$

The rank of the outer product of the residuals is one: $\text{rank}(ee^T) = 1$. The rank of the indicator matrix is at most $n(n-1)/2$, since the indices (i,j) share a member with an arbitrary pair (k,ℓ) if and only if the indices (j,i) do as well. Thus, the column of $\mathbf{1}_{[\{i,j\}\cap\{k,l\}\neq\varnothing]}$ corresponding to (i,j) is the same as that corresponding to (j,i).

For any two square matrices of equal size A and B, $\text{rank}(A \circ B) \leq \text{rank}(A)\text{rank}(B)$. Using the representation of $\widehat{\Omega}_{DC}$ in (B.79), we have that

$$\begin{aligned}\text{rank}(\widehat{\Omega}_{DC}) &\leq \text{rank}(ee^T)\text{rank}(\mathbf{1}_{[\{i,j\}\cap\{k,l\}\neq\varnothing]}) \\ &\leq \frac{n(n-1)}{2}\end{aligned}$$

$\widehat{\Omega}_{DC}$ is therefore not full rank. □

Remark 19. *Theorem 18 does not hold for undirected data when $R = 1$. If the data are undirected, then the bound does not guarantee singularity of $\widehat{\Omega}_{DC}$ since the dimension of $\widehat{\Omega}_{DC}$ is exactly $n(n-1)/2$. In practice, we find that $\widehat{\Omega}_{DC}$ is full rank in this special case.*

Remark 20. *The result of Theorem 18 holds for both directed and undirected data when $R > 1$. In this case, the column in the indicator matrix $\mathbf{1}_{[\{i,j\}\cap\{k,l\}\neq\varnothing]}$ corresponding to the indices (i,j,s) is the same as that column corresponding to (i,j,t) for all values of $t \in \{1,...,R\}$. Thus, again $\widehat{\Omega}_{DC}$ is not full rank.*

B.7 Efficient inversion of the exchangeable covariance matrix

To perform the GEE procedure as described in Section 3.6, we must invert the exchangeable covariance matrix Ω_E as defined in Figure 3.2. For now, we work in the case where $R = 1$. Since Ω_E is a real symmetric matrix, its inverse is real and symmetric as well. However, we can say more about the patterns in the inverse Ω_E^{-1}. Recall that Ω_E has at most six unique terms; call these parameters ϕ. We find that the inverse Ω_E^{-1} has at most six unique terms as well. If we define the parameters in Ω_E^{-1} as $\boldsymbol{p}$, we can write

$$\Omega_E(\boldsymbol{\phi})\Omega_E^{-1}(\boldsymbol{p}) = I \;\text{ for }\; \boldsymbol{\phi}, \boldsymbol{p} \in \mathbb{R}^6 \tag{B.80}$$

where I is the $n(n-1) \times n(n-1)$ identity. Lastly, we make the conjecture that the parameter pattern in Ω_E^{-1} is exactly the same as that in Ω_E; we find this conjecture to be true in practice. One caveat is that the locations in which we assume zeros in Ω_E are not zero in Ω_E^{-1} in general.

We can find the inverse parameters $\boldsymbol{p}$ from $\boldsymbol{\phi}$ without inverting the entire matrix Ω_E by instead solving the following linear system

$$C(\boldsymbol{\phi}, n)\boldsymbol{p} = [1,0,0,0,0,0]^T \text{ for } C(\boldsymbol{\phi}, n) \in \mathbb{R}^{6\times 6}, \tag{B.81}$$

where $C(\boldsymbol{\phi}, n)$ is a set of six linear equations based on the parameters $\boldsymbol{\phi}$ and the number of actors n and is depicted in Figure B.4. Thus, we replace the need to invert the $n(n-1) \times n(n-1)$ matrix Ω_E by the inversion of the 6×6 matrix $C(\boldsymbol{\phi}, n)$. Using this procedure, the computational cost associated with finding the inverse of Ω_E is independent of the number of actors n.

Now consider the case of array data with $R > 1$. Inversion of the exchangeable covariance matrices $\Omega = \mathrm{V}[\boldsymbol{\Xi}_{\boldsymbol{v}}]$ in Figure 3.6 requires consideration of the patterns in the block matrices. Focusing on Figure 3.6(a), note that Ω_E is parametrized by twelve terms. We denote the first six parameters as $\boldsymbol{\phi}^{(1)}$ and the second six $\boldsymbol{\phi}^{(2)}$, corresponding to Ω_1 and Ω_2 respectively. Again the inverse Ω^{-1} has the exact same block matrix pattern as Ω. Thus, the inverse may be parametrized by $\boldsymbol{p}^{(1)}$ and $\boldsymbol{p}^{(2)}$, each with length six, defined by the following linear equations.

$$\begin{aligned} C(\boldsymbol{\phi}^{(1)}, n)\boldsymbol{p}^{(1)} + (R-1)C(\boldsymbol{\phi}^{(2)}, n)\boldsymbol{p}^{(2)} &= [1,0,0,0,0,0]^T \\ C(\boldsymbol{\phi}^{(2)}, n)\boldsymbol{p}^{(1)} + C(\boldsymbol{\phi}^{(1)}, n)\boldsymbol{p}^{(2)} + (R-2)C(\boldsymbol{\phi}^{(2)}, n)\boldsymbol{p}^{(2)} &= 0_{6\times 1} \end{aligned} \tag{B.82}$$

This is twelve linear equations in $\boldsymbol{p}^{(1)}$ and $\boldsymbol{p}^{(2)}$. In this formulation we reduce a $Rn(n-1) \times Rn(n-1)$ inversion to a 12×12 inversion for calculation of Ω_E^{-1}. Again, note that there is no dependence of the complexity of the inversion on the array dimensions n and R. The inverses of the other possible exchangeable covariance matrices in Figure 3.6, while more complex, can be calculated using a similar procedure that again omits dependence on array dimension n.

B.8 Eigenvalues of exchangeable covariance matrix

Since the entries in the exchangeable covariance matrix estimator $\widehat{\Omega}_E$ are empirical averages, it is possible the estimate is not positive definite. Here we briefly investigate the constraints on the parameters that guarantee the resulting covariance matrix is positive definite for $R = 1$. Note that for computing the sandwich estimator variance of $\widehat{\boldsymbol{\beta}}$ and making inference on $\widehat{\boldsymbol{\beta}}$, positive definiteness of $\widehat{\Omega}_E$ is not necessary. However, if a GEE procedure is employed, the inverse of the covariance matrix estimator is required, and hence positive definiteness of $\widehat{\Omega}_E$ is desired.

$$\begin{bmatrix}
\phi_1 & \phi_2 & (n-2)\phi_3 & (n-2)\phi_4 & 2(n-2)\phi_5 & (n-2)(n-3)\phi_6 \\
\phi_2 & \phi_1 & (n-2)\phi_5 & (n-2)\phi_5 & (n-2)(\phi_3+\phi_4) & (n-2)(n-3)\phi_6 \\
\phi_3 & \phi_5 & \phi_1+(n-3)\phi_3 & \phi_5+(n-3)\phi_6 & \phi_2+\phi_4+(n-3)(\phi_5+\phi_6) & (n-3)\left(\phi_4+\phi_5+(n-4)\phi_6\right) \\
\phi_4 & \phi_5 & \phi_5+(n-3)\phi_6 & \phi_1+(n-3)\phi_4 & \phi_2+\phi_3+(n-3)(\phi_5+\phi_6) & (n-3)\left(\phi_3+\phi_5+(n-4)\phi_6\right) \\
\phi_5 & \phi_4 & \phi_2+(n-3)\phi_5 & \phi_3+(n-3)\phi_6 & \phi_1+\phi_5+(n-3)(\phi_4+\phi_6) & (n-3)\left(\phi_3+\phi_5+(n-4)\phi_6\right) \\
\phi_6 & \phi_6 & \phi_4+\phi_5+(n-4)\phi_6 & \phi_3+\phi_5+(n-4)\phi_6 & \phi_3+\phi_4+2\phi_5+2(n-4)\phi_6 & \phi_1+\phi_2+(n-4)\left(\phi_3+\phi_4+2\phi_5+(n-5)\phi_6\right)
\end{bmatrix}$$

Figure B.4: Matrix $C(\boldsymbol{\phi}, n)$.

B.8.1 Undirected relational data

We focus first on the undirected case, where the exchangeable covariance matrix contains two distinct nonzero entries: a variance σ^2 and a parameter ϕ in the off-diagonal representing the correlation between any pairs of relations that share an actor. Below we consider the correlation matrix, rather than the covariance matrix, which contains only nonzero correlation value. We denote this value by a, and note that $a = \phi/\sigma^2$.

Based on a thorough empirical investigation, we conjecture that the exchangeable correlation matrix corresponding to an undirected set of relations among n actors, which has nonzero value a in select off-diagonal entries, has exactly three eigenvalues as given below.

Eigenvalue	Multiplicity
$1 + 2(n-2)a$	1
$1 - 2a$	$\frac{1}{2}n(n-3)$
$1 + (n-4)a$	$n-1$

The correlation matrix is positive definite if and only if all eigenvalues are positive. Thus, if $a \in \left(\frac{-1}{2(n-2)}, \frac{1}{2}\right)$, the correlation matrix is positive definite. Notice that the upper bound on a does not vary with n. Using the relation between a and $\{\sigma^2, \phi\}$, this constraint can be re-expressed as a constraint on the covariance parameters.

B.8.2 Directed relational data

We find empirically that the directed covariance matrix Ω_E has five unique eigenvalues. Further, each of the eigenvalues are contained within the set of six eigenvalues of the matrix C, defined in Figure B.4 and used in computation of the inverse of the exchangeable covariance matrix. As C is a bilinear function of Ω_E, this observation does not appear implausible. We may construct $C = A^T \Omega_E B$ for $A, B \in \mathbb{R}^{n(n-1)\times 6}$ and $A^T B = I$. One such pair is B taken to be the first column of $\mathcal{S}_1$ thought $\mathcal{S}_6$ and A taken to be all zeros except for a single 1 in each column occupying rows $\{1, n, 2n, 2, n+1, n(n-1)\}$, respectively.

In analyzing the eigenvalues of the directed covariance matrix Ω_E, we again focus on the exchangeable correlation matrix which contains four nonzero off-diagonal elements $\{a, b, c, d\}$ corresponding, respectively, in placement to $\{\phi_a, \phi_b, \phi_c, \phi_d\}$ in the exchange covariance matrix Ω_E. Note $a = \phi_a/\sigma^2$, $b = \phi_b/\sigma^2$, and so on. Based on an eigenvalue analysis of C and various empirical studies, we conjecture the eigenval-

ues for the exchangeable correlation matrix associated with a directed set of relations among n actors has exactly five eigenvalues as given below.

Eigenvalue(s)	Multiplicity
$1 + a + (n-2)(b+c) + 2(n-2)d$	1
$1 + a - (b + c + 2d)$	$(n-1)(n-2)/2 - 1$
$1 - (a + b + c) + 2d$	$(n-1)(n-2)/2$
$((n-3)(b+c) - 2d + 2)/2 \pm \sqrt{(\alpha+\beta)}/2$	$n-1$

where $\alpha = (c^2 + b^2)(n^2 - 2n + 1) + 4d^2(n^2 - 6n + 9) + 2bc(1 - n^2 + 2n)$ and $\beta = ad(8n - 24) + (b + c)d(12 - 4n) + 4a(a - (b + c))$. As in the undirected case, these constraints can be re-expressed as constraints on the original five covariance parameters.

B.9 Trade data prediction study

To compare the ability of the proposed exchangeable GEE model and the model from Westveld and Hoff (2011) – which we refer to as the hierarchical, longitudinal mixed effects model (HLMEM)– to represent the trade data, we examined the out-of-sample predictive performance of the estimators. First, we establish some notation for this section. Recall that the trade data set has covariate measures in X that vary by year. Thus, we rewrite the linear model in (1.1) as

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \boldsymbol{\xi}_t, \quad t \in \{1, 2, \ldots, T\}$$

where $\mathbf{y}_t$ represents the $n(n-1)$ vector of relations among the $n = 58$ countries in year t, $\mathbf{X}_t$ is a matrix eight covariates corresponding to year t, and $\boldsymbol{\xi}_t$ is a vector of errors for year t.

We estimated both models on the first 10 and 19 years of data, and then used these estimates to predict the trade at $T = 11$ and $T = 20$. To generate predictions from each model, we computed the conditional expectation $E[\mathbf{y}_T \mid \{\mathbf{y}_t\}_{t=1}^{T-1}]$ based on the assumption that $\mathbf{y}_T$ and $\{\mathbf{y}_t\}_{t=1}^{T-1}$ are jointly normal. The exchangeable GEE model and HLMEM correspond to different models of the variance-covariance matrices of $\boldsymbol{\xi}_t$ and the covariances matrices between vectors $\boldsymbol{\xi}_t$ and $\boldsymbol{\xi}_{t+h}$. As a baseline, we included ordinary least squares (OLS), assuming independence of each year, i.e. $\boldsymbol{\xi}_t$ independent of $\boldsymbol{\xi}_{t+h}$.

We first discuss the OLS estimation and prediction procedure used for this study. We estimated the coefficients in the prediction time period, $\boldsymbol{\beta}_T$, with the coefficients from the previous time period and assumed independent and identically distributed entries in all $\boldsymbol{\xi}_t$. Thus, the OLS estimator of the trade in year T is

$$E[\mathbf{y}_T \mid \{\mathbf{y}_t\}_{t=1}^{T-1}]_{(OLS)} = \mathbf{X}_T \widehat{\boldsymbol{\beta}}_{T-1}.$$

For the GEE procedure, we again set $\widehat{\boldsymbol{\beta}}_T = \widehat{\boldsymbol{\beta}}_{T-1}$. Based on the model underlying the GEE procedure (Figure 6(a)), the variance $V(\mathbf{y}_t) = \boldsymbol{\Omega}_1$ for all $t \in \{1, 2, \ldots, T\}$ and the covariance $Cov(\mathbf{y}_t, \mathbf{y}_{t+h}) = \boldsymbol{\Omega}_2$ for all h. Further, it can be shown that the precision corresponding to the concatenated vector $\mathbf{z}_{T-1}^T := [\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_{T-1}^T]$ is of the same pattern as the variance $V(\mathbf{z}_{T-1})$, which has $\boldsymbol{\Omega}_1$ along the diagonal blocks and $\boldsymbol{\Omega}_2$ in the off-diagonal blocks (as in Figure 6(a)). We define the diagonal blocks of $V(\mathbf{z}_{T-1})^{-1}$ as $\boldsymbol{\Psi}_1$ and the off-diagonal blocks $\boldsymbol{\Psi}_2$. Then, when the relations $\{\mathbf{y}_t\}_{t=1}^T$ are jointly normally distributed, we define the prediction from the GEE procedure is

$$E[\mathbf{y}_T \mid \{\mathbf{y}_t\}_{t=1}^{T-1}]_{(GEE)} = \mathbf{X}_T \widehat{\boldsymbol{\beta}}_{T-1} + \boldsymbol{\Omega}_2(\boldsymbol{\Psi}_1 + (T-2)\boldsymbol{\Psi}_2) \sum_{t=1}^{T-1} \left(\mathbf{y}_t - \mathbf{X}_t \widehat{\boldsymbol{\beta}}_t \right).$$

Finally, to detail predictions from the HLMEM, we first review some terms used in Westveld and Hoff (2011). Their proposed model is

$$y_{ij,t} = \mathbf{x}_{ij,t}^T \boldsymbol{\beta}_t + s_{i,t} + r_{j,t} + g_{ij,t},$$

where $s_{i,t}$ and $r_{j,t}$ are sender and receiver random effects, respectively, and $g_{ij,t}$ is a reciprocal random effect. These effects evolve according to autoregressive order one processes, such that

$$\begin{bmatrix} s_{i,t} \\ r_{i,t} \end{bmatrix} = \boldsymbol{\Phi}_{sr} \begin{bmatrix} s_{i,t-1} \\ r_{i,t-1} \end{bmatrix} + \boldsymbol{\epsilon}_{i,t}, \quad i \in \{1, 2, \ldots, n\}, t \in \{1, 2, \ldots, T-1\},$$

$$\begin{bmatrix} g_{ij,t} \\ g_{ji,t} \end{bmatrix} = \boldsymbol{\Phi}_g \begin{bmatrix} g_{ij,t-1} \\ g_{ji,t-1} \end{bmatrix} + \mathbf{e}_{ij,t}, \quad i, j \in \{1, 2, \ldots, n\}, i \neq j, t \in \{1, 2, \ldots, T-1\},$$

where $\boldsymbol{\Phi}_{sr}$ is a 2×2 autoregressive matrix and $\boldsymbol{\Phi}_g$ is a symmetric autoregressive matrix. The error terms $\boldsymbol{\epsilon}_{i,t}$ and $\mathbf{e}_{ij,t}$ are mean-zero independent bivariate Gaussian random variables. Please see Westveld and Hoff

(2011) for more details of the model. With the notation and model defined, and again setting $\widehat{\boldsymbol{\beta}}_T = \widehat{\boldsymbol{\beta}}_{T-1}$, it is easy to see that the prediction from the HLMEM is

$$E[y_{ij,T} \mid \{\mathbf{y}_t\}_{t=1}^{T-1}]_{(WH)} = \mathbf{x}_{ij,T}^T \widehat{\boldsymbol{\beta}}_{T-1} + \begin{bmatrix} 1 & 0 \end{bmatrix}^T \widehat{\boldsymbol{\Phi}}_{sr} \begin{bmatrix} \widehat{s}_{i,T-1} \\ \widehat{r}_{i,T-1} \end{bmatrix} + \begin{bmatrix} 0 & 1 \end{bmatrix}^T \widehat{\boldsymbol{\Phi}}_{sr} \begin{bmatrix} \widehat{s}_{j,T-1} \\ \widehat{r}_{j,T-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \end{bmatrix}^T \widehat{\boldsymbol{\Phi}}_{g} \begin{bmatrix} \widehat{g}_{ij,T-1} \\ \widehat{g}_{ji,T-1} \end{bmatrix}.$$

To estimate the HLMEM, Westveld and Hoff (2011) generated a Markov chain of length 55,000, of which the first 10,000 were discarded and every 20$^{\text{th}}$ iteration saved, giving 2,250 samples to approximate the posterior distributions of $\boldsymbol{\beta}_{T-1}$, Φ_{sr}, Φ_g, and the random effects at $t = T-1$ in the above equations. The mean of these 2,250 samples from the joint posterior distribution, for each pair $i \neq j$, was used to construct $E[\mathbf{y}_T \mid \{\mathbf{y}_t\}_{t=1}^{T-1}]_{(HLMEM)}$.

We evaluate performance in the predicted year using the mean square prediction error, defined

$$MSPE = \frac{1}{n(n-1)} \left|\left| E[\mathbf{y}_T | \{\mathbf{y}_t\}_{t=1}^{T-1}] - \mathbf{y}_T \right|\right|_2^2,$$

where the expectation is replaced by one of the three prediction estimators.

Appendix C

Regression of binary network data with exchangeable latent errors

C.1 Details of estimation

In this section we supply details of estimation in support of Algorithm 2, beginning with the initialization of ρ. We then provide details of the maximization of $\ell_{\mathbf{y}}$ with respect to $\boldsymbol{\beta}$, the approximations of maximizing $\ell_{\mathbf{y}}$ with respect to ρ, and the handling of missing data in the BC-EM algorithm.

C.1.1 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 $\boldsymbol{\Omega}$, it can be shown that ρ lies in the interval $[0, 1/2)$ when $\boldsymbol{\Omega}$ 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 $\mathcal{A}$ of Θ_2 of size $2n^2$,and estimating ρ using the values of $\mathcal{A}$. 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:

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

We compute the average $\frac{1}{|\mathcal{A}|} \sum_{jk,lm \in \mathcal{A}} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ using the linearization approach described in Section C.1.3.

C.1.2 Implementation of β expectation step

Under general correlation structure, computation of the expectation $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}]$ (step 1.1 in Algorithm 2, 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 $\boldsymbol{\Omega}$ to compute $E[\boldsymbol{\epsilon} \,|\, \mathbf{y}]$ 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

$$E[\epsilon_{jk} \,|\, \mathbf{y}] = E[E[\epsilon_{jk} \,|\, \boldsymbol{\epsilon}_{-jk}, y_{jk}] \,|\, \mathbf{y}], \tag{C.2}$$

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 $\boldsymbol{\Omega}^{-1} = \sum_{i=1}^{3} p_i \mathcal{S}_i$, we may write the untruncated distribution directly as

$$\begin{aligned} \epsilon_{jk} \,|\, \boldsymbol{\epsilon}_{-jk} &\sim \mathrm{N}(\mu_{jk}, \sigma_n^2), \\ \mu_{jk} &= -\sigma_n^2 \mathbf{1}_{jk}^T \left(p_2 \mathcal{S}_2 + p_3 \mathcal{S}_3\right) \widetilde{\boldsymbol{\epsilon}}_{-jk}, \\ \sigma_n^2 &= \frac{1}{p_1}, \end{aligned} \tag{C.3}$$

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 $p_2\mathcal{S}_2 + p_3\mathcal{S}_3$ consists of all zeros so that μ_{jk} is free of ϵ_{jk}.

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

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

where $\widetilde{\eta} := (\eta + \mu)/\sigma$. Now, taking $z = (\epsilon_{jk} \,|\, \boldsymbol{\epsilon}_{-jk})$ and defining $\widetilde{\mu}_{jk} := (\mu_{jk} + \mathbf{x}_{jk}^T \boldsymbol{\beta})/\sigma_n$, we have that

$$E[\epsilon_{jk} \,|\, \boldsymbol{\epsilon}_{-jk}, y_{jk}] = \mu_{jk} + \sigma_n \left(\frac{\phi(\widetilde{\mu}_{jk}) \left(y_{jk} - \Phi(\widetilde{\mu}_{jk})\right)}{\Phi(\widetilde{\mu}_{jk})(1 - \Phi(\widetilde{\mu}_{jk}))} \right). \tag{C.5}$$

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

$$E[\epsilon_{jk} \,|\, \mathbf{y}] = -\sigma_n^2 \mathbf{1}_{jk}^T \left(p_2 \mathcal{S}_2 + p_3 \mathcal{S}_3\right) E[\boldsymbol{\epsilon} \,|\, \mathbf{y}] + \sigma_n E\left[\frac{\phi(\widetilde{\mu}_{jk}) \left(y_{jk} - \Phi(\widetilde{\mu}_{jk})\right)}{\Phi(\widetilde{\mu}_{jk})(1 - \Phi(\widetilde{\mu}_{jk}))} \,\Big|\, \mathbf{y} \right]. \tag{C.6}$$

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

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

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

$$\mathbf{w}_0 := \frac{\phi(\mathbf{X}\boldsymbol{\beta}) \left(\mathbf{y} - \Phi(\mathbf{X}\boldsymbol{\beta})\right)}{\Phi(\mathbf{X}\boldsymbol{\beta})(1 - \Phi(\mathbf{X}\boldsymbol{\beta}))}. \tag{C.8}$$

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

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

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

$$\frac{\partial}{\partial \mathbf{w}^T} g(\mathbf{w}) = \mathbf{B} - \mathbf{I} + \mathbf{D}\mathbf{B}. \tag{C.10}$$

where we define

$$\mathbf{D} = \operatorname{diag} \left\{ \frac{-w_{jk}\phi(w_{jk})(y - \Phi(w_{jk})) - \phi(w_{jk})^2 - \phi(w_{jk})^2 (y - \Phi(w_{jk}))(1 - 2\phi(w_{jk})\Phi(w_{jk}))}{\Phi(w_{jk})(1 - \Phi(w_{jk}))} \right\}_{jk}.$$

The term $\mathbf{DB}$ arises from differentiating $v(\mathbf{w})$ with respect to $\mathbf{w}$. Using the expression in (C.10), we are then able to write the second term in (C.9) as

$$\left(\frac{\partial}{\partial \mathbf{w}^T} g(\widehat{\mathbf{w}})\right)^{-1} g(\widehat{\mathbf{w}}) = (\mathbf{B} - \mathbf{I} + \mathbf{DB})^{-1} \left((\mathbf{B} - \mathbf{I})\mathbf{w} + v(\mathbf{w})\right), \tag{C.11}$$

$$= \mathbf{B}^{-1} \left(\mathbf{I} + \mathbf{D} - \mathbf{B}^{-1}\right)^{-1} \left((\mathbf{B} - \mathbf{I})\mathbf{w} + v(\mathbf{w})\right). \tag{C.12}$$

We notice that the matrix $\mathbf{I} + \mathbf{D}$ is diagonal, but not homogeneous (in which case we compute (C.12) 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

$$\left(\mathbf{I} + \mathbf{D} - \mathbf{B}^{-1}\right)^{-1} = (\mathbf{Q} + \mathbf{M})^{-1} \approx \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}^{-1}, \tag{C.13}$$

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 C.2 (in $O(1)$ time), the following approximation represents an improvement in computation from $O(n^3)$ to $O(n^2)$ time:

$$\left(\frac{\partial}{\partial \mathbf{w}^T} g(\widehat{\mathbf{w}})\right)^{-1} g(\widehat{\mathbf{w}}) \approx \mathbf{B}^{-1} \left(\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}^{-1}\right) \left((\mathbf{B} - \mathbf{I})\mathbf{w} + v(\mathbf{w})\right). \tag{C.14}$$

C.1.3 Approximation to ρ expectation step

The EM update for ρ in relies on the computation of $\gamma_i = E[\boldsymbol{\epsilon}^T \mathbf{S}_i \boldsymbol{\epsilon} \,|\, \mathbf{y}]/|\Theta_i|$, for $i \in \{1, 2, 3\}$ (step 2.2 in Algorithm 2). 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.

Pairwise expectation

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

$$\gamma_1 = \frac{1}{|\Theta_1|}\sum_{jk} E[\epsilon_{jk}^2 \,|\, \mathbf{y}] \approx \frac{1}{|\Theta_1|}\sum_{jk} E[\epsilon_{jk}^2 \,|\, y_{jk}], \tag{C.15}$$
$$\gamma_2 = \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, \mathbf{y}] \approx \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}],$$
$$\gamma_3 = \frac{1}{|\Theta_3|}\sum_{jk,lm\in\Theta_3} E[\epsilon_{jk}\epsilon_{lm} \,|\, \mathbf{y}] \approx \frac{1}{|\Theta_3|}\sum_{jk,lm\in\Theta_3} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}],$$

where Θ_i is the set of ordered pairs of relations (jk, lm) which correspond entries in $\mathbf{S}_i$ that are 1, for $i \in \{1, 2, 3\}$. These approximations are natural first-order approximations: recalling that $y_{jk} = \mathbb{1}[\epsilon_{jk} > -\mathbf{x}_{jk}^T\boldsymbol{\beta}]$, the approximations in (C.15) are based on the notion that knowing the domains of ϵ_{jk} and ϵ_{lm} is significantly more informative for $E[\epsilon_{jk}\epsilon_{lm} \,|\, \mathbf{y}]$ than knowing the domain of, for example, ϵ_{ab}.

The approximations in (C.15) are orders of magnitude faster to compute than the full expectations $E[\epsilon_{jk}\epsilon_{lm} \,|\, \mathbf{y}]$. In particular, when $i \in \{1, 3\}$, the expectations are available in closed form:

$$E[\epsilon_{jk}^2 \,|\, y_{jk}] = 1 - \eta_{jk}\frac{\phi(\eta_{jk})(y_{jk} - \Phi(\eta_{jk}))}{\Phi(\eta_{jk})(1 - \Phi(\eta_{jk}))}, \tag{C.16}$$
$$E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}] = \frac{\phi(\eta_{jk})\phi(\eta_{lm})(y_{jk} - \Phi(\eta_{jk}))(y_{lm} - \Phi(\eta_{lm}))}{\Phi(\eta_{jk})\Phi(\eta_{lm})(1 - \Phi(\eta_{jk}))(1 - \Phi(\eta_{lm}))}, \quad |\{j, k\} \cap \{l, m\}| = 0, \tag{C.17}$$

where we define $\eta_{jk} = \mathbf{x}_{jk}^T\beta$, $\Phi(.)$ is the standard normal cumulative distribution function, and $\phi(.)$ is the standard normal probability distribution function. The approximation to the expectations for $i = 2$ in (C.15) is

$$E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}] = \rho\left(1 - \frac{\bar{\eta}_{jk}\phi(\eta_{jk})}{L_{jk,lm}}\Phi\left(\frac{\bar{\eta}_{lm} - \bar{\rho}\,\bar{\eta}_{jk}}{\sqrt{1-\rho^2}}\right) - \frac{\bar{\eta}_{lm}\phi(\eta_{lm})}{L_{jk,lm}}\Phi\left(\frac{\bar{\eta}_{jk} - \bar{\rho}\,\bar{\eta}_{lm}}{\sqrt{1-\rho^2}}\right)\right) \tag{C.18}$$
$$+ \frac{1}{L_{jk,lm}}\sqrt{\frac{1-\rho^2}{2\pi}}\phi\left(\sqrt{\frac{\eta_{jk}^2 + \eta_{lm}^2 - 2\rho\,\eta_{jk}\eta_{lm}}{1-\rho^2}}\right), \quad |\{j, k\} \cap \{l, m\}| = 1,$$
$$L_{jk,lm} = \mathbb{P}\left((2y_{jk} - 1)\epsilon_{jk} > -\eta_{jk} \cap (2y_{lm} - 1)\epsilon_{lm} > -\eta_{lm}\right),$$

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

Linearization

The computation of $E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ in (C.18) 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 C.1.3 above, $\gamma_2 \approx \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ is approximately linear in ρ. This linearity is somewhat intuitive, as the expectation of $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ is equal to ρ, and is thus linear functions of ρ. As $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ is an average and concentrates around its expectation, it concentrates around a linear function of ρ, and it is reasonable to approximate $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ as a linear function of ρ. To do so, we compute the approximate values of γ_2 if $\rho = 0$ and if $\rho = 1$. In particular,

$$
\begin{aligned}
&\gamma_2 \approx a_2 + b_2\rho, \qquad\qquad (C.19)\\
&a_2 = \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk} \,|\, y_{jk}]E[\epsilon_{lm} \,|\, y_{lm}],\\
&\quad = \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} \frac{\phi(\eta_{jk})\phi(\eta_{lm})(y_{jk}-\Phi(\eta_{jk}))(y_{lm}-\Phi(\eta_{lm}))}{\Phi(\eta_{jk})\Phi(\eta_{lm})(1-\Phi(\eta_{jk}))(1-\Phi(\eta_{lm}))},\\
&c_2 = \frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]\,\Big|_{\rho=1},\\
&b_2 = c_2 - a_2.
\end{aligned}
$$

To compute c_2, we must compute the value of $E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ when $\rho = 1$. Computing $E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ is simple when the values $y_{jk} = y_{lm}$, as in this case $E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}] = E[\epsilon_{jk}^2 \,|\, y_{jk} = y_{lm}]$ 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 dersired expectation may be approximated $E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}] \approx E[\epsilon_{jk}^2 \,|\, \epsilon_{jk} \in U_{jk} \cap U_{lm}]$. 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

$$
E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}] \approx E[\epsilon_{jk}^2 \,|\, \epsilon_{jk} \in U_{jk}]\,\mathbb{P}(\epsilon_{jk} \in U_{jk}) + E[\epsilon_{lm}^2 \,|\, \epsilon_{lm} \in U_{lm}]\,\mathbb{P}(\epsilon_{lm} \in U_{lm}). \qquad (C.20)
$$

To summarize, we compute $c_2 = E[\epsilon_{jk}\epsilon_{lm} \,|\, \mathbf{y}] \Big|_{\rho=1}$ in (C.19) by using the following approximation:

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

C.1.4 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.

Update $\boldsymbol{\beta}$:

To maximize $\ell_{\mathbf{y}}$ with respect to $\boldsymbol{\beta}$ (Step 1 of Algorithm 2) 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 $\frac{1}{|\mathcal{M}^c|} \sum_{lm \in \mathcal{M}^c} x_{lm}^{(1)}$, where the superscript (1) refers to the first entry in $\mathbf{x}_{jk}$ and $\mathcal{M}$ 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 $\bar{\eta} = \frac{1}{|\mathcal{M}^c|} \sum_{lm \in \mathcal{M}^c} \mathbf{x}_{lm}^T \widehat{\boldsymbol{\beta}}$ and we compute $\mathbf{w} = E[\boldsymbol{\epsilon} \,|\, \mathbf{y}]$ using the procedure in Section C.1.2. We initialize this procedure at $\mathbf{w}^{(0)}$, where any missing entries $jk \in \mathcal{M}$ 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 2.

Update ρ:

To maximize $\ell_{\mathbf{y}}$ with respect to ρ (Step 2 of Algorithm 2), we approximate $\{\gamma_i\}_{i=1}^3$ using only observed values. Using the pairwise expressions in (C.15), the expressions for the expectation step under missing data

are

$$\gamma_1 \approx \frac{1}{|\mathcal{M}^c|} \sum_{jk \in \mathcal{M}^c} E[\epsilon_{jk}^2 \,|\, y_{jk}], \tag{C.21}$$

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

$$\begin{aligned}\gamma_3 &\approx \frac{1}{|\sum_{jk,lm\in\Theta_3} \mathbb{1}[jk \in \mathcal{M}^c]\mathbb{1}[lm \in \mathcal{M}^c]|} \sum_{jk,lm\in\Theta_3} E[\epsilon_{jk} \,|\, y_{jk}]E[\epsilon_{lm} \,|\, y_{lm}]\mathbb{1}[jk \in \mathcal{M}^c]\mathbb{1}[lm \in \mathcal{M}^c], \\ &\approx \frac{1}{|\Theta_3|}\left(\left(\frac{|\Theta_1|}{|\mathcal{M}^c|}\sum_{jk\in\mathcal{M}^c} E[\epsilon_{jk} \,|\, y_{jk}]\right)^2 - \frac{|\Theta_1|}{|\mathcal{M}^c|}\sum_{jk\in\mathcal{M}^c} E[\epsilon_{jk} \,|\, y_{jk}]^2 \right. \\ &\qquad \left. - \frac{|\Theta_2|}{|\mathcal{A}^{(s)}|}\sum_{jk,lm\in\mathcal{A}^{(s)}} E[\epsilon_{jk} \,|\, y_{jk}]E[\epsilon_{lm} \,|\, y_{lm}]\right),\end{aligned}$$

where we only subsample pairs of relations that are observed such that $\mathcal{A}^{(s)} \subset \Theta_2 \cap \mathcal{M}^c$. Then, given the values of $\{\gamma_i\}_{i=1}^3$ in (C.21), the maximization of $\ell_{\mathbf{y}}$ with respect to ρ (Step 2 in Algorithm 2) may proceed as usual.

Prediction:

Joint prediction in the presence of missing data is required for out-of-sample evaluation of the BC-EM estimator, for example, for cross validation studies in Section 4.8. 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 4.6 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).

C.2 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 BC-EM algorithm.

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

$$\boldsymbol{\Omega}(\boldsymbol{\phi}) = \sum_{i=1}^{3} \phi_i \mathcal{S}_i. \tag{C.22}$$

We find empirically that the inverse matrix of any undirected exchangeable network covariance matrix has the same form, that is $\boldsymbol{\Omega}^{-1} = \sum_{i=1}^{3} \mathbf{p}_i \mathcal{S}_i$. Using this fact and the particular forms of the binary matrices $\{\mathcal{S}_i\}_{i=1}^{3}$, 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 $\boldsymbol{\phi}$:

$$\mathbf{C}(\boldsymbol{\phi})\mathbf{p} = [1, 0, 0]^T, \text{ where} \tag{C.23}$$

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

We observe empirically that the eigenvalues of $\mathbf{C}(\boldsymbol{\phi})$ are contained within those of $\boldsymbol{\Omega}$, and thus, we may invert $\boldsymbol{\Omega}$ with a 3×3 inverse to find the parameters $\mathbf{p}$ of $\boldsymbol{\Omega}^{-1}$.

The equation in (C.23) allows one to compute the partial derivatives $\left\{\frac{\partial \phi_i}{\partial p_j}\right\}$. First, based on (C.23), 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 (C.23) 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

$$\frac{\partial \boldsymbol{\phi}}{\partial p_j} = -\mathbf{C}(\mathbf{p})^{-1} \mathbf{A}_j^{(n)} \mathbf{C}(\mathbf{p})^{-1} [1, 0, 0]^T, \tag{C.24}$$

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

C.3 Simulation studies

In this section we present details pertaining to the two simulation studies in Section 4.7.

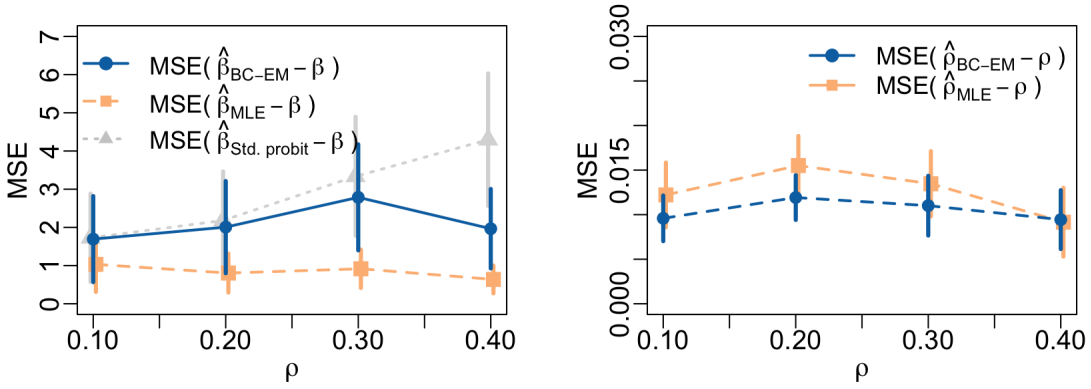


Figure C.1: The left panel depicts the MSE in estimating β using the BC-EM algorithm, MLE, and Ordinary probit regression. The right panel depicts the same for ρ. The MSEs are plotted as a function of the true values of ρ, and solid vertical lines denote Monte Carlo error bars.

C.3.1 Evaluation of BC-EM approximations

See Section 4.7.1 for a description of the simulation study to evaluate the BC-EM algorithm approximations. In Figure C.1, we note that the BC-EM estimator is an improvement over standard probit regression assuming independent errors in estimating β. However, the BC-EM estimator does not attain MSEs quite as low as the MLE for β. The opposite is true in estimating ρ: the BC-EM estimator has lower MSE than the MLE.

C.3.2 Evaluation of estimation of $\boldsymbol{\beta}$

See Section 4.7.2 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 $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]^T = [-1, 1, 1, 1]^T/2$ throughout the simulation study. When generating from the latent eigenmodel in (4.5), we set $\Lambda = \mathbf{I}$, $\sigma_a^2 = 1/6$, $\sigma_u^2 = 1/\sqrt{6}$, and $\sigma_\xi^2 = 1/3$.

To further investigate the source of poor performance of the `amen` estimators of the social relations and latent eigenmodels, we computed the bias and the variance of estimators when generating from the PX model and the latent eigenmodel in Figures C.2 and C.3, respectively. Figures C.2 and C.3 show that the variances of the `amen` estimators of the social relations and latent eigenmodels are similar to the PX model, however, that the bias of the `amen` estimators are substantially larger.

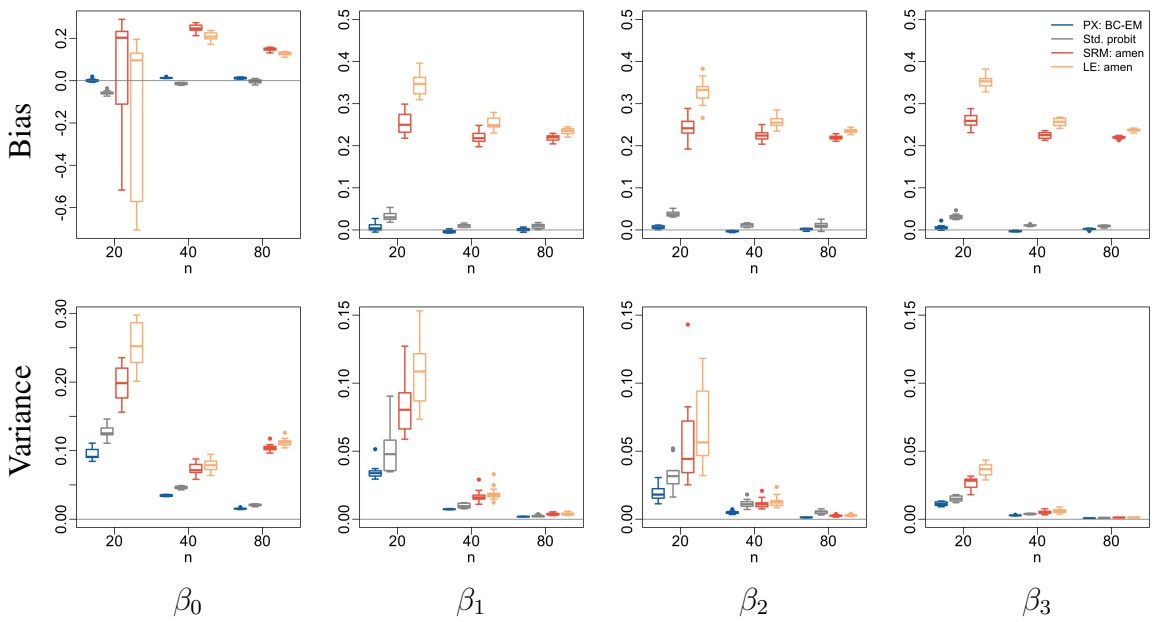


Figure C.2: PX model: Bias and variance of estimators of $\boldsymbol{\beta}$ for a given $\mathbf{X}$ when generating from the PX model. Variability captured by the boxplots reflects variation with $\mathbf{X}$. Note that the intercept, β_0, has biases and variances on different scales than the remaining coefficients.

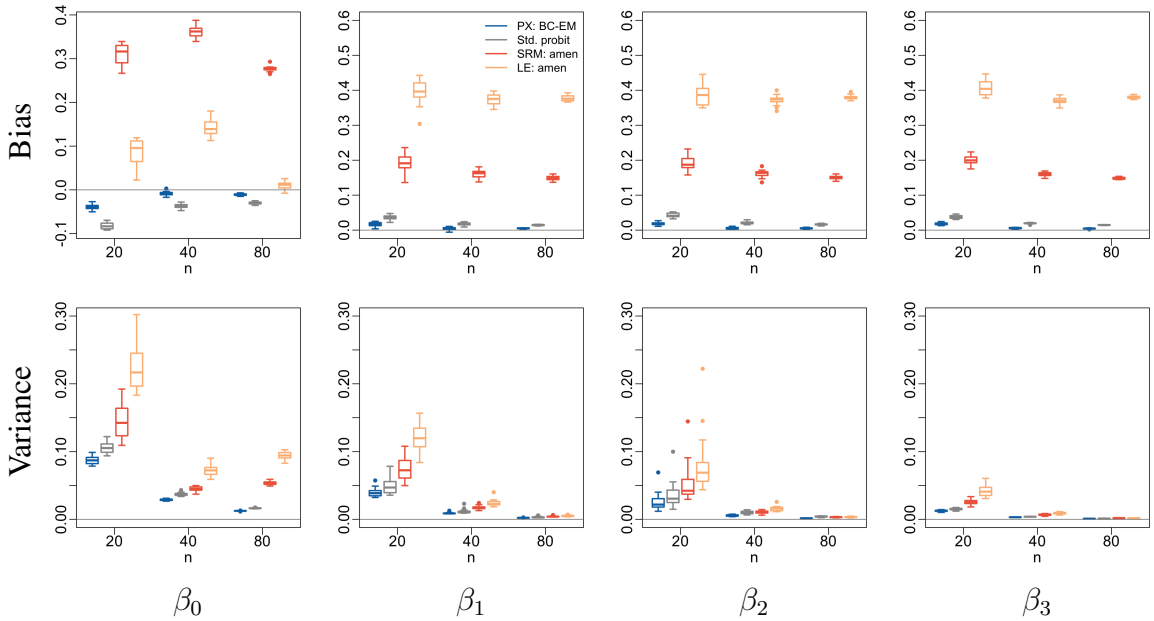


Figure C.3: LE model: Bias and variance of estimators of $\boldsymbol{\beta}$ for a given $\mathbf{X}$ when generating from the latent eigenmodel. Variability captured by the boxplots reflects variation with $\mathbf{X}$. Note that the intercept, β_0, has biases and variances on different scales than the remaining coefficients.

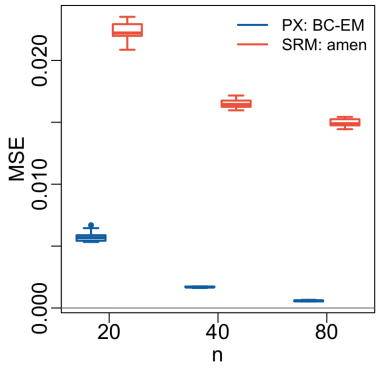


Figure C.4: MSE of the BC-EM estimator and `amen` estimator of the social relations model of ρ when generating from the PX model. Variability captured by the boxplots reflects variation in MSE with $\mathbf{X}$.

Both the BC-EM estimator of the PX model and `amen` estimator of the social relations model provide estimates of ρ. We computed the MSE for each estimator, for each $\mathbf{X}$ realization, when generating from the PX model. In Figure C.4, the MSE plot for $\widehat{\rho}$ shows that the MSE, and the spread of the MSE, decreases with n for the BC-EM estimator, suggesting that the BC-EM estimator of ρ is consistent. As with the $\boldsymbol{\beta}$ parameters, the `amen` estimator displays substantially larger MSE than the BC-EM estimator of ρ.

C.4 Analysis of political books network

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

C.4.1 Prediction performance using ROC AUC

In Section 4.8, we use area under the precision-recall curve to evaluation predictive performance on the political books network data set. Figure C.5 shows the results of the cross validation study, described in Section 4.8, as measured by area under the receiver operating characteristic (ROC AUC). The conclusions are the same as those given in Section 4.8: 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.

C.4.2 Linear approximation in ρ in BC-EM algorithm

In Section 4.5.2, we discuss a series of approximations to the E-step of an EM algorithm to maximize $\ell_{\mathbf{y}}$ with respect to ρ. One of these approximations is a linearization of the sample average

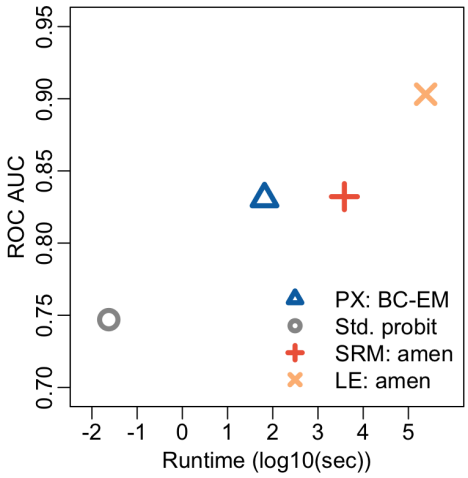


Figure C.5: Out-of-sample performance in 10-fold cross validation, as measured by area under the precision-recall curve (ROC AUC), plotted against mean runtime in the cross validation for Krebs' political books network. The estimators are standard probit assuming independent observations (Std. probit), the proposed PX estimator as estimated by BC-EM (PX: BC-EM), the social relations model as estimated by `amen` (SRM: amen), and the latent eigenmodel as estimated by `amen` (LE: amen).

$\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ with respect to ρ. In Figure C.6, we confirm that this approximation is reasonable for the political books network data set. Figure C.6 shows that the linear approximation to $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \,|\, y_{jk}, y_{lm}]$ (dashed blue line), as described in detail in Section C.1.3, agrees well with the true average of the pairwise expectations (solid orange line).

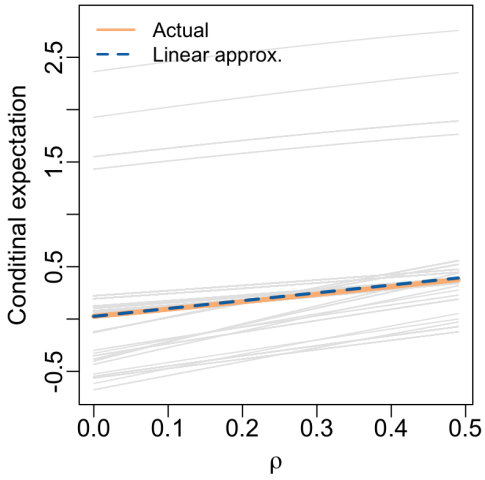


Figure C.6: The average of all pairwise expectations $\frac{1}{|\Theta_2|}\sum_{jk,lm\in\Theta_2} E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ is shown in orange, and the linear approximation to this average, described in Section 4.5, is shown in dashed blue. In addition, pairwise conditional expectations $E[\epsilon_{jk}\epsilon_{lm} \mid y_{jk}, y_{lm}]$ are shown in light gray, for a random subset of 500 relation pairs $(jk, lm) \in \Theta_2$.