©Copyright 2019

Mengjie Pan

Inferring Network Structure From Partially Observed Graphs

Mengjie Pan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Tyler H. McCormick, Chair

Adrian Dobra

Ali Shojaie

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Inferring Network Structure From Partially Observed Graphs

Mengjie Pan

Chair of the Supervisory Committee:
Associate Professor Tyler H. McCormick
Department of Statistics and Department of Sociology

Collecting social network data is notoriously difficult, meaning that indirectly observed or missing observations are very common. In this dissertation, We address two of such scenarios: inference on network measures without any direct network observations, and inference of regression coefficients when actors in the network have latent block memberships. Direct network data is expensive to collect because it requires soliciting connections between all members of the population. Collecting aggregate relational data (ARD) is much more cost effective. In the first two methodological chapters, we show that we can use ARD to estimate individual and global network properties. We connect ARD to a network formation model, which allows us to obtain draws from the posterior distribution over graphs given the ARD response vector. We can then compute network statistics based on these posterior samples. We demonstrate our method using evidence from simulation and replicating results from cases where the complete graph was observed. In the last methodological chapter, we discuss how we make inference on coefficients where the outcome of a linear regression is the interaction between an ordered pair of actors. We propose block-exchangeable errors and algorithms for estimating standard errors. We show that the block-exchangeable estimator is preferable to the exchangeable estimator when latent blocks and observed covariates are dependent.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ARD: Aggregate relational data.

B-E: Block-exchangeable.

DC: Dyadic clustering.

EXCH: Exchangeable.

ACKNOWLEDGMENTS

I am extremely grateful to have received so much guidance and advice from my advisor Tyler H. McCormick, collaborators, faculty and peers in the department of Statistics at the University of Washington. First I would like to thank Tyler H. McCormick for his guidance and encouragement on research ideas, presentation and writing skills, and career choices. I have learned a great amount from Tyler over the past three years. The work in this dissertation would not be possible without the inspiring conversations with Tyler.

I would also like to thank my committee members, Adrian Dobra, Ali Shojaie, and Darryl J. Holman for attending my exams and providing feedback. They have especially motivated me to think deeply about how to tell a story of my dissertation, and where my research stands in literature.

I have been fortunate to work with many wonderful collaborators. My thanks go to Arun Chandrasekhar and Emily Breza for sharing their knowledge in economic applications and motivating discussions, and Bailey Fosdick for sharing her time, enthusiasm, and research ideas during our regular meetings .

There are many individuals in the statistics department who have helped me during my graduate journey. I wish to thank Paul Sampson, Michael D. Perlman, Marina Meila, Thomas Richardson, and Mathias Drton for their influence on my perspective of statistics and help with graduate assistantships. I would like to thank Ellen Reynolds and Eileen Heimer for administrative help, and Asa Sourdiffe for cluster usage help. I also appreciate feedback from Wesley Lee and Austin Schumacher during group meetings and from Fan Xia, Shane Lubold,

Christopher Aicher and Sean Jewell during my practice exam. Outside the University of Washington, I am grateful for my college advisor, Lynne Butler, for encouraging me to pursue a Ph.D. in statistics.

Finally, I would like to thank my parents and grandparents for always supporting me to pursue my dreams, and funding my college education. I would not have made all of the achievements without their love and support.

Chapter 1
INTRODUCTION

1.1 An overview and motivation

Network data represents relations between actors, and is particularly useful for understanding human behaviors. One instance is that network data can be used to identify the most central actors, and seeding information with these central individuals leads to fast spread of information. For example, spreading information on a new technology to the most central farmers in a village is an efficient way to introduce the technology. Another example of using network data is to study patterns of risk sharing. For instance, introducing micro-finance to neighborhoods in Hyderabad, India, decreases rick sharing among actors.

While network data has many applications in economics and other social sciences, collecting network data is notoriously difficult, meaning that indirectly observed or missing observations are very common. My dissertation addresses two of such scenarios: inference on network measures without any network observations and inference of regression coefficients when actors have latent block memberships.

Network data is expensive and difficult to collect. A typical network elicitation exercise requires (1) enumerating every member of the network in a census, (2) asking each subject to name those individuals with whom they have a relationship, and (3) matching each individual's list of social connections back to the census. Moreover, this process needs to be repeated across many networks to conduct convincing inference. A full network survey of 120 villages in Karnataka, India costs around $\$190,000$. These barriers place significant limitations on conducting high-quality work in this space and discouraging research, especially

by those without access to considerable resources.

To address these challenges and ease the burden of collecting network data for field researchers, we propose that researchers collect aggregated relational data (ARD), responses to questions of the form "How many of your social connections have trait k." ARD is considerably cheaper to obtain than full or even partial-network data. Our proposed method builds on prior work by McCormick and Zheng (2015), which shows how the network formation model is related to a likelihood that depends only on ARD. This network formulation approach is related to the latent space model (Hoff et al. (2002)), an often-used model in the statistics literature, where the probability of a connection depends on individual heterogeneity and the positions of nodes in a latent space. Previous literature on inferring sprcific network measures without observing network data include Banerjee et al. (2016b) and Beaman et al. (2016). However, these methods are limited because they only speak to identifying central individuals or focus on proxies. Our approach does not restrict the researcher to make inferences about one specific aspect of the data. Instead we provide a blueprint to recover a distribution over the entire graph at minimal cost.

The second part of my dissertation concerns a scenario where latent blocks in the network impact inference of regression coefficients. Specifically, We focus on a regression case where continuous relations between ordered pairs of actors, or dyads, are modeled as a linear function of observable covariates. These continuous pairwise relations can be represented in a network as directed and weighted edges, where weights of edges follow a continuous distribution. We refer to these relations as relational observations/response. The challenge of getting accurate estimation of the standard error and thus a confidence interval with the correct coverage lies in modeling covariance structure of errors. Relational observations of two dyads are likely correlated when the dyads share a member in common.

One set of approaches to model the covariance structure of errors is to impose parametric

distributional assumptions on errors or model the error covariance structure directly (Hoff (2005) and Hoff (2015)). While these approaches produce interpretable representations of underlying residual structure, they always assume the error structure is consistent with the underlying parametric model. Another set of approaches to model the covariance structure of errors is through non-parametric approaches. However, such approaches either makes no distributional assumptions (see dyadic clustering estimator in Fafchamps and Gubert (2007)), or assume exchangeability of errors (Marrs et al. (2017)). The former approach results in a standard error estimator that is extremely flexible yet extremely variable, whereas the latter approach assumes all actors are identically distributed and results in a relatively restricted estimator. We propose an alternative block-exchangeable estimator that bridges the gap between these two existing approaches. We assume that actors have block memberships and actors within the same block are exchangeable. Heterogeneity based on unobserved variables are quite common in networks, and relational observations between actors in the same block may have different patterns than those observations between actors in different blocks.

For the rest of this chapter, we describe the organization of the rest of the dissertation.

1.2 Organization of the dissertation

In this dissertation, we aim to provide modeling and computational strategies for making inference on network structure from partially observed graphs. Chapter 2 provides a review of the latent surface model, and a brief introduction to the literature on non-parametric approaches for modeling error covariance structure of relational observations. The core methodological chapters (Chapter 3,4, and 5) will address the challenges described before.

Chapter 3 proposes a full framework and estimation algorithm for recovering node- and network- level statistics from ARD collected on a randomly selected subset of actors. This includes a setup of the problem, a theorem with proof on identification of model parameters,

and how we generate graphs from the network formulation model. From generated graphs, we obtain a distribution over graph statistics. Then we apply our methods to two empirical field experiments that collected full or partial network data, and we show that we can replicate the findings by using ARD alone.

Chapter 4 presents theorems, proofs, and simulation evidence that estimates of model parameters and network statistics obtained using ARD are consistent, under certain assumptions. In addition to estimating network statistics in a single large network, we also show that estimates of regression coefficients are consistent, when response or covariate is an estimated network measure from ARD.

Chapter 5 proposes a novel block-exchangeable estimator as well as two estimation algorithms for estimating the standard errors of estimated regression coefficients. With theoretical results and simulation evidence, we show scenarios where our block-exchangeable estimator outperforms its ancestors, by having more accurate confidence interval coverage.

Finally, in Chapter 6 we discuss directions for future research.

Chapter 2
BACKGROUND

2.1 Latent surface model

Latent space models are widely used for analyzing networks in the statistics literature (for example, Hoff et al. (2002)). These models assume that actors form ties conditionally independent given their latent positions, and the propensity to form a tie is inversely related to the distance between their latent positions. By using a distance measure that preserves the triangle inequality, this latent geometry captures dependence structure in the networks, such as transitivity.

However, models such as the latent space model from Hoff et al. (2002) assume complete network data observations, which may be financially difficult to collect for researchers on a budget. To this end, McCormick and Zheng (2015) derive a latent surface representation of social network structure when no network data is available but ARD is collected. By relating the network formulation model to a likelihood that only depends on ARD, ARD allows us to infer the relative location of actors, as well as where subgroups lie relative to one another in the latent space.

For the rest of the section, we describe the latent surface model and its setup, how the likelihood relates to the model formulation parameters, and a Bayesian framework that allows us to make inference on model parameters with observed ARD.

Let $\mathbf{g} = (V, E)$ be an undirected, unweighted graph with vertex set V and edge set E, with $|V| = n$ nodes. We let $g_{ij} = \mathbf{1}\{ij \in E\}$ denote whether an edge exists between a pair of actors i and j, $i, j \in V$. An ARD response is a count y_{ik} to a question "How many of your

social connections have trait k." We can write

$$y_{ik} = \sum_{j \in G_k} g_{ij},$$

where $G_k \subset V$ is the set of actors with trait k. That is, y_{ik} is a count of the number of actors in group k that actor i knows.

McCormick and Zheng (2015) model the underlying network as

$$\Pr(g_{ij} = 1 | \nu_i, \nu_j, \zeta, z_i, z_j) \propto \exp(\nu_i + \nu_j + \zeta z_i' z_j), \tag{2.1}$$

where ν_i are person-specific random effects that capture heterogeneity in linking propensity, $\zeta > 0$ modulates the intensity of the latent component, and z_i are latent positions on the surface of $p+1$ dimensional hypersphere, $\mathcal{Z} = \mathcal{S}^{p+1}$, centered at the origin.

Using a Bayesian framework, McCormick and Zheng (2015) model priors for latent positions on $\mathcal{S}^{p+1}$ as

$$z_i | \upsilon_z, \eta_z = 0 \sim \mathcal{M}(\upsilon_z, 0) \text{ and } z_{j \in G_k} | \upsilon_k, \eta_k \sim \mathcal{M}(\upsilon_k, \eta_k)$$

where $\mathcal{M}$ denotes the von Mises-Fisher distribution across $\mathcal{S}^{p+1}$, υ_k denotes the location on the sphere and η_k is the intensity: $\eta = 0$ means that the location is uniform at random. The $z_{j \in G_k}$ terms describe the latent positions of individuals who have a particular trait k.

McCormick and Zheng (2015) show that the expected ARD response by i for category k can be expressed as

$$\lambda_{ik} = \mathbb{E}[y_{ik}] = d_i b_k \left(\frac{C_{p+1}(\zeta) C_{p+1}(\eta_k)}{C_{p+1}(0) C_{p+1}\left(\sqrt{\zeta^2 + \eta_k^2 + 2\zeta\eta_k \cos(\theta_{(z_i, \upsilon_k)})}\right)} \right), \tag{2.2}$$

where d_i is the respondent degree and b_k is the share of ties made with members of group k, $C_{p+1}(\cdot)$ is the normalizing constant of the von Mises-Fisher distribution, $\theta_{(z_i,\upsilon_i)}$ is the angle between the two vectors (McCormick and Zheng, 2015). The expected number of nodes of type k known by i is roughly its expected degree scaled by the population share of the group, adjusted by a factor that captures the relative proximity of the node to the type in question in latent-space.

A key assumption in this formation model is that the propensities for individuals to form ties are conditionally independent given the latent variables. The likelihood for the formation model, conditional on the latent variables, is a Bernoulli trial for each pair. ARD, then, is the sum of (conditionally) independent Bernoulli trials, which we can approximate with a Poisson distribution. This allows us to compute the distribution of the ARD response, which will be distributed Poisson,

$$y_{ik}|d_i, b_k, \zeta, \eta_k, \theta_{(z_i,\upsilon_k)} \sim \text{Poisson}\left(\lambda_{ik}\right).$$

Though the likelihood above relies only on ARD, it does not uniquely identify the formation model since λ_{ik} estimates on the degree, d_i, rather than the individual heterogeneity parameter ν_i. The expected degree can be computed,

$$d_i = n \exp(\nu_i)\, \mathbb{E}[\exp(\nu_j)] \left(\frac{C_{p+1}(0)}{C_{p+1}(\zeta)}\right). \tag{2.3}$$

McCormick and Zheng (2015) propose Gamma priors for ζ and η_k with conjugate priors

on the hyperparameters. Then if $\boldsymbol{\theta}$ is the shorthand for all parameters, the posterior is

$$\begin{aligned}\boldsymbol{\theta}|y_{ik} \propto & \prod_{k=1}^{K}\prod_{i=1}^{n}\exp(-\lambda_{ik})\lambda_{ik}^{y_{ik}}\prod_{i=1}^{n}\text{Normal}(\log(d_i)|\mu_d,\sigma_d^2) \\ & \times\prod_{k=1}^{K}\text{Normal}(\log(b_k)|\mu_b,\sigma_b^2)\prod_{k=1}^{K}\text{Gamma}(\eta_k|\gamma_{\eta_k},\psi_{\eta_k})\text{Gamma}(\zeta|\gamma_\zeta,\psi_\zeta).\end{aligned}$$

Given the data, we can compute posteriors over degrees of nodes, their unobserved heterogeneity, population shares of categories, intensity of the latent space component in the network formation model, relative locations of categories on the sphere, and how intensely they are concentrated at these locations.

2.2 Non-parametric approaches on modeling error covariance structure for network data

In this section, we present two existing non-parametric approaches on modeling errors for regression on network data. We first introduce the set-up and the regression model. Let n be the observed number of individuals, y_{ij} be the directed relational response from actor i to actor j, and $\boldsymbol{X}_{ij} = [1 \;\; X_{1,ij} \;\; \cdots \;\; X_{(p-1),ij}]$ be a $(p \times 1)$ vector of observable covariates. We assume there is no edge from an actor i to itself. The regression model is

$$y_{ij} = \boldsymbol{\beta}^T\boldsymbol{X}_{ij} + \xi_{ij}, \quad i,j \in \{1,...,n\}, i \neq j, \tag{2.4}$$

Let Ξ be the error vector. We have $\Xi = \begin{bmatrix}\xi_{21} & \xi_{31} & \cdots & \xi_{n1} & \cdots & \xi_{1n} & \cdots & \xi_{(n-1)n}\end{bmatrix}^T \sim N(\mathbf{0},\Omega)$, where $\Omega = \text{Var}(\Xi)$ is a $n(n-1)$ by $n(n-1)$ symmetric matrix.

It is reasonable to assume that dyadic observations are not independent, due to the presence of individual effects for pairs of dyads that involve the same actor. In a linear regression model of form (2.4), there are a number of ways to model Ω, ranging from more flexible assumption yielding more parameters, to more restricted assumption yielding much

fewer parameters.

Fafchamps and Gubert (2007) propose to model Ω such that $\text{Cov}(\xi_{ij}, \xi_{kl}) \neq 0$ if $\{i, j\} \cap \{k, l\} \neq \emptyset$ and $\text{Cov}(\xi_{ij}, \xi_{kl}) = 0$ if $\{i, j\} \cap \{k, l\} = \emptyset$. Without additional assumptions, $\text{Cov}(\xi_{ij}, \xi_{kl})$ is a single parameter for each unique set (i, j, k, l) when $\{i, j\} \cap \{k, l\} \neq \emptyset$.

Let DC denote the shorthand of dyadic clustering, and let Ω_{DC} denote the covariance matrix under the assumption in Fafchamps and Gubert (2007). The dyadic clustering estimator $\widehat{\Omega}_{DC}$ estimates elements in Ω_{DC} by taking a single residual product: $\widehat{\text{Cov}}(\xi_{ij}, \xi_{kl}) = r_{ij} r_{kl}$, where r_{ij} and r_{kl} are the residuals of corresponding dyads.

While the dyadic clustering estimator is extremely flexible because there are no assumptions posed on Ω_{DC} except that non-overlapping dyads are independent, the estimator $\widehat{\Omega}_{DC}$ contains $\mathcal{O}(n^3)$ parameters and the elements in $\widehat{\Omega}_{DC}$ are estimated by a single product of residuals. This makes the variance of the estimator very large, which results in large variance of standard errors.

In order to decrease the number of parameters and the variance of $\widehat{\Omega}_{DC}$, Marrs et al. (2017) propose exchangeability assumption on the error vector Ξ and a simple moment-based estimator. The exchangeability assumption states that the errors in a relational data model are jointly exchangeable if $P(\Xi)$, the probability distribution of the error vector, is invariant under permutation of the rows and columns. Under the exchangeability assumption, Marrs et al. (2017) state that there are five non-zero parameters in Ω_E, and estimates elements in Ω_E by averages of residual products, which greatly reduces the variance of the exchangeable estimator $\widehat{\Omega}_E$ compared to that of $\widehat{\Omega}_{DC}$.

Chapter 3

USING AGGREGATE RELATIONAL DATA TO FEASIBLY IDENTIFY NETWORK STRUCTURE WITHOUT NETWORK DATA[1]

3.1 Introduction

In this chapter, we present a method that recovers node- and network- level statistics using ARD alone without any network observations. ARD is considerably cheaper to obtain than full or even partial-network data. J-PAL South Asia cost estimates show that collecting ARD leads to a 70-80% cost reduction. Our approach makes network research scalable and accessible on a budget.

Our proposed method is intuitive and comes down to the following three simple observations. First, ARD is considerably cheaper and easier to collect than network data. Second, ARD provides the researcher with enough information to identify parameters of an oft-used and standard network formation model in the statistics literature (see e.g. Hoff et al. (2002)). The argument builds on prior work by McCormick and Zheng (2015), which shows how the network formation model is related to a likelihood that depends only on ARD.

Third, this parametric model of network formation is sufficiently rich to capture a number of features of real-world network structures. We provide two examples of recent research where either full or partial network data had been collected. Breza and Chandrasekhar (2019) study how the observation of one's savings behavior by more central individuals in the network leads to greater savings in order to maintain a reputation for being responsible. We

[1]The contents of this chapter are based on the paper Breza et al. (2017)

show with constructed ARD, we can replicate the paper's findings. Banerjee et al. (2016a) use partial network data to study how exposure to microcredit erodes social capital by reducing support. The authors in part collected survey ARD in this sample, and we show we can replicate the findings. Further, the ARD enables conclusions about how microcredit exposure affected the neighborhood-level informal financial network structure. These examples show the effectiveness of our approach across different contexts and how ARD would have helped in policy-relevant empirical work. Researchers could have reached their conclusions without collecting full network data, which also means that the financial barrier to entry for such research would be considerably lower, thereby democratizing in part this research frontier.

For the bulk of the chapter, we consider settings where we have ARD for a randomly-selected subset of nodes in the network and a basic vector of covariates for the full set of nodes. ARD counts the number of links an agent has to members of different subgroups in the population. The core insight of our approach is that by combining ARD with a network formation model, we can derive the posterior distribution for the graph. To do this, we assume a network formation model, where the probability of a connection depends on individual heterogeneity and the positions of nodes in a latent social space Hoff et al. (2002). The distance between nodes in the space is a pair-specific latent variable that is inversely related to the probability of a tie: nodes that are closer together in the latent space are more likely to form ties. The propensity to form ties across pairs is assumed conditionally independent given the latent variables. ARD gives us information on where different subgroups lie relative to one another in this latent space. That is, ARD allows us to triangulate the relative locations of nodes. In prior work, McCormick and Zheng (2015) show how to relate the network formation model to a likelihood that depends only on ARD. We extend that result and show how we can recover the parameters of the network formation model. In our case, this consists of both individual-level effects for every node in the sample as well as the location of all nodes in

the latent-space. Using a Bayesian framework for inference, we show that the choice of prior distribution has minimal impact on our ability to accurately recover moments for a variety of network configurations. We note that, equipped with estimates of the degree distribution as well as the latent space locations in the ARD sample, we can use the demographic covariates for the entire sample to estimate the posterior distributions of the degrees, fixed-effects, and latent locations for the entire population. We can then generate graphs from the posterior distribution over formation model parameters given the ARD response vector and compute network statistics for each generated graph.

Our work contributes to and builds on several literatures. First, there is a nascent literature that seeks to apply the lessons from the economics of networks without having access to network data (e.g., Beaman et al. (2016), Banerjee et al. (2016b), and Chassang et al. (2017)). These methods are limited because they only speak to identifying central individuals or focus on proxies. Prior work shows that proxies such as geography or ethnic divisions do not capture the network well and augmenting sampled network data, which works, can still be expensive (Chandrasekhar and Lewis, 2016). Our approach does not restrict the researcher to inferences about one specific aspect of the data. Instead we provide a blueprint to recover a distribution over the entire graph at minimal cost.

Second, our work builds on a sizable literature on ARD, but expands both the context and inferential quantities of interest. In contrast to our work, most previous work on ARD focused on estimating the size of "hard-to-reach" populations (see e.g. Killworth et al. (1998) or Bernard et al. (2010)). These groups consist of individuals who are outside the sampling frame of most surveys. Rather than needing to reach these individuals directly, using ARD allows researchers to study individuals through their interactions with others who are captured by more traditional sampling strategies. Bernard et al. (1991) use ARD to estimate the number of individuals impacted by an earthquake whereas Kadushin et al. (2006) use ARD

to estimate the number of individuals using heroine.

The closest related work from the ARD literature is McCormick and Zheng (2015) – here, we use the same network formation model and build on derivations that are the key contribution of that work. Specifically, McCormick and Zheng (2015) show that, for a specific formation model, it is possible to arrive at a likelihood that is informed by information in ARD. That is, they interpret and do inference on a likelihood for ARD. While we also have this likelihood, in our work it is merely an intermediate step. In this chapter, we perform inferences about the parameters of the formation model itself. By explicitly making the link to the formation model, we can generate graphs and compute both graph and individual level statistics.

Third, our latent surface model is closely related to the β-model (Holland and Leinhardt, 1981; Hunter, 2004; Park and Newman, 2004; Blitzstein and Diaconis, 2011) and the properties examined in Chatterjee et al. (2010) and Graham (2017). Every node has a fixed-effect. Links form conditionally independently given the fixed effects of the nodes involved, modulated by a function of distance between the nodes in a latent space. Relative to the Graham (2017) and Chatterjee et al. (2010) models, our model places nodes in a latent space (as in Hoff et al. (2002)), which we are trying to estimate, whereas the former only allows for observable covariates, and the latter has none. Whereas previous approaches consider an asymptotic frame based on a growing graph, we consider an explicitly sampling-based framework.

We begin with an overview of our method for an applied researcher in Section 3.2. Section 3.3 presents the full framework, model, and estimation algorithm. In section 3.4, we present a theorem with proof on identification of network formation model parameters. In Section 3.5, we apply our results to two empirical examples. Section 3.6 provides a discussion of how an applied researcher could navigate the model's limitations.

3.2 Overview of method

We begin with a simple overview of the proposed method. Suppose that a researcher is interested in studying networks in a set of rural villages. A village network with n households is given by $\mathbf{g}$, which is a collection of links ij where $g_{ij} = 1$ if and only if households i and j are linked and $g_{ij} = 0$ otherwise. To fix ideas, suppose that the researcher wants to learn how some outcome variable W is related to a network statistic (or a vector of statistics) of interest $S(\mathbf{g})$. Or, perhaps the researcher is interested in how a treatment (such as exposure to microcredit) affects features of network structure, $S(\mathbf{g})$.

Our procedure takes five steps.

I. **Conduct ARD survey:** Sample a share ψ (e.g., 30%) of households. Ask 5-8 ARD questions, such as

> *"How many households among your network list do you know where any adult has had typhoid, malaria, or cholera in the past six months?"*

The ARD response for a household i is

$$y_{ik} = \sum_j g_{ij} \cdot \mathbf{1}\{j \text{ has had one of those diseases in past 6 mo.}\}$$

where trait k denotes the disease question. This just adds up all friends that have had the diseases over the last six months.

II. **Conduct census exercise:** Obtain basic information about the full set of households in the village in a very rapid survey (denoted X_i for all $i = 1, ..., n$).

- Minimal demographics: e.g., GPS coordinates, caste/subcaste.

- ARD traits: e.g., whether the household has had typhoid, malaria, or cholera in the past six months.

III. **Estimate network formation model with ARD:** Use the information from the ARD survey and the population counts from the census to estimate the parameters of a network formation model in (2.1). In this model, the probability that two households i and j are linked depends on household fixed effects (ν_i) and distance in some latent space (latent locations z_i)

- Fit a model to predict ν_i, z_i using the ARD sample.
- Predict ν_i, z_i using X_i for all households in the census but not in the ARD sample.

Equipped with estimated fixed effects and latent locations for all n households in the network, the probability of any network $\mathbf{g}$ being drawn is fully computed.

IV. **Compute network statistics of interest:** Use the estimated probability model (using ζ, fixed effects ν_i and latent locations z_i) to compute $\mathbb{E}[S(\mathbf{g})|\mathbf{Y}]$.

V. **Estimate economic parameter of interest:** E.g., run regressions such as

$$W_v = \alpha + \beta' \, \mathbb{E}[S(\mathbf{g}_v)|\mathbf{Y}_v] + \epsilon_v \text{ or } \mathbb{E}[S(\mathbf{g}_v)|\mathbf{Y}_v] = \alpha + \beta \text{Treatment}_v + \epsilon_v,$$

though clearly one can do more complex exercises once one has estimated the above network formation model.

3.3 Methods and estimation

In this section, we present formally the procedure outlined above. This includes defining ARD, introducing the network formation model, linking explicitly the formation model to the

ARD, and finally, outlining how to generate graphs from that network formation model. The result is a distribution over graphs (and therefore graph statistics) based on the observed ARD.

3.3.1 Setup

We begin by describing the underlying graph and the ARD. Recall that $\mathbf{g} = (V, E)$ is an undirected, unweighted graph with vertex set V and edge set E, with $|V| = n$ nodes. Let $g_{ij} = \mathbf{1}\{ij \in E\}$. We also assume that researchers have a vector of demographic characteristics, X_i for every $i \in V$.

Finally, we assume that the researcher has an ARD sample of $m \leq n$ nodes which are selected uniformly at random (where we define $\psi = \frac{m}{n}$). These could be the whole sample, with $\psi = 1$, or a smaller share, and will depend on the context. It is useful to define V_{ard} to be the ARD sample set and $V_{non} = V \setminus V_{ard}$.

Recall that an ARD response is a count y_{ik} to a question "How many households with trait k do you know?" which we can write as

$$y_{ik} = \sum_{j \in G_k} g_{ij}$$

where $G_k \subset V$ is the set of nodes with trait k. That is, y_{ik} is a count of the number of households in group k that person i knows. Note that throughout we assume that we observe y_{ik} and, in some cases, additional information about the group of people with trait k (e.g., the number of households with this trait in the population), but we do not observe any links in the network.

3.3.2 Latent surface model

The setup and model we use is from McCormick and Zheng (2015), as presented in Section 2.1. Using the latent surface model, we obtain a likelihood that only depends on ARD. And with a Bayesian framework, we obtain posterior distribution of a vector that contains all parameters, denoted by $\boldsymbol{\theta}$. Therefore, with any draw of $(z_1, ..., z_n)'$, $(\nu_1, ..., \nu_n)'$, and η, we can generate a graph from the distribution in (2.1).

3.3.3 From ARD sample to Non-ARD sample

Thus far we only have posteriors for our ARD sample V_{ard}. We now turn to predicting ν_i and z_i for $j \in V_{non}$. We use k-nearest neighbors to draw this distribution. Given demographic covariates X_i for all $i \in V$, we define a distance between nodes in the feature space $d(X_i, X_j)$ for $i, j \in V$. For each $j \in V_{non}$, we pick $i' \in V_{ard}$ such that $d(X_{i'}, X_j)$ is among the k smallest distances. We then take a weighted average of $\nu_{i'}$ and $z_{i'}$ with weights inversely proportional to $d(X_{i'}, X_j)$, to estimate ν_j and z_j, respectively. We normalize z_j such that $|z_j| = 1$ to map it to the surface of the sphere. Thus, we have described a framework that a researcher can use with only ARD data and demographic covariates to take a sample of draws from a network formation latent surface model.

3.3.4 Drawing a graph

We now describe the algorithm used to generate a distribution of graphs $\{\mathbf{g}_s\}_{s=1}^S$. The algorithm for drawing graphs requires specifying the dimension of the latent hypersphere. Throughout the chapter we follow McCormick and Zheng (2015) and use $p = 2$, for a three-dimensional hypersphere. This choice also facilitates visualizing latent structure. The posterior distribution is not available in closed form. We therefore use a Metropolis-within-Gibbs algorithm to obtain samples from the posterior. In the description below the jumping

scale is tuned adaptively throughout the course of sampling. Specifically, every 50 draws we look at the acceptance rate of these draws and then adjust the scale of the jumping distribution. We follow the guidelines given in Gelman et al. (2013) and perform checks to ensure that our sampler has converged.

ALGORITHM **1** (Drawing Graphs)**.** *Input:* $y_{ik} \ \forall i \in V_{ard}$, $X_i \ \forall i \in V$.
Assume ARD groups, $k = 1, ..., K$, *such that* $K \geq p$. *We propose fitting the model as follows (noting that steps 1 & 2 follow from McCormick and Zheng (2015)):*

1. *For a subset of the ARD groups,* $k^{(s)} = 1, ..., K^{(s)}$, *fix* $\boldsymbol{\upsilon}_k^{(s)}$.

2. *Repeat to convergence for* $t = 1, ..., T$

   (a) *For each* i, *update* z_i *using a random walk Metropolis step with proposal* $z_i^* \sim \mathcal{M}(z_i^{(t-1)}, \textit{jumping distribution scale})$. *Use the algorithm proposed by Wood (1994) to simulate proposals implemented in the* `R` *package* `Rfast`*(Papadakis et al., 2017).*

   (b) *Update* $\boldsymbol{\upsilon}_k$ *using a conditionally conjugate Gibbs step* $\boldsymbol{\upsilon}_k \sim \mathcal{M}(\boldsymbol{m}_k/||\boldsymbol{m}_k||_2, ||\boldsymbol{m}_k||_2)$, *where* $\boldsymbol{m}_k = \eta_k \sum_{j \in k} z_j$. *(See e.g. Mardia and El-Atoum (1976); Guttorp and Lockhart (1988); Hornik and Grün (2013); Straub et al. (2015)).*

   (c) *Update* d_i *with a Metropolis step with* $\log(d_i^*) \sim N(\log(d_i)^{(t-1)}, \textit{jumping distribution scale})$.

   (d) *Update* b_k *with a Metropolis step with* $\log(b_k^*) \sim N(\log(b_k)^{(t-1)}, \textit{jumping distribution scale})$.

   (e) *Update* η_k *with a Metropolis step with* $\eta_k^* \sim N(\eta_k^{(t-1)}, \textit{jumping distribution scale})$.

   (f) *Update* ζ *with a Metropolis step with* $\zeta^* \sim N(\zeta^{(t-1)}, \textit{jumping distribution scale})$.

   (g) *Update* $\mu_b \sim N(\hat{\mu}_b, \sigma_b^2)$ *where* $\hat{\mu}_b = \sum_{k=1}^{K} \log(b_k)/K$.

   (h) *Update* $\sigma_b^2 \sim \textit{Inv-}\chi^2(K-1, \hat{\sigma_b^2})$ *where* $\hat{\sigma_b^2} = \frac{1}{K-1} \sum_{k=1}^{K} (\log(b_k) - \mu_b)^2$.

(i) Update $\mu_d \sim N(\hat{\mu}_d, \sigma_d^2)$ *where* $\hat{\mu}_d = \sum_{i=1}^{n} \log(d_i)/n$.

(j) Update $\sigma_d^2 \sim$ *Inv-*$\chi^2(n-1, \hat{\sigma_d^2})$ *where* $\hat{\sigma_d^2} = \frac{1}{n-1}\sum_{i=1}^{n}(\log(d_i) - \mu_d)^2$.

3. Repeat for $t \in \{T/2+1, ..., T\}$

(a) Calculate $\nu_i^t \ \ \forall i \in V_{ard}$ *such that* ν_i^t *satisfies* $(d_i)^t = \exp(\nu_i^t)\sum_i \exp(\nu_i^t)\left(\frac{C_{p+1}(0)}{C_{p+1}(\zeta)}\right)$.

(b) Use method described in Section 3.3.3 to estimate ν_j^t *and* $z_j^t \ \ \forall j \in V_{non}$.

(c) Sample graph $\mathbf{g}_t$ *using the the procedure described below.*

Output: $\{\mathbf{g}_s\}_{s=1}^{S}$

To generate graphs, recall that the formation model has $\Pr(g_{ij} = 1|\nu_i, \nu_j, \zeta, z_i, z_j) \propto \exp(\nu_i + \nu_j + \zeta z_i' z_j)$. We estimate ζ and z_i, z_j using the likelihood derived in McCormick and Zheng (2015). Equation (2.3) relates degree to the unobserved gregariousness parameters, ν_i. If we approximate $\mathbb{E}[\exp(\nu_j)]$ as the average of the ν_i's, then we can view equation (2.3) as a system with n equations and n unknowns and obtain estimates for ν_i for each respondent.

We then normalize the $\exp(\nu_i + \nu_j + \zeta z_i^{t\ \prime} z_j^t)$ terms to produce probabilities. Define

$$\Pr(g_{ij} = 1|z_i, z_j, \nu_i, \nu_j) = \frac{\exp(\nu_i + \nu_j + \zeta z_i' z_j)\sum_i \mathbb{E}[d_i]}{\sum_{i,j}\exp(\nu_i + \nu_j + \zeta z_i' z_j)}.$$

Normalizing in this way ensures $\sum_i \mathbb{E}[d_i] \triangleq \sum_i \sum_j \Pr(g_{ij} = 1|z_i, z_j, \nu_i, \nu_j)$. Since the formation model assumes that the propensities to form a ties between pairs are conditionally independent given the latent variables, we can now generate graphs by taking draws from a Bernoulli distribution for each pair with probability defined by $\Pr(g_{ij} = 1|z_i, z_j, \nu_i, \nu_j)$.

3.3.5 Discussion

We have provided a simple algorithm to go from ARD questions to draws from the posterior distribution of the graph that would have given rise to ARD answers by respondents with

characteristics similar to those we observed in the data. The model leverages a latent surface model similar to Hoff et al. (2002), used in McCormick and Zheng (2015), which is intimately related to the β-model studied in Chatterjee and Diaconis (2011) and Graham (2017). One issue that has arisen from both the Bayesian and frequentest perspectives is the notion of density in the limit, or the rate at which the number of edges grows compared to the number of nodes. The Bayesian paradigm uses the Aldous-Hoover Theorem (Hoover, 1979; Aldous, 1981) for node-exchangeable graphs to justify representing dependence in the network through latent variables, though this theorem only gives the existence of a latent variable representation and not the specific form we use. The exchangeability assumption implies that a graph can be sparse if and only if it is empty (Lovász and Szegedy, 2006; Diaconis and Janson, 2007; Orbanz and Roy, 2015; Crane and Dempsey, 2015). From a frequentist perspective, Chatterjee and Diaconis (2011) show that the individual fixed effects (corresponding to, for example, gregariousness) can only be consistently estimated when the network sequence is dense.

In contrast to this previous work, however, we assume that our sample of egos arises from a population with fixed n. That is, in our paradigm there is a network of finite size, n, and we observe a small m number of actors. We see the reliance on this assumption in, for example, our expression relating degree to the individual heterogeneity parameters, ν_i. Put a different way, there is no asymptotic sequence of networks. The number of edges in a graph still impacts estimation, however. Even when the number of nodes is large, we do not expect d_i to uniformly converge to $\mathbb{E}[d_i]$ if the graph is not dense. This additional variability propagates through the model and inflates the posteriors of ν_i. These may be quite poor in practice, though it is difficult to derive the finite sample distribution. Nonetheless, what this suggests is that in cases where the network is too sparse, the ARD approach may be uninformative, and the researcher will see this plainly. This is the case for two reasons. First, by definition, anyone in the ARD sample will know fewer alters with trait k since the network

has fewer links on average. Second, there will be too much variation in our location estimates and degree estimates, which then will also affect our node heterogeneity estimates. This means that when the researcher faces rather diffuse posteriors, the network may be too sparse to convey much information.

3.4 Identification

In this section, we start with providing a simple intuition of how model parameters in (2.2) are identified, followed by a formal theorem. We present the proof in Section 3.4.1.

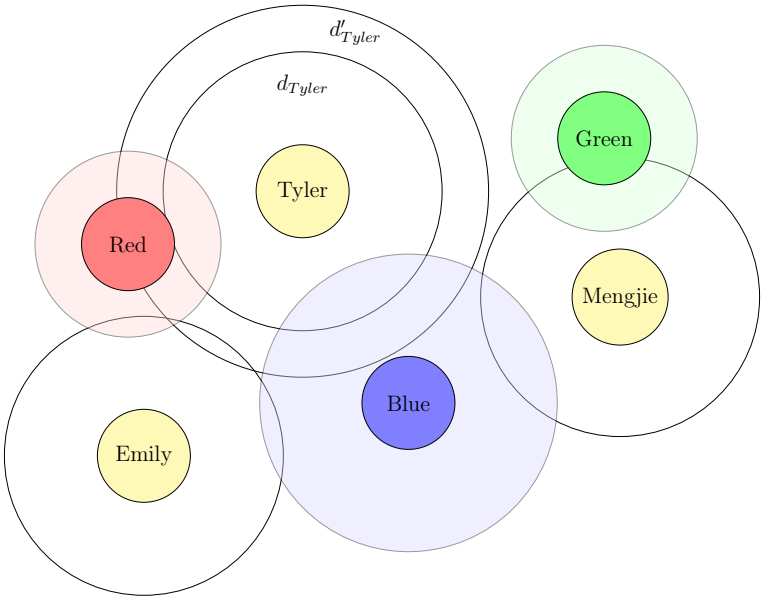


Figure 3.1: Identification of υ_k and η_k for $k \in \{\text{Red}, \text{Blue}, \text{Green}\}$ holding fixed locations and degrees of nodes in the ARD sample. Identification of $\mathbb{E}[d_i]$ holding fixed locations and concentration parameters.

Figure 3.1 shows how the location υ_k and the concentration η_k for category k is intuitively identified assuming the latent geometry is a plane. Holding the location of three nodes fixed (here Tyler, Emily and Mengjie), and holding fixed their degree, the relative locations of categories (here Red, Green, and Blue) can be identified by placing their centers and

controlling the concentration to match the Poisson rates observed in the ARD. To see that the concentrations of the Red, Green, and Blue trait groups are identified, consider what would happen if we changed the concentration of one of the groups. If we increased the concentration of the Blue group (i.e., decreased the variance), then we would need to move Mengjie (and Tyler and Emily) closer to the Blue group to preserve the overlap between Emily's disc and the Blue group. Moving Emily closer to the Blue group, though, necessitates moving her away from the Red group, reducing her overlap with the Red group. We could try to compensate by decreasing the concentration (increasing the variance) of the Red group. We can't do this, though, because doing so would change the overlap between Tyler's disc and the Red group. Similarly the figure shows how the $\mathbb{E}[d_{Tyler}]$ can be identified holding fixed the location and concentration of the various categories, since this affects $\lambda_{Tyler,k}$. Because the likelihood only depends on the latent space through the distances between individuals and groups, we fix the location of the center a small number of groups to address the invariance to distance-preserving rotations.

To see the formal statement, it is useful to recall that we say two points on a sphere are *antipodal* if there are indefinitely many great circles passing through them.

ASSUMPTION **3.4.1.** *$K > 3$ and the centers of the von Mises-Fisher distributions representing three of the alter groups are fixed.*

ASSUMPTION **3.4.2.** *The fixed centers are not all on the same great circle.*

ASSUMPTION **3.4.3.** *For some k, k', $\eta_k \neq \eta_{k'}$.*

ASSUMPTION **3.4.4.** *$\zeta > 0$.*

THEOREM **3.4.1.** *Under Assumptions 3.4.1-3.4.4, for any n by K matrix of ARD responses $\mathbf{Y}$, we have that $\mathcal{L}(d_i, b_k, \zeta, \eta_k, \theta_{(z_i,\upsilon_k)}; \mathbf{Y}) = \mathcal{L}(d_i, b_k, \zeta', \eta'_k, \theta'_{(z_i,\upsilon_k)}; \mathbf{Y})$ only if $\eta_k = \eta'_k$, $\theta_{(z_i,\upsilon_k)} = \theta'_{(z_i,\upsilon_k)}$, $\zeta = \zeta'$, $\nu_i = \nu'_i$ and $z_i = z'_i$.*

Assumption 3.4.1, that $K > 3$ and are fixed, is innocuous. The content of Assumption 3.4.2 is as follows. Let the traits be "red", "blue", and "green". If you know the likelihood of say a "red" and a "blue" type linking on average (i.e., distance between the centers) and you know the likelihood of a "red" and a "green" type linking on average, it does not entirely determine the likelihood of a "blue" and "green" linking. Practically this means that essentially knowing two features (someone having a migrant, someone having a 10th standard pass family member) does not determine the third (on average).

Assumption 3.4.3 requires that at least one trait has a different concentration parameter. In some sense both 3.4.2 and 3.4.3 can be interpreted as ruling out "measure zero" events if one thinks of of trait centers and concentration parameters themselves being generated according to any smooth distribution on a sphere. Assumption 3.4.4 means that the latent space has content for the model (by assumption $\zeta \neq 0$): distance in the space indeed reduces the odds of being linked. Put another way, it means that there is network structure not explained by the individual effects.

3.4.1 Proof

In this section, we present a proof of Theorem 3.4.1. Essentially, we need three latent group centers to be fixed and to have distinct positions on the hypersphere. We also need to know the trait status of at least some individuals and for there to be at least some individuals with more than one trait. This is sufficient to identify the parameters governing the locations of each of the types and the concentration parameters (**Proposition** 3.4.1). If we assume that trait status is unrelated to gregariousness (which is necessary for the derivation of the likelihood anyway) then we can identify the coefficient ζ (**Proposition** 3.4.1). Based on ζ and degree d_i (which is identified as described in McCormick and Zheng (2015) using the latent trait group sizes) we can identify the individual gregariousness parameters (**Proposition**

3.4.2). All that is left are the individual level latent positions, which we show can be identified based on the previously described parameters (**Proposition** 3.4.3).

We begin by defining terms necessary to describe the spherical geometry and then provide the necessary conditions. Throughout the proofs here we will assume a latent sphere centered at the origin.

PROPOSITION **3.4.1.** *Considering the Assumptions 3.4.1-3.4.4, trait centers υ_k for $k = 4, ..., K$, concentration parameters η_k for $k = 1, ..., K$, and ζ are identified.*

Proof The von Mises-Fisher distribution is a symmetric unimodal distribution with probability mass declining in distance from the center, υ, tuned by concentration parameter η. For each individual we know their latent trait group(s). This is a fundamental distinction between our setting and that of McCormick and Zheng (2015), who typically do not assume this information is known. We can think of the positions of each individuals as draws from one or more of the von Mises-Fisher distributions on the sphere. An individual who belongs to two trait group has to be at the intersection of the densities of the two trait groups. Knowing the fraction of individuals who have both traits, therefore, intuitively tells us something about the overlap between the densities of the two trait groups. Throughout this proof keep in mind that we are not using the specific locations of individuals (which we only show is identified in a subsequent proposition), but rather the density defined by the overlap between trait groups.

More formally, define the lens, $\ell(A, B)$, as the expected share of individuals drawn from this distribution who have traits A and B. Equivalently, we can think of this as the volume of the overlap between the densities of the two distributions for all individuals up to a pre-specified, but arbitrary, cumulative probability. In general let $\ell(A_1, ..., A_k)$ denote the expected share of individuals drawn who have all traits. We can treat all lenses as observed in the data because for a large m, we know the traits that every node has.

For notational convenience and without loss of generality, we will assume that the fixed group centers correspond to the first three latent trait groups, $\upsilon_1, \upsilon_2, \upsilon_3$. Observe that this immediately implies all three η_k for $k = 1, ..., 3$ are identified. For the sake of argument assume that η_1 is known. Then from $\ell(1, 2)$ we have that η_2 is identified. Given η_2, from $\ell(2, 3)$, we have η_3 identified. But we can of course identify η_1 similarly from η_3. This logic applies because we can map the overlapping section, $\ell(1, 2)$, into specific values of the cumulative distribution function of the von Mises-Fisher distributions. If we change η_2, then the location of individuals' latent positions that are draws from this distribution must also change. Changing these locations changes the boundary of $\ell(1, 2)$. Similarly, changing the boundary of $\ell(1, 2)$ implies a change in the densities of the von Mises-Fisher distributions for the first and second traits. Since the centers of these distributions are fixed any change in the distribution must come through the concentration parameter.

Further, this solution is unique. To see this, assume that we are at some unique solution η_1, η_2, η_3. Consider an alternative value of any combination of concentration parameters. Clearly all concentration parameters cannot increase because then the lenses would not match the true lenses. Consider then the case where at least one η_k declines. In this case, if $\eta_{k'}$ were not to increase, then $\ell(k, k')$ would not match the expectation observed in the data. Consequently, $\eta_{k'}$ must increase. In this case, should $\eta_{k'}$ increase, then $\eta_{k''}$ must decline to preserve $\ell(k', k'')$. But in this case, the lens $\ell(k, k'')$ must increase as both concentration parameters have declined. Therefore the solution is unique.

To see why ζ is identified, consider any two k, k' with $\eta_k \neq \eta_{k'}$. Because we know the respective von Mises-Fisher distributions for each trait, we can compute the ratios of the expectations of (2.2) conditional on each type k and k', plugging in for d_i from (2.3). Because the individual effects are drawn independently of trait by assumption, all terms that depend on ν_i drop since the distribution of ν_i is independent of trait type, so they have the same

expectations irrespective of k or k'. As such

$$\frac{\mathbb{E}_i[\lambda_{ik}|i \in G_k]}{\mathbb{E}_j[\lambda_{jk}|j \in G_{k'}]} = f(b_k, b_{k'}, \eta_k, \eta_{k'}, \zeta)$$

where the right hand side is a known function that comes from taking these ratios. The only unknown is ζ. There is a unique solution to the equation—we leave the algebra to the reader—but can be seen from the fact that the link probability is monotonically declining in ζ and faster for lower η_k, holding all else fixed, so the ratio term also is monotone in ζ.

PROPOSITION **3.4.2.** *Considering the conditions above, ν_i for $i = 1, ..., m$, individual gregariousness effects for the entire ARD sample, are identified.*

Proof By Proposition 3.4.1, the υ_k and η_k and ζ are identified. By Equation (2.2), d_i can be obtained and by Equation (2.3) we have for every $i = 1, ..., m$ in the ARD sample an equation relating the fixed effect ν_i to the degree d_i. We have m equations and m unknowns.

To see why the solution is unique consider fixing for the moment some ν_1 without loss of generality. In this case, we can write $\nu_i = h_i\nu_1$ for every i, where h_i is the ratio of the degrees between person i and person 1. Then we can write

$$\exp(\nu_1)(\frac{1}{n}\sum_i \exp(h_i\nu_1)) = \frac{d_1}{m \cdot \frac{C_{p+1}(0)}{C_{p+1}(\zeta)}}.$$

This is a monotone function in ν_1 and has a unique solution, which then identifies the remainder of the ν_i as well scaling by h_i.

PROPOSITION **3.4.3.** *Considering the conditions above, the latent locations z_i for $i = 1, ..., m$ for the entire ARD sample, are identified.*

Proof From Propositions 3.4.1 and 3.4.2, we have identified all parameters except for z_i. To show this result, we first state two results from spherical geometry. The proofs of these results are available in standard texts (e.g. Biringer (2015)).

Result: *The great circle between two points is unique unless the points are antipodal.*

Result: *There are exactly three isomorphisms for spherical geometry.*

The first result defines a unique distance from each respondent latent position and at least two of the three latent group means. A respondent position can be antipodal with one of the three fixed groups, but then cannot be with the two others because the three groups are not on the same great circle.

The second result limits the number of possible operations that threaten identifiability. Recall that, if an operation changes the latent distance between an point and the center of a group, then the operation will also change the likelihood. Thus, if we show that we cannot perform any of the three possible distance preserving transformations on the sphere after fixing group centers, then we have also completed the proof.

We consider two cases, the first takes and arbitrary point that is not antipodal to any of the latent centers, whereas the second case considers any point that is antipodal with one latent center.

Case 1. Since we fix three centers which are not on a great circle, we cannot do any reflections of points without changing the distance to one of the centers. For rotations, consider centers υ_1 and υ_2, and a point z_i. Since υ_1 and υ_2 are not antipodes, if we rotate z_i around center υ_1 and keep $d(z_i, \upsilon_1)$ the same, it is possible that $d(z_i, \upsilon_2)$ changes. The points z_i, z_i' such that $d(z_i, \upsilon_1) = d(z_i', \upsilon_1)$ and $d(z_i, \upsilon_2) = d(z_i', \upsilon_2)$ are reflections over the plane that intersects υ_1 and υ_2 in a great circle. z_i and z_i' have equal distance to any point on this great circle, and unequal distance to any point not on this great circle. Since the third center υ_3 is not on this the great circle that intersects υ_1 and υ_2, $d(z_i, \upsilon_3) \neq d(z_i', \upsilon_3)$.

Case 2. When we change the point's position, then the distance between that point and the antipodal latent center decreases.

This completes the proof.

Proof of Theorem 3.4.1 Under the Assumptions 3.4.1-3.4.4, this is a direct corollary to Propositions 3.4.1, 3.4.2, and 3.4.3.

3.5 Empirical Applications

We now present two empirical applications that use ARD techniques. They build upon prior work by the authors, in part. The goal is to illustrate here that a researcher could have done this sort of economic analysis using ARD only, equipped with our method.

The first example looks at what would have happened if the researchers had obtained ARD for an experiment on savings and reputation. The second example actually looks at a setting where survey ARD was collected.

3.5.1 Encouraging savings behavior in rural Karnataka

Our first application builds on Breza and Chandrasekhar (2019). The authors study social reputation through the lens of savings. In a field experiment, savers set 6-month targets for themselves. They do so knowing they may be assigned a "monitor," a villager who will be notified biweekly about their progress. Progressing towards a self-set target exhibits more responsibility, providing an avenue for the saver to build reputation with the monitor and others in the community. In 30 villages, monitors are randomly assigned to a subset of savers. This generates variation in the position of the monitor in the network. Because the monitor is free to talk to others, information about the saver's progress and reputation may spread. A signaling model on a network guides the analysis: if the saver is more central, information can spread more widely, and if the saver is more proximate to the monitor, information likely

spreads to those with whom the saver is more likely to interact in the future. For saver i and monitor j, the model shows that the network matters for signaling through the quantity.

$$q_{ij} = \frac{1}{n}\text{Monitor Centrality} \times \text{Saver Centrality} + n \cdot \text{Proximity of Saver-Monitor}.$$

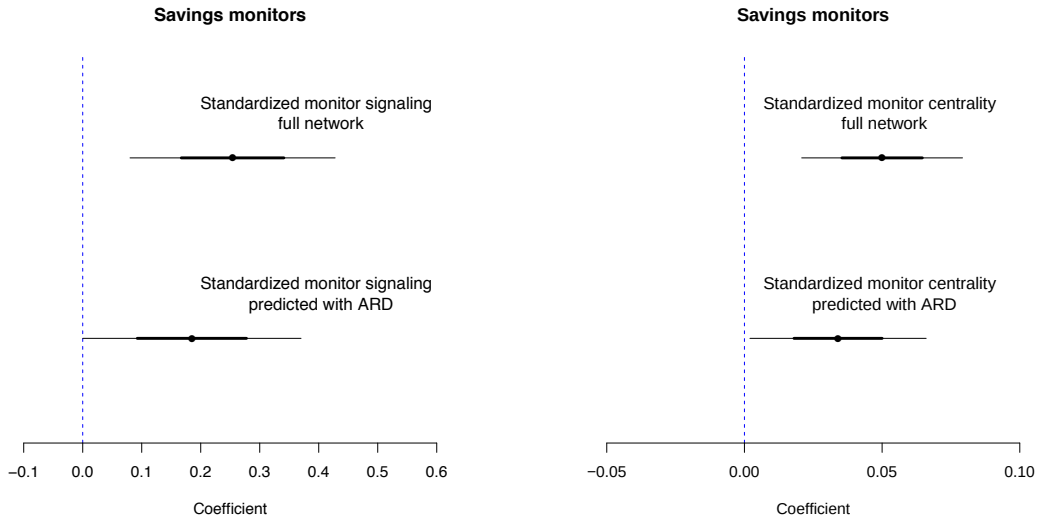
Formally, Breza and Chandrasekhar (2019) show

$$q_{ij} = \frac{1}{n}\sum_k p_{jk} \sum_k p_{ik} + n \cdot (p_{\cdot i}.p_{\cdot j})$$

Here $p_{ij} \propto \left[\sum_{t=1}^{T}(\theta g)^t\right]$ is the probability that a unit of information that begins with i is sent to j, where transmission across each link happens with probability θ. Banerjee et al. (2016b) shows that for sufficiently high T, $\sum_k p_{jk}$ converges to the eigenvector centrality of j. Breza and Chandrasekhar (2019) shows that in equilibrium, only when q_{ij} is sufficiently high does the saver actually save.

Breza and Chandrasekhar (2019) have near-full network data (from the Banerjee et al. (2016b) sample), allowing them to calculate $q_{i,j}$. They find that randomly-selected monitors increase household savings across all accounts by 35%. Consistent with the model, a one-standard deviation increase in q_{ij} leads to an additional 29.6% increase in total savings. Additionally, 15 months after the end of our savings period, they show that reputational information spread: randomly selected individuals surveyed about savers in the study were more likely to have updated correctly about a saver's responsibility when the saver was randomly assigned a more central monitor. Moreover, the savings increase persisted, and in the intervening 15 months, monitored savers were better able to cope with shocks.

How would our conclusions have changed if Breza and Chandrasekhar (2019) only had access to ARD and not the full network maps? Table 3.1 presents regressions of the log of total household savings across all household accounts against the model-based measure of



(a) Monitor signaling **(b)** Monitor centrality

Figure 3.2: Estimates of regression coefficients plus/minus one and two standard errors are plotted. The left plot is for a regression of total log savings across all household accounts on monitor signaling and the right plot is for a regression of monitor's belief about saver's responsibility on monitor centrality value. With all 95% confidence intervals on the right side of the zero vertical line, we show that we would have reached the same conclusions using estimated monitor centrality and signaling from ARD, as researchers who have collected network data.

	(1) Log Total Ending Savings	(2) Log Total Ending Savings
Signaling value of monitor with full network data (q_{ij}), Standardized	0.254 (0.0869)	
Predicted signaling value of monitor with ARD (q_{ij}), Standardized		0.185 (0.0925)
Observations	422	422
Number of villages	30	30

Notes: Standard deviation of village-level block bootstrap in parentheses.

Table 3.1: Log total savings across all household accounts regressed on monitor signaling value

how much signaling value the monitor provides the saver, q_{ij}. Top line of Figure 3.2a shows the 95% credible intervals of regression coefficient using true q_{ij}, while the bottom line is the interval with estimated q_{ij} from ARD. We construct ARD estimates by taking samples from

the posterior distribution and then using the average estimated q_{ij} across those posterior draws. In the experiment we showed that a one standard deviation increase in q_{ij} due to random assignment of the monitor led to a 25.4% increase in total household savings (column 1). In column 2 we show that even if we did not have the network data, if we had ARD alone for a 30% sample, we would have had a very similar conclusion, inferring that a one standard deviation increase in predicted q_{ij} corresponds to a 18.5% increase in total household savings across all accounts. With both 95% confidence intervals on the right side of the zero vertical line in Figure 3.2a, we show that we would have reached the same conclusions using estimated monitor centrality and signaling from ARD, as researchers who have collected network data. Said differently, we could have used ARD questions to easily pick good monitor-saver pairs.

	(1) Belief about saver's responsibility	(2) Belief about saver's responsibility
Monitor centrality with full network data, Standardized	0.0500	
	(0.0146)	
Predicted monitor centrality with ARD, Standardized		0.0340
		(0.0160)
Observations	4,743	4,743
Number of villages	30	30

Notes: Standard deviation of village-level block bootstrap in parentheses. "Responsibility" is constructed as 1(Saver reached goal)*1(Respondent indicates saver is good or very good at meeting goals) + (1-1(Saver reached goal))*1(Respondent indicates saver is mediocre, bad or very bad at meeting goals). See Breza and Chandrasekhar (2019) for further details.

Table 3.2: Beliefs about savers and monitor centrality

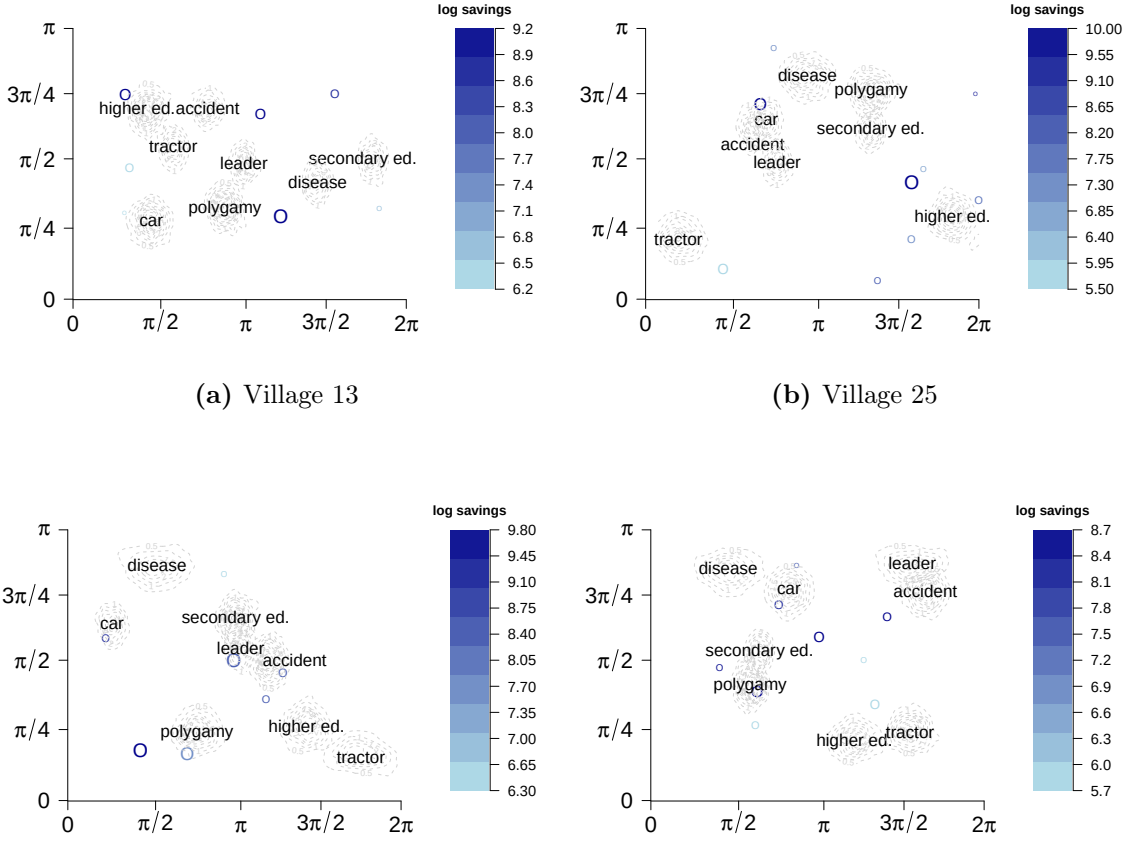
As a further examination of our approach, we repeat the same exercise using another specification from Breza and Chandrasekhar (2019). Table 3.2 shows the results of a regression where we the outcome is the respondent's belief about the saver's responsibility and the regressor is the monitor's centrality. Observing the complete network, a unit increase in the monitor's centrality corresponds to about a 5% increase respondent's belief about saver responsibility. Using ARD, we would estimate an increase of about 3.4%, leading (as in the previous example) to the same substantive conclusions. Similarly in Figure 3.2b, we see that both 95% confidence intervals lie right of the zero vertical line, indicating the same conclusion.

This application also gives us an opportunity to visualize how network characteristics map to the latent space representation. In Figure 3.3, we plot the locations and concentrations of the ARD traits for four sample villages that were part of the Breza and Chandrasekhar (2019) savings study. We then overlay the positions in the latent space of the individuals participating in the experiment as monitors, depicted as rings. The size of the ring depicts the monitor's eigenvector centrality. Finally, we color the monitor rings to indicate the savings performance of the saver to whom each monitor was randomly allocated – darker shades depict higher levels of savings.

As Breza and Chandrasekhar (2019) find, there appears to be a relationship between monitor centrality (here denoted by larger rings) and the saver's performance (here given by darker colors). This is consistent with the theory that more central monitors under the signaling model generate larger incentives for the saver to save. Furthermore, the visualization demonstrates that the larger rings tend to be located closer to the centers of traits or between centers of traits. That is, they are closer to the center of masses of clusters of types of individuals. This makes sense as this means that the latent location of a central monitor will tend to be closer to many more other individuals.

3.5.2 Impact of microfinance in Hyderabad

The goal of our final example is to demonstrate to the reader a context in which we collected and use only ARD survey questions in our analysis. We first demonstrate that the researcher could have obtained the same conclusions using the ARD instead of the network data that was collected in this study. But because the network data was incomplete (specifically the authors only measured degree – the number of links but not the identities – and support – how many links had a friend in common), the researchers could not ask how their intervention impacted the network more generally. Using ARD techniques, we show what conclusions the



(a) Village 13 **(b)** Village 25

(c) Village 54 **(d)** Village 55

Figure 3.3: Sample latent locations of randomly assigned monitors by centrality and the savings of the their respective savers. Monitors with higher eigenvector centrality have larger rings. The color of the ring indicates the savings performance of the saver to whom each monitor was randomly assigned, with darker colors indicating higher savings levels. This illustrates the pattern that more central monitors corresponded to higher levels of savings.

researchers could have learned about how the network was affected by the intervention only using the ARD survey data and estimates from the surveys of each neighborhood's average degree.

This example concerns the introduction of microfinance in Hyderabad, India. In Banerjee

et al. (2015), the authors study a randomized controlled trial where microfinance was introduced randomly to 52 out of 104 neighborhoods in Hyderabad. Banerjee, Breza, Duflo, and Kinnan (2016a) look at longer run outcomes 6 years after the intervention. This example is useful for two reasons. First, it is an urban setting where the researchers have no hope of obtaining full network data. Second, it shows how we may measure the effect of economic interventions on social network structure, as predicted by theory, despite not having network data.

In the original paper, Banerjee et al. (2016a) measure each node's within-neighborhood degree and support, defined as the fraction of links between the respondent and a connection such that there exists a third person who is linked to both nodes in the pair. They find that both degree and support decrease with the treatment. Note that they did not get any subgraph data since the links were not matched to a household listing: degree and support can be thought of as just two numbers.

Banerjee et al. (2016a) also collected ARD data, which we use here. In particular, a sample of approximately 55 nodes in every neighborhood was surveyed and demographic covariates as well as ARD were collected for this entire sample. As before, we fit a network formation model using the ARD data and this sample of nodes. In this application we use the survey responses for degree and input each graph's estimated average degree directly into the model.

We explore whether microfinance affects network structure by regressing

$$y_v(g) = \alpha + \beta \text{Treatment}_v + \epsilon_v$$

where v indexes neighborhood and Treatment_v is a dummy for treatment neighborhoods. Our outcome variable $y_v(g)$ of interest is the rate of support.

Theory is silent on whether density should increase or reduce, whether triadic closure

(clustering or support) should increase or reduce, which can depend on a number of things: for instance, whether relending or autarky forces affect the incentives to maintain risk-sharing links Jackson et al. (2012).

	(1) Percent Supported (Data)	(2) Percent Supported (Estimate)	(3) Graph-level Proximity (Estimate)
Treatment Neighborhood	-0.0655	-0.0901	-0.0515
	(0.0317)	(0.0551)	(0.0139)
Constant	0.4427	0.4364	0.4536
	(0.0633)	(0.094)	(0.0082)
Mean of the response variable	0.3893	0.3120	0.4267
Observations	3,458	3,539	61

Notes: Standard deviation of village-level block bootstrap in parentheses. Sample includes neighborhoods with estimated sampling rate $\geq 20\%$. For large number of excluded low sampling rate neighborhoods, the population count is top-coded at 500 households. For these very large neighborhoods, we calculate the sampling rate using a population of 500. The outcome variable of columns 1 and 2 is the share of links that are supported and in column 3 it is the average proximity in the graph.

Table 3.3: Network statistics regressed on treatment

Table 3.3 reports the regression results. Column 1 replicates the specification from Banerjee et al. (2016a) that past exposure decreased support. Column 2 presents the same regression, but using estimated support. The estimates of the treatment effects along with the levels of support (the regression constant) are quite similar. Figure 3.4 shows estimates of treatment effect on percent supported, plus/minus one and two standard errors, where we also see that treatment effects are similar, with more variation using estimated support as the outcome of the regression. We view this exercise as a "validation" of the ARD-based model. The fact that estimated support matches measured support quite well is especially reassuring given that triadic closure is exactly the type of network statistic that the Hoff model may have a hard time replicating.

Given that the estimated treatment effect looks quite similar using the different support measures, in Column 3, we present the results of a graph-level regression, using proximity (the average inverse path length in the network) as the outcome variable. Note that it was not

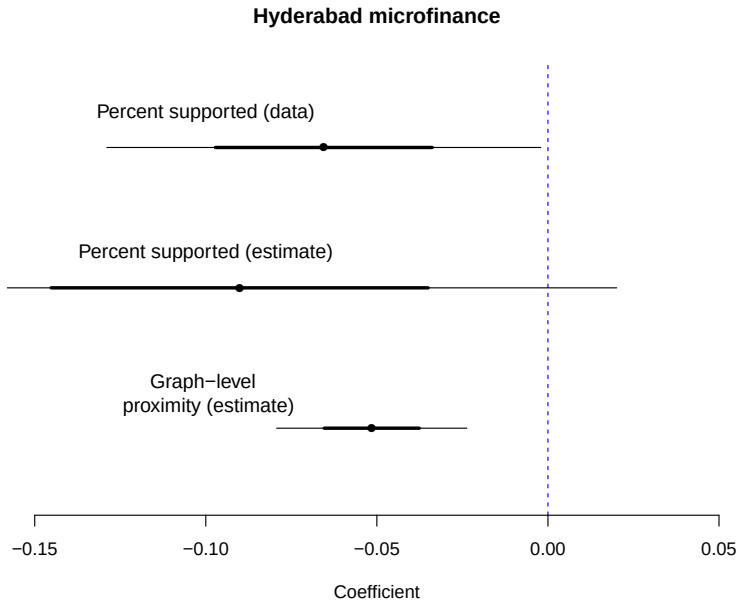


Figure 3.4: Estimates of treatment effect, plus/minus one and two standard errors are plotted. The top line represents when measured percent supported is used in regression, the middle line represents when estimated percent supported is used, and the bottom line represents estimated graph-level proximity is used as the outcome variable.

possible for the authors to collect such a statistic using their surveys. We find that estimated proximity decreases, meaning that the decline in links due to microfinance exposure leads to larger average distances between households in the community. This treatment effect is significant since we see the 95% confidence interval is on the left of zero vertical line in Figure 3.4. This exercise demonstrates how our method may be useful to researchers seeking to study the evolution of networks, without requiring full network data.

3.6 Discussion

Our method is not without limitations, and we have highlighted two issues that should be considered when using our method for applied research. While a detailed theoretical study is

beyond the scope of the chapter, we discuss briefly how applied researchers might navigate these limitations.

First, the method produces a distribution of networks that are consistent with the estimated network formation model – we do not learn about the specific realization that generated the observed graph. For example, of course we can never say whether a given link exists. This means that network features that rely on the existence of specific links will not be captured well. If the research question requires knowledge of specific links, then the researchers should ask about these relationships directly when possible. This intuition also suggests that features such as betweenness centrality – which rely on specific paths – may be hard to capture with the method. That being said, our Savings Monitors example shows that the model can do well at capturing more recursive notions of centrality such as diffusion and eigenvector centrality. Thus, the method should still do well in cases where betweenness centrality is highly correlated with these other measures. Finally, appealing to the result of Chandrasekhar and Lewis (2016), if inference is being conducted across many independent networks, then these issues are of much less concern. In fact, in Chapter 4, we present theoretical and simulation evidence showing that working with the expected graph- or node-level measure is no different than working with the actual realized measure.

Second, the method relies upon a parametric network formation model. If that model is not a good representation of the network of interest, then the resulting ARD estimates may be biased. As mentioned above, one might be particularly concerned about the ability of the Hoff model to capture the level of clustering. However, as we show in our Hyderabad microfinance example, the model actually does well in practice at predicting the change in the level of link support, a related notion. We recommend that applied researchers follow this empirical example and also elicit network support directly from survey respondents. The researcher can then "validate" the ARD method for the specific applied context by estimating

support using ARD and comparing the estimates to the true values.

Finally, in Chapter 4, we study how the quality of the estimation of network features varies by statistic. To summarize these results briefly, we find that the method works quite well for many empirically relevant network features both at the node and network-level. At the network level it performs well when we look at average path length, maximal eigenvalue of the adjacency matrix, graph-level clustering, whereas it does poorer when estimating the number of components in the network. At the node level, degree, eigenvector centrality, among other features perform well, whereas node-level clustering and betweenness centrality performs worse, and existence of a link performs worst.

Chapter 4

CONSISTENT ESTIMATION OF GRAPH STATISTICS USING AGGREGATED RELATIONAL DATA[1]

4.1 Introduction

Following the method presented in Chapter 3, we now present theoretical results and simulation evidence on what node- and network- level measures can be recovered well using ARD alone without any network observation. We have shown in Section 3.5 that using ARD, we would have reached the same conclusions as field economists who have collected network data. In the empirical examples, we focused on individual measures such as centrality and support, and network level measures such as proximity. However, researchers may be interested in other measures as well. Therefore, the aim of this chapter is to offer a guideline to practitioners on what measures can be estimated well using ARD alone.

In section 4.2, we present a set-up of the problem, the network formation model, and briefly discuss our estimation aims. In section 4.3, we present a theorem with proof, which states that the maximum likelihood estimator of network formation model parameters is consistent. This sets the ground for all our propositions in Section 4.4, where we assume the parameters are known. Our theoretical results concerns two contexts: estimating features of the underlying, unobserved network structure itself from a single large network, and estimating how changes in network features correspond to changes in socio-economic outcomes when multiple independent networks are observed. In Section 4.5, we present simulation evidence that is consistent with our theoretical results. Finally, in Section 4.6 we discuss limitations of

[1]The contents of this chapter are based on the paper Breza et al.

our approach and future directions for improvements.

4.2 Overview

In this section, we restate our setup of ARD and the network formation model. Then we briefly describe our estimation aims in two contexts.

4.2.1 Aggregated Relational Data

Recall that an undirected, unweighted graph, $\mathbf{g} = (V, E)$ consists of a vertex set V and edge set E, with $n = |V|$ nodes, with $g_{ij} = \mathbf{1}\{ij \in E\}$ denoting existence of a link.

Researchers have a sample of Aggregated Relational Data (ARD) from $m \leq n$ nodes. The sample is selected uniformly at random. An ARD response addresses a question of the form "How many nodes with trait k are you linked to?", given by y_{ik} and assumed to be $y_{ik} = \sum_{j \in G_k} g_{ij}$ where $G_k \subset V$ consists of all nodes with trait k. We assume there are $K > 3$ such traits. Henceforth let $\mathbf{Y}$ denote the $m \times K$ matrix of ARD responses.

4.2.2 Latent Surface Model

We use the latent surface model from McCormick and Zheng (2015) (see also Hoff et al. (2002) among others). The network g is drawn from a distribution given by.

$$\Pr(g_{ij} = 1 | \nu_i^0, \nu_j^0, \zeta^0, z_i^0, z_j^0) \propto \exp(\nu_i^0 + \nu_j^0 - \zeta^0 d(z_i^0, z_j^0)). \tag{4.1}$$

Equation (4.1) is equivalent to Equation (2.1), with superscript 0 denoting the true parameters. Here $\nu_i^0 \in \mathcal{V}$ are person-specific random effects that capture heterogeneity in linking propensity and $\mathcal{V} \subset \mathbb{R}$ is compact. The latent positions of nodes are on the surface of

p dimensional hypersphere, $\mathcal{S}^p$. In what follows, we let $\boldsymbol{\theta}^0 = (\nu_1^0, \ldots, \nu_n^0, z_1^0, \ldots, z_n^0, \zeta^0)$. In Section 4.3, we show that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^0$ is consistent.

4.2.3 Estimation Aims

We are interested in when researchers are able to accurately estimate network features or parameters of interest. We investigate this both theoretically and empirically in two contexts:

1. Can researchers consistently estimate features of the underlying, unobserved network? Examples include centrality measures or clustering for nodes. This analysis studies the case of a single large network.

2. Can researchers consistently estimate how changes in network features correspond to changes in socio-economic outcomes, or how an intervention might affect the structure of the network. This analysis studies the case of many independent networks.

4.3 Consistency Of MLE For Model Formation Parameters

In this section, we show that a researcher who observes ARD can expect to consistently recover the parameters of the network formation in (4.1) with a sufficiently large graph. Our argument builds on work by Shalizi and Asta (2017) who show that, when the complete graph is observed, it is possible to consistently estimate the latent components of the formation model. We extend their result in two ways. First, we show that the consistently argument applies not just to the latent component but to the entire parameter vector, $\boldsymbol{\theta}^0$ (Lemma 4.3.1). Second, we adapt the proof so that the consistency extends to cases where, rather than observing the entire graph, a researcher observes ARD (Theorem 4.3.1).

LEMMA **4.3.1.** *Suppose that we observe data generated from a formation model noted as a*

Continuous Latent Space (CLS) model in Shalizi and Asta (2017). Specifically, let

$$\Pr(G_{ij} = 1|\nu^0, z^0, \beta^0) \propto \exp(\nu_i^0 + \nu_j^0 + \beta^{0\prime} X_{ij} - d_{\mathcal{M}^p}(z_i, z_j))$$

Let $X_{ij} \in \mathbb{R}^h$ *so* $\beta \in \mathbb{R}^h$. *Let* $V^n \subset (-\infty, 0)^n$ *a compact subset. Then, under the same conditions as Shalizi and Asta (2017), we have*

$$(\hat{\nu}, \hat{z}, \hat{\beta}) \xrightarrow{p} (\nu^0, z^0, \beta^0)$$

where

$$(\hat{\nu}, \hat{z}, \hat{\beta}) = \underset{\nu,z,\beta \in V^n \times \mathcal{M}^p \times \mathbb{R}^h}{\operatorname{argmax}} \ell(\nu, z, \beta),$$

the maximum likelihood estimates.

Proof of Lemma 4.3.1

Recall

$$\Pr(G_{ij} = 1|\nu^0, z^0, \beta^0) \propto \exp(\nu_i^0 + \nu_j^0 + \beta^{0\prime} X_{ij} - d_{\mathcal{M}^p}(z_i, z_j))$$

and our goal is to show that $(\hat{\nu}, \hat{z}, \hat{\beta}) \xrightarrow{p} (\nu^0, z^0, \beta^0)$, where $(\hat{\nu}, \hat{z}, \hat{\beta})$ are the maximum likelihood estimates. Our proof mirrors Shalizi and Asta (2017), with the exception that we include individual effects, ν and coefficients β. For simplicity, throughout the proof we use single variable notation to denote equivalence classes based on latent distances. As Shalizi and Asta (2017) note, we can only identify equivalence classes up to isometries.

We want to show the following:

1. Identification: the true parameters maximize the expected likelihood

$$(\nu^0, z^0, \beta^0) = \underset{\nu,z,\beta \in V^n \times \mathcal{M}^p \times \mathbb{R}^h}{\operatorname{argmax}} \mathbb{E}[\ell(\nu, z, \beta)]$$

2. Uniform convergence of the observed likelihood to its expectation:

$$\sup_{\nu,z,\beta} |\ell(\nu, z, \beta) - \mathbb{E}[\ell(\nu, z, \beta)]| \xrightarrow{p} 0.$$

We first establish identification. Our proof here follows Lemma 13 of Shalizi and Asta (2017). The likelihood is

$$\Pr(G|\theta) = \prod_{i<j} p_{ij}(\nu, z, \beta)^{G_{ij}} (1 - p_{ij}(\nu, z, \beta))^{1-G_{ij}}$$

where we denote $p_{ij}(\nu, z, \beta) \propto \exp(\nu_i + \nu_j - d_{\mathcal{M}^p(\kappa)}(z_i, z_j) + \beta' X_{ij})$. The log-likelihood is then

$$\begin{aligned}
\ell(\nu, z, \beta) &= \binom{n}{2}^{-1} \sum_{i<j} G_{ij} \log(p_{ij}(\nu, z, \beta) + (1 - G_{ij}) \log(1 - p_{ij}(\nu, z, \beta)) \\
&= \binom{n}{2}^{-1} \sum_{i<j} \log(1 - p_{ij}(\nu, z, \beta)) + G_{ij} \log\left[\frac{p_{ij}(\nu, z, \beta)}{1 - p_{ij}(\nu, z, \beta)}\right]
\end{aligned}$$

The last term in the above expression corresponds to $\lambda_n(\nu_i, \nu_j, z_i, z_j, \beta)$ given in (17) of Shalizi and Asta (2017).

We will establish identification using the cross-entropy. For further description of cross-entropy and the decomposition used below, we refer the reader to Cover and Thomas (2012). We first show that the expected log likelihood is equal to the cross-entropy. First, cross-entropy for two random variables p and q with observations x is defined as as $H(p, q) = \sum_{x \in \mathcal{X}} p(x) \log q(x)$. We now show the expected log likelihood matches this form, specifically

$$\begin{aligned}\mathbb{E}[\ell(\nu, z, \beta)] =& \binom{n}{2}^{-1} \sum_{i<j} p_{ij}(\nu^0, z^0, \beta^0) \log(p_{ij}(\nu, z, \beta)) \\ &+ (1 - p_{ij}(\nu^0, z^0, \beta^0)) \log(1 - p_{ij}(\nu, z, \beta)) \\ =& \binom{n}{2}^{-1} \sum_{i<j} \sum_{a \in \{0,1\}} \Pr(G_{ij} = a|\nu^0, z^0, \beta^0) \log(\Pr(G_{ij} = a|\nu, z, \beta)),\end{aligned}$$

where the last expression matches the form of cross-entropy for each dyad. As in Shalizi and Asta (2017), we further define

$$\begin{aligned}&- \sum_{a \in \{0,1\}} \Pr(G_{ij} = a|\nu^0, z^0, \beta^0) \log(\Pr(G_{ij} = a|\nu, z, \beta)) \\ &= H(\Pr(G_{ij}|\nu^0, z^0, \beta^0)) + D(\Pr(G_{ij}|\nu^0, z^0, \beta^0) || \Pr(G_{ij}|\nu, z, \beta))\end{aligned}$$

where $D(\cdot)$ denotes the KL divergence and $H(\cdot)$ denotes the entropy. The left hand side is minimized when $\Pr(G_{ij}|\nu^0, z^0, \beta^0) = \Pr(G_{ij}|\nu, z, \beta)$. These results hold only up to an equivalence class defined by distance (see condition 1 in Definition 1 in Shalizi and Asta (2017)). Previous work using latent distance models (e.g. Hoff et al. (2002)) discuss identification to the equivalence class. Leveraging conditional independence given latent positions and noting entropy and KL divergence are both additive over independent random variables gives the result.

We now move to the second part of the Lemma, uniform convergence. The uniform convergence argument will proceed in two steps. Pointwise convergence by establishing a concentration inequality and then a move to uniformity by passing to the supremum over all parameters to show there is a concentration inequality that applies jointly. We follow the

arguments of Lemmas 14 and 15 of Shalizi and Asta (2017) to establish pointwise convergence and Theorem 16 for the extension to uniform convergence.

As in Shalizi and Asta (2017), we begin with a concentration inequality. Recall from the likelihood above that the data enter the likelihood only through a single term and the latent random variables we are conditioning on consist of ν, z and X and is a non-random triangular array. Further, define $\lambda_n(\nu_i, \nu_j, z_i, z_j, \beta) := \log\left[\frac{p_{ij}}{1-p_{ij}}\right]$. Now, from the form of the likelihood we see that the maximum change in the likelihood that results from changing one G_{ij} and leaving the rest the same is, as in Shalizi and Asta (2017), bounded by $\frac{2}{n(n-1)}\lambda_n(\nu_i, \nu_j, z_i, z_j, \beta)$. This bound arises from the form of the likelihood and is not altered by the additional parameters for individual effects and coefficients.

We appeal to the bounded difference theorem (McDairmid's inequality) for the sum. This gives us, denoting $p_{ij}(\nu, z, \beta)$ as p_{ij} for simplicity,

$$\begin{aligned}\Pr\left(|\ell(\nu, z, \beta) - \mathbb{E}[\ell(\nu, z, \beta)]| > \epsilon\right) &\leq \text{const.} \times \exp\left(-\frac{2\epsilon^2}{\sum_{p<q} c_{pq}^2}\right) \\ &= \text{const.} \times \exp\left(-\frac{2\epsilon^2 \binom{n}{2}^2}{\sum_{p<q} \lambda_n^2}\right)\end{aligned}$$

Note by assumption that since p_{ij} has a lower and upper bound, this is actually converging to zero because the numerator is a factor n^2 than the denominator since λ_n is bounded above and below. The next result immediately follows by logit boundedness in any case, which is implied by the assumptions (after all the boundedness of the link function implies that $v_n = o(n)$ since it is actually order constant).

We can proceed simply using a constant then as an upper bound

$$\Pr\left(|\ell(\nu, z, \beta) - \mathbb{E}[\ell(\nu, z, \beta)]| > \epsilon\right) \leq \text{const.} \times \exp\left(-\text{Const.}\binom{n}{2}\epsilon^2\right).$$

The above logic follow directly from Lemmas 14 and 15 from Shalizi and Asta (2017). Equipped with this concentration inequality we want to show that this is uniform over the parameter space.

To pass to uniformity, we use an argument based on complexity. For a normed space (with norm $\|\cdot\|$ and a subset of the space Θ, an ϵ-covering is finding a union of balls of radius ϵ which covers the subset: $\Theta \subset \cup_{i=1}^{n} B_i(\epsilon)$. The covering number is the minimal number n^* of such balls needed to cover Θ. Consistent with Shalizi and Asta (2017), we denote this as $\mathcal{N}(\Theta, \|\cdot\|, \epsilon)$. Further, let $\mathcal{L}_n$ denote the class of log-likelihood functions, so we are interested in covering $\mathcal{L}_n$: $\mathcal{N}(\mathcal{L}_n, L_1, \epsilon)$. The goal is to argue that the covering number is slowly growing relative to the concentration inequality term, thereby allowing for a uniformity result.

The space $\mathcal{L}_n$ is C^∞ and has dimension $n \dim(\mathcal{M}^p)$ for latent effects (present in Shalizi and Asta (2017)) and $n \dim \mathbb{R}$ for individual ν_i effects (not present in Shalizi and Asta (2017)). For simplicity, we exclude the regression parameters, β, here, though our argument directly extends to the case where they are present. Shalizi and Asta (2017) establish a bound on the pseudo-dimension of $\mathcal{L}_n$ using the number of connected components (Proposition 11 and Theorem 12). This argument applies directly here, yielding a bound of

$$2 \log_2 B_{\mathcal{M}^p} + 2(n(p+1)) \log_2 e.$$

where $B_{\mathcal{M}^p}$ is the number of connected components of the isometry of $\mathcal{M}^p$. The result follows directly the inequality argument of Shalizi and Asta (2017).

Shalizi and Asta (2017) consider Euclidean space and a 2-hyperbolic space. We use a sphere to represent the latent space and, thus, extend the proof for spheres. First, observe that $\text{isom}(\mathbf{S}^p) = O(p+1)$, where $O(p+1)$ is the orthogonal group in dimension $p+1$. The orthogonal group, $O(p+1)$, is a subgroup of the Euclidean group $\mathbb{E}(p+1)$ (see Theorem 1.12 in Parkkonen (2012)). Further, $O(p+1)$ has two connected components with positive and

negative determinants. Let $O(p+1,1)$ be the orthogonal group of the Minkowski bilinear form (see Parkkonen (2012)). Denote $O^+(p+1,1)$ as the index 2 subgroup of $O(p+1,1)$ consisting of those transformations which preserve the components. By Theorem 1.3.1 in Paupert (2016) the Hyperbolic space, $\mathbf{H}^p$, has the following relationship with $O(p+1,1)$: $\text{isom}(\mathbf{H}^p) = O^+(1,p)$ (see also Theorem 1.12 in Parkkonen (2012)) The result follows because the latter has a finite number of connected components.

The above completes the proof of the Lemma.

In Lemma 4.3.1, we extend the results by Shalizi and Asta (2017) to include not only latent parameters but also individual specific effects. This result shows consistency of all formation model parameters for cases when the entire graph is observed. Now, we show that there is sufficient information in ARD to consistently estimate these parameters.

THEOREM **4.3.1.** *Consider a sequence of graphs with* $\mathbf{g}_n$ *drawn according to distributions in* (4.1) *with* $(z_i^0, \nu_i^0)_{i=1}^n$ *being drawn independent and identically distributed according to an absolutely continuous distribution on* $\mathcal{S}^p \times \mathbb{R}$. *The researcher only observes ARD* $\mathbf{Y}_{n\times K}$. *Then* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^0$ *as* $n \to \infty$.

Proof of Theorem 4.3.1

The proof again adapts Shalizi and Asta (2017).The distinction is that the ARD case uses a Poisson likelihood since the observed data are counts rather than binary. The proof of Theorem 4.3.1 relies on the likelihood

$$l(z_{1:n}, \mathbf{Y}) = \frac{1}{n^2 K} \sum_{p=1}^{n} \sum_{k=1}^{K} Poisson(\lambda_{pk}) = \frac{1}{n^2 K} \sum_{p=1}^{n} \sum_{k=1}^{K} -\lambda_{pk} + y_{pk} \log \lambda_{pk} - \log(y_k!)$$

where $\lambda_{pk} = N_k \cdot \mathbb{E}[\Lambda(d(z_p, z_q)|q \in G_k)]$ where $\Lambda(\cdot)$ is the link function. We extend their arguments about identification and uniform convergence of the likelihood. First we show that

the argmax of the likelihood is the true vector of locations (Lemma 13 of Shalizi and Asta (2017)):

$$[z^0_{1:n}] = \underset{z_{1:n} \in M^n}{\operatorname{argmax}} \bar{l}(z_{1:n})$$

where $\bar{l}(z_{1:n}) = \mathbb{E}[l(z_{1:n})]$. Following the same argument, let $\pi_{pk}(a) = \Pr(y_{pk} = a | z_p, z_{q \in G_k})$, $\pi^*_{pk}(a) = \Pr(y_{pk} = a|\mu)$ so $\bar{l}(z_{1:n}, G) = \frac{1}{n^2 K} \sum_{p=1}^{n} \sum_{k=1}^{K} \sum_{a=0}^{N_k} \pi^*_{pk}(a) \log \pi_{pk}(a)$. Since we can write the summand in terms of the entropy and KL-divergence, $-\sum_{a=0}^{N_k} \pi^*_{pk}(a) \log \pi_{pk}(a) = H[\pi^*_{pq}] + D[\pi^*_{pq} || \pi_{pq}]$, it follows that $-\bar{l}(z_{1:n}) = H[\pi^*] + D(\pi^* || \pi)$, and so $[z^0_{1:n}] = \underset{z_{1:n} \in M^n}{\operatorname{argmax}} \bar{l}(z_{1:n})$ by taking $\pi = \pi^*$.

Second, we show uniform convergence of the sample likelihood to its expectation (Theorem 16 of Shalizi and Asta (2017))

$$\sup_{z_{1:n}} |l(z_{1:n}) - \bar{l}(z_{1:n})| \xrightarrow{p} 0$$

by developing a concentration inequality (adapting Lemmas 14 and 15 to the ARD setting). We first show a concentration inequality for the likelihood for a given parameter vector and then we pass to uniformity using a complexity argument. The difference in our argument relative to that in Shalizi and Asta (2017) is that we have a Poisson random variable rather than a sub-Gaussian random variable. The likelihood for ARD, therefore, does not satisfy bounded differences well-enough to apply McDairmid's inequality. Instead, we appeal to inequalities for sub-exponential random variables.

The difference of interest is

$$l(z_{1:n}) - \bar{l}(z_{1:n}) = \frac{1}{n^2 K} \sum_{p=1}^{n} \sum_{k=1}^{K} (y_{pk} - \lambda_{pk}) \log \lambda_{pk} - (\log(y_k!) - \mathbb{E}[\log(y_k!)]).$$

We can use Stirling's approximation $\log(y_{pk}!) \approx y_{pk} \log(y_{pk}) - y_{pk}$ and so in addition to a

summand that is linear in y_{pk} we will have a term $y_{pk}\log(y_{pk})$ and its expectation. It is easy to check that $y_{pk}\log(y_{pk}) - \mathbb{E}[y_{pk}\log(y_{pk})]$ is sub-exponential (by condition 2 of Theorem 2.13 in Wainwright (2019)). Using the sub-exponential inequality we can calculate

$$\begin{aligned}\Pr\left(\left|\frac{1}{n^2K}\sum_{p=1}^{n}\sum_{k=1}^{K}(y_{pk}-\lambda_{pk})\log\lambda_{pk}\right|\geq\epsilon\right) &= \Pr\left(\left|\sum_{p=1}^{n}\sum_{k=1}^{K}(y_{pk})-nK\lambda\right|\geq\frac{\epsilon n^2K}{\log\lambda}\right)\\ &\leq 2\exp\left(-\frac{\epsilon^2n^3K}{2\log\lambda(\lambda\log\lambda+n\epsilon)}\right).\end{aligned}$$

Similarly, letting $S_n := y_{pk}\log(y_{pk})$ and $\mathbb{E}[S_n] = nK\,\mathbb{E}[y\log y]$ we have

$$\begin{aligned}\Pr\left(\left|\frac{1}{n^2K}\sum_{p=1}^{n}\sum_{k=1}^{K}(S_n-\mathbb{E}[S_n])\right|\geq\epsilon\right) &= \Pr\left(e^{\theta(S_n-\mathbb{E}[S_n])}\geq e^{\theta\epsilon n^2K}\right)\\ &\leq e^{-\theta\epsilon n^2K}\,\mathbb{E}[e^{\theta(S_n-\mathbb{E}[S_n])}]\\ &= e^{-\theta\epsilon n^2K}\prod_{i=1}^{nK}\mathbb{E}[e^{\theta(Z_i-E[Z_i])}]\\ &\leq e^{-\theta\epsilon n^2K}\prod_{i=1}^{nK}e^{\frac{\nu^2\theta^2}{2}} = \exp\left(-\theta\epsilon n^2K+nK\frac{\nu^2\theta^2}{2}\right).\end{aligned}$$

We notice that $-\theta\epsilon n^2K + nK\frac{\nu^2\theta^2}{2}$ is a quadratic function of θ, and is minimized at $\theta = \frac{\epsilon n}{\nu^2}$. Therefore a tighter bound is

$$\Pr\left(\frac{1}{n^2K}\left|S_n-\mathbb{E}[S_n]\right|\geq\epsilon\right)\leq\exp\left(-\frac{\epsilon^2n^3K}{2\nu^2}\right).$$

Finally, to pass to uniformity, the result follows from the proof of Theorem 16 in Shalizi and Asta (2017) which itself parallels the argument in Theorem 17.1 in Anthony and Bartlett (1999), where instead of the concentration inequality developed using bounded differences, we

use the sub-exponential based inequality developed above. It immediately follows that

$$\Pr\left(\sup_{z_{1:n}} |l(z_{1:n}) - \bar{l}(z_{1:n})| \geq \epsilon\right) \leq 4\mathcal{N}_1(\mathcal{L}_n, \epsilon/16) \exp\left(-\frac{\epsilon^2 n^3 K}{16(2\log\lambda(\lambda\log\lambda + n\epsilon))}\right)$$

where $\mathcal{N}_1(\mathcal{L}_n, \epsilon/16)$ is the covering number of the space of likelihoods $\mathcal{L}_n$ with balls of size $\epsilon/16$. By the same argument the covering number is $O(n\log 1/\epsilon)$ whereas the exponential term exponentially declines at rate n^2 so overall the probability tends to 0 as $n \to \infty$.

4.4 When ARD Works

For theoretical results, we assume that data arise from a formation model of the form presented in (4.1) and that the ARD procedure tightly identifies the model parameters. These assumptions allow us to focus on when the expectation of the network statistic is sufficiently informative about any given graph realization. Under these assumptions, let $p_{ij}^{\boldsymbol{\theta}}$ denote the probability that nodes i and j are linked under the data generating process with parameter vector $\boldsymbol{\theta}$.

We separate our discussion into two cases: (1) the researcher has a single large network with n nodes (or a handful of networks); (2) the researcher has many independent networks.

4.4.1 Single Large Network

We first consider the case where there is a single large network, and the researcher is interested in measuring a specific network statistic, $S_i(\mathbf{g})$ for node i computed on graph $\mathbf{g}$. For the purposes of this argument, there is one actual realization of the graph, $\mathbf{g}^*$. This realization is what we would have observed if we had collected information about all actual connections between members of the population, rather than collecting ARD. Importantly, the researcher collecting ARD cannot observe $\mathbf{g}^*$. This actual network realization does, however, come from a generative model that has parameters that can be estimated from the ARD. The researcher

can, therefore, simulate graph realizations from the underlying data generating process under the true parameter vector $\boldsymbol{\theta}$, and construct an estimate for $\mathbb{E}\left[S_i\left(\mathbf{g}\right)|\boldsymbol{\theta}\right]$. This expectation is over the possible graphs generated from the model with parameters $\boldsymbol{\theta}$. Recall, in practice, we will observe a $n \times K$ matrix of ARD, $\mathbf{Y}_n$, rather than $\boldsymbol{\theta}$ (for simplicity here we set $m = n$). This expectation, then, is $\mathbb{E}\left[S_i\left(\mathbf{g}\right)|\mathbf{Y}_n\right]$. As we describe in Section 4.3, the ARD data, $\mathbf{Y}_n$, are sufficient to identify the generative parameters, θ. To simplify notation, we will omit the conditioning for the remainder of this section.

To recap, if a researcher collected information about all links in the population, she could compute $S_i(\mathbf{g}^*)$ directly. With ARD, however, she can recover an expectation over graphs generated with a given set of parameters, $\mathbb{E}\left[S_i\left(\mathbf{g}\right)\right]$. We are interested in cases in which knowing $\mathbb{E}\left[S_i\left(\mathbf{g}\right)\right]$ is sufficient for learning about $S_i(\mathbf{g}^*)$. That is, cases where, if we can get a good estimate for $\mathbb{E}\left[S_i\left(\mathbf{g}\right)\right]$ using ARD, we can say with confidence that we have recovered a statistic that is very similar to the statistic the researcher would have observed had she collected data on the entire graph. More formally, for any realized graph, $\mathbf{g}$, does $S_i\left(\mathbf{g}\right) \xrightarrow{p} \mathbb{E}\left[S_i\left(\mathbf{g}\right)\right]$?

If this condition holds, then when the population of individuals, n, is large, the statistic of interest, $S_i\left(\mathbf{g}\right)$, will be close to its expectation for any realization of the graph, including the one that is the researcher's population of interest, $\mathbf{g}^*$. And the MSE between $S_i\left(\mathbf{g}^*\right)$ and $\mathbb{E}\left[S_i\left(\mathbf{g}\right)\right]$. The key feature of the result is that we do not need to know the exact structure of the graph that the researcher would have observed using a network census, $\mathbf{g}^*$. Instead, we rely on the notion that the statistic will be close to its expectation for a sufficiently large graph and that this is true for any realization of the graph from a given generative process.

We formalize this intuition using Corollary 4.4.1. Though the corollary is uncomplicated to prove, it cements the condition required of the statistic of interest for us to reasonably expect that our ARD estimates will be similar to what a researcher would have observed by directly

computing the statistic from the fully-elicited graph. Further, it serves to demystify how ARD can work to recover network statistics with such limited information on the graph. The information in ARD, by the arguments in Section 4.3, is sufficient to estimate the parameters of the formation model. After we state and proce the corollary, we provide examples of statistics where ARD should and should not perform well. We prove that $S_i(\mathbf{g}) \xrightarrow{p} \mathbb{E}[S_i(\mathbf{g})]$ for density, diffusion centrality, and node-level clustering. We confirm our intuition through simulations in Section 4.5.

COROLLARY **4.4.1.** *Assume $\boldsymbol{\theta}$ is known. Let $S_i(\mathbf{g}^*)$ be the (unobserved) statistic of the underlying network and $S_i(\mathbf{g})$ be the same statistic computed from graph $\mathbf{g}$, drawn from the distribution with parameters $\boldsymbol{\theta}$. If $S_i(\mathbf{g}) \xrightarrow{p} \mathbb{E}[S_i(\mathbf{g})]$, Then the MSE is*

$$\mathbb{E}\left[\left(\mathbb{E}[S_i(\mathbf{g})] - S_i(\mathbf{g}^*)\right)^2\right] = o_p(1).$$

Proof of Corollary 4.4.1 If $S_i(\mathbf{g}) \xrightarrow{p} \mathbb{E}[S_i(\mathbf{g})]$ for any realization $\mathbf{g}$, then $S_i(\mathbf{g}^*) \xrightarrow{p} \mathbb{E}[S_i(\mathbf{g})]$ for the true unobserved realization $\mathbf{g}^*$. Assume $|S_i(\mathbf{g})| < M \in \mathbb{R}$, then convergence in probability implies convergence in L^p. Taking $p = 2$, this completes the proof for MSE.

To clarify when this applies and when this fails, we provide several pedagogical examples. Our first example is existence of a link. This is when Corollary 4.4.1 does not apply since $S_i(\mathbf{g}) \overset{p}{\not\to} \mathbb{E}[S_i(\mathbf{g})]$.

COROLLARY **4.4.2.** *Given an (unobserved) graph of interest, $\mathbf{g}^*$, and non-degenerate linking probabilities $0 < p_{ij}^{\boldsymbol{\theta}} < 1$, then the MSE for $\mathbb{E}[S_i(\mathbf{g})] = \mathbb{E}[g_{ij}]$, expected connectivity of any single link g_{ij} is given by*

$$\mathbb{E}\left[\left(g_{ij} - g_{ij}^*\right)^2\right] = p_{ij}^{\boldsymbol{\theta}}\left(1 - p_{ij}^{\boldsymbol{\theta}}\right)$$

Note that irrespective of n, this cannot tend to zero. When a link exists, the mean-squared error is $(1-p_{ij}^{\boldsymbol{\theta}})^2$ and when it does not, the MSE is $p_{ij}^{\boldsymbol{\theta}^2}$.

PROPOSITION **4.4.1.** *Let $S_i(\mathbf{g})$ be a statistic of graph $\mathbf{g}$ drawn from teh distribution with parameter vector $\boldsymbol{\theta}$. $S_i(\mathbf{g}) \xrightarrow{p} \mathbb{E}[S_i(\mathbf{g})]$ for the following statistics:*

1. *Density (normalized degree): $d_i(\mathbf{g})/n := \sum_j g_{ij}/n$.*
2. *Diffusion centrality (nests eigenvector centrality and Katz-Bonacich centrality): For parameter sequence $q_n = \frac{C}{n}$ and any T, $DC_i(\mathbf{g}; q_n, T) := \sum_j \left[\sum_{t=1}^{T} (q_n \mathbf{g})^t\right]_{ij}$.*
3. *Clustering: $\text{clustering}_i(\mathbf{g}) := \frac{\sum_{j,k\in N(i)} g_{jk}}{|N(i)|\cdot(|N(i)|-1)}$ where $N(i) := \{j : g_{ij} = 1\}$.*

Proof of Proposition 4.4.1 For part 1, density, we have

$$\sum_{j\in\{1,\ldots,n\}, j\neq i} \frac{\text{var}(g_{ij})}{(n-1)^2} = \sum_{j\in\{1,\ldots,n\}, j\neq i} \frac{p_{ij}^{\theta}\left(1-p_{ij}^{\theta}\right)}{(n-1)^2} \leq \sum_{j\in\{1,\ldots,n\}, j\neq i} \frac{1}{(n-1)^2} = \frac{1}{n-1} \to 0$$

so the Kolmogorov condition is satisfied and

$$\frac{d_i}{n} - \frac{\mathbb{E}[d_i]}{n} \to_{a.s.} 0$$

which satisfies the conditions of Proposition 4.4.1.

In part 2 we turn to diffusion centrality. Recall that.

$$DC_i(\mathbf{g}; q_n, T) := \sum_j \left[\sum_{t=1}^{T} (q_n \mathbf{g})^t\right]_{ij} = \sum_j \sum_{t=1}^{T} \frac{C^t}{n^t} \sum_{j_1,\ldots,j_{t-1}} g_{ij_1} \cdots g_{j_{t-1}j}.$$

For any t, we have

$$\text{var}\left(\frac{1}{n^t}\sum_j \sum_{j_1,\ldots,j_{t-1}} g_{ij_1}\cdots g_{j_{t-1}j}\right) = \frac{1}{n^{2t}}\sum_j \sum_{j_1,\ldots,j_{t-1}} \text{var}(g_{ij_1}\cdots g_{j_{t-1}j}) + \frac{1}{n^{2t}}\sum_j \sum_{j_1,\ldots,j_{t-1}} \sum_k \sum_{k_1,\ldots,k_{t-1}} \text{cov}(g_{ij_1}\cdots g_{j_{t-1}j}, g_{ik_1}\cdots g_{k_{t-1}k})$$

where $j_0 = k_0 = i$ and $j_s = j, k_s = k$. $\text{var}(g_{ij_1}\cdots g_{j_{t-1}j})$ has variance

$$\prod_{s=1}^{t} p_{j_{s-1}j_s}\left(1 - \prod_{s=1}^{t} p_{j_{s-1}j_s}\right) \leq 1$$

and $\text{cov}(g_{ij_1}\cdots g_{j_{t-1}j}, g_{ik_1}\cdots g_{k_{t-1}k}) \leq 1$. In order for $\text{cov}(g_{ij_1}\cdots g_{j_{t-1}j}, g_{ik_1}\cdots g_{k_{t-1}k}) \neq 0$, $g_{ij_1}\cdots g_{j_{t-1}j}$ and $g_{ik_1}\cdots g_{k_{t-1}k}$ need to have at least one edge in common. Notice that $g_{ij_1}\cdots g_{j_{t-1}j}$ has n^t combinations since i is given. Therefore, given a fixed common edge that $g_{ij_1}\cdots g_{j_{t-1}j}$ and $g_{ik_1}\cdots g_{k_{t-1}k}$ share, $g_{ij_1}\cdots g_{j_{t-1}j}$ has n^{t-2} free choices of actors in the path, and $g_{ik_1}\cdots g_{k_{t-1}k}$ also has n^{t-2} free choices of actors in the path. Therefore, for a given fixed common edge, there are $n^{2(t-2)}$ non-zero covariance terms. Since there are n^2 choices of a common edge, there are a total of n^{2t-2} non-zero covariance terms. Therefore,

$$\text{var}(\frac{1}{n^t}\sum_j \sum_{j_1,\ldots,j_{t-1}} g_{ij_1}\cdots g_{j_{t-1}j}) \leq \frac{n^t + n^{2t-2}}{n^{2t}}.$$

Let $DC_{i,t} = \frac{1}{n^t}\sum_j \sum_{j_1,\ldots,j_{t-1}} g_{ij_1}\cdots g_{j_{t-1}j}$, we have

$$\Pr(|DC_{i,t} - E[DC_{i,t}]| \geq \epsilon) \leq \frac{n^t + n^{2t-2}}{n^{2t}\epsilon^2} \text{ by Chebyshev's inequality}$$

$$\Pr(|DC_{i,t} - E[DC_{i,t}]| < \epsilon) \geq 1 - \frac{n^t + n^{2t-2}}{n^{2t}\epsilon^2} \to 1 \text{ as } n \to \infty$$

Therefore,

$$DC_{i,t} \xrightarrow{P} E[DC_{i,t}] \text{ as } n \to \infty$$

By continuous mapping theorem,

$$DC_i(\mathbf{g}; q_n, T) = \sum_{t=1}^{T} C^t \cdot DC_{i,t} \xrightarrow{P} E[DC_i(\mathbf{g}; q_n, T)].$$

For part 3, clustering, the argument is identical to the convergence of clustering in Erdos-Renyi graphs because every link is conditionally edge independent. Let $N(i)$ denote the set of neighbors of actor i and $|N(i)|$ denote the size of neighbors, then

$$\text{clustering}_i(\mathbf{g}) = \frac{\sum_{j,k \in N(i)} g_{jk}}{|N(i)| \cdot (|N(i)| - 1)}$$

Similar to the proof for density, we have

$$\sum_{j,k \in N(i)} \frac{\text{var}(g_{jk})}{(|N(i)| \cdot (|N(i)| - 1))^2} = \sum_{j,k \in N(i)} \frac{p_{jk}^{\theta}\left(1 - p_{jk}^{\theta}\right)}{(|N(i)| \cdot (|N(i)| - 1))^2}$$
$$\leq \sum_{j,k \in N(i)} \frac{1}{(|N(i)| \cdot (|N(i)| - 1))^2} = \frac{1}{|N(i)| \cdot (|N(i)| - 1)} \to 0$$

so the Kolmogorov condition is satisfied and

$$\text{clustering}_i(\mathbf{g}) \longrightarrow_p \mathbb{E}_{z_j, \nu_j, z_k, \nu_k | j,k \in N(i)} \left[\Pr(g_{jk} = 1 | \nu_j, \nu_k, z_j, z_k)\right].$$

A few remarks are worth mentioning. First, diffusion centrality is a more general form which nests eigenvector centrality when $q_n \geq \frac{1}{\lambda_1^n}$, and because the maximal eigenvalue is on the order of n, this meets our condition. It also nests Katz-Bonacich centrality. In each of these, $T \to \infty$. It also captures a number of other features of finite-sample diffusion processes

that have been used in applied work. Each of these notions relate to the eigenvectors of the network—objects that are ex-ante not obviously captured by the ARD procedure but ex-post work because the models are such that in large samples the statistics converge to their limits.

These results give us two practical extreme benchmarks. Our procedure should not perform well at all for estimating a realization of any given link in the network. In contrast, it should perform quite well for statistics such as degree or eigenvector centrality. Other statistics may fall somewhere in the middle of this spectrum. For example, a notion of centrality such as betweenness, which relies on the specifics of the exact realized paths in the network, is unlikely to work particularly well because even for large n, the placement of specific nodes may radically change its value. Section 4.5.1 explores these predictions empirically using simulations.

4.4.2 Many Independent Networks

Now consider the setting where the researcher has R independent networks each of size n_r. We'll take $n_r = n$ for simplicity, though the results presented here do not require this. We also have an ARD sample $\mathbf{Y}_{n,r}$ for every network $r = 1, ..., R$. Every network is generated from a network formation process with true parameter $\boldsymbol{\theta}_r$. In this case of many networks, we consider how well the ARD procedure performs when the researcher wants to learn about network properties, aggregating across the R graphs. This is the case in a large literature (Cai et al., 2013; Beaman et al., 2016).

Let $S_r^* := S(\mathbf{g}_r^*)$ be a network statistic from the R unobserved graphs generating the ARD. For any given graph from the data generating process, define $S_r := S(\mathbf{g}_r)$. For notational simplicity, we consider network-level statistics, but the argument can easily be extended to node, pair, or subset-based statistics.

Assume the goal of the researcher is to estimate some model

$$y_r = \alpha + \beta S_r^* + \epsilon_r$$

where y_r is some socio-economic outcome of interest and and the parameter of interest is β. As before, S_r^* is unobserved because $\mathbf{g}_r^*$ is unobserved and the researcher must make do with ARD, $\mathbf{Y}_r$. The researcher instead estimates the expectation of the statistic given using ARD, $\bar{S}_r := \mathbb{E}(S_r)$. The regression then becomes:

$$y_r = \alpha + \beta \bar{S}_r + u_r.$$

Under standard regularity conditions, we can consistently estimate β. The intuition is that the deviation of the conditional expectation $\bar{S}_r$ from S_r is by definition orthogonal to the conditional expectation and independent across r. So one can think of the conditional expectation as an instrumental variable for the true S_r where the first-stage regression has a coefficient of 1.

Similarly, we can consider the network feature to be the outcome of interest and study how it responds to an intervention given by T_r:

$$S_r^* = \alpha + \gamma T_r + \epsilon_r.$$

Instead estimating

$$\bar{S}_r = \alpha + \gamma T_r + \epsilon_r.$$

yields consistent estimates for γ.

PROPOSITION **4.4.2.** *As $R \to \infty$, (1) assume the design matrix has full rank, $\mathbb{E}[YS] < \infty, \mathbb{E}[Y] < \infty, \mathbb{E}[S] < \infty$, $\hat{\beta} \xrightarrow{p} \beta$*

and (2) assume the design matrix has full rank, $\mathbb{E}[S] < \infty$*,* $\hat{\gamma} \xrightarrow{p} \gamma$*.*

Proof of Proposition 4.4.2 For (1), we show that β is still consistently estimated when using $\bar{S}_r$ as a regressor rather than S_r. First, expand the error term,

$$y_r = \alpha + \beta S_r^* + \epsilon_r = \alpha + \beta \bar{S}_r + \left\{ \epsilon_r + \beta \left(S_r^* - \bar{S}_r \right) \right\}.$$

By iterated expectations we can see that

$$\mathbb{E}\left[\bar{S}_r \left(S_r^* - \bar{S}_r \right) \right] = \mathbb{E}\left[\mathbb{E}\left[\bar{S}_r \left(S_r^* - \bar{S}_r \right) | \theta_r \right] \right] = \mathbb{E}\left[\bar{S}_r \left(\mathbb{E}\left[S_r^* | \theta_r \right] - \bar{S}_r \right) \right] = \mathbb{E}\left[\bar{S}_r \left(\bar{S}_r - \bar{S}_r \right) \right] = 0.$$

The result immediately follows.

$$\hat{\beta} \xrightarrow{P} \frac{Cov(y, \bar{S})}{Var(\bar{S})} = \beta + \frac{Cov(\bar{S}, \beta \left(S^* - \bar{S} \right))}{Var(\bar{S})} = \beta$$

For (2), we see that

$$S_r^* = \alpha + \gamma T_r + \epsilon_r$$

which transforms to

$$\bar{S}_r = \alpha + \gamma T_r + \epsilon_r + \bar{S}_r - S_r^*$$

$$\widehat{\gamma} \rightarrow_p \frac{\operatorname{cov}\left(\bar{S}, T \right)}{\operatorname{var}\left(T \right)} = \frac{\operatorname{cov}\left(S^*, T \right)}{\operatorname{var}\left(T \right)} + \frac{\operatorname{cov}\left(\bar{S} - S^*, T \right)}{\operatorname{var}\left(T \right)} = \gamma$$

where we use that the estimation error is independent of the treatment assignment since $\mathbb{E}[(\bar{S} - S^*)T] = \mathbb{E}[\mathbb{E}[(\bar{S} - S^*)T|T]] = \mathbb{E}[T \cdot \mathbb{E}[(\bar{S} - S^*)|T]] = \mathbb{E}[T \cdot 0] = 0$.

To illustrate this practically, let us take the most extreme example of a single link, where we know its presence cannot be identified in a single large network. This means that even if

we were interested in a regression of

$$y_{12,r} = \alpha + \beta g_{12,r} + \epsilon_r,$$

where whether nodes 1 and 2 are linked affects some outcome variable of interest, and we are interested in this across all R networks, we can use $p_{12}^{\theta,r} := \mathbb{E}\left[g_{12,r}|\mathbf{Y}_r\right]$ instead in the regression to consistently estimate β. Note that in contrast to the single network case, where we were interested in recovering g_{12} itself, and even with large n the MSE would not tend to zero, here simply having the conditional expectation is enough to be able to estimate the economic slope of interest, β. Therefore, with many graphs, the ARD procedure should work well regardless of the properties of the given network statistic.

4.5 Simulation Results

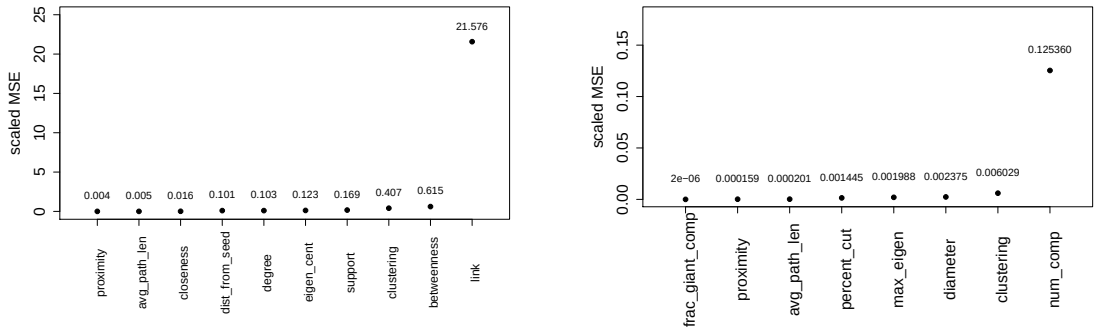
4.5.1 Single Large Graph

We explore the results for a single large graph through simulations. For this simulation, we use graphs with 250 nodes, which is a similar size to the Karnataka data we describe in Section 3.5.1, simulated from the data generating process in Equation 4.1. In Figure 4.1, we plot the mean squared errors of our estimation procedure across a range of network statistics which are commonly used in applied economics. In order to make the MSEs comparable across statistics, we scale by $\frac{1}{\mathbb{E}[S_i(\mathbf{g})]^2}$. Figure 4.1a focuses on node level statistics while Figure 4.1b focuses on graph-level statistics.

The node level statistics are as follows: (1) degree (the number of links); (2) eigenvector centrality (the ith entry of the eigenvector corresponding to the maximal eigenvalue of the adjacency matrix for node i); (3) betweenness centrality (the share of shortest paths between all pairs j and k that pass through i); (4) closeness centrality (the average inverse distance

from i over all other nodes); (5) clustering (the share of a node's links that are themselves linked); (6) support (as defined in Jackson et al. (2012) – whether linked nodes ij have some k as a link in common); (7) whether link ij exists; (8) closeness; (9) average path length; and (10) the average distance from a randomly chosen "seed" (as in an information diffusion experiment).

The graph level statistics are as follows: (1) diameter; (2) average path length; (3) average proximity (average of inverse of shortest paths); (4) share of nodes in the giant component; (5) number of components; (6) maximal eigenvalue; (7) clustering; and (8) the share of links across the two groups relative to within the two groups where the cut is taken from the sign of the Fiedler eigenvector (this reflects latent homophily in the graph and is denoted as percent cut in graph labels).



(a) Node-level **(b)** Graph-level

Figure 4.1: Scaled MSE of node-level and graph-level network features. Each point in the figure represents the MSE across 250 simulations using graphs of size 250, a size comparable to the data we examine in Section 3.5.1. These results corroborate the theoretical intuition developed in Section 4.4.1.

Figure 4.1a shows that the scaled MSEs in our simulations are quite small for most network statistics, including degree and (eigenvector) centrality, as predicted. The scaled MSE for the estimates of the existence of a link is extremely large. Moreover, betweenness

also performs worse than the other statistics.

Figure 4.1b considers graph level statistics. The scaled MSEs tend to be small for all but one network statistic – the number of components in the graph. The number of components depends crucially on the existence of a small number of specific link realizations, calling upon the same intuition as the node-level existence of a link.

4.5.2 Many Independent Networks

We now explore results on regression coefficients and treatment effects when researchers collect ARD in multiple networks, through simulations. For this simulation, we use graphs with 250 nodes, and we assign graph level treatment randomly to half of the graphs. Graphs in the control group have expected degree generated from $\mathcal{N}(\mu = 15, \sigma = 5)$, while graphs in the treatment group have expected degree generated from $\mathcal{N}(\mu = 25, \sigma = 5)$. All graphs have a minimum expected degree of 5 and a maximum expected degree of 35. Due to the association between density and treatment, we expect treatment effects on graph-level statistics, such as average path length and diameter. The average sparsity over all graphs is 20/250=0.08, which is a value similar to Karnataka data discussed in Section 3.5.1. We set the number of network used in one regression to be $R = 50, 100$ and 200. For individual measures, 50 actors are randomly selected in each network. For link measure between a pair of actors, 1000 pairs are randomly selected in each network. For network level measures, there is one measure per network, so the regression consists of R data samples.

Figure 4.2 shows the distribution of $\hat{\beta}$ for β in regression $y_{ij,r} = \alpha + \beta \bar{S}_{ij,r} + \epsilon_r$, where $S_{ij,r}$ and $\bar{S}_{ij,r}$ represent a true and mean individual-level measure, respectively. The top panel represent results when we use true model formation parameter $\boldsymbol{\theta}$ to get $\bar{S}_{ij,r}$, while the bottom panel represent results when we use estimated model formation parameter $\hat{\boldsymbol{\theta}}$ to get $\bar{S}_{ij,r}$. The middle line of each boxplot is the medium $\hat{\beta}$, and the borders of boxes denote first and third

quartiles. All boxplots have outliers removed. Each boxplot represents the distribution of $\hat{\beta}$ for one individual level measure when $R = 50, 100$ or 200 networks are used in regression. The x-axis labels suggest the type of the individual measure and number of networks used. The red line denotes the true $\beta = 1$ used to generate $y_{ij,r} = \alpha + \beta S^*_{ij,r} + \epsilon_r$ in the simulation. We generate ϵ_r from a normal distribution with zero mean, and $Var(\epsilon_r) = Var(S^*_{ij,r})$ so that we maintain a 0.5 noise to signal ratio.

When using true model parameters to get $\bar{S}_{ij,r}$, $\hat{\beta}$ for all individual measures are unbiased, with whiskers of boxes crossing over the true value. When using estimated model parameters to get $\bar{S}_{ij,r}$, we see slight over estimation of $\hat{\beta}$ for betweenness and diffusion centrality, while all other measures (degree, eigenvector centrality, clustering, proximity, support, closeness, link, average path length, and distance from seed) have good $\hat{\beta}$. Increasing the number of networks in one regression decreases the variance of $\hat{\beta}$.

Figure 4.3 shows the distribution of $\hat{\beta}$ for β in regression $y_r = \alpha + \beta \bar{S}_r + \epsilon_r$, where S_r and $\bar{S}_r$ represent a true and mean network-level measure, respectively. Each box represents the distribution of $\hat{\beta}$ for one measure and use of R=50, 100 or 200 networks in regression. The red line denotes the true $\beta = 1$ used to generate $y_r = \alpha + \beta S^*_r + \epsilon_r$ in the simulation. Similarly we generate ϵ_r from a normal distribution with zero mean, and $Var(\epsilon_r) = Var(S^*_r)$ so that we maintain a 0.5 noise to signal ratio. We see that using $\bar{S}_r|\boldsymbol{\theta}$ and $\bar{S}_r|\hat{\boldsymbol{\theta}}$ yield similar results, which suggests that the estimated model parameters can recover these network measures very well. All measures have unbiased $\hat{\beta}$, and variance of $\hat{\beta}$ decreases with increasing R.

Figure 4.4 shows the distribution of percentage errors of $\hat{\gamma}$ for γ in regression $\bar{S}_r = \alpha + \gamma T_r + \epsilon_r$. The percentage error is defined as $(\hat{\gamma} - \gamma)/\gamma$. The red vertical line sits at zero, which means no error. We see that using $\bar{S}_r|\boldsymbol{\theta}$ and $\bar{S}_r|\hat{\boldsymbol{\theta}}$ yield similar results for this estimation of treatment effect. We also see that percent cut and diameter has large variation of percent errors than the other measures. This is due to the fact that the treatment effect,

density differences between treatment and control groups, has smaller effect on percent cut and diameter than on other measures. To see this, the average percent of variation explained by treatment in S_r for percent cut and diameter is around 0.3, while it is around 0.5 for other measures.

In Section 4.3 we present a theorem on consistency of $\boldsymbol{\theta}$, and in Section 4.4 we present theoretical results on consistency of regression coefficient and treatment effect. The simulation results in this Section using true model parameters to get $\bar{S}_r$ or $\bar{S}_{ij,r}$ is validating propositions and corollaries in Section 4.4. While we present our theoretical results in two parts, simulation results in this Section using estimated model parameters validate that empirically regression coefficient and treatment effect are consistent when we get $\bar{S}_r$ or $\bar{S}_{ij,r}$ conditioned on $\hat{\boldsymbol{\theta}}$.

4.6 Discussion

In this chapter, we present consistency results on recovering network measures from a single large network and recovering regression coefficients and treatment coefficients for multiple independent networks. After we state and prove a theorem on consistency of MLE for network formation model parameters, we present and prove our regression related results using expected network measures given true model parameter in place of true network measure. We have shown that even though it is not possible to recover the existence of a particular link from a single large network, we can however recover regression coefficients when link is the independent variable. We present simulation results on both scenarios for a variety of individual and network level measures commonly used in economics.

Our theoretical results and simulation procedure assume that the network formulating process is consistent with the latent surface model, and ARD is not under- or over- recalled. Incorrect recall may happen in field experiments, especially when the subpopulation is of moderate to large size. This may impact recovering individual random effects ν_i and therefore

recovering network measures and regression coefficients. In that case, researchers can develop a calibration curve for ARD, as well as collect additional information to adjust recall bias. Additionally, respondents may not give an exact count but give a range of numbers as the answer. In that case, an extension of this work on censored data can be considered.

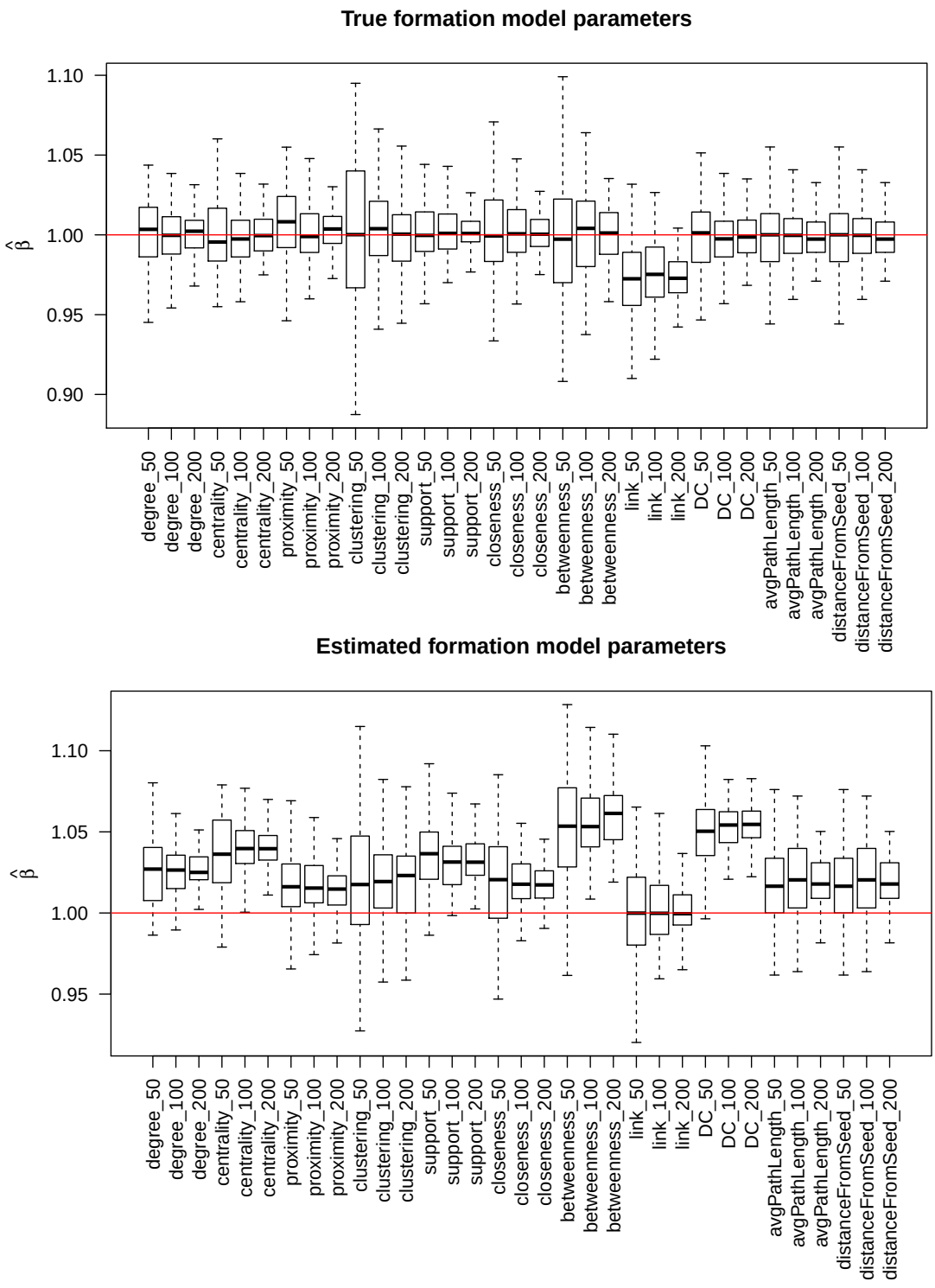


Figure 4.2: Boxplot of $\hat{\beta}$ for β in regression $y_{ij,r} = \alpha + \beta \bar{S}_{ij,r} + \epsilon_r$, where $S_{ij,r}$ and $\bar{S}_{ij,r}$ represent a true and mean individual-level measure, respectively. Each box represents the distribution of $\hat{\beta}$ for one measure and use of R=50, 100 or 200 networks in regression. 50 actors and 1000 pairs (for link) are randomly selected for each network. The middle line of the boxplot denotes medium, and borders of the boxes denote first and third quartile. The red line denotes the true $\beta = 1$ used to generate $y_{ij,r} = \alpha + \beta S^*_{ij,r} + \epsilon_r$ in the simulation. These results corroborate the theoretical intuition developed in Section 4.4.2.

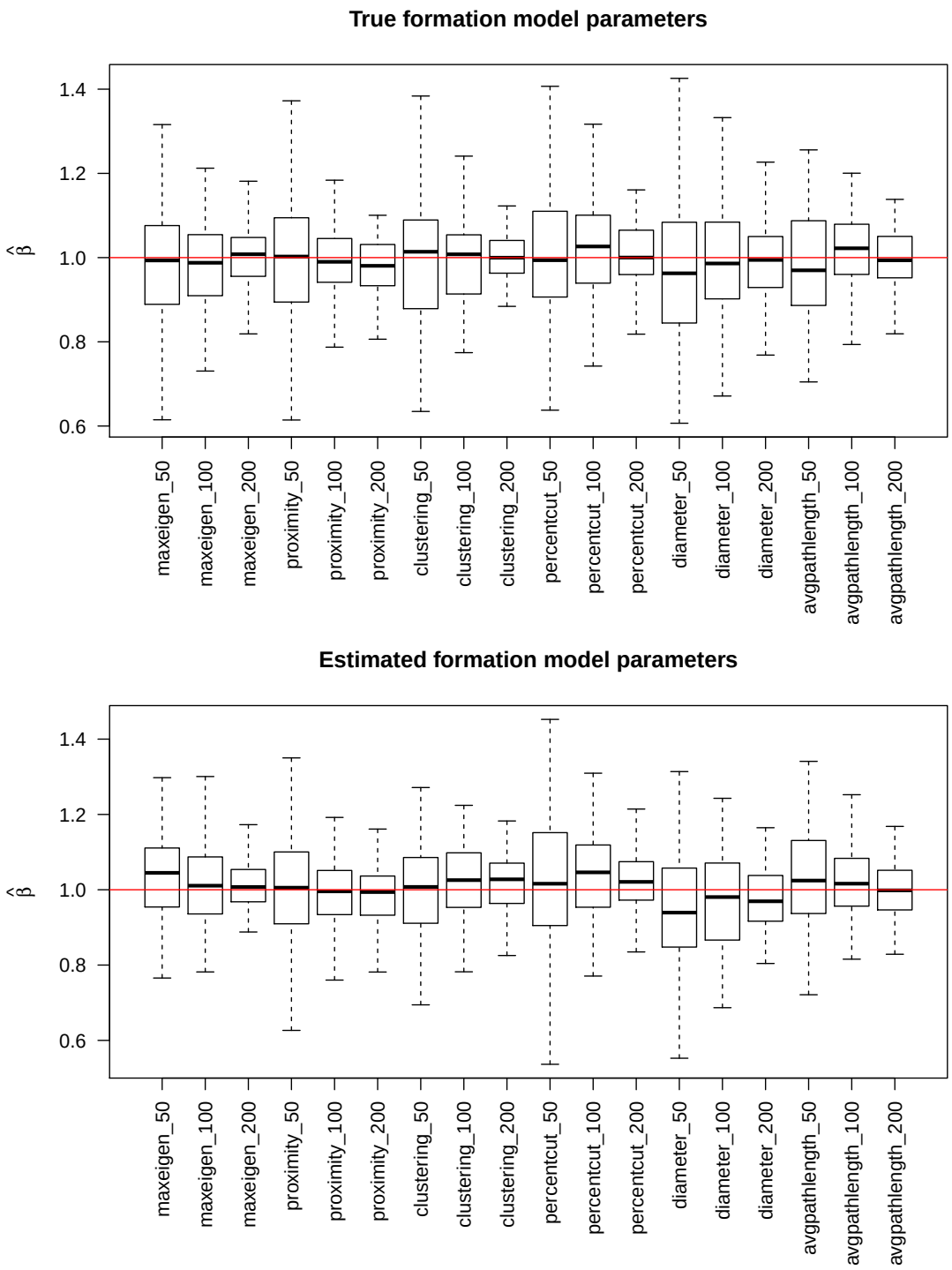


Figure 4.3: Boxplot of $\hat{\beta}$ for β in regression $y_r = \alpha + \beta \bar{S}_r + \epsilon_r$, where S_r and $\bar{S}_r$ represent a true and mean network-level measure, respectively. Each box represents the distribution of $\hat{\beta}$ for one measure and use of R=50, 100 or 200 networks in regression. The middle line of the boxplot denotes medium, and borders of the boxes denote first and third quartile. The red line denotes the true $\beta = 1$ used to generate $y_r = \alpha + \beta S_r^* + \epsilon_r$ in the simulation. These results corroborate the theoretical intuition developed in Section 4.4.2.

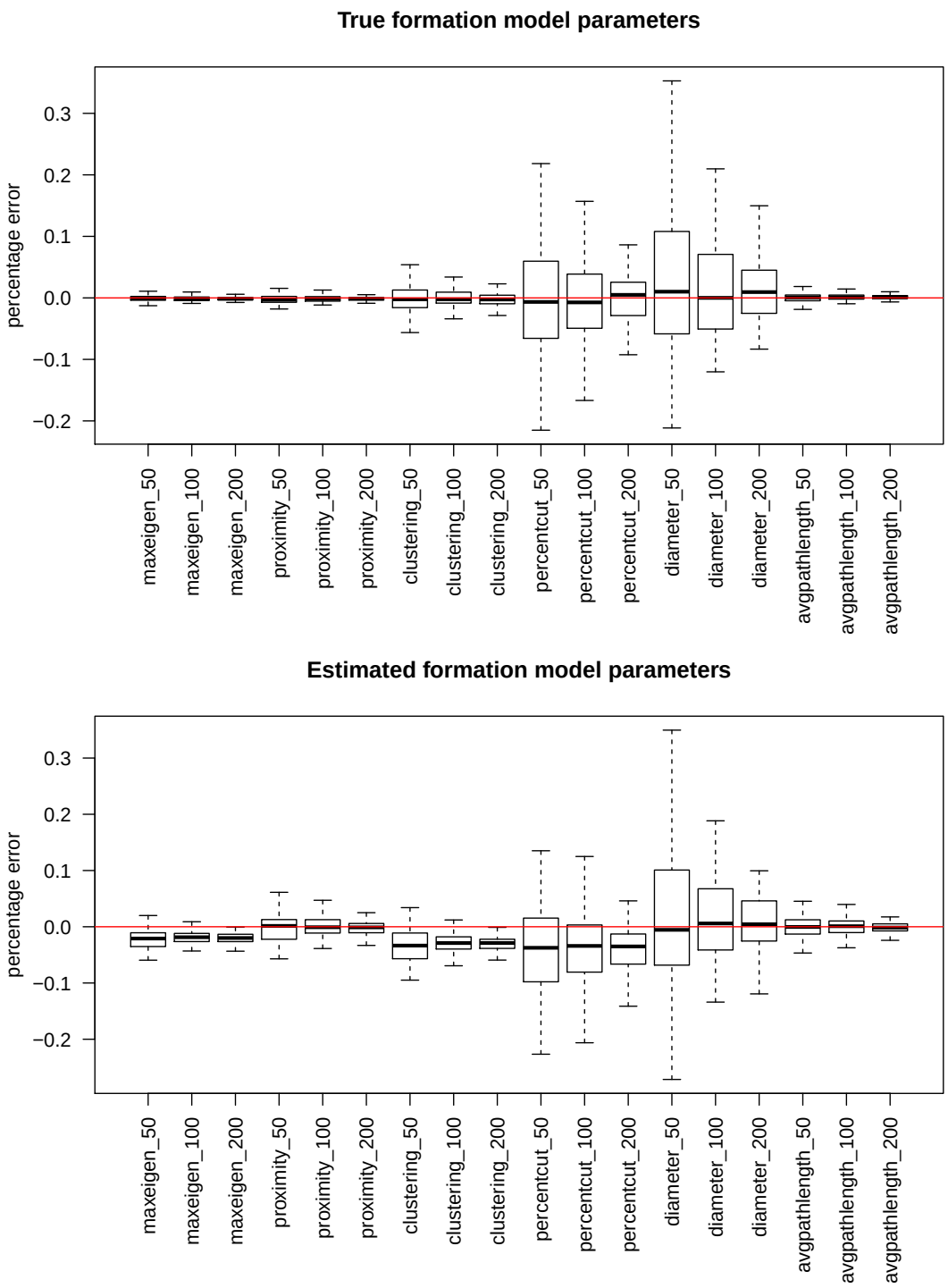


Figure 4.4: Boxplot of percentage errors of $\hat{\gamma}$ for γ in regression $\bar{S}_r = \alpha + \gamma T_r + \epsilon_r$, where S_r and $\bar{S}_r$ represent a true and mean network-level measure, respectively. Each box represents the distribution of percentage errors for one measure and use of R=50, 100 or 200 networks in regression. The middle line of the boxplot denotes medium, and borders of the boxes denote first and third quartile. These results corroborate the theoretical intuition developed in Section 4.4.2.

Chapter 5

BLOCK EXCHANGEABLE STANDARD ERRORS FOR NETWORK REGRESSION[1]

5.1 Introduction

Researchers are often interested in how relations between pairs of actors are affected by observable covariates, such as demographic, sociological and geographic factors. An ordered pair of actors is called a dyad. For example, Ward and Hoff (2007) examine political and institutional effects on international trade and find that the domestic political framework of the expoerter and importer are important factors of the trade; Aker (2010) explore the impact of mobile phones on the price difference of grain between a pair of markets and they find that the introduction of mobile phone service explains a reduction in grain price dispersion; Fafchamps and Gubert (2007) examine the role of geographic proximity on risk sharing among villagers and find that intra-village mutual insurance links are largely deter- mined by social and geographical proximity and are only weakly the result of purpose and diversification of income risk.

In this chapter, we focus on a case where continuous relations between pairs of actors are modeled as a linear function of observable covariates. Continuous pairwise relations can be represented in a network as directed and weighted edges, where weights of edges follow a continuous distribution. We refer to these relations as relational observations/response. We also have observed covariates of the actors and dyads and we want to study the the association between the relational response and covariates. The main contribution of this chapter is

[1]The contents of this chapter are based on the paper Pan et al.

a novel non-parametric block-exchangeability assumption on the covariance structure of the error vector, when there are hidden blocks in the network. When the underlying error structure of data satisfies the model assumptions, we have correct inference on the coefficients, which means that we have the correct confidence interval coverage. Additionally we present algorithms that estimate the standard errors of the error vector and the latent blocks. Our goal of the chapter is to provide a new approach to model and estimate the dependence between dyads, which bridges the gap between the existing non-parametric estimators.

We consider a linear regression model discussed in Section 2.2 and presented in Equation (2.4). For example, if we are interested in how geographical and demographic factors affect number of mobile calls between actors, then y_{ij} is the number of mobile calls from actor i to actor j and X_{ij} may include the actors' geographical distances and their mobile plans. Making inference on $\boldsymbol{\beta}$ then provides insights into how a change in geographical distance or mobile plans is associated with a change in the number of mobile calls.

In order to get accurate estimation of the standard error and thus a confidence interval with the correct coverage, we need to pose assumptions on the error structure that is satisfied by the data. The challenge is to model $\Omega = \mathrm{Var}(\Xi)$ since ξ_{ij} and ξ_{kl} are likely correlated if these relations share a member, i.e.$\{i, j\} \cap \{k, l\} \neq \emptyset$ (Kenny et al. (2006)) . For example, in a mobile traffic network, we may be interested in how the volume of traffic mobile calls is affected by geographical distance between two individuals and their mobile plans. Then the residuals of number of mobile calls from actor A to actor B and from actor A to actor C are likely correlated because they both involve actor A. The residuals represent variation in relational observations not accounted by observable covariates, and two residuals which both involve actor A may be affected by actor A's individual effects. For example, if the residual of number of mobile calls from actor A to actor B is negative, we may expect number of mobile calls from actor A to actor C is likely also negative, because actor A does not use

mobile calls a lot. Another example is the case of reciprocal edges (Miller and Kenny (1986)). The residuals of number of mobile calls from actor A to actor B and from actor B to actor A are also likely correlated, because they involve the same pair of actors. We would expect part of the variation in the number of phone calls made between them is attributed to interactions between actors, and not accounted for in covariates and random noise. For example, actors A and B stay in touch regularly using mobile calls, then if the residual of number of mobile calls from actor A to actor B is positive, we likely expect residual of number of mobile calls from actor B to actor A is positive.

One set of approaches to model the covariance structure Ω is to impose parametric distributional assumptions on the error vector or model the error covariance structure directly (Hoff (2005), Ward and Hoff (2007), Hoff et al. (2011), Hoff (2015)). While these approaches produce interpretable representations of underlying residual structure, they always assume the error structure is consistent with the underlying parametric model.

Another set of approaches to model the covariance structure Ω is through non-parametric approaches. However, such approaches either makes no distributional assumptions and estimate $\mathcal{O}(n^3)$ parameters (see dyadic clustering estimator in Fafchamps and Gubert (2007)), or assume exchangeability of the error vector Ξ and estimate five parameters (Marrs et al. (2017)). The former approach results in a standard error estimator for β that is extremely flexible yet extremely variable, whereas the latter approach assumes all actors are identically distributed and results in a relatively restricted estimator. The former approach is appealing when researcher do not have any information on the error structure, have a large network data and are not interested in Bernoulli covariates; the later approach is appealing when researchers are certain that errors are exchangeable and thus can enjoy the simplicity of the error structure and a fixed number of covariance parameters. Although both approaches appeal to researchers in some ways, there are likely cases where researchers have some

information about the error structure, but errors are not exchangeable. This is when an approach that bridges the gap between these two existing approaches is appealing.

We propose an alternative block-exchangeable estimator that assumes that actors have block memberships and actors within the same block are exchangeable (i.e. have relations that follow the same distribution). Heterogeneity based on unobserved variables are quite common in networks, and relational observations between actors in the same block may have different patterns than those observations between actors in different blocks. The stochastic block model (Holland et al. (1983), Snijders and Nowicki (1997)) and degree-correct stochastic block model (Karrer and Newman (2011)) have been proposed to model connectivity between actors based on latent block membership and actor degree heterogeneity. Spectral clustering algorithms (Rohe et al. (2011), Qin and Rohe (2013)) also have been proposed to estimate the hidden block membership for these models. By imposing the exchangeability assumption on the error vector conditioned on block membership of the actors, we take into account the possible block structure in the network and how block structure may affect the error patterns. We have developed an algorithm that estimates the covariance matrix $\hat{\Omega}$ given the block membership, as well as a second algorithm to estimate the block membership using spectral clustering. We also present theory and simulation results in how much block-exchangeable estimator outperforms exchangeable estimator when the errors are block-exchangeable but not exchangeable. The intuition is that, if the distribution of the covariates is dependent on block memberships, we see a larger difference in standard errors using block-exchangeable estimator and using exchangeable estimator.

This chapter is organized as follows. We describe our block-exchangeable estimator in Section 5.2 and present a simple network to demonstrate the covariance structure. We present our estimation algorithms in Section 5.3 and we state our theoretical results on the estimator and present proof in Section 5.4. We present our simulation results in Section 5.5. Finally, in

Section 5.6, we discuss limitations of our approach and future directions for improvements.

5.2 Block-exchangeability

In Section 2.2, we see that dyadic cluster estimator makes a single assumption but yields too many parameters, while the exchangeable estimator makes relatively restrictive assumptions. To bridge the gap between these two estimators, we propose block-exchangeability assumption that leverages between imposing assumptions on error vector and model complexity. In this section we present the definition of block-exchangeability and give a simple example of four-node network to illustrate the cases of edge directed pairs and covariance matrix.

In a network of B latent blocks, let g_i denote the block assignment of node i, $g_i \in \{1, ..., B\}$. We propose the following definition of **block-exchangeability** assumption by assuming conditional exchangeability (Lindley et al. (1981)) of Ξ given g:

DEFINITION **5.2.1.** *The errors in a relational data model are jointly block-exchangeable if $P(\Xi)$, the probability distribution of the error vector, is invariant under permutation of the rows and columns within each block:*

$$P(\Xi) = P(\prod(\Xi)) \text{ such that } g_i = g_{\pi(i)} \text{ and } g_j = g_{\pi(j)},$$

where $\prod(\Xi) = \{\xi_{\pi(i)\pi(j)}\}$ is the residual array with its rows and columns reordered according to permutation operator π.

Block-exchangeability assumption in regression settings has been discussed in McCullagh (2005), where a block-exchangeable process means that the distribution of samples is invariant under those permutations that preserve the block structure,i.e. permutations π such that $B(i,j) = B(\pi(i), \pi(j)) \forall i, j$, where $B(i,j) = 1$ if $g_i = g_j$, and $B(i,j) = 0$ otherwise. There are two differences between this assumption and the assumption we propose. One is that block-exchangeability in McCullagh (2005) is on the samples, where we propose block-exchangeability on the errors. The other is that when there are more than two blocks, the

permutation in our definition requires that $g_i = g_{\pi(i)}$ and $g_j = g_{\pi(j)}$, but the permutation in McCullagh (2005) only requires that $B(i,j) = B(\pi(i), \pi(j)) = 0$.

Under the block-exchangeability assumption and conditioned on block membership, for any four nodes $\{i, j, k, l\}$, the covariance between the errors ξ_{ij} and ξ_{kl} take one of the following six values depending on $g_i, g_j, g_k,$ and g_l:

- $\text{Var}(\xi_{ij}) = \sigma^2_{(g_i,g_j);}$
- $\text{Cov}(\xi_{ij}, \xi_{kj}) = \phi_{C,(g_j,\{g_i,g_k\});}$
- $\text{Cov}(\xi_{ij}, \xi_{ji}) = \phi_{A,\{g_i,g_j\};}$
- $\text{Cov}(\xi_{ij}, \xi_{ki}) = \phi_{D,(g_i,g_j,g_k);}$
- $\text{Cov}(\xi_{ij}, \xi_{il}) = \phi_{B,(g_i,\{g_j,g_l\});}$
- $\text{Cov}(\xi_{ij}, \xi_{kl}) = 0,$

where $\{\}$ denotes unordered set and $()$ denotes ordered set.

Figure 5.1 shows configurations of edge directed pairs under block-exchangeability assumption in a simple network of four nodes A, B, C, and D. We assume nodes A and B are in Block 1 (indicated by purple color) and nodes C and D are in Block 2 (indicated by light coral color). Each circle represents one dyad configuration, and the elements inside the circle represent block membership pairs/triplets for that particular dyad configuration, and each pair/triplet is associated with a unique parameter in Ω_B. The parameters following an arrow outside the circle means that all parameters in the circle would share the same value under exchangeability assumption, which is the parameter in Ω_E. The total number of parameters in Ω_B for this simple network is the sum of parameters over all circles, while the total number of parameters in Ω_B is five, the number of circles. The top left circle shows the variance parameters: $\sigma^2_{(1,1)}, \sigma^2_{(1,2)}, \sigma^2_{(2,1)},$ and $\sigma^2_{(2,2)}$. The variance of ξ_{ij} is determined by the block membership of sender i and receiver j. For example, in the topmost row, since both the sender A and the receiver B are in Block 1, $\text{Var}(\xi_{AB}) = \sigma^2_{(1,1)}$. On the contrary, under the exchangeability assumption, all variance term share the same parameter value σ^2. The

Covariance term	Number of parameters
$\text{Var}(\xi_{ij}) = \sigma^2$	B^2
$\text{Cov}(\xi_{ij}, \xi_{ji}) = \phi_A$	$B(B+1)/2$
$\text{Cov}(\xi_{ij}, \xi_{il}) = \phi_B$	$B^2(B+1)/2$
$\text{Cov}(\xi_{ij}, \xi_{kj}) = \phi_C$	$B^2(B+1)/2$
$\text{Cov}(\xi_{ij}, \xi_{ki}) = \phi_D$	$B^2(B+1)$

Table 5.1: Number of parameters for each covariance term under the block-exchangeability assumption.

bottom-middle circle in Figure 5.1 shows the covariance between reciprocal edges y_{ij}, y_{ji}. There are a total of three parameters: $\phi_{A,\{1,1\}}, \phi_{A,\{1,2\}},$ and $\phi_{A,\{2,2\}}$. The topright circle shows four parameter values of edge directed pairs of configuration (y_{ij}, y_{kj}) can take. For example, in the topleft corner of the circle, the common receiver actor B is in Block 1, and one sender A is in Block 1 and the other sender C is in Block 2. Therefore $\text{Cov}(\xi_{AB}, \xi_{CB}) = \phi_{B,(1,\{1,2\})}$. Because we only have two actors in one block, the case when $\text{Cov}(\xi_{ij}, \xi_{kj}) = \phi_{B,(1,\{1,1\})}$ is not shown in Figure 5.1 since it would require actors $i, j,$ and k are all in Block 1.

Figure 5.2 shows a visualization of Ω_B, a 12×12 matrix. Under both exchangeability and block-exchangeability assumption, the blank entries indicate a covariance value of zero between non-overlapping dyads (y_{ij}, y_{kl}) where $\{i, j\} \cap \{k, l\} = \emptyset$. Under block-exchangeability assumption, each color denotes a dyad configuration and conditioned on the color, each symbol denotes a parameter. So entries with the same color and symbol share the same parameter value. For example, $\text{Var}(\xi_{CA}) = \text{Var}(\xi_{DA}) = \text{Var}(\xi_{CB}) = \text{Var}(\xi_{DB}) = \sigma^2_{21}$ because the sender is in Block 2 and the receiver is in Block 1. This corresponds to the grid in the third row in the topleft circle in Figure 5.1. On the contrary, under exchangeability assumption, entries with the same color share the same value. Therefore, compared to Ω_E, Ω_B has more parameters than Ω_E, while maintaining the same places for zero-values entries.

Table 5.1 shows the number of parameters that each covariance term can take. Therefore

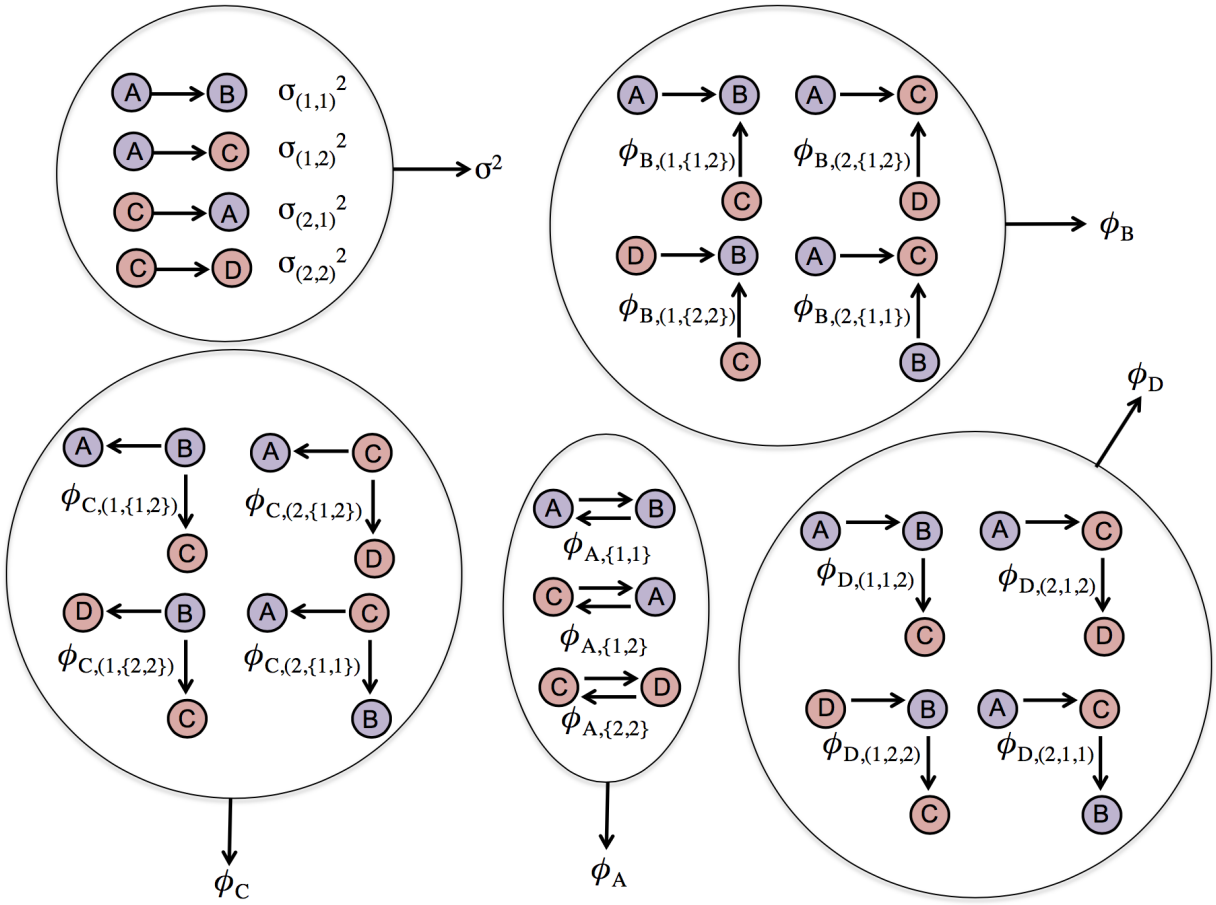


Figure 5.1: Configurations of edge directed pairs under block-exchangeability assumption in a simple network of four nodes A, B, C, and D. Nodes A and B are in one block (indicated by purple color) and nodes C and D are in the other (indicated by light coral color). Each circle represents one dyad configuration, and the parameters with in the circle are those under the block-exchangeability assumption. The parameters following an arrow outside the circle means that all parameters in the circle would share the same value under exchangeability assumption.

the number of parameters in Ω_B is on the order of $\mathcal{O}(B^3)$. This is greater than the number of parameters under exchangeability assumption, which is five regardless of network size. However, this number is significantly smaller than the number of parameters for dyadic clustering, which is on the order of $\mathcal{O}(n^3)$. Our block-exchangeablility estimator leverages

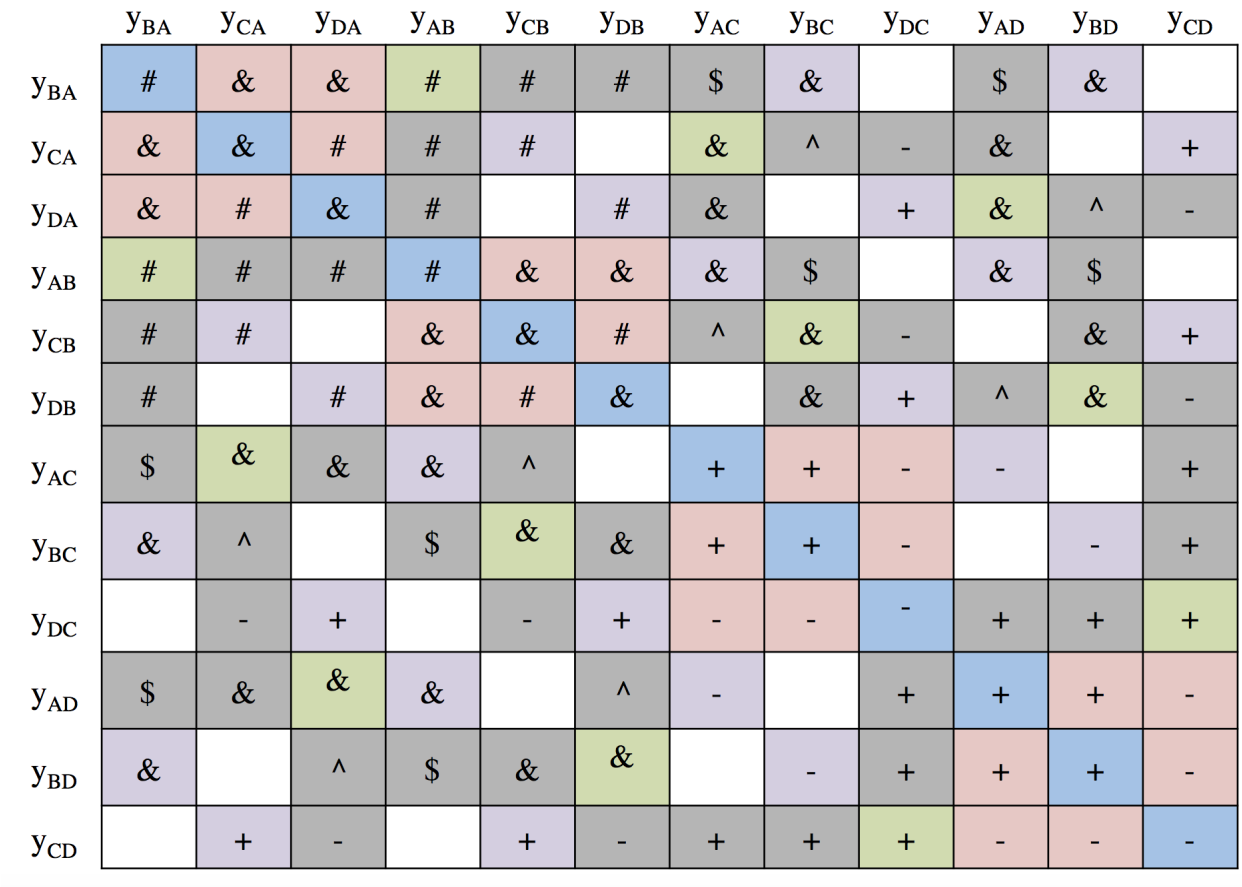


Figure 5.2: Visualization of covariance matrix for a simple network of four nodes. Under the block-exchangeabilty assumption, entries shaded with the same color and symbol share the same parameter value. While under the exchangeability assumption, entries with the same color share the same value.

between imposing assumptions on error vector and model complexity, in an attempt to model the covariance matrix with reasonable number of parameters while keeping the assumptions feasible for real world applications.

5.3 Network Regression with Block-exchangeable Errors

Assuming the errors are block-exchangeable and there are B blocks, we now present algorithms that produce standard error estimation for the coefficients $\boldsymbol{\beta}$ in a linear regression model (2.4). Section 5.3.1 presents an algorithm that estimates the the covariance values by average of residual products when block memberships are known. Section 5.3.2 presents an algorithm that estimates unobserved blocks using residuals, when block memberships are unknown. As a preliminary step for both algorithms, let $\boldsymbol{X}$ be the design matrix, we compute OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$ and calculate residuals $r_{ij} = y_{ij} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

5.3.1 Known blocks

When given block memberships, this algorithm get residual pairs that share the same covariance value. Since the expected values of residuals are zero, we estimate the covariance value by taking averages of products of these residual pairs.

Let $[B] = \{1, ..., B\}$ and $[n] = \{1, ..., n\}$. Let $\mathcal{M}$ denote the set of dyad configurations $\{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, and Q_M denote the set of block pairs/triplets for dyad configuration $M \in \mathcal{M}$ given $[B]$. Then $Q_{\sigma^2} = \{(u, v), u, v \in [B]\}$, and we present the definitions of Q_{ϕ_A}, Q_{ϕ_B}, Q_{ϕ_C} and Q_{ϕ_D} in the appendix A.1. Let $M_q, q \in Q_M$ denote the parameters in Ω_B. For example, for variance component with both actors in block one, we take $M = \sigma^2$, and $q = (1, 1)$. Then the parameters $M_q = \sigma^2_{(1,1)}$.

Given:

Residuals $r_{ij} \ \forall\, i \neq j$, number of blocks B, and block membership $g_i \,\forall\, i$.

Output: $\widehat{\Omega}_B$

Algorithm 1

1. Let $\Phi_{M,q}$, where $q \in Q_M$, denote the set of edge directed pairs that have the configuration M and block specification q. Then $\Phi_{\sigma^2,(u,v)} = \{[(i,j),(i,j)] : i,j \in [n], i \neq j, g_i = u, g_j = v\}$, and we present the definitions of $\Phi_{\phi_A,\{u,v\}}$, $\Phi_{\phi_B,(u,\{v,w\})}$, $\Phi_{\phi_C,(u,\{v,w\})}$ and $\Phi_{\phi_D,(u,\{v,w\})}$ in the appendix A.1.

2. Calculate the set of corresponding residual products for $\Phi_{M,q}$, where $q \in Q_M$:

$$\boldsymbol{R}_{M,q} = \{r_{jk} r_{mn} : [(j,k),(m,n)] \in \Phi_{M,q}\}$$

3. Compute point estimates of parameters in Ω_B using average of residual products:

$$\widehat{M_q} = \frac{\sum_{t:t\in \boldsymbol{R}_{M,q}} t}{|\boldsymbol{R}_{M,q}|}$$

where $\widehat{M_q}$ denote the estimator for parameter M_q with dyad configuration M and block specification q where $q \in Q_M$.

5.3.2 Unknown blocks

When block membership is unknown, we use spectral clustering to estimate block membership by constructing a similarity matrix from residuals. Specifically for each actor i, we extract all pairs of residuals that involve actor i as an overlapping actor, for each dyad configuration M. Then we use Kolmogorov-Smirnov statistic between the distribution of residual products that involve actor i and such distribution that involve actor j, to create a similarity value between a pair of actors i and j.

Given:

Residuals $r_{ij}\ \forall\ i \neq j$ and number of blocks B

Output:

Estimated block membership $\hat{g}_i$

Algorithm 2

1. For $M \in \mathcal{M}$, let $\Phi_{M,i}$ denote the set of dyads that involve a specific node i : $\Phi_{\sigma^2,i} = \{[(i,j),(i,j)] : j \in [n], i \neq j\} \cup \{[(j,i),(j,i)] : j = 1, ..., n, i \neq j\}$; $\Phi_{\phi_B,i} = \{[(i,j),(i,k)] : j \in [n], k \in [n], i \neq j \neq k\}$. We put the rest of definitions in Appendix A.1.

2. For each i and M, calculate the set of corresponding residual products for $\Phi_{M,i}$:

$$\boldsymbol{R}_{M,i} = \{r_{ab}r_{cd} : [(a,b),(c,d)] \in \Phi_{M,i}\}$$

3. Let $F_{i,M}$ be the empirical distribution function for $\boldsymbol{R}_{M,i}$. For each pair of actors i and j and for each M, get Kolmogorov–Smirnov statistic $KS_{i,j,M} = \sup_x |F_{i,M}(x) - F_{j,M}(x)|$.

4. For each pair of actors i and j, define

$$s_{ij} = 1 - \left(\sum_{M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}}^{5} KS_{i,j,M} \right) /5.$$

5. Let $W = (w_{ij})_{i,j=1,...,n}$ denote the weighted adjacency matrix,where

$$w_{ij} = w_{ji} = \begin{cases} s_{ij}, & \text{if } i \in KNN(j) \text{ or } j \in KNN(i) \\ 0, & \text{otherwise} \end{cases}$$

6. Perform unnormalized spectral clustering (Von Luxburg (2007)) on weighted graph W to get estimated blocks $\hat{g}_i$, where $\hat{g}_i \in [B]\ \ \forall i$.

When block membership is known, we apply Algorithm 1 to get $\widehat{\Omega}_B$. When block membership is unknown, we apply Algorithm 2 to get $\hat{g}_i$ and apply Algorithm 1 using $\hat{g}_i$ as an input to get $\widehat{\Omega}_B$. We keep K as a tuning parameter in Step 5 of Algorithm 2. Maier et al. (2007) prove that choosing $K = c_1 n - c_2 \log(n/K) + c_3$, where $c_1, c_2 \geq 0$ and c_3 are all constants, provides an optimal choice of K. we have found that for our simulation setting $K = 0.2n$ works the best. When block membership is known, we find that the computational cost to get $\widehat{\Omega}_B$ is quite inexpensive because the algorithm just extracts all dyad pairs with the same covariance and take average of residual products. The computational time to get $\widehat{\Omega}_B$ is five seconds when $n = 80$, and is 20 seconds when $n = 160$. When the block membership is unknown, Step 2 and 3 of Algorithm 2 may be expensive if the network size is large. Therefore we suggest a modification of Step 3. Instead of letting $F_{i,M}$ be the empirical distribution function for $\boldsymbol{R}_{M,i}$, we modify $F_{i,M}$ to be the empirical distribution function for quantiles of $\boldsymbol{R}_{M,i}$. This reduces the size of the set $\boldsymbol{R}_{M,i}$, which decreases the storage cost as well as the computational cost of getting Kolmogorov-Smirnov statistic.

5.4 Theoretical analysis of estimator

If the data satisfies the assumptions we pose on the error structure, then we get an accurate estimation of the standard errors, and confidence intervals constructed with such standard errors have the correct coverage. This is why an accurate estimation of standard errors is important in inference on the coefficients.

After we get $\widehat{\Omega}_B$ from Algorithm 1, we use sandwich estimator to get the estimated standard errors of $\hat{\boldsymbol{\beta}}$:

$$\hat{V}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\Omega}_B\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{5.1}$$

We observe that entries in $\hat{V}(\hat{\boldsymbol{\beta}})$ are entries in $\widehat{\Omega}_B$ weighted by functions of $\boldsymbol{X}$. It is possible that even with the wrong structure in Ω, we still get a good standard error estimation of $\widehat{\Omega}_B$

because the difference $\widehat{\Omega}_B - \widehat{\Omega}_E$ is evened out by $\boldsymbol{X}$. The goal of this section is to quantify the difference in standard errors of $\hat{\boldsymbol{\beta}}$ using exchangeable estimator an block-exchangeable estimator, as a function of covariate $\boldsymbol{X}$, block assignment g_i, and parameters in Ω. More precisely, the difference converges in probability to weighted average of differences between true parameters and parameters when the error structure is mis-specified as exchangeable.

We show our theoretical results in a simple linear regression model with only one covariate:

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + \xi_{ij}, \tag{5.2}$$

where y_{ij} is the observed relational array, X_{ij} is a single covariate, ξ_{ij} is the error term, β_0 is the intercept, and β_1 is the slope. In addition, we assume that there are two blocks in the network, with block size n_1 and n_2 respectively, and we do not specify the distribution of the covariate. Under the assumption that the error vector is block-exchangeable, we show that the difference in $\hat{V}(\hat{\boldsymbol{\beta}})$ using exchangeable estimator and block-exchangeable estimator converges in probability to a matrix that depends on the distribution of the covariate, block assignment and parameters in Ω_B.

Let $\hat{V}_B(\hat{\boldsymbol{\beta}})$ be the estimated standard error of $\hat{\boldsymbol{\beta}}$ using block-exchangeable estimator and let $\hat{V}_E(\hat{\boldsymbol{\beta}})$ be the estimated standard error of $\hat{\boldsymbol{\beta}}$ using exchangeable estimator.

THEOREM **5.4.1.** *Assume*

1. *there are 2 blocks of size n_1 and n_2, respectively;*

2. *the error vector satisfies the block-exchangeability assumption with with covariance structure as in Figure 5.2, and parameters $\sigma^2_{(u,v)}$, $\phi_{A,\{u,v\}}$, $\phi_{B,(u,\{v,w\})}$, $\phi_{C,(u,\{v,w\})}$, $\phi_{D,(u,\{v,w\})}$, where $u, v, w \in \{1, 2\}$;*

3. $\boldsymbol{X}$ *is full rank, where* $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{21} & \boldsymbol{X}_{31} & \cdots & \boldsymbol{X}_{n1} & \cdots & \boldsymbol{X}_{1n} & \cdots & \boldsymbol{X}_{(n-1)n} \end{bmatrix}^T$, *and* $\boldsymbol{X}_{ij} = \begin{bmatrix} 1 & X_{ij} \end{bmatrix}$;

4. *the covariate* X_{ij} *are independent and identically distributed;*

5. *the fourth moment of the errors and covariartes are bounded:* $\mathbb{E}[(\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T)^2] \leq C < \infty$ *where the square is taken element-wise on* $\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T$, *and* $\mathbb{E}[\xi_{jk}^4] \leq C' < \infty$,

6. *the errors* Ξ *and* $\boldsymbol{X}$ *are independent;*

7. *the number of blocks* B *is* $\mathcal{O}(1)$.

we show that

1. *As* $n_1 \to \infty, n_2 \to \infty$, *and* $n_1/n_2 \to$ *a constant*

$$n\left(\hat{V}_B(\hat{\boldsymbol{\beta}}) - \hat{V}_E(\hat{\boldsymbol{\beta}})\right) \xrightarrow{p} c(\boldsymbol{X}). \tag{5.3}$$

$c(\boldsymbol{X})$ is a weighted linear combination of the differences between the block exchangeable parameters and the exchangeable parameter, where the weights are specified by functions of $\boldsymbol{X}$:

$$\begin{aligned} c(\boldsymbol{X}) &= \sum_{M,q \in Q_M} f_{M,q}(M_q - M) \\ &= \sum_{u,v \in \{1,2\}} f_{\sigma^2,(u,v)}(\sigma^2_{(u,v)} - \sigma^2) + \sum_{u,v \in \{1,2\}} f_{\phi_A,\{u,v\}}(\phi_{A,\{u,v\}}) - \phi_A) + ... \end{aligned} \tag{5.4}$$

where $f_{M,q}$ are functions of $\boldsymbol{X}$. More specifically, given M and q, $h_{M,q}$ is a function of

elements in the set $\{[X_{ij}, X_{kl}]|[(i,j),(k,l)] \in \Phi_{M,q}\}$. *The parameter*

$$\sigma^2 = \frac{n_1(n_1-1)\sigma^2_{(1,1)} + n_2(n_2-1)\sigma^2_{(2,2)} + n_1 n_2(\sigma^2_{(1,2)} + \sigma^2_{(2,1)})}{n(n-1)} \tag{5.5}$$

We discuss $f_{M,q}$ *in more detail and present definitions of* $c(\boldsymbol{X})$ *in Section 5.4.1.*

2. *When* X_{ij} *is independent of* g_i *and* g_j*, we show that the quantity on the right-hand-side is zero. More specifically, we show in the Appendix that each of the five terms in Equation* (5.4) *is zero.*

The insight from the theorem is that $SE(\hat{\beta}_1)_B - SE(\hat{\beta}_1)_E$ converges in probability to weighted sum of a difference between a true parameter and a parameter if the exchangeability assumption is satisfied. In terms of the variance, σ^2 is a weighted average of $\sigma^2_{(1,1)}, \sigma^2_{(1,2)}, \sigma^2_{(2,1)}$ and $\sigma^2_{(2,2)}$, and we can interpret σ^2 as the variance term if the exchangeability on the error vector is satisfied. When $\sigma^2_{(1,1)}, \sigma^2_{(1,2)}, \sigma^2_{(2,1)}$ and $\sigma^2_{(2,2)}$ are not all equal, the differences $\sigma^2_{(1,1)} - \sigma^2, \sigma^2_{(1,2)} - \sigma^2, \sigma^2_{(2,1)} - \sigma^2, \sigma^2_{(2,2)} - \sigma^2$ may be nonzero. In fact, since σ^2 is a weighted average, some of the differences will be positive, and some will be negative. And the differences are adjusted by their weights $f_{\sigma^2,(u,v)}$ to determine the difference in standard errors. We can use the same logic for the rest four configurations, and finally we can see that the magnitude and sign of the difference in standard errors using block-exchangeable estimator and exchangeable estimator are determined by a sum of weighted differences on all five configurations. Therefore, we can conclude that whether exchangeable estimator has over- or under- coverage depends on parameters in Ω, g_i, and X_{ij}.

Part 2 of the theorem says that even if the differences $\sigma^2_{(1,1)} - \sigma^2, \sigma^2_{(1,2)} - \sigma^2, \sigma^2_{(2,1)} - \sigma^2, \sigma^2_{(2,2)} - \sigma^2$ are nonzero, as long as X_{ij} is independent of g_i and g_j, on average the difference will disappear after adjusted by weights. This is an important insight, because we see that in order for

block-exchangeable estimator to have lower bias than exchangeable estimator, we need:

1. the error vector satisfies block exchangeability but not exchangeability.

2. the distribution of X_{ij} is independent of g_i and g_j

If either one of the conditions is not satisfied, we will not see that the block-exchangeable estimator has more accurate coverage than the exchangeable estimator.

5.4.1 Proof

In this section, we present a proof of Theorem 5.4.1.

$$
\begin{aligned}
& n\left(\hat{V}_B(\hat{\boldsymbol{\beta}}) - \hat{V}_E(\hat{\boldsymbol{\beta}})\right) \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T(\widehat{\Omega}_B - \widehat{\Omega}_E)X(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \frac{n}{n^2(n-1)^2}\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n(n-1)}\right)^{-1}\left(\sum_{M\in\mathcal{M}}\sum_{q\in Q_M}\frac{\sum\limits_{(j,k),(m,n)\in\Phi_{M,q}}\boldsymbol{X}_{jk}\boldsymbol{X}_{mn}^T\left(\widehat{M}_q - \widehat{M}\right)|\Phi_{M,q}|}{|\Phi_{M,q}|}\right)\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n(n-1)}\right)^{-1} \\
&= \sum_{M\in\mathcal{M}}\sum_{q\in Q_M}\frac{|\Phi_{M,q}|}{n(n-1)^2}\left(\widehat{M}_q - \widehat{M}\right)\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n(n-1)}\right)^{-1}\left(\frac{\sum\limits_{(j,k),(m,n)\in\Phi_{M,q}}\boldsymbol{X}_{jk}\boldsymbol{X}_{mn}^T}{|\Phi_{M,q}|}\right)\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n(n-1)}\right)^{-1} \\
&= \sum_{M\in\mathcal{M}}\sum_{q\in Q_M}\frac{c_{M,q}\cdot|\Phi_M|}{n(n-1)^2}\left(\widehat{M}_q - \widehat{M}\right)h_{M,q}(\boldsymbol{X}) \\
&= \sum_{M\in\mathcal{M}}\sum_{q\in Q_M}c'_M c_{M,q}\left(\widehat{M}_q - \widehat{M}\right)h_{M,q}(\boldsymbol{X}) \qquad (5.6)
\end{aligned}
$$

where $c'_M = \frac{|\Phi_M|}{n(n-1)^2}$, $c_{M,q}$ is the proportion of dyad pairs with configuration M and block specification q over all dyad pairs with configuration M, and $h_{M,q}$ contains the remaining terms which are functions of $\boldsymbol{X}$. Because we assume B is $\mathcal{O}(1)$, each $|\Phi_M|$ is at most $\mathcal{O}(n^3)$, so each $c'_M \to d_M$ for some constant d_M. Marrs et al. (2017) (Eq.27) show that

$h_{M,q}(\boldsymbol{X}) \xrightarrow{p} h'_{M,q}(\boldsymbol{X}) =$

$$\begin{cases} \mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T]^{-1}\,\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T|(j,k)\in\Phi_{\sigma^2,q}]\,\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T]^{-1}, & \text{for } M=\sigma^2 \\ \mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T]^{-1}\,\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{mn}^T|(j,k),(m,n)\in\Phi_{M,q}]\,\mathbb{E}[\boldsymbol{X}_{jk}\boldsymbol{X}_{jk}^T]^{-1}, & \text{for } M\in\mathcal{M}\setminus\sigma^2 \end{cases}$$

We have shown $c_{M,q}$ and $h_{M,q}$ both converge in probability to constants. So the only part left in Equation 5.6 is $\left(\widehat{M_q}-\widehat{M}\right)$. Marrs et al. (2017) have shown that

$$\widehat{M_q} \xrightarrow{p} M_q \text{ and } \widehat{M} \xrightarrow{p} M, \tag{5.7}$$

where

$$M = \frac{\sum\limits_{q\in Q_M} M_q\cdot|\Phi_{M,q}|}{\sum\limits_{q\in Q_M}|\Phi_{M,q}|} = \sum_{q\in Q_M} M_q\cdot c_{M,q} \tag{5.8}$$

Therefore, by Slutsky's theorem,

$$n\left(\hat{V}_B(\hat{\boldsymbol{\beta}})-\hat{V}_E(\hat{\boldsymbol{\beta}})\right) \xrightarrow{p} \sum_{M\in\mathcal{M}}\sum_{q\in Q_M}(M_q-M)\,f_{M,q}(\boldsymbol{X}), \tag{5.9}$$

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

$$\begin{aligned} n\left(\hat{V}_B(\hat{\boldsymbol{\beta}})-\hat{V}_E(\hat{\boldsymbol{\beta}})\right) &\xrightarrow{p} \sum_{M\in\mathcal{M}} d_M f_M(\boldsymbol{X}) \sum_{q\in Q_M} c_{M,q}\,(M_q-M) \\ &= \sum_{M\in\mathcal{M}} d_M f_M(\boldsymbol{X}) \left(\sum_{q\in Q_M} c_{M,q}M_q - \sum_{q\in Q_M} c_{M,q}M\right) \\ &= \sum_{M\in\mathcal{M}} d_M f_M(\boldsymbol{X})\,(M-M)=0 \end{aligned} \tag{5.10}$$

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

In the case of two blocks,

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

where

- $\sigma^2 = \dfrac{n_1(n_1-1)\sigma^2_{(1,1)} + n_2(n_2-1)\sigma^2_{(2,2)} + n_1n_2(\sigma^2_{(1,2)} + \sigma^2_{(2,1)})}{n(n-1)}$
- $\phi_A = \dfrac{n_1(n_1-1)\phi_{A,\{1,1\}} + n_2(n_2-1)\phi_{A,\{2,2\}} + 2n_1n_2\phi_{A,\{1,2\}}}{n(n-1)}$
- $\phi_B = \dfrac{n_1(n_1-1)(n_1-2)\phi_{B(1,\{1,1\})} + 2n_1(n_1-1)n_2\phi_{B(1,\{1,2\})} + n_1n_2(n_2-1)\phi_{B(1,\{2,2\})}}{n(n-1)(n-2)}$
  $+\dfrac{+n_2(n_2-1)(n_2-2)\phi_{B(2,\{2,2\})} + 2n_2n_1(n_2-1)\phi_{B(2,\{1,2\})} + n_2n_1(n_1-1)\phi_{B(2,\{1,1\})}}{n(n-1)(n-2)}$
- $\phi_C = \dfrac{n_1(n_1-1)(n_1-2)\phi_{C(1,\{1,1\})} + 2n_1(n_1-1)n_2\phi_{C(1,\{1,2\})} + n_1n_2(n_2-1)\phi_{C(1,\{2,2\})}}{n(n-1)(n-2)}$
  $+\dfrac{+n_2(n_2-1)(n_2-2)\phi_{C(2,\{2,2\})} + 2n_2n_1(n_2-1)\phi_{C(2,\{1,2\})} + n_2n_1(n_1-1)\phi_{C(2,\{1,1\})}}{n(n-1)(n-2)}$
- $\phi_D = \dfrac{n_1(n_1-1)(n_1-2)\phi_{D(1,1,1)} + n_1(n_1-1)n_2\phi_{D(1,1,2)} + n_1(n_1-1)n_2\phi_{D(1,2,1)}}{n(n-1)(n-2)}$
  $+\dfrac{n_1n_2(n_2-1)\phi_{D(1,2,2)} + n_2(n_2-1)(n_2-2)\phi_{D(2,2,2)} + n_2n_1(n_2-1)\phi_{D(2,1,2)}}{n(n-1)(n-2)}$
  $+\dfrac{n_2n_1(n_2-1)\phi_{D(2,2,1)} + n_2n_1(n_1-1)\phi_{D(2,1,1)}}{n(n-1)(n-2)}.$

5.5 Simulations

In this section we present the error generating model in Section 5.5.1 and discuss its connection to the block-exchangeability assumption. Then we results on confidence interval coverage in a variety of simulation cases when the error is block-exchangeable in Section 5.5.2. We also include our simulation results on how well we estimate the latent block membership with varying strength of block structure in Appendix A.2.

5.5.1 Error generating Model

While the block exchangeability assumption does not specify a parametric form for the error vector, in our simulation study, we use a generative model to generate errors that satisfy the block-exchangeability assumption. The error generating model is adapted from the bilinear mixed effects model in Hoff (2005) by adding block membership to the parameters. We consider a linear regression model in Equation (5.2) and the errors are generated that

$$\xi_{ij} = a_i + b_j + z_i^T z_j + \gamma_{(ij)} + \epsilon_{ij},$$

where

$$(a_i, b_i)|g_i \sim N_2(0, \Sigma_{ab,g_i});\ \Sigma_{ab,g_i} = \begin{pmatrix} \sigma^2_{a,g_i} & \rho_{ab}\sigma_{a,g_i}\sigma_{b,g_i} \\ \rho_{ab}\sigma_{a,g_i}\sigma_{b,g_i} & \sigma^2_{b,g_i} \end{pmatrix};$$

$$z_i|g_i \sim N_d\left(0, \sigma^2_{z,g_i} I_d\right);\ \epsilon_{ij} \sim N(0, \sigma^2_\epsilon);$$

$$\gamma_{(ij)} = \gamma_{(ji)}|g_i, g_j \sim (0, \sigma^2_{\gamma,\{g_i,g_j\}}).$$

Under the generative model, the variance and covariances are:

- $\mathrm{Var}(\xi_{ij}) = \sigma^2_{a,g_i} + \sigma^2_{b,g_j} + d\sigma^2_{z,g_i}\sigma^2_{z,g_j} + \sigma^2_{\gamma,\{g_i,g_j\}} + \sigma^2_{\epsilon;}$

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

We recognize that the error vector satisfies the block-exchangeability by making the observation that $\text{Cov}(\xi_{ij}, \xi_{kl}) = \text{Cov}(\xi_{\pi(i)\pi(j)}, \xi_{\pi(k)\pi(l)})$ with $g_i = g_{\pi(i)}, g_j = g_{\pi(j)}, g_k = g_{\pi(k)},$ and $g_l = g_{\pi(l)}$. However, this does not correspond to the most general form of the covariance matrix Ω_B that satisfy block-exchangeability. For example, under the error generating model, $\text{Cov}(\xi_{ij}, \xi_{il})$ takes B parameters, compared to $B^2(B+1)/2$ in the most general form in Table 5.1.

Figure 5.3 shows a visualization of the covariance matrix Ω_B under the error generating model. Entries shaded with the same color share the same covariance value. Compared to Figure 5.2, the error generative model does not correspond to the most general formualtion of block-exchangeability covariance structure. For example, $cov(\xi_{ij}, \xi_{ik})$ can take B values under the error generating model, but on the order of B^3 with the most general formulation.

5.5.2 Evaluating Block-exchangeable Estimator

5.5.2.1 Generating Covariates

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

1. $X_{ij,1} = \mathbf{1}_{X_i = X_j}$, where $X_i \sim \text{Bernoulli}(p_{g_i})$ and
   (a) p_{g_i} is uncorrelated with g_i, i.e., p_{g_i} is a fixed number

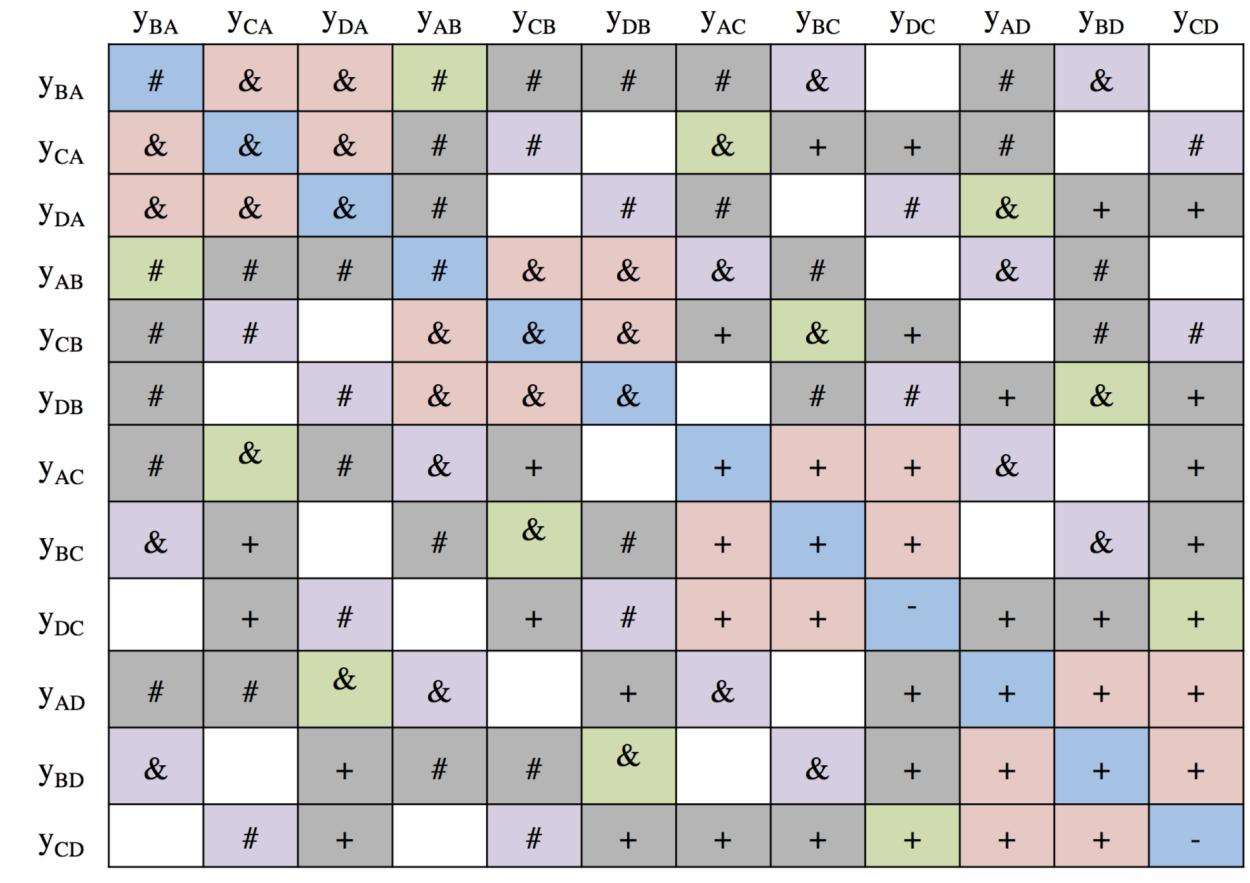


Figure 5.3: Visualization of covariance matrix Ω under the error generating model used in simulation. Entries shaded with the same color and symbol share the same parameter value, and a white box indicates a covariance of zero.

(b) $p_{g_i}|g_i = 2 > p_{g_j}|g_j = 1 > 0.5$, which suggests that high $\text{Var}(X_{ij,1})$ is associated with high $\text{Var}(\xi_{ij})$

(c) $p_{g_i}|g_i = 1 > p_{g_j}|g_j = 2 > 0.5$, which suggests that high $\text{Var}(X_{ij,1})$ is associated with low $\text{Var}(\xi_{ij})$

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

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

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

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

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

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

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

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

5.5.2.2 Noise To Signal Ratio

In all simulations we set $\beta_0 = 1$ and $\beta_1 = 1$. We set the parameters for generating covariates such that the noise to signal ratio, which is defined as the ratio of sum of squared errors over total sum of squares, is consistent across all three scenarios. Let NTS denote the noise-to-signal ratio, then

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

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

Therefore, for all three types of covariates:

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

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

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

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

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

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

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

5.5.2.3 Simulation Results

In this section, we present how confidence interval coverages compare among different estimators. Figure 5.4 shows the coverage of 95% confidence interval for β_1 for all nine scenarios of simulation cases. The first column represents the cases where the covariate X_{ij} is uncorrelated with block membership, the second column represents the cases where high variance in X_{ij} is correlated with low variance in ξ_{ij}, and the third column represents the cases where high variance in X_{ij} is correlated with high variance in ξ_{ij}. The rows represent different covariates: the first row is $X_{ij,1} = \mathbf{1}_{X_i = X_j}$, the second row represents $X_{ij,2} = |X_i - X_j|$, and the third row represents $X_{ij,3} \sim N(0, a^2_{g_i,g_j})$. We present the coverage of 95% confidence interval for β_1 over 1000 simulated errors given each of 500 simulations of the covariate and block membership, using four estimators of the Ω. We perform each simulation on network of size $20, 40, 80,$ and 160. The red box shows the coverage using the block-exchangeable estimator with the true block membership, the blue box shows the coverage using the block-exchangeable estimator with the estimated block membership, the yellow box shows the coverage using exchangeable etsimator, and the purple box shows the coverage using the dyad clustering estimator. For boxplots, the middle line indicates the median coverage, the top and bottom

boundaries indicate the 90% and 10% percentile, and the top and bottom whiskers indicate the 2.5% and 97.5% percentile.

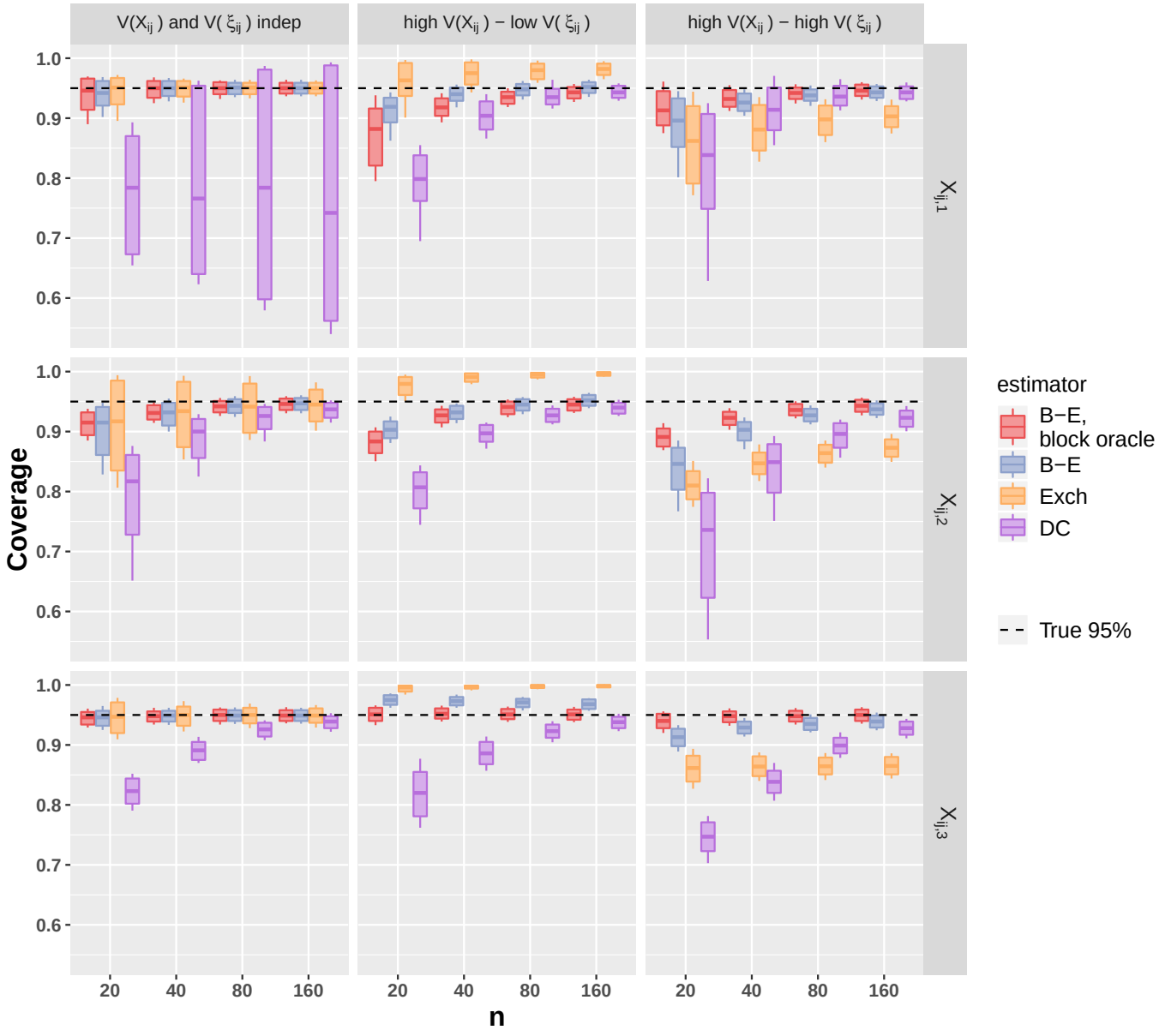


Figure 5.4: Coverage of 95% confidence interval for β_1 for three scenarios using block-exchangeable estimator with block oracle, block-exchangeable estimator with estimated blocks, excheangeable estimator, and dyadic clustering estimator.

Our main takeaway from Figure 5.4 is that the block-exchangeable estimator performs similarly as the excheangeable estimator, when the covariate is uncorrelated the errors, while block-exchangeable estimator outperforms exchangeable estimator greatly when the covariate is correlated the errors in both ways. This is consistent with our theory in Section 5.4. When high variance in X_{ij} is correlated with low variance in ξ_{ij}, we observe that the exchangeable estimator has over-coverage, and the bias increases with increasing network size. On the contrary, the bias of block-exchangeable estimator decreases with increasing network size. When high vairiance in X_{ij} is correlated with high variance in ξ_{ij}, the exchangeable estimator has under-coverage, and its performance does not improve much with increasing network size. At $n = 160$, its coverage is worse than dyadic clustering estimator, which is evidence that estimators with strict assumptions perform worse than distribution-free estimators when the assumptions are violated. In addition, we observe that the differences between the block-estimator using true block membership and the block-estimator using estimated block membership decreases with increasing network size, which suggests that our block estimation gets better with increasing network size.

5.6 Discussion

In this chapter, we propose a novel block-exchangeable estimator to estimate the standard errors of regression coefficients, assuming block-exchangeability. Our proposed estimator bridges the gap between the existing dyadic clustering estimator where no distributional assumptions are made, and the exchangeable estimator where the joint distribution of errors are assumed to be exchangeable. Through theory and simulations, we have shown that when latent blocks are dependent on the generative process of covariates, our block-exchangeable estimator outperforms exchangeable estimator by having less bias of coverage, and outperforms dyadic clustering estimator by having less variance. Although our simulations are done with

simple linear regression with two latent blocks, these observations can be extended to multiple linear regression and more blocks.

There are a few limitations of our work, we discuss them here. Our Algorithm 2 assumes the number of blocks as an input, which may not be the case empirically. In this scenario, information score or cross validation can be used as a criteria for model selection. Additionally, if the block sizes are unbalanced, the variance of the estimated parameters associated with the smallest block may be large, and comparing the performances of different estimators at various levels of unbalanced block size can be studied for future research. In this chapter, we consider linear regression and continuous relational observations on a fully connected network, and we assume nodes are sampled randomly. Some of the future directions for this work include extending it to respondent-driven sample, considering binary or count data with block-exchangeable errors, or working with a non fully connected network.

Chapter 6
DISCUSSION AND FUTURE WORK

In this dissertation, we present techniques and develop new methods for inferring network structure from partially observed graphs. The development of our methods was mainly motivated by the financial difficulty of collecting network data, as well as latent information such as unobserved blocks that researchers cannot possibly collect but plays an important role on inference.

The first two main chapters concern a scenario where ARD alone is used to make inference of individual- and network- level measures, as well as inference of regression coefficient and treatment effect. Traditional network survey collects whether a link exists between every pair of actors, while collecting ARD only requires collecting counts of connections between an actor and a subpopulation, which leads to up to 80% cost reduction. Through theoretical results, a batter of simulations with results on a variety of network measures, and two empirical example, we have shown that our inference of network structure using ARD is similar to that using true network data. We have shown that our method works well on recovering a variety of network features, such as degree, eigenvector centrality, proximity, average path length and maximum eigenvalue. The two empirical examples are reassuring that our method performs well on real networks.

Although the cost-saving advantage of using ARD instead of traditional network is appealing, our method is not without limitations. First, for measures that involve specific paths such as betweenness, our method does not recover those measures very well. This is due to the fact that our method does not recover the actual network, but rather recovers a

distribution of the network. One of the future directions is that, if researchers can collect partial observations of links in addition to ARD, then they can put constraints on simulating distributions of networks from estimated model parameters, so that the distribution of networks recovered has a small variance and may improve the estimation of some path related measures.

Second, our method assumes that the network of interest is consistent with the network formulation model, and ARD is correctly recalled. The other future direction for this work is adjusting recall bias. Researchers may need to collect additional information, such as density of the network, to adjust bias. Additionally, respondents may not have an exact count as the answer. Instead they give a range of numbers as the answer. In that case, an extension of this work on censored data can be considered.

The last chapter concerns a scenario where latent blocks in the network impact inference of regression coefficients. A common challenge in network regression is modeling the dependence structure in errors, due to the fact that relational observations between dyads with a shared actor are likely dependent, after taking into account observable covariates. A number of parametric and non-parametric approaches have been proposed in the literature to address dependency structure in errors. Our proposed block-exchangeable estimator bridges the gap between the two existing non-parametric approaches. We have shown through theory and simulation evidence that when latent block are dependent of covariates, our block-exchangeable estimator outperforms its ancestors.

Future directions of this work include developing a block-exchangeable estimator for binary or count data, or more generally within the generalized linear model framework. Additionally we assume a simple random sample of actors are selected, while respondent-driven sampling is quite common in network data collection, and can be a future direction to benefit practitioners in social fields.

BIBLIOGRAPHY

Jenny C Aker. Information from markets near and far: Mobile phones and agricultural markets in niger. *American Economic Journal: Applied Economics*, 2(3):46–59, 2010.

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Martin M Anthony and Peter Bartlett. Learning in neural networks: theoretical foundations. 1999.

A. Banerjee, E. Breza, E. Duflo, and C. Kinnan. Do credit constraints limit entrepreneurship: Heterogeneity in the returns to microfinance. *Working Paper*, 2016a.

Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1):22–53, 2015.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. Using gossips to spread information: Theory and evidence from two randomized controlled trials. *National Bureau of Economic Research Working Paper*, 2016b.

Lori Beaman, Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak. Can network theory based targeting increase technology adoption? *Working Paper*, 2016.

H Russell Bernard, Eugene C Johnsen, Peter D Killworth, and Scott Robinson. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social science research*, 20(2):109–121, 1991.

H Russell Bernard, Tim Hallett, Alexandrina Iovita, Eugene C Johnsen, Rob Lyerla, Christopher McCarty, Mary Mahy, Matthew J Salganik, Tetiana Saliuk, Otilia Scutelniciuc, et al. Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections*, 86(Suppl 2):ii11–ii15, 2010.

Ian Biringer. *Geometry in Two Dimensions*, 2015. URL https://www2.bc.edu/ian-p-biringer/Geom2Dim.pdf.

Joseph Blitzstein and Persi Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics*, 6(4):489–522, 2011.

Emily Breza and Arun G. Chandrasekhar. Social networks, reputation and commitment: Evidence from a savings monitors experiment. *Econometrica*, 87(1):175–216, 2019.

Emily Breza, Arun G Chandrasekhar, Tyler H McCormick, and Mengjie Pan. Consistent estimation of graph statistics using aggregated relational data.

Emily Breza, Arun G Chandrasekhar, Tyler H McCormick, and Mengjie Pan. Using aggregated relational data to feasibly identify network structure without network data. Technical report, National Bureau of Economic Research, 2017.

J. Cai, A. deJanvry, and E. Sadoulet. Social networks and the decision to insure. *University of Michigan Working Paper*, 2013.

A. Chandrasekhar and R. Lewis. Econometrics of sampled networks. Stanford Working Paper, 2016.

Sylvan Chassang, Pascaline Dupas, Catlan Reardon, and Erik Snowberg. Selective trials for technology evaluation and adoption. *Working Paper*, 2017.

S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *Arxiv preprint arXiv:1102.2650*, 2011.

S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. *Arxiv preprint arXiv:1005.1136*, 2010.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Harry Crane and Walter Dempsey. A framework for statistical network modeling. *arXiv preprint arXiv:1509.08185*, 2015.

Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*, 2007.

Marcel Fafchamps and Flore Gubert. The formation of risk sharing networks. *Journal of development Economics*, 83(2):326–350, 2007.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC Press, 2013.

Bryan S Graham. An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4):1033–1063, 2017.

Peter Guttorp and Richard A Lockhart. Finding the location of a signal: a Bayesian analysis. *Journal of the American Statistical Association*, 83(402):322–330, 1988.

P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:460:1090–1098, 2002.

Peter D Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the american Statistical association*, 100(469):286–295, 2005.

Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169, 2015.

Peter D Hoff et al. Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196, 2011.

Paul W Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

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

Kurt Hornik and Bettina Grün. On conjugate families and Jeffreys priors for von Mises-Fisher distributions. *Journal of Statistical Planning and Inference*, 143(5):992–999, 2013.

David R Hunter. MM algorithms for generalized bradley-terry models. *Annals of Statistics*, pages 384–406, 2004.

Matthew O. Jackson, Tomas R. Rodriguez-Barraquer, and Xu Tan. Social capital and social quilts: Network patterns of favor exchange. *American Economic Review*, 102(5):1857–1897, 2012.

Charles Kadushin, Peter D Killworth, H Russell Bernard, and Andrew A Beveridge. Scale-up methods as applied to estimates of heroin use. *Journal of Drug Issues*, 36(2):417–440, 2006.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

David A Kenny, Deborah A Kashy, and William L Cook. *Dyadic data analysis*. Guilford press, 2006.

Peter D Killworth, Christopher McCarty, H Russell Bernard, Gene Ann Shelley, and Eugene C Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. *Evaluation Review*, 22(2):289–308, 1998.

Dennis V Lindley, Melvin R Novick, et al. The role of exchangeability in inference. *The Annals of Statistics*, 9(1):45–58, 1981.

László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

Markus Maier, Matthias Hein, and Ulrike Von Luxburg. Cluster identification in nearest-neighbor graphs. In *International Conference on Algorithmic Learning Theory*, pages 196–210. Springer, 2007.

K V Mardia and S A M El-Atoum. Bayesian inference for the von Mises-Fisher distribution. *Biometrika*, 63(1):203–206, 1976.

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

Tyler H McCormick and Tian Zheng. Latent surface models for networks using aggregated relational data. *Journal of the American Statistical Association*, 110(512):1684–1695, 2015.

Peter McCullagh. Exchangeability and regression models. 2005.

Lynn C Miller and David A Kenny. Reciprocity of self-disclosure at the individual and dyadic levels: A social relations analysis. *Journal of Personality and Social Psychology*, 50 (4):713, 1986.

Peter Orbanz and Daniel M Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37 (2):437–461, 2015.

Mengjie Pan, Tyler H. McCormickm, and Bailey Fosdick. Block exchangeable standard errors for network regression.

M Papadakis, M Tsagris, M Dimitriadis, I Tsamardinos, M Fasiolo, G Bor-boudakis, and J Burkardt. Rfast: Fast r functions. *R package version*, 1(5), 2017.

Juyong Park and Mark EJ Newman. Statistical mechanics of networks. *Physical Review E*, 70(6):066117, 2004.

Jouni Parkkonen. Hyperbolic geometry, 2012. URL http://users.jyu.fi/~parkkone/RG2012/HypGeom.pdf.

J Paupert. Introduction to hyperbolic geometry, 2016. URL https://math.la.asu.edu/~paupert/HyperbolicGeometryNotes.pdf.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Cosma Rohilla Shalizi and Dena Asta. Consistency of maximum likelihood for continuous-space network models. *arXiv preprint arXiv:1711.02123*, 2017.

Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic

blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

Julian Straub, Trevor Campbell, Jonathan P How, and John W Fisher. Small-variance nonparametric clustering on the hypersphere. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 334–342, 2015.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Michael D Ward and Peter D Hoff. Persistent patterns of international commerce. *Journal of Peace Research*, 44(2):157–175, 2007.

A. T. A. Wood. Simulation of the von mises fisher distribution. *Communications in Statistics, Simulation and Computation*, 23:157–64, 1994.

Appendix A

APPENDIX FOR CHAPTER 5

A.1 Definitions of notations

Q_M is defined as:

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

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

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

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

- $\Phi_{\sigma^2,i} = \{[(i,j),(i,j)] : j \in [n], i \neq j\} \cup \{[(j,i),(j,i)] : j = 1, ..., n, i \neq j\}$
- $\Phi_{\phi_A,i} = \{[(i,j),(j,i)] : j \in [n], i \neq j\}$
- $\Phi_{\phi_B,i} = \{[(i,j),(i,k)] : j \in [n], k \in [n], i \neq j \neq k\}$
- $\Phi_{\phi_C,i} = \{[(j,i),(k,i)] : j \in [n], k \in [n], i \neq j \neq k\}$
- $\Phi_{\phi_D,i} = \{[(i,j),(k,i)] : j \in [n], k \in [n], i \neq j \neq k\}$

A.2 Evaluating Block Membership Estimation

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

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

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

A.2.1 Simulation Parameters

Based on the error generating model in Section 5.5.1, we set parameters as follows:

- $[\sigma_{a,1} \;\; \sigma_{a,2}] = [\sqrt{2}\alpha_1 \;\; \sqrt{2}r\alpha_1]$
- $[\sigma_{b,1} \;\; \sigma_{b,2}] = [\alpha_1 \;\; r\alpha_1]$
- $[\sigma_{z,1} \;\; \sigma_{z,2}] = [\alpha_1 \;\; r\alpha_1]$
- $[\sigma_{\gamma,\{1,1\}} \;\; \sigma_{\gamma,\{1,2\}} \;\; \sigma_{\gamma,\{2,2\}}] = [\alpha_1 \;\; \sqrt{r}\alpha_1 \;\; r\alpha_1]$

- $\sigma_\epsilon = \alpha_1$, $\rho = 0.5$, and $d = 2$.

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

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

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

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

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

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

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

- $\mathrm{Var}(\xi_{ij})|g_i = 1, g_j = 1 > \mathrm{Var}(\xi_{ij})|g_i = 1, g_j = 2 > var(\xi_{ij})|g_i = 2, g_j = 1 > \mathrm{Var}(\xi_{ij})|g_i = 2, g_j = 2$
- $\mathrm{Cov}(\xi_{ij}, \xi_{ji})|g_i = 1, g_j = 1 > \mathrm{Cov}(\xi_{ij}, \xi_{ji})|g_i = 1, g_j = 2 = \mathrm{Cov}(\xi_{ij}, \xi_{ji})|g_i = 2, g_j = 1 > \mathrm{Cov}(\xi_{ij}, \xi_{ji})|g_i = 2, g_j = 2$
- $\mathrm{Cov}(\xi_{ij}, \xi_{il})|g_i = 1 > \mathrm{Cov}(\xi_{ij}, \xi_{il})|g_i = 2$
- $\mathrm{Cov}(\xi_{ij}, \xi_{kj})|g_j = 1 > \mathrm{Cov}(\xi_{ij}, \xi_{kj})|g_j = 2$
- $\mathrm{Cov}(\xi_{ij}, \xi_{ki})|g_i = 1 > \mathrm{Cov}(\xi_{ij}, \xi_{ki})|g_i = 2$

A.2.2 Simulation Results

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

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

Figure A.2 shows the distribution of KS statistic $KS_{i,j,M}$ calculated in Algorithm 2 Step 3 on each of the five cases at different values of r. Each column represents one of the five cases $M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, and each row represents a given r value. The red curves represent

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

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

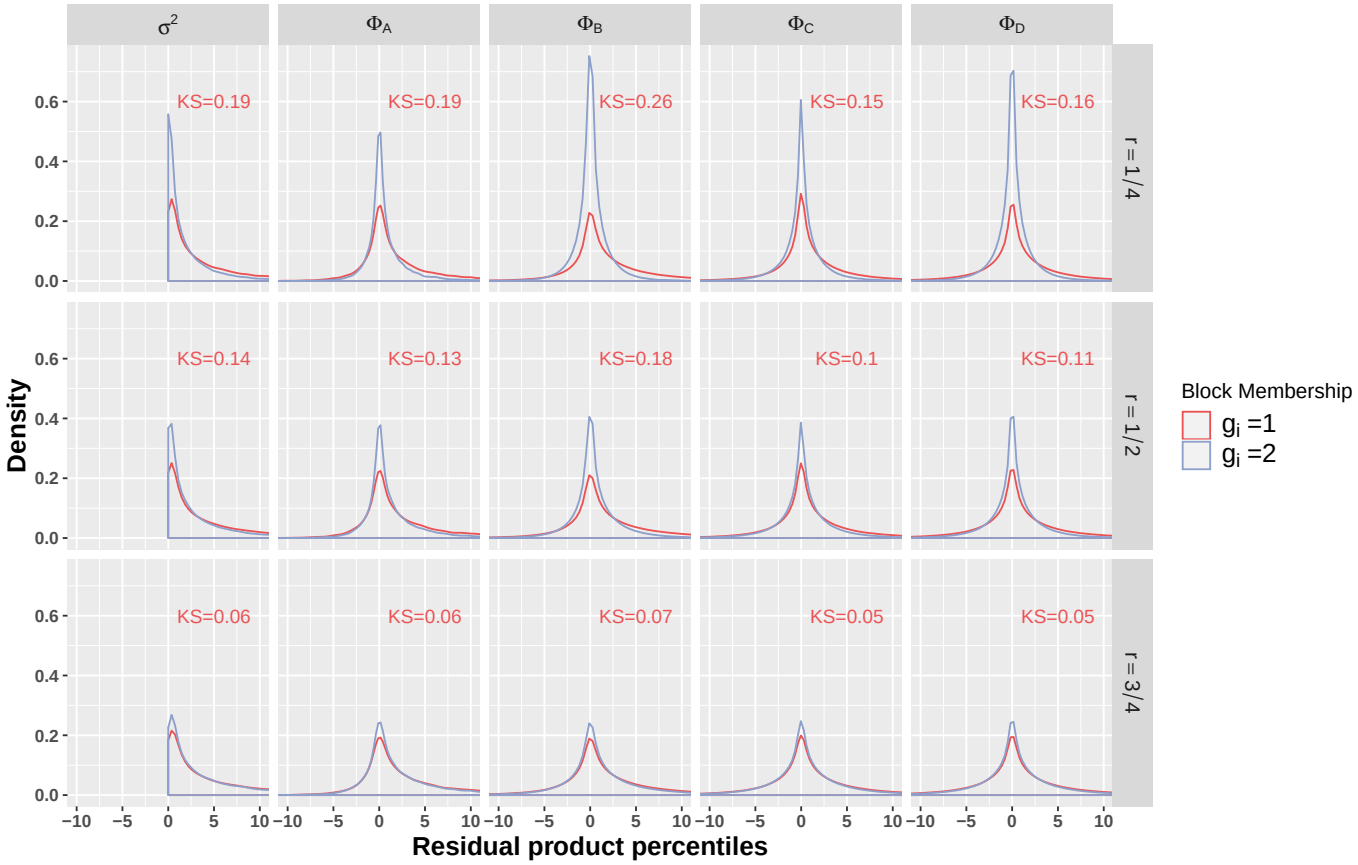


Figure A.1: Residual products of five dyads from 10 simulation of $n = 80$. Each column represents one of the five cases $M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, and each row represents a given r value. The red and blue curves represent the distribution in Block 1 and Block 2, respectively.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

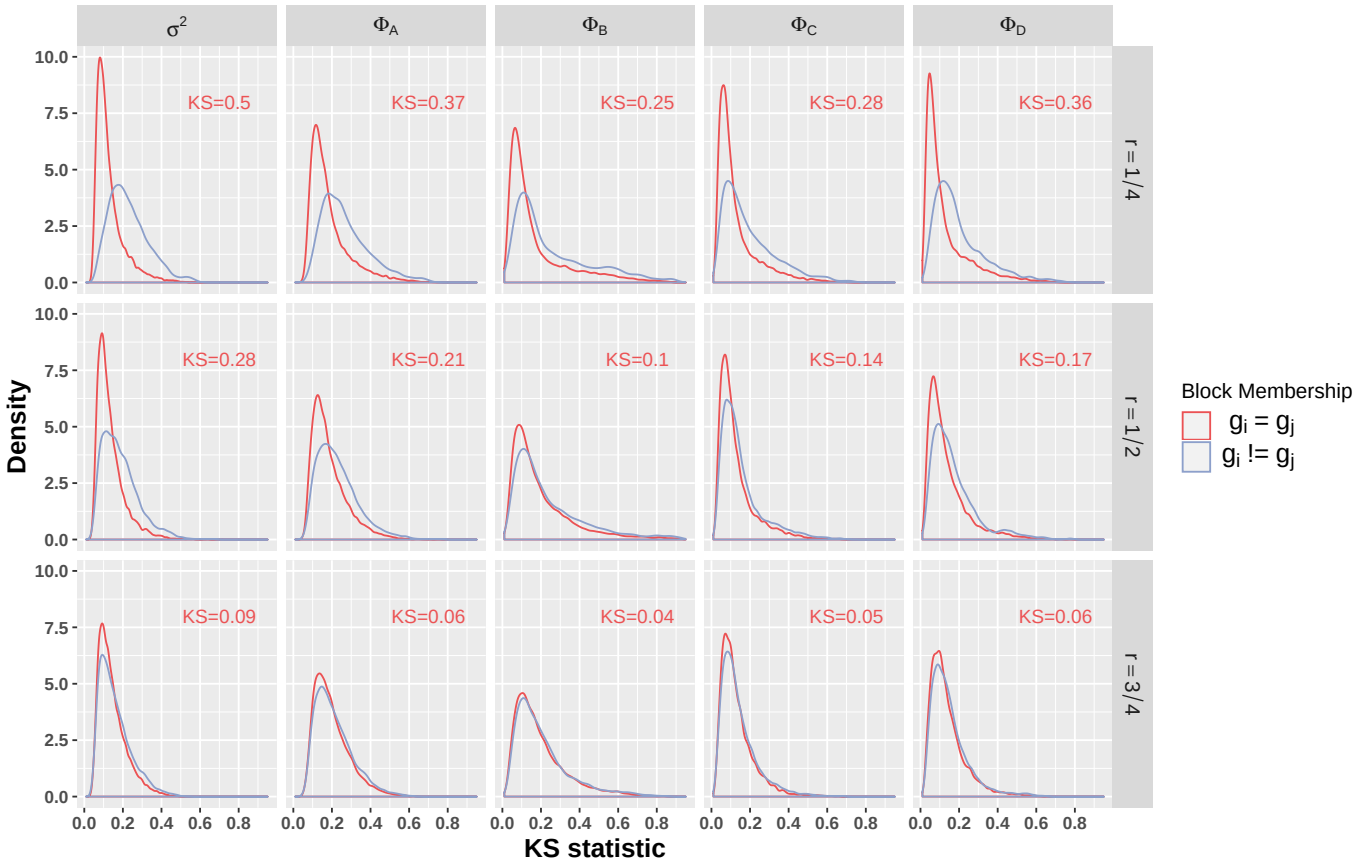


Figure A.2: Distribution of KS statistic between residual products of five dyads from 10 simulation of $n = 80$. Each column represents one of the five cases $M \in \{\sigma^2, \phi_A, \phi_B, \phi_C, \phi_D\}$, and each row represents a given r value. The red curves represent the distribution where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

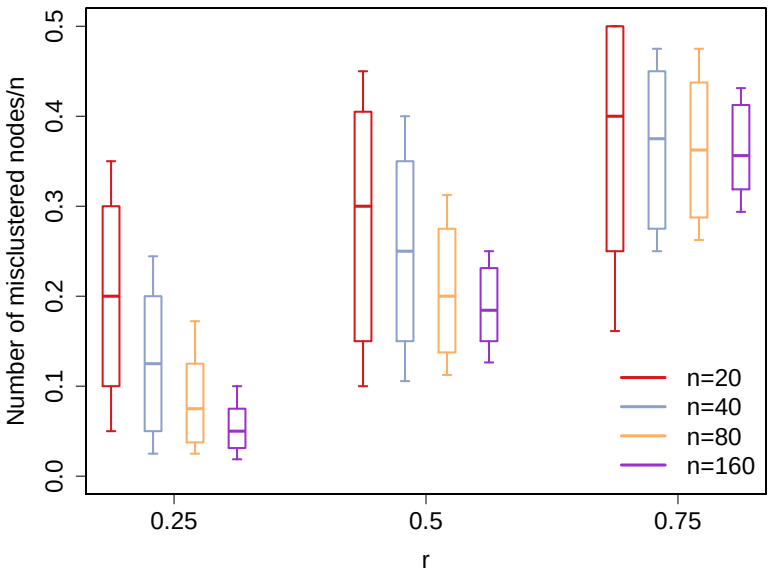


Figure A.3: Number of misclustered nodes over n at different r.