©Copyright 2024
Steven Wilkins-Reeves

Statistical Inference with Missing and Latent Data: Methods for Data Harmonization, Network Curvature Estimation and Experimentation Under Interference

Steven Wilkins-Reeves

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

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Tyler McCormick, Chair
Yen-Chi Chen, Chair
Carlos Cinelli

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Statistical Inference with Missing and Latent Data: Methods for Data Harmonization, Network Curvature Estimation and Experimentation Under Interference

Steven Wilkins-Reeves

Co-Chairs of the Supervisory Committee:
Tyler McCormick
Department of Statistics and Sociology

Yen-Chi Chen
Department of Statistics

This dissertation explores several statistical challenges involving inference problems where the object of interest is a latent phenomenon or involves missing data. Effective modeling of the latent processes or missing data is crucial for accurate inference in such scenarios. We delve into issues of missing and latent data across three distinct settings. The first project addresses missing outcomes resulting from changes in neuropsychological test battery versions, where each version represents different testing models and scales. The second project focuses on inference for causal parameters using partially measured network data, also highlighting the experimental design challenges associated with such problems. The final project presents a nonparametric method for estimating network curvature from distance matrices. This approach emphasizes network models and introduces tests for constant curvature, providing a clearer understanding of the underlying network structure.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank everyone who has supported me in reaching this point. First and foremost, I want to express my gratitude to my advisors, Yen-Chi and Tyler. Yen-Chi, thank you for helping me find joy in studying statistics across various fields. Tyler, thank you for helping me see the bigger picture in research. I am also grateful to Gary Chan and Arun Chandrasekhar for being excellent collaborators and for helping me become a better scientist. Additionally, I would like to acknowledge the numerous group members in both Tyler's and Yen-Chi's research groups over the years, especially Shane Lubold, who I worked with on one of the chapters in this work.

I would also like to thank all the friends I have made during my time here at the University of Washington who have helped me enjoy my time outside of work in Seattle.

Lastly, I want to thank my wonderfully supportive partner, Mary O'Sullivan, as well as my parents, for their unwavering support during stressful times.

DEDICATION

To Mary, my parents John and Kathryn, and my brother Dean.

Chapter 1

INTRODUCTION

This dissertation explores the challenges and strategies of working with incomplete or partially observable data, a common scenario in many real-world applications. The data that researchers aspire to collect and the data actually available often diverge significantly due to limitations in data collection processes. These datasets frequently contain missing values and require careful consideration of the assumptions used to manage this missingness. Additionally, many studies aim to analyze properties that are not directly observable but are latent in nature. Such structures are prevalent in models like item response theory and latent variable models [McGrory et al., 2014] designed to describe specific network structures [Hoff et al., 2002c].

Chapter 2 introduces a framework for data harmonization, which seeks to equate different measurements within a common domain. Motivated by dementia research, this chapter presents a nonparametric latent trait model for harmonizing various neuropsychological tests that measure similar cognitive abilities, such as memory or attention. We describe the development of a unique regularized maximum likelihood estimator, demonstrate the convergence of a nonparametric EM algorithm to this estimator, and highlight its computational advantages over traditional methods. This chapter also addresses model selection and goodness-of-fit assessments—often overlooked in mixing distribution estimation—and introduces uncertainty-quantified score conversion techniques. We introduce a nonparametric EM algorithm [Laird, 1978, Lindsay, 1995] and prove the weak convergence of the nonparametric EM algorithm to the maximizer. These methodologies are applied to the National Alzheimer's Coordination Center Uniform Data Set, showing superior performance over standard dementia research techniques.

Chapter 3 discusses how to handle network interference when the network facilitating this interference is unknown. It challenges the stable unit treatment value assumption by acknowledging that treatments can affect individuals beyond the direct recipients. This chapter explores the use of partially or indirectly observed network data, such as subsamples and aggregated relational data, to mitigate the high costs and logistical challenges of complete network data collection. We propose a structural causal model framework for estimating and adjusting treatment effects with partial network data and detail experimental design and analysis enhancements for these contexts. We extend upon the framework for missing data inference of Chandrasekhar and Lewis [2011] to the case of single network asymptotics, by leveraging a dependent data central limit theorem result [Chandrasekhar et al., 2023] . Our methodology is validated through simulated experiments on actual networks, with applications to information diffusion in India and Malawi.

Chapter 4 investigates the estimation of curvature in latent space models of network data, which often depict nodes and edges representing complex, high-dimensional structures. By embedding these graphs in low-dimensional geometric spaces, we can infer network characteristics like the tendency to form triangles or exhibit tree-like structures. We take an alternative view of the problem of [Lubold et al., 2023], where we do not test relative to a fixed set of constant curvatures but take a local approach to the problem, this modularity also is useful for applications beyond the network setting. This chapter develops an estimating function for curvature based on triangle side lengths and midpoints, relying only on a distance matrix. We also introduce a novel latent distance matrix estimator and an efficient algorithm for computing these estimates, applying the method to detect security threats in network data from Los Alamos National Laboratory and to uncover non-constant latent curvature in physics co-authorship networks.

Chapter 5 offers concluding remarks and discusses potential avenues for future research.

Chapter 2

DATA HARMONIZATION

2.1 Introduction

Data harmonization is the process by which measurements from different sources and methods are combined into a single variable in a dataset for further analysis. In our application, it often involves converting scores between two different measurements of the same domain. For example, in neuropsychological testing for Alzheimer's Disease, collections of tests, known as batteries, are used to measure various cognitive domains such as attention, language, episodic memory, and visual-spatial ability, among others. Different studies, or even in the same study, may use different validated instruments to measure a certain domain. As an increasing number of data sets are becoming available for public use, data harmonization efforts are common in order to create harmonized variables in the combined data set.

This work is motivated by the different versions of cognitive measurements in the Neuropsychological Test Battery of the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS). The NACC UDS has the largest number of participants for studying the progression to mild cognitive impairment and dementia in the United States [Weintraub et al., 2009, Besser et al., 2018]. The batteries used for neuropsychological assessment of cognition change over time. Since 2015, the third version of the Uniform Data Set utilized a non-proprietary neuropsychological battery while prior to that time, a different set of proprietary neuropsychological batteries was given. Additionally, for comparison across tests, a small group of cognitively normal individuals received both the old and new battery in a crosswalk study [Monsell et al., 2016], which will serve as a validation sample of various harmonization methods.

Our primary contributions are two fold. The first is appropriate uncertainty quantifica-

tion for the task of score conversion using a semiparametric model which incorporates the inherent measurement uncertainty present in item response models and a flexible distribution of latent traits. In this process, we introduce a regularization method and an algorithm for estimating the latent trait distribution. We also include goodness-of-fit tests for the model, and finally, discuss a bootstrap method for converting scores with multiple imputations for downstream statistical tasks involving multiple score versions. Secondly, this uncertainty quantification allows for the study of long-term cognitive outcomes. Due to the version changes of the diagnostic tools, long-term studies become infeasible in the NACC dataset. For instance, there are no individuals in the NACC dataset who have the same test version over an 8-10 year span. Our methods allow for appropriate uncertainty quantification of the downstream analyses of harmonized scores, which is not possible with deterministic score conversions alone.

Our modeling choices for uncertainty quantification are motivated by the structure of the NACC datasets. Data harmonization requires two main steps. The first of which is to determine whether two scores $Y \in \{0, 1, 2, \ldots, N_Y\}$ and $Z \in \{0, 1, 2, \ldots, N_Z\}$ measure the same cognitive domain. That is, whether their distributions can be commonly represented by a trait pair γ, ζ for which there is monotone mapping between the two. This first step may require expert knowledge or studies of individuals completing both tests. In our case, the change in battery and their correspondence in measuring cognitive domains were guided by the Clinical Task Force (CTF), a group formed by the National Institute on Aging to develop standardized methods for collecting longitudinal data that would encourage and support collaboration across the Alzheimer's Disease Research Centers. The second step is to propose a joint model of the scores, frequently using a latent variable representation [Griffith et al., 2013, van den Heuvel et al., 2020]. These methods, however, tend to primarily rely on fully parametric assumptions [Griffith et al., 2013] and develop task-specific joint models for studying trends in variables over time. We instead propose a flexible semiparametric framework that can be used for either prediction or imputation of the missing scores from the observed ones, which allows a practitioner to study associations on a single scale of

interest rather than develop task-specific joint models.

Our approach is connected to mixing distribution estimation, for which there is a rich history. One of the earliest versions of the problem arose from educational testing where scores for each individual are assumed to follow a binomial distribution with parameters N, p_i [Lord, 1965, 1969], with p_i following a particular distribution. The binomial case is a particularly well-studied application of mixing distribution estimation [Lindsay, 1983a, Wood, 1999, Tian et al., 2017, Vinayak et al., 2019]. Estimation of mixing distribution, in general, has numerous other applications such as in positron emission tomography [Vardi et al., 1985, Silverman et al., 1990], portfolio optimization [Cover, 1984] and ecology [Bell et al., 2000]. In cases where the outcome is continuous, de-convolution kernels are a common approach [Leonard and Carroll, 1990, Basulto-Elias et al., 2021] as is similar in structure to measurement error problems.

The nonparametric EM algorithm [Laird, 1978] is a classical approach to compute the nonparametric maximum likelihood estimator (NPMLE) [Kiefer and Wolfowitz, 1956]. Laird initially assumed that the maximum likelihood estimator should be a discrete distribution. This was formalized by Lindsay [1983b], who proved that under some very mild conditions, there always exists a discrete maximum likelihood estimator of the mixing distribution, even if the underlying mixing distribution is continuous. This phenomenon is further studied by Lindsay [1983a,b], which describes the set identifiability of these models in general. Recently, in special cases such as the binomial model, de-convolution rates have been explored for general latent distributions [Tian et al., 2017, Vinayak et al., 2019]. When the goal is data harmonization, a discrete estimated mixture distribution may be undesirable since, in data harmonization, it is often assumed that a continuous latent map exists between latent traits, and a discrete latent distribution would violate this assumption [van den Heuvel et al., 2020]. We will address this problem with a regularization approach and illustrate the convergence properties of a nonparametric EM algorithm.

The remainder of the paper will outline our statistical approach to this data harmonization problem. First, we introduce a graphical representation of our problem. We then

discuss nonparametric estimation of the latent trait distributions. We further focus on the aspects of picking the conditional distribution (measurement error) assumption required for modeling and testing the assumption's feasibility. We illustrate how this model can be used for prediction and inference tasks and highlight how this method will automatically handle missing-at-random outcomes, unlike current methods. Code available for replicating the analysis is available at https://github.com/SteveJWR/dataharmonize and a corresponding `R` package is available at https://github.com/SteveJWR/dnoiseR. We refer to our method using the abbreviation DNOISe referring to **D**ata Harmonization by **No**nparametric **I**mputation of **S**cores in the applications and simulations.

2.2 Definitions and structural assumptions

Consider a set of *i.i.d.* random variables (X_i, Y_i, Z_i, V_i), where $X_i \in \mathbb{R}^d$ is a set of covariates and $Y_i \in \{0, \dots, N_Y\}$ is the version 1 score (old test) and $Z_i \in \{0, \dots, N_Z\}$ is the version 2 score (new test). We drop the subscript on N if it is clear from the context. We focus on discrete test scores as the neuropsychological tests considered in the application are all discrete. Lastly, $V_i \in \{0, 1\}$ is an indicator variable denoting which score is observed. In this setting, only one score is observed for each individual; denote $V_i = 0$ if score Y_i is observed and $V_i = 1$ if score Z_i is observed. Our motivation comes from heterogeneous measurements of a common trait. In our application, we think of this as two test scores measuring the same cognitive domain, such as memory or attention. As is common in item response theory, we assume that our observed outcomes (scores) (Y_i, Z_i) are noisy measurements of some underlying traits (γ_i, ζ_i). The test-specific traits are not identical as the tests may have different difficulty is thus measuring the common domain using different scales. However, since they are measuring the same domain,we assume a monotone mapping between the conditional quantiles of one latent trait to another. Explicitly,

$$\begin{aligned} \phi_{\gamma \to \zeta}(q|x) &= G^{-1}\left(F(q|x)|x\right) \ , \\ \phi_{\zeta \to \gamma}(q|x) &= F^{-1}\left(G(q|x)|x\right) = \phi^{-1}_{\gamma \to \zeta}(q|x) \end{aligned} \tag{2.1}$$

where $F(\cdot|x)$ and $G(\cdot|x)$ are the conditional CDF's of γ and ζ. Note that equation (2.1) is related to the Skorohod representation in quantile regression models [Chernozhukov and Hansen, 2006].

More explicitly, we assume the following hierarchical model for data harmonization:

$$\begin{aligned} \Omega_i &\sim U(0,1), & X_i &\sim P_X, \\ \gamma_i &= F^{-1}(\Omega_i|X_i), & \zeta_i &= G^{-1}(\Omega_i|X_i), \\ Y_i|\gamma_i &\sim P_{A_Y}(\cdot|\gamma_i), & Z_i|\zeta_i &\sim P_{A_Z}(\cdot|\zeta_i), \\ V_i &\sim P_{V|X} \end{aligned} \tag{2.2}$$

where $U(0,1)$ denotes the uniform distribution on $[0,1]$, Ω_i is the relative rank, P_X represents the distribution of the covariates, $F^{-1}(\cdot|x), G^{-1}(\cdot|x)$ represent the inverse of the conditional cdfs of γ and ζ respectively, and P_{A_Y} (or P_{A_Z}) represents the corresponding distribution of the observed score, given the latent trait (i.e., the measurement error). Additionally, we restrict the scale of the latent traits γ, ζ to be on the unit interval $[0,1]$. This model automatically encodes a missing-at-random assumption [Rubin, 1976], which states

$$V \perp\!\!\!\perp (\Omega, Y, Z)|X. \tag{2.3}$$

We do not include a dependence between Ω and V as this would represent a selection bias where only either high or low rank individuals within a certain covariate group X are observed. One can easily see that the model in equation (2.2) implies the conditions in equation (2.1). The model in equation (2.2) is similar to rank-preserving models though we do not assume a constant additive effect [Robins and Tsiatis, 1991, White et al., 1997, 1999].

We summarize the generative process of this hierarchical model using Figure 2.1, which is not a conventional of partially directed acyclic graphs in causal inference. The variables denoted with squares are deterministic functions of their parent variables in the graph. The generative model in equation (2.2) can be interpreted as follows. Ω is the relative rank of an individual's cognitive ability within a domain in the population, thus defines an ordering of the trait for individuals with the same covariate. Therefore, Ω is defined to be independent

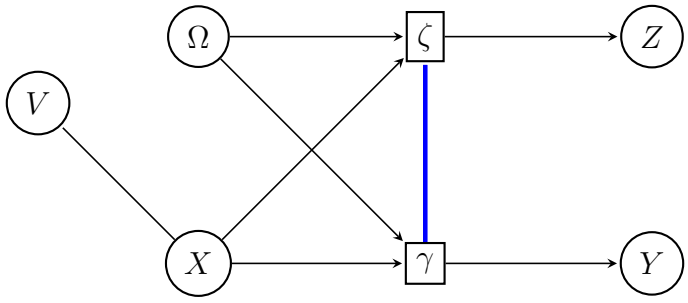


Figure 2.1: Score Generation Model. Deterministic relationships between latent trait are shown in blue. The trait variables which are a deterministic function of their parent variables are indicated with squares.

of X. Thus, an individual's rank Ω_i together with their covariates X_i determine the two test specific latent traits γ_i, ζ_i. The observed scores Y and Z measurements of the latent traits, therefore have a conditional distribution that depends only on their latent traits γ and ζ, respectively. The conditional independence (2.3) is motivated by the fact that the score conversion in the NACC data occurred in 2015, which is not relevant to any individual's ability but it may be relevant to the year of visit/age, which is part of the covariate X).

Although we only observe either $(X_i, Y_i, V_i = 0)$ or $(X_i, Z_i, V_i = 1)$, the generative model in equation (2.2) implies that the observed data is determined by $p(y|\gamma), p(z|\zeta), p(\gamma|x), p(\zeta|x)$, $p(v|x)$ and $p(x)$. For data harmonization, we do not need to model $p(v|x)$ or $p(x)$, so we will focus on modeling the first four distributions. The first two distributions $p(y|\gamma), p(z|\zeta)$ are the measurement assumptions, so we denote them as $p_A(y|\gamma), p_A(z|\zeta)$. The two distributions $p(\gamma|x), p(\zeta|x)$ are the latent trait models, so we denote them as $p_M(\gamma|x), p_M(\zeta|x)$. In what follows, we describe how we model p_A and p_M. To simplify the notations, we focus on $p_A(y|\gamma)$ and $p_M(\gamma|x)$; the case of $p_A(z|\zeta)$ and $p_M(\zeta|x)$ can be modeled in a similar manner. In the following section 2.3, we will discuss the estimation of the p_M model, then follow this up by the model selection of the measurement assumption model p_A in section 2.4.

Remark. We contrast our model with the approach of [van den Heuvel et al., 2020],

in which they assume that the distribution of $Y|\gamma, X$ may depend on X but the mapping between latent traits $\phi_{\gamma\to\zeta}$ is independent of X. Our measurement assumption is similar to [Meredith, 1993] in which the observed score only depends on the latent trait $p(y|\gamma, x) = p(y|\gamma)$ and similarly for Z, corresponding to non-differential measurement error assumptions, but allow the distribution of latent traits and the mapping between them to depend on X.

Remark. We will focus our study of this model on problems that involve study on a single outcome for each test version. This will include predicting the score on the new scale and inference on the new scale. When the latent traits are unknown, this procedure can also be interpreted as an empirical Bayes problem. One can consider the analogs as selecting the measurement assumption as choosing f modelling, whereas estimating the latent trait is g modeling [Efron, 2019]. We next discuss model selection for the measurement model in Section 4 and estimation of the latent trait in Section 3.

2.3 Estimation of Latent Trait Distribution

In the following Section 2.3.1, we focus on the estimation of the latent trait distribution, i.e. we first assume that the measurement model $p_A(y|\gamma)$ is given. Later in Section 2.4.3, we discuss how to empirically validate the measurement model.

2.3.1 Nonparametric latent trait estimation

In many applications, parametric models for a latent distribution may be considered restrictive. We want to relax such assumptions and consider a nonparametric method, and we propose a method of estimating this model via the nonparametric EM algorithm [Laird, 1978]. To simplify the derivation, we omit the covariate X, but later in Section 2.3.5, we will include a simple, yet general, technique for incorporating covariates. Estimating the latent model $P_M(\gamma)$ (with density $p_M(\gamma)$) now clearly becomes a mixing distribution problem with log-likelihood:

$$\ell_n[P_M] = \frac{1}{n}\sum_{i=1}^{n} \log\left(\int p_A(y = Y_i|\gamma)p_M(\gamma)d\gamma\right), \quad P_M \in \mathcal{P}_{[0,1]} \tag{2.4}$$

with $\mathcal{P}_{[0,1]}$ denotes the space of all probability measures over $[0,1]$. We use the following shorthand notation to denote distributional quantities derived from the model: $p_{MA}(y,\gamma) = p_A(y|\gamma)p_M(\gamma), p_{MA}(\gamma|y) = \frac{p_{MA}(y,\gamma)}{p_{MA}(y)}$ and $p_{MA}(y) = \int_0^1 p_{MA}(y,\gamma)d\gamma$. We also highlight that P_M refers to the measure with a corresponding generalized density function p_M for which we may allow point masses. This problem involved solving the NPMLE [Kiefer and Wolfowitz, 1956] and was the original inspiration for the nonparametric EM algorithm (NPEM, also known as the functional EM) [Laird, 1978].

With a basic rearrangement, we can rewrite equation (2.4), the mixture likelihood, to the following:

$$\ell_n[P_M] = \sum_{y=0}^{N} \widehat{p}(y) \log \left(\int p_A(y|\gamma)p_M(\gamma)d\gamma \right) \tag{2.5}$$

where $\widehat{p}(y)$ is the empirical distribution of y

$$\widehat{p}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = y). \tag{2.6}$$

Naturally, one can derive the standard nonparametric EM algorithm [Laird, 1978]

$$p_M^{(j+1)}(\gamma) = \mathbb{E}_{\widehat{p}(y)}[p_M^{(j)}(\gamma|y)] = \sum_{y=0}^{N} \frac{p_A(y|\gamma)\widehat{p}(y)p^{(j)}(\gamma)}{p_{MA}^{(j)}(y)} \tag{2.7}$$

where $p_{MA}^{(j)}(y) = \int p_A(y|\gamma)p_M^{(j)}(\gamma)d\gamma$.

However a problem exists in that the NPMLE (maximizer of equation (2.5)) may not be unique. Because of this, the EM algorithm will not necessarily converge to a fixed value. This problem is particularly serious when we want to use the NPMLE to perform a score conversion. In this next section 2.3.2, we address this problem through regularization.

Remark. [Lindsay, 1983b] studied the NPMLE in-depth and proved that if Y has a support size of $N+1$, then there will always exist a discrete NPMLE with at most $N+1$ masses, even when the true latent model is continuous. This phenomenon is explained by the non-identifiability of this model, as the likelihood functional $\ell_n[P_M]$ is only concave in P_M, but not strictly concave. Namely, there might be multiple distributions that maximize

$\ell_n[P_M]$. Due to the non-identifiability of a maximizer, the convergence properties of the nonparametric EM algorithm are far less studied than in the parametric case. As far as we know, Chung and Lindsay [2015] is the only work that proved the convergence of nonparametric EM algorithm in terms of likelihood values. Under the conditions of Chung and Lindsay [2015], the update in equation (2.7) leads to an improved estimator of P_M in the sense that the implied marginal distribution of Y, $p_{MA}(y)$, is closer to the empirical distribution in relative entropy after each iteration.

2.3.2 Regularized nonparametric latent trait estimation

To deal with the non-identifiability issue, we introduce a regularized likelihood and the corresponding NPEM algorithm leading to a unique maximizer. Let P_U be the measure of a uniform random variable over [0,1]. We introduce a regularization term via the KL-divergence from the uniform distribution, i.e., $\mathcal{D}(P_U||P_M)$. Though one can replace P_U with any other reference distribution, the uniform distribution is convenient in practice. One reason for regularizing to a uniform density is due to its reasonable choice as an uninformative latent distribution, much like the choice of regularizing coefficients to 0 in the LASSO, though, in principle, any distribution can be selected for regularization, analogous to the possibility to shrink to any value in the parameter space in the LASSO [Tibshirani, 1996].

We define the regularized likelihood as

$$\ell_{n,\mu}[P_M] = \ell_n[P_M] - \mu\mathcal{D}(P_U||P_M).$$

Due to the strict convexity of $\mathcal{D}(P_U||P_M)$ in each parameter, the regularized objective has a unique maximizer whenever $\mu > 0$. Additionally, since any discrete latent distribution P_M will have infinite KL-divergence from the uniform $\mathcal{D}(P_U||P_M)$, regularizing toward a uniform distribution will smooth the estimate to a continuous distribution. Moreover, this choice of regularization penalty leads to computational convenience provide an EM algorithm to estimate the regularized NPMLE (the maximizer of $\ell_{n,\mu}$).

We introduce a regularized NPEM algorithm involving an update step that is a mixture

between the unregularized NPEM algorithm and the uniform distribution. First we expand the expression for $\ell_{n,\mu}$ as

$$\begin{aligned}\ell_{n,\mu}[P_M] &= \sum_{y=0}^{N} \widehat{p}(y) \log\left(\int p_A(y|\gamma) p_M(\gamma) d\gamma\right) - \mu \mathcal{D}(P_U||P_M) \\ &= \sum_{y=0}^{N} \widehat{p}(y) \left(\log\left(p_{MA}(y,\gamma)\right) - \log\left(p_{MA}(\gamma|y)\right)\right) - \mu \mathcal{D}(P_U||P_M)\end{aligned}$$

Since $p_{MA}(y) = \frac{p_{MA}(y,\gamma)}{p_{MA}(\gamma|y)}$ is not dependent on γ, $\log(p_{MA}(y)) = \log(p_{MA}(y,\gamma)) - \log(p_{MA}(\gamma|y))$ is a constant function of γ, and we can integrate over any probability density, namely $p_{MA}^{(j)}(\gamma|y)$:

$$\ell_{n,\mu}[P_M] = \sum_{y=0}^{N} \widehat{p}(y) \int \left(p_{MA}^{(j)}(\gamma|y) \left(\log\left(p_{MA}(y,\gamma)\right) - \log\left(p_{MA}(\gamma|y)\right)\right) d\gamma - \mu \mathcal{D}(P_U||P_M)\right) .$$

Therefore,

$$\ell_{n,\mu}[P_M] = Q[P_M||P_M^{(j)}] + H[P_M||P_M^{(j)}] - \mu \mathcal{D}(P_U||P_M),$$

where

$$\begin{aligned}Q[P_M||P_M^{(j)}] &= \sum_{y=0}^{N} \widehat{p}(y) \int p_{MA}^{(j)}(\gamma|y) \log\left(p_{MA}(y,\gamma)\right) d\gamma, \\ H[P_M||P_M^{(j)}] &= -\sum_{y=0}^{N} \widehat{p}(y) \int p_{MA}^{(j)}(\gamma|y) \log\left(p_{MA}(\gamma|y)\right) d\gamma\end{aligned}$$

By Gibbs' Inequality $H[P_M||P_M^{(j)}] \geq H[P_M^{(j)}||P_M^{(j)}]$ with equality only if $P_M = P_M^{(j)}$. Therefore maximizing $Q[P_M||P_M^{(j)}] - \mu \mathcal{D}(P_U||P_M)$ will automatically increase in the likelihood with each iteration since

$$\begin{aligned}\ell_{n,\mu}[P_M^{(j+1)}] - \ell_{n,\mu}[P_M^{(j)}] &= \underbrace{\left(H[P_M||P_M^{(j)}] - H[P_M^{(j)}||P_M^{(j)}]\right)}_{\geq 0} \\ &+ \underbrace{\left(Q[P_M^{(j+1)}||P_M^{(j)}] - \mu \mathcal{D}(P_U||P_M^{(j+1)})\right) - \left(Q[P_M^{(j)}||P_M^{(j)}] - \mu \mathcal{D}(P_U||P_M^{(j)})\right)}_{\geq 0}.\end{aligned}$$

The EM algorithm update is the solution to the following problem

$$P_M^{(j+1)} = \arg\max_{P_M \in \mathcal{P}_{[0,1]}} Q[P_M||P_M^{(j)}] - \mu\mathcal{D}(P_U||P_M).$$

In order to compute $P_M^{(j+1)}$ directly, observe that

$$\begin{aligned}
&Q[P_M||P_M^{(j)}] - \mu\mathcal{D}(P_U||P_M) \\
&= \int \sum_{y=0}^{N} \widehat{p}(y) \frac{p_A(y|\gamma)p_M^{(j)}(\gamma)}{p_{MA}^{(j)}(y)} \log(p_A(y|\gamma)p_M(\gamma))d\gamma + \mu \int \log(p_M(\gamma))d\gamma \\
&= (1+\mu) \int \left(\sum_{y=0}^{N} \widehat{p}(y) \frac{p_A(y|\gamma)p_M^{(j)}(\gamma)}{(1+\mu)p_{MA}^{(j)}(y)} + \frac{\mu}{1+\mu} \right) \log(p_M(\gamma))d\gamma + C.
\end{aligned}$$

Once again, by Gibbs' inequality, we can uniquely maximize $Q[P_M||P_M^{(j)}] - \mu\mathcal{D}(P_U||P_M)$ by setting

$$p_M^{(j+1)}(\gamma) = \frac{1}{1+\mu}\mathbb{E}_{\widehat{p}(y)}[p_M^{(j)}(\gamma|y)] + \frac{\mu}{1+\mu} = \sum_{y=0}^{N} \frac{p_A(y|\gamma)\widehat{p}(y)p_M^{(j)}(\gamma)}{(1+\mu)p_{MA}^{(j)}(y)} + \frac{\mu}{1+\mu}. \tag{2.8}$$

When $\mu = 0$, this reduces to the nonparametric EM algorithm in equation (2.7). We can summarize the important convergence properties of the algorithm in Theorem 2.3.1

Theorem 2.3.1. *Denote the unique global solution* $P_{M,\mu}^* = \arg\max_{P_M \in \mathcal{P}_{[0,1]}} \ell_{n,\mu}[P_M]$.

Then consider a sequence of latent trait distributions $\{P_{M,\mu}^{(t)}\}_{t=0}^{\infty}$ *generated by the EM algorithm for the regularized likelihood. If* $\mu > 0$ *and* $p_A(y|\gamma)$ *is continuous in* γ *for each* y*, then*

$$\ell_{n,\mu}[P_{M,\mu}^{(t)}] \xrightarrow{t\to\infty} \ell_{n,\mu}[P_{M,\mu}^*].$$

and

$$P_{M,\mu}^{(t)} \xrightarrow[w]{t\to\infty} P_{M,\mu}^*$$

where $\longrightarrow_w$ *denotes weak convergence of measures.*

The proofs are in the supplementary material and rely on techniques for proving the convergence of the unregularized EM algorithm [Chung and Lindsay, 2015], as well as convex optimization in infinite dimensional vector spaces [Kosmol and Müller-Wichards, 2011].

With the inclusion of a small regularization in the likelihood, we gain the uniqueness of a maximizer and weak convergence of an EM algorithm no matter the measurement assumption model $p_A(y|\gamma)$, a very desirable property. This analysis is presented for a continuous latent distribution; however, in practice, we must use an approximation to the regularized EM algorithm by binning the latent distribution to make the problem computationally feasible which we illustrate in Section 2.3.3.

2.3.3 Computation of the NPEM Algorithm

In order to compute the EM algorithm in practice, one must make an approximation of the latent distribution since we cannot evaluate the continuous distribution exactly. By partitioning the distribution into bins, the NPEM algorithm can be computed using simple matrix operations and without the need for sampling at each iteration. We bin the latent density uniformly on the interval $[0, 1]$ using a fixed number of bins R. The latent density now can be simplified to the following form:

$$p_M(\gamma) = R\theta_r \quad \gamma \in \left[\frac{r-1}{R}, \frac{r}{R}\right) \quad r \in \{1, 2, \ldots, R\}.$$

where θ_r is the weight of the mixing density between the points $[\frac{r-1}{R}, \frac{r}{R})$. Let $\widetilde{A} \in \mathbb{R}^{(N+1)\times R}$ be a matrix where $\widetilde{A}_{ij} = \int_{\frac{j-1}{R}}^{\frac{j}{R}} p_A(i|\gamma)d\gamma$ and $\widetilde{A}_{y\cdot} \in \mathbb{R}^R$ be the row vector of $\widetilde{A}$ at position y.The number of bins R will govern the computational approximation of the latent distribution. We include an algorithm for selecting the number of bins in Section 2.2 of the appendix.

Under this binned approximation, the NPEM algorithm can be computed as follows. Let $\widetilde{A}$ be the conditional distribution matrix, then $\widetilde{A}_{ij} = p(Y = i - 1|\gamma \in [\frac{j-1}{R}, \frac{j}{R}])$. We can

re-express the iterations of the NPEM algorithm in a closed form:

$$\theta_j^{(t+1)} = \frac{1}{1+\mu}\sum_{y=0}^{N} p_{MA}(\gamma \in [\frac{j-1}{R}, \frac{j}{R}]|y)\widehat{p}(y) + \frac{\mu}{1+\mu}$$
$$= \frac{1}{1+\mu}\sum_{y=0}^{N} \frac{e_y^\top \widetilde{A}_j \theta_j^{(t)}}{e_y^\top \widetilde{A}\theta^{(t)}}\widehat{p}(y) + \frac{\mu}{1+\mu} \ .$$

Alternatively,

$$\theta^{(t+1)} = \frac{1}{1+\mu}\sum_{y=0}^{N} \frac{\widetilde{A}^\top e_y \odot \theta^{(t)}}{e_y^\top \widetilde{A}\theta^{(t)}}\widehat{p}(y) + \frac{\mu}{1+\mu}$$

where $\odot$ refers to the Hadamard product.

Remark. We remark that after discretizing the latent distribution, the maximum likelihood operation can be expressed as a geometric program, which can be solved using generic solvers such as the splitting conic solver (SCS) algorithm [O'Donoghue et al., 2016] or using proprietary solvers such as MOSEK[Andersen and Andersen, 2000]. However, due to the closed form expression, our regularized NPEM approach is much simpler and can be computed much faster. Details of the geometric program and comparisons of computation time is given in the Supplementary Materials.

2.3.4 Selection of regularization parameter

We consider the selection of the regularization parameter using cross-validation. Given a measurement assumption model A we partition the data into K folds $\{\mathcal{I}_k\}_{k=1}^K$. For each k, we construct the training data using all but the k^{th} fold $\mathcal{I}_k^c = \cup_{k' \neq k}\mathcal{I}_{k'}$, and validate using the k^{th} fold $\mathcal{I}_k$. The training loss can be computed as follows $\ell_{train}(P_m) = \frac{1}{|\mathcal{I}_k^c|}\sum_{i\in\mathcal{I}_k^c} \log\left(\int p_A(y_i|\gamma)dP_M(\gamma)\right) + \mu\mathcal{D}(P_U||P_M)$ while the validation loss will not include the regularization, $\ell_{val}(P_m) = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \log\left(\int p_A(y_i|\gamma)dP_M(\gamma)\right)$.

2.3.5 Incorporating covariates

Our framework can incorporate covariates easily by considering a weighting function $w(x, x')$ representing the similarity of any two covariate vectors x, x'. This can include simple smooth-

ing methods such as a kernel function $K(x, x')$ and more general similarity measures such as kernel functions or tree-based methods. Based on a choice of w, we can construct a conditional likelihood is expressed as follows:

$$\ell_{n,\mu}[P_M|x; w] = \frac{1}{n}\sum_{i=1}^{n} w(x, X_i) \log\left(\int p_A(y_i|\gamma)dP_M(\gamma|x)\right) + \mu\int_0^1 \log(p_M(\gamma|x))d\gamma.$$

Note that $p_M(\gamma|x)$ now depends on x as well, so the binning approach in Section 2.3.3 will require replacing the parameter θ for each target x by $\theta(x)$.

For example, if we use a tree-based partition for weighting, the weight function is

$$w(x, x') = \begin{cases} 1 & \text{if } (x, x') \in B_s \text{ for some } s \\ 0 & \text{otherwise} \end{cases}$$

where $\{B_s : s = 1, \cdots, K\}$ is a partition of the covariate space. For example, we can partition the age into $\{[50, 55), [55, 60), [60, 65), \ldots, \}$. For a smoothing kernel approach, the weight function is chosen to be

$$w(x, x') = \exp(-\frac{1}{2\sigma^2}||x - x'||^2),$$

where $\sigma > 0$ is the smoothing parameter that can be selected using the cross-validation procedure as in section 2.3.4.

2.3.6 Conditional Distribution Computation

By applying the above procedure to both $(X, Y, V = 0)$ and $(X, Z, V = 1)$, we obtain estimates of $p_M(\gamma|x)$ and $p_M(\zeta|x)$. These quantities can be used to create the conversion mappings $\phi_{\gamma\to\zeta}(\cdot|x)$ and $\phi_{\zeta\to\gamma}(\cdot|x)$.

Many tasks will be desired on the new scale Z, most of which will involve the computation of the model-based conditional distribution $Z|Y, X$. This can be computed as follows. Let $\widehat{F}(\gamma|x)$ be the estimated conditional distribution of γ and $\widehat{G}(\zeta|x)$ that of ζ. Then

$$\widehat{p}(z|y, x) = \frac{\int_0^1 p(z|\widehat{G}^{-1}(\omega|x))p(y|\widehat{F}^{-1}(\omega|x))d\omega}{\int_0^1 p(y|\gamma)dF(\gamma|x)}$$

If one is interested in a point prediction, a natural method is to take the mean, median or mode of $\widehat{p}(z|y,x)$. If an analyst is interested in a prediction interval of the missing test, then one can take the highest probability z values to create a prediction interval.

In the supplement, we describe specific details how to compute the conditional distribution for score conversion when using the binned approximation.

2.3.7 Multiple Conversion and Uncertainty Quantification

The main goal of constructing a joint model of the two outcomes is to use it for uncertainty quantification in later downstream tasks. Without properly accounting for uncertainty, score conversion may lead to spurious conclusions such as inflated type 1 error rates. In this section, we highlight an example of this task, i.e. fitting a generalized linear model (GLM) to the whole population to understand the relationship between a covariate (e.g. age) and a standard outcome of interest (e.g. neuropsychological test score) while leveraging another outcome measurements. The proper way to do the conversion is converting the score by sampling from the conversion model, e.g., if we are missing z, we convert the score of y to z by sampling from $\widehat{p}(z|y,x)$. To reduce the Monte Carlo error due to the randomness of the conversion (sampling), we will perform conversion multiple times for each individual and aggregate the results together to fit the model. This shares a similar structure and intuition as multiple imputation in the missing data problem, so we call it multiple conversion.

Complete Data Inference for GLMs

We are interested in the inference in a possibly misspecified GLM, which can be used to summarize the relationship between a dependent set of variables X_i and the measurement scale of interest Z_i (i.e. fitting a model $g(\mathbb{E}[Z_i|X_i]) = \beta^\top X_i$). In this setting, g is some pre-specified link function, and β is the parameter which is the maximizer of the population likelihood associated with the GLM $l(Z_i, g^{-1}(\beta^T X_i))$, $\beta^* = \text{argmax}_\beta \mathbb{E}[l(Z_i, g^{-1}(\beta^T X_i))]$.

We must consider three sources of uncertainty. First, the sampling uncertainty of the data; second, the finite sample imputation uncertainty; and lastly, the uncertainty in the

estimated conditional model itself. The first two can be handled by a standard multiple imputation approach for estimating the mean and variance of a parameter using Rubin's rules [Rubin, 1996]. However, this can lead to under-coverage in inference because it does not account for the uncertainty in estimating the conversion model itself (i.e. the estimate of $p(z|x, y)$). To properly handle the uncertainty, we propose a bootstrap procedure to account for the variability parameter estimate, which accounts for all three types of uncertainty.

Bootstrap Multiple Conversion.

We combine the standard bootstrap procedure with multiple imputation and illustrate the procedure in Algorithm 1. There are two types of resampling involved: firstly, multiple imputation resampling M-times for the conversion of scores. Secondly, we resample the observed data B-times to incorporate uncertainty in the conversion function $p(z|y, x)$.

Let $\widehat{\beta}_M$ be the estimated parameter using the conditional distribution $\widehat{p}(z|y, x)$ estimated from the full data (Step 2. in Algorithm 1). We then resample with replacement the original data to account for the uncertainty in the conditional model $\{\widehat{p}^{(b)}(z|y, x)\}_{b=1}^{B}$. For each conversion model, we use multiple imputation to estimate the target parameters while accounting for this conversion model uncertainty $\widehat{\beta}_M^{(1)}, \cdots, \widehat{\beta}_M^{(B)}$. (Step 3. in Algorithm 1)

Lastly, we use a standard formula to compute the estimated variance-covariance matrix of $\widehat{\beta}_M$, i.e., $\widehat{\mathbb{V}}_{model} = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\beta}_M^{(b)} - \widehat{\beta}_M)(\widehat{\beta}_M^{(b)} - \widehat{\beta}_M)^\top$ (Step 4. in Algorithm 1). For further details on the use of the bootstrap with multiple imputation see Schomaker and Heumann [2018]."

Accounting for the uncertainty in the multiple imputation using Algorithm 1 will allow us to construct proper coverage confidence intervals for the regression parameters. In practice, we can choose $M = 50$, $B = 50$ to be default parameters which tend to be sufficient in our simulations.

Remark. Though we present these results for estimating the parameters of a GLM, this procedure is easily replicated for many statistical procedures where we want to infer some parameter on the full data using the imputed test version.

2.4 Measurement Assumption Model and Feasibility tests

In this next section we turn our attention to model selection and testing the measurement model $p_A(y|\gamma)$. Data driven selection of the measurement assumption model will typically require multiple observations per individual. We label these multiple outcomes $Y_{i1}, Y_{i2}, \ldots$ and $Z_{i1}, Z_{i2}, \ldots$ respectively.

2.4.1 Choice of measurement assumption models

When studying an ordinal score outcome Y, a conventional choice for modeling p_A is the binomial model [Lindsay, 1983a, Wood, 1999, Tian et al., 2017, Vinayak et al., 2019]. However, this may be a restrictive assumption as it inherently depends on equal difficulties for each test question with independent responses conditional on the latent trait. If we had access to all the question items, it would be possible to learn a conditional model with the difficulties for each question using a Rasch model [Rasch, 1966, Lindsay et al., 1991] or a more elaborate item response model. In the NACC UDS application, however, the individual binary question items are unavailable. One option is to let a domain expert select the measurement assumption model, however, we can also test for its feasibility as we present in Section 2.4.2 and we also present a method for selecting a model from the data in Section 2.4.3.

We first consider a flexible method to construct a measurement assumption which we denote the measurement kernel model (MKM). Specifically, we model the measurement assumption as

$$p_A(y|\gamma) \propto K\left(\frac{y - N\gamma}{h}\right), \tag{2.9}$$

where K is the measurement kernel and $h > 0$ is the smoothing bandwidth. This model enables great flexibility regarding the shape of the error distribution, controlled by K, and the spread of the distribution, controlled by h. We denote the conditional distribution $Y|\gamma \sim MKM(K, h, \gamma)$. The kernel function defines the relative probability of the conditional distribution $Y|\gamma$, assigning a higher probability for y near $N\gamma$. K defines the decay of the relative probability as y is further from $N\gamma$

Due to the fact we allow for great flexibility in the latent model, the measurement assumption will not be identifiable in general. However, the conditional model P_A provides constraints on the distribution of score Y. This allows us to construct tests of whether a particular measurement model is compatible with the data, and can be thought of as a type of falsification test under partial identifiability.

When we only have one observation per person, it is difficult to check if P_A is compatible with the observed score. This is due to the fact that the only observable distribution is that of the scores $\widehat{p}(y)$. However, the NACC UDS is longitudinal, and we have multiple observations per individual that can be leveraged for model checking.

Remark. Under the measurement kernel model, letting $h \to 0$ will allow all univariate distributions to be expressed under this model. However, this places strong restrictions on the bivariate distribution.

2.4.2 Feasibility tests of the measurement assumption model

Here, we propose two simple tests to examine if the measurement assumption is reasonable. The first one is based on a single observation per individual, so we call it the first-order feasibility test. However, selecting a measurement model is challenging with a single observation per individual, and only over-dispersed models can be ruled out. The second approach leverages information from two subsequent observations, and we call it the second-order feasibility test. The goal is to examine whether the measurement assumption $p_A(y|\gamma)$ is compatible with the observed data. For a fixed measurement assumption, the collection of possible observed distributions (over $\boldsymbol{y}$) is denoted as $\mathcal{M}_A$ and can be understood as the space of all mixtures over γ of $P_A(y|\gamma)$. In sections 2.4.2 and 2.4.2 we describe the geometry of this model in more detail.

Our feasibility test involves comparing the empirical distribution $\widehat{p} \in \mathcal{S}_{\widetilde{T}(N)}$ to the fitted model $\widehat{p}_{MA} \in \mathcal{S}_{\widetilde{T}(N)}$ where $\mathcal{S}$ is the probability simplex and $\widetilde{T}(N)$ is the number of cells for the possible discrete distributions. For example, in the case of a first order feasibility test, then this set will be $(N+1)$ and for the second order test, this will be $(N+1)^2$.

To produce a test for the null hypothesis $H_0 : p_0 \in \mathcal{M}_A$ against the alternative $H_A : p_0 \notin \mathcal{M}_A$, let T_n and $\widehat{T}_n$ be the following test statistics:

$$T_n = 2n\mathcal{D}(\widehat{p}_n||p_0) \ ,$$
$$\widehat{T}_n = \arg\min_{p\in\mathcal{M}_A} 2n\mathcal{D}(\widehat{p}_n||p).$$

By definition, $\widehat{T}_n \leq T_n$. Since $\widehat{p}_{MA}$ is the marginal distribution corresponding to the unregularized solution of equation (2.4), then $\widehat{T}_n = 2n\mathcal{D}(\widehat{p}_n||\widehat{p}_{MA})$. We have for any $p_0 \in \mathcal{M}_A$, $P(\widehat{T}_n > t|H_0) \leq P(T_n > t|H_0)$. Therefore, any bound on the tail probability of T_n (i.e. $\mathbb{P}(T_n > t|H_0)$) will be immediately applicable to $\widehat{T}_n$.

We will use the recent finite sample concentration result in [Guo and Richardson, 2021] to compute an upper bound for $P(\widehat{T}_n > t|H_0)$. This bound requires a non-convex univariate optimization procedure but is implemented in the author's corresponding `multChernoff R` package. We can also use the slightly looser bound of Mardia et al. [2018], which has a closed form. A failure of this test indicates gross misspecification of the measurement model. In the first-order feasibility test, the cardinally of the space of the discrete distribution is $(N + 1)$. We note that for a fixed number of questions N and a growing n, that these bounds are asymptotically tight. However, for growing N, improvements on these bounds are an active area of research. We outline this procedure in Algorithm 2.

We can also consider second (and higher order) versions of the test for which we replace $\widehat{p}_n$ with the empirical distribution over two scores sampled from the same individual in a short time span. This is to quantify the variability of two subsequent observations using a single γ. In practice, we consider individuals with subsequent visits < 2 years apart without a change in clinical diagnosis score to have a stable cognitive score, and thus we can consider the implied distribution of bivariate scores over this time. To conduct the second order test, we denote the empirical distribution of pairs of scores of individuals over two subsequent visits $\widehat{p}_{2,n} \in \mathcal{S}_{(N+1)^2}$. We can define the measurement assumption model by the product of the two conditional distributions, $p_A(y_1, y_2|\gamma) = p_A(y_1|\gamma)p_A(y_2|\gamma)$, and use the same estimation procedure for computing the NPMLE. We can repeat this process using an unregularized

NPMLE to compute the corresponding observed data distribution $\widehat{p}_{2,MA}$, and test statistics, $T_{2,n} = 2n\mathcal{D}(\widehat{p}_{2,n}||p_0)$, $\widehat{T}_{2,n} = 2n\mathcal{D}(\widehat{p}_{2,n}||\widehat{p}_{2,MA})$. In the following subsections, we discuss the geometric aspects of this test in more detail.

Remark. We use the term feasibility test rather than a hypothesis test since for a finite k, we will not, in general, be able to discern all models $p_A(y|\gamma)$ as $n \to \infty$, only rule out models which do not meet the feasibility test. This test would have no power in these situations, a phenomenon similar to falsification tests in partially identified problems [Kang et al., 2013, Wang et al., 2017].

First-order feasibility test

Given a measurement model and a latent distribution, we can compute the implied marginal distribution $p_{MA}(y) = \int p_A(y|\gamma)dP_M(\gamma)$. The marginal distribution $p_{MA}(y)$ can be compared to the empirical distribution $\widehat{p}_n$, the empirical observed distribution. We can use the KL divergence between $\widehat{p}_n$ and $\widehat{p}_{MA}$ to investigate whether such a measurement assumption $p_A(y|\gamma)$ is feasible.

Geometrically, the first-order feasibility test can be understood as follows. Let $W = (W_0, \cdots, W_N) \in \mathbb{R}^{N+1}$ be a probability vector, i.e., $\sum_{j=0}^{N} W_j = 1$ and $W_j \geq 0$. Clearly, $W \in \mathcal{S}_{N+1}$, where $\mathcal{S}_{N+1}$ is the $(N+1)$-simplex. The empirical distribution $\widehat{p}_n$ is a point $\widehat{p}_n = (\widehat{p}_n(0), \cdots, \widehat{p}_n(N)) \in \mathcal{S}_{N+1}$. At a given γ, the measurement assumption $p_A(y|\gamma)$ is also an element $p_A(\cdot|\gamma) = (p_A(0|\gamma), \cdots, p_A(N|\gamma)) \in \mathcal{S}_{N+1}$. By the same construction, the implied marginal distribution $p_{MA} \in \mathcal{S}_{N+1}$. While a different latent distribution $p_M(\gamma)$ leads to a different marginal p_{MA}, it is easy to see that the collection of all possible observed distributions from a measurement assumption p_A is the convex combination of all distributions for a fixed γ. We let Γ_1 denote the path of all just distributions parameterized by γ

$$\Gamma_1 = (p_A(\boldsymbol{y}|\gamma) : \gamma \in [0, 1]) \subset \mathcal{S}_{N+1}.$$

Then $\mathsf{conv}(\Gamma)$, the convex hull of this path is the set of all possible distributions represented through any mixture under the measurement model $p_A(y|\gamma)$. We illustrate this in Figure 2.2.

We test for population feasibility, i.e. whether $p_0 \in \text{conv}(\Gamma)$ by checking the closeness of $\widehat{p}_n$ to conv(Γ). This can be done by fitting the NPMLE. Though the estimated mixing distribution $\widehat{P}_M$ may not be unique, the marginal implied distribution $\widehat{p}_{MA}$ will be unique. This is due to the fact maximum likelihood estimation is equivalent to finding a particular mixing distribution $\widehat{P}_M$, which minimizes the relative entropy distance between $\widehat{p}$ and $p_{MA} \in$ conv(Γ) while $\mathcal{D}(\widehat{p}||\cdot)$ is strictly convex. See Figure 2.2 for an illustration of the path Γ and the convex hull. For many choices of conditional distribution, the dimension of this convex hull will be equal to that of the simplex. Therefore, we will exploit sequential observations to narrow the conditional model further.

Second-order feasibility test

We can generalize the above procedure when we have two observations of the same individual which occurs in the NACC data. We assume an individual has a constant trait γ between measurements (in practice we make this assumption by selecting two subsequent visits within a time frame where no other cognitive changes have occurred). Our model then describes a restriction on the distribution of the pairs of measurements. Similarly let $W_{(2)} = (W_{0,0}, W_{0,1}, \ldots, W_{N,N})$ be a probability vector $W_{(2)} \in \mathcal{S}_{(N+1)^2}$. We define $\widehat{p}_{2,n} = (\widehat{p}(0,0), \widehat{p}(0,1), \ldots, \widehat{p}(N,N)) \in \mathcal{S}_{(N+1)^2}$ as the empirical distribution of the pairs of observations. When two observations are generated from an individual with a single γ, we can fit a latent distribution on two observations with the following mixture likelihood:

$$\ell_{2,n}[P_M] = \frac{1}{n}\sum_{i=1}^{n} \log\left(\int p_A(y_{i1}|\gamma)p_A(y_{i2}|\gamma)dP_M(\gamma)\right) \quad P_M \in \mathcal{P}_{[0,1]}$$

We obtain an analogous implied marginal distribution on the distribution of pairs (Y_{i1}, Y_{i2}), $\widehat{p}_{2,MA}$ from a fitted model where $\widehat{P}_{2,M} \in \arg\max \ell_{2,n}[P_M]$

$$\widehat{p}_{2,MA}(y_1, y_2) = \int p_A(y_1|\gamma)p_A(y_2|\gamma)d\widehat{P}_{2,M}(\gamma)d\gamma$$

As before, the collection of all bivariate distributions generated by this model with a fixed γ can be expressed as conv(Γ_2) where $\Gamma_2 \subset \mathcal{S}_{(N+1)^2}$ is the path defined by the set of conditional

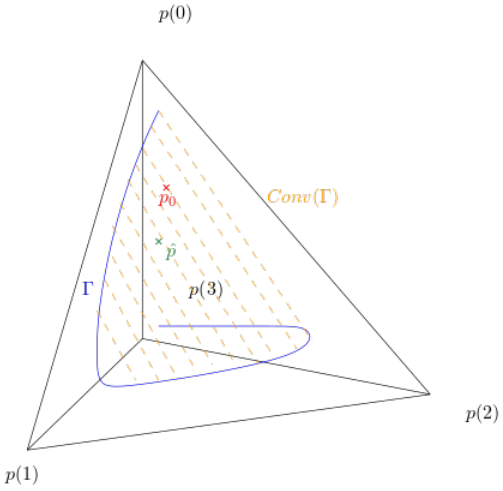


Figure 2.2: Example for $N = 3$ of the first order population feasibility of p_0 and $\widehat{p}$. As both are in the interior of conv(Γ_1)

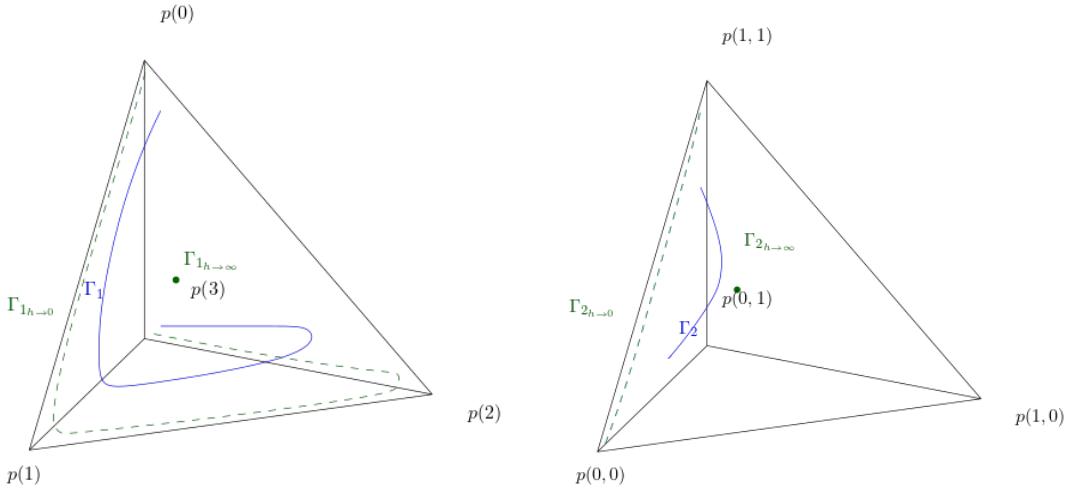
distributions defined as

$$\Gamma_2 = (p_A(\boldsymbol{y}_1|\gamma) \otimes p_A(\boldsymbol{y}_2|\gamma) : \gamma \in [0,1]) \subset \mathcal{S}_{(N+1)^2}$$

and $\otimes$ is the outer product.

The reason to consider a second-order feasibility test is that the first-order test may not be able to detect under-dispersion in the measurement model. In particular in the measurement kernel model when $h \to 0$. The measurement assumption model becomes sharper and approaches a path on the boundary of $\mathcal{S}_{(N+1)}$, meaning all population score distributions can be represented by this model $p_A(y|\gamma)$. Also, as $h \to \infty$, the measurement assumption path converges to a single point $\Gamma \to p_U(\boldsymbol{y})$ where $p_U(\boldsymbol{y})$ is the uniform distribution over $N+1$ points and conv(Γ) shrinks to that point.

When 2 observations are collected per individual, under the measurement kernel model as $h \to 0$, Γ_2 approaches the edge of the simplex moving from $p(0,0)$ to $p(1,1), \ldots, p(N,N)$, and therefore, this model can be falsified. This is because, under this model, we assume no variability given an individual's γ score. See Figure 2.3 for an outline. Once again, all the



(a) First order population feasibility region (b) Second order population feasibility region

Figure 2.3: Though conv (Γ_1) tends to grow as h gets small, this places restrictions on conv(Γ_2) approaches a line on the boundary of $\mathcal{S}_{(N+1)^2}$ connecting the distributions with point masses on $p_0(0,0), p_0(1,1), \ldots, p_0(N,N)$.

bivariate distributions expressible by a single mixing distribution can be denoted as conv(Γ_2) and a second order population feasibility test can be interpreted as to whether $\widehat{p}_{2,MA}(\boldsymbol{y}_1, \boldsymbol{y}_2)$ is sufficiently close to $\widehat{p}_{2,n}(\boldsymbol{y}_1, \boldsymbol{y}_2)$. As indicated previously, the distribution $\widehat{p}_{2,MA}$ is the unique closest point in conv(Γ_2) to $\widehat{p}_{2,n}$ in terms of the KL-divergence.

The computation of the finite sample valid p-values follows immediately in the second-order case. This idea can be generalized easily to the k-th order feasibility test for settings where k observations from a common latent trait are available.

2.4.3 Selection of the Measurement Assumption model using consecutive observations

Beyond just testing whether a model is feasible, we would like to select the model that best fits the data. Here we describe a simple data-driven procedure of choosing a measurement assumption model based on two consecutive observations in longitudinal data. This will be presented for two consecutive observations but can be generalized to many observations.

Let $\mathcal{A}$ be a discrete collection of measurement assumption models. One can think of the element $A \in \mathcal{A}$ as a particular choice of measurement assumption model with kernel K and bandwidth h. Our model selection procedure is simple and considers how well the measurement assumption model fits the observed data from any distribution. This could involve several choices of h and a given K. We express the likelihood as a function of the measurement model and the latent distribution P_M. Given a dataset which consists of bivariate outcomes $\{(y_{i1}, y_{i2})\}_{i=1}^n$ where y_{i2} may not be observed, then we can define the likelihood $\ell_{(1,2),n}[A, P_M]$ as follows

$$\ell_{(1,2),n}[A, P_M] = \sum_{i=1}^{n} I(y_{i2} \text{ observed}) \log \left(\int p_A(y_{i1}|\gamma) p_A(y_{i2}|\gamma) dP_M(\gamma) \right)$$
$$+ I(y_{i2} \text{ not observed}) \log \left(\int p_A(y_{i1}|\gamma) dP_M(\gamma) \right)$$

Then we consider the collapsed version by plugging in the maximizer for a given A

$$\ell_{(1,2),n}[A] = \sup_{P_M \in \mathcal{P}_{[0,1]}} \ell_{(1,2),n}[A, P_M]$$
$$\widehat{A} = \arg\max_{A \in \mathcal{A}} \ell_{(1,2),n}[A].$$

This procedure is essentially a profile likelihood strategy for selecting the model A. We select the optimal $\widehat{A}$ via grid-search as $\mathcal{A}$ will typically not be large. If there are multiple maxima, then any of these fit the data equally well and cannot distinguish between them. We highlight the efficacy of this method to correctly select the conditional model in the supplement through a simulation study.

2.5 Simulations

We conduct simulations to illustrate the efficacy of our methods. We first describe the simulated joint model, then we illustrate our methods first on a prediction task, and multiple imputation inference tasks.

In the supplement, we also include additional simulations of the performance cross-validation procedures for choosing the regularization parameter. We also highlight the improved computational speed of our algorithm against off-the-shelf geometric program solvers, and illustrate the consistency of our model selection procedure.

2.5.1 Simulated Data

Assume we only observe one version of the measurements. Consider the joint model simulated independently and identically for each i. We construct X to be a covariate representing age and simulate from the following model:

$$\begin{aligned} X_i &\sim \text{Uniform}\{55, 56, \ldots, 79, 80\}, \\ m_i(X_i) &= -((1/10)(X_i - 67.5))^2 + 1/2, \\ \gamma_i &= \text{logistic}(m_i(X_i) + \epsilon_i), \qquad \epsilon_i \sim N(0,1), \qquad \zeta_i = \sqrt{\gamma_i}, \\ Y_i &\sim \text{Binomial}(N_Y, \gamma_i), \\ Z_i &\sim \text{Binomial}(N_Z, \zeta_i), \\ V_i &\sim p(V|X_i). \end{aligned}$$

where the missingness mechanism depends on the covariate value

$$p(V = 1|X_i) = \begin{cases} 0.8 & \text{if } X_i < 61 \\ 0.2 & \text{otherwise} \end{cases}.$$

2.5.2 Simulations: Prediction

We conduct simulation to predict the missing score Z_i for $V_i = 0$ for $N_y = N_z = 30$ and $n \in \{104, 195, 494, 1001, 2002, 5005\}$. This seemingly unusual choice is so that all samples are divisible by 26, the number of grid points in the x distribution, however this is not strictly necessary for the method since we use smoothing weights over the covariates.

For smoothing over similar covariates, we use kernels for which $\sigma_n = 5/n^{1/5}$ for smoothing $w_{\sigma_n}(x, x') = \exp(-\frac{(x-x')^2}{\sigma_n^2})$. We also let $\mu_x = \mu_y = 50/n$. These choices are driven by

the understanding that the rate $n^{-1/5}$ represents the best smoothing rate for nonparametric regression. Additionally, reducing the regularization terms μ_x and μ_y at a rate of n^{-1} guarantees that as the sample size increases, the impact of these regularization terms is comparable to a constant number of observations and becomes insignificant as the sample size continues to expand.

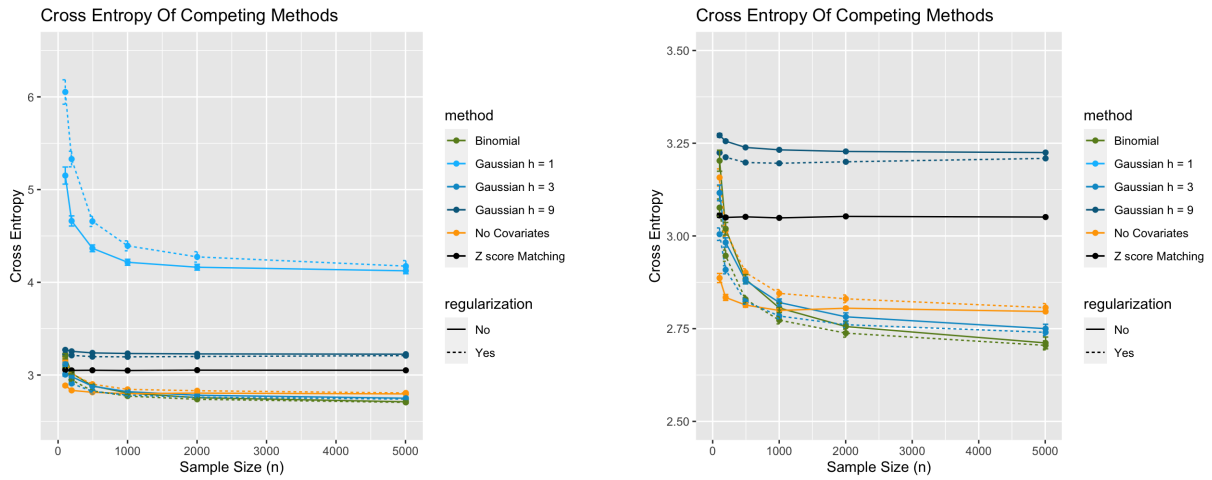
We compare the conversion with the correct measurement models against several misspecified measurement kernel models. We also use a set of Gaussian measurement kernel for K with $h \in \{1, 3, 9\}$ as in equation (2.9). These bandwidth choices correspond to under-dispersed, well-fitting and over-dispersed cases. The value of $h = 3$ was chosen as the best among these as it achieves the lowest cross-entropy in Figure 2.4(b), though the true model, the binomial distribution performs the best.

We also compare against a simple conversion model which is commonly used in practice, for which normal noise is assumed, and we round to the nearest point in the data using a linear transformation (i.e., we take μ_w, σ_w for $w \in \{y, z\}$ to be the corresponding mean and variance, then for a given y we predict z using $\widetilde{z} = \frac{\sigma_z}{\sigma_y}(y - \mu_y) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_z)$ we then round ϵ_i to make the score and make the prediction). We then evaluate the methods on the cross-entropy C_n on the missing score

$$CE_n[\widehat{p}(z|y, x)] = \sum_{i=1}^{n} -\log\left(\widehat{p}(z_i|y_i, x_i)\right) I(V_i = 0).$$

A lower cross entropy denotes a better estimate of the conditional distribution.

In Figure 2.4, we see that when the conditional model is correct, the best-performing model is the covariate smoothed version for large n. However, for small n, using our method without including covariates performs better. Additionally, we see that the severely misspecified measurement models (ones for which $h = 1, 9$) perform poorly; however, the best model in that set $h = 3$ still performs well. We also see that the benefit of including the regularization term is more dramatic for smaller sample sizes and when the covariates are included as the regularization stabilizes the estimates.



(a) Comparison of cross-entropy of conversion methods.

(b) Zoomed in cross-entropy to discern the binomial and $h = 3$ cases.

Figure 2.4: Conversion cross-entropy. True data generated from Binomial conditional models, which tend to perform the best. The Gaussian MKM conditional model with $h = 3$ also performs well.

2.5.3 Simulations: Inference

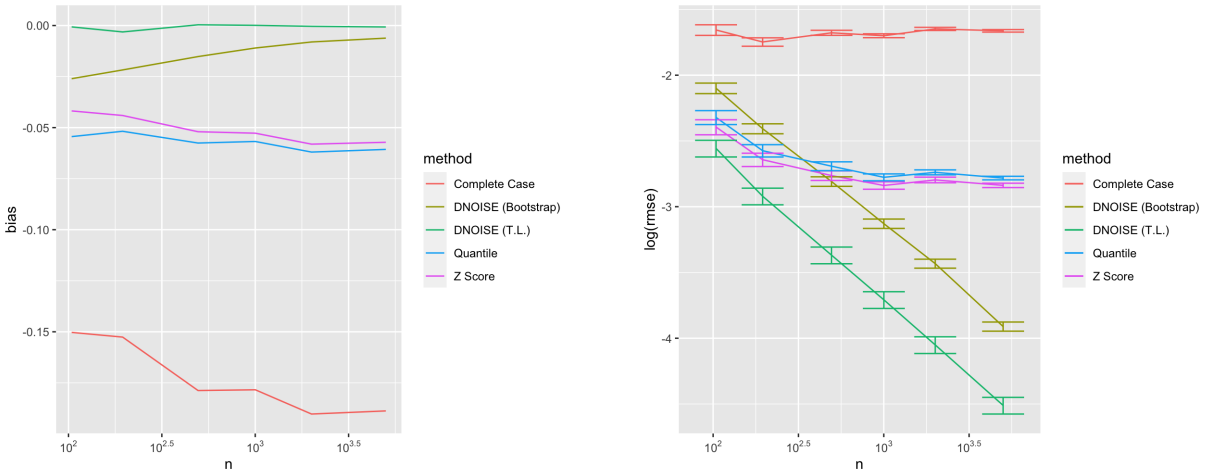
Using the data generating process, we consider a simple regression problem. We want to understand the misspecified linear regression of $\mathbb{E}[Z|X] = \beta_0 + \beta_1 X_{age}$. We note that the optimal regression parameter for this model is $\beta_1 = 0$. This is due to the fact that $\gamma|X$ is symmetric around our domain of X, thus the best fitting linear model has value $\beta_1 = 0$. However, due to the missing at-random assumption, the Z only complete case regression model will be biased, as the tendency is for Z to be observed ($V = 1$) under smaller values of X.

To account for the uncertainty in the estimate of the latent trait distributions, we apply nonparametric bootstrap to account for this additional source of variation.

We use robust standard errors and compare the corresponding regression coefficients for our method, with a z-score matching approach, a quantile matching approach, and our meth-

ods, including and not including covariate adjustment. We impute 50 scores for each missing data point and let $B = 200$ be the number of bootstraps used to compute the additional variance. For each bootstrap re-sample, we are re-estimating the latent distribution to account for this source of uncertainty.

We first compare the root mean squared error (RMSE) and bias of each of the estimates in Figure 2.5. The bias is minimal in the bootstrap and covariate-adjusted versions. Each of these methods tends to decrease RMSE when sample size increases, whereas when the covariates are not adjusted for, as in the naive methods, there is a plateau in the RMSE due to the asymptotic bias of the estimators.



(a) Bias of estimate of $\widehat{\beta}_1$

(b) RMSE of estimate of $\widehat{\beta}_1$

Figure 2.5: Bias and RMSE of regression parameter estimation. T.L. refers to an oracle using the true latent distribution, while the bootstrap method estimates the latent distribution with uncertainty our bootstrap procedure.

We also plot the coverage in Figure 2.6 and observe that the proper coverage is obtained when using the bootstrap-adjusted version. We consider coverage at the $\alpha = 0.05$ level. We see that the nominal coverage is consistently too small when not accounting for the sampling variability with the latent trait. However, this can be corrected by using the bootstrap to

obtain the nominal coverage rate. The oracle method uses the exact latent traits tends to over-cover. The remaining methods suffer from under-coverage due to lack of uncertainty propagation.

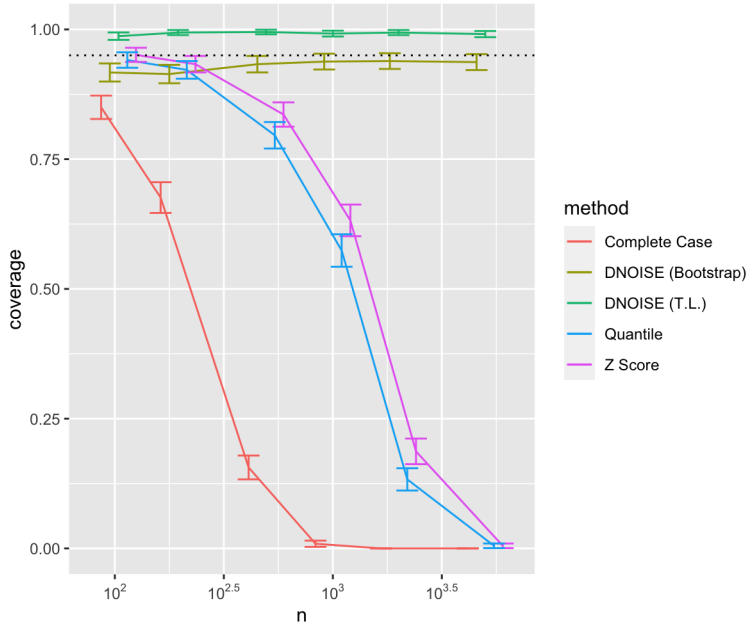


Figure 2.6: Coverage of imputation methods. T.L. refers to an oracle using the true latent distribution, while the bootstrap method estimates the latent distribution with uncertainty our bootstrap procedure.

2.6 Application to the NACC data

Our primary motivation has been to use this method for converting scores in the NACC Uniform dataset (UDS data freeze No. 47, Obtained July 2020). We consider the conversion between the proprietary C1 battery score, the Mini-Mental State Examination (MMSE), and the non-proprietary C2 battery score, the Montreal Cognitive Assessment (MOCA). Both scores have a range of $\{0, \ldots, 30\}$. We have a sample of 11194 individuals having a recorded MMSE score and 6898 with a MOCA. We consider a training set of first visits

and a validation set of second visits within 500 days of the first as a bivariate outcome set used for model selection. We have 7614 and 4051 follow-up visits within each test, respectively. Lastly, we have 760 individuals as part of the crosswalk dataset, a much smaller group of individuals with both scores measured [Monsell et al., 2016]. Since the crosswalk dataset is comparably very small, learning the joint distribution with this dataset is infeasible. Instead, we utilize the harmonizability assumptions, which allow us to use the whole training dataset to estimate the latent distribution. We reserve the crosswalk dataset to verify the performance of converting scores.

Using our model selection procedure outlined in section 2.4.3, we select the binomial model for both the MMSE and the MOCA measurement assumption model. However, the Gaussian ($h = 3$) and binomial models achieved similar likelihood values for the MOCA and MMSE. Both models obtained a p-value of 1 for the feasibility tests. The results in the following section using the Gaussian measurement assumption model were nearly identical and were therefore omitted for brevity. Using the cross-validation we obtain a smoothing kernel of $w(x, x') = \exp\left(-\left(\frac{x-x'}{\sigma}\right)^2\right)$ with $\sigma = 4$ and $\mu_Y = 0.01, \mu_Z = 0.01$. We next consider prediction and inference tasks for this dataset.

2.6.1 Prediction

We then use this model to compute the cross-entropy on the crosswalk dataset. We compare this against the naive method of score conversion outlined in Section 2.5 by matching Z-scores and adding the appropriate noise.

To validate our method against alternatives, we consider a prediction problem in the crosswalk study, a small group of 760 individuals who completed both the MOCA and MMSE cognitive tests [Monsell et al., 2016]. We label these observations $i \in \{1, \ldots, n_{cw}\}$ with observations $(X_i^{cw}, Y_i^{cw}, Z_i^{cw})$. We compute the cross-entropy as a function of the regularization parameter of each model and plot the results in Figure 2.7. We see that the optimal μ_Y and μ_Z selected from the marginal data only is very close to that of the optimal value with the bivariate crosswalk data.

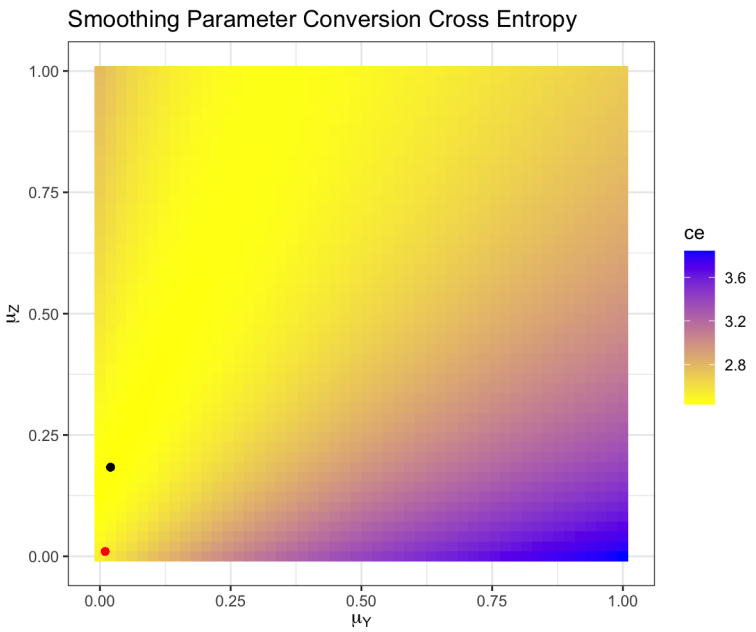


Figure 2.7: Cross-entropy as a function of the regularization parameter of each model. The red dot indicates the regularization parameters chosen without looking at the crosswalk data, and the black dot indicates the regularization parameters chosen from the crosswalk dataset.

We can compare these results to the cross-entropy found for the naive conversion method in Table 2.6.1. We find that compared to z-score matching, we can reduce the conversion cross-entropy by 9.55% by using the learned parameters, and if the crosswalk data is used to select the optimal regularization parameters, then this can be decreased by 12.66%.

Method	Z-Score Matching	DNOISe	DNOISe(CW opt)
CE	2.805	2.547	2.450

Table 2.1: Comparison in cross-entropy of Z-score matching to our methods. CW opt refers to the regularization parameters selected from the full crosswalk dataset, while the other DNOISe column selects the tuning parameters using each of the scores separately without considering the crosswalk dataset.

2.6.2 Inference

The cognitive reserve is a popular concept in dementia research. It refers to the fact that individuals with higher education levels tend to cognitively decline more slowly over time compared to those with a lower education level [Stern, 2009, 2012, Cheng, 2016, Meng and D'arcy, 2012].

In this section, we study how the cognitive reserve may interact with genetic characteristics. It is known that the APOe4 allele is a gene most prominently associated with Alzheimer's disease, with individuals having a substantially elevated risk with these alleles [Sienski et al., 2021]. A particular challenge of studying this progression is the long-term nature of the changes, often taking many years to manifest. We will study a regression model to detect whether we observe an interaction between APOe4 allele and education.

Our outcome of interest which measures cognitive decline is the difference of scores from an initial visit to a followup visit. We consider two time spans, $3-4$ years and $8-10$ years. The first one refers to a short-term effect, whereas the second one refers to a long-term effect. We compare the results to the MOCA-only version and the naive conversion methods.

The number of complete cases with MOCA scores for the 3 year conversion is relatively small (376), whereas we have numerous (7999) observations, most commonly with a single MOCA score and an MMSE score or a pair of MMSE scores. Because we have very few complete cases, any parameter of interest using MOCA-only subsample may be difficult to estimate, and may suffer from bias due to this missing data structure. We will show that for the 3 year conversion, we can leverage the information from the other tests, providing a sound propagation of uncertainty through the model, whereas the Z-score methodology does not fully account for this uncertainty.

For the long-term effect ($8-10$ years), we have no complete cases. Therefore, any inference will require a conversion of scores. In this setting, we have $n = 2240$ individuals with visits matching in this window.

Subpopulation	Age	Sex (Female)	Education $\geq$ 16 years	CDR = 0.5	CDR $\geq$ 1	Num e4 = 1	Num e4 = 2
3-4 Year	73.03 (6.39)	0.576	0.613	0.312	0.098	0.307	0.056
8-10 Year	73.10 (6.48)	0.628	0.650	0.231	0.038	0.289	0.038

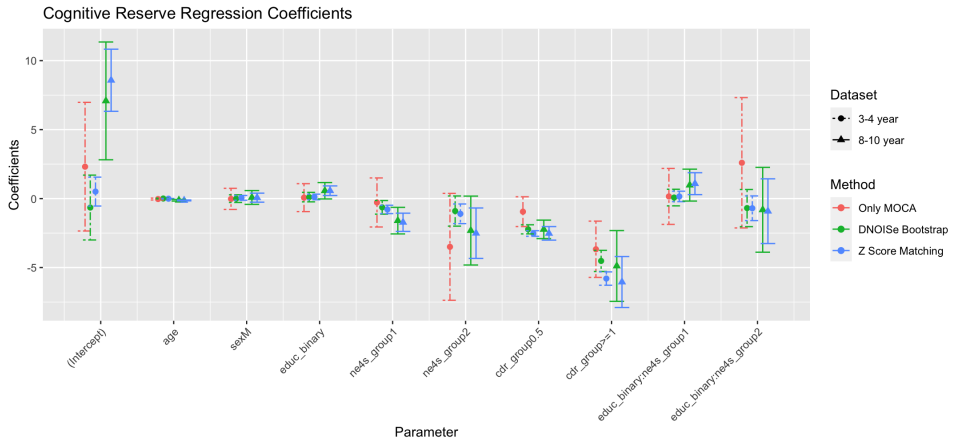
Table 2.2: Population summary statistics for groups used in MOCA decline. Mean (sd) or proportion displayed in the table.

Specifically, we run a linear model according to the following equation:

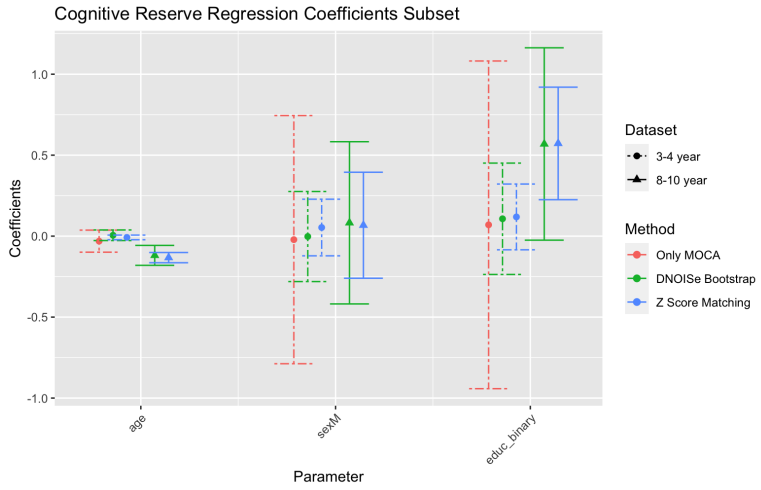
$$\begin{aligned}\mathbb{E}[\Delta Z|X] = \beta_0 + \beta_{age}X_{age} + \beta_{educ}I(Educ \geq 16) + \beta_{sex}I(Sex = Female) \\ + \beta_{cdr,0.5}I(CDR = 0.5) + \beta_{cdr,\geq 1}I(CDR \geq 1) + \beta_{e4,1}I(e4 = 1) + \beta_{e4,2}I(e4 = 2) \\ + \beta_{educ,e4,1}I(e4 = 1, Educ \geq 16) + \beta_{educ,e4,2}I(e4 = 2, Educ \geq 16).\end{aligned}$$

where $\Delta Z = Z_2 - Z_1$ The outcome Y is the difference of scores from baseline to the followup score either $3-4$ or $8-10$ years later. The covariates are measured at the time of the baseline test date. The CDR (cognitive decline rating) is a clinical assessment of an individual's cognitive ability. A CDR score of 0 refers to cognitively normal individuals, a score of 0.5 refers to Mild Cognitive Impairment, and scores of $1, 2, 3$ refers to dementia (with a different severity). We use the baseline CDR score here.

The complete case data has a limited sample size, so we cannot make conclusions about these interaction effects. The naive conversion methods (Z-score) also may not capture effects that are known to hold, such as a faster rate of decline for those with Mild Cognitive Impairment (CDR score 0.5) when compared to the normal controls. In Figure 2.8(a), we see that when considering the interaction between education and those with a single $e4$ allele, this effect is significant if using the naive methods; however, when uncertainty is adequately accounted for using the bootstrap, we see that these conclusions may not be as significant as initially thought. This difference highlights the importance of correctly accounting for uncertainty, lest we be wary of spurious results. Additionally, the main effect for education is observed to be larger in the $8 - 10$ year analysis consistent with the known long-term



(a) Regression coefficient estimates for cognitive decline using both time-frame datasets and all conversion methods. Note the larger uncertainty for the DNOISe method compared to the naive Z-score matching.



(b) Subset of the age, sex, and education covariates for the decline over each time frame. Note that the age effect is only significant in the 8-10 year time frame, indicating an accelerating decline compared to the baseline decline as captured through the intercept only.

Figure 2.8: Regression coefficients from cognitive decline regression

progression of disease. Moreover, we see the importance of propagating uncertainty of the conversions at the worry of false-positive evidence of interaction effects.

2.7 Discussion

In this paper, we introduced a framework for data harmonization inspired by a problem in Alzheimer's Disease and dementia research. We connect the problem to the existing literature in nonparametric mixing density estimation.

When MAR assumption $V \perp\!\!\!\perp (\Omega, Y, Z)|X$ is not satisfied, the latent distribution would be of $p_{M_Y}(\gamma|x, V = 0)$ and $p_{M_Z}(\zeta|x, V = 1)$ which depends on the observed V. This would correspond to a setting where which test is observed may depend on an individual's relative rank Ω. At the NACC, the test change was due to switching from proprietary to non-proprietary methods at a short calendar time window, so we believe that the MAR assumption is reasonable, as the design change is independent of the cognitive ability of the participants.

Our harmonizability assumption requires an invertible transformation between latent traits. This assumption is vital because the two latent traits measure the same cognitive domain. In practice, these traits may not be co-monotonic but are instead highly correlated. In this case, the mapping we are learning is the optimal transport map (under the $c(x, y) = |x-y|$ cost function) between the continuous latent trait models p_{M_Y} and p_{M_Z} [Villani, 2003]. This may help explain the strong performance of converting scores on real-world data.

Our methods draw similarities to semi-parametric factor analysis [Jöreskog and Moustaki, 2001, Gruhl et al., 2013] and item response theory [Johnson, 2007, Woods and Thissen, 2006, Paganin et al., 2021]. These approaches often use parametric models for a latent trait distribution and a nonparametric link function, relating the observed data to the latent data (or vice versa). Our approach is more flexible as we allow a nonparametric model of the latent trait distribution p_M. However, this requires specifying how a conditional distribution of the observed variables is given the latent trait $p_A(y|\gamma)$. Though in nonparametric mixture models with discrete observations, we have a non-identifiability problem [Lindsay, 1995], this has not proven to be an issue when using these imputation methods. However, one could correct this using a similar procedure to Ignatiadis and Wager [2019a]. In a theoretical study

of a similar problem focused on the deconvolution of the true latent distribution, Tian et al. [2017] and Vinayak et al. [2019] proved that the NPMLE and a moment matching method recover the true latent distribution within $\mathcal{O}_P(\frac{1}{N})$ in terms of the Wasserstein 1 distance [Tian et al., 2017, Vinayak et al., 2019] in the case where the conditional distribution p_A is a binomial distribution.

One can interpret our method as an empirical Bayesian procedure, where the latent trait distributions are the priors for the population traits. For an overview, see Efron [2019]. We provide a general framework for including covariates in our problem. However, alternative approaches, such as the location shift models used in recent empirical Bayes problems, may be an avenue for future research [Ignatiadis and Wager, 2019b].

There are many extensions of potential interest. In particular to longitudinal and multivariate data harmonization, as well as theoretical questions involving recovering a latent distribution P_M under other conditional models p_A. Such a model will be extremely important for further statistical applications such as imputation of missing scores, change point detection of mild cognitive impairment, and classification of neuro-degenerative disease via mixture models on the test scores.

For the harmonization of multivariate scores, a naive approach may be to match all the quantiles separately. If the copulae between two multivariate distributions are the same and given a particular cost function, Alfonsi and Jourdain [2014] find that this is the solution to an optimal transport problem. However, this will not hold in general, and a deeper investigation into the use of optimal transport methods may be necessary.

Algorithm 1 Bootstrap Multiple Conversion Procedure

Inputs: B (number of bootstraps), M (number of imputations), $(X_i, Y_i, Z_i)_{i=1}^n$ (observed data).
Estimate the parameter β using the full data.
Estimate the conversion model $\widehat{p}(z|x,y)$ from $(X_i, Y_i, Z_i)_{i=1}^n$.
for $m = 1$ to M **do**
    for each y_i **do**
        Simulate M samples from $\widehat{p}(z|x,y)$ and denote these by $\widehat{z}_{i,m}$.
        Estimate the regression coefficients $\widehat{\beta}^{(m)}$ using the m^{th} sample of $\widehat{z}_{i,m}$.
    end for
end for
Obtain the parameter estimate by averaging over multiple imputation draws $\widehat{\beta}_M = \frac{1}{M}\sum_{m=1}^M \widehat{\beta}^{(m)}$.
for $b = 1$ to B **do** ▷ Bootstrap estimates of the parameter
    Resample with replacement from the observed data to generate $(X_i^{(b)}, Y_i^{(b)}, Z_i^{(b)})_{i=1}^n$.
    Estimate the conversion model $\widehat{p}^{(b)}(z|x,y)$ from $(X_i^{(b)}, Y_i^{(b)}, Z_i^{(b)})_{i=1}^n$.
    for $m = 1$ to M **do**
        for each $y_i^{(b)}$ **do**
            Simulate M samples from $\widehat{p}^{(b)}(z|x,y)$ and denote these by $\widehat{z}_{i,m}^{(b)}$.
            Estimate $\widehat{\beta}^{(b,m)}$ using the m^{th} sample of $\widehat{z}_{i,m}^{(b)}$.
        end for
    end for
    Obtain a bootstrap sample of the parameter by averaging over the multiple imputations: $\widehat{\beta}_M^{(b)} = \frac{1}{M}\sum_{m=1}^M \widehat{\beta}^{(b,m)}$.
end for
Estimate the bootstrap variance $\widehat{\mathbb{V}}_{model} = \frac{1}{B-1}\sum_{b=1}^B (\widehat{\beta}_M^{(b)} - \widehat{\beta}_M)(\widehat{\beta}_M^{(b)} - \widehat{\beta}_M)^\top$.
Return parameter estimate and variance estimate $\widehat{\beta}_M, \widehat{\mathbb{V}}_{model}$.

Algorithm 2 Feasibility Test for a Conditional Model $p_A(y|\gamma)$

1: Select a conditional model $p_A(y|\gamma)$.

2: Compute the empirical distribution of observed scores $Y_i \in \{0, 1, 2, \ldots, N\}$: $\widehat{p}_n$.

3: Estimate the unregularized NPMLE $\widehat{P}_M$, as in equation (2.4).

4: Compute the observed data distribution corresponding to the NPMLE, $\widehat{p}_{MA}$.

5: Compute the observed test statistic $\widehat{t}_{obs} = 2n\mathcal{D}(\widehat{p}_n||\widehat{p}_{MA})$.

6: Compute the upper bound $\widetilde{p}$ where $\mathbb{P}(T_n \geq \widehat{t}_{obs}) \leq \widetilde{p}$. ▷ Using finite sample concentration results such as Guo and Richardson [2021]

7: If $\widetilde{p} \leq \alpha$ for some specified type 1 error threshold, then reject p_A.

Chapter 3

INTERFERENCE WITH PARTIALLY OBSERVED NETWORK DATA

3.1 Introduction

Interference occurs when one individual's treatment status impacts others' outcomes. Interference, also known as "spillover effects," is occurs in multiple scientific domains, including the study of infectious diseases [Hudgens and Halloran, 2008, Tchetgen and VanderWeele, 2012], studying peer influence [Manski, 1993, Bramoullé et al., 2009, De Giorgi et al., 2010, Epple and Romano, 2011, Goldsmith-Pinkham and Imbens, 2013], public policy [Malani et al., 2021, Imai et al., 2021], information diffusion [Banerjee et al., 2013b, 2019], technology adoption [Beaman et al., 2021], online platforms [Saveski et al., 2017, Pouget-Abadie et al., 2018, 2019] and online marketplaces [Ha-Thuc et al., 2020, Johari et al., 2022], among other domains.

Interference violates the stable unit treatment value assumption (SUTVA), which states that an individual's outcome is not impacted by the treatment status of their peers. When SUTVA is violated, each potential outcome, the counterfactual outcome under a given treatment assignment, could depend on all treatment assignments within the population. Valid inference for treatment effects under SUTVA violations is an active area of research, with solutions typically depending on a combination of exposure maps and structural causal models. Exposure maps categorize respondents according to their network characteristics and the vector of treatment statuses [Aronow and Samii, 2017], while structural causal models identify specific pathways for influence between individuals [van der Laan, 2012, Ogburn et al., 2022].

Despite recent advances, estimating causal effects under interference typically requires

complete network data, which is expensive and onerous to collect or may not be available due to privacy constraints. Partially observed network data takes many forms: subgraph samples where a researcher observes the presence/absence for only a subset of possible connections, egocentric sampling using either specific links or aggregates, or network-based sampling methods such as snowball sampling or respondent-driven sampling. In each case, incomplete network information introduces miss-measurement in the exposure map–a person may have treated peers, but if links to those peers are not observed the researcher will think their outcome is totally orthogonal to the treatment.

This paper introduces a framework for estimation and inference of causal effects under *partial* network data arising from a single graph. Partial here means that we may observe some or no links or aggregate summaries of links which we will formalize later. With such data, we recover multiple estimands including various conditional or average treatment effects. To do this, we define a broad class of structural causal models that are amenable to estimation using partial data. This class covers many empirically relevant schemas for interference, such as diffusion and its generalizations.

Estimation leverages a dual approach: first, by using an iterated expectation method for de-biased estimation of model parameters with partial network data, and second, by managing the dependence of exogenous noise in the outcomes. Our method applies when the underlying graph has features captured by the class of node-exchangeable formation models, which we commonly see in practice and connect this to the problem of estimating effects of experiments. Chandrasekhar and Lewis [2011] introduced a similar strategy for cases when multiple networks are available and data are independent across networks, while we tackle the more challenging inference task of single network asymptotics. We also consider is the experimental design associated with network exposure and discuss design for variance reduction of these estimands.

Our approach also significantly aids empirical researchers in experimental design, particularly in scenarios where obtaining pristine network data ahead of randomized controlled trials (RCTs) is challenging. By collecting partial network data and employing a Bayesian

optimization algorithm, we optimize experimental designs that efficiently maximize treatment saturation tailored to specific estimands of interest. Our results demonstrate that this methodology not only surpasses traditional methods like inverse probability weighted (IPW) estimators in estimating global average treatment effects but also facilitates innovative seeding strategies that leverage the unique characteristics of partial network data. In addition, we extend the existing methodology on estimating stochastic blockmodels from aggregated relational data to a version that is more practical as it does not require mutually distinct traits, a challenge to its usability in practice until now.

The remainder of the paper is structured as follows. We begin with a review of related work (Section 3.1.1. Section 3.2 defines the necessary background, then Section 3.3 describes the procedure for estimation and inference. Section 3.4 describes experimental design using partial network data and Section 3.5 provides empirical examples. We conclude in Section 3.6.

3.1.1 Related Work

Complete network data collection can be prohibitively expensive and restricted by privacy concerns [Breza et al., 2020]. Researchers typically work with partial network data derived from various sources such as survey samples, coarse geographic data, kinship information from censuses, or aggregated financial transactions. Comprehensive reviews of methods for handling network data can be found in De Paula [2017], Graham [2020], and discussions on identification in network and related models are provided in Manski [2009].

A direct approach is node subsampling, selecting a portion of nodes from the population and mapping the entire graph among them. If random sampling of nodes is infeasible, or if populations are sensitive or stigmatized, techniques like snowball or respondent-driven sampling offer a limited but focused view of the graph [Heckathorn, 1997, Goel and Salganik, 2009, 2010, Baraff et al., 2016, Green et al., 2020].

When complete edge enumeration among node subsets is impractical, researchers adopt standard survey methods such as Aggregated Relational Data (ARD) collection. The main

intuition is that each of the partial network designs mentioned above can be used to estimate a breakdown of each respondent's network in terms of observable characteristics. In ARD surveys, respondents are asked, "How many people do you know with trait X?" for various traits. Additional conditions may be added in addition to collect the type of connection that is relevant [Feehan et al., 2016]. Originally designed to estimate hard-to-reach populations like HIV-positive men in the US [Killworth et al., 1998b, Scutelniciuc, 2012, Jing et al., 2014], has been extended to a variety of other settings such as financial contagion models [Acemoglu et al., 2015] as well as more general network scale up methods utilized (NSUM) [Killworth et al., 1998a, Kadushin et al., 2006, Feehan and Salganik, 2016, McCormick, 2020] and is notably 70 to 80% less costly than full network data collection [Breza et al., 2020].

Another standard survey method, egocentric sampling, asks respondents to consider specific individuals in their networks and provide detailed information about them, unlike the aggregate focus of ARD and is commonly used in applications such as contact tracing Potter et al. [2011], violence perpetration Bond and Bushman [2017] or adolescent substance measurement Huang et al. [2014]. Clustering based on observed characteristic groups allows for estimating latent network types and subsequently, the mixing across these latent types [Breza et al., 2023], although this approach primarily focuses on network statistics like centrality, not interference.

The first task in causal inference problems is defining the target estimand such as the global average treatment effect (GATE), which assesses the impact of treating everyone versus treating no one, considering peer effects Ugander et al. [2013]. Other interests might include the effect of specific treatment allocations, like identifying influential individuals Kempe et al. [2003], Banerjee et al. [2019], often limited by policy constraints (e.g., subsidies for the ultra-poor as in Anderson and Feder [2007]) or due to non-monotone peer effects dynamics [Banerjee et al., 2018]: treating everyone may change interaction dynamics in equilibrium. More generally Aronow and Samii [2017] compare average treatment effects between two exposure configurations. A distinct but related line of work seeks to detect whether interference is present at all [Athey et al., 2018].

Much of the literature on SUTVA violations and exposure maps assumes a fully observed graph. A recent line of literature address imperfect or incompletely sampled graphs under certain conditions and for specific average causal effects [Hardy et al., 2019, Yu et al., 2022, Cortez et al., 2022]. Models for peer influence like contagion [Jackson et al., 2008, Banerjee et al., 2013b, Beaman et al., 2021, He and Song, 2023] or hearing models [Banerjee et al., 2019] structure interference analysis and we highlight how these can be learned with partial network data. Auerbach and Tabord-Meehan [2021] also explore these effects through structural causal models focusing on nonparametric estimation, while our work emphasizes estimation, inference, and design using partial network data.

3.2 Environment

Let $i \in \{1, 2, \ldots, n\} = \mathscr{V}$ denote a populations of interacting individuals and let $\mathscr{G} = \mathscr{V} \times \mathscr{E}$ be the network by which interference is propagated; where $\mathscr{V}$ is the set of node vertices and $\mathscr{E} \subset \mathscr{V} \times \mathscr{V}$ is a set of edges (either directed or undirected). We can also extend this to weighted graphs, however binary networks are presented for simplicity. We can represent this graph by the adjacency matrix $G \in \{0,1\}^{n \times n}$. We consider binary treatments denoted by a treatment vector $\boldsymbol{a} \in \{0,1\}^n$ and let denote the potential outcome $Y_i(\boldsymbol{a}) \in \mathbb{R}$, under a treatment assignment $\boldsymbol{a}$, and Y_i denote the actual observed outcome. Lastly, we assume that we have access to pre-treatment node-level covariates $X_i \in \mathbb{R}^m$.

In the remainder of the paper let O and o denote the usual big and little oh notation and O_P and o_P denote the stochastically bounded and convergence to 0 in probability for sequences of random variables. We use $\widetilde{O}$ if we are suppressing logarithmic factors in the rate. Let $||\cdot||_p$ denote and p-norm, and let $||\cdot||_F$ denote the Frobenius norm.

3.2.1 A structural causal model

We use the framework of structural causal models, a nonparametric extension of structural equation models [Pearl, 2009]. Similar approaches have been studied by Ogburn et al. [2022] and Auerbach and Tabord-Meehan [2021] in the case of fully observed networks. We derive

a model that is amenable to estimation with partial data.

Let $Y_i(\boldsymbol{a})$ denote the potential outcome of Y_i under a treatment allocation $\boldsymbol{a}$. The exposure mapping V_i is represented as a function f_V such that $V_i = f_V(\boldsymbol{a}, \varphi_i(G)) \in \mathbb{R}^{p_V}$ where φ_i is the relevant graph information for individual i relative to their position with respect to treated individuals. We also allow for the potential outcome to be modulated by some additional confounder $S_i = f_S(\boldsymbol{X}, \vartheta_i(G)) \in \mathbb{R}^{p_S}$. We model the potential outcomes Y_i as a function of the exposure, type-value S_i and some additional noise $\boldsymbol{\varepsilon}_Y$

$$Y_i = f_Y(S_i, V_i, \boldsymbol{\varepsilon}_Y) \tag{3.1}$$

The benefits of structural causal models are that they allow for the characterization of all causal effects in a system, as well as the distributions of counterfactuals. However, they require correct specification of the causal process, i.e. correct specification of the exposure map and the relevant confounders. Even if one can propose a model for interference, estimation is not straightforward due to the fact that we only observe partial graph information in G^*. Many common models of interference can be expressed as structural causal models, and can be thought of as parameterizations of $f_Y(S_i, V_i, \boldsymbol{\varepsilon}_Y) = f_Y(S_i, V_i, \boldsymbol{\varepsilon}_Y; \beta_0)$. This then reduces the challenge to estimating β_0 using partially observed data. The exogenous noise, ϵ_Y, within our model is likely influenced by the graph structure, as interactions and peer effects can induce correlations in outcomes that extend beyond individual exposures. This complexity suggests that the noise, even if initially considered as external to the model, is intertwined with the network dynamics, reflecting the propagation and interference effects inherent in our structural causal framework.

We differentiate the two types of target parameters. The first are the **outcome model** parameters which parameterize the distribution of the outcome, exposure, and confounder (Y, S, V). Specifically, $f_Y(S_i, V_i, \epsilon_Y) = f_Y(S_i, V_i, \epsilon_Y; \beta_0)$ under some parameterization $\beta \in \mathbb{R}^p$. We denote the true model parameters $\beta_0 \in \mathbb{R}^p$. Such parameters may be identified through a moment equation m, $\mathbb{E}[m(Y_i, S_i, V_i, \beta_0)] = 0$ or more explicitly through a regression parameterization. In a model of simple diffusion, this is simply the probability of infecting

a neighbouring node $q \in [0, 1]$.

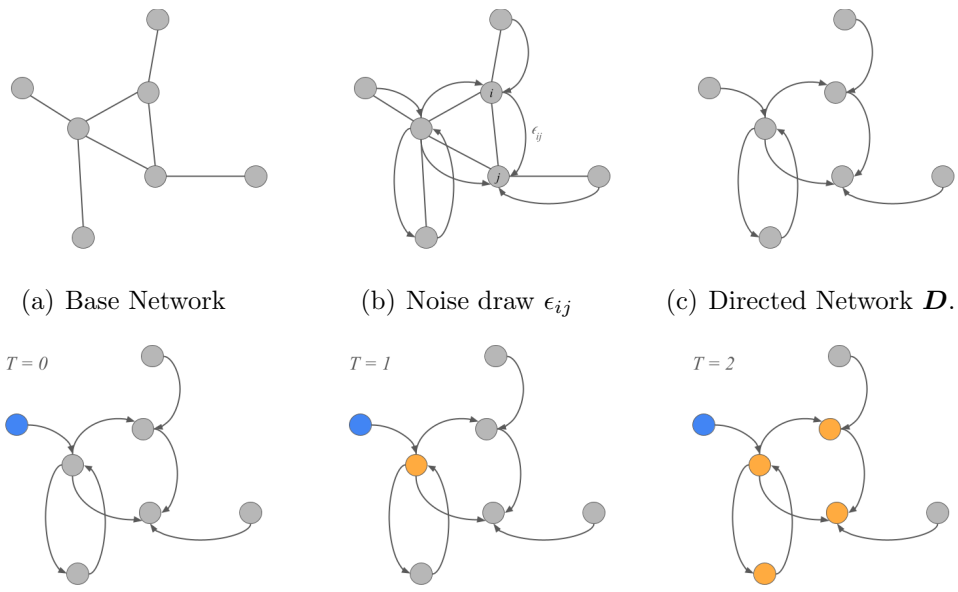
The second set of parameters we consider are the **causal** parameters, those involving the distributions of the counterfactuals. The main causal parameter we will consider is the expected average potential outcome on the complete network G, $\Psi(\boldsymbol{a}|G) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i(\boldsymbol{a})]$, though these can also be made conditional on a covariate x: $\Psi(\boldsymbol{a}|x, G)$. Leveraging the structural causal model, we can define the these causal effects in terms of the structural causal model. We illustrate conditions for identification of these causal effects in Section 3.2.3. While we focus on defining causal quantities through conditional means, the nonparametric identification can also apply to other functionals like quantiles.

Inference for the causal relationship between Y_i and V_i amounts to learning the relationship between Y_i and S_i, V_i. We consider settings where the assignment of treatments can be manipulated by an experimenter, which we discuss in Section 3.4. If one leverages this model, either through assumption or estimation, of the generation of the outcomes then one can use a structural causal model to generate expected potential outcomes under different treatment assignments $f_Y(S_i, V_i, \boldsymbol{\varepsilon}_Y)$, which is precisely what is done in the case of seeding. A contrast of these frameworks is included in the Appendix in Section B.2. The applicability of a model to a new population parallels challenges in distribution shift, as explored in Shimodaira [2000], Wilkins-Reeves et al. [2024].

Adding structure to the potential outcomes model is standard in fields like economics, where researchers often propose models to explain how information or behaviors spread across networks. Many of these models include a temporal element. In our setting, we consider outcomes at a fixed time T, i.e., $Y_i(\boldsymbol{a}) = Y_{i,T}(\boldsymbol{a})$. For instance, Banerjee et al. [2013b] explore a latent diffusion process in micro-lending, Banerjee et al. [2019] study a hearing model for information diffusion, and Beaman et al. [2021] analyze behavior adoption in agriculture through complex contagion. Additionally, Centola and Macy [2007] differentiate the spread of information, often through single links, from behaviors that require multiple neighbors for network propagation. Our framework incorporates linear in means models [Manski, 1993], extending them to identify effects like the global average treatment effect [Chin, 2019].

Example: Contagion as a structural causal model

A foundational model of information diffusion is based on simple contagion, and generalizations of SIR (Susceptible-Infected-Recovered) models on networks Kermack and McKendrick [1927], Giles [1977]. These models have been further studied and extended in various settings [Jackson and Yariv, 2006, Aral et al., 2009, Romero et al., 2011, Chierichetti et al., 2011, Banerjee et al., 2013b, 2019]. Here we illustrate how the base model, under which many extensions are built, can be interpreted as a structural causal model. This interpretation can also be applied to complex contagion settings Centola and Macy [2007], Beaman et al. [2021], Cencetti et al. [2023].



(a) Base Network (b) Noise draw ϵ_{ij} (c) Directed Network $\boldsymbol{D}$.

(d) Contagion process at time period $T = 0$. (e) Time period $T = 1$. (f) Time period $T = 2$.

Figure 3.1: Contagion process where a single node is seeded in time $T = 0$ in blue, and infected nodes displayed in orange at times $T = 1$ and $T = 2$.

Consider a scenario where initially infected (treated) seeds $\boldsymbol{a}$ transmit the infection to each neighbor with probability q at each time-step $t \in 1, 2, \ldots, T$, after which they are no

longer infectious. An infection status at time t is denoted as $Y_{it} = 1$. The overall outcome $Y_i = 1$ indicates whether a node was infected at any time up to T. For a simple case with $T = 2$, we model the transmission using Bernoulli random variables $\epsilon_{ij} \sim \text{Bernoulli}(q)$, representing potential infection from node i to node j. Let $\boldsymbol{E}_{ij} = \epsilon_{ij}$ and $\boldsymbol{D} = \boldsymbol{E} \odot G$. This setup is depicted in Figure 3.1. Given a random sample of the directed graph $\boldsymbol{D}$, one can characterize what would have happened if a node were treated, which is precisely the counterfactual. For instance in Figure 3.1 we seed the left most which proceeds to propagate in steps 1 and 2. Additionally, one can construct the relevant exposure map for any fixed number of time steps T.

Examples of Exposure Maps

We consider several examples of exposure maps, though this list is not exhaustive.

Example (Local Interaction Effects). Simple examples of local network effects may include the total number of treated neighbors, $V_i = \sum_j G_{ij} a_j$, or the treated fraction of one's neighbors $V_i = \sum_j \frac{G_{ij}}{d_i} a_j$, where $d_i = \sum_j G_{ij}$ is degree.

Example (Risk-Sharing Networks [Ambrus et al., 2014]). Equilibrium risk sharing is that the graph consists of C mutually exclusive communities such that any endowment vector within the community is aggregated and shared evenly. Let treatment $\boldsymbol{a}$ be an "endowment" and let $\bar{\boldsymbol{a}}_c = \sum_{j \in c} a_j$ be the sum of the endowment vector for community c, with $|c|$ denoting its size. Then, $V_i = f_V(\boldsymbol{a}, \varphi_i) = \bar{\boldsymbol{a}}_c \cdot |c|^{-1}$. That is, the exposure is just a function of the total endowment of the community and nothing more.

Example (Hearing Information [Banerjee et al., 2019]). Many phenomena, like the spread of diseases, information, or social behaviors, can be effectively modeled as contagion processes. These models show how such phenomena spread through networks [Keeling and Rohani, 2008, Centola and Macy, 2007, Barrat et al., 2008, Pastor-Satorras et al., 2015, Cencetti et al., 2023].

Banerjee et al. [2019] introduces a message-passing model based on such a contagion pro-

cess. The treatments, denoted by $\boldsymbol{a}$, represent a seed piece of information disseminated over a series of time steps, from 1 to T. After T time steps, no further message spreading occurs. We define a "hearing matrix" $\boldsymbol{H}_0$, which calculates the expected number of times person j hears information from person i after T time steps, based on transmission probabilities.

The expected total number of messages that person j hears by time T is represented by V_j (the exposure) which affects their response Y_i through a link function Λ:

$$\mathbb{E}[Y_i|V_i] = \Lambda(\beta_0 + \beta_1 V_i).$$

A common assumption is propose a single transmission probability for each individual q, thus giving structure to the exposure map:

$$V_i = (\boldsymbol{H}\boldsymbol{a})_i \text{ the } i^{th} \text{ element of this vector}$$
$$\text{where } \boldsymbol{H} = \sum_{t=1}^{T} q^t G^t$$

It is straightforward to generalize this to include heterogeneity in the diffusion time steps β_t and illustrate this model in equation (3.2):

$$\mathbb{E}[Y_i|V_i] = \Lambda(\beta_0 + \sum_{t=1}^{T} \beta_t \boldsymbol{e}_i^T G^t \boldsymbol{a}) = h(S_i, V_i; \beta). \tag{3.2}$$

Furthermore, we can relax this model to allow for additional heterogeneity through graph-level statistics S_i, which may include node-level covariates X_i or individual graph-level information such as the degree d_i.

3.2.2 Examples of Partially Measured Network Data

In our setting, we do not have access to the full graph G, but rather, have access to some summarizing function the graph $G^* = \zeta(G)$. Tsiatis [2006] uses the term coarsened data to refer to such partial measurements of missing data in general, not necessarily in the network setting. A non-exhaustive set of examples of partially measured network data include induced subgraphs or egocentric sampling [Freeman, 1982, Almquist, 2012], respondent driven

sampling [Heckathorn, 1997], aggregated relational data [Killworth et al., 1998b], respondent driven sampling [Heckathorn, 1997, Goel and Salganik, 2009, 2010, Green et al., 2020] and more.

Example (Induced subgraph). We sample $m \leq n$ of nodes in the graph randomly, with at least one node from each of the K communities. Let $G^* = G_{I_m, I_m}$ be the sub-graph induced by these m nodes where $I_m \subset \{1, 2, \ldots, n\}$ are the set of nodes that are sub-sampled from the whole population.

Example (Respondent Driven Sampling). Let $i \in I_m \subset \{1, 2, \ldots, n\}$ denote the indices of a sample of individuals obtained through respondent driven sampling. An initial number of individuals are recruited as seeds, and subsequent individuals are recruited via referrals from the others in a population. Under this process we receive a subgraph of connected individuals G_{I_m, I_m} as well as the list of connections to additional nodes $I_{n \setminus m} = \{1, 2, \ldots, n\} \setminus I_m$ $G^* = G_{I_m, I_m}, G_{I_m, I_{n \setminus m}}$.

Example (Aggregated Relational Data). Aggregated relational data consists of aggregated sums of connections to nodes of a given trait. Typically this is collected from a survey consisting of questions such as "How many many people do you know with [X] trait?". For a set of T traits, this consists of $X^*_{it} = \sum_{i=1}^{n} G_{ij} I(t_j = t)$.

In order to infer about the distribution of the missing part of the graph, we propose that $G \sim \theta_0$ where we assume that $\theta_0 \in \Theta$ denotes the parameters of a random graph model. In this case, for each i, there is an a latent ξ_i parameter such that

$$P(G_{ij} = 1 | \xi_i, \xi_j) = \widetilde{g}(\xi_i, \xi_j)$$

for some function symmetric, measurable $\widetilde{g}$ known as a graphon Lovász and Szegedy [2006], Orbanz and Roy [2015]. Many common graph models, such as latent space models Hoff et al. [2002a], **?**], Lubold et al. [2023], Wilkins-Reeves and McCormick [2022], are included in this category. Graphons are appealing in this context because, following Airoldi et al. [2013], Gao et al. [2015], they can be approximated arbitrarily well using latent types assigned to each

node. Said another way, graphons introduce complex dependence in the network-generating mechanism through clustering induced by latent types associated with each node. In our inferential procedures in Section 3.3, the general procedures involve estimation from a missing data perspective. This will involve estimating the graph model $\widehat{\theta} := \widehat{\theta}(G^*)$ then inferring about the distribution $G|G^*, \widehat{\theta}$. Further details for estimating the graph model are included in Section 3.3.4.

3.2.3 Nonparametric Identification of Causal Effects

Our initial aim is to identify the causal parameters without further model specification. These are analogous to standard identification assumptions, adapted to our framework.

Definition 3.2.1 (Exposure Weak Ignorability). We say that an exposure assignment is **weakly ignorable** if the following holds:

$$Y_i(v) \perp\!\!\!\perp \{V_i = v\}|S_i$$

Conditioning on the graph statistic, S_i, ensures that potential outcomes are independent of actual exposure, with all confounding accounted for by S_i. In simple contagion models, nodes are equivalent, and this independence occurs naturally without conditioning. Generally, conditioning should occur on variables that influence outcome heterogeneity. Section 3.5.1 discusses an example from Ugander and Yin [2023] where conditioning on node degree suffices for any randomization.

Definition 3.2.2 (Exposure Consistency). Exposure consistency holds if

$$V_i = v \implies Y_i = Y_i(v) = Y_i(\boldsymbol{a})$$

where $Y_i(v)$ is the potential outcome of individual i for the exposure v.

This assumption can be simply understood as the exposure is correct.

Definition 3.2.3 (Conditional Independence of the Graph and Outcome). We assume that the outcome is conditionally independent of the outcome conditional on the exposure and

the graph generative parameters

$$Y_i(\boldsymbol{a}) \perp\!\!\!\perp G | V_i, S_i.$$

This assumption states that once we have adjusted for V_i and S_i, then the potential outcomes are independent of the network G. Under these assumptions, the causal effects can be identified through conditional distributions of the observed data, allowing for the leveraging of models for estimating causal effects.

$$\begin{aligned}
P(Y_i(\boldsymbol{a}) = y | S_i = s, G) &= P(Y_i(v) = y | S_i = s, G) \text{ By the exposure mapping} \\
&= P(Y_i(v) = y | V_i = v, S_i = s, G) \text{ By weak ignorability} \\
&= P(Y_i = y | V_i = v, S_i = s, G) \text{ By consistency} \\
&= P(Y_i = y | V_i = v, S_i = s) \text{ Graph conditional independence} \\
\implies \Psi(\boldsymbol{a}|G) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i | V_i = f_V(\boldsymbol{a}, \varphi_i), S_i = f_S(\boldsymbol{X}, \vartheta_i)]
\end{aligned}$$

For brevity, we denote the true conditional mean $\mathbb{E}[Y_i | V_i = v, S_i = s] = h_0(s, v)$ and a model class $h(s, v; \beta)$ that will be used to model the outcome $h_0(s, v)$.

Given a network model θ_0, observed graph data G^*, and a conditional model $h(s, v; \beta)$ we can also define the expected average treatment effect

$$\Psi(\boldsymbol{a}|\beta, G^*, \theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[h(f_V(\boldsymbol{a}, \boldsymbol{X}; \varphi_i), f_S(\boldsymbol{X}; \vartheta_i); \beta) | \boldsymbol{a}, \boldsymbol{X}, G^*, \theta] \tag{3.3}$$

where under the correct model conditional model and graph model $\mathbb{E}[\Psi(\boldsymbol{a}|G)|\boldsymbol{a}, \boldsymbol{X}, \theta_0] = \Psi(\boldsymbol{a}|\beta_0, G^*, \theta_0)$. In Appendix B.1.4, we illustrate when this population average effect under any draw of the network $\Psi(\boldsymbol{a}|G)$ will be close to the average over the model class $\Psi(\boldsymbol{a}|\beta_0, G^*, \theta_0)$; allowing for the construction of plug-in estimators $\widehat{\Psi}(\boldsymbol{a}|G) = \Psi(\boldsymbol{a}|\widehat{\beta}, G^*, \widehat{\theta})$.

3.3 Inference

We outline our method for estimating parameters with partial network data. Developing these results requires two theoretical tools: a fast estimation rate for network model parameters θ_0, and a suitable central limit theorem for scenarios with correlated outcomes.

3.3.1 Outcome Model Parameters and Estimators

Next we consider estimating the outcome model parameters β_0. We present two methods for estimating such parameters, instrumentation in a linear model, and Z estimators. The iterated expectation procedure for estimating such parameters was introduced in Chandrasekhar and Lewis [2011], however, we extend inference to the single network setting. Similar approaches exist for peer effects models Boucher and Houndetoungan [2020].

Estimation in Linear Models

We first illustrate identification of the conditional model under a linear model assumption.

$$Y_i = \beta_0^T \widetilde{h}(S_i, V_i) + \varepsilon_i$$

where $\mathbb{E}[\varepsilon_i] = 0$ and there can be general correlation $\text{Var}[\boldsymbol{\varepsilon}] = \Sigma$. Without access to the network data, one can recover the model parameters through conditional expectation

$$\mathbb{E}[\mathbb{E}[Y|S(G), V(G), G, \boldsymbol{a}, \boldsymbol{X}]|\boldsymbol{a}, \boldsymbol{X}, G^*, \theta_0] = \beta_0^T \mathbb{E}[\widetilde{h}(S(G), V(\boldsymbol{a}, G))|\boldsymbol{a}, \boldsymbol{X}, G^*, \theta_0]$$

where we create a new set of features $\widetilde{H}_i = \mathbb{E}[\widetilde{h}(S_i(G), V_i(\boldsymbol{a}, G))|\boldsymbol{a}, \boldsymbol{X}, G^*, \theta_0]$ by averaging over the network model. Here identification comes from the variation of these averaged features $\widetilde{H}_i$ over the population. More clearly, letting $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{n \times p}$ denote the design matrix of this model, identification comes from the linear independence of the columns of $\widetilde{\boldsymbol{H}}$.

Z estimators

In other cases, parameters may be defined through a moment equation, and can be used to construct a Z-estimator. These may include parameters in a moment equation such as generalized linear models (GLMs), $\mathbb{E}[Y|S, V] = \Lambda^{-1}(\beta_0^T \widetilde{h}(S, V))$. These parameters can be identified using an estimating equation approach where given a moment function $\widetilde{m}(Y_i, S_i, V_i; \beta)$ such that $\mathbb{E}[\widetilde{m}(Y_i, S_i, V_i; \beta)|\boldsymbol{a}, \boldsymbol{X}, G] = 0$ if and only if $\beta = \beta_0$. Through the use of iterated

expectations, we can define a new estimating equation, by marginalizing over the draws of the graph model

$$m_i(Y_i; \beta, \boldsymbol{a}, \boldsymbol{X}, G^*, \theta) := \mathbb{E}[\widetilde{m}(Y_i, S_i, V_i; \beta) | Y_i, \boldsymbol{a}, \boldsymbol{X}, G^*, \theta] \tag{3.4}$$

then applying iterated expectations

$$\mathbb{E}[m_i(Y_i, S_i, V_i; \beta_0, \theta_0) | G^*, \boldsymbol{a}, \boldsymbol{X}, \theta_0] = \mathbb{E}[\mathbb{E}[\widetilde{m}(Y_i, S_i, V_i; \beta_0) | \boldsymbol{a}, \boldsymbol{X}, G] | G^*, \boldsymbol{a}, \boldsymbol{X}, \theta_0] = 0.$$

Identification in this case comes from the variation of the exposure and the confounders, such that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n m_i(Y_i, S_i, V_i; \beta, \theta_0) \middle| G^*, \boldsymbol{a}, \boldsymbol{X}, \theta_0\right] = 0$ if and only if $\beta = \beta_0$. Exact conditions depend on the parameter, but GLMs can use a similar identification strategy as linear models.

3.3.2 Inference with partially measured data.

We introduce a general procedure for estimating the outcome model parameters. We also illustrate inference for estimation of a causal target parameter on a particular graph G. We present a pseudo-code approach to the procedure as follows. Let $\widetilde{Z}_i = (Y_i, S_i, V_i)$ denote the full (including unobserved) data, and let $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{a}, \boldsymbol{X}, G^*)$ denote the observed data.

Algorithm 3 Z-estimation overview

1: Define an model for the relationship of $\boldsymbol{Y}$ given the exposures $\boldsymbol{V}$ and confounders $\boldsymbol{S}$ (for instance, a regression model $\mathbb{E}[Y|V,S] = h(v, s; \beta_0), \beta \in \mathcal{B} \subset \mathbb{R}^p$ with parameters which can be estimation via the estimating function $\widetilde{m}_n(\widetilde{\boldsymbol{Z}}; \beta)$. Let $\widetilde{m}_n(\widetilde{\boldsymbol{Z}}; \beta) = \frac{1}{n}\sum_{i=1}^n m(\widetilde{Z}_i; \beta)$ denote the empirical estimating function.

2: Estimate a model of the network, using the node-level covariates $\widehat{\theta} := \widehat{\theta}(G^*)$.

3: Estimate $\widehat{\beta}$ by solving the estimating equation $m_n(\boldsymbol{Y}; \widehat{\beta}, G^*, \widehat{\theta}) = 0$, where $m_n(\boldsymbol{Y}; \widehat{\beta}, G^*, \widehat{\theta}) = \frac{1}{n}\sum_{i=1}^n m_i(Y_i; \beta, \boldsymbol{a}, \boldsymbol{X}, G^*, \theta)$ where m_i is defined in equation (3.4).

4: (optional) Plug in $\widehat{\beta}$ to $\Psi(\boldsymbol{a}|\widehat{\beta}, G^*, \widehat{\theta})$.

Step 1 asks the practitioner to propose a response model given the treatment, i.e. the causal model in Section 3.2.1. Step 2 estimates the generative model given the partial network

data and the node covariates observed. We give theoretical results where the formation model is a stochastic blockmodel, then give rate estimation relative to the more general graphon approach. Step 3 estimates the parameter by marginalizing the estimating function over the graph model. Lastly, Step 4 is optional if the target parameter is a plug-in estimator of the causal parameter using the regression model. We discuss inference for the plug-in estimate of causal parameters using a delta method argument in the Appendix. We next give our asymptotic results, then provide an example of this algorithm in Section 3.5.1.

3.3.3 Asymptotic Results

The asymptotic results for both the Z-estimator and the linear model will depend on being able to establish a central limit theorem based on the exogenous noise. To establish asymptotic properties for our outcomes on a network, we extend the application of the central limit theorem (CLT) to structures not commonly associated with traditional time series or spatial dependencies. Nonetheless when the exogenous noise is correlated, we will need a method of handling the central limit theorem. Specifically, we utilize a general version of the CLT for dependent data from Chandrasekhar et al. [2023]. For brevity in presentation, we leave the full detail of this central limit theorem to the appendix.

We denote $g_i(\boldsymbol{Z};\beta) = m_i(\boldsymbol{Y};\boldsymbol{a},\boldsymbol{X},\beta,G^*,\theta_0)$ to be the moment function evaluated using the true generative model and correspondingly $g_n(\boldsymbol{Z};\beta) = \frac{1}{n}\sum_{i=1}^n g_i(\boldsymbol{Z};\beta)$. Further, define the (normalized) random vector of the estimating function evaluated at the correct model parameters $\mathcal{E}_i = \frac{1}{n} g_i(\boldsymbol{Z};\beta_0)$. And lastly let $D_n(\boldsymbol{Z};\beta_0) = \nabla_\beta g_n(\boldsymbol{Z};\beta_0) \in \mathbb{R}^{p\times p}$ denote the gradient of the estimating equation $g_n(\boldsymbol{Z};\beta)$.

To develop valid inference, we must estimate the graph model quickly enough to disregard the graph estimation component during inference. We will next present the theorem and discuss the assumptions further.

Assumption 3.3.1 (Regularity Conditions for Z-Estimation)**.** *Suppose the following conditions hold for all n.*

Consistency for a Z estimator

A1. $\mathbb{E}[g_n(\boldsymbol{Z};\beta)] = 0$ *for* $\beta = \beta_0$ *and for all* $\epsilon > 0$, $inf_{||\beta-\beta_0||>\epsilon}\mathbb{E}[g_n(\boldsymbol{Z};\beta)] > 0$

A2. $\sup_{\beta\in\mathcal{B}}\left|\left(\frac{\partial}{\partial\beta}\right)^l g_n(\boldsymbol{Z};\beta) - \left(\frac{\partial}{\partial\beta}\right)^l \mathbb{E}[g_n(\boldsymbol{Z};\beta)]\right| = o_P(1)$ *for* $l \in \{0,1,2\}$

Graph Model Regularity conditions

B1. $\widehat{\theta}$ *is an* $s(n)$*-consistent estimate of the graph parameters* $||\widehat{\theta} - \theta_0|| = o_P(s(n))$

B2. $\sup_{\beta\in\mathcal{B}}|m_n(\boldsymbol{Z};\beta,\theta) - m_n(\boldsymbol{Z};\beta,\theta')| \leq b_n(\boldsymbol{Z})||\theta - \theta'||$ *where* $b_n(\boldsymbol{Z}) = O_P(1)$ *(that is,* $b_n(\boldsymbol{Z})$ *is stochastically bounded).*

Central Limit Theorem (CLT)

C1. The random vectors $\mathcal{E}_{1:n}$ *satisfy the affinity set conditions of Chandrasekhar et al. [2023] (restated as Theorem B.1.2 in the appendix) with corresponding covariance matrix* $\Gamma_n = Var[\sum_{i=1}^n \mathcal{E}_i]$. *Where* $r(n) := \sqrt{\lambda_{min}(\Gamma_n)}$.

Theorem 3.3.2 (Single Network Z-estimator Asymptotics)**.** *Suppose that Assumptions 3.3.1 hold and that* $s(n) = o(r(n))$. *Then:*

$$\Gamma_n^{-1/2} D(\beta_0)(\widehat{\beta} - \beta) \to_d N(0, I_p) \tag{3.5}$$

Where $\mathbb{E}[\nabla_\beta g_n(\boldsymbol{Z};\beta)|\boldsymbol{a},\boldsymbol{X},G^*,\theta_0]\Big|_{\beta=\beta_0} = D(\beta_0)$

The first set of assumptions ensures the consistency of Z-estimators, typically derived from uniform laws of large numbers as discussed in Andrews [1987] or Newey and McFadden [1994]. The second set involves conditions that make the graph model's estimation negligible, requiring the estimating functions to be smooth with respect to the graph parameters.

The final set of assumptions, stated in C1, are utilized so that $\mathcal{E}_i$ satisfy a central limit theorem Chandrasekhar et al. [2023]. This assumption is required if the data exhibit further dependence after controlling for graph parameters (if, for example, there are latent factors

that impact both outcomes and the propensity to form ties). The main idea of Chandrasekhar et al. [2023] is to represent dependence in terms of "affinity sets" where the majority of dependence structure is captured within sets, leaving little between sets. In the modelling of social behaviours beyond just considering outcomes as a function of the exposure observed, outcomes may be further correlated, beyond examples of spatial dependence or heteroskedasticity. In practice we can include these dependencies through correlation terms matching the generative graph model, such as between blocks of a stochastic blockmodel or via latent positions in a latent space model.

Under conditions which we give explicitly in the Appendix, Chandrasekhar et al. [2023] derive a CLT which we can apply here to extend our results to settings with dependent observations. Here, $r(n)$ describes the effective rate at which the variance converges. For the estimation of the graph model θ_0 to be considered negligible, it must occur more rapidly than $r(n)$. In cases of independent or minimally dependent noise, it is typical for $r(n) \approx n^{-1/2}$. Alternatively, in different scenarios, $\mathcal{E}_i$ might exhibit correlation within densely connected blocks of the network, such as during a diffusion process in a stochastic blockmodel with k_n densely linked blocks (refer to Chandrasekhar et al. [2023], section 4.4, for detailed descriptions). In such cases, $r(n)$ is generally on the order of $k_n^{-1/2}$. If both $r(n)$ and $s(n)$ approach zero, yet the ratio $\frac{s(n)}{r(n)}$ diverges or stabilizes at a nonzero constant, it remains possible to obtain a consistent estimator for the parameters of the outcome model. However, its asymptotic distribution may be dominated by how the graph model is estimated, thus requiring a tailored inference approach based on the chosen graph model estimation technique.

An analogous argument follows when conducting inference using a linear model. For the sake of brevity and avoiding repetition, we include it in the Appendix in Section B.1.2. In Theorem 3.3.3 we present a summary.

Theorem 3.3.3. *Let* $\widetilde{H}_i(\theta) = \mathbb{E}[\widetilde{h}(S_i(G), V_i(G))|\boldsymbol{a}, \boldsymbol{X}, G^*; \theta]$. *The OLS estimator uses the model averaged coefficients* $\widetilde{H}_i(\theta)$ *in place of the true unobserved coefficients* $\widetilde{h}_i$. *Let* $\mathsf{H}_n(\theta) =$

$\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\theta)\widetilde{H}_i^T(\theta)$. *Given an estimate of the model parameters* $\widehat{\theta}$*, we define the*

$$\widehat{\beta}_{ols} = \mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})Y_i$$

Let $u_i = (\widetilde{h}(S_i(G), V_i(G)) - \widetilde{H}_i(\theta_0))\beta_0 + \epsilon_i$ *and let* $\Gamma_n = Var\left[\frac{1}{n}\sum_{i=1}^{n} u_i\right]$ *Suppose the conditions of Theorem B.1.3 in the Appendix hold. Then*

$$\Gamma_n^{-1/2}\mathsf{H}_n(\widehat{\theta})(\widehat{\beta}_{ols} - \beta_0) \to_d N(0, I_p)$$

3.3.4 Network Model Estimation

We next discuss the estimation of the generative model for the network using a variety of data types. Breza et al. [2020] and Breza et al. [2023] consistently estimate a generative model for ARD with mutually exclusive traits. We extend this work in several ways. We introduce a novel method for estimating the stochastic blockmodel with non-mutually exclusive traits using constrained least squares approach as well as give approximation rates for graphons. We summarize the resulting rates using ARD for a variety of model classes in the appendix in Table B.1.3.

Additionally, through techniques previously existed for estimating stochastic blockmodels using aggregated relational data, this required mutually distinct traits Breza et al. [2023]. Instead we introduce method for estimating the stochastic blockmodel without requiring that the surveyed traits of the alters be mutually distinct, which greatly improves the feasibility of the method to be applied using standard ARD surveys.

In the main text, we concentrate on estimating the stochastic blockmodel using ARD. In the Appendix in Section B.1.3 we estimate generative models using partial network data such as subgraph sampling and develop similar rates for the stochastic blockmodel for subgraph sampling and reference a similar result for respondent driven sampling.

SBM Estimation with ARD

Recall that X_{it}^* represents a set of ARD response vectors. Breza et al. [2023] show that we can consistently estimate the connection probabilities between latent types, however, we present an improved version of the SBM estimator which allows for an non-mutually exclusive traits. Let n_t denote the total number of individuals of trait type t. Let N_k' denote the nodes in our sample in group k, and let n_k denote the number of nodes in the graph in group k. We cluster the node memberships according to Algorithm 4.

Algorithm 4 ARD SBM clustering procedure

1: Count the number of individuals with each trait n_t
2: Denote the normalized ARD responses $X_{it}^\dagger = X_{it}^*/n_t$.
3: Cluster the normalized ARD response vectors $\{X_i^\dagger\}_{i=1}^T$ into K groups using hierarchical agglomerative clustering into a set of clusters $\widehat{k}_i \in \{1, 2, \ldots, K\}$

After we obtain a clustering, we can estimate the stochastic blockmodel. Let $\widehat{\omega}_{kt} = \widehat{N}_{kt}/N_t$ where N_{kt} are the number of traits in the estimated group k and with trait t, and N_t are the number of individuals with trait t, and $\omega_{kt} = N_{kt}/N_t$, the analogous population quantity. We next define the probability matrix of observing a connection of group k with a trait t. $\widetilde{\mathbf{P}}_{kt} = \sum_{k'} \mathbf{P}_{kk'}\omega_{k't}$, where $\widetilde{\mathbf{P}}_{kt} = P(G_{ij} = 1 | k_j = k, t_i = t)$. This relationship can be expressed in a linear system $\widetilde{\mathbf{P}} = \Omega\mathbf{P}$ where $\Omega \in \mathbb{R}^{T\times K}$ and $\Omega_{kt} = \omega_{kt}$. If Ω is of full column rank, then a unique solution will exist. Given an estimate of the latent communities, one can estimate $\widehat{\Omega}$.

$$\widehat{\mathbf{P}}_{kk'} = \left(\widehat{\Omega}^\intercal\widehat{\Omega}\right)^{-1}\widehat{\Omega}^\intercal\widehat{\widetilde{\mathbf{P}}} \quad \text{where } \widehat{\widetilde{\mathbf{P}}}_{kt} = \frac{1}{n_k n_t}\sum_{i\in\widehat{N}_k} X_{it}^*.$$

In general, one can symmetrize $\widehat{\mathbf{P}}_{kk'}$ after the estimate to ensure the constraints of an undirected stochastic blockmodel are satisfied. Alternatively, once can also minimize the constrained least squares objective which can be implemented using standard convex solvers

such as `CVX` [Fu et al., 2020]

$$\widehat{\boldsymbol{P}} = \underset{0 \leq \boldsymbol{P} \leq 1 : \boldsymbol{P} = \boldsymbol{P}^\intercal}{\arg\min} \sum_{i=1}^{n} \sum_{t=1}^{T} (\widetilde{X}_{it} - \sum_{k'} \widehat{\Omega}_{k't} P_{k',k_i})^2.$$

Breza et al. [2023] develop a procedure for consistently estimating the stochastic blockmodel, but we extend their result and obtain a rate on the estimation of the model parameters in Lemma 3.3.4 and relax the assumption that groups are mutually exclusive. We differentiate between the cross-group probabilities in which the clusters that are estimated $\mathbf{P}^{(\widehat{\boldsymbol{k}})}$ with the cross-group probabilities under known membership $\mathbf{P}^{(\boldsymbol{k})}$.

Lemma 3.3.4. *Suppose that we use the clustering strategy outlined in Section 3.3.4 to cluster the observations based on aggregated relational data. Let* $Z_k = (\widetilde{\mathbf{P}}_{k1}, \ldots \widetilde{\mathbf{P}}_{kT})$ *and* $\widetilde{\mathbf{P}}_{kt} = P(G_{ij} = 1 | k_i = k, t_j = t)$. *Assume also that* $\inf_{k,k'} ||Z_k - Z_k||_2 > 0$ *and that* $T \geq K$ *where* T *is the number of discrete traits asked about and* K *is the true number of clusters.*

Let $\widehat{\boldsymbol{k}}$ *denote the estimated cluster memberships and let* $\widehat{\mathbf{P}}^{(\widehat{\boldsymbol{k}})}$ *be the corresponding estimate of the cross block probabilities. Let* $\Omega_{kt} = N_{kt}/N_t$ *denote the matrix which involves the fraction of the individuals in cluster* k *who also have trait* t, *and* $\widehat{\Omega}$ *the estimated counterpart based on membership clusters. Let* $C_\Omega = \lambda_{max}((\Omega^T \Omega)^{-1})$ *and* $\lambda_{max}(\cdot)$ *denotes the largest eigenvalue of a matrix and* $C_\Omega > 0$.

Then with probability at least $1 - \delta - \frac{1}{n}$

$$||\widehat{\mathbf{P}}^{(\widehat{\boldsymbol{k}})} - \mathbf{P}^{(\boldsymbol{k})}||_1 \leq C_\Omega \frac{KT}{n} \sqrt{\frac{\log(2/\delta) \log(KT)}{2}}$$

We contrast our results to the optimal estimation rate for a stochastic blockmodel reported in Gao et al. [2015] as $\widetilde{O}_P(n^{-1/2})$. While our rate initially seems faster, a key difference stems from the complexity of our clustering problem compared to theirs. Our clustering is aided by node-level observed traits providing extra information. As the network expands, the normalized ARD vector for each individual converges to its mean, making the clustering progressively easier, hence the faster rate.

Misspecification of the Graph Model

We use a stochastic blockmodel as it effectively approximates a general graphon class. Even if θ_0 belongs to a smooth graphon class rather than a stochastic blockmodel, we can still bound the bias in estimating the relevant model parameters. Consider a scenario where edges are generated under a true graphon model. A graphon is a function $\widetilde{g} : [0,1]^2 \to [0,1]$ that assigns pairwise conditions based on a sample of $[0,1]$ random variables.

$$\eta_{ij} = \widetilde{g}(\xi_i, \xi_j) = P(G_{ij} = 1|\boldsymbol{\xi}) \quad \text{where } \boldsymbol{\xi} \sim_{iid} P_{\boldsymbol{\xi}} \in [0,1].$$

Let $\mathcal{H}_\alpha(M)$ denote a smooth graphon class defined via the α-M-Hölder class as follows. Let $\mathcal{D} = [0,1]^2 \cap x \leq y$ denote the domain of (x,y). We define the norm $||\widetilde{g}||_{\mathcal{H}_\alpha}$ as:

$$||\widetilde{g}||_{\mathcal{H}_\alpha} = \max_{j+k \leq \lfloor \alpha \rfloor} \sup_{x,y \in \mathcal{D}} |\nabla_{jk} \widetilde{g}(x,y)| + \max_{j+k = \lfloor \alpha \rfloor} \sup_{(x,y) \neq (x'y') \in \mathcal{D}} \frac{\nabla_{jk}\widetilde{g}(x,y) - \nabla_{jk}\widetilde{g}(x',y')}{(|x-x'| + |y-y'|)^{\alpha - \lfloor \alpha \rfloor}}$$

and the Hölder class corresponding to this norm as

$$\mathcal{H}_\alpha(M) = \{||\widetilde{g}||_{\mathcal{H}_\alpha} \leq M : \widetilde{g}(x,y) = \widetilde{g}(y,x); 0 \leq \widetilde{g}(x,y) \leq 1\}.$$

Prior work has focused on the approximability of a stochastic blockmodel to any element of a smooth graphon class. In particular there will always be some assignment of block memberships such that we can bound the 2-norm probability deviation from the true model.

Lemma 3.3.5. *Suppose that θ_* corresponds to a true graphon model and θ_0 a corresponding approximating stochastic blockmodel satisfying the conditions of B.3.1. Denote the population estimating function, as a function of the model parameters*

$$L_n(\beta, \theta) = \mathbb{E}[\widetilde{m}_n(\widetilde{\boldsymbol{Z}}; \beta)|\boldsymbol{a}, \boldsymbol{X}, \theta]$$

where $L_n(\beta_0, \eta_0) = 0$ defines the population parameter β_0 under the misspecified model θ_0, and let $L_n(\beta_*, \theta_*) = 0$ define the population solution β_* to the correctly specified graph model θ_*. Let η_0 and η_* be the pairwise edge probabilities corresponding to the models θ_0, θ_* respectively. Finally assume that:

F1. $\mathcal{B}$ *is compact*

F2. $\sup_{\beta\in\mathcal{B}} |L_n(\beta,\eta) - L_n(\beta,\eta_*)| \leq L||\eta-\eta_*||_2/n$

F3. $\min_j \frac{\partial}{\partial\beta_j} L_n(\beta,\eta_*)\Big|_{\beta=\beta_*} = \lambda > 0$

Then the approximation error under the graph misspecification is bounded by the rate:

$$||\beta_0 - \beta_*|| = O(\lambda^{-1}K^{-\alpha\wedge 1}) \quad \textit{where } a\wedge b = min(a,b). \tag{3.6}$$

In practice, since we do not directly select clusters, but rather the misspecified clusters are related to the observed traits, we cannot guarantee achieving this bound. However, this is a worst-case bound, and in fact, may be overly conservative to the bias that we observe in practice. This therefore suggests a possibility of future work that involves the sensitivity analysis of both the response function and the latent graph model.

3.4 Experimental Design

Our focus so far has been on estimating model parameters given a treatment assignment $\boldsymbol{a}$. We now explore experimental design methods that leverage partial network data when determining $\boldsymbol{a}$ to reduce the variance of our estimands.

We consider saturation randomization experiments, which divide the dataset into J clusters of size n_j. A proportion τ_j of each cluster is assigned the treatment, totaling $n_t = \sum_{j=1}^{J} \tau_j n_j$, and generally will not be the same "blocks" as those in a graph model, for example a stochastic blockmodel. Practically, due to budget constraints, the set of possible saturation levels $\boldsymbol{\tau}$ is limited to $\mathcal{T} \subset [0,1]^J$. For example, this could be due to limited resources like a finite number of vouchers in a vaccine trial.

3.4.1 Bayesian Optimization of Asymptotic Regression Estimators

Our goal is to optimize the asymptotic variance of a function the model parameter $\widehat{\beta}$ in Section 3.3. We highlight this by optimizing the variance of the estimates of linear contrasts

of the parameters $\phi^T\beta$. When using the stochastic block model for the network model these treatment blocks could align with the model blocks, however this need not be the case. They could, instead, be based on observed characteristics (e.g. geography, classrooms).

Denote the variance of the target contrast parameter conditional on the treatment assignment as $\boldsymbol{a}$: $\upsilon^{\phi}(\boldsymbol{a};\theta) = \text{Var}(\phi^T\widehat{\beta}|\boldsymbol{a},\theta)$. Ideally, the goal is to find a treatment assignment a^* that minimizes the variance of the contrast: $a^* = \arg\min_{\boldsymbol{a}\in\{0,1\}^n} \upsilon^{\phi}(\boldsymbol{a};\theta)$.

Without added structure, optimizing treatment assignments is NP-hard, requiring a search over 2^n possible assignments. By changing the objective to one where we optimize over a set of saturation levels over a set of groups $\boldsymbol{\tau} \in [0,1]^J$, we simplify the problem so that it is no longer NP-hard in the sample complexity and is therefore tractable. The distribution of treatment assignments, $\boldsymbol{a}$, under $\boldsymbol{\tau}$ is denoted by $P_{\boldsymbol{\tau}}$, and we aim to minimize:

$$\text{Var}(\boldsymbol{\tau};\theta) = \mathbb{E}_{\boldsymbol{a}\sim P_{\boldsymbol{\tau}}}[\upsilon^{\phi}(\boldsymbol{a};\theta_0)].$$

In Algorithm 5, we present a method for evaluating the variance of a linear model using a generic feature map $\widetilde{h}$ for a given treatment assignment $\boldsymbol{a}$ and a graph model θ. A general approach for Z-estimators is detailed in the Appendix. Algorithm 5 operates under specific assumptions about the covariance matrix Σ, which may include correlations within densely connected network components. We will present our algorithm for minimizing this variance using Bayesian optimization, which accounts for the uncertainty in the outcome, given a graph model. In the appendix we give an extension which also which incorporates network model uncertainty $\widehat{\theta}$ (Section B.1.5).

Bayesian Optimization. Calculating the average variance $\text{Var}(\boldsymbol{\tau};\widehat{\theta})$ in Algorithm 5 is computationally intensive to evaluate and often non-convex. Since the number of cluster saturation tends to be relatively small, this suggests that Bayesian optimization is an appropriate method for minimizing this saturation variance.

We provide a description of our procedure using a Bayesian optimization procedure for variance reduction. Let $\text{Var}(\tau) := \text{Var}(\tau;\widehat{\theta})$ denote our objective function of the variance evaluated using an estimate of the network model $\widehat{\theta}$. Given a set of pilot points

Algorithm 5 Saturation Randomized Design Variance.

Inputs: Variance structure $\text{Var}[\boldsymbol{u}] = \Sigma$, Model estimate $\widehat{\theta}$.

Sample L draws from the graph model $\{\widehat{G}^{(l)}\}_{l=1}^{L} \sim \widehat{\theta}|G^*$

Sample R treatments $\{\boldsymbol{a}_r\}_{r=1}^{R}$ according to the block saturation levels $\boldsymbol{\tau}$.

for $r \leftarrow 1$ **to** R **do**

Compute the averaged features over draws from the graph model $\{\widehat{G}^{(l)}\}_{l=1}^{L}$,

$$\widehat{H}_{ir}(\boldsymbol{a}) = \frac{1}{L}\sum_{l=1}^{L} \widetilde{h}(S_i(\widehat{G}^{(l)})V_i(\boldsymbol{a}_r; \widehat{G}^{(l)}))$$

Compute the Hessian $\widehat{\mathsf{H}}_n(\boldsymbol{a}_r) = \frac{1}{n}\sum_{i=1}^{n} \widehat{H}_{ir}(\boldsymbol{a})\widehat{H}_{ir}^T(\boldsymbol{a})$.

Compute the design matrix $\widehat{H}_r^T(\boldsymbol{a}) \in \mathbb{R}^{n\times p}$ where each row is $\widehat{H}_{ir}(\boldsymbol{a})$.

Compute the variance for a single draw of the treatment vector $\boldsymbol{a}_r$:

$$\upsilon^{\phi}(\boldsymbol{a}_r; \widehat{\theta}) = \phi^T \widehat{\mathsf{H}}_n^{-1}(\boldsymbol{a}_r)\widehat{H}_r^T(\boldsymbol{a})\Sigma\widehat{H}_r(\boldsymbol{a})\widehat{\mathsf{H}}_n^{-1}(\boldsymbol{a}_r)\phi$$

end for

Average over each of the draws $\text{Var}(\boldsymbol{\tau}; \widehat{\theta}) = \sum_{r=1}^{R} \upsilon^{\phi}(\boldsymbol{a}_r; \widehat{\theta})$

$\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_{n_0}$. we propose a Gaussian process prior satisfying

$$\text{Var}(\boldsymbol{\tau}_{1:n_0}) \sim N(\mu_0(\boldsymbol{\tau}_{1:n_0}, \Sigma_0(\boldsymbol{\tau}_{1:n_0}, \boldsymbol{\tau}_{1:n_0})))$$

where $\text{Cov}[\text{Var}(\boldsymbol{\tau}_i), \text{Var}(\boldsymbol{\tau}_j)] = \Sigma_0(\boldsymbol{\tau}_i, \boldsymbol{\tau}_j)$ where Σ_0 is a positive semidefinite kernel function. As a default, we use the Gaussian kernel $\Sigma_0(x, x') = \alpha_0 \exp(-||x - x'||^2)$. We can then use this prior to define a posterior over remainder of the design space $\mathcal{T}$

$$\begin{aligned}
\text{Var}(\boldsymbol{\tau})|\text{Var}(\boldsymbol{\tau}_{1:n_0}) &\sim N(\mu_n(\boldsymbol{\tau}), \sigma_n^2(\boldsymbol{\tau})) \\
\mu_n(\boldsymbol{\tau}) &= \Sigma_0(\boldsymbol{\tau}, \boldsymbol{\tau}_{1:n_0})\Sigma_0(\boldsymbol{\tau}_{1:n_0}, \boldsymbol{\tau}_{1:n_0})^{-1}(\text{Var}(\boldsymbol{\tau}) - \mu_0(\boldsymbol{\tau}_{1:n_0})) + \mu_0(\boldsymbol{\tau}) \\
\sigma_n^2(\boldsymbol{\tau}) &= \Sigma_0(\boldsymbol{\tau}, \boldsymbol{\tau}) - \Sigma_0(\boldsymbol{\tau}, \boldsymbol{\tau}_{1:n_0})\Sigma_0(\boldsymbol{\tau}_{1:n_0}, \boldsymbol{\tau}_{1:n_0})^{-1}\Sigma_0(\boldsymbol{\tau}_{1:n_0}, \boldsymbol{\tau}).
\end{aligned}$$

From this posterior, we define an acquisition function $A(\boldsymbol{\tau})$. As a default, we choose the upper confidence bound (UCB) acquisition function $A(\boldsymbol{\tau}) = \mu_n(\boldsymbol{\tau}) - \kappa\sigma_n(\boldsymbol{\tau})$ for a chosen κ (where we set $\kappa = 2$). This method is implemented in the `R` package `rBayesianOptimization`, which uses `GPfit` [R Core Team, 2021, Yan, 2021, MacDonald et al., 2015]. For a detailed review of Bayesian optimization techniques, refer to Frazier [2018]. We evaluate the complete Bayesian optimization procedure in Algorithm 6, where we apply the procedure for N_0 iterations.

Algorithm 6 Bayesian Optimization Procedure

1: **Inputs:** Graph model $\widehat{\theta}$ and partial graph information G^*. Kernel function Σ_0.
2: Sample $\boldsymbol{\tau}_{1:n_0}$ uniformly from $\mathcal{T}$, as a pilot sample of the design points.
3: Update the posterior on $\text{Var}(\boldsymbol{\tau})$.
4: **for** $i \leftarrow n_0 + 1$ **to** $n_0 + N_0$ **do**
5: Update the posterior on $\text{Var}(\boldsymbol{\tau})|\text{Var}(\boldsymbol{\tau}_{1:(i-1)})$.
6: Let $\boldsymbol{\tau}_i$ be the minimizer of the acquisition function $A(\boldsymbol{\tau})$ (UCB).
7: Evaluate $\text{Var}(\boldsymbol{\tau}_i)$ using Algorithm 5.
8: **end for**
9: Return the point $\boldsymbol{\tau}_{1:(n_0+N_0)}$ with the smallest $\text{Var}(\boldsymbol{\tau})$

The quality of optimization over N_0 iterations depends on the smoothness of $\text{Var}(\boldsymbol{\tau})$. Since variance might diverge under some settings (e.g., as $\boldsymbol{\tau} \to 0$), a simple alternative is to maximize $\exp(-\text{Var}(\boldsymbol{\tau}))$ instead. The closeness of the maximizer after N_0 iterations hinges on the smoothness of $\exp(-\text{Var}(\boldsymbol{\tau}))$, which we assume belongs to a reproducing kernel Hilbert space, $\mathcal{H}$, with a bounded kernel $\Sigma_0(x, x') \leq B$. This function's smoothness affects the approximation rate, detailed in Srinivas et al. [2009]. For instance, with Gaussian kernel Σ_0, the approximation error is $\exp(-\text{Var}(\boldsymbol{\tau}^*)) \geq \frac{1}{N_0}\sum_{m=1}^{N_0}\exp(-\text{Var}(\boldsymbol{\tau}_m)) + O_P(\frac{B\sqrt{\log(N_0)^{K+1}}+\log(N_0)^{K+1}}{\sqrt{N_0}})$. Similar findings apply to Matern and linear kernels per Srinivas et al. [2009].

3.4.2 Designs for Optimal Seeding

Given a model of the potential outcomes, we may also leverage this model for optimal seeding, a task that is NP-hard [Kempe et al., 2003] in general. Many contagion models are exchangeable given an exposure, and with only block information available, then we can reduce our search space to that over block saturation. In our case, where exact network structures are unknown, we determine the optimal blocks for seeding. When $K \ll n$, this structure significantly reduces computational efforts, and we only need to decide how many seeds to allocate to each of the K clusters.

The model leveraged for the outcome $f_Y(V_i, S_i, \boldsymbol{\varepsilon}_Y)$ could be a predefined model based on domain knowledge, such as complex contagion used by Beaman et al. [2021]. In other scenarios, this might be estimated (e.g., simulation using $f_Y(V_i, S_i, \boldsymbol{\varepsilon}_Y; \widehat{\beta})$ in place of $f_Y(V_i, S_i, \boldsymbol{\varepsilon}_Y)$). This is demonstrated in Algorithm 7 (line 5).

When the total number of seeds Algorithm 7 (line 3) is constrained to be small, then it is computationally feasible to implement exactly; however, we could also use a Bayesian optimization procedure if we wanted to control the treatment over saturation levels.

Algorithm 7 Optimal Seeding With Partial Network Data

Inputs: Number of seeds N, Model estimate $\widehat{\theta}$, number of graph draws L.
Sample L draws from the graph model $\{\widehat{G}^{(l)}\}_{l=1}^{L} \sim \widehat{\theta}|G^*$
for $\tau \in \mathcal{T}$ **do**
    Sample L treatments $\{\boldsymbol{a}_l\}_{l=1}^{L}$ according to the block saturation levels $\boldsymbol{\tau}$.
    Compute the outcomes $Y_i^{(l,\boldsymbol{a}_l)}$ according to the outcome model $f_Y(V_i, S_i, \boldsymbol{\varepsilon}_Y)$.
    Compute the average (and standard error) over draws of the network $\bar{Y}^{(\boldsymbol{\tau})} = \frac{1}{L}\sum_{l=1}^{L} Y_i^{(l,\boldsymbol{a}_l)}$
end for
Return saturation level $\boldsymbol{\tau}$ with the largest value of $\bar{Y}^{(\boldsymbol{\tau})}$.

3.5 Data Analysis

In this section, we present three empirical examples to illustrate our framework's utility in estimating causal effects, designing experiments, and implementing seeding strategies. We adopt a semi-synthetic approach in our examples, where the outcomes are simulated based on processes derived from real networks. The networks analyzed pertain to observational and experimental studies focused on information diffusion in rural villages in India and Malawi, as discussed in [Banerjee et al., 2013b, 2019, Beaman et al., 2021]. These networks consist of 30-400 households per village. To ensure continuity across the examples, we generate ARD as the partial data type and model the networks using stochastic blockmodels for each case, however the use of other network generative models and partial network datatype are applicable in these cases.

When covariates are available for all nodes, we use them to construct ARD. If covariates are missing, we apply the Leiden algorithm Traag et al. [2019] in `igraph` Csardi and Nepusz [2006] to cluster the network and treat these clusters as traits. Table 3.1 details which datasets used actual traits versus clustering to manage trait numbers in our simulations.

Network Dataset	Traits
Banerjee et al. [2013b]	Leiden Cluster $K \in [4, 16]$
Banerjee et al. [2019]	Observed Traits (Section B.6.2)
Beaman et al. [2021]	Leiden Cluster K = 8

Table 3.1: Summary of synthetic traits vs. real traits in the simulation and real data analysis settings.

3.5.1 Causal Effect Estimation

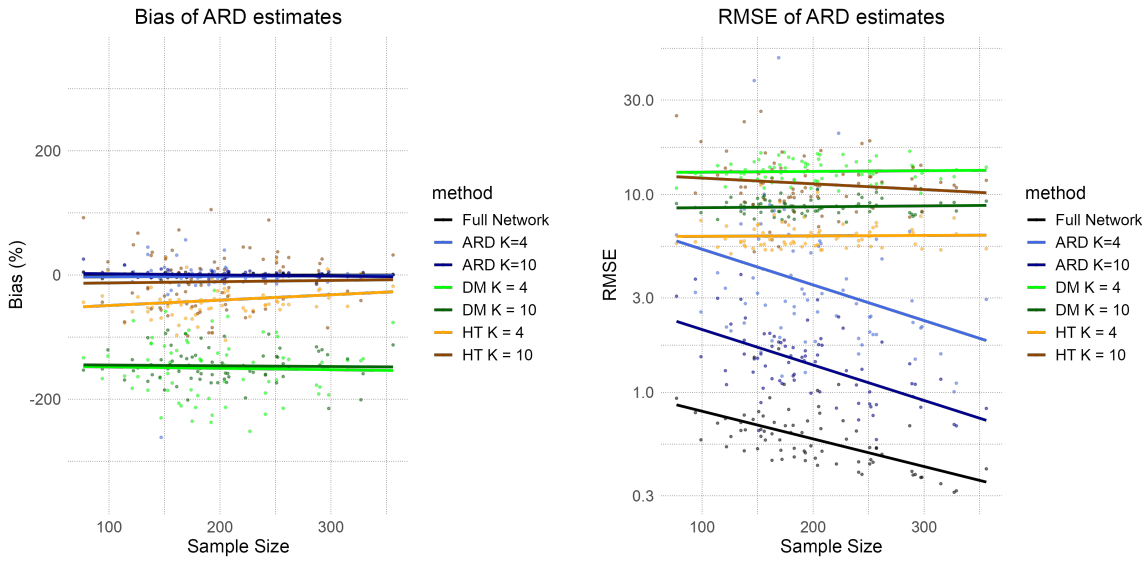
In this example, we aim to estimate the global average treatment. We consider the example from Ugander and Yin [2023] and generate a set of potential outcomes according to the following model

$$Y_i(\mathbf{0}) = \frac{d_i}{\bar{d}} \cdot (\alpha + bX_i + \sigma\epsilon_i), \quad Y_i(\boldsymbol{a}) = Y_i(\mathbf{0}) \cdot \left(1 + \delta a_i + \gamma \frac{\sum_{j\in[n]} G_{ij}a_j}{d_i}\right)$$

where $\epsilon_i \sim_{iid} N(0,1)$ is some independent noise, and X_i is a covariate that varies throughout the network, d_i is the degree of individual i and $\bar{d}$ is the average degree across the network. We set $\alpha = 1$, $b = 1$, $\delta = 1$, $\sigma = 0.5$ and $\gamma = -0.5$. The global average treatment effect in this model is $\frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{0})(\delta + \gamma) = \Psi(\boldsymbol{a} = 1|G) - \Psi(\boldsymbol{a} = 0|G)$. The exposure is the individual treatment in conjunction with the average treatment of neighbors, and the graph confounder include the degree ratio and node level covariates

$$f_V(\boldsymbol{a}; \varphi_i(G)) = \left(a_i, \frac{\sum_{j\in[n]} G_{ij}a_j}{\bar{d}}\right), \quad f_S(\boldsymbol{X}; \vartheta_i(G)) = \left(\frac{d_i}{\bar{d}}, X_i\right).$$

We evaluate the effectiveness of graph cluster randomization by comparing a Horvitz-Thompson estimator Ugander et al. [2013] to a difference in means estimator under a cluster randomized design. In this design, half of the clusters receive no treatment (saturation of 0) and the other half receive full treatment (saturation of 1). We vary the number of clusters from 4 to 16 but display results only for 4 and 10 clusters in Figure 3.2 for clarity.



(a) Bias (b) RMSE

Figure 3.2: Comparison of GATE estimators. ARD denotes our method using aggregated relational data. The "Full Network" method uses a regression approach with the full data available. DM is the difference in means and HT is the Horvitz-Thompson estimator.

Figure 3.2 shows that the full data regression model performs the best, as it leverages more information than the ARD approaches. However, the ARD version still effectively minimizes bias (Figure 3.2(a)) and RMSE (Figure 3.2(b)). In our simulations of dense graphs with few clusters, the Horvitz-Thompson Estimator faces challenges as the network grows—almost all nodes have at least one neighbor with a treatment different than their own. The difference in means estimator shows consistent bias, due to not using heterogeneous covariate information. While regression with complete data is most effective, using partial network data still yields comparably good results.

3.5.2 Experimental Design

We next highlight aspects of experimental design using an information diffusion example based on the hearing model referenced in Section 3.2.1. At each time step the previously

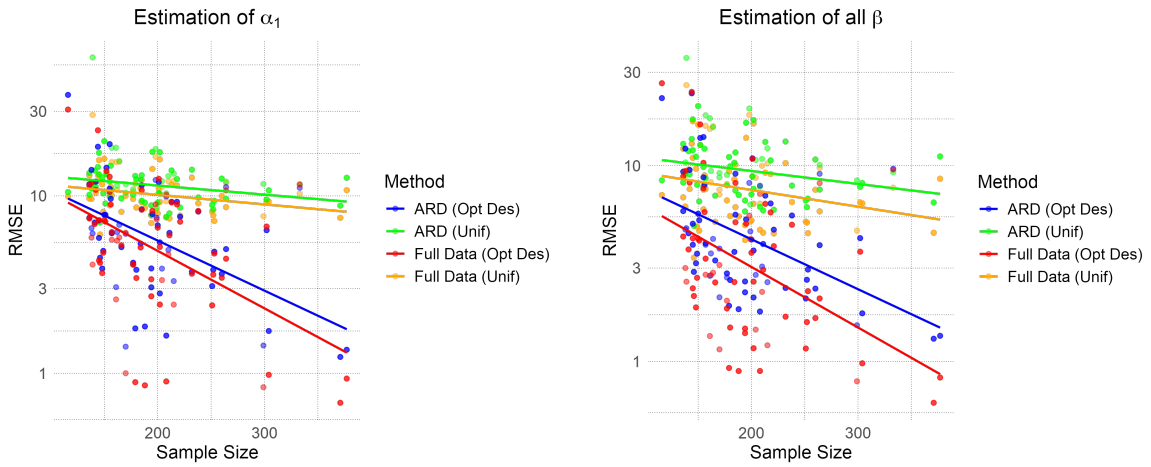
infected nodes are susceptible again the nodes infected in the last round will infect their neighbors with probability q_{t+1}. We repeat this for $T = 3$ rounds. Let N_i denote the total number of infections after the process. We then sample some binary response $P(Y_i = 1|N_i) = \text{logit}(\alpha_0 + \alpha_1 N_i)$ where α_0 and α_1.

In this case, $V_i = \mathbb{E}[N_i|\boldsymbol{a}] = \sum_{t=0}^{3} \beta_t \boldsymbol{a}(G^t)_i$ where $\beta_t = \prod_{j=1}^{t} q_j$. We estimate the coefficients in each of these cases letting $V_i = \mathbb{E}[N_i|\boldsymbol{a}]$ be the exposure mapping. We then generate the outcomes according to the exposure received

$$\mathbb{E}[Y_i|S_i, V_i] = \Lambda(\alpha_0 + \alpha_1(\sum_{t=0}^{3} \beta_t (G^t)_i \boldsymbol{a}))$$

where $\Lambda(\cdot)$ is the logistic function. For our experiments, we set $\beta = (0, 0.5, 0.05, 0.005)$.

In the dataset, seeds are assigned uniformly with either 3 or 5 seeds per network. Following our procedure in Section 3.4, we compute the optimal seed allocations, ensuring no cluster receives more seeds than available in the actual experiment (either 3 or 5). In practice our Bayesian optimization procedure starts by randomly sampling the target space 20 times, followed by 20 iterations to refine saturation. We simulate this process 500 times for each village in the dataset. We then compare the estimates for α_1 and all model parameters as shown in Figure 3.3. The results indicate that a more strategically designed experiment generally yields more significant gains than directly using the graph parameters. On average, using optimized designs rather than uniform random designs when collecting network data significantly reduced RMSE. Specifically, for estimating α_1, the optimized design decreased RMSE by 38% (12%) compared to 11% (2%) with complete data. For all parameters, the optimized design resulted in a 45% (10%) reduction in RMSE, versus an 18% (2%) reduction with complete data.



(a) Estimation of α_1.

(b) Estimation of all model parameters.

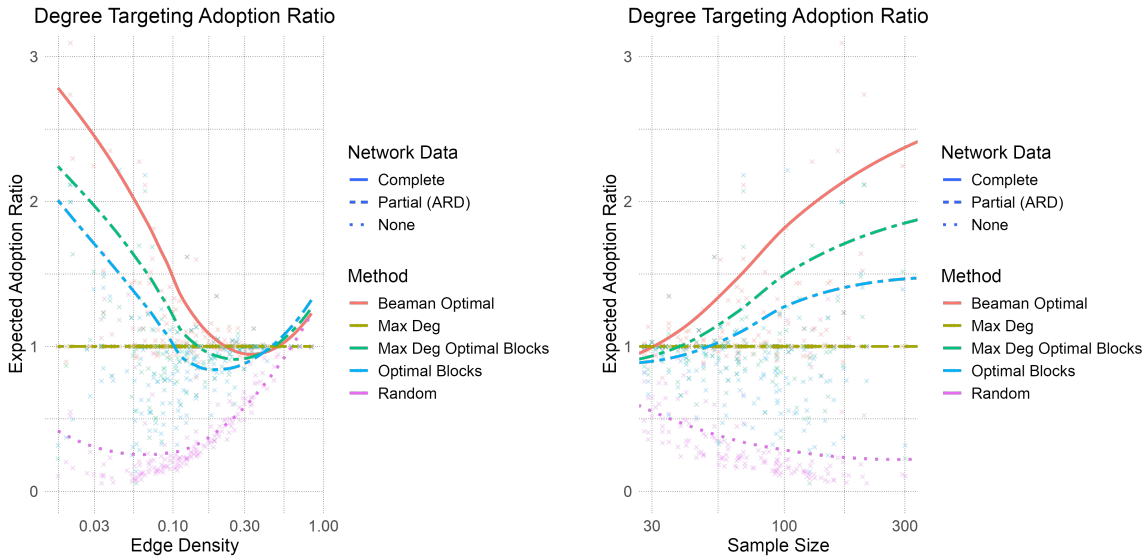
Figure 3.3: Estimation of parameter α_1 and all model parameters β using the naive and optimized seeding. We observe that the potential gain found using a more efficient design is much greater than simply collecting complete network data.

3.5.3 Optimal Seeding

We apply our methodology to the seeding problem described in Beaman et al. [2021], where the diffusion of pit-planting technology among Malawian farmers follows a complex contagion process. The outcome model is defined as $Y_i = f_Y(S_i, V_i, \boldsymbol{\epsilon}_Y)$, with individuals having a threshold $\varsigma_i \sim N_{[0,\infty)}(\lambda, 0.1)$ for spreading infection based on neighbor infections from the previous time (where $N_{(a,b)}(\mu, \sigma)$ refers to the μ, σ normal distribution truncated on the interval (a, b)) . This process is simulated over three time periods to align with their experimental design, setting $\lambda = 2$ and repeating 2000 times for $K = 8$ clusters to determine optimal seeding groups.

We explore two seeding strategies: randomly assigning seeds to the top two members of optimal clusters, and seeding the nodes with the highest degrees within these clusters. We compare these strategies to common degree targeting, noting that our max degree method typically yields the highest adoption rates, especially in larger, sparser villages, as illustrated

in Figure 3.4. However, in very small or dense networks, the performance differences between strategies are negligible. Across all graphs we find the optimal seeding strategy to increase adoption by 1.50 (0.16) times relative to degree seeding, while the optimal blocks was 1.13 (0.12) times and optimal degree within blocks increased adoption by 1.28 (0.13) times.



(a) Adoption ratio as a function of edge density. (b) Adoption ratio as a function of village size.

Figure 3.4: Comparison of different seeding methods under complex contagion. Model-based targeting of optimal blocks generally outperforms degree seeding, especially when targeting the highest degree nodes within those blocks.

3.6 Conclusions

We introduce a framework that identifies causal effects under interference using a structural causal model, facilitating inference with partial network data. The framework is general and can be applied using broad class of outcome models and graph models. Our outcome modelling approach leveraging node-level heterogeneity and exposure mappings allow for the estimation of all causal effects, rather that other methods which tend to focus on a single

causal effect like the GATE. Demonstrations through semi-synthetic problems highlight its effectiveness, matching or surpassing fully observed data methods in certain scenarios.

Our method highlights that directly modeling interference mechanisms offers several advantages, including leveraging transportability of outcome models for seeding and inference for experimental designs when estimating effects under interference.

Future studies might consider semiparametric approaches to estimation with partial data like those in Auerbach [2022]. Additional structured assumptions on potential outcomes as suggested in Belloni et al. [2022] could also be explored. Currently, our focus has been on analyzing problems at a single time point. However, future research could extend to designing experiments with panel data and staggered rollouts. It would also be worthwhile to develop classes of outcome models that more explicitly incorporate this temporal structure.

Chapter 4

LATENT CURVATURE ESTIMATION

4.1 Introduction

Networks or graphs, $G = (V, E)$, are widely used in multiple scientific fields. These objects characterize the relations between a set of n nodes (or vertices) $V = \{1, 2, \ldots, n\}$ via a set of edges $E \subset \{(i, j)|i \neq j, i \in V, j \in V\}$ representing the connections between the nodes. Social networks, where nodes often represent individuals, are a common application of these models (e.g., Borgatti et al. [2009]). Networks are also commonly seen in models of biological and physical sciences, such as nodes representing cells in neuroscience [Bassett et al., 2018] and particles [Papadopoulos et al., 2018]. Complex and high-dimensional structure is an inherent feature of network data, which poses challenges in modeling and representation.

One common modeling approach represents the graph through an embedding into a lower-dimensional geometric space, where both the properties of the geometric space and the positions of points within it provide insights into graph structure. Models that use this embedding representation rely on latent distance matrices [Hoff et al., 2002b, Handcock et al., 2007, Hoff, 2007, Smith et al., 2017], where the distances in the low-dimensional manifold are inversely proportional to the propensity to form a connection. We refer to this broad class as latent distance models. The geometry of the underlying manifold has substantial implications for the types of connections we expect to see in the network. Manifolds with positive curvature (hyper-spheres) tend to encourage triangles to close and produce more group structure, whereas negatively curved manifolds make it easier to form trees with long paths.

Our focus is on estimating the curvature of a manifold based on a set of (noisy) distance measurements. Our results rely on the fundamental, but profound, observation that the

properties of triangles depend on the curvature of the manifold on which they're embedded. In particular, we leverage the fact that the distance between a vertex in a triangle and the midpoint of the side opposite of that vertex is varies based on the curvature of the manifold. Lubold et al. [2023] study a similar problem of hypothesis testing the geometric class of various network models. However, we offer an alternative approach to estimating curvature with several distinct advantages. Notably, our method requires only four distances, derived from lengths on a triangle to estimate curvature, making it a local approach that can be used to develop tests for constant curvature. Furthermore, our estimating function is smooth, allowing a single equation to estimate the curvature and derive interpretable asymptotic results, unlike Lubold et al. [2023], which requires an eigenvalue equation that is generally not smooth and does not lead to interpretable asymptotics.

Our contributions include the following. We develop a smooth estimating equation to estimate curvature using a noisy distance matrix, leveraging triangles and their median lengths. We call the triangle median the distance from one vertex to the midpoint of the other two (not to be confused with the statistical median). These results are general and applicable to any noisy distance matrix, though in this paper we apply them specifically to social networks. We next consider various aspects of working with surrogate midpoints when a true midpoint is not observed in the data. We establish upper and lower bounds for the curvature when collecting a set of distances that does not contain the midpoint of another pair of points. We also establish "good conditions" under which surrogate midpoints form arbitrarily close to the actual midpoint of other points. Next, we turn to the specifics of the latent distance model. In this case, we present a curvature estimator and demonstrate that, under the typical assumptions used to fit latent distance models, it is asymptotically normal. We show that we can further improve estimation using a constrained estimator that reflects the triangle inequality among distance constraints. Lastly, we demonstrate that our estimator is a basis for the development of new methodology in sociology and cybersecurity by testing for changes in curvature.

The remainder of the paper continues as follows. First, we include a literature review

in Section 4.1.1. Next, in Section 4.2, we introduce our aforementioned methodological contributions. We next illustrate the efficacy of these methods through a simulation study. Then, we discuss downstream statistical tasks such as testing whether the curvature of a noisy distance matrix is constant in Section 4.4 and detecting changepoints in Section 4.5. We further elaborate on these with applications to co-authorship networks in physics and an application in cybersecurity.

4.1.1 Literature Review

The use of distance matrices for data analysis is prevalent across numerous fields. Originating from applications in psychometrics, multidimensional scaling (MDS) [Torgerson, 1952] pioneered the use of distance-based methods and has been explored in various domains, including the analysis of protein shapes [Havel and Wüthrich, 1985], image classification Tenenbaum et al. [2000], and natural language processing Kusner et al. [2015]. Notably, Hoff et al. [2002b] introduced this idea in a model of social network formation, which has since been expanded in numerous ways, such as model-based clustering Handcock et al. [2007], multi-view networks Salter-Townshend and McCormick [2017], and dynamic networks Kim et al. [2018]. Some models use mixtures of a block structure to model only at the individual level within a cluster of the network [Fosdick et al., 2016, Lok et al., 2021]. Latent distance models have been applied to problems like modeling social influence [Sweet and Adhikari, 2020], social media relationships of politicians [Lok et al., 2021], and neuron connectivity [Aliverti and Durante, 2019], among others.

We focus on the properties of the geometric space underlying a latent distance model, particularly the notion of curvature. The sectional curvature of a latent space is broadly defined as the deviation from a flat (Euclidean) space via the growth of the circumference of small circles as a function of their radius. An important class of manifolds are those that are simply connected and have constant curvature. A classical result from Killing [1891] characterizes these as the spherical (positive curvature), Euclidean (0 or flat curvature), and hyperbolic (negative curvature) spaces. Though importantly, these do not represent the

entire class of such manifolds.

Classically, the choice of the embedding space was at the discretion of the analyst. Notably, latent spherical and Euclidean spaces were used in Hoff et al. [2002b]. However, other metric spaces, particularly spherical and hyperbolic spaces, have been found to better represent many network data types Smith et al. [2017]. The authors also provide a simulation-based approach to compare the eigenspectrum of the graph Laplacian to models under spherical, hyperbolic, and Euclidean geometry. They show how the latent embedding space influences observed properties of the network, notably the degree distribution and clustering of the network, which can influence the behavior of network contagion processes (i.e., SIR models) [Volz et al., 2011].

Our work bears the closest resemblance to that of Lubold et al. [2023], which discusses hypothesis testing of the latent space among a class of models and the estimation of the related distance matrix. Our work is distinct in several ways. Firstly, our method of estimating the curvature of the latent space is novel and useful for deriving interpretable asymptotic results. This is due to the fact that our method allows for an estimating equation approach to identifying curvature, which leads to desirable properties. Importantly, their approach tests whether the geometry can be embedded globally in each of the canonical spaces, whereas we provide a local approach derived from triangle distances. Furthermore, we also provide an improved latent distance estimator which allows for the construction of an asymptotically normal distance matrix, based on cliques (fully connected subgraphs) in a network. Our approach for curvature estimation is modular and can be applied to general distance matrices. Consequently, we illustrate how to test for constant curvature within an embedding space.

An alternative definition of graph curvature worth discussing includes the Ollivier-Ricci curvature [Ollivier, 2007] and extensions such as Haantjes-Ricci curvature [Saucan et al., 2020] and Forman-Ricci curvature [Leal et al., 2018]. These definitions of curvature are derived from metrics arising from graph distances (i.e., integer-valued shortest path distances) rather than distances on a smooth latent space. As such, it is not apparent what these estimates will converge to (or if they converge at all) when a network is studied as a random

object, with the exception of van der Hoorn et al. [2020], who study a problem where connections are governed by a small radius on a latent space. The authors study the convergence of a modified Ollivier-Ricci curvature to the Ricci curvature of the underlying space in random geometric graphs under the limit of the connection radius shrinking to 0. However, these discrete curvatures have also been applied to various settings, such as financial network instability [Sandhu et al., 2016, Samal et al., 2021], network sampling [Barkanass et al., 2022], cancer detection in gene regulatory networks [Sandhu et al., 2015], functional neuroscience [Farooq et al., 2019], and community detection [Sia et al., 2019, Ni et al., 2019].

4.2 Methods

We begin by formally introducing our environment, including defining the properties of manifolds covered by our method. Next, we propose an estimator of curvature based on noisy distance measurements from triangle midpoints. We will begin with an ideal estimator when the true midpoint is measured and then follow up with a study of using points that are nearly midpoints, which we call surrogate midpoints, in their place. These methods are general and apply in any setting where we have measured noisy distances. We then turn to our setting–social networks–and describe in detail how to construct distance estimators, and subsequent estimates of curvature, for the latent distance model.

4.2.1 Geometric Environment

We begin by defining the geometric environment. We assume points lie on a Riemannian manifold $\mathcal{M}^p$ of dimension p, equipped with a corresponding metric tensor g. The metric tensor can be used to define the sectional curvature of the manifold at a point $q \in \mathcal{M}^p$. For our purposes, we assume that this manifold is connected and that the curvature is both upper and lower bounded. In our problem, we exclusively work with distances and thus consider the metric space induced by the Riemannian manifold $\mathfrak{M} = (\mathcal{M}^p, d)$. We include the related definitions of the metric tensor and the distance on the manifold in the appendix in Section C.4.

We further assume the manifold is a member of the class of simply connected Riemannian manifolds with constant sectional curvature (κ). This assumption is consistent with work on the latent distance model in social networks [Hoff et al., 2002b]. These include the classical Euclidean $\mathbb{E}^p$ ($\kappa = 0$), spherical $\mathbb{S}^p(\kappa)$ ($\kappa > 0$), and, more recently, hyperbolic space $\mathbb{H}^p(\kappa)$ ($\kappa < 0$). The celebrated Killing-Hopf theorem states that these are the only manifolds of this type [Killing, 1891, Hopf, 1926]. Therefore, we refer to these manifolds spaces as the **canonical manifolds**, and we introduce common representations of these manifolds later in this subsection.

To identify the curvature, we rely on a simple geometric insight, relating the side lengths of a triangle and the length of the triangle's median; the line segment connecting a vertex to the midpoint of the opposite side.

In general, we require the manifold to satisfy two properties.

- (A1) (**Algebraic Midpoint Property**) $\mathfrak{M}$ satisfies the algebraic midpoint property. For any $x, y \in \mathcal{M}^p$, there exists a point z such that $d_{\mathcal{M}^p}(z, x) = d_{\mathcal{M}^p}(z, y) = \frac{1}{2} d_{\mathcal{M}^p}(x, y)$.

- (A2) (**Locally Euclidean**) For all $q \in \mathcal{M}^p$, there exists some $\delta > 0$ and some functions c_p, C_p such that for all $\epsilon \leq \delta$:

$$c_p(\epsilon) \leq \frac{\text{Vol}(B_{\mathcal{M}^p}(\epsilon, q))}{\text{Vol}(B_{\mathbb{E}^p}(\epsilon, 0))} \leq C_p(\epsilon) \text{ and } \lim_{\epsilon \to 0} \frac{\text{Vol}(B_{\mathcal{M}^p}(\epsilon, q))}{\text{Vol}(B_{\mathbb{E}^p}(\epsilon, 0))} = 1.$$

Here, $B_{\mathcal{M}^p}(\epsilon, q)$ is the ϵ-ball on $\mathcal{M}^p$ centered at a point q, which we abbreviate to $B(\epsilon, q)$, and $B_{\mathbb{E}^p}(\epsilon, 0)$ is the ϵ-ball in Euclidean space. These conditions will hold under our mild assumptions on the manifold. See Section C.5 in the supplementary materials for details. Next, we consider an explicit set of representations of the canonical manifold.

Canonical Manifolds.

Each of the canonical manifolds can be represented using a set of positions with real-valued vectors and a corresponding distance function, allowing for closed-form computation of the distances. We include definitions for Euclidean, spherical, and hyperbolic spaces for completeness. We emphasize the difference between the intrinsic and extrinsic geometry

here. Though each of these canonical manifolds is embedded in $\mathbb{R}^p$, only the Euclidean space uses the standard 2-norm to construct the distances. The curved canonical manifolds $\mathbb{S}^p$ and $\mathbb{H}^p(\kappa)$ can be embedded in $\mathbb{R}^{p+1}$ along with a properly defined metric. In the spherical example, we compute the length according to the path length on the surface of the sphere, rather than the Euclidean distance through the sphere. We next highlight each of these models.

The **Euclidean manifold**: $\mathbb{E}^p$ can be described using a set of points in $\mathbb{R}^p$ with the standard 2-norm.

$$d_{\mathbb{E}^p}(x, y) = \sqrt{\sum_{k=1}^{p} (x_k - y_k)^2}$$

The **spherical manifold**: $\mathbb{S}^p$ with curvature $\kappa > 0$ is equivalent to the sphere of radius $r = \frac{1}{\kappa^2}$. We express this using a set of coordinates $x \in \mathbb{R}^{p+1}$, such that $\sum_{k=0}^{p} x_k^2 = 1$. The distance on the sphere can be computed using the quadratic form $B_{\mathbb{S}^p}$ defined below:

$$B_{\mathbb{S}^p}(x, y) = \sum_{k=0}^{p} x_k y_k, \quad d_{\mathbb{S}^p}(x, y) = \frac{1}{\sqrt{\kappa}} \arccos(B_{\mathbb{S}^p}(x, y)).$$

The **hyperbolic manifold**: $\mathbb{H}^p$ with curvature $\kappa < 0$ can be constructed using the hyperboloid model (Minkowski model), which corresponds to a set of points $x \in \mathbb{R}^{p+1}$, such that $x_0^2 - \sum_{k=1}^{p} x_k^2 = 1, x_0 > 0$. An analogous quadratic form $B_{\mathbb{H}^p}$ exists for the hyperbolic embedding and can be used to compute the distances:

$$B_{\mathbb{H}^p}(x, y) = x_0 y_0 - \sum_{k=1}^{p} x_k y_k, \quad d_{\mathbb{H}^p}(x, y) = \frac{1}{\sqrt{-\kappa}} \operatorname{arccosh}(B_{\mathbb{H}^p}(x, y)).$$

4.2.2 Identifying Curvature

A number of methods exist to verify whether a particular set of distances can be embedded in a space of constant curvature, typically based on the zeros of eigenvalues of a transformation of the distance matrix. These include methods by Schoenberg [1935] and Begelfor and Werman [2005], as well as Cayley-Menger determinants [Blumenthal and Gillam, 1943]. Lubold et al. [2023] previously used such a criterion to identify whether a set of distances

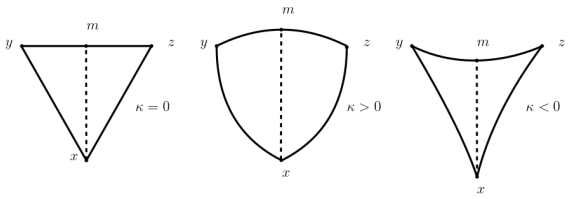


Figure 4.1: Midpoint distances and curvature of the space with equilateral triangles. The length of the triangle median d_{xm} is an increasing function of the curvature κ for fixed other triangle side lengths.

could be globally embedded in a particular curvature space. In this work, we take a different approach where we identify curvature based on a minimal set of points. We also estimate a specific curvature, rather than performing a test on the sign of the curvature, as presented by Lubold et al. [2023].

We rely on a simple geometric observation. Consider a set of three points that form a triangle and the length of the median. This length reveals the sectional curvature of the manifold, where a smaller distance corresponds to a more negatively curved space and a larger distance corresponds to a more positively curved space. This is visualized in Figure 4.1 for Euclidean, spherical, and hyperbolic triangles.

This use of midpoints helps to identify the curvature uniquely, when it may not be identified in other situations. For example, consider 4 points placed equidistant from each other. This set of distances can be either represented in $\mathbb{E}^3$ as the tetrahedron, or using equidistant points on the sphere $\mathbb{S}^2(\kappa)$ for $\kappa > 0$, making it impossible to identify the curvature of the space from this collection of distances alone.

We instead take a more direct approach to eliciting curvature by leveraging distances in a triangle and the triangle median (which, recall, is the distance from one vertex to the midpoint of the other two and distinct from the statistical median). This allows us to identify curvature with only four points in total. We now formalize this intuition. For any

three points x, y, z which lie in an unknown canonical manifold of dimension $p \geq 2$ with constant sectional curvature κ, (x, y, z) can be isometrically embedded in a submanifold of dimension 2. We illustrate this through the use of submanifolds that contain their *geodesics*, the paths on a manifold that minimize the path length between two points, and thus define a distance. Simply stated, even if the manifold's dimension $p > 2$, the curvature can be identified through the totally geodesic submanifold containing the triangle.

Definition 4.2.1. A submanifold $\widetilde{\mathcal{M}} \subset \mathcal{M}$ is *totally geodesic* if every geodesic in $\widetilde{\mathcal{M}}$ is also a geodesic in $\mathcal{M}$.

Some simple examples include the Euclidean plane, within the three-dimensional Euclidean space ($\mathbb{E}^2$ is a totally geodesic submanifold of $\mathbb{E}^3$). However, this is not always the case; consider the two-dimensional sphere $\mathbb{S}^2(\kappa)$, which also resides within three-dimensional Euclidean space but does not contain all of its geodesics, as these geodesics in $\mathbb{E}^3$ pass through the center of the sphere. This distinction highlights the differences between intrinsic and extrinsic notions of distance, as our object of study is the former. We will use the fact that a totally geodesic submanifold contains all points along the geodesic, including the midpoint.

Lemma 4.2.2. *If any $x, y, z \in \mathcal{M}^p(\kappa)$ where $p \geq 2$ and x, y, z are not co-linear. Then $x, y, z, m \in \mathcal{M}^2(\kappa)$ where $\mathcal{M}^2(\kappa)$ is a totally geodesic submanifold of dimension 2 with constant sectional curvature κ and m is the midpoint of points y and z.*

The intuition behind this lemma is that, regardless of the ambient dimension of the latent manifold p, we can determine the curvature from a two-dimensional submanifold. This submanifold is constructed from the geodesics of a given triangle. As we will see in Theorem 4.2.3, the side lengths and the length of the triangle's median are sufficient to identify the curvature of the manifold. The proof is straightforward and in Section C.1.1. The main implication here is that the totally geodesic submanifold allows us to look at the distances in a subspace of dimension 2 which will be useful in the following theorem for identification. The three points x, y, z will fall into one of these sub-manifolds, as well as

m which lies on the geodesic between y and z. Since geodesics determine the distance, and geodesics on the submanifold are the same as geodesics on the manifold.

We now use this fact to derive an equation which will relate the curvature κ to the set of distances between the points (x, y, z, m), which we denote $\boldsymbol{d}^{\triangle} = (d_{xy}, d_{xz}, d_{yz}, d_{xm})$.

Theorem 4.2.3 (Midpoint Curvature Equation)**.** *Suppose that points $x, y, z \in \mathcal{M}^p(\kappa)$ an unknown Riemannian manifold of dimension $p \geq 2$ of constant sectional curvature κ. Let m denote the midpoint between y, z. The following equation holds for $\kappa \in \mathbb{R}$.*

$$g(\kappa, \boldsymbol{d}^{\triangle}) = Re\left[\frac{2\cos(d_{xm}\sqrt{\kappa})}{\kappa} - \frac{\sec(\frac{d_{yz}}{2}\sqrt{\kappa})(\cos(d_{xy}\sqrt{\kappa}) + \cos(d_{xz}\sqrt{\kappa}))}{\kappa}\right] = 0 \tag{4.1}$$

Where d_{jk} denoted the distance between points j, k and $Re[]$ denotes the real part of the equation.

The proof first leverages the fact that by Lemma 4.2.2, we can construct a submanifold of dimension 2 that contains the midpoint of points on a triangle. We then use this to derive an equation that relates the side lengths of the triangle, the triangle median length, and the curvature of the space. For cases when $\kappa < 0$, we take the real part of equation (4.1), which is equivalent to replacing the trigonometric functions with their hyperbolic analogues. Though this does require that $p \geq 2$, this covers most manifolds of interest. The proof is found in the supplementary materials in Section C.1.2. It will also be convenient to express the length of the triangle median $d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z))$ as a function of the curvature κ and a set of triangle side lengths $\boldsymbol{d}^{\triangle}(x, y, z)$.

We remark on the a similar criterion used by Gu et al. [2018] for identifying the number of components of positive and negative curvature in a product-space embedding. Gu et al. [2018] derive the parameter θ_T from Toponogov's theorem (which can be found in the appendix Theorem C.1.3) and its sign can be used to identify the sign of the curvature of the manifold. In fact, it is straightforward to show $\lim_{\kappa \to 0} g(\kappa, \boldsymbol{d}^{\triangle}) = \theta_T(\boldsymbol{d}^{\triangle})$.

If the sectional curvature of $\mathcal{M}^p$ is positive, then $\theta_T(\boldsymbol{d}^{\triangle}) > 0$, and if the curvature is negative, $\theta_T(\boldsymbol{d}^{\triangle}) < 0$. In Euclidean space, $\theta_T(\boldsymbol{d}^{\triangle}) = 0$, and this reduces to the parallelogram

law. Our method is distinct as it directly identifies the curvature using this same set of distances, rather than the sign of the curvature.

If the true manifold that generates the distances is non-constant, then the value κ is the curvature of the 2-dimensional canonical manifold which can isometrically embed a set of triangle distances $\boldsymbol{d}^{\triangle}$.

Our method differs from those of Schoenberg [1935] and the Cayley-Menger determinants [Blumenthal and Gillam, 1943] because it requires midpoint information, which theirs do not. Their methods do not uniquely identify curvature but only determine if a space of constant curvature can embed an given distance matrix. In the appendix, we illustrate the smoothness of this equation, which is also a desirable property for plug-in estimators (Section C.8.2).

4.2.3 Estimating Curvature

In this subsection, we describe how to estimate the curvature from a noisy estimate of a set of distances using Theorem 4.2.3. We begin by introducing an estimator based on triangle distances and its median length. In practice, we are often given an estimate of a distance matrix between an arbitrary set of K points $\boldsymbol{D} \in \mathbb{R}^{K \times K}$. In this setting, it is not guaranteed that there is a midpoint of two other points among the observed points in the distance matrix. We illustrate how one can bound the curvature in this setting. Lastly, we introduce a result that characterizes the formation of points arbitrarily close to the midpoint of other points.

We first consider an idealized scenario. In this setting, we suppose that we are given an estimate of the triangle distances $\widehat{\boldsymbol{d}}^{\triangle}$. We return to the problem of estimating distances in our model in Section 4.2.5. We can use such a set of noisy or estimated distances to estimate the curvature ($\widehat{\kappa}$) by solving for the value of κ which is the solution to equation 4.2.

$$g(\widehat{\kappa}, \widehat{\boldsymbol{d}}^{\triangle}) = 0 \tag{4.2}$$

The advantage of this method is that the smoothness of g allows for the derivation of explicit asymptotics for $\widehat{\kappa}$, which is not possible for the method used by Lubold et al. [2023].

We present this result below in Theorem 4.2.4.

Theorem 4.2.4. *Suppose there exist points $x, y, z \in \mathcal{M}^p$. Let m denote the midpoint between y, z where these points are fixed. Let $\widehat{\boldsymbol{d}}^{\triangle} = (\widehat{d}_{xy}, \widehat{d}_{xz}, \widehat{d}_{yz}, \widehat{d}_{xm})$ be the estimated distances and $\boldsymbol{d}^{\triangle} = (d_{xy}, d_{xz}, d_{yz}, d_{xm})$ be their true, unknown counterparts.*

Assume we have a distance estimator $\widehat{\boldsymbol{d}}$ such that

- *(B1)* $r(n)\left(\widehat{\boldsymbol{d}}^{\triangle} - \boldsymbol{d}^{\triangle}\right) \to N(0, \Sigma)$
- *(B2)* $\kappa < \left(\frac{\pi}{\max\{d_{xy}, d_{xz}, d_{yz}, d_{xm}\}}\right)^2$
- *(B3)* $\frac{d}{d\kappa'} d_{xm}(\kappa'; \boldsymbol{d}^{\triangle}(x, y, z))\Big|_{\kappa'=\kappa} > 0$

Where $r(n)$ is the rate of convergence. Let $\widehat{\kappa}$ be the solution to $g(\kappa, \widehat{d}) = 0$. Then

$$r(n)(\widehat{\kappa} - \kappa) \to_d N\left(0, \left(\frac{\partial g(\kappa, d)}{\partial \kappa}\right)^{-2} \left(\nabla_d g(\kappa, d)^\top \Sigma \nabla_d g(\kappa, d)\right)\right) \tag{4.3}$$

where $\to_d$ refers to convergence in distribution. If $(B1)$ is replaced by a consistency, i.e. $||\widehat{\boldsymbol{d}}^{\triangle} - \boldsymbol{d}^{\triangle}|| = o_P(r(n))$, then $\widehat{\kappa} - \kappa = o_P(r(n))$.

The proof is found in the supplementary materials in Section C.1.3 and is an application of the implicit function theorem together with the delta method. Assumption (B1) is mild as it only requires asymptotic normality of the distance estimator. Assumption (B2) is trivial since it simply requires that the true distances are not greater than the maximum allowable distances on the sphere of curvature κ. Assumption (B3) tends to hold unless the three points x, y, z are collinear. For a more in-depth discussion of the non-decreasing property of the midpoint, see Lemma C.1.2 in the appendix. In Section 4.2.5, we illustrate the asymptotic normality of a distance estimator based on cliques. We next discuss the bias associated when a set of pairwise distances are observed, where no point is necessarily a midpoint of another pair of points.

In general, rather than distances from a triangle median, we may only observe distances in the form of a distance matrix $\boldsymbol{D} \in \mathbb{R}^{K \times K}$. There might not be a midpoint between two

other points within this set. In this case, given a triangle, we can use a point nearby to the midpoint as a *surrogate midpoint* m' taking the role of a midpoint between y, z.

We let κ' be the solution to equation (4.2) where $d_{xm'}$ takes the place of d_{xm}. In this case, we can approximate the bias of the curvature estimate using a Taylor series expansion.

$$\begin{aligned} 0 &= g(\kappa', \boldsymbol{d}^{\triangle'}) - g(\kappa, \boldsymbol{d}^{\triangle}) \\ &= \nabla_\kappa g(\kappa, \boldsymbol{d}^{\triangle})(\kappa - \kappa') + \nabla_{d_{xm}} g(\kappa, \boldsymbol{d}^{\triangle})(d_{xm} - d_{xm'}) + o(|\kappa - \kappa'| + |d_{xm} - d_{xm'}|) \\ \implies |\kappa - \kappa'| &\approx (\nabla_\kappa g(\kappa, \boldsymbol{d}^{\triangle}))^{-1} \nabla_{d_{xm}} g(\kappa, \boldsymbol{d}^{\triangle}) |d_{xm} - d_{xm'}| \\ &\leq (\nabla_\kappa g(\kappa, \boldsymbol{d}^{\triangle}))^{-1} \nabla_{d_{xm}} g(\kappa, \boldsymbol{d}^{\triangle}) d_{mm'}. \end{aligned}$$

Therefore, the bias will scale approximately linearly as a function of $d_{mm'}$ for small values of $d_{mm'}$.

If four points are within a manifold of constant curvature $x, y, z, m' \in \mathcal{M}^p(\kappa)$, then using the curvature value κ, one can compute the distance from the midpoint m to m' by letting m' take the place in equation (4.1) and solving for $d_{mm'}$. As such, we denote this $d_{mm'}(\kappa) := d_{m'm}(\kappa; \boldsymbol{d}^{\triangle}(y, z, m'))$ as a function of the curvature. Using the triangle inequality, we can then upper and lower bound the curvature by replacing d_{xm} with upper and lower bounds in equation (4.1) and solving the corresponding equations. We illustrate this in Theorem 4.2.5.

For a given κ, let $\boldsymbol{d}^{\triangle,+}(\kappa) = (d_{xy}, d_{xz}, d_{yz}, d_{xm'} + d_{mm'}(\kappa))$ and $\boldsymbol{d}^{\triangle,-}(\kappa) = (d_{xy}, d_{xz}, d_{yz}, d_{xm'} - d_{mm'}(\kappa))$.

Theorem 4.2.5 (Curvature Bounds)**.** *Let $x, y, z, m' \in \mathcal{M}^p(\kappa)$. Then let d_{jk} denote the distance between points $j, k \in \{x, y, z, m'\}$. Let κ_u and κ_l denote the solutions*

$$g(\kappa_u, \boldsymbol{d}^{\triangle,+}(\kappa_u)) = 0, \quad g(\kappa_l, \boldsymbol{d}^{\triangle,-}(\kappa_l)) = 0$$

then $\kappa_l \leq \kappa \leq \kappa_u$.

When the surrogate midpoint and the true midpoint are the same, then $d_{mm'} = 0$ and the upper and lower bounds converge. Similar to the curvature estimate $\widehat{\kappa}$, given a noisy

estimate of the distances, we can estimate the upper and lower bounds of the curvature. In Section 4.4, we leverage these bounds to develop a test of constant curvature. We next address the formation of surrogate midpoints arbitrarily close to the midpoints of other pairs of points.

We next provide an outline involving how fast we can expect surrogate midpoints to form. In order to do so, we first introduce a useful definition. A subset $A \subset \mathcal{M}^p$ is *geodesically convex* if the geodesic between any two points in A is contained within A itself. Here, convexity on a manifold will refer to geodesic convexity. In Theorem 4.2.6, we let $h_{m,3}(m_1, m_2)$ denote the joint density function of a pair of midpoints with a shared endpoint and $h_{m,4}(m_1, m_2)$ denote the joint density of two midpoints without a shared endpoint. These two densities will be functions of the unknown manifold $\mathcal{M}$ and the distribution of the positions Z', $G_{Z'}$.

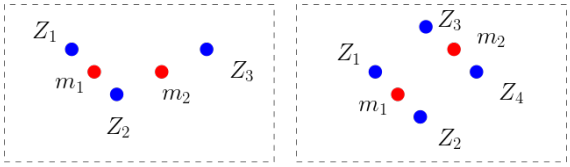


Figure 4.2: Left side illustrates two midpoints with a shared endpoint with joint density $h_{m,3}$, while the right side illustrates a joint density with no shared endpoints $h_{m,4}$. Endpoints, i.e., sampled positions of Z_i are shown in blue while midpoints of the pairs are shown in red.

Theorem 4.2.6. *Suppose that $Z_i' \overset{iid}{\sim} H_{Z'}$ are K points sampled iid from a distribution on a simply connected manifold $\mathcal{M}^p$. Denote this set of points $\{Z_i'\}_{i=1}^K = \mathcal{D}_K$. Suppose there exists a convex region A for which*

- *(C1)* $h(z) \geq \alpha > 0$, *for all* $z \in A$
- *(C2)* $\dim(A) = p \geq 2$
- *(C3)* $h_{m,3}(m_1, m_2), h_{m,4}(m_1, m_2) \leq \alpha_m < \infty$

for all $z \in A$ where f is the density function corresponding to G_Z. Let $m(y, z)$ denote the midpoint between two points y and z. Define the statistic

$$\Phi(\mathcal{D}_K) := \min_{x,y,z \in \mathcal{D}: x \neq y, x \neq z, y \neq z} d\Big(m(y, z), x\Big)$$

which is the minimum distance from an observed point to the midpoint of another pair of points. Then

$$\Phi(\mathcal{D}_K) = \mathcal{O}_P(K^{-3/p}). \tag{4.4}$$

Our approach to demonstrating the above result draws on the work of Cai et al. [2013] and Brauchart et al. [2015], who discuss the convergence of the minimum distance between any two points sampled uniformly on a hypersphere. These authors show that $\min_{i,j} d(Z_i, Z_j) = \mathcal{O}_P(K^{-2/p})$, and they also derive an exact distribution for $\min_{i,j} d(Z_i, Z_j)$ under the assumption of uniformity on the sphere. They use a technique that recursively computes the probability that a point is at least a radius ϵ away from the previous K points. In our approach, at each placement of a new point, there are $\binom{K}{2}$ midpoints instead of K current points, leading to the faster rate we observe.

Assumption (C1) ensures the existence of geodesically convex regions for which midpoints can form. Assumption (C2) ensures that this region has the same dimension as the ambient space. Lastly, Assumption (C3) is relatively mild as long as we have a smooth manifold and a continuous density. The proof is detailed in Section C.1.5 of the supplementary materials and relies on a result regarding medians of arbitrarily correlated random variables, which may be of independent interest (Theorem C.1.4).

4.2.4 Reducing Bias and Variance in Curvature Estimation

When given an estimate of a distance matrix and to later apply our method, as in many statistical problems, we are concerned with the bias and variance of our estimator. The variance, in general, will be driven by the shape of the triangle used to estimate the curvature, while the bias will be driven by the closeness of the surrogate midpoint to the true midpoint

of a triangle. We first visualize this phenomenon and then follow up by providing some practical strategies for constructing good estimators.

We can visualize the theoretical variance by using our asymptotic result in Theorem 4.2.4. In Figure 4.3, we plot the variance of the curvature estimate for a distance estimator with identity variance (i.e., $\widehat{\boldsymbol{d}}^{\triangle} \sim N(\boldsymbol{d}^{\triangle 0}, I_{4\times 4})$) for a variety of choices of κ. In this plot, we change the position of the vertex x of the triangle (which we may also refer to as the reference point). The smallest variance reference points across all of these curvatures are the ones that form nearly equilateral triangles with the points y, z. Additionally, the variance of the estimator tends to be larger in a given reference location as the curvature κ decreases. We plot all reference points x within a ball of radius 2. Secondly, we illustrate, for an equilateral triangle, the bias of the estimate of the curvature when moving the location of the surrogate midpoint m'. The spherical and hyperbolic spaces are shown using a projection where the distance to the center and relative angle are mapped onto Euclidean space.

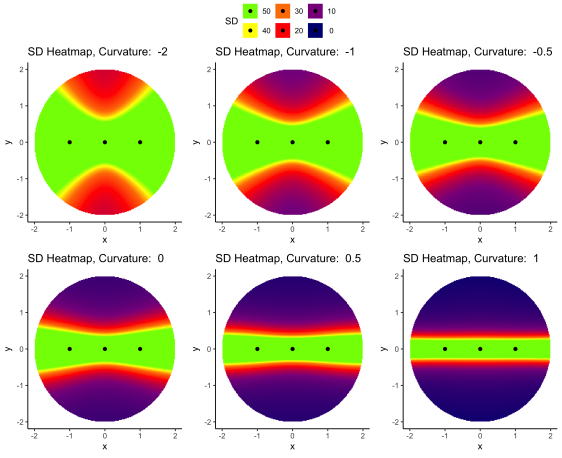


Figure 4.3: Variance as a function of x position. True points (y, m, z) in black.

We next illustrate some practical choices to minimize the bias and variance of a curvature estimate from a distance matrix $\widehat{\boldsymbol{D}}$.

Choosing the Best Midpoint. In practice, we would like to search over the space

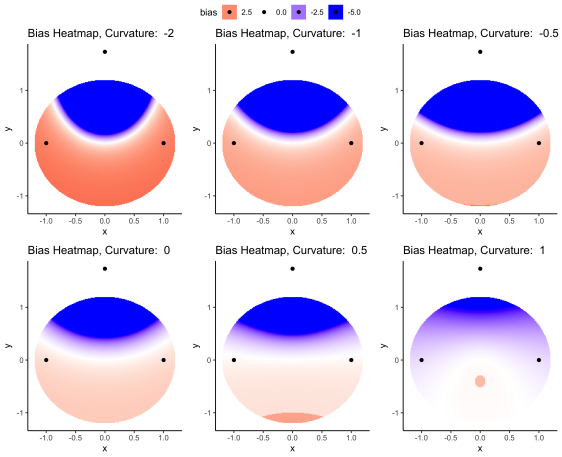


Figure 4.4: Bias as a function of surrogate midpoint m' position. True points (x, y, z) in black. For visualization purposes, the bias is capped at a magnitude of 5.

to find the best midpoint between two other points in the distance matrix. The midpoint between any two points y, z is also known as the Fréchet mean, which is defined as follows:

$$m^* = \arg\min_{m \in \mathcal{M}^p} d_{ym}^2 + d_{zm}^2.$$

In practice we search over the space of candidate entries of a distance matrix D to find the best midpoint available. We also want to ensure that the points y, z are not too close to each other since this tends to lead to a high variance estimator. A reasonable option is to normalize this quantity by d_{yz}. Given a true midpoint m^*, then $d_{ym^*} = d_{zm^*} = \frac{1}{2} d_{yz}$, then

$$\frac{d_{ym^*}^2 + d_{zm^*}^2}{d_{yz}^2} = \frac{1}{2}. \tag{4.5}$$

We can also add a term $\frac{|d_{ym} - d_{zm}|}{d_{yz}}$ which aids in balancing the lengths of the distance to each point.

Therefore, to compute the best surrogate midpoint set, we solve the following problem:

$$\widehat{y}, \widehat{z}, \widehat{m} = \arg\min_{y,z,m} \left(\frac{d_{ym}^2 + d_{zm}^2}{d_{yz}^2} + \frac{|d_{ym} - d_{zm}|}{d_{yz}} \right). \tag{4.6}$$

In cases where we are interested in measuring the curvature across multiple surrogate midpoint sets, we remove this set of points and solve equation (4.6) for the remaining indices to construct a collection of surrogate midpoint set $\{(y^{(j)}, z^{(j)}, m^{(j)})\}_{j=1}^{J}$.

Selecting the Best Triangles. Given a surrogate midpoint set (y, m', z), we seek to find choices for the reference point x which provide the lowest variance. A good rule of thumb is to search for triangles x, y, z that are nearly equilateral.

We exploit this by considering a scaled version of the triangle inequality. Let $C_\triangle \in [1, 2]$ be a constant that determines the flatness of the allowed triangles x, y, z. Then we select only the x such that

$$d_{ij} + d_{jk} \geq C_\triangle d_{ik} \quad \forall (i, j, k) \in (x, y, z). \tag{4.7}$$

If one believes that the curvature is constant across a surrogate midpoint set, we can estimate a single curvature by taking the median across the values of x. This tuning parameter allows us to pick the triangles closest to equilateral, which tend to give the best estimates of the curvature. Letting $C_\triangle = 1$ allows for all triangles, no matter how flat they are, and $C_\triangle = 2$ will only permit exact equilateral triangles. Setting $C_\triangle$ too large results in no triangles being found, and setting $C_\triangle$ too small will result in using triangles that are very flat, often suffering from high variance in the corresponding $\widehat{\kappa}$ estimates. In practice, we find that a value of $C_\triangle \in [1.05, 1.7]$ is effective, and a good default choice is 1.3.

From an estimated distance matrix $\widehat{\boldsymbol{D}}$, we let $\boldsymbol{X}_{\widehat{\boldsymbol{D}}, C_\triangle} = \{x : \text{eq } (4.7)\}$ denote the set of reference points used for curvature estimation. We let $\widehat{\kappa}_x = \kappa(\widehat{\boldsymbol{D}}, \widehat{\boldsymbol{d}}^{\triangle}(x, y, z, m'))$, where $\widehat{\boldsymbol{d}}^{\triangle}(x, y, z, m')$ denotes the estimated distances of the triangle (x, y, z) with surrogate midpoint m'. Let $\widehat{\kappa}_{med} = \text{median}\{\widehat{\kappa}_x\}$. Given $\boldsymbol{X}$, we can construct the median estimator $\widehat{\kappa}_{med,\boldsymbol{X}}$ of the curvature in equation (4.8).

$$\widehat{\kappa}_{med,\boldsymbol{X}} = \text{median}_{x \in \boldsymbol{X}} \kappa(\widehat{\boldsymbol{D}}, y, z, m, x) \tag{4.8}$$

We now turn to estimating the distance matrix $\widehat{\boldsymbol{D}}$ in a latent distance model.

4.2.5 Distance Estimation in Latent Distance Models for Social Networks

We consider the random undirected network corresponding to a graph $G = (V, E)$ where $|V| = n$, with adjacency matrix $A \in \{0,1\}^{n\times n}$ such that $A_{ij} = 1$ iff $(i, j) \in E$. We motivate our method through the latent distance model of network formation Hoff et al. [2002b]. This is one specific choice of distance model that we will focus on, but it is not the only option. In this model, locations Z_i, Z_j are most generally equipped to some metric $d(\cdot,\cdot)$, forming a metric space, $\mathfrak{M}$. As before we consider the metric space generated from a manifold $\mathcal{M}^p$ of dimension p, where the probability of forming an edge is inversely proportional to the distance on the latent metric, $d_{\mathcal{M}}(Z_i, Z_j)$,

$$\Lambda\bigg(P\bigg(A_{ij} = 1\bigg)\bigg) \propto -d_{\mathcal{M}}(Z_i, Z_j)$$

where $\Lambda(\cdot)$ is a link function. Where convenient, we condense $d_{\mathcal{M}}(Z_i, Z_j)$ to d_{ij} for brevity to denote the distance between two indices i and j. A specific form we consider, is to include random effects (representing node level-gregariousness, or a degree correction), is as follows:

$$P(A_{ij} = 1|\nu, Z) = \exp\bigg(\nu_i + \nu_j - d_{\mathfrak{M}}(Z_i, Z_j)\bigg) \tag{4.9}$$

where

$$Z_i \sim_{iid} F_Z \quad \nu_i \sim_{iid} F_\nu$$

with a corresponding measures F_ν having support on the non-positive real line $\text{supp}(F_\nu) \subset (\infty, 0]$. The latent position measure F_Z has support on some the latent manifold $\mathcal{M}^p$. Here *iid* refers to sampling identically and independently from the latent distribution.

This model relates the latent distances of the generative model to the probability of a connection between two points. Although not discussed in this paper, simple generalizations can be derived for directed networks. This paper aims to estimate the curvature of $\mathcal{M}^p$. The $\exp(\cdot)$ link function is used as an example for ease of explanation in deriving the localization of positions in the latent space (this will be further discussed later in this section). However, in the supplementary materials (Sections C.3.1, C.3.2), we discuss the extension to the

expit$(\cdot)$ model, which was the original link function proposed by Hoff et al. [2002b], as well as additional link functions. As we illustrate in Section 4.2.5, this link function will be convenient for describing the formation and localization of latent positions within cliques, which are useful for estimating latent distances.

In this section, we illustrate a procedure for estimating a distance matrix based on latent positions in a latent distance model. We improve upon the results of Lubold et al. [2023] to construct an asymptotically normal distance estimator using *cliques*, fully connected subgraphs of the graph G. This approach has two advantages. First, since cliques represent multiple nodes, we have multiple opportunities for connections between two cliques, meaning we can use the fraction of realized ties between two cliques as an estimate for the probability of connection between the two cliques. Second, cliques correspond to points near each other in space. As the size of the clique grows, the maximal distance between latent positions must be small (in fact, the maximum distance between nodes in a clique will converge exponentially fast with respect to the size of the clique). This means that we can consider cliques as "points" on the manifold that can be used to identify a distance matrix. We visualize this in Figure 4.5.

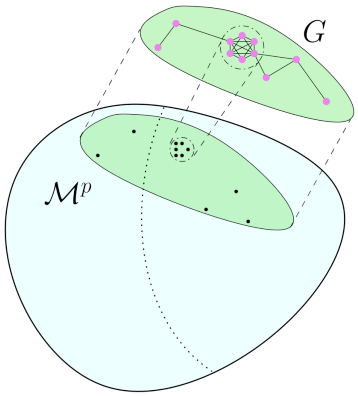


Figure 4.5: Illustration of the localization of the latent positions within a clique. Nodes are shown in magenta, while the position on the latent space is shown in black.

We discuss asymptotics for a fixed number of cliques (K) of size ℓ where the clique size grows, $\ell \to \infty$. In the appendix (Section C.3), we discuss the rate of clique formation as well as other alternatives to the clique-based estimators, which can be tailored to sparse graphs. Under this model, as $n \to \infty$, the expected number of cliques of size $\ell = \mathcal{O}(n^{1/(p+2)-\epsilon})$ for $\epsilon > 0$ grows to infinity as well. Since the formation of cliques requires nodes to be extremely close to one another in the latent space, the primary driver is the intrinsic dimension of the latent space p, rather than the curvature. A formal statement of this is found in the appendix (Section C.3.3).

An important aspect we consider is the maximum radius of a set of positions conditional on a clique. Unless latent positions are nearly (or exactly) in the same location, large cliques are rare in latent distance models. Lemma 4.2.7 illustrates a rate of convergence of these latent locations relative to the size of a clique. Additionally, the nodes which form cliques will also have random effects approaching 0, Lemma C.1.5.

Under this model, the nodes within a clique have nearly the same latent position. For illustration purposes, we first assume that this holds and later illustrate the rate at which the diameter of the set of latent positions within a clique converges. Let $X, Y \in \{1, 2, \ldots, n\}$ denote sets of nodes representing non-overlapping cliques. Then the average probability of connection can be used to identify the latent distance if we can also estimate the average of random effects (ν). Let $|W|$ denote the size of $W \in \{X, Y\}$.

Lemma 4.2.7. *Assume the latent distance model as in equation (4.9) and let C_ℓ denote the event that a collection of nodes indexed by $i \in \{1, \ldots, \ell\}$ form a clique of size ℓ.*

Let $\mu_d := \mathbb{E}[d(Z_i, Z_j)]$ denote the average distance between any two sampled latent positions Z_i and Z_j which are sampled independently from F_Z. If this latent density satisfies the following condition

- *(D1) F_Z admits a continuous density (f_Z)*

then for any $0 < \widetilde{\mu}_d < \mu_d$,

$$\max_{i,j} d(Z_i, Z_j) | C_\ell = o_P(\exp(-\widetilde{\mu}_d \ell / p)).$$

See the supplementary materials Section C.1.6 for the proof of this lemma. The main implication is that it is reasonable to treat the latent positions as a single point when nodes are within a clique. The assumption of F_Z admitting a smooth continuous density (f_z) is extremely mild. In fact, one could derive even faster rates if the latent density contained point masses.

An similar result can be derived for the node-level random effects.

Lemma 4.2.8. *Assume the latent distance model as in equation (4.9) and let C_ℓ denote the event that a collection of nodes indexed by $i \in \{1, \ldots, \ell\}$ form a clique of size ℓ. Suppose that*

1. *(E1) F_ν admits a continuous density (f_ν) on $(-\infty, 0]$ and this density function is positive at 0, $f_\nu(0) > 0$*

Let $\mu_\nu := E[\nu]$. Then for any $\mu_\nu < \widetilde{\mu}_\nu < 0$

$$\min_i \nu_i | C_\ell = o_P\left(\exp\left(-|\widetilde{\mu}_\nu| \ell\right)\right) \tag{4.10}$$

This theorem states that the nodes we find in a clique tend to have near-zero random effects. This will be useful as the random effects will converge to zero, allowing one to use between-clique connections to estimate distances. We next leverage each of these results to construct an estimator for a set of distances on the underlying manifold.

Due to the localization of the latent positions and random effects within a clique, one can estimate a set of distances using the average connection probability across cliques. Let $\mathcal{X}, \mathcal{Y} \subseteq V$ be subsets of vertices that denote nodes that form cliques respectively. We define the average probability of connection across cliques $\mathcal{X}, \mathcal{Y}$ by $p_{\mathcal{X}\mathcal{Y}}$. Based on the results of Lemma 4.2.7, we note that the maximum distance between a set of two points within the same clique is $o_P(\exp(-\widetilde{\mu}_d \ell))$, therefore we let $d_{\mathcal{X}\mathcal{Y}}$ denote the average distance in the latent

space between latent positions of the cliques. For any $x, y \in \mathcal{X}, \mathcal{Y}$, $d_{xy} = d_{\mathcal{X}\mathcal{Y}} + o_P(\exp(-\widetilde{\mu}_d \ell)$ and hence the pairwise distances between nodes in a clique are nearly identical. We can therefore relate the average connection probability to the average distance across cliques:

$$\begin{aligned}
p_{\mathcal{X}\mathcal{Y}} &:= \frac{1}{|\mathcal{X}||\mathcal{Y}|} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{xy} \\
&= \frac{1}{|\mathcal{X}||\mathcal{Y}|} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \exp(\nu_x + \nu_y - d_{xy}) \\
&= \exp(-d_{\mathcal{X}\mathcal{Y}}) \frac{1}{|\mathcal{X}||\mathcal{Y}|} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \exp(\nu_x + \nu_y) \left(1 + o_P(\exp(-\widetilde{\mu}_d \ell))\right) \\
d_{\mathcal{X}\mathcal{Y}} &= -\log(p_{\mathcal{X}\mathcal{Y}}) + \gamma_{\mathcal{X}} + \gamma_{\mathcal{Y}} + o_P(\exp(-\widetilde{\mu}_d \ell)) \\
\text{Where } \gamma_{\mathcal{W}} &:= \log\left(\frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \exp(\nu_w)\right) \qquad \text{for } \mathcal{W} \in \{\mathcal{X}, \mathcal{Y}\}.
\end{aligned}$$

Since there are ℓ^2 possible connections between a pair of cliques of size ℓ, we can construct an asymptotically normal estimator of $p_{\mathcal{X}\mathcal{Y}}$ using the sample mean of the connections across cliques. By Lemma 4.2.7, the distance between latent positions within a clique is negligible. Our goal is to develop an asymptotically normal estimate of $d_{\mathcal{X}\mathcal{Y}}$. We can achieve this by estimating $\gamma_{\mathcal{X}}$ and $\gamma_{\mathcal{Y}}$ at sufficiently fast rates (i.e., at least $o_P(\frac{1}{\ell})$) so that these are negligible compared to the estimation of $p_{\mathcal{X}\mathcal{Y}}$, for which we can leverage a central limit theorem.

For $\mathcal{W} \in \{\mathcal{X}, \mathcal{Y}\}$, in order to estimate $\gamma_{\mathcal{W}}$, we consider the average connection probability to any node in the network, conditional on a node being in a clique of size ℓ:

$$\begin{aligned}
P(A_{ik} = 1 | C_\ell) &= \exp(\nu_i) \int \exp(\nu_k + d(z_i, z_k))\, dF_Z(z_k | C_\ell)\, dF_\nu(\nu_k | C_\ell) \\
\frac{P(A_{ik} = 1 | C_\ell)}{P(A_{jk} = 1 | C_\ell)} &= \exp(\nu_i - \nu_j) + o_P(\exp(-\widetilde{\mu}_d \ell)).
\end{aligned}$$

By Lemma 4.2.7 $d(Z_i, Z_j) = o_P(\exp(-\widetilde{\mu}_d \ell))$. Estimation is straightforward by considering the density of connections, namely: $\widehat{P}(A_{ik} = 1 | C_\ell) = \frac{d_i}{n}$. Since we only use the ratio to compute the estimates of $\exp(\nu_i - \nu_j)$, the total number of nodes n is not necessary. Given $\ell \ll n$, where n is the number of nodes in the network, this ratio can be estimated easily using

the degree ratio. We define $\Delta\nu_i := \nu_i - \nu_{\mathcal{W}}$, allowing us to identify $\Delta\nu_i$, where $\nu_{\mathcal{W}} = \max_{i\in\mathcal{W}} \nu_i$

$$\gamma_{\mathcal{W}} = \nu_{\mathcal{W}} + \log\left(\frac{1}{|\mathcal{W}|}\sum_{i\in\mathcal{W}} \exp\left(\Delta\nu_i\right)\right).$$

The remaining question is to estimate $\nu_{\mathcal{W}}$. Fortunately, as we have previously discussed in Lemma 4.2.8, within a clique, the random effects approaches 0 at an exponentially fast rate. Hence, an estimator for the random effect can be constructed by setting the largest degree node's random effect to 0:

$$\widehat{\gamma}_{\mathcal{W}} = \log\left(\frac{1}{|\mathcal{W}|}\sum_{i\in\mathcal{W}} \frac{d_i}{\max_{j\in\mathcal{W}} d_j}\right). \tag{4.11}$$

This in turn can be used to estimate the distances:

$$\widehat{d}_{\mathcal{X}\mathcal{Y}} = -\log(\widehat{p}_{\mathcal{X}\mathcal{Y}}) + \widehat{\gamma}_{\mathcal{X}} + \widehat{\gamma}_{\mathcal{Y}} \tag{4.12}$$

where $\widehat{p}_{\mathcal{X}\mathcal{Y}} = \frac{1}{|\mathcal{X}||\mathcal{Y}|}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} A_{xy}$. The asymptotic distribution is next described in Theorem 4.2.9.

Theorem 4.2.9. *Let $\widehat{d}_{\mathcal{X}\mathcal{Y}}$ the distance estimator as per equation (4.12). Suppose that $\mathcal{X}, \mathcal{Y}$ are cliques of size ℓ and $X \cap Y = \emptyset$. If the following conditions hold:*

- *(F1)* $\ell = o(\sqrt{n})$
- *(F2)* $\lim_{\ell\to\infty} \frac{\sum_{x\in X}\sum_{y\in Y}(A_{xy}-p_{xy})^2 I(|A_{xy}-p_{xy}|>\epsilon\ell^2\sigma_\ell)}{\sigma_\ell} = 0$ *for all* $\epsilon > 0$

and we denote

$$p_{\mathcal{X}\mathcal{Y}} = \frac{1}{\ell^2}\sum_{x\in X}\sum_{y\in Y} p_{xy} \tag{4.13}$$

$$\sigma_\ell = \frac{1}{\ell^2}\sum_{x\in X}\sum_{y\in Y} p_{xy}(1-p_{xy}). \tag{4.14}$$

Then

$$\sqrt{\ell^2\frac{\sigma_\ell}{p_{\mathcal{X}\mathcal{Y}}}}\left(\widehat{d}_{\mathcal{X}\mathcal{Y}} - d_{\mathcal{X}\mathcal{Y}}\right) \to_d N(0,1) \tag{4.15}$$

The condition (F1) simply states that the size of the cliques should grow at a rate slower than $\sqrt{n}$, the total size of the network, which is a very mild assumption. This is to ensure that the asymptotics associated with the estimation of $\gamma_{\mathcal{W}}$ are negligible. Condition (F2) is the standard Lindeberg CLT condition.

If the node-level probabilities of connection p_{xy} is bounded away from 0 and 1, then this will hold (this may not hold if $d_{x\mathcal{Y}} \to -\infty$). The full proof is found in Section C.1.8. The main implication is that the asymptotic distribution is driven by the estimator $\widehat{p}_{x\mathcal{Y}}$. Since the random effects will approach 0, for large cliques $p_{x\mathcal{Y}} \approx \exp(-d_{x\mathcal{Y}})$, thus allowing for simplified expressions for σ_ℓ and $p_{x\mathcal{Y}}$.

Constrained Estimation. Pairwise estimation of the distances according to equation (4.12) is one method of estimating a distance matrix, however, this does not restrict the final estimate to be a metric. When cliques are not connected to one-another, this may result in distance estimates which are $-\infty$.

For example, we construct an enumeration of the cliques of size 7 or greater from the General Relativity co-authorship network of Leskovec et al. [2007] in Figure 4.6. Any pair of cliques which does not share an edge will inherently be estimated to have infinite distance, which prevents the estimation of curvature.

To address this challenge, we posit a similar estimation problem, while respecting the triangle inequality for each triplet. The following estimation problem is posed below. Let $\mathcal{C}$ denote the set of cliques in an observed graph. Let $\mathcal{X} \subset \mathcal{C}$ be a set of indices corresponding to a clique. We estimate the random effects within a clique $i \in \mathcal{X}$ according to $\widehat{\nu}_i = \log(\frac{d_i}{\max_{j\in\mathcal{X}} d_j})$. Given a set of fixed effects, we can propose a maximum likelihood optimization problem for the distance matrix $D \in \mathbb{R}^{K\times K}$. As by Lemma 4.2.7 cliques have approximately a common latent position. From our latent distance model, we define the following likelihood function for the connections within our set of cliques $\mathcal{C}$, $\mathcal{L}_{\mathcal{C}}$:

$$\mathcal{L}_{\mathcal{C}}(D;\boldsymbol{\nu}) := \sum_{\mathcal{X},\mathcal{Y}\in\mathcal{C},i\in\mathcal{X},j\in\mathcal{Y}} A_{ij}\left(\nu_i+\nu_j-d_{x\mathcal{Y}}\right) + \sum_{\mathcal{X},\mathcal{Y}\in\mathcal{C},i\in\mathcal{X},j\in\mathcal{Y}} (1-A_{ij})\log\left(1-\exp\left(\nu_i+\nu_j-d_{x\mathcal{Y}}\right)\right).$$

We can now define the maximum likelihood optimization problem, after estimating a set of

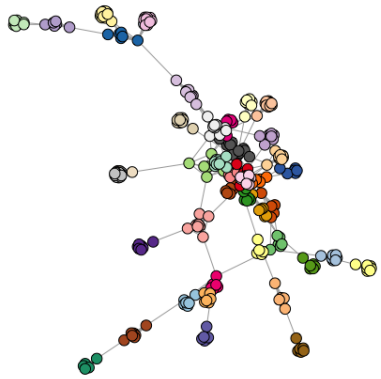


Figure 4.6: Clique subgraph of co-authorship network in ArXiv General Relativity. Cliques of size 7 and greater are shown by color.

random effects $\widehat{\boldsymbol{\nu}}$

$$\widehat{\boldsymbol{D}} = \sup_{D \in \mathbb{R}^{K \times K}} \mathcal{L}_{\mathcal{C}}(D; \widehat{\boldsymbol{\nu}})$$
$$\text{s.t. } d_{xy} \geq 0, \quad \text{Diag}(D) = 0, \quad \text{tr}(E_s^\top D) \geq 0 \quad \forall s \in \mathcal{S}$$

where $\mathcal{C}$ is a list of clique indices. E_s are matrices which define the triangle inequalities and contain mostly 0's other than 3 indices, i, j, k for which $E[i,j] = E[i,k] = 1 = -E[j,k]$. We define $\mathcal{S}$ to be an enumeration of all such matrices E_s. There are $3\binom{K}{3}$ such restrictions in total. The set of feasible distance matrices that satisfy these constraints form a polyhedron of interior dimension $\binom{K}{2}$, therefore therefore the constraints do not reduce the dimensionality of the space of distance matrices, but reduce the volume of the space to a smaller polyhedron. This optimization procedure is a *convex* and in practice, we may use `CVXR` to solve this system [Fu et al., 2020]. For a greater gain in computational speed, we use the `MOSEK` solver for the constrained optimization [Andersen and Andersen, 2000].

Though in its current form, the problem is numerically challenging to solve due to the sheer number of constraints. A natural option is to use the Newton method and approximate the likelihood using a second order Taylor expansion and solve this problem successively. Since the Hessian is diagonal, $\frac{\partial^2}{\partial d_{ij}\partial d_{kl}}\mathcal{L}_{\mathcal{C}}(D;\boldsymbol{\nu}) = 0$ if $d_{ij} \neq d_{kl}$, this can be made into a more efficient quadratic program which can be solved faster in `CVXR`. Let $\widetilde{g}(D; D_0; \boldsymbol{\nu})$ be the second order Taylor series approximation to $\mathcal{L}_{\mathcal{C}}$ about a matrix D_0. We can successively solve for $\widehat{\boldsymbol{D}}_{t+1}$ using the following constrained optimization problem

$$\begin{aligned} \widehat{\boldsymbol{D}}_{t+1} = & \sup_{D \in \mathbb{R}^{K\times K}} \widetilde{g}(D, \widehat{\boldsymbol{D}}_t; \widehat{\boldsymbol{\nu}}) \\ \text{s.t.} d_{\mathcal{X}\mathcal{Y}} \geq 0, & \quad \text{Diag}(D) = 0, \quad \text{tr}(E_s^\top D) \geq 0 \quad \forall s \in \mathcal{S} \end{aligned} \tag{4.16}$$

which can be iteratively computed until $\mathcal{L}_{\mathcal{C}}$ increases less than some threshold. Each iteration is a linear constrained quadratic program, a common convex problem for which standard software has implemented faster solutions. We implement this in `CVXR`. For further details, see the supplementary materials in Section C.2.2.

In practice, running the optimization step for many iterations can be computationally costly. This is particularly problematic in a later application we discuss which involves a subsampling approach to testing for constant curvature (Section 4.4). It is well known in maximum likelihood estimation, one only needs to perform one Newton step for asymptotic efficiency [Le Cam, 1956]. As a result, one can start with any consistent estimator of the distance matrix D and apply a single Newton step from equation (4.16) and obtain the same asymptotic distribution, and therefore in practice, only one step is needed. In the appendix C.3.2 we include an alternative approach for estimating distance matrices under an alternative framework, without the need for cliques. In practice, we can use the method in equation (4.12) to form an initial estimate, then modify the entries so that it is a metric. One example of an algorithm which can be used for this purpose is the Floyd-Warshall Algorithm. This algorithm modifies as few entries as possible, solving a problem also known as sparse metric repair [Gilbert and Jain, 2017]. For full details, see the appendix Section C.2.2.

We now present the entire procedure for estimating latent curvature from a network in Algorithm 8.

Algorithm 8 Algorithm for estimation $\widehat{\kappa}$.

Require: G, Steps T, Triangle constant $C_\triangle$, e.g. $C_\triangle = 1.4$. Minimum clique size ℓ.

1: Find a set of $\mathcal{C}$ where $K = |\mathcal{C}|$ and $C \in \mathcal{C}$ where $C \subset G$ such that $|C| \geq \ell$.

2: Estimate the initial distance matrix estimate as per equation (4.12) and denote this $\widehat{\boldsymbol{D}}^{(-1)}$

3: Apply the sparse metric repair algorithm (i.e. the Floyd-Warshall algorithm to modify $\widehat{\boldsymbol{D}}^{(-1)}$ so that it is a metric and denote this $\widehat{\boldsymbol{D}}^{(0)}$.

4: Estimate the distance matrix $\widehat{\boldsymbol{D}}$ by iterating equation (4.16) T times.

5: Rank the surrogate midpoint sets and identify the best midpoint set $\widehat{y}, \widehat{m}, \widehat{z}$ by applying equation (4.6).

6: Select values of $x, \mathbf{X}$ such that equation (4.7) is satisfied.

7: Estimate $\widehat{\kappa}$ by taking the median over the set $\boldsymbol{X}$ as per equation (4.8).

8: **return** $\widehat{\kappa}$

In the appendix (Section C.3.2), we also illustrate an approach to estimating distance matrices in a similar setting in the absence of cliques. We now continue with an empirical study of our estimators.

4.3 Simulation Study

We construct a number of simulations to illustrate the performance of our curvature estimator under the full model complexity. This involves first sampling the parameters of a latent position cluster model to draw locations and variances independently, then sampling positions and random effects from the random latent position cluster model. This is to illustrate how midpoints may naturally align from random draws of the centers of the cluster model. We

simulate from the following model:

$$\mu_k \sim F_\mu^{\mathcal{M}^p(\kappa)}(\boldsymbol{O}, R) \qquad k \in \{1, \ldots, \lceil \sqrt{\rho} 50 \rceil\}$$
$$\sigma_k^2 \sim F_{\sigma^2} \qquad \boldsymbol{\pi} \sim Dirichlet(\mathbf{2})$$
$$\rho \in \{\frac{1}{\sqrt{2}}, 1, 2\} \qquad \kappa \in \{-2, -1, -0.5, 0, 0.5, 1\}$$

where $F_\mu^{\mathcal{M}^p(\kappa)}(\boldsymbol{O}, R)$ denotes the prior distribution on the latent positions, which is a uniform ball with radius R for $\kappa > 0$, and two concentric balls, one with radius R and the other with $R/2$, with equal probability for $\kappa \leq 0$. This setup facilitates forming midpoints since the volume of a ball grows exponentially as κ decreases. In all simulations, we set $R = 2.5$. This process determines a random latent position cluster model (LPCM) similar to Handcock et al. [2007]. The locations of μ_k as well as the relative sizes of σ_k^2 determine where cliques are most likely to form in the latent space. The parameter ρ refers to the scale of the network simulations, allowing for the number of centers to grow with ρ, $\boldsymbol{\pi}$ refers to the vector of cluster probabilities in the mixture model, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ refer to the cluster mean scale parameters, and κ the true curvature of the space. The parameters $\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\pi}$ determine a latent position cluster model, where the positions, $Z_i \sim F_Z := F_Z(\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\pi}; \kappa)$. The mixture components correspond to the heat kernels in spherical space (Von-Mises Fischer distribution), Gaussian distribution in Euclidean space, and the wrapped normal distribution in hyperbolic space [Nagano et al., 2019]. In all cases, we let $p = 3$. The randomness in the latent position cluster model incorporates the fact that we are not assigning good midpoints exactly but finding them in the data each time we simulate a matrix. We let F_{σ^2} denote the gamma distribution with shape parameter 1/16.

We then sample the networks according to the draw of the latent position cluster model:

$$n = 5000\rho, \quad Z_i \sim F_Z, \quad \nu_i \sim F_{\nu,\rho}$$
$$p_{ij} = \exp(\nu_i + \nu_j - d(Z_i, Z_j))$$
$$A_{ij} \sim \text{Bern}(p_{ij})$$

where $F_Z = F_Z(\boldsymbol{\pi}, \boldsymbol{\sigma^2}, \boldsymbol{\mu})$ is a particular draw of the random latent position cluster model. The random effects distribution $F_{\nu,\rho}$ is a trimmed normal distribution with mean -3ρ and standard deviation 3ρ trimmed at 0 and $\log(2/\sqrt{n})$ so that ν_i remain non-negative and to prevent isolated nodes. We let the minimum clique size used in our estimator be $\ell = (8+4\log(\rho))$, which tends to generate $30-40$ cliques. We repeat this 500 times for each scale and curvature setting. In practice, we find the cliques using the `maximal_cliques` function in the `igraph` R package [Csardi and Nepusz, 2006]. See Section C.6 for additional graph statistics summaries over the simulations. We provide the values of the tuning parameter $C_\triangle$ in the appendix in Section C.7.

4.3.1 Consistency of Simulation Curvature Estimates

We now explore the consistency of our curvature estimate as a function of the curvature of the latent space. In each of the simulations, we limit the number of cliques used in the estimator to 35 (50 for $\kappa \leq -1$) for computational convenience for numerous simulations; though in real data applications, this number can be larger.

We plot the results in Figure 4.7. We see that, as the clique size increases, there is a reduction in bias and variance. We note observe that as the curvature more negative, the estimator has greater variability. This is for two reasons. Firstly, as we saw in Figure 4.3, the variance of the estimate is simply larger when κ is negative in nearly all regions of the space, a small variation in the length of a triangle median corresponds to a large variation in the corresponding curvature. Secondly, due to the vastness of the negatively curved spaces, we tend to have poorer quality midpoints form as well as fewer reference points x which form nearly equilateral triangles.

4.4 Applications: Testing and Detecting Differences in Curvature

We now return to the problem of testing whether the latent space is one of constant curvature, i.e. one of our canonical manifolds. Formulated as a statistical test, we write this null

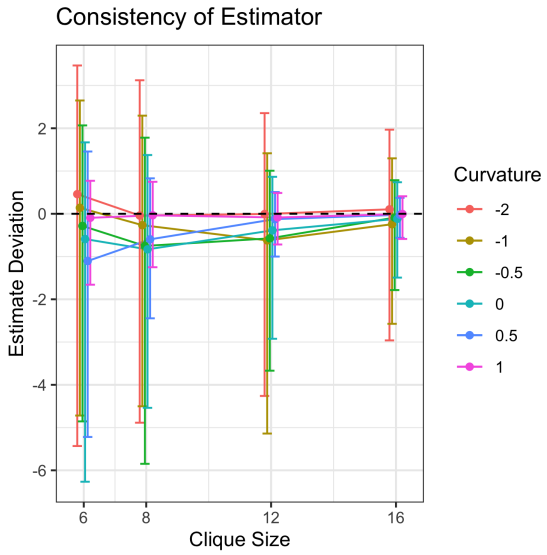


Figure 4.7: Consistency of Curvature Estimator. Upper error bars indicate the 0.95 quantile of simulations and lower indicate the 0.05 quantile. Central dots indicate the mean after trimming outliers larger than ± 100 (0.107% of the observations).

hypothesis as follows,

$$H_0 : \mathcal{M} = \mathcal{M}(\kappa).$$

This test could provide a model diagnostic (e.g. testing whether a latent variable model that assumes a single constant manifold is appropriate) or provide meaningful insights into heterogeneity in the structure of the graph.

Our test of constant curvature will rely on the upper and lower bounds of the curvature, as developed in Section 4.2.3. We do this by leveraging a set of midpoint complexes (i.e. $(y^{(j)}, m^{(j)}, z^{(j)})$), and computing a set of corresponding upper and lower bounds on the curvature at each set.

Our test will require the sampling distribution of the distance matrix estimator $\widehat{\boldsymbol{D}}$. We consider a method of subsampling from the distribution of the cliques in order to approximate the sampling distribution of the distance matrix estimator. This is based on the subsampling

approach of Politis and Romano [1994], which can be used to approximate the distribution of a random variable using subsampling under conditions weaker than the bootstrap, similar to the strategy in Lubold et al. [2023]. This is highlighted in Algorithm 9. For simplicity, we illustrate the algorithm where we use subsamples of size $\ell - 1$ of each of the cliques.

Let $\mathcal{I} = \{\mathbf{I}_k\}_{k=1}^K$ denote a set of indices corresponding to the cliques, where $\mathbf{I}_k \cap \mathbf{I}_{k'} = \emptyset$ when $k \neq k'$. Let $\widehat{\nu}_{\mathbf{I}_k}$ denote the set of estimated random effects corresponding to the clique $\mathbf{I}_k$. In Algorithm 9 f_0 denotes the distance estimation by each of the distance separately

Algorithm 9 Sub-sampling Procedure to Approximate the Sampling Distribution of $\widehat{\boldsymbol{D}}$

Require: G, $B \geq 1$ and $\{\mathbf{I}_k\}_{k=1}^K$

1: **for** $b \in \{1, 2, \ldots, B\}$ **do**
2: **for** $k \in \{1, 2, \ldots, K\}$ **do** Sample $|\boldsymbol{I}_k| - 1$ nodes without replacement from $\boldsymbol{I}_k$ and denote this set $\boldsymbol{I}_k^{(b)}$
3: **end for**
4: Denote the set $\mathcal{I}^{(b)} = \{\mathbf{I}_k^{(b)}\}_{k=1}^K$
5: Let $\nu^{(b)} = \nu_{\mathcal{I}^{(b)}}$ denote the corresponding random effects estimates.
6: Let $\widehat{\boldsymbol{D}}_0^{(b)} = f_0(A, \mathcal{I}^{(b)}, \nu^{(b)})$ Initial Estimate
7: Let $\widehat{\boldsymbol{D}}_1^{(b)} = f_1(\widehat{\boldsymbol{D}}_0^{(b)}; A, \mathcal{I}^{(b)}, \nu^{(b)})$ One-step Estimate
8: **end for**
9: **return** $\{\widehat{\boldsymbol{D}}^{(b)}\}_{b=1}^B = \{\widehat{\boldsymbol{D}}_1^{(b)}\}_{b=1}^B$

as per equation 4.12 and subsequent sparse modification of the distances using the Floyd-Warshall Algorithm so that $\widehat{\boldsymbol{D}}_0$ is a metric. The subsequent step f_1 refers to applying a one-step estimation procedure of equation (4.16).

We now operationalize our test of constant curvature. Consider a set of midpoint sets $y'^{(j)}, j \in \{1, 2, \ldots, J\}$ with corresponding reference points $\boldsymbol{X}^{(j)}$. We test for whether the curvature is constant across the latent space across each of these midpoint sets

$$H_0 : \kappa_j = \kappa_{j'} \text{ for all } j, j' \in \{1, 2, \ldots, J\}$$

using Algorithm 10. In order to choose the corresponding locations we utilize the selection of midpoint sets for J collections of midpoints as illustrated in 4.2.4.

Algorithm 10 Constant Curvature Test

Require: $\{\widehat{\boldsymbol{D}}^{(b)}\}_{b=1}^{B}$ and $\{y^{(j)}, z^{(j)}, m'^{(j)}, \boldsymbol{X}^{(j)}\}_{j=1}^{J}$

for $b \in \{1, 2, \ldots, B\}$ **do**
  for $j \in \{1, 2, \ldots, J\}$ **do**
    for $x \in \{\boldsymbol{X}^{(j)}\}$ **do**
      $\widehat{\kappa}_{u,x}^{(b,j)} = \kappa_u(\widehat{\boldsymbol{D}}^{(b)}; y^{(j)}, z^{(j)}, m'^{(j)}, x)$
      $\widehat{\kappa}_{l,x}^{(b,j)} = \kappa_l(\widehat{\boldsymbol{D}}^{(b)}; y^{(j)}, z^{(j)}, m'^{(j)}, x)$
    end for
    Compute $\widehat{\kappa}_u^{(b,j)} = \text{median}_{x \in \boldsymbol{X}^{(j)}} \widehat{\kappa}_{u,x}^{(b,j)}$
    Compute $\widehat{\kappa}_l^{(b,j)} = \text{median}_{x \in \boldsymbol{X}^{(j)}} \widehat{\kappa}_{u,x}^{(b,j)}$
  end for
  Compute $\widehat{\kappa}_u^{(b)} = \min_{j \in \{1,2,\ldots,J\}} \widehat{\kappa}_u^{(b,j)}$
  Compute $\widehat{\kappa}_l^{(b)} = \max_{j \in \{1,2,\ldots,J\}} \widehat{\kappa}_l^{(b,j)}$
end for
Let $\widehat{\kappa}_{u,(m)}$ be the m^{th} order statistic of $\{\widehat{\kappa}_u^{(b)}\}_{b=1}^{B}$
Let $\widehat{\kappa}_{l,(m)}$ be the m^{th} order statistic of $\{\widehat{\kappa}_l^{(b)}\}_{b=1}^{B}$
Let $m = \min\{m : \widehat{\kappa}_{u,(m)} \geq \widehat{\kappa}_{l,(B-m)}\}$
return p-value: $\min\{2m/B, 1\}$

The constant curvature test involves analyzing a sampling distribution of distance matrices $\widehat{\boldsymbol{D}}^{(b)}{}_{b=1}^{B}$ derived from Algorithm 9. It also uses collections of surrogate midpoint sets $\{y^{(j)}, z^{(j)}, m'^{(j)}\}_{j=1}^{J}$ and the corresponding favorable triangle reference points $\{\boldsymbol{X}^{(j)}\}_{j=1}^{J}$. These sets help estimate the upper and lower bounds of curvature. Practically, each set $\boldsymbol{X}^{(j)}$ is chosen using the method described in equation (4.7). To reduce the variance of these upper and lower bound estimates across surrogate midpoint sets, we use the median across the reference points $\boldsymbol{X}^{(j)}$. This then gives us a collection of upper and lower bound estimates of the curvature across regions $\{\widehat{\kappa}_l^{(j)}, \widehat{\kappa}_u^{(j)}\}$. If the minimum of the upper bounds is less than the maximum of the lower bounds then this corresponds to a distance matrix that cannot

be represented by a single curvature. This process is repeated across the set $b \in 1, 2, \ldots, B$ and the proportion of times these quantities cross can then be interpreted as a p-value for the constant curvature test.

In this test there exists an inherent trade-off: achieving tightly aligned midpoint sets provides better upper and lower bounds on curvature estimates, yet it is also essential to sample from regions of the latent space sufficiently distant to potentially reveal differences in curvature. The analysis primarily focuses on optimizing the choice of midpoint sets on the latent space, as the sampling adequacy in regions with varying curvature typically lies outside of the analyst's direct control.

4.4.1 Simulations: Type 1 Error Control

Under the same simulation setup as in Section 4.3 we can illustrate the coverage of the constant curvature test as a function of clique size. For computational convenience, we restrict these to a maximum of size 12. We see in Figure 4.8 that this test tends to be overly conservative in small sample sizes, but returns to nearly nominal coverage when cliques are larger. This is due to the fact that a poorly aligning midpoints lead to more conservative bounds of κ_l, κ_u, however, better aligning midpoints are found in larger networks with more cliques.

4.4.2 Multiplex Networks

In this next simulation, we consider a model of multiplex (or multiview) networks. Several methods exist for modelling multiplex networks via extensions of the latent distance model, however, most assume the same geometry latent space among views [Salter-Townshend and McCormick, 2017, MacDonald et al., 2022]. For example, Salter-Townshend and McCormick [2017] model the multiple relationships between individuals in the Banerjee et al. [2013a] diffusion of microfinance dataset using Euclidean spaces. We illustrate a simulated example where this is not the case, and how our method can be used to detect this.

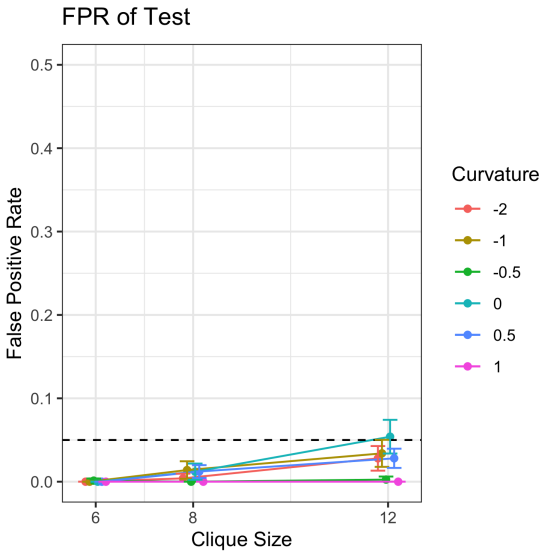


Figure 4.8: False positive rate for constant curvature test.

We first construct a latent position model for which multiple views are drawn from a common set of latent positions, however, these positions are common coordinates of spheres of curvature $(\kappa_1, \kappa_2) = (0.5, 1.5)$ respectively. Additional details for the simulation are identical to the consistency simulations in Section 4.3.

In this example, we simulate a multiplex network for which latent spherical positions are the same for this network set, however, they are embedded in two spheres of different radii and thus different curvatures. We simulate 200 draws of these networks and test the curvature difference in curvature using Algorithm 10. We plot the power of the test in Figure 4.9. In each view we compute the optimal surrogate midpoint set $y^{(j)}, z^{(j)}, m^{(j)}$ from each view's distance matrix $\widehat{\boldsymbol{D}}^{(1)}$ or $\widehat{\boldsymbol{D}}^{(2)}$ using equation 4.6 and subsample each distance matrix accordingly. In Figure 4.9 we observe that the power of the test grows with the clique size.

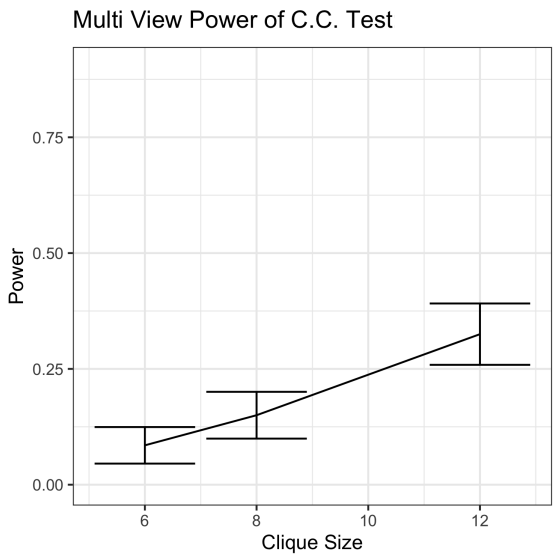


Figure 4.9: Constant curvature test power applied to the multiview network example.

4.4.3 Noncanonical Manifolds

We now demonstrate how our method can detect non-constant curvature in the latent space. Given that many prevalent latent distance models inherently assume constant curvature, it becomes crucial to confirm if this assumption aligns with the actual data observed.

We construct a latent space consisting of two adjacent spheres. For any two points in the same sphere the distance is straightforward to compute. For any two points (x, y) in opposite spheres, the distance can be computed using the distance to the origin $(1, 0, 0)$ in each of the spheres. Since any geodesic must pass through the connecting point, i.e. the origin we can compute these distances as follows

$$d_{\mathcal{M}}(x, y) = d_1(x, o) + d_2(o, y).$$

This manifold was chosen as distances were straightforward to compute but come from a space without constant curvature. These spheres have curvature $\kappa_1 = 1, \kappa_2 = 1.5$ respectively.

The latent cluster locations are sampled according to uniform distributions centered

on opposite poles. We again simulate 200 draws from the latent position cluster model and test for constant curvature by finding the best three non-overlapping sets minimizing equation (4.6). We plot the corresponding power in Figure 4.10 which once again, increases along with the clique size.

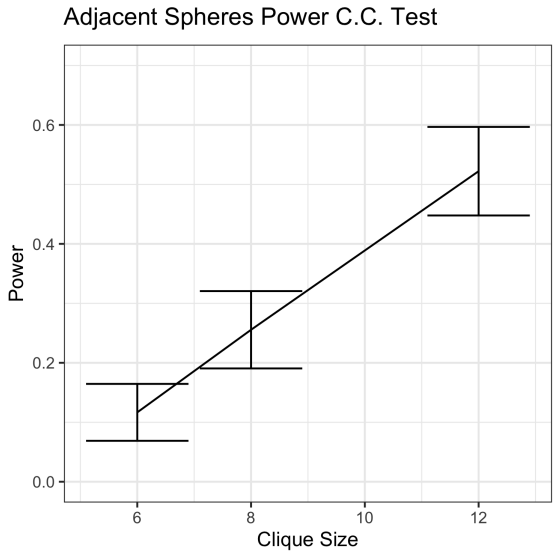


Figure 4.10: Power of the constant curvature test applied to the adjacent spheres simulation, an example of a non-canonical manifold, which does not have constant curvature.

4.4.4 Data Analysis: Geometry of Coauthorship

We apply our method to a collection of co-authorship networks in physics introduced in Leskovec et al. [2007] and available on the `Stanford Network Analysis Project (SNAP)` repository. These consist of citation networks from High Energy Particle Physics, General Relativity, Astrophysics, and Condensed Matter Physics, with sizes of each of the networks seen in Table 4.1. These networks consist of authors as nodes where an edge exists whether any pair of authors has co-authored a paper on ArXiv in any of the specified subject areas between 1993 and 2003. When these data were collected, these were the top 5 most common

subject areas in Physics.

In previous machine learning applications, hyperbolic network embeddings have been successful in tasks such as link-prediction or node in citation networks [Nickel and Kiela, 2017, Chami et al., 2019, Chamberlain et al., 2017]. This is due to the fact that the hierarchical structure (tree-like) structure generally can be more easily embedded in a negatively curved space. We wish to answer the question "under a latent distance model, could the data be generated from a space of constant curvature?"

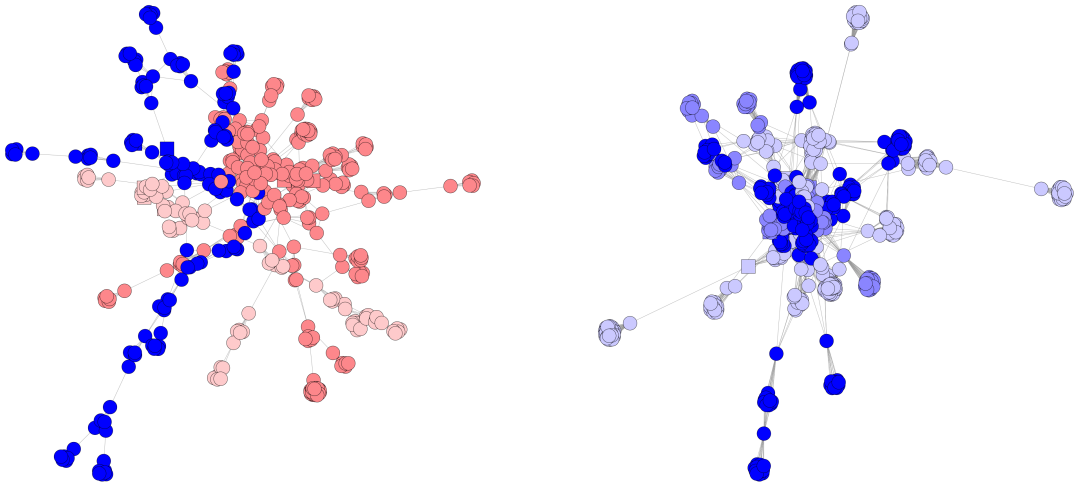
For each of these networks, we construct an estimate of the distance matrix with random effects, followed by an estimate of the curvature. For each of these, we use a minimum clique size seen in Table 4.2. We lastly apply our test to see if the difference in curvature is present across the networks. We estimate the curvature, at the best midpoint set for each of the networks, along with the following p-values for tests of whether a network has constant curvature.

Physics Sub-field	Number of Nodes (n)	Number of Edges ($\|E\|$)
Astrophysics	18771	396160
Condensed Matter Physics	23133	186936
General Relativity	5241	28980
High Energy Particle Physics	12006	237010
High Energy Particle Physics (Theory)	9875	51971

Table 4.1: Physics Co-authorship network sizes.

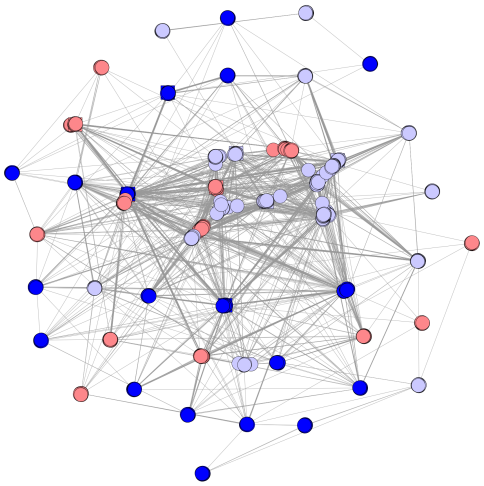
One concern a reader may have is whether good surrogate midpoints form in real data. In practice, we find good midpoints can be found in our real datasets. For example, the Fréchet mean objective function (4.5) has a value at the true midpoint of 0.5. In the 5 physics co-authorship networks this was found to be 0.5000003, 0.5039, 0.50049, 0.50012, 0.50043, suggesting good surrogate midpoints in practice.

We see that Astrophysics and General Relativity are estimated to have a large negative



(a) High Energy Particle Physics (Theory) clique curvature labels.

(b) High Energy Particle Physics clique curvature labels.



(c) Astrophysics clique curvature labels.

Figure 4.11: Cliques colored to nearest labeled curvature value. Blue, negative and red positive, with p values of constant curvature decreasing from left to right.

Physics Sub-field	Min Clique Size (ℓ)	Number of Cliques of Size $\geq \ell$ (K)	Largest Clique Size
Astrophysics	19	57	57
Condensed Matter Physics	12	52	26
General Relativity	7	44	44
High Energy Particle Physics	14	42	239
High Energy Particle Physics (Theory)	7	42	32

Table 4.2: Clique size and number of cliques used to estimate distance matrix.

Physics Sub-field	Curvature Estimate	Constant Curve Test p-value
Astrophysics	-0.01378	0.030
Condensed Matter Physics	0.107	1.000
General Relativity	0.0989	1.000
High Energy Particle Physics	-0.986	1.000
High Energy Particle Physics (Theory)	0.1674	0.240

Table 4.3: Curvature estimates from best midpoints and p-values for constant curvature test $J = 3$.

curvature, which however, may not be constant in the case of the Astrophysics citation network. In the Astrophysics network, at the best 3 surrogate midpoint sets, we estimate curvature to be $(-0.01378, -\infty, 0.306)$. The estimate of $-\infty$ comes from the fact there is a minimum distance of d_{xm} given the other 3, d_{xy}, d_{yz}, d_{xz}, and since the midpoints are not exact, sometimes this length can be too short to embed in any of the hyperbolic spaces. If the estimated distance is below this value we call the estimate $-\infty$, and similarly ∞ if it is too large. This highlights the fact that this network appears to have negatively curved,

positively curved and flat regions, and therefore models which reflect only a single curvature, may not capture the individual level behavior of the network very well. In contrast, condensed matter physics and HEP physics seem to have a slight curvature, though both are nearly flat networks. We remark on the large p-values for 3 of the networks in Table 4.3. In these settings, the second and 3rd midpoint sets proved to have poorer alignment and therefore were quite conservative when constructing the constant curvature test, leading to larger p-values.

4.5 Application: Multiple Change Point Detection

Another natural question one may seek to answer is whether a change in latent curvature has occurred. This is distinct from a problem of whether the latent positions may change. Standard multiple change point algorithms (for example that of Harchaoui and Lévy-Leduc [2010]) are not immediately well-suited to this problem due to the possibility of large outliers, which may occur in our setting, in particular for largely negative values. We apply the method of Fearnhead and Rigaill [2016] which proposes a multiple changepoint detection algorithm under the presence of outliers. For a particular network, we may measure a collection of estimates of the curvature $\{\widehat{\kappa}_t\}_{t=1}^T$. Under this model, we assume that a network has a constant curvature within a single time point t. We construct an objective function for the changepoint problem

$$L(\theta) = \sum_{t=1}^{T} \widetilde{\ell}(\widehat{\kappa}_t, \theta_t)$$

In our application, we let $\widetilde{\ell}$ be the bi-weight loss function, however, more general loss functions such as absolute deviation or Huber loss are also available (see Fearnhead and Rigaill [2016] for a more in depth discussion of robust change point detection). In multiple change point detection algorithms a penalty of β for the number of segments included is also applied. Let $J(\theta)$ denote a function of the number of breaks in the sequence θ. Then the full loss function is

$$L(\theta; \beta) = L(\theta) + \beta J(\theta). \tag{4.17}$$

Since we often care about understanding where the changes of curvature occur, we can apply a monotone transformation to the estimates of curvature. The function $c\tanh(\cdot/c)$ function which smoothly truncates the extreme values under a monotone transformation. In all simulations and applications where this is applied, we set $c = 10$.

We next apply this to a simulation setting where we construct a sequence of networks with latent positions evolving according to the following process

$$Z_i^{(t)} = \mathcal{F}(Z_i^{(t-1)}, \epsilon_i^{(t)})$$

where $Z_{(t-1)i}$ is the location's previous position, $\epsilon_i^{(t)}$ is a noise random variable sampled from the true cluster's density, and $\mathcal{F}$ stands for the mean on the sphere (the Fréchet mean). We set three different curvatures $\kappa_1 = 1.0 \quad \kappa_2 = 0.15 \quad \kappa_3 = 1.3$ to occur changepoints at $t_1 = 18, t_2 = 35$ and with the final time $T = 50$.

We use the implementation of the changepoint method in the `robseg` package introduced in Fearnhead and Rigaill [2016] using the default regularization parameters. We see that as clique size ℓ increases, we are able to consistently estimate the true curvature function. We plot the mean of the absolute deviation of the curvature estimate over 200 simulations and plot the results in Figure 4.12 showing consistency of the true curvature with respect to the mean absolute deviation.

4.5.1 Application: Cybersecurity

For our curvature changepoint application, we utilize our network curvature estimates on the Los Alamos Unified Network and Host Dataset to demonstrate that variations in curvature can help identify a red team attack; a controlled exercise where a cybersecurity team simulates an infiltration on a device network to test its security.

This dataset encompasses 89 days of directed communication events among 27436 devices at the Los Alamos National Laboratory. It records 56 normal operational days followed by a red team attack that spans from day 57 to day 89.

Anomaly detection holds significant importance in cybersecurity, and recent studies, such

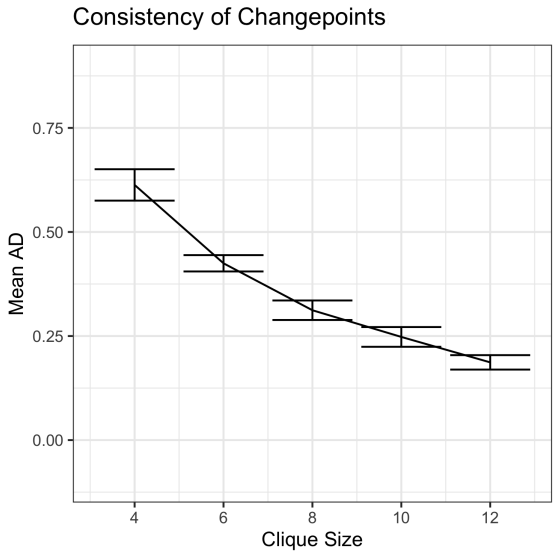
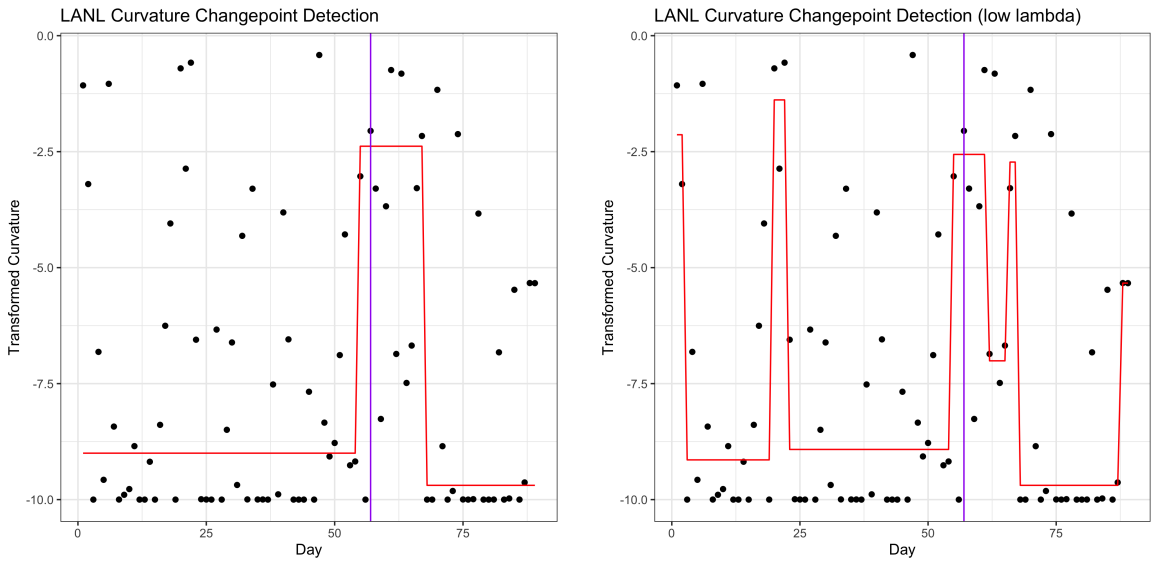


Figure 4.12: Mean absolute deviation of curvature estimates over time window.

as [Lee et al., 2019], have highlighted latent distance models as a promising method for detecting changes in node properties. In contrast, our research adopts changepoint methods applied to sequential curvature estimates within this dataset, underscoring the utility of curvature as a comprehensive indicator of network behavior.

Edges are defined in this dataset as messages passed between nodes during a particular time period. In order to maintain enough connections to find cliques, we consider a connection to be formed if any message was passed in the previous 4 days. We then compute the curvature values at each time point and take the median over the time-point. We scale each estimated value by $c\tanh(\cdot/c)$ for $c = 10$ in order to limit the influence of extreme negative outliers. We then apply the off the shelf change point algorithm of Fearnhead and Rigaill [2016]. We estimate the curvature at each time step. We consider a minimum clique size of $\ell = 5$. Due to the small number of available cliques, and relative sparsity of the dataset, we restrict the random effects to be 0 and compute the corresponding distance matrix.

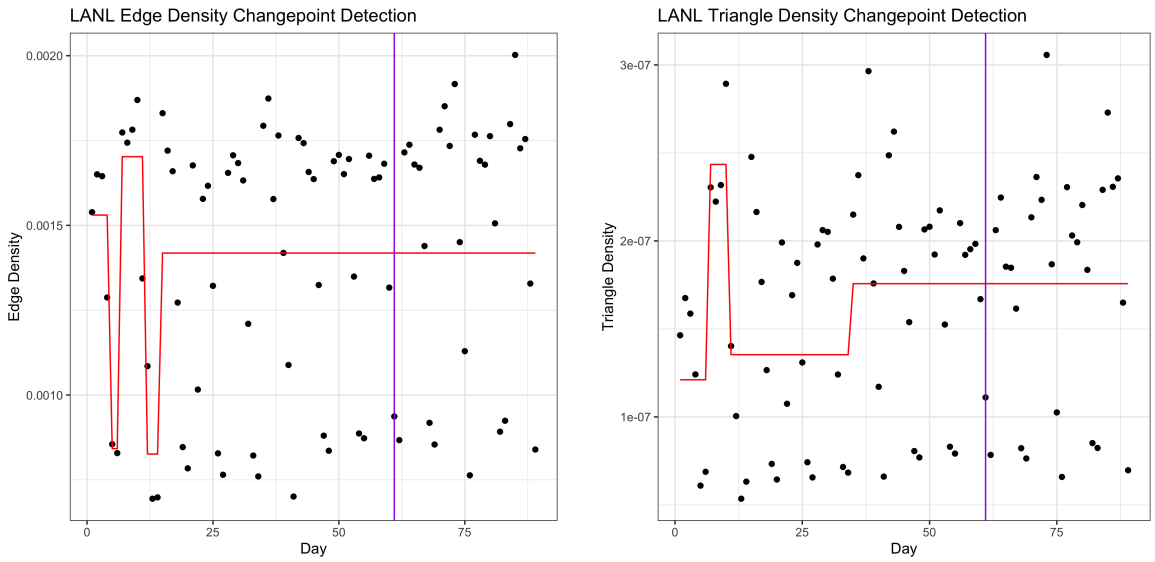
In Figure 4.13 we show that the most substantial changepoint in curvature occurs at the time of the red team attack. In contrast, these changes are much less substantial in



(a) High regularization changes in curvature estimates.

(b) Low regularization changes in curvature estimates.

Figure 4.13: Two regularization values for LANL Netflow changepoint dataset. True red team attack time is illustrated in purple.



(a) Edge density changepoints.

(b) Tri. density changepoints.

Figure 4.14: Changepoints of daily LANL measurements using simple graph motifs. Red team attack illustrated in purple.

Figure 4.14 when using simple graph motifs from the daily averages.

Since the time of detection after the event is most important, we wish to investigate the time after detection as a function of the number of events involved (alarm rate). We show that in Figure 4.15 that our method achieves a much smaller detection delay given any alarm rate.

This suggests that models accounting for changes in network curvature may be a promising avenue for the development of specialized models to detect anomalous events in the online setting.

4.6 Discussion

Riemannian sectional curvature is a fundamental property of a manifold, and we present a novel method to estimate it from a noisy distance matrix. Though our motivating example involves estimating the distances of a latent distance model from a random network, the

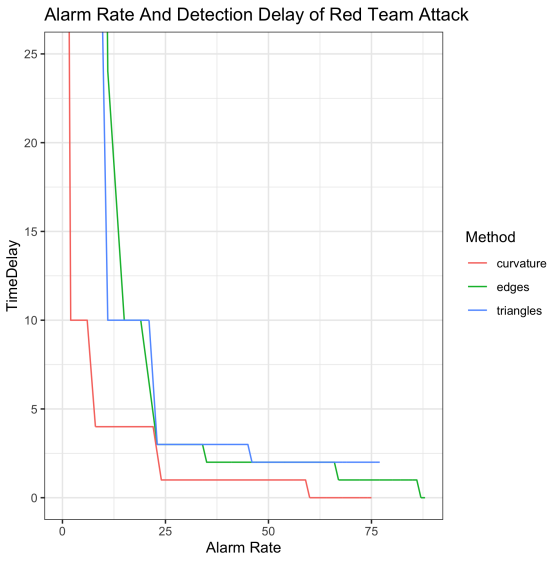


Figure 4.15: Time to first changepoint in days (TimeDelay) as a function of alarm rate.

curvature estimates (and constant curvature tests) in this paper are more general and can be applied whenever one can either estimate a distance matrix (or bootstrap or subsample their distance matrices).

We develop a test for detecting whether the curvature is constant on the latent manifold. A natural followup question is what should a practitioner do if they find that their data are not represented well by this model. One might instead use non-geometric models such as stochastic block models and their variants, however, this also suggests the development of latent distance models which are not restricted to constant curvature. The development of such a model class as something which will scale to large networks is of further interest. One promising approach we plan to investigate in future work is that of product spaces for latent distance models. This geometry has seen considerable success in representation learning tasks such as Gu et al. [2018] or Zhang et al. [2021].

One might question the interpretation of the curvature parameter κ when the set of points does not reside within a canonical manifold. Our approach specifically fits a constant curvature manifold for each quartet (x, y, z, m), embedding these four distances. By taking the

median, we estimate the median curvature of these interpolating spaces. This concept echoes the manifold learning methods proposed by Li et al. [2017], Li and Dunson [2019], which use spherelets to approximate manifolds rather than using locally linear approximations of tangent spaces.

Future extensions may include methods for fitting models in the product space geometry (as is done in Gu et al. [2018]) or other non-constant curvature spaces. Other methodology may include statistically consistent node-level definitions of network curvature as well as more application-focused development of anomaly detection incorporating curvature into the models. Additional work may include using our local definition of curvature to understand its role in relation to notions of brokerage in the sociology literature [Burt, 1992, Buskens and Van de Rijt, 2008], as well as considering how curvature interacts with other quantities of interest across other sciences. Furthermore, additional applications may use non-parametric distance estimators to identify behaviors that form good midpoints in networks, such as overlapping sub-fields within physics in our applications.

Chapter 5

CONCLUDING REMARKS

In this dissertation, we explored several challenges associated with inference problems involving missing data. In the first chapter, we addressed the issue of imputing test scores to accommodate common version changes in Alzheimer's research. While our approach focused on the univariate problem, a logical progression would be to extend this method to jointly harmonize multiple scores and include time-varying effects.

In the second chapter, we investigated network interference under conditions of partial network data. Future research could explore the semiparametric theory applicable to such methods. Often, the partial network data collected can benefit from semiparametric estimation approaches like those discussed in Auerbach [2022]. Additionally, structured assumptions on potential outcomes, as suggested in Belloni et al. [2022] for estimating average direct effects, could be further examined.

Lastly, we addressed the estimation of latent properties by providing a smooth estimator for the curvature of distance matrices, particularly within latent space models. Future work may involve advancing embedding methods for distance matrices exhibiting non-constant curvature.

BIBLIOGRAPHY

Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Systemic risk and stability in financial networks. *The American Economic Review*, 105(2):564–608, 2015.

Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Advances in Neural Information Processing Systems*, 26, 2013.

Aurélien Alfonsi and Benjamin Jourdain. A remark on the optimal transport between two probability measures sharing the same copula. *Statistics & Probability Letters*, 84:131–134, 2014.

Hossein Alidaee, Eric Auerbach, and Michael P Leung. Recovering network structure from aggregated relational data using penalized regression. *arXiv preprint arXiv:2001.06052*, 2020.

Emanuele Aliverti and Daniele Durante. Spatial modeling of brain connectivity data via latent distance models with nodes clustering. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):185–196, 6 2019. ISSN 1932-1872. doi: 10.1002/SAM.11412. URL https://onlinelibrary.wiley.com/doi/full/10.1002/sam.11412.

Zack W Almquist. Random errors in egocentric networks. *Social networks*, 34(4):493–505, 2012.

Attila Ambrus, Markus Mobius, and Adam Szeidl. Consumption risk-sharing in social networks. *American Economic Review*, 104(1):149–182, 2014.

Erling Bernhard Andersen. Asymptotic properties of conditional maximum-likelihood esti-

mators. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(2):283–301, 1970.

Erling D. Andersen and Knud D. Andersen. The Mosek Interior Point Optimizer for Linear Programming: An Implementation of the Homogeneous Algorithm. pages 197–232. Springer, Boston, MA, 2000. doi: 10.1007/978-1-4757-3216-0{_}8. URL https://link.springer.com/chapter/10.1007/978-1-4757-3216-0_8.

Jock R Anderson and Gershon Feder. Agricultural extension. *Handbook of agricultural economics*, 3:2343–2378, 2007.

Donald WK Andrews. Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica: Journal of the Econometric Society*, pages 1465–1471, 1987.

Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, 11(4):1912–1947, 2017. ISSN 1932-6157. doi: 10.1214/16-AOAS1005. URL https://projecteuclid.org/euclid.aoas/1514430272.

Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

Eric Auerbach. Identification and estimation of a partially linear regression model using network data. *Econometrica*, 90(1):347–365, 2022.

Eric Auerbach and Max Tabord-Meehan. The local approach to causal inference under network interference. *arXiv preprint arXiv:2105.03810*, 2021.

Abhijit Banerjee, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. The Diffusion of Microfinance. *Science*, 341(6144), 2013a. ISSN 10959203. doi: 10.1126/SCIENCE.1236498.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013b.

Abhijit Banerjee, Emily Breza, Arun G Chandrasekhar, and Benjamin Golub. When less is more: Experimental evidence on information delivery during india's demonetization. Technical report, National Bureau of Economic Research, 2018.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, 2019.

Aaron J Baraff, Tyler H McCormick, and Adrian E Raftery. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proceedings of the National Academy of Sciences*, 113(51):14668–14673, 2016.

Vladislav Barkanass, Urgen Jost, and Edwin Hancock. Geometric sampling of networks. *Journal of Complex Networks*, 10(4), 6 2022. ISSN 2051-1310. doi: 10.1093/COMNET/CNAC014. URL https://academic.oup.com/comnet/article/10/4/cnac014/6644814.

Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

Danielle S. Bassett, Perry Zurn, and Joshua I. Gold. On the nature and use of models in network neuroscience. *Nature Reviews Neuroscience 2018 19:9*, 19(9):566–578, 7 2018. ISSN 1471-0048. doi: 10.1038/s41583-018-0038-8. URL https://www.nature.com/articles/s41583-018-0038-8.

Guillermo Basulto-Elias, Alicia L. Carriquiry, Kris De Brabanter, and Daniel J. Nordman. Bivariate Kernel Deconvolution with Panel Data. *Sankhya B*, 83(1):122–151, 5 2021. ISSN 09768394. doi: 10.1007/s13571-020-00226-x. URL https://link.springer.com/article/10.1007/s13571-020-00226-x.

Lori Beaman, Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak. Can network theory-based targeting increase technology adoption? *American Economic Review*, 111(6):1918–43, 2021.

Evgeni Begelfor and Michael Werman. The World is not always Flat or Learning Curved Manifolds. *School of Engineering and Computer Science, Hebrew University of Jerusalem., Tech. Rep*, 3(8), 2005.

Graham Bell, Martin J. Lechowicz, and Marcia J. Waterway. Environmental heterogeneity and species diversity of forest sedges. *Journal of Ecology*, 88(1):67–87, 2 2000. ISSN 0022-0477. doi: 10.1046/j.1365-2745.2000.00427.x. URL http://doi.wiley.com/10.1046/j.1365-2745.2000.00427.x.

Alexandre Belloni, Fei Fang, and Alexander Volfovsky. Neighborhood adaptive estimators for causal inference under network interference. *arXiv preprint arXiv:2212.03683*, 2022.

M Berger. An extension of rauch's metric comparison theorem and some applications. *Illinois Journal of Mathematics*, 6(4):700–712, 1962.

Lilah Besser, Walter Kukull, David S. Knopman, Helena Chui, Douglas Galasko, Sandra Weintraub, Gregory Jicha, Cynthia Carlsson, Jeffrey Burns, Joseph Quinn, Robert A. Sweet, Katya Rascovsky, Merilee Teylan, Duane Beekly, George Thomas, Mark Bollenbeck, Sarah Monsell, Charles Mock, Xiao Hua Zhou, Nicole Thomas, Elizabeth Robichaud, Margaret Dean, Janene Hubbard, Mary Jacka, Kristen Schwabe-Fry, Joylee Wu, Creighton Phelps, and John C. Morris. Version 3 of the national Alzheimer's coordinating center's uniform data set, 2018. ISSN

08930341. URL /pmc/articles/PMC6249084//pmc/articles/PMC6249084/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6249084/.

Patrick Billingsley. *Convergence of probability measures* . John Wiley \& Sons, 2013. URL https://www.bibsonomy.org/bibtex/2657f92e619abe605188197d74b27f572/peter.ralph.

L. M. Blumenthal and B. E. Gillam. Distribution of Points in n-Space. *The American Mathematical Monthly*, 50(3):181, 3 1943. ISSN 00029890. doi: 10.2307/2302400.

Robert M Bond and Brad J Bushman. The contagious spread of violence among us adolescents through social networks. *American journal of public health*, 107(2):288–294, 2017.

Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2 2009. ISSN 00368075. doi: 10.1126/SCIENCE.1165821/ASSET/2421E2F8-2DC7-4CAD-8A84-471B43A3C443/ASSETS/GRAPHIC/323_892_F5.JPEG. URL https://www.science.org/doi/10.1126/science.1165821.

Vincent Boucher and Aristide Houndetoungan. *Estimating peer effects using partial network data*. Centre de recherche sur les risques les enjeux économiques et les politiques . . . , 2020.

Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. 2005.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009.

J. S. Brauchart, A. B. Reznikov, E. B. Saff, I. H. Sloan, Y. G. Wang, and R. S. Womersley. Random Point Sets on the Sphere — Hole Radii, Covering, and Separation. *Experimental Mathematics*, 27(1):62–81, 12 2015. ISSN 1944950X. doi: 10.1080/10586458.2016.1226209. URL https://arxiv.org/abs/1512.07470v2.

Jennifer Brennan, Vahab Mirrokni, and Jean Pouget-Abadie. Cluster randomized designs for one-sided bipartite experiments. *arXiv preprint arXiv:2210.16415*, 2022.

Emily Breza, Arun G Chandrasekhar, Tyler H McCormick, and Mengjie Pan. Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 2020.

Emily Breza, Arun G Chandrasekhar, Shane Lubold, Tyler H McCormick, and Mengjie Pan. Consistently estimating network statistics using aggregated relational data. *Proceedings of the National Academy of Sciences*, 120(21):e2207185120, 2023.

Ronald S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, Cambridge, MA, 1992.

Vincent Buskens and Arnout Van de Rijt. Dynamics of networks if everyone strives for structural holes. *American Journal of Sociology*, 114(2):371–407, 2008.

Tony Cai, Jianqing Fan, and Tiefeng Jiang. Distributions of Angles in Random Packing on Spheres. *Journal of Machine Learning Research*, 14:1837–1864, 2013.

Giulia Cencetti, Diego Andrés Contreras, Marco Mancastroppa, and Alain Barrat. Distinguishing simple and complex contagion processes on networks. *Physical Review Letters*, 130(24):247401, 2023.

Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

Benjamin Paul Chamberlain, James R Clough, and Marc Peter Deisenroth. Neural Embeddings of Graphs in Hyperbolic Space. In *13th international workshop on mining and learning from graphs held in conjunction with KDD*, 2017.

Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic Graph Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 32, 10 2019.

ISSN 10495258. doi: 10.48550/arxiv.1910.12933. URL https://arxiv.org/abs/1910.12933v1.

Arun Chandrasekhar and Randall Lewis. Econometrics of sampled networks. *Unpublished manuscript, MIT.[422]*, 2011.

Arun G Chandrasekhar, Matthew O Jackson, Tyler H McCormick, and Vydhourie Thiyageswaran. General covariance-based conditions for central limit theorems with dependent triangular arrays. *arXiv preprint arXiv:2308.12506*, 2023.

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

Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. *The Annals of Applied Probability*, pages 1400–1435, 2011.

Mingli Chen, Kengo Kato, and Chenlei Leng. Analysis of networks via the sparse β-model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):887–910, 2021.

Sheung-Tak Cheng. Cognitive reserve and the prevention of dementia: the role of physical and cognitive activities. *Current psychiatry reports*, 18:1–12, 2016.

Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 6 2006. ISSN 03044076. doi: 10.1016/j.jeconom.2005.02.009.

Flavio Chierichetti, David Liben-Nowell, and Jon Kleinberg. Reconstructing patterns of information diffusion from incomplete observations. *Advances in neural information processing systems*, 24, 2011.

Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2), 2019.

Yeojin Chung and Bruce G. Lindsay. Convergence of the EM algorithm for continuous mixing distributions. *Statistics and Probability Letters*, 96:190–195, 1 2015. ISSN 01677152. doi: 10.1016/j.spl.2014.09.021.

Mayleen Cortez, Matthew Eichhorn, and Christina Yu. Staggered rollout designs enable causal inference under interference without network knowledge. In *Advances in Neural Information Processing Systems*, 2022.

Thomas M. Cover. An Algorithm for Maximizing Expected Log Investment Return. *IEEE Transactions on Information Theory*, 30(2):369–373, 1984. ISSN 15579654. doi: 10.1109/TIT.1984.1056869.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

Giacomo De Giorgi, Michele Pellizzari, and Silvia Redaelli. Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics*, 2(2):241–275, 2010.

Aureo De Paula. Econometrics of network models. In *Advances in economics and econometrics: Theory and applications, eleventh world congress*, pages 268–323. Cambridge University Press Cambridge, 2017.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Fabrizio Durante, Juan Fernández-Sánchez, and Carlo Sempi. A topological proof of Sklar's theorem. *Applied Mathematics Letters*, 26(9):945–948, 9 2013. ISSN 0893-9659. doi: 10.1016/J.AML.2013.04.005.

Bradley Efron. Bayes, Oracle Bayes and Empirical Bayes. *Statistical Science*, 34(2):177–201, 5 2019. ISSN 0883-4237. doi: 10.1214/18-STS674. URL

https://projecteuclid.org/journals/statistical-science/volume-34/issue-2/Bayes-Oracle-Bayes-and-Empirical-Bayes/10.1214/18-STS674.fullhttps://projecteuclid.org/journals/statistical-science/volume-34/issue-2/Bayes-Oracle-Bayes-and-Empirical-Bayes/10.1214/18-STS674.short.

Dennis Epple and Richard E Romano. Peer effects in education: A survey of the theory and evidence. In *Handbook of social economics*, volume 1, pages 1053–1163. Elsevier, 2011.

Hamza Farooq, Yongxin Chen, Tryphon T. Georgiou, Allen Tannenbaum, and Christophe Lenglet. Network curvature as a hallmark of brain structural connectivity. *Nature Communications 2019 10:1*, 10(1):1–11, 10 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12915-x. URL https://www.nature.com/articles/s41467-019-12915-x.

Paul Fearnhead and Guillem Rigaill. Changepoint Detection in the Presence of Outliers. *Journal of the American Statistical Association*, 114(525):169–183, 9 2016. ISSN 1537274X. doi: 10.48550/arxiv.1609.07363. URL https://arxiv.org/abs/1609.07363v2.

Dennis M Feehan and Matthew J Salganik. Generalizing the network scale-up method: a new estimator for the size of hidden populations. *Sociological methodology*, 46(1):153–186, 2016.

Dennis M Feehan, Aline Umubyeyi, Mary Mahy, Wolfgang Hladik, and Matthew J Salganik. Quantity versus quality: A survey experiment to improve the network scale-up method. *American journal of epidemiology*, 183(8):747–757, 2016.

Bailey K. Fosdick, Tyler H. McCormick, Thomas Brendan Murphy, Tin Lok James Ng, and Ted Westling. Multiresolution network models. *Journal of Computational and Graphical Statistics*, 28(1):185–196, 8 2016. ISSN 15372715. doi: 10.48550/arxiv.1608.07618. URL https://arxiv.org/abs/1608.07618v5.

Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Maurice Frechet. Sur la distance de deux lois de probabilite. *CR. Acad Sci. Paris*, 244, 1957.

Linton C Freeman. Centered graphs and the structure of ego networks. *Mathematical Social Sciences*, 3(3):291–304, 1982.

Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 11 2020. ISSN 15487660. doi: 10.18637/jss.v094.i14. URL https://CRAN.R-project.

Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian Geometry*. Universitext. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-20493-0. doi: 10.1007/978-3-642-18855-8. URL http://link.springer.com/10.1007/978-3-642-18855-8.

Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, pages 2624–2652, 2015.

Anna C Gilbert and Lalit Jain. If it ain't broke, don't fix it: Sparse metric repair. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 612–619. IEEE, 2017.

P Giles. The mathematical theory of infectious diseases and its applications. *Journal of the Operational Research Society*, 28(2):479–480, 1977.

Sharad Goel and Matthew J Salganik. Respondent-driven sampling as markov chain monte carlo. *Statistics in medicine*, 28(17):2202–2229, 2009.

Sharad Goel and Matthew J Salganik. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15):6743–6747, 2010.

P. Goldsmith-Pinkham and G. Imbens. Social networks and the identification of peer effects. *Journal of Business and Economic Statistics*, 31:3:253–264, 2013.

Bryan S Graham. Network data. In *Handbook of econometrics*, volume 7, pages 111–218. Elsevier, 2020.

AKB Green, TH McCormick, and AE Raftery. Consistency for the tree bootstrap in respondent-driven sampling. *Biometrika*, 107(2):497–504, 2020.

L Griffith, E van den Heuvel, I Fortier, S Hofer, P Raina, N Sohel, H Payette, C Wolfson, and S Belleville. Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis. In *Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis*. Agency for Healthcare Research and Quality (US), Rockville (MD), 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23617017.

Geoffrey R Grimmett and Colin JH McDiarmid. On colouring random graphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 77, pages 313–324. Cambridge University Press, 1975.

Mikhail Gromov, editor. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhäuser Boston, 2007. doi: 10.1007/978-0-8176-4583-0.

Jonathan Gruhl, Elena A. Erosheva, and Paul K. Crane. A semiparametric approach to mixed outcome latent variable models: Estimating the association between cognition and regional brain volumes. *Annals of Applied Statistics*, 7(4): 2361–2383, 12 2013. ISSN 19326157. doi: 10.1214/13-AOAS675. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-7/issue-4/A-semiparametric-approach-to-mixed-outcome-latent-variable-models/10.1214/13-AOAS675.fullhttps://projecteuclid.org/journals/annals-of-applied-statistics/volume-7/issue-4/A-semiparametric-approach-to-mixed-outc 10.1214/13-AOAS675.short.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature rep-

resentations in product spaces. In *International Conference on Learning Representations*, 2018.

F. Richard Guo and Thomas S. Richardson. Chernoff-Type Concentration of Empirical Probabilities in Relative Entropy. *IEEE Transactions on Information Theory*, 67(1):549–558, 1 2021. ISSN 15579654. doi: 10.1109/TIT.2020.3034539.

Viet Ha-Thuc, Avishek Dutta, Ren Mao, Matthew Wood, and Yunli Liu. A counterfactual framework for seller-side a/b testing on marketplaces. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2296, 2020.

Paul R. Halmos. *Measure Theory*, volume 18. Springer, 2013. doi: 10.1007/978-1-4684-9440-2{_}1.

Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 3 2007. ISSN 1467-985X. doi: 10.1111/J.1467-985X.2007.00471.X. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-985X.2007.00471.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2007.00471.xhttps://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2007.00471.x.

Zaıd Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

Morgan Hardy, Rachel M. Heath, Wesley Lee, and Tyler H. McCormick. Estimating spillovers using imprecisely measured networks. *arXiv*, 3 2019. URL http://arxiv.org/abs/1904.00136.

Timothy F Havel and Kurt Wüthrich. An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *Journal of molecular biology*, 182(2):281–294, 1985.

Xiaoqi He and Kyungchul Song. Measuring Diffusion Over a Large Network. *The Review of Economic Studies*, page rdad115, 12 2023. ISSN 0034-6527. doi: 10.1093/restud/rdad115. URL https://doi.org/10.1093/restud/rdad115.

Douglas D Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, 44(2):174–199, 1997.

Wassily Hoeffding. Masstabvariarte Korrelationstheorie. *Schrijl Math. Inst. Univ. Berlin*, 5 (6), 1940.

Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in neural information processing systems*, 20, 2007.

Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent Space Approaches to Social Network Analysis. *https://doi.org/10.1198/016214502388618906*, 97(460):1090–1098, 12 2002a. ISSN 01621459. doi: 10.1198/016214502388618906. URL https://www.tandfonline.com/doi/abs/10.1198/016214502388618906.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002b.

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

H Hopf. Zum clifford-kieinschen raumproblem. *Mathematische Annalen*, 95:313–339, 1926.

Grace C Huang, Jennifer B Unger, Daniel Soto, Kayo Fujimoto, Mary Ann Pentz, Maryalice Jordan-Marsh, and Thomas W Valente. Peer influences: the impact of online and offline friendship networks on adolescent smoking and alcohol use. *Journal of Adolescent Health*, 54(5):508–514, 2014.

Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Nikolaos Ignatiadis and Stefan Wager. Confidence Intervals for Nonparametric Empirical Bayes Analysis. *Journal of the American Statistical Association*, 2 2019a. ISSN 1537274X. doi: 10.48550/arxiv.1902.02774. URL https://arxiv.org/abs/1902.02774v4.

Nikolaos Ignatiadis and Stefan Wager. Covariate-Powered Empirical Bayes Estimation. *Advances in Neural Information Processing Systems*, 32, 6 2019b. ISSN 10495258. doi: 10.48550/arxiv.1906.01611. URL https://arxiv.org/abs/1906.01611v2.

Kosuke Imai, Zhichao Jiang, and Anup Malani. Causal inference with interference and noncompliance in two-stage randomized experiments. *Journal of the American Statistical Association*, 116(534):632–644, 2021.

Matthew O Jackson and Leeat Yariv. Diffusion on social networks. *Economie publique/Public economics*, (16), 2006.

Matthew O Jackson et al. *Social and economic networks*, volume 3. Princeton university press Princeton, 2008.

Liwei Jing, Chengyi Qu, Hongmei Yu, Tong Wang, and Yuehua Cui. Estimating the sizes of populations at high risk for HIV: a comparison study. *PloS ONE*, 9(4):e95601, 2014.

Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Y Weintraub. Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 68(10):7069–7089, 2022.

Matthew S. Johnson. Modeling dichotomous item responses with free-knot splines. *Computational Statistics and Data Analysis*, 51(9):4178–4192, 5 2007. ISSN 01679473. doi: 10.1016/j.csda.2006.04.021.

Karl G. Jöreskog and Irini Moustaki. Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 36(3):347–387, 2001.

ISSN 00273171. doi: 10.1207/S15327906347-387. URL https://www.tandfonline.com/action/journalInformation?journalCode=hmbr20.

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.

Hyunseung Kang, Benno Kreuels, Ohene Adjei, Ralf Krumkamp, Jürgen May, and Dylan S. Small. The causal effect of malaria on stunting: A Mendelian randomization and matching approach. *International Journal of Epidemiology*, 42(5):1390–1398, 10 2013. ISSN 03005771. doi: 10.1093/ije/dyt116. URL https://pubmed.ncbi.nlm.nih.gov/23925429/.

Matthew James Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

J. Kiefer and J. Wolfowitz. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 12 1956. ISSN 0003-4851. doi: 10.1214/aoms/1177728066. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-27/issue-4/Consistency-of-the-Maximum-Likelihood-Estimator-in-the-Presence-of/10.1214/aoms/1177728066.fullhttps://projecteuclid.org/

journals/annals-of-mathematical-statistics/volume-27/issue-4/Consistency-of-the-Maximum-Likelihood-Estimator-in-the-Presence-of/10.1214/aoms/1177728066.short.

Wilhelm Killing. Ueber die Clifford-Klein'schen Raumformen. *Mathematische Annalen 1891 39:2*, 39(2):257–278, 6 1891. ISSN 1432-1807. doi: 10.1007/BF01206655. URL https://link.springer.com/article/10.1007/BF01206655.

Peter D Killworth, Eugene C Johnsen, Christopher McCarty, Gene Ann Shelley, and H Russell Bernard. A social network approach to estimating seroprevalence in the United States. *Social Networks*, 20(1):23–50, 1998a.

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, 1998b.

Bomin Kim, Kevin H Lee, Lingzhou Xue, and Xiaoyue Niu. A review of dynamic network models with latent variables. *Statistics surveys*, 12:105, 2018.

Wilhelm Klingenberg. Riemannian Geometry. *Riemannian Geometry*, 12 1995. doi: 10.1515/9783110905120.

Peter Kosmol and Dieter Müller-Wichards. *Optimization in Function Spaces : With Stability Considerations in Orlicz Spaces*, volume 13. De Gruyter, Würzburg, Germany, 2 2011.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Nan Laird. Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association*, 73(364):805, 12 1978. ISSN 01621459. doi: 10.2307/2286284.

Lucien Le Cam. On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 129–157. University of California Press, 1956.

Wilmer Leal, Guillermo Restrepo, Peter F. Stadler, and Jürgen Jost. Forman-Ricci Curvature for Hypergraphs. *Advances in Complex Systems*, 24(1), 11 2018. doi: 10.13140/RG.2.2.27347.84001. URL http://arxiv.org/abs/1811.07825http://dx.doi.org/10.13140/RG.2.2.27347.84001.

Wesley Lee, Tyler H Mccormick, Joshua Neil, Microsoft Cole, Sodja Microsoft, and Yanran Cui. Anomaly Detection in Large Scale Networks with Latent Space Models. *Technometrics*, pages 1–23, 11 2019. ISSN 0040-1706. doi: 10.1080/00401706.2021.1952900. URL https://arxiv.org/abs/1911.05522v2.

Stefanski Leonard and Raymond J. Carroll. Deconvoluting Kernel Density Estimators. *Statistics*, 21(2):169–184, 1 1990. ISSN 10294910. doi: 10.1080/02331889008802238. URL https://www.tandfonline.com/doi/abs/10.1080/02331889008802238.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 3 2007. ISSN 15564681. doi: 10.1145/1217299.1217301. URL https://dl.acm.org/doi/abs/10.1145/1217299.1217301.

Didong Li and David B Dunson. Geodesic distance estimation with spherelets. *arXiv preprint arXiv:1907.00296*, 2019.

Didong Li, Minerva Mukhopadhyay, and David B Dunson. Efficient manifold and subspace approximations with spherelets. *arXiv preprint arXiv:1706.08263*, 2017.

Bruce Lindsay, Clifford C. Clogg, and John Grego. Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class Model

for Item Analysis. *Journal of the American Statistical Association*, 86(413):96, 3 1991. ISSN 01621459. doi: 10.2307/2289719.

Bruce G. Lindsay. The Geometry of Mixture Likelihoods, Part II: The Exponential Family. *The Annals of Statistics*, 11(3):783–792, 9 1983a. ISSN 0090-5364. doi: 10.1214/aos/1176346245. URL https://projecteuclid.org/journals/annals-of-statistics/volume-11/issue-3/The-Geometry-of-Mixture-Likelihoods-Part-II--The-Exponential/10.1214/aos/1176346245.full.

Bruce G. Lindsay. The Geometry of Mixture Likelihoods: A General Theory. *The Annals of Statistics*, 11(1):86–94, 3 1983b. ISSN 0090-5364. doi: 10.1214/aos/1176346059. URL https://projecteuclid.org/euclid.aos/1176346059.

Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

Tin Lok, James Ng, Thomas Brendan Murphy, Ted Westling, Tyler H Mccormick, and Bailey Fosdick. Modeling the social media relationships of Irish politicians using a generalized latent space stochastic blockmodel. *Annals of Applied Statistics*, 15(4):1923–1944, 12 2021. ISSN 1932-6157. doi: 10.1214/21-AOAS1483.

Frederic M. Lord. A strong true-score theory, with applications. *Psychometrika*, 30(3):239–270, 9 1965. ISSN 00333123. doi: 10.1007/BF02289490. URL https://link.springer.com/article/10.1007/BF02289490.

Frederic M. Lord. Estimating true-score distributions in psychological testing (an empirical bayes estimation problem). *Psychometrika*, 34(3):259–299, 9 1969. ISSN 00333123. doi: 10.1007/BF02289358. URL https://link.springer.com/article/10.1007/BF02289358.

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

Shane Lubold, Arun G Chandrasekhar, and Tyler H McCormick. Identifying the latent space geometry of network models through analysis of curvature. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):240–292, 2023.

Blake MacDonald, Pritam Ranjan, and Hugh Chipman. Gpfit: An r package for fitting a gaussian process model to deterministic simulator outputs. *Journal of Statistical Software*, 64:1–23, 2015.

Peter W MacDonald, Elizaveta Levina, and Ji Zhu. Latent space models for multiplex networks with shared structure. *Biometrika*, 109(3):683–706, 2022.

Anup Malani, Phoebe Holtzman, Kosuke Imai, Cynthia Kinnan, Morgen Miller, Shailender Swaminathan, Alessandra Voena, Bartosz Woda, and Gabriella Conti. Effect of health insurance in india: a randomized controlled trial. Technical report, National Bureau of Economic Research, 2021.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

Charles F Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

Jay Mardia, Jiantao Jiao, Ervin Tánczos, Robert D. Nowak, and Tsachy Weissman. Concentration Inequalities for the Empirical Distribution. 9 2018. URL https://arxiv.org/abs/1809.06522v3.

Tyler H McCormick. The network scale-up method. *The Oxford Handbook of Social Networks*, page 153, 2020.

Sarah McGrory, Jason M. Doherty, Elizabeth J. Austin, John M. Starr, and Susan D. Shenkin. Item response theory analysis of cognitive tests in people with dementia: A systematic review. *BMC Psychiatry*, 14(1):47, 2 2014. ISSN 1471244X. doi: 10.1186/1471-244X-14-47. URL /pmc/articles/PMC3931670/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3931670/.

Xiangfei Meng and Carl D'arcy. Education and dementia in the context of the cognitive reserve hypothesis: a systematic review with meta-analyses and qualitative analyses. *PloS one*, 7(6):e38268, 2012.

William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 12 1993. ISSN 00333123. doi: 10.1007/BF02294825. URL /record/1994-11989-001.

Sarah E. Monsell, Hiroko H. Dodge, Xiao Hua Zhou, Yunqi Bu, Lilah M. Besser, Charles Mock, Stephen E. Hawes, Walter A. Kukull, Sandra Weintraub, Steven Ferris, Joel Kramer, David Loewenstein, Po Lu, Bruno Giordani, Felicia Goldstein, Dan Marson, John Morris, Dan Mungas, David Salmon, and Kathleen Welsh-Bohmer. Results from the NACC Uniform Data Set Neuropsychological Battery Crosswalk Study. *Alzheimer Disease and Associated Disorders*, 30(2):134–139, 2016. ISSN 08930341. doi: 10.1097/WAD.0000000000000111. URL https://pubmed.ncbi.nlm.nih.gov/26485498/.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A Wrapped Normal Distribution on Hyperbolic Space for Gradient-Based Learning. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:8242–8251, 2 2019. doi: 10.48550/arxiv.1902.02992. URL https://arxiv.org/abs/1902.02992v2.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Chien Chun Ni, Yu Yao Lin, Feng Luo, and Jie Gao. Community Detection on Networks with Ricci Flow. *Scientific Reports 2019 9:1*, 9(1):1–12, 7 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-46380-9. URL https://www.nature.com/articles/s41598-019-46380-9.

Maximillian Nickel and Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. *Advances in Neural Information Processing Systems*, 30, 2017.

Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016. URL http://stanford.edu/~boyd/papers/scs.html.

Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. Causal inference for social network data. *Journal of the American Statistical Association*, pages 1–15, 2022.

Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 1 2007. ISSN 00221236. doi: 10.48550/arxiv.math/0701886. URL https://arxiv.org/abs/math/0701886v4.

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.

Sally Paganin, Christopher J. Paciorek, Claudia Wehrhahn, Abel Rodriguez, Sophia Rabe-Hesketh, and Perry de Valpine. Computational methods for Bayesian semiparametric Item Response Theory models. 1 2021. URL http://arxiv.org/abs/2101.11583.

Lia Papadopoulos, Mason A. Porter, Karen E. Daniels, and Danielle S. Bassett. Network analysis of particles and grains. *Journal of Complex Networks*, 6(4):485–565, 8 2018. ISSN 20511329. doi: 10.1093/COMNET/CNY005. URL https://academic.oup.com/comnet/article/6/4/485/4959635.

Panagiotis Papastamoulis. label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271*, 2015.

Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Xavier Pennec. Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements. In *International Workshop on Nonlinear Signal and Image Processing*, pages 194–198, Antalya, Turkey, 6 1999. URL http://www-sop.inria.fr/epidaure/personnel/pennec/pennec.html.

Dimitris N. Politis and Joseph P. Romano. Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions. *https://doi.org/10.1214/aos/1176325770*, 22 (4):2031–2050, 12 1994. ISSN 0090-5364. doi: 10.1214/AOS/1176325770.

Gail E Potter, Mark S Handcock, Ira M Longini Jr, and M Elizabeth Halloran. Estimating within-household contact networks from egocentric data. *The annals of applied statistics*, 5(3):1816, 2011.

Jean Pouget-Abadie, Vahab Mirrokni, David C Parkes, and Edoardo M Airoldi. Optimizing cluster-based randomized experiments under monotonicity. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2090–2099, 2018.

Jean Pouget-Abadie, Guillaume Saint-Jacques, Martin Saveski, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. Testing for arbitrary interference on experimentation platforms. *Biometrika*, 106(4):929–940, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Georg Rasch. An Individualistic Approach to Item Analysis. Technical report, Readings in Mathematical Social Science, 1966.

James M. Robins and Anastasios A. Tsiatis. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications*

in Statistics - Theory and Methods, 20(8):2609–2631, 1 1991. ISSN 1532415X. doi: 10.1080/03610929108830654. URL https://www.tandfonline.com/doi/abs/10.1080/03610929108830654.

Sebastien Roch and Karl Rohe. Generalized least squares can overcome the critical threshold in respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 115(41): 10299–10304, 2018.

Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704, 2011.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B. Rubin. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489, 6 1996. ISSN 1537274X. doi: 10.1080/01621459.1996.10476908.

Michael Salter-Townshend and Tyler H. McCormick. Latent space models for multiview network data. *https://doi.org/10.1214/16-AOAS955*, 11(3):1217–1244, 9 2017. ISSN 1932-6157. doi: 10.1214/16-AOAS955. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-11/issue-3/Latent-space-models-for-multiview-network-data/10.1214/16-AOAS955.fullhttps://projecteuclid.org/journals/annals-of-applied-statistics/volume-11/issue-3/Latent-space-models-for-multiview-network-data/10.1214/16-AOAS955.short.

Areejit Samal, Hirdesh K. Pharasi, Sarath Jyotsna Ramaia, Harish Kannan, Emil Saucan, Jürgen Jost, and Anirban Chakraborti. Network geometry and market instability. *Royal*

Society Open Science, 8(2), 2 2021. ISSN 20545703. doi: 10.1098/RSOS.201734. URL https://royalsocietypublishing.org/doi/10.1098/rsos.201734.

Romeil Sandhu, Tryphon Georgiou, Ed Reznik, Liangjia Zhu, Ivan Kolesov, Yasin Senbabaoglu, and Allen Tannenbaum. Graph Curvature for Differentiating Cancer Networks. *Nature Scientific Reports*, 5(1):1–13, 7 2015. ISSN 2045-2322. doi: 10.1038/srep12323. URL https://www.nature.com/articles/srep12323.

Romeil S. Sandhu, Tryphon T. Georgiou, and Allen R. Tannenbaum. Ricci curvature: An economic indicator for market fragility and systemic risk. *Science Advances*, 2(5), 5 2016. ISSN 23752548. doi: 10.1126/SCIADV.1501495/SUPPL{_}FILE/1501495{_}SM.PDF. URL https://www.science.org/doi/10.1126/sciadv.1501495.

Emil Saucan, Areejit Samal, and Jürgen Jost. A Simple Differential Geometry for Complex Networks. *Network Science*, 9(S1):S106–S133, 4 2020. ISSN 20501250. doi: 10.48550/arxiv.2004.11112. URL https://arxiv.org/abs/2004.11112v2.

Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1027–1035, 2017.

I. J. Schoenberg. Remarks to Maurice Frechet's Article "Sur La Definition Axiomatique D'Une Classe D'Espace Distances Vectoriellement Applicable Sur L'Espace De Hilbert. *The Annals of Mathematics*, 36(3):724, 7 1935. ISSN 0003486X. doi: 10.2307/1968654.

Michael Schomaker and Christian Heumann. Bootstrap inference when using multiple imputation. *Statistics in medicine*, 37(14):2252–2266, 2018.

O Scutelniciuc. Network scale-up method experiences: Republic of kazakhstan. *Consultation on estimating population sizes through household surveys: Successes and challenges (New York, NY)*, 2012.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Jayson Sia, Edmond Jonckheere, and Paul Bogdan. Ollivier-Ricci Curvature-Based Method to Community Detection in Complex Networks. *Nature, Scientific Reports 2019 9:1*, 9 (1):1–12, 7 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-46079-x. URL https://www.nature.com/articles/s41598-019-46079-x.

Grzegorz Sienski, Priyanka Narayan, Julia Maeve Bonner, Nora Kory, Sebastian Boland, Aleksandra A Arczewska, William T Ralvenius, Leyla Akay, Elana Lockshin, Liang He, et al. Apoe4 disrupts intracellular lipid homeostasis in human ipsc-derived glia. *Science translational medicine*, 13(583):eaaz4564, 2021.

B. W. Silverman, M. C. Jones, J. D. Wilson, and D. W. Nychka. A Smoothed Em Approach to Indirect Estimation Problems, with Particular Reference to Stereology and Emission Tomography. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(2): 271–303, 1 1990. ISSN 00359246. doi: 10.1111/j.2517-6161.1990.tb01788.x. URL http://doi.wiley.com/10.1111/j.2517-6161.1990.tb01788.x.

SKLAR and M. Fonctions de repartition a n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959. URL https://ci.nii.ac.jp/naid/10011938360.

Anna L. Smith, Dena M. Asta, and Catherine A. Calder. The Geometry of Continuous Latent Space Models for Network Data. *Statistical Science*, 34(3):428–453, 12 2017. ISSN 21688745. doi: 10.48550/arxiv.1712.08641. URL https://arxiv.org/abs/1712.08641v2.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Yaakov Stern. Cognitive reserve. *Neuropsychologia*, 47(10):2015–2028, 2009.

Yaakov Stern. Cognitive reserve in ageing and alzheimer's disease. *The Lancet Neurology*, 11(11):1006–1012, 2012.

Tracy Sweet and Samrachana Adhikari. A Latent Space Network Model for Social Influence. *Psychometrika*, 85(2):251–274, 6 2020. ISSN 18600980. doi: 10.1007/S11336-020-09700-X/FIGURES/11. URL https://link.springer.com/article/10.1007/s11336-020-09700-x.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Kevin Tian, Weihao Kong, and Gregory Valiant. Learning Populations of Parameters. Technical report, 2017.

Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1996.tb02080.xhttps://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.xhttps://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x.

Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17 (4):401–419, 1952.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

Viet Chi Tran and Thi Phuong Thuy Vo. Estimation of dense stochastic block models visited by random walks. *Electronic Journal of Statistics*, 15(2):5855–5887, 2021.

Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

Johan Ugander and Hao Yin. Randomized graph cluster randomization. *Journal of Causal Inference*, 11(1):20220014, 2023.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F128815:329–337, 8 2013. doi: 10.1145/2487575.2487695.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 10 1998. doi: 10.1017/cbo9780511802256. URL /core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D.

Edwin R. van den Heuvel, Lauren E. Griffith, Nazmul Sohel, Isabel Fortier, Graciela Muniz-Terrera, and Parminder Raina. Latent variable models for harmonization of test scores: A case study on memory. *Biometrical Journal*, 62(1):34–52, 1 2020. ISSN 15214036. doi: 10.1002/bimj.201800146. URL https://onlinelibrary.wiley.com/doi/full/10.1002/bimj.201800146https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201800146https://onlinelibrary.wiley.com/doi/10.1002/bimj.201800146.

Pim van der Hoorn, William J. Cunningham, Gabor Lippner, Carlo Trugenberger, and Dmitri Krioukov. Ollivier-Ricci curvature convergence in random geometric graphs. *Physical Review Research*, 3(1), 8 2020. doi: 10.1103/PhysRevResearch.3.013211. URL http://arxiv.org/abs/2008.01209http://dx.doi.org/10.1103/PhysRevResearch.3.013211.

Mark J van der Laan. Causal inference for networks. 2012.

Aad van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 10 1998. doi: 10.1017/cbo9780511802256. URL /core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D.

Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985. ISSN 1537274X. doi: 10.1080/01621459.1985.10477119.

Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 3 2003. ISBN 9780821833124. doi: 10.1090/gsm/058. URL http://www.ams.org/gsm/058.

Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham M. Kakade. Maximum Likelihood Estimation for Learning Populations of Parameters. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:11217–11226, 2 2019. URL http://arxiv.org/abs/1902.04553.

Davide Viviano. Experimental design under network interference. *arXiv preprint arXiv:2003.08421*, 2020.

Erik M. Volz, Joel C. Miller, Alison Galvani, and Lauren Meyers. Effects of Heterogeneous and Clustered Contact Patterns on Infectious Disease Dynamics. *PLOS Computational Biology*, 7(6):e1002042, 2011. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1002042. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002042.

Linbo Wang, James M. Robins, and Thomas S. Richardson. On falsification of the binary instrumental variable model, 3 2017. ISSN 14643510. URL https://academic.oup.com/biomet/article/104/1/229/2938060.

Sandra Weintraub, David Salmon, Nathaniel Mercaldo, Steven Ferris, Neill R. Graff-Radford, Helena Chui, Jeffrey Cummings, Charles DeCarli, Norman L. Foster, Douglas Galasko, Elaine Peskind, Woodrow Dietrich, Duane L. Beekly, Walter A. Kukull, and John C. Morris. The Alzheimer's Disease Centers' Uniform Data Set (UDS): The neuropsychologic test battery. *Alzheimer Disease and Associ-*

ated Disorders, 23(2):91–101, 4 2009. ISSN 08930341. doi: 10.1097/WAD.0b013e318191c7dd. URL /pmc/articles/PMC2743984//pmc/articles/PMC2743984/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2743984/.

Ian R. White, Sarah Walker, Abdel G. Babiker, and Janet H. Darbyshire. Impact of treatment changes on the interpretation of the Concorde trial. *AIDS*, 11(8):999–1006, 1997. ISSN 02699370. doi: 10.1097/00002030-199708000-00008. URL https://pubmed.ncbi.nlm.nih.gov/9223734/.

Ian R. White, Abdel G. Babiker, Sarah Walker, and Janet H. Darbyshire. Randomization-based methods for correcting for treatment changes: Examples from the Concorde trial. *Statistics in Medicine*, 18(19):2617–2634, 10 1999. ISSN 02776715. doi: 10.1002/(SICI)1097-0258(19991015)18:19⟨2617::AID-SIM187⟩3.0.CO;2-E. URL https://europepmc.org/article/MED/10495460.

Steven Wilkins-Reeves and Tyler McCormick. Asymptotically normal estimation of local latent network curvature. *arXiv preprint arXiv:2211.11673*, 2022.

Steven Wilkins-Reeves, Xu Chen, Qi Ma, Christine Agarwal, and Aude Hofleitner. Multiply robust estimation for local distribution shifts with multiple domains. *arXiv preprint arXiv:2402.14145*, 2024.

G. R. Wood. Binomial mixtures: geometric estimation of the mixing distribution. *The Annals of Statistics*, 27(5):1706–1721, 10 1999. ISSN 0090-5364. doi: 10.1214/aos/1017939148. URL https://projecteuclid.org/journals/annals-of-statistics/volume-27/issue-5/Binomial-mixtures-geometric-estimation-of-the-mixing-distribution/10.1214/aos/1017939148.full.

Carol M. Woods and David Thissen. Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71(2):281–301, 6 2006.

ISSN 00333123. doi: 10.1007/s11336-004-1175-8. URL https://link.springer.com/article/10.1007/s11336-004-1175-8.

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

Yachen Yan. *rBayesianOptimization*, 2021. URL https://github.com/yanyachen/rBayesianOptimization.

Christina Lee Yu, Edoardo M Airoldi, Christian Borgs, and Jennifer T Chayes. Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119, 2022.

Shuai Zhang, Yi Tay, Wenqi Jiang, Da-cheng Juan, and Ce Zhang. Switch spaces: learning product spaces with sparse gating. *arXiv preprint arXiv:2102.08688*, 2021.

Appendix A

SUPPLEMENTARY MATERIAL FOR PROJECT 1

A.1 Proof of Main Paper Theorem

Theorem A.1.1. *(Listed as Theorem 3.1 in the main paper.) Denote the unique global solution* $P^*_{M,\mu} = \arg\max_{P_M \in \mathcal{P}_{[0,1]}} \ell_{n,\mu}[P_M]$.

Then consider a sequence of latent trait distributions $\{P^{(t)}_{M,\mu}\}_{t=0}^{\infty}$ *generated by the EM algorithm for the regularized likelihood. If* $\mu > 0$ *and* $p_A(y|\gamma)$ *is continuous in* γ *for each* y*, then*

$$\ell_{n,\mu}[P^{(t)}_{M,\mu}] \xrightarrow{t\to\infty} \ell_{n,\mu}[P^*_{M,\mu}].$$

and

$$P^{(t)}_{M,\mu} \xrightarrow[w]{t\to\infty} P^*_{M,\mu}$$

where $\longrightarrow_w$ *denotes weak convergence of measures.*

Before proving the main theorem, we prove a series of lemmas that will be useful in characterizing a maximizer.

Lemma A.1.2. *If* $\mu > 0$ *and* $p_A(y|\gamma)$ *is continuous in* γ *for each* y *then the maximizer of* $\ell_{n,\mu}[\cdot]$*,* P^*_M *is a continuous distribution. Specifically,* P^*_M *and* P_U *are mutually absolutely continuous i.e.* $P^*_M \ll P_U$ *and* $P_U \ll P^*_M$

Proof. Firstly, if $P_U \not\ll P^*_M$ then $\mathcal{D}(P_U||P^*_M) = \infty$ and clearly P^*_M is not a maximizer. The proof now just considers the opposite direction. Next consider the Lebesgue decomposition theorem [Halmos, 2013]. Consider an arbitrary probability measure Q which we compare to

P_U (which on the interval is equivalent to the Lebesgue measure). Unless otherwise stated, singular and continuous will be with respect to this measure P_U. There exist measures Q_c and Q_s such that:

$$Q = Q_c + Q_s$$
$$\text{Where } Q_c \ll P_U$$
$$Q_s \perp P_U$$

I.e. Q_c is absolutely continuous with respect to P_U and Q_s is singular with respect to P_U. Next, we consider the KL divergence between P_U and Q. Since Q is a probability measure, we can rescale the measures and define $Q = (1-\lambda)Q_c + \lambda Q_s$ where Q_c and Q_s are probability measures as well. Then

$$\begin{aligned}\mathcal{D}(P_U||Q) &= \int \log\left(\frac{dP_U}{dQ}\right) dP_U \\ &= \int \log\left(\frac{dP_U}{(1-\lambda)dQ_c + \lambda dQ_s}\right) dP_U\end{aligned}$$

Denote $\mathcal{X}_s \subset [0,1]$ the subset of the interval where the singular part of the measure is defined. Since $P_U(\mathcal{X}_s) = 0$ on this subset, $\frac{dP_U}{(1-\lambda)dQ_c+\lambda dQ_s}(x) = 0$ for all $x \subset \mathcal{X}_s$. Hence the KL divergence can be computed directly using *only* the absolutely continuous part, though the value of λ contributes according to the following form.

$$\begin{aligned}\mathcal{D}(P_U||Q) &= \int \log\left(\frac{dP_U}{(1-\lambda)dQ_c + \lambda dQ_s}\right) dP_U \\ &= \int_{[0,1]} \log\left(\frac{dP_U}{(1-\lambda)dQ_c + \lambda dQ_s}(x)\right) dP_U(x) \\ &= \int_{[0,1]} \log\left(\frac{dP_U}{(1-\lambda)dQ_c}(x)\right) dP_U(x) \\ &= -\log(1-\lambda) + \int_{[0,1]} \log\left(\frac{dP_U}{dQ_c}(x)\right) dP_U(x)\end{aligned}$$

Secondly, we verify that $\ell_{n,\mu}[P_M]$ is a concave function of P_M. We can add a constant to

$\ell_{n,\mu}[P_M]$ so that

$$\ell_{n,\mu}[P_M] - \sum_y \widehat{p}(y)\log(\widehat{p}(y)) = -\mathcal{D}(\widehat{p}||p_{MA}) - \mu\mathcal{D}(p_U||p_M).$$

Since p_{MA} is linear in p_M and the KL divergence is a strictly concave function, then this regularized likelihood is clearly a strictly concave function in p_M.

Next, we consider Theorem 3.4.3 in [Kosmol and Müller-Wichards, 2011] which is a sufficient and necessary condition for the optimality of a measure P_M^* to be the optimizer of a convex objective.

Theorem A.1.3 (Theorem 3.4.3 in [Kosmol and Müller-Wichards, 2011]). *If K is a convex subset of a vector space X and $f : X \to \mathbb{R}$ a convex function, f has a minimal solution $x_0 \in K$ if and only if for all $x \in K$*

$$\dot{f}_+(x_0; x - x_0) \geq 0 \tag{A.1}$$

Where

$$\dot{f}_+(x_0; z) = \lim_{t\to 0^+} \frac{f(x_0 + tz) - f(x_0)}{t}$$

In our case, we define X to be the vector space of signed measures and K probability measures over $[0, 1]$. We can also use this to show if x is a direction that adds a singular measure to a particular $\widetilde{x}_0$ then the Gateaux derivative will be negative, and hence any measure which increases the likelihood cannot have any singular component. Since our log-likelihood functional $\ell_{n,\mu}[\cdot]$ is concave, the theorem above will by defining $f = -\ell_{n,\mu}$.
Denote, for any measure H, $p_A(y; H) = \int p_A(y|\gamma)dH(\gamma)$.

Let Q be a continuous measure, and define a direction $G = Q_s - Q$ where Q_s is a singular

measure. Let q be the corresponding density function to Q.

$$\begin{aligned}
\ell_{n,\mu}[Q+\lambda G] &= \sum_y \widehat{p}(y)\log\left(p_A(y;Q)+\lambda p_A(y;G)\right) - \mu\int_0^1 \log(q(\gamma))d\gamma \\
&\quad + \mu\log(1-\lambda) \\
&= \sum_y \widehat{p}(y)\left(\log\left(p_A(y;Q)\right) + \log\left(1+\lambda\frac{p_A(y;G)}{p_A(y;Q)}\right)\right) - \mu\int_0^1 \log(q(\gamma))d\gamma \\
&\quad + \mu\log(1-\lambda) \\
&= \ell_{n,\mu}[Q] + \sum_y \widehat{p}(y)\log\left(1+\lambda\frac{p_A(y;G)}{p_A(y;Q)}\right) + \mu\log(1-\lambda) \\
&= \ell_{n,\mu}[Q] + \lambda\sum_y \widehat{p}(y)\frac{p_A(y;G)}{p_A(y;Q)} + \mu\log(1-\lambda) + O(\lambda^2) \\
\implies \dot{\ell}_{n,\mu}(Q;G) &= \sum_y \widehat{p}(y)\frac{p_A(y;Q_s)}{p_A(y;Q)} - 1 - \mu
\end{aligned}$$

and hence by Theorem A.1.3, $\dot{\ell}_{n,\mu}(Q;G)$ [1] We consider 2 cases, whether or not if there exists a continuous measure Q^* which maximizes the unregularized problem, i.e. when we set $\mu = 0$.

Case 1: *There exists Q^* which is continuous which maximizes the unregularized likelihood ($\ell_{n,\mu=0}$).*

If this holds then

$$\sum_y \frac{\widehat{p}(y)p_A(y|\gamma)}{p_A(y;Q^*)} - 1 < \sum_y \frac{\widehat{p}(y)p_A(y|\gamma)}{p_A(y;Q^*)} - 1 - \mu \leq -\mu < 0.$$

And hence any direction which adds a singular measure will decrease the likelihood.

---

[1] A similar condition holds in Lindsay [1995] for the unregularized case, Q^* is a maximizer of $\ell_{n,\mu=0}[\cdot]$ if and only if

$$\sum_y \frac{\widehat{p}(y)p_A(y|\gamma)}{p_A(y;Q^*)} - 1 \leq 0$$

for all γ. When integrating this over any measure Q' it will also hold that:

$$\sum_y \frac{\widehat{p}(y)p_A(y;Q')}{p_A(y;Q^*)} - 1 \leq 0$$

Case 2: *Any maximizer Q^* of the unregularized likelihood ($\ell_{n,\mu=0}$) is not continuous.*

If the only maximizers of the unregularized problem are not continuous, then by the fundamental theorem of mixture models there exists a maximizer with at most $N+1$ discrete support points [Lindsay, 1995]. If $p_A(y|\gamma)$ is continuous in γ for each y then we can define a measure $\widetilde{Q}_\delta$ such that $\widetilde{Q}_\delta$ places the same point mass uniformly within a region of $\pm\delta$ of each of the point masses (some of which may only be $+$ or $-$ if the mass is near the boundary 0 or 1) then for all $\delta > 0, \widetilde{Q}_\delta$ is continuous measure. Since $p_A(y|\gamma)$ is continuous, $p_A(y;Q^*) = p_A(y;\widetilde{Q}_\delta) + \varepsilon(y;\delta)$ and for all $\epsilon > 0$ there exists a $\delta > 0$ such that $|\varepsilon(y;\delta)| < \epsilon$ for all y.

Therefore, computing the derivative in the direction of any singular measure from $\widetilde{Q}_\delta$

$$\begin{aligned}
\dot{\ell}_{n,\mu}(\widetilde{Q}_\delta; G) &= \sum_y \widehat{p}(y)\frac{p_A(y;Q_s)}{p_A(y;\widetilde{Q}_\delta)} - 1 - \mu \\
&= \sum_y \widehat{p}(y)\frac{p_A(y;Q_s)}{p_A(y;\widetilde{Q}_\delta)} - 1 - \mu \\
&= \sum_y \widehat{p}(y)\frac{p_A(y;Q_s)}{p_A(y;Q^*)}\left(\frac{1}{1+\frac{\varepsilon(y;\delta)}{p_A(y;Q^*)}}\right) - 1 - \mu
\end{aligned}$$

We choose ϵ such that $\epsilon < \frac{1}{2}\inf_y p_A(y;Q^*)$ then

$$\left|\frac{1}{1+\frac{\varepsilon(y;\delta)}{p_A(y;Q^*)}}\right| \leq 1 + 2\frac{|\varepsilon(y;\delta)|}{p_A(y;Q^*)}.$$

Now we also ensure $\epsilon \leq \min\{\mu, \frac{1}{2}\}\inf_y p_A(y;Q^*)$ and therefore:

$$\left|\frac{1}{1+\frac{\varepsilon(y;\delta)}{p_A(y;Q^*)}}\right| \leq (1+\mu)$$

Therefore since there exists $\delta > 0$ such that the above holds, $\widetilde{Q}_\delta$ is continuous and the Gateaux derivative in the direction of any singular measure is negative, therefore the maximizer must be a continuous measure ($P_M^* \ll P_U$), and therefore the maximizer does not include a singular component. □

We next introduce another lemma, which will justify that the density function of the maximizer will be bounded below

Lemma A.1.4. *The maximizer of $\ell_{n,\mu}$, P_M^* satisfies*

$$\operatorname{ess\,inf} p_M^*(\gamma) \geq \frac{\mu}{1+\mu} \tag{A.2}$$

where p_M^* is the corresponding density function of the measure P_M^*, and $\operatorname{ess\,inf} f$ *is the essential infimum of f, the maximum value α such that the set $\{x : f(x) < \alpha\}$ has (Lebesgue) measure* 0*.*

Proof. We prove this by contradiction and using the monotonicity of the EM algorithm. Suppose $\widetilde{P}_M$ is a maximizer such that $\operatorname{ess\,inf}_\gamma \widetilde{P}_M(\gamma) < \frac{\mu}{1+\mu}$. Denote $\widetilde{P}_{M_2}(\gamma)$ as the density function from taking one regularized NPEM step from $\widetilde{P}_M(\gamma)$. We can show

$$\ell_{n,\mu}[\widetilde{P}_{M_2}] - \ell_{n,\mu}[\widetilde{P}_M] \geq (1+\mu)\mathcal{D}(\widetilde{P}_{M_2}||\widetilde{P}_M)$$

But clearly by the monotonicity of the EM algorithm.

$$\begin{aligned}\widetilde{p}_{M_2}(\gamma) &= \left(\frac{1}{1+\mu}\right)\sum_y \frac{\widehat{p}(y)p_A(y|\gamma)\widetilde{p}_M(\gamma)}{\widetilde{p}_{MA}(y)} + \left(\frac{\mu}{1+\mu}\right) \\ &\geq \frac{\mu}{1+\mu}\end{aligned}$$

And if $\widetilde{p}_{M_2} \neq \widetilde{p}_M$ except for on a region of Lebesgue measure 0 then $\mathcal{D}(\widetilde{P}_{M_2}||\widetilde{P}_M) > 0$ and we guarantee $\ell_{n,\mu}[\widetilde{P}_{M_2}] > \ell_{n,\mu}[\widetilde{P}_M]$ and hence $\widetilde{P}_M$ is not a maximizer, a contradiction. □

We now introduce our final lemma

Lemma A.1.5. *A distribution (measure) P_M^* is a maximizer of $\ell_{n,\mu}$ if and only if:*

$$\sum_y \frac{\widehat{p}(y)p_A(y|\gamma)}{p_{MA}^*(y)} + \frac{\mu}{p_M^*(\gamma)} - 1 - \mu \leq 0 \quad a.e. \tag{A.3}$$

Where p_M^* is the corresponding density function to P_M^*

Proof. Recalling Theorem A.1.3 we have a sufficient and necessary condition for the optimality of a measure P_M^*.

We can define a direction $G = \widetilde{P}_M - P_M^*$ and by lemma A.1.2 we only need to consider the directions G which admit a density function and therefore can express the directions in terms of the difference of densities g. Then

$$\begin{aligned}
\ell_{n,\mu}[P_M^* + tG] &= \sum_y \widehat{p}(y) \log\left(\int p_A(y|\gamma)\left(p_M^*(\gamma) + tg(\gamma)\right)d\gamma\right) \\
&\quad + \mu \int \log\left(p_M^*(\gamma) + tg(\gamma)\right)d\gamma \\
&= \sum_y \widehat{p}(y) \log\left(\int p_A(y|\gamma)p_M^*(\gamma)d\gamma\right) + \mu\int \log\left(p_M^*(\gamma)\right)d\gamma + \\
&\sum_y \widehat{p}(y)\log\left(1 + t\int \frac{p_A(y|\gamma)g(\gamma)}{p_{MA}^*(y)}d\gamma\right) + \mu\int \log\left(1 + t\frac{g(\gamma)}{p_M^*(\gamma)}\right)d\gamma \\
&= \ell_{n,\mu}[P_M^*] + t\int\left(\sum_y \frac{\widehat{p}(y)}{p_{MA}^*(y)}p_A(y|\gamma) + \mu\frac{1}{p_M^*(\gamma)}\right)g(\gamma)d\gamma + O(t^2) \\
\implies \dot{\ell}_{n,\mu}(Q;G) &= \int\left(\sum_y \frac{\widehat{p}(y)}{p_{MA}^*(y)}p_A(y|\gamma) + \mu\frac{1}{p_M^*(\gamma)}\right)g(\gamma)d\gamma \\
&= \sum_y \frac{\widehat{p}(y)}{p_{MA}^*(y)}\int p_A(y|\gamma)p_M(\gamma)d\gamma + \mu\int\frac{\widetilde{p}_M(\gamma)}{p_M^*(\gamma)}d\gamma - 1 - \mu
\end{aligned}$$

Since this holds for any $\widetilde{P}_M$ which are continuous, at each point on $\gamma \in [0,1]$ this will hold for a sequence of continuous measures converging weakly to a point mass for each point on the interval. Since P_M^* is a continuous measure, it has a corresponding density function which has discontinuities at most on a set of Lebesgue measure 0. Therefore

$$\sum_y \frac{\widehat{p}(y)}{p_{MA}^*(y)}p_A(y|\gamma) + \mu\frac{1}{p_M^*(\gamma)} - 1 - \mu \leq 0 \quad a.e.$$

We note that with a minor abuse of notation, if $g \in L^2_{[0,1]}$ then we could express the Gateaux derivative using the inner product of a functional gradient element $\nabla\ell_{n,\mu}[P_M]$ and the direction g.

$$\nabla \ell_{n,\mu}[P_M] = \sum_y \frac{\widehat{p}(y)}{p_{MA}(y)} p_A(y|\cdot) + \mu \frac{1}{p_M(\cdot)}$$

and

$$\dot{f}_+(P_M; G) = \langle \nabla \ell_{n,\mu}[P_M], g \rangle.$$

This functional gradient itself is a function of γ and is useful for defining the NPEM algorithm

$$p_M^{(t+1)}(\gamma) = \frac{\nabla \ell_{n,\mu}[P_M](\gamma)}{1+\mu} p_M^{(t)}(\gamma),$$

as well as the optimality criterion

$$\nabla \ell_{n,\mu}[P_M](\gamma) \leq 1 + \mu \quad a.e.$$

Remark: As we proved in Lemma A.1.4, $1/p_M^*(\gamma)$ will be bounded. We compare this result to the optimality criterion of [Lindsay, 1995] a similar criteria is developed for the unregularized mixing distribution estimation problem, for which p_M^* is a maximizer if and only if

$$\sum_y \frac{\widehat{p}(y) p_A(y|\gamma) p_M^*(\gamma)}{p_{MA}^*(y)} - 1 \leq 0 \quad \forall \gamma$$

which is equal to our condition when $\mu = 0$. The main difference is due to the possibility of discontinuity in $\frac{1}{p_M^*(\gamma)}$ Lemma A.1.5 only holds up to a set of Lebesgue measure 0 while the unregularized case holds for all γ. □

Now we have all the required lemmas to prove the main part of the theorem.

Proof of Theorem A.1.1. We use a similar technique as in the convergence of the mixing distribution. [Chung and Lindsay, 2015]. Denote the sequence of measures generated by the EM algorithm as $\{P_M^{(t)}\}_{t=0}^{\infty}$. Since these are all measures defined on a compact domain i.e. $[0,1] \subset \mathbb{R}$, this sequence is tight. By Prokorov's theorem [Billingsley, 2013], this sequence is also sequentially compact.

Since this sequence is sequentially compact, there must exist a convergent sub-sequence $\{P_M^{(t_k)}\}_{k=0}^{\infty}$, where with limit P_M^{**}. Note that since $\ell_{n,\mu}[P_M^{(t)}]$ is monotonely increasing, and

$\ell_{n,\mu}[P_M^{(t+1)}] - \ell_{n,\mu}[P_M^{(t)}] \geq \mathcal{D}(P_M^{(t)}||P_M^{(t+1)})$ every sub-sequence must have their likelihood converge to the same likelihood value i.e. $\ell_{n,\mu}[P_M^{**}]$. Suppose that P_M^{**} is not the global optimum for $\ell_{n,\mu}$, then by Lemma A.1.5 there must be some region $G^* \subset [0,1]$ with non-zero Lebesgue measure where $\nabla \ell_{n,\mu}[P_M^{**}](\gamma) \geq \delta' > 1 + \mu$ for $\gamma \in G^*$. Since $\{P_M^{(t_k)}\}_{k=0}^{\infty}$ converges weakly to P_M^{**} then $\nabla \ell_{n,\mu}[P_M^{(t_k)}](\gamma) \geq \delta > 1 + \mu$ for $\gamma \in G^*$ for all k greater than some K.

Thus,

$$p_M^{(t_{k+1})}(\gamma) = \left(\frac{\nabla \ell_{n,\mu}[P_M](\gamma)}{1+\mu}\right)^{t_{k+1}-t_k} p_M^{(t_{k+1})}(\gamma) \geq \frac{\delta}{1+\mu} p_M^{(j)}(\gamma)$$

which further implies since $t_{k+1} - t_k \geq 1$

$$p_M^{(t_{k+1})}(\gamma) \geq \left(\frac{\delta}{1+\mu}\right) p_M^{(t_k)}(\gamma).$$

Therefore, when $k \to \infty$, $p_M^{(t_k)}(\gamma)$ diverges and is not a probability density function, hence the sequence converges uniquely to p_M^*. As a result,

$$\ell_{n,\mu}[P_M^{(t)}] \overset{t\to\infty}{\longrightarrow} \ell_{n,\mu}[P_M^*]$$

Furthermore, if we consider any sub-sequence of $\{P_M^{(t)}\}_{t=0}^{\infty}$, $\{P_M^{(t_k)}\}_{k=0}^{\infty}$, then the sub-sequence must have a further sub-sub sequence which converges to P_M^* weakly by the same argument as above. Thus, since every sub-sequence has a sub-sub sequence which converges to the maximizer, the primary sequence $\{P_M^{(t)}\}_{t=0}^{\infty}$ must converge weakly to P_M^*. □

A.2 Additional Computational Details

In this section we highlight several computational aspects not covered in the main paper.

A.2.1 Representation as a geometric program

Given a discrete approximation to the latent distribution, the maximum likelihood procedure can be expressed as a geometric program.

$$\begin{aligned} \sup_{\theta} \quad & \sum_{y=0}^{N} \widehat{p}(y) \log(\widetilde{A}_{y\cdot}^{\top}\theta) + \mu \frac{1}{R} \sum_{r=1}^{R} \log(R\theta_r) \\ \text{subject to} \quad & 1^{\top}\theta = 1 \\ & \theta \succeq 0. \end{aligned} \tag{A.4}$$

This form allows us to implement competing optimizers using a standard software in CVXR Fu et al. [2020].

A.2.2 Computing the discretized joint latent distribution

Recall $\theta_R(x), \vartheta_R(x) \in \Delta_R$, the discretized probability We compute the joint distribution according to Algorithm A.2.2. Since the joint distribution maps the quantiles from $\widehat{p}_M(\gamma|x)$ and $\widehat{p}_M(\zeta|x)$ respectively, the matrix $p(y, z|x)$ in Algorithm 12 will be sparse since there is a one-to-one mapping of quantiles in the joint distribution.

Algorithm 11 Joint Distribution Matrix Computation

1: Input: θ, ϑ, M(gridsize)
2: Define a uniform grid over $[0, 1] : \Omega$, let $M = |\Omega|$
3: Initialize $P = \{0\}^{M \times M}$
4: **for** $\omega \in \Omega$ **do**
5: $\quad n = Q(\omega, \theta)$
6: $\quad m = Q(\omega, \vartheta)$
7: $\quad P[n, m] = P[n, m] + \frac{1}{M}$
8: **end for**
9: **return** P (The discrete joint distribution)

Algorithm 12 Conditional distribution computation

1: Input: $\theta(x), \vartheta(x), m(\text{grid size}), p_A(y|\gamma), p_A(z|\zeta)$

2: Compute the discretized conditionals $A^{(y)} \in [0,1]^{N_1+1\times R_y}$, $A^{(z)} \in [0,1]^{N_1+1\times R_z}$

3: Compute the joint latent distribution matrix P_x using Algorithm A.2.2 with inputs $\theta(x), \vartheta(x), m$

4: Compute the joint distribution $p(y,z|x) = A^{(y)} P_x A^{T,(z)}$

5: Normalize each row of $p(y,z|x)$ to get $p(y|x) = \sum_z p(y,z|x)$

6: **return** $\widehat{p}(z|y,x)$

A.2.3 Selecting the number of bins

In practice, we would like to select a number of bins to discretize the latent distribution. Given a specified threshold of improvement on the likelihood, (we set the default to 10^{-5}), we can iterate through a sequence of bin sizes until the difference in the likelihood gained at the next bin is less than the threshold used in the Nonparametric EM algorithm. We illustrate this idea in Algorithm 13. In practice, we find that setting $R = 1000$ tends to work well for our applications and simulations.

Algorithm 13 Selecting R (The number of bins)

Input: $\widehat{p}$, $p_A(Y|\gamma)$, R_{min}, R_{max}, R_{step}, thresh.
Set: $R = R_{min}$
Compute: $A \in \mathbb{R}^{N+1\times R}$ from $p_A(y|\gamma)$
Estimate $\widehat{\theta}_R \in \mathbb{R}^R$ via the NPEM algorithm
Evaluate like_old $= \ell(A, \widehat{\theta}_R) = -\mathcal{D}(\widehat{p}||A\widehat{\theta}_R) - \mu\mathcal{D}(P_U||\widehat{\theta}_R)$
Set: diff $= \infty$
while $R < R_{max}$ **do**
    $R = R + R_{step}$
    Compute $A \in \mathbb{R}^{N+1\times R}$
    Estimate $\widehat{\theta}_R \in \mathbb{R}^R$ via the NPEM algorithm
    Evaluate like_new $= \ell(A, \widehat{\theta}_R) = -\mathcal{D}(\widehat{p}||A\widehat{\theta}_R) - \mu\mathcal{D}(P_U||\widehat{\theta}_R)$
    diff = like_new − like_old
    if diff < thresh **then**
        break
    end if
end while
return R

A.3 Additional simulations

In this supplement, we include simulations highlighting the steps required for score conversion. We begin with a comparison of the NPEM algorithm to the off-the-shelf convex solver, highlighting the improved computational speed of using the NPEM algorithm. We next show the consistency of the model selection procedure as n increases for various choices of the conditional distribution. We next show the finite sample valid feasibility tests can be used to reject poorly fitting models of the measurement assumption model. We next illustrate the selection of the tuning parameter μ in the regularization and highlight the corresponding approximation error of the latent distribution. We also give a heuristic for selecting the number

of bins for the use of the latent distributions in practice. Lastly, we include an alternative mean response function to replicate the conversion and prediction simulations of the main paper.

A.3.1 Computational comparison of EM algorithm vs CVXR (SCS)

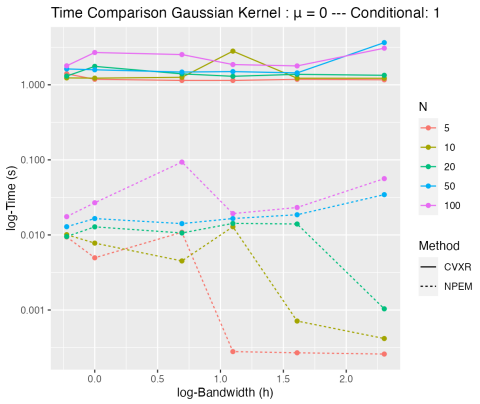
We first consider comparing the off-the-shelf convex solver (the Splitting Conic Solver in CVXR) to the nonparametric EM algorithm. We consider four different observed distributions to which we will compute a regularized NPMLE. We do this for varying sizes of the observed distribution N, the bandwidth of the conditional model (h), and the regularization parameter μ.

We use the following 4 distributions to simulate the observed scores Y

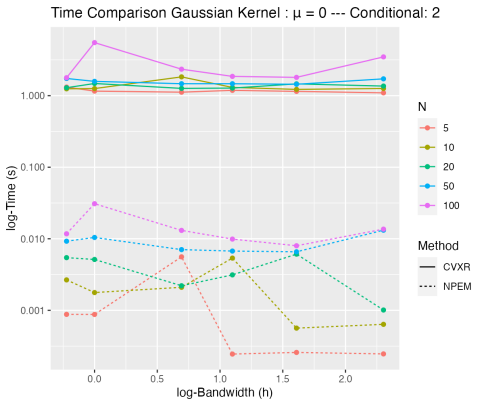
$$p_0(y) \propto \begin{cases} 1/(N+1) & 1\text{: Uniform} \\ -(y-N/2)^2 + 2*(N/2)^2 & 2\text{: Quadratic} \\ \exp(-y^2) + \exp(-(y-N)^2)) & 3\text{: Bimodal} \\ \sin(y4\pi/N) + 1 & 4\text{: Periodic} \end{cases}$$

We consider $N \in \{5, 10, 20, 50, 100\}$ and p_A given by the measurement kernel model where $K(x) = \exp(-x^2)$ (Gaussian) or $\exp(-|x|)$ (Exponential) and $h \in \{0.8, 1, 2, 3, 5, 10\}$, and $\mu \in \{0, 2\}$.

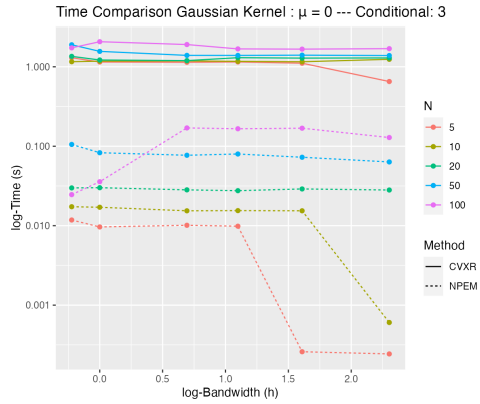
As we see in Figure A.1, the computational time is greatly improved by using the NPMLE compared to the off-the-shelf solver. In this case, the NPEM is consistently an order of magnitude faster for the same problem. In nearly all of these situations, we can achieve a likelihood value at least as good as the naive convex solver. These and the remaining simulations for different values of μ we repeat this for $\mu = 2$ in Figure A.3 and find the same trends hold.



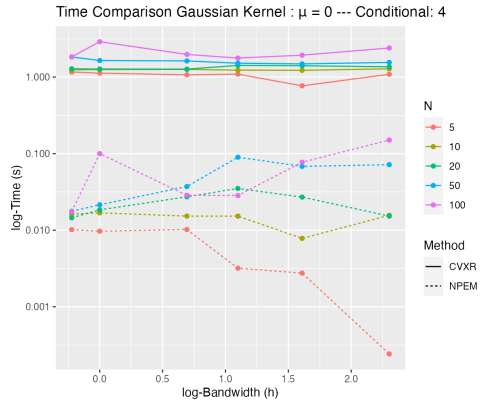
(a) Marginal 1: Uniform



(b) Marginal 2: Quadratic



(c) Marginal 3: Bimodal



(d) Marginal 4: Periodic

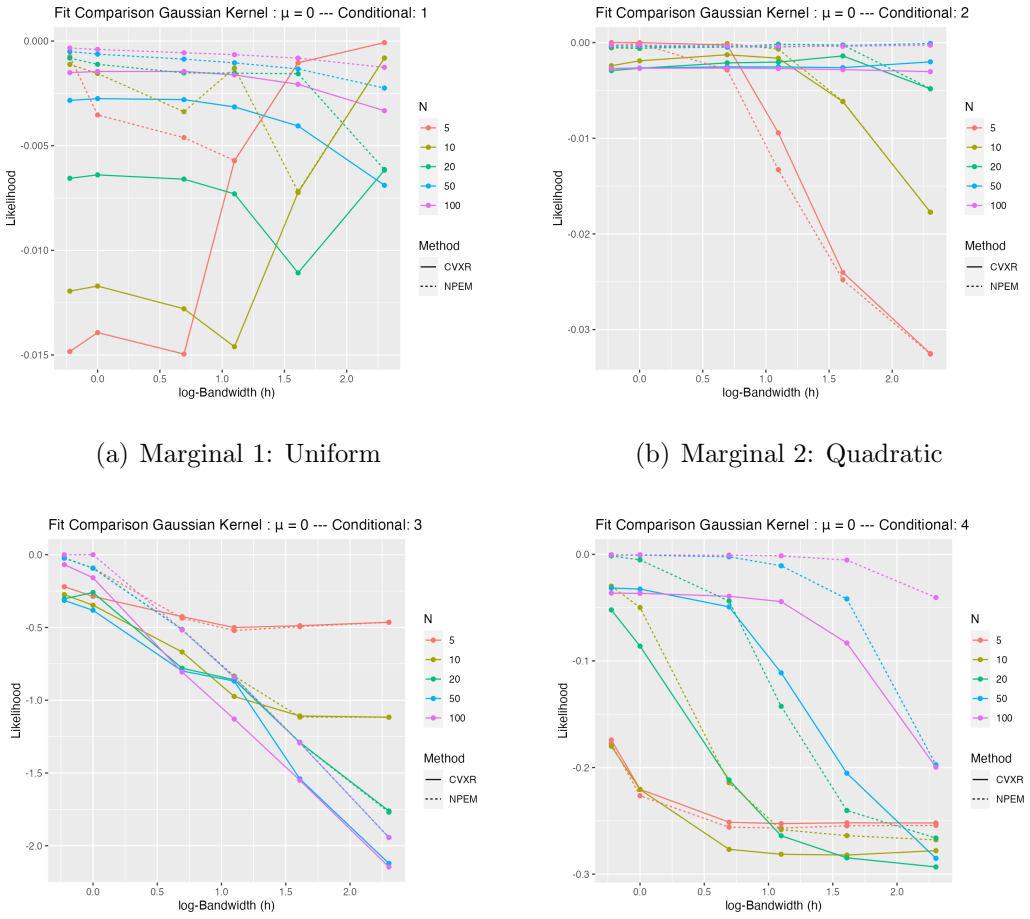
Figure A.1: Time Comparison of CVXR and NPEM for $\mu = 0$. The NPEM achieves a better fit in less time in nearly all settings.

A.3.2 Model Selection

We illustrate the model selection procedures for the measurement assumption model using the following latent model.

$$P_\gamma = \frac{2}{3}\text{Beta}(\alpha_1, \alpha_2) + \frac{1}{3}\text{Beta}(\alpha_2, \alpha_1)$$

where $\alpha_1 = 1.2$, $\alpha_2 = 3$, and $N = 30$. We illustrate the consistency of the model selection procedure while varying the true conditional model P_A. We sample $n \in$



(a) Marginal 1: Uniform

(b) Marginal 2: Quadratic

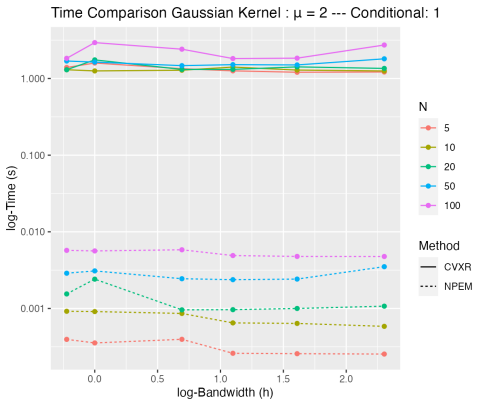
(c) Marginal 3: Bimodal

(d) Marginal 4: Periodic

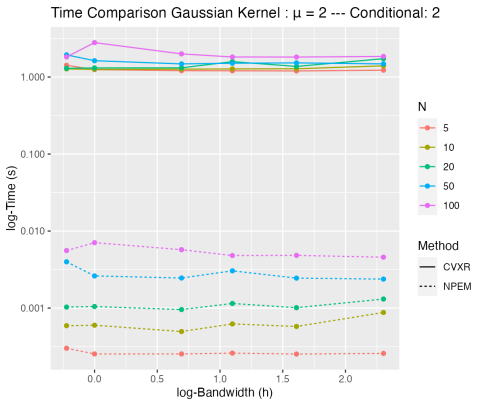
Figure A.2: Fit Comparison of CVXR and NPEM for $\mu = 0$. Larger values indicate a better fit in the model. The NPEM achieves a better fit in less time in nearly all settings.

$\{100, 500, 1000, 5000, 10000\}$ where $n_1 = n_2 = n$ and n_1 refers to the number of samples with a single observation and n_2 the number of samples with 2 observations. We repeat this for $n_{sims} = 500$ times for each setting.

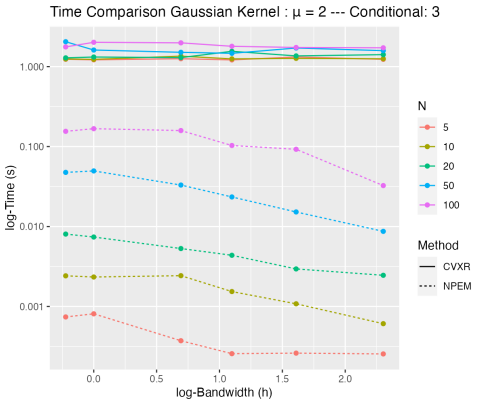
We consider a Gaussian and Exponential measurement kernel model where the true model has a bandwidth in the set $h \in \mathcal{H} = \{1.0, 2.0 \ldots, 5.0\}$. We then consider the average deviation of the estimated h^* from the true model h (i.e. $|h^* - h|$) as a function of the



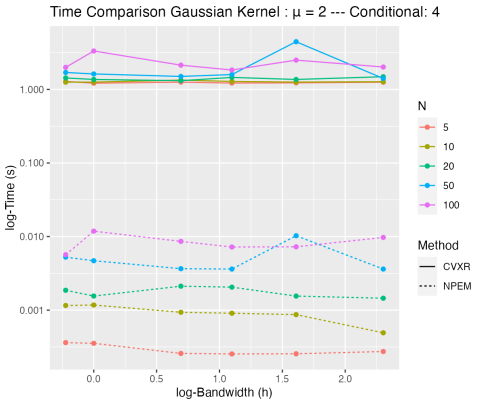
(a) Marginal 1: Uniform



(b) Marginal 2: Quadratic



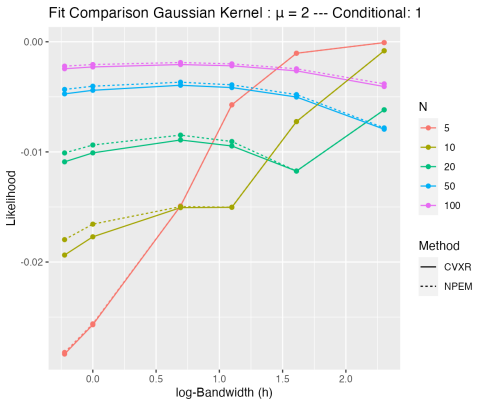
(c) Marginal 3: Bimodal



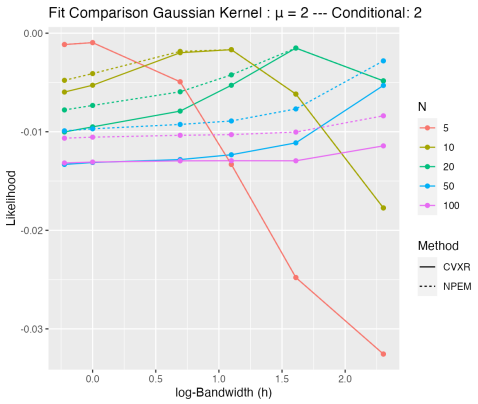
(d) Marginal 4: Periodic

Figure A.3: Time Comparison of CVXR and NPEM for $\mu = 2$. The NPEM achieves a better fit in less time in nearly all settings.

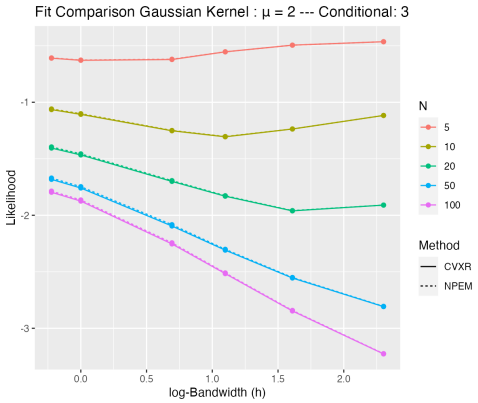
sample size. We see in Figure A.5 as we increase the sample size, in each case, the procedure eventually picks the correct model, and even for small sample sizes, the average deviation tends to be small.



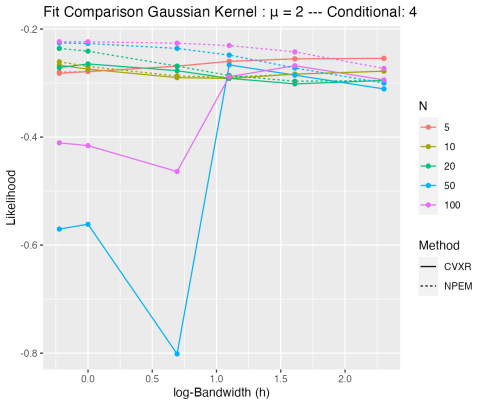
(a) Marginal 1: Uniform



(b) Marginal 2: Quadratic



(c) Marginal 3: Bimodal



(d) Marginal 4: Periodic

Figure A.4: Fit Comparison of CVXR and NPEM for $\mu = 2$. Larger values indicate a better fit in the model. The NPEM achieves a better fit in less time in nearly all settings.

A.3.3 Feasibility Tests

Along with a model selection procedure, we also have a finite sample valid feasibility test. We highlight the feasibility test using the same Bimodal Beta distribution as in the model selection simulations in Section A.3.2.

We set the true conditional distributions to be measurement error kernels with $h = 3$ with both Gaussian and Exponential Kernels. In both Figures A.6(a) and A.7(a) we see

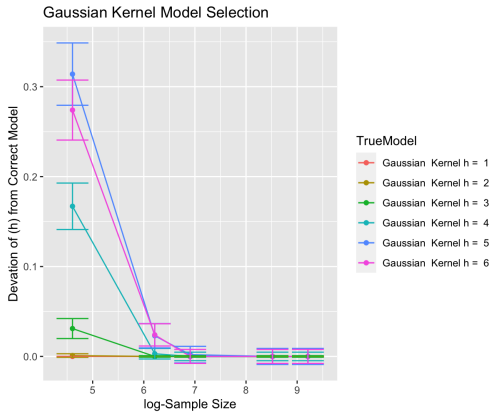
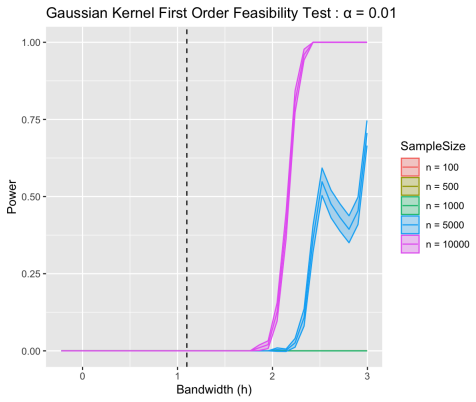


Figure A.5: Consistency of model selection of the conditional model with a Gaussian measurement kernel model.

that the first order test only has power when h is much too large, however, the second order test has power to reject measurement kernel models which have h too small. We see that as the sample size increases, the region of rejection grows (Figures A.6(b), A.7(b)). As in the model selection (Section A.3.2), we observe that selecting the proper conditional model requires more than a single observation, as this describes the measurement variability when holding γ fixed. Otherwise, for single observations of y_i per observational unit, a small enough h can fit any univariate distribution, and thus only rejects when h is too large.

A.3.4 Selecting the regularization tuning parameter

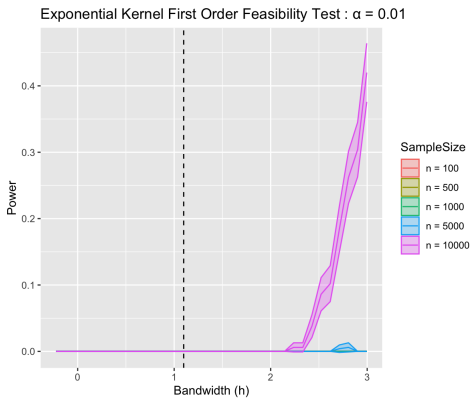
After selecting a measurement assumption model, we will select the regularization parameter in order to better fit the observed data. We also assume a Gaussian and Exponential Measurement assumption model with $h = 3$ for a Gaussian and Exponential measurement kernel model $P_A(y|\gamma)$. We illustrate the cross-validation procedure for $N = 30$ and $n \in \{100, 500, 1000, 5000\}$ and simulate 5000 times for each configuration. Figure A.8 shows that as the sample size increases, the optimal μ used for prediction shrinks and is similar for each conditional model.



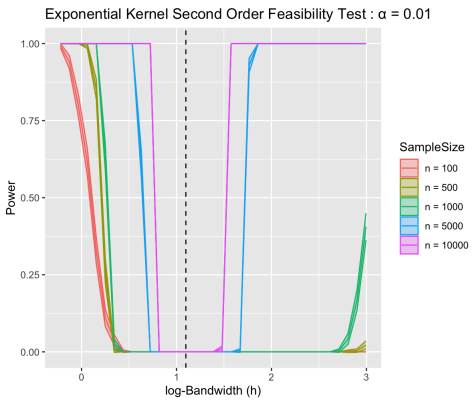
(a) First order feasibility tests Gaussian $h = 3$

(b) Second order feasibility tests Gaussian $h = 3$

Figure A.6: Feasibility test in simulations. We find the first order test can discriminate when h is too large, while the second order test can reject a model when h is too small.



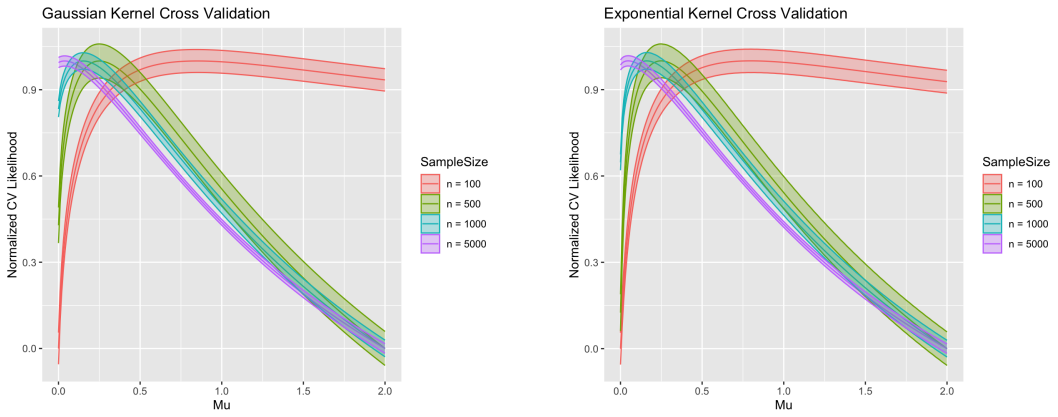
(a) First order feasibility tests Exponential $h = 3$



(b) Second order feasibility tests Exponential $h = 3$

Figure A.7: Feasibility test in simulations. We find the first order test can discriminate when h is too large, while the second order test can reject a model when h is too small.

Beyond just generalization to the observed distribution of Y, we can also plot the estimated distribution's distance to the true generative distribution. We find in Figure A.9 that



(a) Gaussian Measurement Kernel (b) Exponential Measurement Kernel

Figure A.8: Cross validation of the likelihood value as a function of the regularization parameter μ

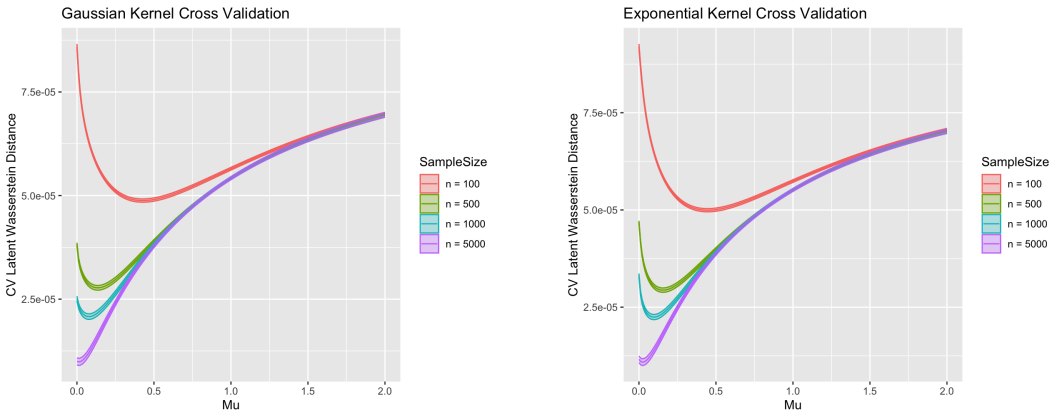
using the regularized estimates can improve the Wasserstein-1 distance to the true estimate. This effect is most pronounced for small sample sizes, where the variance is larger. The optimal μ for each latent approximation at each sample size also tends to be near the optimal μ for cross-validation on the observed data, both shrinking as the sample size increases. Though the true latent distribution is not identifiable, due to N being finite (Lindsay [1983a,b]), we still can improve on the approximation of the latent trait through regularization.

A.3.5 Robustness of the mean model

In this section, we illustrate the robustness of the simulation in the mean model through a different conditional model $m(X_i)$. In this case, we let:

$$m(x) = 0.5 + \min(-(1/100)(x - 71), -(1/5)(x - 71))$$

to represent a piece-wise linear model which slowly decreases up to 71, the proceeds with a more rapid decline. The remaining simulation parameters are the same as in the main paper.



(a) Gaussian Measurement Kernel (b) Exponential Measurement Kernel

Figure A.9: Latent Wasserstein-1 distance of the estimate to the true data generating latent distribution.

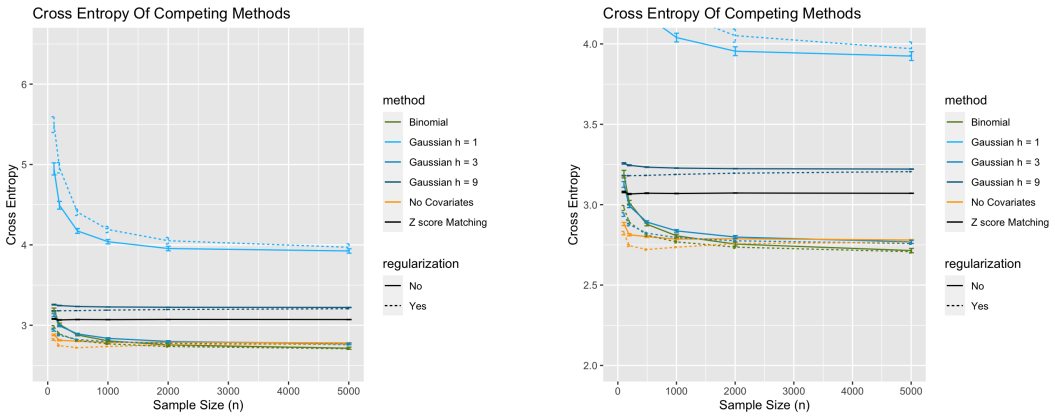
Simulations: Prediction

We once again compare the cross entropy of the various conversion methods under the piece-wise linear model. As in the main paper, in Figure A.10, we see that when the conditional model is correct, the best-performing model is the covariate smoothed version for large n. However, for small n, using our method without including covariates performs better. Additionally, we see that the severely misspecified models (ones for which $h = 1, 9$) perform poorly; however, the best model in that set $h = 3$ still performs well.

Simulations: Inference

Using the same setup, as the main paper, but with the new piece-wise linear response model, we consider the same problem. In this case the optimal $\beta_1 = -0.403403$ under the full data regression. We once again compare the observed bias, RMSE and coverage of each of the models using the multiple-imputation estimator with the bootstrap corrected confidence intervals.

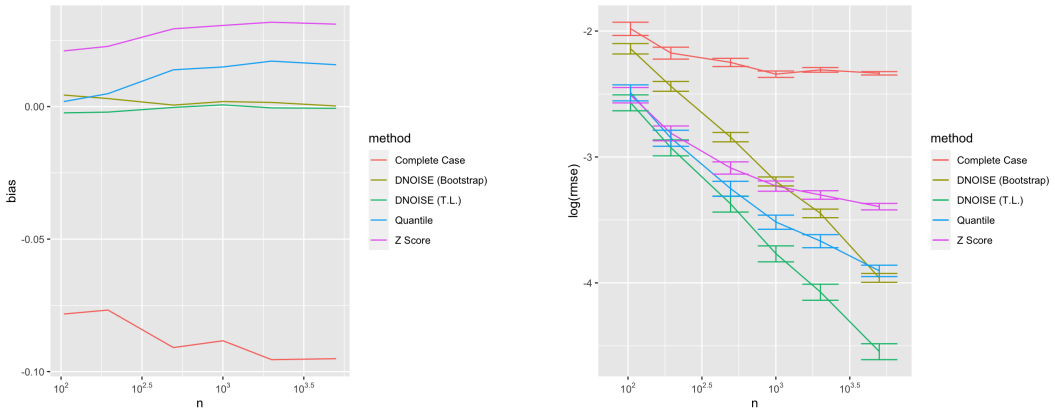
Again, we see in Figure A.12 that the proper coverage is obtained when using the



(a) Comparison of cross-entropy of conversion methods

(b) Zoomed in cross-entropy

Figure A.10: Piece-wise linear model conversion cross entropy. Data generated from a Binomial conditional models.



(a) Bias of the regression

(b) RMSE of the regression

Figure A.11: Piece-wise linear model, bias and RMSE of regression parameter estimation. All methods which do not account for covariates end up biased.

bootstrap-adjusted version. For this mean response function, the quantile-based conversion method tends to perform more competitively with the nonparametric latent trait models,

however, the performance becomes noticeably poorer under large sample sizes.

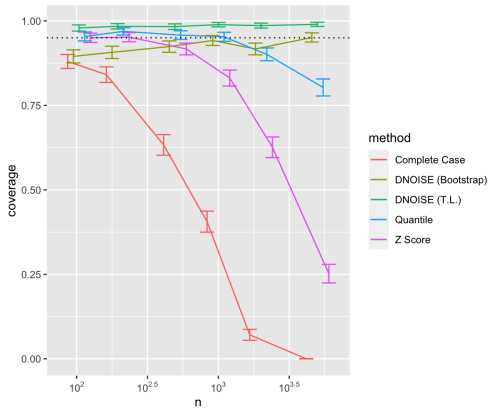


Figure A.12: Piece-wise linear model, coverage of imputation methods. Methods that do not include covariates undercover. Proper coverage is obtained only with our bootstrap procedure.

Appendix B

SUPPLEMENTARY MATERIAL FOR PROJECT 2

B.1 Additional Methodological Details

In this section we discuss extensions to several aspects of the paper with respect to the paper and fill in related additional mathematical constructions.

B.1.1 A central limit theorem for dependent data.

Although network models do not neatly fit into conventional time series or spatial dependency categories, we provide a general framework by satisfying the necessary conditions through common dependence assumptions such as M-dependence. This includes scenarios characterized by α-, ϕ-, or ρ-mixing [Bradley, 2005]. Our approach begins by defining affinity sets, (sets for which there is high correlation with an outcome) that form the foundational framework for applying the CLT, setting the stage for demonstrating its relevance and utility in analyzing network data.

Definition B.1.1 (Affinity sets). Denote a triangular array of mean 0 random vectors $W_{1:n}^{(n)}$ with dimension p. Let $\mathcal{A}_{(i,d)}^{(n)}$ denote an affinity set which contains all of the variables in the triangular array which are highly correlated with $W_{i,d}^{(n)}$, the d^{th} dimension of the i^{th} random variable.

The affinity sets can be used to construct a matrix which contains the bulk of the covariance across observations and dimensions. The regularity conditions can be understood as control of the covariance within affinity sets (B.1), control of the covariance across affinity sets (B.2) and control of the covariance outside of the affinity sets (B.3). We collectively refer to these as the affinity set conditions. The affinity sets can be used to construct a covariance

matrix $\Gamma_{n,dd'} = \sum_{i=1}^{n} \sum_{(j,d') \in \mathcal{A}_{(i,d)}^{(n)}} \text{cov}(W_{i,d}^{(n)}, W_{j,d'}^{(n)})$.

$$\sum_{(i,d):(j,d'),(k,d'')} \mathbb{E}[W_{i,d}|W_{j,d'}W_{k,d''}] = o(||\Gamma_n||_F^{3/2}), \tag{B.1}$$

$$\sum_{(i,d),(j,d');(k,d''),(l,\widehat{d})} \text{cov}(W_{i,d}W_{k,d''}, W_{j,d'}W_{l,\widehat{d}}) = o(||\Gamma_n||_F^2), \tag{B.2}$$

$$\sum_{(i,d)} \mathbb{E}[|\boldsymbol{W}_{-i,d}\mathbb{E}[W_{i,d}\boldsymbol{W}_{-i,d}]|] = o(||\Gamma_n||_F). \tag{B.3}$$

Theorem B.1.2 (Theorem 1 from Chandrasekhar et al. [2023]). *Denote a mean* 0 *triangular array of random vectors* $W_{1:n}^{(n)}$. *If a collection of affinity sets* $\mathcal{A}_{(i,d)}^{(n)}$ *satisfy the conditions of equations* (B.1), (B.2) *and* (B.3). *Then*

$$\Gamma_n^{-1/2} S_n \to_d N(0, I_p)$$

The authors illustrate several examples under which these conditions are sufficient for the this central limit theorem to hold. We next proceed with our main asymptotic results.

B.1.2 Inference for the OLS Estimator

Here we first give the full theorem and regularity conditions with respect to the OLS model.

Theorem B.1.3. *Let* $\widetilde{H}_i(\theta) = \mathbb{E}[\widetilde{h}(S_i(G), V_i(G))|\boldsymbol{a}, \boldsymbol{X}, G^*; \theta]$. *The OLS estimator uses the model averaged coefficients* $\widetilde{H}_i(\theta)$ *in place of the true unobserved coefficients* $\widetilde{h}_i$. *Let* $\mathsf{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{H}_i(\theta)\widetilde{H}_i^T(\theta)$. *Given an estimate of the model parameters* $\widehat{\theta}$, *we define the*

$$\widehat{\beta}_{ols} = \mathsf{H}_n^{-1}(\widehat{\theta}) \frac{1}{n} \sum_{i=1}^{n} \widetilde{H}_i(\widehat{\theta}) Y_i$$

Let $u_i = (\widetilde{h}(S_i(G), V_i(G)) - \widetilde{H}_i(\theta_0))\beta_0 + \epsilon_i$. *Suppose the following conditions hold for all* n.

Model Regularity conditions

D1. $\widehat{\theta}$ *is a* $s(n)$*-consistent estimate of the graph parameters* $||\widehat{\theta} - \theta_0|| = o_P(s(n))$

D2. $|\mathsf{H}_n(\theta) - \mathsf{H}_n(\theta')| \leq b_n(\boldsymbol{Z})||\theta - \theta'||$ *where* $b_n(\boldsymbol{Z}) = O_P(1)$ *(that is,* $b_n(\boldsymbol{Z})$ *is stochastically bounded).*

D3. $\max_i ||\widetilde{H}_i(\theta) - \widetilde{H}_i(\theta')|| \leq b_n(\boldsymbol{Z})||\theta - \theta'||$

D4. $||H_i(\theta)|| \leq M < \infty$

D5. $\left|\frac{1}{n}\sum_{i=1}^n |u_i| - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|u_i|]\right| = o_P(1)$

Lastly, let Γ_n *denote a matrix that satisfies the following central limit theorem for the estimating function*

Central Limit Theorem

E1. For the array of random variables $\mathcal{G}_i = \frac{1}{n}H_i(\theta_0)u_i$*, there exists a set of affinity sets* $\mathcal{A}^n_{(i,d)}$ *such that* (B.1)*,* (B.2) *and* (B.3) *are satisfied with a corresponding matrix* Γ_n*, where* $\sqrt{\lambda_{min}(\Gamma_n)} = r(n)$*.*

Then if $r(n) = o(s(n))$

$$\Gamma_n^{-1/2}\mathsf{H}_n(\widehat{\theta})(\widehat{\beta}_{ols} - \beta_0) \to_d N(0, I_p)$$

B.1.3 Estimation of Network Models

We next discuss the estimation of generative models of network formation using several datatypes. We summarize the information for using ARD in Table B.1.3 as discuss similar rates for other datatypes.

Estimation of the Stochastic Blockmodel Using Sampled Data

We illustrate that it is possible to estimate the stochastic blockmodel using a diverse set of partial and sampled network data types. In each case, $\mathbf{P}_{kk'}$ refer to the cross-block probabilities, while $k_i \in \{1, 2, \ldots, K\}$ denote the node memberships. We consider *partial network data* to be any subset of the network data which can be used to generate an estimate of the generative model $\widehat{\theta}$.

Example (Induced subgraph). We sample $m \leq n$ of nodes in the graph randomly, with at least one node from each of the K communities. Let G' be the sub-graph induced by these

Network Model	Norm	ARD Rate
SBM	$\sum_{k,k'} \lvert\widehat{P}_{kk'} - P_{kk'}\rvert$	$\widetilde{O}_P(K/n)$
Latent Space	$\sup_{i\in\{1,2,\ldots,n\}} \lvert\widehat{\theta}_i - \theta_i\rvert$	$O_P(\sqrt{log(n)/n})$
Beta Model	$\sup_{i\in\{1,2,\ldots,n\}} \lvert\widehat{\theta}_i - \theta_i\rvert$	$O_P(\sqrt{log(n)/n})$
Low-Rank Graphon	$\frac{1}{n^2}\lvert\lvert\widehat{\eta} - \eta_0\rvert\rvert_2$	$\widetilde{O}_P(1/T)$

Table B.1: Summary of estimation rates with respect to model classes. The norms used for the latent space and beta models are with respect to their individual parameters θ_i. We let $\eta_{0,ij} = P(G_{ij} = 1|\theta_0)$ denote the probability of two nodes connecting in the graphon model. Rates for the latent space and beta models are derived in Breza et al. [2023] and the low-rank graphon in Alidaee et al. [2020].

m nodes. Let N'_k denote the set of sampled nodes in community k, assumed to be positive for each k. Let

$$\widehat{\mathbf{P}}_{kk'} = \frac{1}{|N'_k||N'_{k'}|} \sum_{i\in N'_k} \sum_{j\in G'_{k'}} G'_{ij} \ .$$

Example (Edges missing). Suppose that edges are missing according to some distribution. Let G' be the observed graph, and suppose that $P(G'_{ij} = 1|X_{ij} = x)$ is the probability of observing the edge G'_{ij}, given dyad-level covariates X and the edge G_{ij}. Suppose that we have a consistent estimator of this conditional response. Then,

$$\widehat{\mathbf{P}}_{kk'} = \frac{1}{|N'_k||N'_{k'}|} \sum_{i\in N'_k} \sum_{j\in G'_{k'}} \frac{G'_{ij}}{\widehat{P}(G'_{ij} = 1|X_{ij})} \ .$$

Lemma B.1.4 (Rates for induced subgraph and Edges Missing)**.** *Consider an estimate for a stochastic blockmodel cross probabilities based on either the induced subgraph or the edges missing example of $m \leq n$. Let $m_k = |N_k| = \rho_k m$ for some $\rho_k \in (0, 1)$. Then with probability at least $1 - \delta$*

$$|\widehat{\mathbf{P}}_{kk'} - \mathbf{P}_{kk'}| \leq \frac{1}{\rho_k \rho_{k'} m} \sqrt{\frac{\log(2/\delta)}{2}} \tag{B.4}$$

Further, suppose that $\sup_x |\widehat{P}(G_{ij} = 1|X_{ij} = x) - P(G_{ij} = 1|X_{ij} = x)| = o_P(m^{-1})$ with $P(G_{ij} = 1|X_{ij} = x) \geq \lambda > 0$. Then for large enough m, equation B.4 holds for the missing

edges example as well.

Lastly, we discuss respondent driven sampling. In this setting, community membership can be defined based on a partition of the covariates, thus allowing for an observable trait in the graph, a similar strategy is adopted by Roch and Rohe [2018].

Example (Respondent driven sampling). Let $i \in \{1, 2, \ldots, m\}$ denote the indices of a sample of individuals obtained through respondent driven sampling. An initial number of individuals are recruited as seeds, and subsequent individuals are recruited via referrals from the others in a population. Tran and Vo [2021] develop a consistent estimator for the model parameters of the stochastic blockmodel.

Let $\widetilde{G}_m$ be the subgraph of G_n sampled from a set of nodes $\{1, 2, \ldots, m\}$. Let M_k denote the number of individuals in the subsample of type k and let $M_{kk'}^{\leftrightarrow}$ denote the number of connected individuals in the subgraph $\widetilde{G}_m$.

The cross-type probabilities can be estimated as follows:

$$\widehat{\boldsymbol{P}}_{kk'} = \begin{cases} \frac{M_{kk'}^{\leftrightarrow}}{M_k M_{k'}} & \text{When } k \neq k' \\ \frac{M_{kk}^{\leftrightarrow}}{M_k(M_k - 1)} & \text{otherwise} \end{cases}$$

Tran and Vo [2021] illustrate the consistency of these parameters (Theorem 4.2 in their paper), in particular $|\widehat{\boldsymbol{P}}_{kk'} - \boldsymbol{P}_{kk'}| = O_P(m^{-1})$

Estimation of Other Network Models

Though we emphasise the estimation of the stochastic blockmodel, there are several other methods available for estimation of the network formation model. These include the beta model of Chatterjee and Diaconis [2011], in which the graph generation model consists of two model parameters ν_i, ν_i possibly altered through some additional dyadic covariates X_{ij}^*

$$P(G_{ij} = 1|\theta_0) = \widetilde{f}(\nu_i + \nu_j + \beta^T X_{ij}^*)$$

where $\widetilde{f}$ is a link function. Alternatively one can consider the latent space model of Hoff et al. [2002a] which include latent positions on some unobserved manifold $\mathcal{M}^p$.

$$P(G_{ij} = 1|\theta_0) = \widetilde{f}(\nu_i + \nu_j + d(Z_i, Z_j))$$

In each of these cases Breza et al. [2023] illustrate consistent estimation rates in the $||\widehat{\theta} - \theta_0||_\infty = \mathcal{O}_P\left(\sqrt{\frac{\log(n)}{n}}\right)$ with the use of aggregated relational data. Since this represents the coarsest datatype we expect similar rates to hold for subgraph sampling and respondent driven sampling. Though this rate is too slow for the to ignore the effect of the estimation of the graph model, in examples where one expect a high level of correlation among the outcomes it can be practical to use these methods.

B.1.4 Plug-in estimates of the Causal parameter

For many problems, the parameter of interest is a causal query conditional on the complete graph G as described in Section 3.2.3. For example, one may care about the expected number of adoptions after seeding an individual in block k v.s. block k'. In this section, we illustrate how to construct an estimate of the causal parameter $\Psi(\boldsymbol{a}|G)$ using our conditional model estimation procedure.

Let $\Psi(\boldsymbol{a}|\theta_0) = \mathbb{E}[\Psi(\boldsymbol{a}|G)|\boldsymbol{a}, \boldsymbol{X}, \theta_0]$ be the average causal effect of policy $\boldsymbol{a}$ over all draws of the graph model θ_0. We will establish conditions under which these two quantities are close to one another.

Recall the true conditional mean function $\mathbb{E}[Y|S_i = s, V_i = v] = h_0(s, v)$. Under a correctly specified conditional model, $h_0(s, v) = h(s, v; \beta_0)$, and $\Psi(\boldsymbol{a}|\theta_0) = \Psi(\boldsymbol{a}|\beta_0, \theta_0)$ where

$$\Psi(\boldsymbol{a}|\beta, \theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(V_i, S_i; \beta)|\boldsymbol{a}, \boldsymbol{X}, \theta]. \tag{B.5}$$

In order to estimate $\Psi(\boldsymbol{a}|G)$ we plug-in the estimates for the mean model and network model $\Psi(\boldsymbol{a}|\widehat{\beta}, \widehat{\theta})$. We next discuss the asymptotics of the plug-in estimate.

Lemma B.1.5 (Inference for a plug-in causal parameter)**.** *Assume the conditions of 3.3.1.*

Further, assume:

$$\sup_{\beta} |\mathbb{E}[h(S_i(\boldsymbol{X}; G), V_i(\boldsymbol{a}|G); \beta)|\boldsymbol{a}, \boldsymbol{X}, \theta] - \mathbb{E}[h(S_i(\boldsymbol{X}; G), V_i(\boldsymbol{a}|G); \beta)|\boldsymbol{a}, \boldsymbol{X}, \theta']| \leq b_i||\theta - \theta'|| \tag{B.6}$$

where $b_i \leq M < \infty$. *Denote*

$$Q_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta'} \mathbb{E}[h(S_i(\boldsymbol{X}; G), V_i(\boldsymbol{a}|G); \beta')|\boldsymbol{a}, \boldsymbol{X}, \theta_0]\Big|_{\beta'=\beta} \in \mathbb{R}^{1\times p}$$

and

$$\widetilde{\omega}_n := Q_n(\beta_0) D_n(\beta_0) \Gamma_n D_n(\beta_0)^T Q_n(\beta_0)^T.$$

If $s(n) = o(\sqrt{\widetilde{\omega}_n})$. *Then*

$$\widetilde{\omega}_n^{-1/2}(\Psi(\widehat{\beta}, \widehat{\theta}) - \Psi(\beta_0, \theta)) \to_d N(0, 1) \tag{B.7}$$

This lemma is essentially an application of the delta method, with the additional caveat that we estimate θ before the plug-in estimate. As before, this requires a fast estimate of the graph generative model parameter, but we add the slightly different assumption (B.6) that the smoothness in the model class is over the conditional response models $\mathbb{E}[h(S_i, V_i; \beta)|\theta]$, rather than the estimating function $\widetilde{m}(Y, S, V|\beta, \theta)$.

Convergence of the causal parameter to the average over graphs

As we have previously discussed, we can only hope to estimate $\Psi(\boldsymbol{a}|\theta_0)$ as we do not have access to the full graph G. We next introduce a simple conditions under which the parameter $\Psi(\boldsymbol{a}|G)$ is close to its average over draws of the graph $G \sim \theta_0$, $\Psi(\boldsymbol{a}|\theta_0)$.

Assumption B.1.6 (v_n-response dependence)**.** *For any graph draw* G *let* $G'^{(ij)}$ *denote the graph* G *with the* ij *entry swapped from* 0 *to* 1 *or vice versa. Let* $c_{ij,n}$ *denote the bounds of the differences such that*

$$\left|\frac{1}{n}\sum_{i=1}^{n} h_0(S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G)) - h_0(S_i(\boldsymbol{X}, G'^{(ij)}), V_i(\boldsymbol{a}, G'^{(ij)}))\right| \leq c_{ij,n} \tag{B.8}$$

And let $v_n^2 = \sum_{ij:i\neq j} c_{ij,n}^2$

Lemma B.1.7. *Under Assumption B.1.6*

$$\Psi(\boldsymbol{a}|G) - \Psi(\boldsymbol{a}|\theta_0) = O_P(v_n)$$

The proof is a one-line application of McDiarmid's inequality. Previous related work such as Breza et al. [2023] typically assume that such a quantity is consistent, however here we quantify the rate here. We next highlight an example;

Example (Conditional Mean Function Example). We abbreviate $G = G$ and $G' = G'^{(kl)}$. Let $h_0(S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G)) = \beta_0 + \beta_1 a_i + \beta_2 X_i + \beta_3 \sum_{l \neq i} \frac{X_l G_{kl}}{n} + \beta_4 \sum_{l \neq k} \frac{a_l G_{il}}{n}$ denote a linear response function dependent on the density of connected neighbors. Suppose that the covariate values are bounded $|X_i| \leq M < \infty$. Then:

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{k=1}^{n} h_0(S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G)) - h_0(S_i(\boldsymbol{X}, G'), V_i(\boldsymbol{a}, G')) \right| \\
&= \left| \frac{1}{n} \sum_{k=1}^{n} \sum_{l \neq k} \beta_3 \frac{X_l (G_{kl} - G'_{kl})}{n} + \beta_4 \frac{a_l (G_{kl} - G'_{kl})}{n} \right| \\
&\leq \left| \frac{1}{n} \frac{X_j + X_i}{n} + \frac{|a_j| + |a_i|}{n} \right| \\
&\leq (2M + 2)/n^2
\end{aligned}$$

Applying B.1.7 illustrates that: $\Psi(\boldsymbol{a}|G) - \Psi(\boldsymbol{a}|\theta_0) = O_p(n^{-1})$. Hence in order to estimate the expected average outcome, all we need is a consistent estimate of the model parameters β_0.

B.1.5 Experimental Design Variance Minimization With Model Uncertainty

As an extension of our variance minimizing procedure, we can incorporate the uncertainty in our estimates of the model parameters. For instance, consider the following parametric bootstrap approach for estimating the model parameters of the stochastic blockmodel when using ARD.

For example, consider a scenario where we utilize the stochastic blockmodel and we collect ARD. Denote $\widehat{\theta} = (\{\widehat{Z}_i\}_{i=1}^n, \widehat{\boldsymbol{P}})$ the initial estimate of the model as computed from

Lemma 3.3.4. We can construct a sampling distribution of $\widehat{\theta}^{(b)}$ using the following procedure. Let X_{it}^* denote the ARD responses of the number of connections individual i has to someone of trait t and let $T_i \in \{0,1\}^T$ denote the trait memberships of the corresponding individuals.

1. Estimate $\widehat{\theta}$ from $\boldsymbol{X}^*$
2. For $b \in \{1, 2, \ldots, B\}$
   (a) Sample $G^{(b)} \sim \widehat{\theta}$
   (b) Construct the ARD vector based on the resampled responses $X_{it}^{*(b)}$ using counts according to connections of $G^{(b)}$ to the nodes with corresponding traits $\{T_i\}_{i=1}^n$
   (c) Estimate $\widehat{\theta}$ from $\boldsymbol{X}^{*(b)}$

This approach can work for any procedures which can allow for a sampling distribution of the model parameters $\{\widehat{\theta}^{(b)}\}_{b=1}^B$. For example Baraff et al. [2016] considers a nonparametric bootstrap for respondent driven sampling.

In all such cases where $\widehat{\theta}$ is modeled with uncertainty, we apply Algorithm 5 to each of the b draws. Since the model is equivalent under cluster permutations, we choose the permutation for each $\{\widehat{Z}_i^{(b)}\}_{i=1}^n$ which minimizes the classification error with respect to $\{\widehat{Z}_i\}_{i=1}^n$. This is implemented using standard software, for example in the `label.switching` `R` package Papastamoulis [2015]. Thus far we have discussed assigning seeds or treatments from the perspective of designing more efficient experiments, however, in many applications, one may wish to select nodes which will maximize some outcome over the network, such as diffusion processes.

B.2 A discussion on the frameworks of interference

We contrast the approaches of a fixed outcome approach as in Aronow and Samii [2017] to a structural causal model approach. In the former approach, each individual has a distinct

outcome under an exposure v, $Y_i(v)$. Though such an approach is robust for learning parameters such as average treatment effects $\frac{1}{n}\sum_{i=1}^{n} Y_i(v)$, the information in an individual i's potential outcome is completely distinct from individual j. This important details has important downstream implications.

Consider the simple contagion model from the example in Section 3.2.1 which takes place in a single time period ($T = 1$). Consider the nodes i, j in Figure B.1 with seeded nodes in blue. Suppose that at time $T = 1$, that each neighbour of a treated node is infected with probability q. Since each one has only a single treated neighbor the distribution of the infection probability $P(Y_i = 1|\boldsymbol{a}, G)$ i and j are equivalent as their exposures are identical (i.e. they are each connected to a single seed node). However, in the finite sample framework the potential outcomes of any two nodes with a single treated neighbor can be arbitrarily different ($Y_i(v) \neq Y_j(v)$).

This nonparametric structure imposed on the potential outcomes later imposes restrictions on the degree of influence of others a node can have for estimation, thereby limiting this framework to examples with local dependencies (a phenomena also seen in Ogburn et al. [2022]).

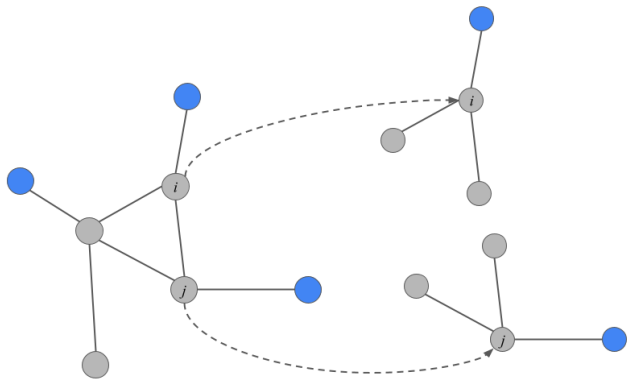


Figure B.1: Equivalence of distribution of potential outcomes of nodes i and j are equivalent under this given treatment assignment as all of the rooted networks are equivalent.

B.2.1 Why not IPW estimators?

In many nonparametric approaches to estimating causal quantities under interference, inverse probability weighted (IPW) estimates can be developed given a randomization scheme $P_{\boldsymbol{A}}$ [Aronow and Samii, 2017]. This is useful as it can be used to develop estimators for causal effects any exchangeability assumptions on the potential outcomes. However when V_i is not observed directly, we must leverage additional structure in order to estimate any causal effects.

Our objective is to understand the model's structure and often apply it to tasks such as seeding. Thus, we rely on a correct model specification. The challenge with developing an IPW estimator arises when exposure is not observed. In such cases, it becomes impossible to determine which potential outcome was observed, violating the causal consistency assumption. Specifically, we don't know which potential outcome Y_i represents (i.e., which exposure $v, Y_i = Y_i(v)$).

B.3 Proofs of paper theorems:

B.3.1 Proof of Lemma B.1.4

Proof. The proof is straightforward application of Hoeffding's inequality. Given an m node subsample of the full graph, and given their known types. Since $\widehat{\mathbf{P}}_{kk'} = \frac{1}{\rho_{kk'}m}\sum_{i,j} G_{ij} I(k_i = k, k_j = k')$, then the final result is a direct application of Hoeffding's inequality.

For the missing data case, we can plug-in the estimate of the edge sampling $P(G_{ij} = 1|X_{ij} = x)$ in order to correct for the missingness of the edges. If $\sup_x |\widehat{P}(G_{ij} = 1|X_{ij} = x) - P(G_{ij} = 1|X_{ij} = x)| = o_P(m^{-1})$ then the estimation of the propensity is negligible and we can correct for the missingness of edges. □

B.3.2 Proof of Lemma 3.3.4

Proof. Under the stochastic blockmodel assumption, the true latent traits are some discrete type $k_i \in \{1, 2, \ldots, K\}$. Then the mean connection probability Z_{ck} is simply a mixture over

the connection probabilities, weighted by $P(k_j = k'|t_j = t)$. Let N_k denote the set of individuals with group k membership. Furthermore, let $n = |N_k|$. Denote analogous quantities for the trait memberships N_t as well as the intersection of k and t by N_{kt}. When we have a correct clustering. Denote $\widehat{P}_{kt} = \frac{1}{n_k}\sum_{i\in N_k}\frac{1}{n}\widetilde{Y}_{it}$. Assuming independent samples conditional on the graph clusters, let $P_{kt} = \frac{1}{n_t}\sum_{k'\in[K]}\sum_{i\in N_{tk'}} P_{kk'}$ denote the mean probability of connection averaged over the clusters conditional on their latent traits. Let $\omega_{kt} = \frac{n_{kt}}{n_t}$.

We can express $\widetilde{P}_{kt} = P(G_{ij} = 1|k_i = k, t_j = t)$ as a weighted sum of the connection probabilities from the constituent distributions. If the true clusters are known, then these proportions ω_{kt} are known exactly from the data. Then

$$\begin{aligned}
\widetilde{P}_{kt} &= P(G_{ij} = 1|k_i = k, t_j = t) \\
&= \sum_{k'=1}^{K} P(G_{ij} = 1|k_i = k, k_j = k', t_j = t)P(k_j = k'|t_j = t) \\
&= \sum_{k'=1}^{K} P(G_{ij} = 1|k_i = k, k_j = k', t_j = t)\omega_{k't} \\
&= \sum_{k'=1}^{K} P(G_{ij} = 1|k_i = k, k_j = k')\omega_{k't} \\
&= \sum_{k'=1}^{K} P_{kk'}\omega_{k't}
\end{aligned}$$

Expressing this relationship over the whole set of matrices, we have:

$$\widetilde{P} = \Omega P$$

Where $\Omega_{tk} = \frac{n_{tk}}{n_k}$,

We can solve this system as long as the columns of Ω are linearly independent. Therefore:

$$P = (\Omega^T\Omega)^{-1}\Omega^T\widetilde{P}$$

We next bound the estimation error in Frobenius norm of the true cross-cluster proba-

bilities

$$\begin{aligned}
||\widehat{P} - P||_F &= ||(\Omega^T\Omega)^{-1}\Omega^T(\widehat{\widetilde{P}} - \widetilde{P})||_F \\
&\leq ||(\Omega^T\Omega)^{-1}\Omega^T||_F||(\widehat{\widetilde{P}} - \widetilde{P})||_F \\
&\leq \sqrt{||(\Omega^T\Omega)^{-1}\Omega^T||_F^2}||(\widehat{\widetilde{P}} - \widetilde{P})||_F \\
&\leq \sqrt{\mathrm{Tr}\left((\Omega^T\Omega)^{-1}\Omega^T\Omega(\Omega^T\Omega)^{-1}\right)}||(\widehat{\widetilde{P}} - \widetilde{P})||_F \\
&= \sqrt{\mathrm{Tr}\left((\Omega^T\Omega)^{-1}\right)}||(\widehat{\widetilde{P}} - \widetilde{P})||_F
\end{aligned}$$

Since we assume that Ω's column's are linearly independent, then $\Omega^T\Omega$ is invertible. Therefore, what remains is bounding the Frobenius norm of $||(\widehat{\widetilde{P}} - \widetilde{P})||_F$.

For each element, let

$$\begin{aligned}
\widehat{\widetilde{P}}_{tk} &= \frac{1}{n_k n_t}\sum_{i\in N_k}\widetilde{Y}_{ik} \\
&= \frac{1}{n_k n_t}\sum_{i\in N_k}\sum_{j\in N_t} G_{ij}
\end{aligned}$$

Therefore, applying Hoeffding's inequality

$$P(|\widehat{\widetilde{P}}_{tk} - \widetilde{P}_{tk}| \geq \epsilon) \leq 2\exp\left(-2\epsilon^2 n_k n_t\right)$$

Letting $\rho_k = \frac{n_k}{n}, \rho_t = \frac{n_t}{n}$, then

$$P(|\widehat{\widetilde{P}}_{tk} - \widetilde{P}_{tk}| \geq \epsilon) \leq 2\exp\left(-2\epsilon^2 \rho_k\rho_t n^2\right)$$

Therefore, by a union bound,

$$\begin{aligned}
&P(\max_{k,t}|\widehat{\widetilde{P}}_{tk} - \widetilde{P}_{tk}| \geq \epsilon) \leq 2KT\exp\left(-2\epsilon^2\rho_k\rho_t n^2\right) \\
\implies\ &P(\sum_{k,t}|\widehat{\widetilde{P}}_{tk} - \widetilde{P}_{tk}| \geq KT\epsilon) \leq 2KT\exp\left(-2(KT)^2\epsilon^2\rho_k\rho_t n^2\right)
\end{aligned}$$

Therefore,

$$||\widehat{\widetilde{P}}_{tk} - \widetilde{P}_{tk}||_1 = \mathcal{O}_P(\frac{KT\sqrt{\log(KT)}}{n})$$

Hence

$$||\widehat{P} - P||_2 = \mathcal{O}_P(\frac{KT\sqrt{\log(KT)}}{n})$$

Lastly, we show that as n grows, the probability of achieving a correct clustering of the true block memberships approaches 1. Recall that $n_t = \rho_t n$, and let $\underline{\rho_T} = \min_t \rho_t$. By Hoeffding's inequality: $P(||X_i^\dagger - Z_{k_i}|| > \epsilon_n) \leq 2\exp(-2\epsilon_n^2 n/\underline{\rho_T})$. Taking a union bound over all response vectors, $P(\max_i ||X_i^\dagger - Z_{k_i}|| > \epsilon_n) \leq 2n2\exp(-2\epsilon_n^2 n/) \to 0$ for all $\epsilon_n = o(\sqrt{\log(n)/n})$.

Therefore, as n grows, the normalized response vectors in each cluster become well separated, and once $\epsilon_n < \min ||Z_k - Z_{k'}||/2$, then all clusters will be well separated and naively hierarchical agglomerative clustering will consistently group the blocks together for K clusters. Therefore for example, if we let $\epsilon_n = \log(n)n^{-1/2}$,then $P(\max_i\{\widehat{k} \neq k\} = O(\frac{1}{n}))$. Of course the labels learned are only consistent up to permutation. We exploit the fact that as referred to in Breza et al. [2023], the clustering problem gets easier as the sample size grows. Let $\mathcal{E}$ be the event that $\widehat{k}_i = k_i$ up to permutation for all $i \in \{1, 2, \ldots, n\}$, i.e. $P(\max_i |\widehat{k}_i = k_i| > 0) = 1 - P(\mathcal{E}) \leq \frac{1}{n}$. Since the estimators are not necessarily independent of the event of perfect classification.

$$\begin{aligned}
P(||\widehat{P}_{\widehat{\boldsymbol{k}}} - P_{\boldsymbol{k}}|| > \epsilon) &= P(||\widehat{P}_{\widehat{\boldsymbol{k}}} - P_{\boldsymbol{k}}|| > \epsilon|\mathcal{E})P(\mathcal{E}) + P(||\widehat{P}_{\widehat{\boldsymbol{k}}} - P_{\boldsymbol{k}}|| > \epsilon|\mathcal{E}^c)P(\mathcal{E}^c) \\
&\leq P(||\widehat{P}_{\widehat{\boldsymbol{k}}} - P_{\boldsymbol{k}}|| > \epsilon, \mathcal{E}) + P(\mathcal{E}^c) \\
&\leq P(||\widehat{P}_{\widehat{\boldsymbol{k}}} - P_{\boldsymbol{k}}|| > \epsilon, \mathcal{E}) + \frac{1}{n} \\
&= P(||\widehat{P}_{\boldsymbol{k}} - P_{\boldsymbol{k}}|| > \epsilon, \mathcal{E}) + \frac{1}{n} \text{ Since } \mathcal{E} \text{ indicates the correct classification} \\
&\leq P(||\widehat{P}_{\boldsymbol{k}} - P_{\boldsymbol{k}}|| > \epsilon) + \frac{1}{n} \\
&\leq \sqrt{\text{Tr}\left((\Omega^T\Omega)^{-1}\right)}2KT\exp\left(-2(KT)^2\epsilon^2\rho_k\rho_t n^2\right) + \frac{1}{n}
\end{aligned}$$

Therefore

$$||\widehat{P}_{\widehat{\boldsymbol{k}}} - P_{\boldsymbol{k}}|| = \mathcal{O}_P(\frac{KT\sqrt{\log(KT)}}{n})$$

□

B.3.3 Proof of Theorem 3.3.2

Proof. We emphasise that in general, the outcomes $\boldsymbol{Y}$ may be dependent, and this is reflected through correlation in the estimating functions (or the residuals in the case of OLS). We will partition the proof into two sections. First, we will prove the consistency of the estimator $\widehat{\beta}$ and secondly, we will prove the central limit theorem.

Consistency: The following result hinges on a typical consistency proof for the M or Z estimators using a structure similar to those found in Chapter 5 of Vaart [1998]. First, we denote that:

$$\begin{aligned} m_n(\boldsymbol{Z};\widehat{\beta},\widehat{\theta}) - g_n(\boldsymbol{Z};\widehat{\beta}) &= m_n(\boldsymbol{Z};\widehat{\beta},\widehat{\theta}) - m_n(\boldsymbol{Z};\widehat{\beta},\theta_0) \\ &\leq b_n(\boldsymbol{Z})||\widehat{\theta} - \theta_0|| \\ &= O_P(1)o_P(s(n)) \\ &= o_P(s(n)) \end{aligned}$$

Next, we can see that, based on this expansion,

$$\begin{aligned} & m_n(\boldsymbol{Z};\widehat{\beta},\widehat{\theta}) = 0 \\ \implies & 0 = (m_n(\boldsymbol{Z},\widehat{\beta},\widehat{\theta}) - g_n(\boldsymbol{Z},\widehat{\beta})) + g_n(\boldsymbol{Z};\widehat{\beta}) \\ & = o_P(s(n)) + g_n(\boldsymbol{Z};\widehat{\beta}) \text{ By A2} \end{aligned}$$

At this point, we can now treat this as a standard Z-estimation problem. Therefore, by A2 and A1, then $\widehat{\beta}$ is a solution to the estimating function g and is therefore consistent by an application of Theorem 5.9 of Vaart [1998].

Asymptotic Normality: We illustrate asymptotic normality through a Taylor series

expansion argument. As we saw in the consistency part of the proof

$$g_n(\boldsymbol{Z}; \widehat{\beta}) = o_P(s(n))$$

For brevity in notation, we suppress the dependence on $\boldsymbol{Z}$, which is implicit for functions, with the subscript n. Using a Taylor expansion around β_0, and let $\widetilde{\beta}_j \in [\beta_{0,j}, \widehat{\beta}_j]$ for $\beta_{0,j} \leq \widehat{\beta}_j$ and $\widetilde{\beta}_j \in [\widehat{\beta}_j, \beta_{0,j}]$ otherwise.

$$\begin{aligned} g_n(\widehat{\beta}) &= g_n(\beta_0) + D_n(\boldsymbol{Z}; \beta_0)(\widehat{\beta} - \beta_0) + \sum_{jk} \frac{\partial^2}{\partial \beta_j \beta_k} g_n(\boldsymbol{Z}; \widetilde{\beta})(\widehat{\beta}_j - \beta_{0,j})(\widehat{\beta}_k - \beta_{0,k}) \\ &= g_n(\beta_0) + D_n(\boldsymbol{Z}; \beta_0)(\widehat{\beta} - \beta_0) + o_P(s(n) + ||\widehat{\beta} - \beta_0||) \end{aligned}$$

by the application of the consistency and A2. Therefore, we focus on main terms. By Assumption C1.

Therefore:

$$\Gamma_n^{-1/2} D_n(\boldsymbol{Z}; \beta_0)(\widehat{\beta} - \beta_0) = \Gamma_n^{-1/2} g_n(\beta_0) + o_p(\frac{s(n)}{r(n)})$$

Noting that $D_n(\beta_0) - D(\beta_0) = o_P(1)$, by an application of Slutsky's lemma:

$$\Gamma_n^{-1/2} D(\beta_0)(\widehat{\beta} - \beta_0) \to_d N(0, I_p)$$

and therefore, the proof is complete. □

B.3.4 Proof of Theorem B.1.3

Proof. We first we expand the form of the OLS estimator.

$$\begin{aligned}
\widehat{\beta}_{ols} &= \mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})Y_i \\
&= \mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})(\widetilde{H}_i^T(\theta_0)\beta_0 + u_i) \\
&= \mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})(\widetilde{H}_i^T(\widehat{\theta})\beta_0 + (\widetilde{H}_i^T(\theta_0) - \widetilde{H}_i^T(\widehat{\theta}))\beta_0 + u_i) \\
&= \beta_0 + \underbrace{\mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\widehat{\theta})(\widetilde{H}_i^T(\theta_0) - \widetilde{H}_i^T(\widehat{\theta}))\beta_0}_{(A)} \\
&+ \underbrace{\mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_i(\widehat{\theta}) - \widetilde{H}_i(\theta_0))u_i}_{(B)} + \underbrace{\mathsf{H}_n^{-1}(\widehat{\theta})\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_i(\theta_0)u_i}_{(C)}
\end{aligned}$$

We next bound terms (A) and (B) after which, the asymptotic distribution of (C) will be apparent.

We note that the Hessian evaluated at the true model parameters can be evaluated $\mathsf{H}_n(\widehat{\theta}) = \mathsf{H}_n(\theta_0) + o_P(s(n))$ by assumptions D2 and D1. By the continuous mapping theorem $\mathsf{H}_n(\widehat{\theta}) = \mathsf{H}_n(\theta_0) + o_P(s(n))$. We see that $(A) = o_P(s(n))$ by assumpions D4, D2 and D1. Next, by the stochastic boundedness of the error D5 and applying Hölder's inequality.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_i(\widehat{\theta}) - \widetilde{H}_i(\theta_0)) &\leq (\frac{1}{n}\sum_{i=1}^{n}|u_i|)\max_i||\widetilde{H}_i(\widehat{\theta}) - \widetilde{H}_i(\theta_0)|| \\
&= o_P(s(n))
\end{aligned}$$

Therefore:

$$\Gamma_n^{-1/2}\mathsf{H}_n(\widehat{\theta})(\widehat{\beta}_{ols} - \beta_0) = \Gamma_n^{-1/2}\sum_{i=1}^{n}\widetilde{H}_i(\theta_0)u_i + o_P\left(\frac{s(n)}{r(n)}\right) \to_d N(0, I_p)$$

by E1 and Slutsky's Lemma, completing the proof. □

B.3.5 Proof of Lemma B.1.5

Proof. The proof follows from an application of the delta method, with the additional caveat that we must account for the estimation of the model parameters θ_0 In this case:

$$\begin{aligned}|\Psi(\widehat{\beta},\widehat{\theta}) - \Psi(\widehat{\beta},\theta_0)| &\leq \frac{1}{n}\sum_{i=1}^{n} b_i||\widehat{\theta} - \theta_0|| \\ &= o_P(s(n))\end{aligned}$$

The remainder of the proof follows from a simple application of the delta method using the plug-in estimator $\Psi(\widehat{\beta},\theta_0)$. See Theorem 3.1 of Vaart [1998]. □

B.3.6 Proof of Lemma 3.3.5

We first include a useful lemma for bounding the approximation of the error of the graphon model.

Lemma B.3.1 (Lemma 2.1 of Gao et al. [2015])**.** *Denote $k_i \in \{1, 2, \ldots, K\}$ are the block memberships of a stochastic-blockmodel with average connection probabilities across blocks $\bar{\eta}_{ij} = P_{k,k'} = \frac{1}{n_k n_{k'}} \sum_{i,j:k_i=k,k_j=Z_i} \sum_{l:Z_l=Z_j} \eta_{kl}$. If the true graphon $g \in \mathcal{H}_\alpha(M)$, then, there exists some membership vector $\boldsymbol{k}$ and corresponding average across block probabilities P_0 such that:*

$$\frac{1}{n^2}\sum_{ij}(\eta_{ij} - \bar{\eta}_{ij})^2 \leq M^2\left(\frac{1}{K^2}\right)^{\alpha\wedge 1}$$

We now proceed with a the proof of the lemma.

Proof. We firstly use a Taylor expansion of $L_n(\beta_0, \eta_*)$ where $\widetilde{\beta}$ is an element-wise intermediate value of β and $\widetilde{\beta}$

$$\begin{aligned}L_n(\beta_0,\eta_*) &= L_n(\beta_*,\eta_*) + \frac{\partial}{\partial\beta}L_n(\beta,\eta_*)\Big|_{\beta=\beta_*}(\beta_0 - \beta_*) \\ &\quad + \sum_{jk}\frac{\partial^2}{\partial\beta_j\partial\beta_k}L_n(\widetilde{\beta},\eta_*)(\beta_{0j} - \beta_{*j})(\beta_{0k} - \beta_{*k}) \\ \frac{\partial}{\partial\beta}L_n(\beta,\eta_*)\Big|_{\beta=\beta_*}(\beta_0 - \beta_*) &= -L_n(\beta_0,\eta_*) + \sum_{jk}\frac{\partial^2}{\partial\beta_j\partial\beta_k}L_n(\widetilde{\beta},\eta_*)(\beta_{0j} - \beta_{*j})(\beta_{0k} - \beta_{*k})\end{aligned}$$

Since we assume $L_n(\beta, \eta_*)$ is twice continuously differentiable in β, and $\mathcal{B}$ is compact, then $\frac{\partial^2}{\partial\beta_j \partial\beta_k} L_n(\widetilde{\beta}, \eta_*)$ is bounded. Therefore,

$$||\beta_0 - \beta_*||_2 \leq \frac{|L_n(\beta_0, \eta_*)|}{\lambda\sqrt{p}} + O(||\beta_0 - \beta_*||_2^2)$$

Lastly, by our continuity assumptions, $|L_n(\beta_0, \eta_*)| \leq L||\eta_0 - \eta_*||_2/n \leq LMK^{-(\alpha\wedge 1)}$. After applying this, our proof is complete.

□

B.4 Additional Implementation details

B.4.1 An EM algorithm for Logistic Regression

Here we elaborate on the computation of a Z estimator. In general, an estimator may require specific implementation, we provide an illustrative example with logistic regression. Recall the characterization of the average estimating function $m_i(Y_i, \boldsymbol{a}, \boldsymbol{X}; \beta, \theta) = \mathbb{E}[\widetilde{m}(Y_i, S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta)|\boldsymbol{Y}, \boldsymbol{a}, \boldsymbol{X}; \theta]$. Under this model, $P(Y_i = 1|S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G)) = \Lambda(\widetilde{h}(S_i, V_i)^T\beta)$.

In order to compute the new estimating function, we need to be able to consider the distribution of the graph, conditional on the observed outcome Y_i. Specifically.

$$\begin{aligned} P(G|Y_i, \boldsymbol{a}, \boldsymbol{X}, \beta, \theta) &= \frac{P(Y_i|G, \boldsymbol{a}, \boldsymbol{X}; \beta)P(G|\boldsymbol{a}, \boldsymbol{X}, \theta)}{P(Y_i|\boldsymbol{a}, \boldsymbol{X}, \beta, \theta)} \\ &= \frac{P(Y_i|S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta)P(G|\theta)}{P(Y_i|\boldsymbol{a}, \boldsymbol{X}, \beta, \theta)} \end{aligned}$$

In a standard missing data problem, one would impute the missing covariates directly, however, due to the dependence through the graph, this can be very difficult to achieve in practice. However, it will be straightforward to sample from the graph model $P(G|\theta)$. Using a simple approach, we can compute the maximizer exploiting standard software methods using an EM algorithm [Dempster et al., 1977, Wu, 1983]. Suppose that we draw a sample of graphs from the generative model $\{G^{(l)}\}_{l=1}^L \sim_{iid} P(G|\theta)$.

Let $w_i(Y_i, G; \beta)$ define the weight of an observation.

$$w(Y_i, G; \beta) = \frac{P(Y_i|S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta)}{P(Y_i|\boldsymbol{a}, \boldsymbol{X}, \beta, \theta)}$$
$$\approx \frac{P(Y_i|S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta)}{\frac{1}{L}\sum_{l=1}^{L} P(Y_i|S_i(\boldsymbol{X}, G^{(l)}), V_i(\boldsymbol{a}, G^{(l)}); \beta)}$$

The EM algorithm can now be defined as follows.

1. Sample $\{G^{(l)}\}_{l=1}^{L} \sim_{iid} P(G|\widehat{\theta})$ denote a sample from the graph model and initialize parameters $\widehat{\beta}^{(0)}$

2. For $t \in \{1, 2, \ldots, T\}$

   (a) (E-step) Compute the weighted empirical estimating function

   $$m_n^{(t)}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{X}, \beta, \widehat{\theta}) = \frac{1}{L}\frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n} \widetilde{m}(Y_i, S_i(\boldsymbol{X}, G^{(l)}), V_i(\boldsymbol{a}, G^{(l)}); \beta) w(Y_i, G^{(l)}; \widehat{\beta}^{(t-1)})$$

   (b) (M-step) Solve the new estimating function by solving:

   $$m_n^{(t)}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{X}, \widehat{\beta}^{(t)}, \widehat{\theta}) = 0$$

In practice, this allows for one to use standard solvers for the (M-step), after sampling a single time with the (E-step).

Additionally, one can include correlations across the observations Y_i through the use of a generalized estimating equation approach. In other generalized linear models, additional assumptions may be required in order to model the full conditional distribution $P(Y_i|S_i(\boldsymbol{X}, G), V_i(\boldsymbol{a}, G); \beta)$ such as a dispersion component.

B.4.2 Optimal design for a Z-estimator

Here we illustrate the optimal design approach for Z-estimators. In this example, the variance itself may depend on the a parameter β, and thus one can include a working candidate for the parameter β'.

Algorithm 14 Saturation Randomized Design Variance.

Inputs: Working covariance Γ_n, model estimate $\widehat{\theta}$, working parameter β'

Sample L draws from the graph model $\{\widehat{G}^{(l)}\}_{l=1}^{L} \sim \widehat{\theta}$

Sample R treatments $\{\boldsymbol{a}_r\}_{r=1}^{R}$ according to the block-saturations $\boldsymbol{\tau}$.

for $r \leftarrow 1$ **to** R **do**

  Compute $\widehat{D}_r(\boldsymbol{a}) = \frac{1}{nL}\sum_{l=1}^{L}\sum_{i=1}^{n} \nabla_\beta m_i(Y_i, S_i V_i; \widehat{G}^{(l)}, \beta')$

  Compute the variance for a single draw $\boldsymbol{a}_r$:

$$\upsilon^{\phi}(\boldsymbol{a}_r; \widehat{\theta}) = \phi^T \widehat{D}_r(\boldsymbol{a})^{-1} \Gamma_n \widehat{D}_r(\boldsymbol{a})^{-1T} \phi$$

end for

Average over each of the draws $\bar{\upsilon}(\boldsymbol{\tau}; \widehat{\theta}) = \sum_{r=1}^{R} \upsilon^{\phi}(\boldsymbol{a}_r; \widehat{\theta})$

B.5 Additional Simulations

B.5.1 Coverage of the GATE

In our simulation setup in Section 3.5.1 we can also compute confidence intervals based on the regression $Y_i = \beta^T \mathbb{E}[\widetilde{h}(S_i, V_i)] + \epsilon_i$ where we apply the Eicker-Huber-White sandwich estimator of the variance. We then compute the corresponding plug-in estimator of the variance using the covariates observed and Lemma B.1.5. Since the covariates in the true regression model behave like averages over the graph, we expect Lemma B.1.7 to hold and therefore the difference between the GATE for any one draw of the graph, and the true GATE is very small. We see in Figure B.2 that the coverage tends to be larger than the nominal 95%, though in general, due to model misspecification of the true-graph, there can be additional uncertainty due to the misspecification of the graph model. However, we see in this simple example that the coverage performs well with an off-the-shelf implementation.

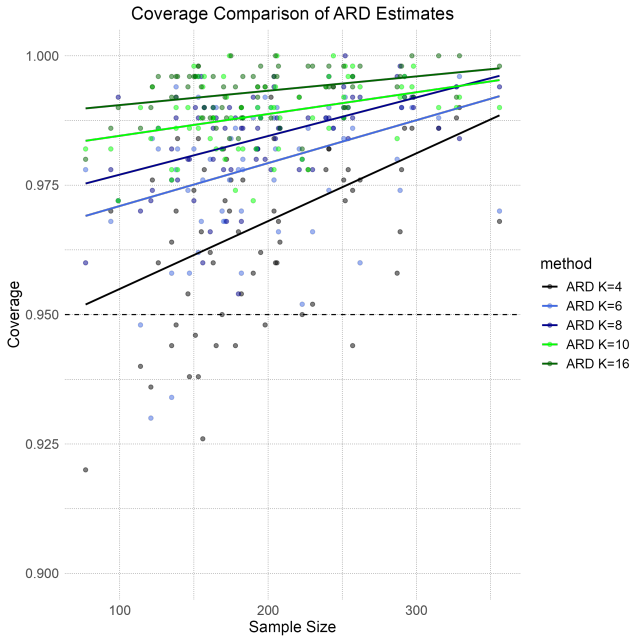


Figure B.2: Coverage of the GATE using Eicker-Huber-White estimates of the variance.

B.5.2 Experimental Design: Local Diffusion

We next consider an example using a local diffusion process. We suppose that seed nodes are placed at time 0 and that outcomes are measured at time $T = 1$, allowing for diffusion to only take place to the immediate neighbors with a fixed probability q. In this case, for non-seed nodes the probability of infection is related to the total number of treated neighbors through the following link function. Under this model let $V_i \in \{0, 1\}$ denote the exposure as to whether one of their neighbors have received the treatment, i.e. $V_i = I(\sum_j G_{ij} a_j > 0)$. Then

$$\mathbb{E}[Y_i | V_i, S_i] = qV_i$$

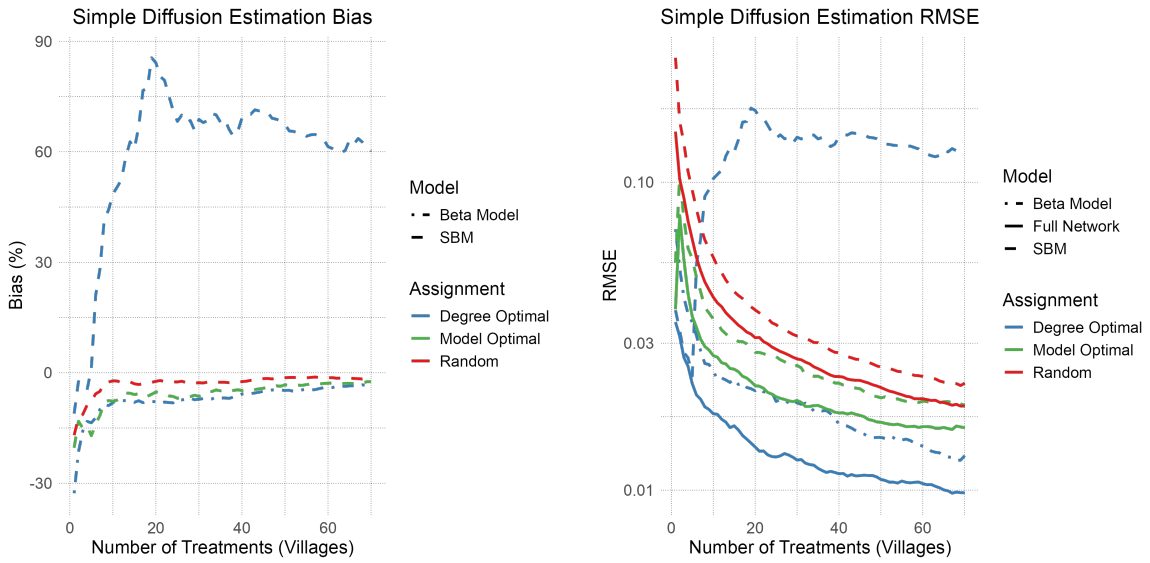
In this experiment, a single individual is seeded in each network. Our goal is to identify the best individuals in each of the network to seed and rank them by the expected variance of the estimator. We compare this to random seeding of individuals in the network as well as

seeding by only the highest degree nodes. We use the networks constructed by the union of all connections of Banerjee et al. [2019]. We construct estimates of the stochastic blockmodel as the partial data example using $K = 3$ in each case. We construct the traits using ARD responses based on number of connections with the following traits outlined in the Appendix in Section B.6.2. We also include an alternative where a beta-model [Chatterjee and Diaconis, 2011] is used in place of the SBM for the degree seeding where further details on estimation are included in Section B.6.1. We then draw samples of the graph using the parametric bootstrap to obtain a resampled distribution of ARD $\{\boldsymbol{X}^{*(b)}\}_{b=1}^{B}$ for $B = 1000$. We identify the optimal treatment block for each parameter according to Section B.1.5. We simulate 1000 draws of the draws in the diffusion process for each true, and plot the associated bias and RMSE of the seeding strategies in Figure B.3 with a true diffusion parameter $q = 0.2$.

In the full data case, the optimal strategy would be to seed the highest degree node in each of the networks and measure whether each of their neighbors are infected at time $T = 1$. However, this poses a problem for the stochastic blockmodel as we are essentially picking an outlier to seed, which is different than a typical member of the block over draws of the process. This can be corrected for using a model which accounts for degree heterogeneity, in our case, the beta model. In our optimal seeding strategy, we find that the RMSE is lower in both the degree optimized strategy with the beta model, as well as the block optimized strategy with the SBM, than even the full data version with a completely randomized allocation, hence highlighting the role of the interplay of the model of the graph and the experimental design. This behavior is observed in Figure B.3.

B.5.3 Estimated outcome model

In this example, we consider a problem of optimal treatment assignment after the outcome model is estimated. We consider an example where an outcome model is estimated and transported to a new population. In this example we suppose that there is some benefit $\beta_1 > 0$ to receiving a treatment, and some smaller benefit based on the fraction of the neighbors treated $0 < \beta_2 < \beta_1$. We wish to assign treatments in a way that will maximize



(a) Bias of Full and Partial Data Diffusion Parameter Estimates

(b) RMSE of Full and Partial Data Diffusion Parameter Estimates

Figure B.3: RMSE and bias of estimating parameter q using random seeding, and the optimal seed for each village.

the expected outcome $\Psi(\boldsymbol{a}|G)$ for each network.

$$Y_i = \beta_0 + \beta_1 a_i + \beta_2 q_1 + \epsilon_i$$

Where $q_i := \frac{1}{d_i}\sum_{j=1}^{n} G_{ij}a_j$ denotes the normalized number of treated neighbors. We simulate the data with $\beta_0 = 1$, $\beta_1 = 1$ and $\beta_2 = 1/2$ with $\sigma_i \sim N(0,1)$. We choose this form of a response function since it will be simple to solve with an off the shelf mixed-integer programming approach using CVXR [Fu et al., 2020].

We suppose that in each example there is only a budget for $B \in \{10, 20, 40, 80\}$ treatments for each of the villages. The goal is to maximize the overall expected outcome. We consider the following competing procedures. In this case, we suppose that we have a single pilot network where we can learn the model and the goal is to maximize the benefit on the remaining networks. We use the same gossip diffusion networks as in sections B.5.2 and 3.5.2.

We compare the following seeding strategies.

1. Random assignment to all individuals in the network

2. Equal assignment amongst clusters.

3. Assign treatments ordered by the highest degree of the nodes.

4. Maximize the total expected outcome by maximizing $\max_{\boldsymbol{a};||a||_1 \leq B} \Psi(\boldsymbol{a}; \widehat{\beta}, \widehat{\theta})$

Let $\mathbb{E}[Y_i|\boldsymbol{a}] = \beta_0 + \beta_1 a_i + \beta_2(1-a_i)\sum_{k'=1}^{K} \widehat{P}_{\widehat{k}_i k'} n_{t,k}$ and let $n_{t,k} = \sum_{j:k_j=k} a_j$. Therefore, the objective function.

$$\Psi(\boldsymbol{a}|\beta,\theta) = \beta_0 + \frac{1}{n}\beta_1 \mathbf{1}^T \boldsymbol{n}_t + \frac{1}{n}\beta_2 \zeta^T \boldsymbol{n}_t$$

where $\zeta = \frac{1}{d_i}\sum_{i=1}^{n} \mathbf{P}_{k_i,\cdot}$ and $\boldsymbol{n}_t = (n_{t,1}, n_{t,2}, \ldots, n_{t,K})$. In general, given a conditional model, one may fine tune the optimization approach to the particular challenges of evaluating the optimal treatment allocation. We partition each network into 6 blocks.

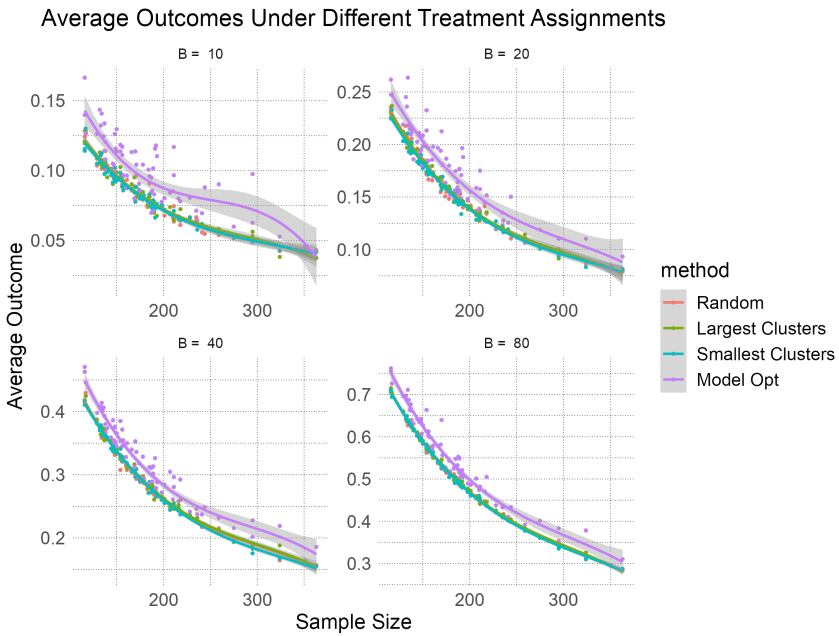


Figure B.4: Our method, model based optimal treatment allocation (Model Opt) compared to random assignment and assignment to largest and smallest clusters respectively. The larger the values represent larger average outcomes in each of the networks. Curves are fit using cubic splines. The model based optimal design tends to give a higher value at each of the sample sizes at each treatment budget. For example, at a sample size of 150 and a treatment budget of 10, our methods leads to a 30% increase in the average outcome.

We plot the expected average outcome under each of the treatment allocations for the remaining 68 networks after learning a model from the first pilot network. We repeat this for the total number of treatments $B \in \{10, 20, 40, 80\}$.

In Figure B.4 we find that based on our method, we can achieve higher average outcomes than simple models based on the block positioning alone, emphasizing the importance of considering the potential outcome model when optimal targeting.

B.5.4 Inference for evidence of complex contagion with partial network data

We can also replicate the results of Beaman et al. [2021]'s study on the evidence of pitplanting. They consider 3 measures of information diffusion. Firstly, if an individual has heard of pitplanting, second, if they know how to pitplant, and thirdly whether they adopt pitplanting in their practice. In order to control for one's position in the network, the authors consider the distance between the optimal seeds using two other targeting methods, simple diffusion, and geo-targeting as well as complex contagion. They then compare the increased odds of con

$$Y_{iv} = \alpha + \beta_1 I(1TSeeds) + \beta_2 I(2TSeeds) + \beta_3 I(1Simple)_{iv} + \beta_4 I(2Simple)_{iv} + \beta_5 I(1Complex)_{iv} + \beta_6 I(2Complex)_{iv} + \beta_7 I(1Geo)_{iv} + \beta_8 I(2Geo)_{iv} + \delta_v + \epsilon_{iv}$$

Again, we generate synthetic covariates and apply a stochastic blockmodel in order to estimate $K = 8$ blocks within each of the networks. We plot the coefficients for the connection to exactly 1 seed, 2 seeds and within radius 2 of at least 1 seed in Figure B.5. We note that we run the same regression as in Beaman et al. [2021], however, some since the full network data includes some additional noise top preserve anonymity, we do not have the exact same estimates of the coefficients as in their paper, however, the conclusions are substantively the same.

B.6 Additional Experimental Details

To aid in reputability, we include additional details regarding the implementation of our methods as well as competing methods.

B.6.1 Beta Model Estimation

Another common model utilized for random graph formation is the beta model coined by Chatterjee et al. [2011]. Namely these are a class of models that can be learned based on

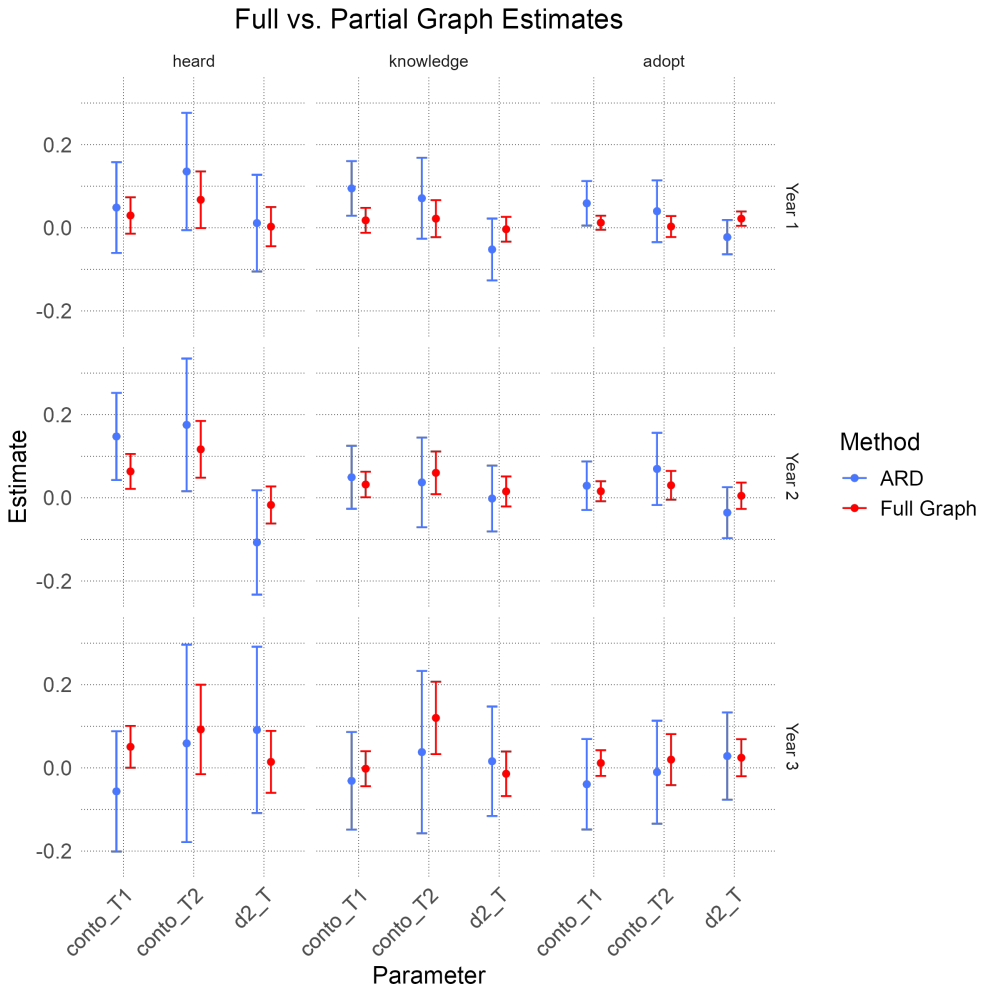


Figure B.5: Replication of regression coefficients using aggregated relational data and associated 95% confidence intervals.

their degree sequence. We consider a version where each node has an affinity parameter ν_i and the probability of connection between each pair of nodes is $P(G_{ij} = 1) = \nu_i \nu_j$. Let $\nu_n = \sum_{i=1}^{n} \nu_i$ Therefore, $\mathbb{E}[d_i = d] = \sum_{j \neq i} P(G_{ij} = 1) = \nu_i(\nu_n - \nu_i)$. The set of parameters $\{\nu_i\}_{i=1}^{n}$ can be estimated using an iterative solution to the fixed point equation:

$$\nu_i^{(t+1)} = d_i / (\nu_n^{(t)} - \nu_i^{(t)})$$

B.6.2 ARD Questions

We utilize the measured traits to construct responses for ARD questions for each individual for the networks in Banerjee et al. [2019]. The constructed ARD include traits which ask "How many people do you know ..."

- that are in each sub-caste?
- that are Farmers, Shop owners, Domestic workers etc. ?
- that own their house?
- that have a house with at least 3 rooms?
- that have access to electricity?

For the estimation of the GATE using Banerjee et al. [2013b], we use Leiden clustering and denote the clusters traits. When replicating the results of Beaman et al. [2021], only a subset of nodes have available covariate. As was done in our examples with Banerjee et al. [2013b], we construct synthetic traits using the clusters observed from Leiden clustering for $K = 10$. ARD is then constructed based on the connections to nodes of each trait.

B.6.3 GATE Estimators

The two estimators we compare for estimation of the global average treatment effect are the difference in means estimator $\widehat{\tau}_{DM}$ and the Horvitz-Thompson estimator $\widehat{\tau}_{HT}$. Let E_{i0} and

E_{i1} denote the events that all neighbours of i are untreated (including i themselves) and treated respectively.

$$\widehat{\tau}_{DM} = \frac{1}{n_1}\sum_{i=1}^{n} Y_i a_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-a_i)$$
$$\widehat{\tau}_{HT} = \frac{1}{n}\sum_{i=1} \frac{Y_i I(E_{i1})}{P(E_{i1})} - \frac{Y_i I(E_{i0})}{P(E_{i0})}$$

In general, the Horvitz-Thompson estimator will be unbiased, however, it can often suffer from high variance for two reasons. Firstly, the probabilities of the events that all nodes are treated may be exceedingly low, inflating this variance, and also, relatively few nodes receive the exposures under which all of their neighbours are treated or none of them are.

In the case where the spillover effects are relatively mild, often a difference in means approach to the estimator is preferred. The effect of cluster randomization on the MSE of this estimator has been further studied in the complete networkBrennan et al. [2022], Viviano [2020].

Appendix C

SUPPLEMENTARY MATERIAL FOR PROJECT 3

C.1 Proofs of Theorems

C.1.1 Proof of Lemma 4.2.2

In order to prove this lemma, we first introduce a useful result.

Theorem C.1.1 (Theorem 1.10.15 in Klingenberg [1995]). *If $f : (\bar{M}, \bar{g}) \to (\bar{M}, \bar{g})$ is an isometry on a Riemannian manifold, then the fixed point set of f forms a totally geodesic submanifold.*

Using this theorem, if we can construct an isometry, a bijective isomorphism between two metric spaces (in this case, the submanifold within the canonical manifold which contains the triangle, as well as the canonical manifold of dimension 2), then the fixed points of the set will form a totally geodesic submanifold which will be useful for constructing our fixed point equation.

Proof. We will prove this by first considering constructing a change of basis to parameterize the sub-manifold using the first 3 coordinates, then we will construct the isometry between the manifold of dimension $p \geq 2$ and that of dimension 2.

Consider a set of 3 points $(x, y, z) \in \mathbb{S}^p(\kappa)$ which are not co-linear. We can construct an orthogonal matrix (rotation matrix) $Q \in \mathbb{R}^{p+1 \times p+1}$ which allows us to construct a rotational isometry.

WLOG, in our coordinate system, we place x at the origin, i.e. $x = (1, 0, \ldots, 0)$. We construct an orthogonal rotation matrix Q which rotates the coordinates into the first 3 indices, with the rest of them being 0. Let $\widetilde{q}_i$ denote the i^{th} un-normalized column vector of

Q, and $q_i = \widetilde{q}_i/||\widetilde{q}_i||_2$. We the define the following first 3 basis functions as the normalized projections of the components of y and then z:

$$\begin{aligned}
\widetilde{q}_1 &= [1, 0, \ldots, 0]^\top \\
\widetilde{q}_2 &= y - (y^\top q_1)q_1 \\
\widetilde{q}_3 &= z - (z^\top q_2)q_2 - (z^\top q_1)q_1.
\end{aligned}$$

Since x, y, z are not colinear, then q_1, q_2, q_3 are independent and are the normalization's of $\widetilde{q}_1, \widetilde{q}_2, \widetilde{q}_3$ respectively and are orthogonal. The remaining columns of Q are the completion of the orthonormal basis with any remaining orthogonal basis vectors of $\mathbb{R}^p$. Thus we have constructed an orthogonal matrix Q and $Q^\top Q = I = QQ^\top$. This matrix Q can be used to construct an isometry $f_1(\cdot) : \mathbb{S}^p \mapsto \mathbb{S}^p$ where

$$\begin{aligned}
f_1(x) &= Q^\top x \\
d(f_1(x), f_1(y)) &= \frac{1}{\sqrt{\kappa}}\text{acos}\left(f_1(x)^\top f_1(y)\right) \\
&= \frac{1}{\sqrt{\kappa}}\text{acos}\left(x^\top QQ^\top y\right) \\
&= \frac{1}{\sqrt{\kappa}}\text{acos}\left(x^\top y\right) \\
&= d(x, y).
\end{aligned}$$

Hence, this rotation is an isometry. Therefore, we can equivalently use the parameterization where the points x, y, z have non-zero coordinates only in the first three indices.

Next, let $f_2(x) = (x_0, x_1, x_2, -x_3, -x_4, \ldots, -x_p)$ denote an second mapping. Under this transformation

$$\begin{aligned}
f_2(x)^\top f_2(y) &= \sum_{i=0}^{2}(x_i)(y_i) + \sum_{i=3}^{p}(-x_i)(-y_i) \\
&= x^T y \\
\implies d(f_2(x), f_2(y)) &= d(x, y)
\end{aligned}$$

and thus f_2 is also an isometry. Since the composition of isometries is also an isometry. By the composition of this rotation and sign flip of coordinates, we can construct the

corresponding totally geodesic submanifold $\mathcal{M}^2(\kappa)$ as follows. Lastly the totally geodesic submanifold of dimension 2 can be constructed using the set of points satisfying $\{Qv\}$ for $v = [v_0, v_1, v_2, 0, 0, \ldots, 0]^\top$ for any v_0, v_1, v_2 such that $v_0^2 + v_1^2 + v_2^2 = 1$, which is simply a reparameterization of $\mathbb{S}^2(\kappa)$ mapped into $\mathbb{S}^p(\kappa)$. By construction of this rotation matrix, x, y, z are all fixed points of this isometry $(f_2 \circ f_1)$. When translating between coordinate positions and the distances, one must use the same curvature value κ, and thus the totally geodesic manifold of dimension 2 exists with the same curvature, and thus contains its midpoints.

The proofs for $\mathbb{E}^p$ and $\mathbb{H}^p$ follows this argument identically and thus the proof is complete.

□

C.1.2 Proof of Theorem 4.2.3

Proof. In order to develop our identifying equation for the curvature using triangle distances, we note that any three points (x, y, z) can be embedded isometrically in a totally geodesic submanifold of dimension 2 when $\mathcal{M}^p(\kappa)$ is a canonical manifold, as stated in Lemma 4.2.2. Since this is a totally geodesic submanifold, the distance to the midpoint parameterized by coordinates in $\mathcal{M}^2(\kappa)$ will be the same as in $\mathcal{M}^p(\kappa)$. We continue with the proof by embedding the triangle xyz in a canonical manifold of dimension 2. This will provide an implicit equation for the curvature, determined by the side lengths of the triangle as well as the length of the triangle median. In each case, we use the representations of the canonical manifolds outlined in Section 4.2.1.

Case 1: Spherical. For convenience of derivation, we derive an implicit equation in a coordinate system where the midpoint of points y and z is placed at the origin ($m = (1, 0, 0)$ in $\mathbb{S}^2$) and the line between y and z define the axis $(0, 1, 0)$. Given the distance d_{yz} we place point y at the point $(y_0, y_1, 0)$ where $y_0 = \cos(\sqrt{\kappa}\frac{d_{yz}}{2})$. Next we place point z at $z = (z_0, z_1, 0)$ where $z_0 = y_0$ and $z_1 = -y_1$. Given this parameterization of the manifold embedding of y, z and the distances, d_{xy}, d_{xz}, we solve for the coordinates of x.

In general, $x = (x_0, x_1, x_2)$. Since the midpoint is placed at $(1, 0, 0)$, then $x_0 =$

$\cos(\sqrt{\kappa}\frac{d_{xm}}{2})$. Next, x_1 can be solved for by considering its distance to y:

$$\begin{aligned}\cos(\sqrt{\kappa}d_{xy}) &= x_0y_0 + x_1y_1 \\ &= \cos(\sqrt{\kappa}d_{xm})\cos(\sqrt{\kappa}d_{yz}/2) + x_1\sin(\sqrt{\kappa}d_{yz}/2).\end{aligned}$$

Similarly the distance to z can be computed as:

$$\begin{aligned}\cos(\sqrt{\kappa}d_{xz}) &= x_0z_0 + x_1z_1 \\ &= \cos(\sqrt{\kappa}d_{xm})\cos(\sqrt{\kappa}d_{yz}/2) - x_1\sin(\sqrt{\kappa}d_{yz}/2).\end{aligned}$$

Solving for x_1 leads to the expression

$$\frac{\cos(\sqrt{\kappa}d_{xy}) - \cos(\sqrt{\kappa}d_{xm})\cos(\sqrt{\kappa}d_{yz}/2)}{\cos(\sqrt{\kappa}d_{xz}) - \cos(\sqrt{\kappa}d_{xm})\cos(\sqrt{\kappa}d_{yz}/2)} = -1$$

which can be further rearranged as follows:

$$\begin{aligned}\cos(\sqrt{\kappa}d_{xy}) - \cos(\sqrt{\kappa}d_{xm})\cos(\sqrt{\kappa}d_{yz}/2) &= -(\cos(\sqrt{\kappa}d_{xz}) - \cos(\sqrt{\kappa}d_{xm})\cos(\sqrt{\kappa}d_{yz}/2)) \\ \implies \cos(\sqrt{\kappa}d_{xy}) + \cos(\sqrt{\kappa}d_{xz}) &= 2\cos(\sqrt{\kappa}d_{xm})\cos(\sqrt{\kappa}d_{yz}/2) \\ \implies \mathrm{Sec}(\frac{d_{yz}}{2}\sqrt{\kappa})(\cos(\sqrt{\kappa}d_{xy}) + \cos(\sqrt{\kappa}d_{xz})) &= 2\cos(\sqrt{\kappa}d_{xm}).\end{aligned}$$

This finally leads to the expression:

$$2\cos(d_{xm}\sqrt{\kappa}) - \mathrm{Sec}(\frac{d_{yz}}{2}\sqrt{\kappa})(\cos(d_{xy}\sqrt{\kappa}) + \cos(d_{xz}\sqrt{\kappa})) = 0.$$

We then normalize this by the curvature value $\frac{1}{\kappa}$, which allows for $g(\kappa, d)$ to be a continuous function of κ from the hyperbolic $\kappa < 0$ to spherical space $\kappa > 0$.

In the spherical case, we also require that the distances themselves satisfy $d_{pq} \leq \frac{\pi}{\sqrt{\kappa}}$, due to the fact that this corresponds to the maximum possible distance on the sphere, however this restriction is not present in the $\mathbb{E}^p$ and $\mathbb{H}^p$.

Case 2: Hyperbolic. The proof for deriving this method in the hyperbolic case follows an identical method to the spherical case and is left out for brevity. The resultant estimating

function is of the following form

$$\frac{1}{-\kappa}\Bigg(2\cosh(d_{xm}\sqrt{-\kappa}) \\ -\operatorname{Sech}(\frac{d_{yz}}{2}\sqrt{-\kappa})(\cosh(d_{xy}\sqrt{-\kappa})+\cosh(d_{xz}\sqrt{-\kappa}))\Bigg)=0$$

Remark: Under the limit as $\kappa \to 0$ we find that $\lim_{\kappa\to 0} g(\kappa, \boldsymbol{d}^{\triangle}) = \frac{1}{2}d_{xy}^2 + \frac{1}{2}d_{xz}^2 - \frac{1}{4}d_{yz}^2 - d_{xm}^2$ which gives exactly the parallelogram law in Euclidean space. This also highlights the necessity that the term $\frac{1}{\kappa}$ plays in maintaining a smooth equation as a function of κ through 0. □

C.1.3 Proof of Theorem 4.2.4

Proof. We will use the implicit function theorem to construct a function for which we can later apply a delta method. In order to do so, we must ensure that $\frac{d}{d\kappa}g(\kappa, \boldsymbol{d}^{\triangle}) \neq 0$ for κ such that $g(\kappa, \boldsymbol{d}^{\triangle}) = 0$. We first note that $g(\kappa, \boldsymbol{d}^{\triangle})$ is continuously differentiable when $\sqrt{\kappa} < \frac{\operatorname{acos}(\pi)}{d_{yz}}$. This boundary corresponds to the maximum distance allowed on a sphere, as given by two anti-polar points. Therefore to apply the implicit function theorem, what remains is

$$\frac{\partial}{\partial\kappa}g(\kappa, \boldsymbol{d}^{\triangle}) \neq 0.$$

which will hold by a brief application of (B3). As we have derived in Theorem 4.2.3, $d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z))$, the length of the triangle median as a function of the triangle lengths $\boldsymbol{d}^{\triangle}(x, y, z)$ is a differentiable, continuous, increasing function of κ.

By definition, this function satisfies

$$g(\kappa, d_{xy}, d_{xz}, d_{yz}, d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z))) = 0$$

Let d_{xm} denote the fixed value of the triangle median length at value of the true curvature κ, and let $d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z)))$ denote the median length as a function of the triangle side length. We can now write the estimating function g as a function of the equivalent exact

midpoint $d_{xm}(\kappa) := d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x,y,z))$[1]

$$g(\kappa, d_{xy}, d_{xz}, d_{yz}, d_{xm}) = \frac{2\cos(d_{xm}\sqrt{\kappa})}{\kappa} - \frac{2\cos(d_{xm}(\kappa)\sqrt{\kappa})}{\kappa}$$

We compute the derivative as a function of κ

$$\begin{aligned}
&\frac{d}{d\kappa'} g(\kappa'; d_{xy}, d_{xz}, d_{yz}, d_{xm})\Big|_{\kappa'=\kappa} \\
&= 2\frac{-d_{xm}\sqrt{\kappa}\sin(d_{xm}\sqrt{\kappa}) + 2\cos(d_{xm}\sqrt{\kappa})}{2\kappa^2} \\
&+ 2\frac{(d_{xm}(\kappa)\sqrt{\kappa} + \kappa^{3/2} d'_{xm}(\kappa))\sin(d_{xm}\sqrt{\kappa}) - 2\cos(d_{xm}(\kappa)\sqrt{\kappa})}{2\kappa^2}
\end{aligned}$$

Since $g(\kappa, \boldsymbol{d}^{\triangle}) = 0 \implies d_{xm} = d_{xm}(\kappa)$,

$$\begin{aligned}
&\frac{d}{d\kappa'} g(\kappa'; d_{xy}, d_{xz}, d_{yz}, d_{xm})\Big|_{\kappa'=\kappa} \\
&= 2d'_{xm}(\kappa)\frac{\sin(d_{xm}(\kappa)\sqrt{\kappa})}{\sqrt{\kappa}}
\end{aligned}$$

Since $\kappa \leq \frac{\pi^2}{d_{xm}^2}$ then if $d'_{xm}(\kappa) > 0$ then $\frac{d}{d\kappa} g(\kappa, \boldsymbol{d}^{\triangle}) > 0$ and hence $g(\kappa, \boldsymbol{d}^{\triangle})$ has positive derivative $\frac{d}{d\kappa} g(\kappa, \boldsymbol{d}^{\triangle}) > 0$ for $\kappa : g(\kappa; \boldsymbol{d}^{\triangle}) = 0$, where we use the notation $g(\kappa; \boldsymbol{d}^{\triangle})$ when we are specifying the g as a function of κ for a fixed set of distances.

Next, by the implicit function theorem, there exists an open neighborhood $\widetilde{\mathcal{D}} \in \mathbb{R}^4$ and an "implicit" function $\widetilde{f}(d)$ such that: $\kappa = \widetilde{f}(d)$ and $g(\widetilde{f}(d), d) = 0$ for $d \in \widetilde{\mathcal{D}}$. Furthermore the gradient satisfies:

$$\nabla_d \widetilde{f}(d) = -\left(\frac{\partial g(\kappa, \boldsymbol{d}^{\triangle})}{\partial \kappa}\right)^{-1} \left[\nabla_d g(\kappa, \boldsymbol{d}^{\triangle})\right].$$

Therefore by the delta method we arrive at the asymptotic distribution of $\widehat{\kappa}$

$$\sqrt{r(n)}(\widehat{\kappa} - \kappa) = \sqrt{r(n)}(\widetilde{f}(\widehat{d}) - \widetilde{f}(d)) \to_D N(0, \nabla\widetilde{f}(d)^\top \Sigma \nabla \widetilde{f}(d))$$

---

[1] We drop triangle lengths dependence $\boldsymbol{d}^{\triangle}(x,y,z)$ for brevity

where $\to_D$ refers to convergence in distribution. Lastly, by the implicit function theorem:

$$\nabla \widetilde{f}(d)^\top \Sigma \nabla \widetilde{f}(d) = \left(\frac{\partial g(\kappa, \boldsymbol{d}^{\triangle})}{\partial \kappa} \right)^{-2} \left[\nabla_d g(\kappa, \boldsymbol{d}^{\triangle})^\top \Sigma \nabla_d g(\kappa, \boldsymbol{d}^{\triangle}) \right].$$

If instead we only have consistency, rather than asymptotic normality of $\widehat{d}$, $||\widehat{d} - d||_2 = o_P(r(n)^{-1/2})$, then we can use the continuous mapping theorem instead and we have $|\widehat{\kappa} - \kappa| = o_P(r(n)^{-1/2})$. **Remark:** This form is very similar to the usual standard asymptotic normality proofs for Z estimators as in van der Vaart [1998]. However, the main difference is in the fact that we are not averaging the estimating function, but rather plugging in a distance estimate which is asymptotically normal, meaning that we develop this delta method argument for a plug-in estimator.

We find that in practice, condition (B3) holds quite generally, unless the points x, y, z are co linear. In fact, in the Euclidean case, we can derive this exactly according to the Taylor series expansion at $\kappa = 0$

$$g(\kappa, \boldsymbol{d}^{\triangle}) = \left(\frac{d_{xy}^2}{2} + \frac{d_{xz}^2}{2} - d_{xm}^2 - \frac{d_{yz}^2}{4} \right) + \left(\frac{d_{xm}^2}{12} - \frac{5d_{yz}^2}{192} + \frac{1}{16} d_{yz}^2 (d_{xy}^2 + d_{xz}^2) - \frac{1}{24} (d_{xy}^2 + d_{xz}^2) \right) \kappa + \mathcal{O}(\kappa^2).$$

Clearly at $\kappa = 0$ the solution reduces to the parallelogram law ($\frac{d_{xy}^2}{2} + \frac{d_{xz}^2}{2} - d_{xm}^2 - \frac{d_{yz}^2}{4} = 0$). When we substitute in the corresponding solution at $\kappa = 0$

$$\begin{aligned} \frac{\partial}{\partial \kappa} g(\kappa, \boldsymbol{d}^{\triangle}) \Big|_{\kappa=0} &= -\frac{1}{48} \left(d_{yz}^4 - 2d_{yz}^2 (d_{xz}^2 + (d_{xy}^2) + (d_{xz}^2 - d_{xy}^2)^2 \right) \\ &= -\frac{1}{48} \Big((d_{yz} + d_{xz} + d_{xy})(d_{yz} - d_{xz} - d_{xy}) \\ &\quad \times (d_{yz} - d_{xz} + d_{xy})(d_{yz} + d_{xz} - d_{xy}) \Big). \end{aligned}$$

As long as the triangle inequality is satisfied strictly for x, y, z, then $\frac{\partial}{\partial \kappa} g(\kappa; \boldsymbol{d}^{\triangle}) \Big|_{\kappa=0} > 0$. However, if the triangle inequality is not strict, i.e. x, y, z are co-linear, and $d_{ij} = d_{ik} + d_{kj}$ for $(i, j.k)$ being some permutation of (x, y, z), then $\frac{\partial}{\partial \kappa} g(\kappa; \boldsymbol{d}^{\triangle}) \Big|_{\kappa=0} = 0$. For $\kappa \neq 0$, we do

not have simple closed-form expressions for $\frac{\partial}{\partial\kappa} g(\kappa; \boldsymbol{d}^{\triangle})\Big|_{\kappa=0}$ and thus we leave (B3) as an assumption, however, we believe this pattern to hold in the other canonical manifolds.

□

Lemma C.1.2. *If $x, y, z \in \mathcal{M}^p(\kappa)$, let m denote the midpoint between y and z. Then let $d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z))$ denote the distance to the midpoint m from x as a function of the curvature of the latent space κ.*

Then $d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z))$ is a non-decreasing function in κ.

Proof. The proof follows from Toponogov's triangle theorem and the negative curvature extension [Berger, 1962] which we include immediately following this proof in Theorem C.1.3. Abbreviated for our purposes in Theorem C.1.3, we simply let $p = z, x = r, q = m$ and compare two manifolds $\kappa \leq \kappa'$.

Then it immediately follows

$$d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z)) \leq d_{xm}(\kappa'; \boldsymbol{d}^{\triangle}(x, y, z)).$$

□

Theorem C.1.3 (Toponogov's Triangle Comparison Theorem (Theorem 3 of Berger [1962])**.** *Let $\mathcal{M}$ be a complete Riemannian manifold, and let $\mathcal{M}(\kappa)$ be the simply connected manifold of constant curvature $\kappa \in \mathbb{R}$. Let $\triangle_{zyx}$ denote a geodesic triangle with points $z, y, x \in \mathcal{M}$. Let $\widetilde{\triangle}_{zyx}$ is the geodesic triangle with side lengths $d_{zy} = \widetilde{d}_{zy}$ and $d_{zx} = \widetilde{d}_{yx}$ on $\mathcal{M}(\kappa)$. If $\underline{K}(\mathcal{M})$ is the minimum Riemannian sectional curvature on the manifold, and if $\kappa \leq \underline{K}(\mathcal{M})$, then*

$$\widetilde{d}_{yx} \leq d_{qr}.$$

This theorem suggests that assumption (B3) is therefore very mild and that we will always have a non-decreasing d_{xm} function of κ but that we only have to assume that the increasingness is strict. In practice, we observe this is only not strict when the points x, y, z

are co-linear, and therefore the distance to the midpoint does not change as a function of the curvature since, effectively, these points lie along a single geodesic.

C.1.4 Proof of Theorem 4.2.5

Proof. Recall that $\boldsymbol{d}^{\triangle}(x, y, z)$ denotes the vector of the distances of the triangle with vertices (x, y, z), and let $\boldsymbol{d}^{\triangle}(m', y, z)$ be the distances in triangle (m', y, z). We also measure the distance to the surrogate midpoint, $d_{xm'}$. Though we do not have access to the distance to the true midpoint d_{xm}, this is going to be a function of the curvature of the space, and the distances of the triangle (x, y, z) and the median length $d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z))$. For any curvature κ and a surrogate midpoint m', so that $m', x, y, z, \in \mathcal{M}^p(\kappa)$, we can upper and lower bound the triangle median length. By the triangle inequality these upper and lower bounds are

$$d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z)) \leq d^{+}_{xm}(\kappa) := d_{xm'} + d_{m'm}(\kappa; \boldsymbol{d}(m', y, z))$$
$$d_{xm}(\kappa; \boldsymbol{d}^{\triangle}(x, y, z)) \geq d^{-}_{xm}(\kappa) := d_{xm'} - d_{m'm}(\kappa; \boldsymbol{d}^{\triangle}(m', y, z))$$

By Lemma C.1.2 each d_{xm} function is monotone in κ. The upper and lower bounds $\kappa^{\pm}$ can then be computed by plugging in this value to g, i.e. $g(\kappa, \boldsymbol{d}^{\triangle,\pm}(\kappa)) = 0$ and solving for κ. By this monotonicity, the number of solutions in κ to either equation will typitcally be $\{0, 1\}$.

If there is a solution to $g(\kappa, \boldsymbol{d}^{\triangle,\pm}(\kappa)) = 0$, then we have found and upper and lower bounds. If there are no solutions, then the upper and lower bounds are $\pm\infty$. In the event that there are colinear points, then this can lead to a situation where $(\kappa, \boldsymbol{d}^{\triangle,\pm}(\kappa)) = 0$ has multiple solutions in κ. In this case, we can take the minimum of the upper bounds and maximum of the lower bounds respectively. □

C.1.5 Proof of Theorem 4.2.6

Before proving Theorem 4.2.6, we first introduce a useful theorem for the proof.

Theorem C.1.4. *Consider a set of continuous random variables $X_i \in \mathbb{R}$ which are marginally identical but have an arbitrary dependency. Let $\widehat{\mathbb{M}}[X]$ denote the sample median of the set of these random variables. Then*

$$P(\widehat{\mathbb{M}}[X] \leq t) \leq 2P(X \leq t) \tag{C.1}$$

We find the proof in the supplementary materials in Section C.1.10. This bound is only useful up to $t = \mathbb{M}[X]$ at which point the upper bound is 1. The implication here is that we only need most of the midpoints between any two points to be reasonably separated. We later illustrate how we will find good midpoints and so we can verify their existence in any observed dataset.

Proof. Let $\Xi(Z_{\{1,\ldots,i\}})$ denote the set of distances from one midpoint to another as a function of their endpoints: Z_j, Z_k, Z_l, Z_r, i.e. $\Xi(Z_{\{1,\ldots,i\}}) = \{d(m(Z_j, Z_k), m(Z_l, Z_r))\}_{(j,k,l,r)\in\{1,\ldots,i\}}$ where $m(Z_j, Z_k)$ refers to the midpoint of Z_j and Z_k. Denote the set of events $M_i(t) = \widehat{\mathbb{M}}[\Xi(Z_{\{1,\ldots,i\}})] > t$ the median of these distances after sampling i endpoint positions is at least t, and let $G := G(t) = \cap_{i=1}^{K} M_i(t)$ be the intersection of a growing sequence of these medians all satisfy this bound. We can upper bound the probability that $\Phi(\mathcal{D}_K) \leq t$, that the smallest distance to the midpoint of the pair of endpoints by the following induction argument:

$$\begin{aligned}
&P(\Phi(\mathcal{D}_K) \leq t) \\
&= P(\Phi(\mathcal{D}_K) \leq t | \Phi(\mathcal{D}_{K-1}) \leq t, G) P(\Phi(\mathcal{D}_{K-1}) \leq t | G) P(G) \\
&+ P(\Phi(\mathcal{D}_K) \leq t | \Phi(\mathcal{D}_{K-1}) > t, G) P(\Phi(\mathcal{D}_{K-1}) > t | G) P(G) \\
&+ P(\Phi(\mathcal{D}_K) \leq t | G^c) P(G^c) \\
&\leq \underbrace{P(\Phi(\mathcal{D}_K) \leq t | \Phi(\mathcal{D}_{K-1}) \leq t, G)}_{(I)} P(\Phi(\mathcal{D}_{K-1}) \leq t | G) + \underbrace{P(G^c)}_{(II)}.
\end{aligned}$$

We first consider (I). Since we condition on G, at each step at least half of the midpoints are at a distance at least t away from one another. By the locally Euclidean assumption

(A2) and the continuity of the latent distribution on a convex region (C1), then since there are at least $\binom{K}{2}/2$ midpoints at a distance at least t away from one another:

$$\begin{aligned}(I) &= P(\Phi(\mathcal{D}_K) \leq t | \Phi(\mathcal{D}_{K-1}) \leq t, G) \\ &\leq \max\{0, 1 - \binom{K}{2}\alpha\frac{C_p}{2}t^p\} \\ &\leq \exp(-\binom{K}{2}\alpha\frac{C_p}{2}t^p)\end{aligned}$$

By recursion, we can bound this over a growing sequence of samples of midpoints $k \in \{1, 2, \ldots, K\}$

$$\begin{aligned}\prod_{k=1}^{K} P(\Phi(\mathcal{D}_K) \leq t | \Phi(\mathcal{D}_{K-1}) > t, G) &\leq \exp(-\sum_{k=1}^{K}\binom{k}{2}\alpha\frac{C_p}{2}t^p) \\ &= \exp(-\binom{K+1}{3}\alpha\frac{C_p}{2}t^p) \\ &\leq \exp(-(K+1)^3\alpha\frac{C_p}{12}t^p)\end{aligned}$$

Secondly, let us consider term (II). We will also demonstrate that this term is generally negligible compared to the first term. By a union bound arguement:

$$\begin{aligned}P(G^c) &= P(\cup_{k=1}^{K}\{\widehat{\mathbb{M}}[\Xi(Z_{\{1,\ldots,k\}})] \leq t\}) \\ &\leq \sum_{k=1}^{K} P(\widehat{\mathbb{M}}[\Xi(Z_{\{1,\ldots,k\}})] \leq t) \\ &\leq K(P(\widehat{\mathbb{M}}[\Xi(Z_{\{1,\ldots,k\}})] \leq t))\end{aligned}$$

We note that this in fact describes a set of marginally identical, yet correlated random variables, allowing up to leverage Theorem C.1.4.

$$\begin{aligned}P(\{\widehat{\mathbb{M}}[\Xi(Z_K)] \leq t\}) &\leq 2P(\Xi(Z_K) \leq t\}) \\ &= 2P(\cup_{m' \neq m} d(m, m') \leq t\}) \\ &\leq 2\sum_{m' \neq m} P(d(m, m') \leq t\})\end{aligned}$$

And by assumption (C3) in our theorem $P(d(m,m') \leq t) \leq \alpha_m C_p t^p$. Since there are $K\binom{K}{2}$ terms in this bounds, then we add a factor of $K\binom{K}{2}$ to construct the bound for (II). Hence we combine these bounds to obtain:

$$P(\Phi(\mathcal{D}_K) \leq t) \leq \exp(-(K+1)^3 \alpha_m \frac{C_p}{12} t^p) + 2\alpha_m C_p K \binom{K}{2} t^p$$

Letting $t = \frac{C_2}{K^{3/p}}$ will bound $\lim_{K\to\infty} P(\Phi(\mathcal{D}_K) \leq t_K)$ by any constant and therefore

$$\Phi(\mathcal{D}_K) = \mathcal{O}_P(K^{-3/p}).$$

□

C.1.6 Proof of Lemma 4.2.7

The proof draws on a similar structure to that of the consistency result of Lubold et al. [2023], we extend this to illustrate the rate of convergence.

Proof. Let $\epsilon_\delta := \{\max_{i,j} d(Z_i, Z_j) < \delta\}$. We wish to show that:

$$\lim_{\ell\to\infty} \frac{P(\epsilon_\delta | C_\ell)}{P(\epsilon_\delta^C | C_\ell)} \to \infty$$

for $\delta = O(\exp(-\widetilde{\mu}_d \ell/p))$.

Firstly we derive the probability of a clique forming, conditional on ϵ_δ. For convenience, we denote $\bar{\nu} = \frac{1}{\ell}\sum_{i=1}^{\ell} \nu_i$ and $\bar{d}(\boldsymbol{z} = \binom{\ell}{2}^{-1} \sum_{i<j}^{\ell} d(Z_i, Z_j)$

$$\begin{aligned} P(C_\ell | \epsilon_\delta) &= \int\int_{\epsilon_\delta} \exp(-\binom{\ell}{2}\bar{d}(z)) \exp\left(-\binom{\ell}{2}\bar{\nu}\right) dP_{Z|\epsilon_\delta}(z) dP_{\boldsymbol{\nu}}(\nu) \\ &= \Psi(P_\nu) \int_{\epsilon_\delta} \exp(-\binom{\ell}{2}\bar{d}(z)) dP_{Z|\epsilon_\delta}(z) \\ \text{where } \Psi(P_\nu) &= \int \exp\left(-\binom{\ell}{2}\bar{\nu}\right) dP_{\boldsymbol{\nu}}(\nu). \end{aligned}$$

Here we factor out the dependence on the random effects into the multiplicative term $\Psi(P_\nu)$ This can be developed analogously for $P(C_\ell | \epsilon_\delta^c)$:

$$\begin{aligned} P(C_\ell | \epsilon_\delta^c) &= \int\int_{\epsilon_\delta} \exp(-\binom{\ell}{2}\bar{d}(z)) \exp\left(-\binom{\ell}{2}\bar{\nu}\right) dP_{Z|\epsilon_\delta}(z) dP_{\boldsymbol{\nu}}(\nu) \\ &= \Psi(P_\nu) \int_{\epsilon_\delta^c} \exp(-\binom{\ell}{2}\bar{d}(z)) dP_{Z|\epsilon_\delta^c}(z) \end{aligned}$$

This allows us to ignore the dependence on the random effects when taking the ratio

$$\frac{P(\epsilon_\delta|C_\ell)}{P(\epsilon_\delta^C|C_\ell)} = \frac{P(C_\ell|\epsilon_\delta)}{P(C_\ell|\epsilon_\delta^C)}\frac{P(\epsilon_\delta)}{P(\epsilon_\delta^c)}.$$

We now proceed with the remainder of the proof. Consider the ball of radius $\frac{\delta}{2}$ at a point q $B(\delta, q)$. This describes one set of points for which all points have a maximal distance of δ. Thus $\left\{\{Z_i\}_{i=1}^\ell \in B(\delta, q)\right\} \subset \epsilon_\delta$ for any q. Since f is continuous, then for some $\widetilde{\delta}$ if $f(q)$ has positive probability $f(q) > 0$ for some q then for all $\delta \leq \widetilde{\delta}$, $f(x) \geq c_2; \forall x \in B(q, \delta)$ for some constant c_2. Therefore $P(\epsilon_\delta) \geq P(\{Z_i\}_{i=1}^\ell \in B(\delta, q)) \geq \left(c_2\left(\frac{\delta}{2}\right)^p\right)^\ell$.

Next we exploit Lemma A.1 in Lubold et al. [2023] states that if Z_i are drawn iid from a latent distribution with funute mean μ_d, then

$$P(C_\ell|\epsilon_\delta^c) \leq \exp\left(-\binom{\ell}{2}\mu_d\right)$$

Putting this all together, for $P(\epsilon_\delta) \leq \frac{1}{2}$

$$\begin{aligned}
\frac{P(\epsilon_\delta|C_\ell)}{P(\epsilon_\delta^C|C_\ell)} &= \frac{P(C_\ell|\epsilon_\delta)}{P(C_\ell|\epsilon_\delta^C)}\frac{P(\epsilon_\delta)}{1-P(\epsilon_\delta)} \\
&\geq \frac{P(C_\ell|\epsilon_\delta)}{P(C_\ell|\epsilon_\delta^C)}\frac{1}{2}P(\epsilon_\delta) \\
&\geq \frac{1}{2}\frac{P(C_\ell|\epsilon_\delta)}{P(C_\ell|\epsilon_\delta^C)}c_2^\ell\left(\frac{\delta}{2}\right)^{p\ell} \\
&\geq \frac{1}{2}\frac{\exp(-\binom{\ell}{2}\delta)}{P(C_\ell|\epsilon_\delta^C)}c_2^\ell\left(\frac{\delta}{2}\right)^{p\ell} \\
&\geq \frac{1}{2}\frac{\exp(-\binom{\ell}{2}\delta)}{\exp(-\binom{\ell}{2}\mu_d)}c_2^\ell\left(\frac{\delta}{2}\right)^{p\ell} \\
&= \frac{1}{2}\exp\left(-\binom{\ell}{2}\delta + \binom{\ell}{2}\mu_d + \ell\log(c_2) + \log(\delta/2)p\ell\right)
\end{aligned}$$

Therefore letting $\delta = 2\exp(-\widetilde{\mu}_d\frac{\ell-1}{p})$ for $\widetilde{\mu}_d < \mu_d$. Then

$$\lim_{\ell\to\infty}\frac{P(\epsilon_\delta|C_\ell)}{P(\epsilon_\delta^C|C_\ell)} \to \infty$$

and therefore the proof is complete. □

C.1.7 Proof of Lemma 4.2.8

At a high level, the proof structure is nearly identical to Lemma 4.2.7, however, we swap the roles of ν and $d(Z_i, Z_j)$.

Proof. Let $\varepsilon_\delta = \left\{\min_{i\in\{1,2,\dots,\ell\}} \nu_i \geq -\delta\right\}$

If $f_\nu(\nu)$ is continuous around 0 then for small enough $\delta, F_\nu(\nu > -\delta) \geq c_3\delta$ for some c_3. As in the proof of Lemma 4.2.7, we similarly take the ratio and integrate out $\Psi(P_{\bar{d}})$ since

$$\begin{aligned} P(C_\ell|\varepsilon_\delta) &= \int_{\varepsilon_\delta}\int \exp(-\binom{\ell}{2}\bar{d}(z)) \exp\left(-\binom{\ell}{2}\bar{\nu}\right) dP_Z(z) dP_{\boldsymbol{\nu}|\varepsilon_\delta}(\nu) \\ &= \Psi(P_{\bar{d}}) \int_{\varepsilon_\delta} \exp\left(-\binom{\ell}{2}\bar{\nu}\right) dP_{\boldsymbol{\nu}|\varepsilon_\delta}(\nu) \\ \text{where } \Psi(P_{\bar{d}}) &= \int \exp(-\binom{\ell}{2}\bar{d}(\boldsymbol{z}))dP_Z(\boldsymbol{z}) \end{aligned}$$

Then consider the set $[-\delta, 0]$. Then $P(\varepsilon_\delta) \geq (c_3(\delta))^\ell$.

Next we consider the probability of a clique, conditional on the fact that the locations did not occur in set ε_δ^c.

$$P(C_\ell|\varepsilon_\delta^c) \leq \int \exp\left(2\binom{\ell}{2}\bar{\nu}\right) dP_{\boldsymbol{\nu}}(\boldsymbol{\nu}).$$

Let $\mu_\nu = \mathbb{E}[\nu] < 0$. This is due to the fact ε_δ denotes the event $\{\min_i \nu_i < -\delta\}$ it is reasonable that $P(C_\ell|\epsilon_\ell^c) \geq P(C_\ell)$ since we are excluding the events most likely to create a clique. Furthermore by the same argument as in Lemma A.1 in Lubold et al. [2023] we can show

$$P(C_\ell|\varepsilon_\delta^c) \leq \exp(-2\binom{\ell}{2}|\mu_\nu)$$

This can be achieved as follows:

$$\begin{aligned}
P(C_\ell|\varepsilon_\delta^c) &= E\left[\prod_{i<j}\exp(\nu_i+\nu_j)\middle|\varepsilon_\delta^c\right] \\
&\leq E\left[\prod_{i<j}\exp(\nu_i+\nu_j)\right] \\
&\leq \prod_{i<j}\left(E\left[\exp(\binom{\ell}{2}(\nu_i+\nu_j))\right]\right)^{1/\binom{\ell}{2}} \\
&\leq \exp(\binom{\ell}{2}2\mu_\nu)\prod_{i<j}\left(E\left[\exp(\binom{\ell}{2}(\eta_i+\eta_j))\right]\right)^{1/\binom{\ell}{2}} \\
&\leq \exp(\binom{\ell}{2}2\mu_\nu)\times 1 \\
&= \exp(-\binom{\ell}{2}|\mu_\nu|)
\end{aligned}$$

due to the fact that the probability of connection is higher within ε_δ than outside of ε_δ^c. The remaining steps follow from an application of Holder's generalized inequality and $x_i = \mu_\nu + \eta_i$ for some noise η_i.

Returning to the computation of the ratio of probabilities,

$$\begin{aligned}
\frac{P(\varepsilon_\delta|C_\ell)}{P(\varepsilon_\delta^C|C_\ell)} &= \frac{P(C_\ell|\varepsilon_\delta)}{P(C_\ell|\varepsilon_\delta^C)}\frac{P(\varepsilon_\delta)}{1-P(\varepsilon_\delta)} \\
&\geq \frac{1}{2}\frac{P(C_\ell|\varepsilon_\delta)}{P(C_\ell|\varepsilon_\delta^C)}P(\varepsilon_\delta) \\
&\geq \frac{1}{2}\frac{P(C_\ell|\varepsilon_\delta)}{P(C_\ell|\varepsilon_\delta^C)}c_3^\ell\,(\delta)^\ell \\
&\geq \frac{1}{2}\frac{\exp(-2\binom{\ell}{2}\delta)}{P(C_\ell|\varepsilon_\delta^C)}c_3^\ell\,(\delta)^\ell \\
&\geq \frac{1}{2}\exp\left(-2\binom{\ell}{2}\delta + 2\binom{\ell}{2}|\mu_\nu| + \ell\log(c_3) + \log(\delta)\ell\right)
\end{aligned}$$

therefore we can let $\delta = \exp(-|\widetilde{\mu}_\nu|2(\ell-1))$ for any $|\widetilde{\mu}_\nu| < |\mu_\nu|$ and therefore the proof is complete.

□

C.1.8 Proof of Theorem 4.2.9

In order to prove the derive the asymptotic distribution of the distance estimator, we introduce a useful theorem. This theorem will illustrate the rate of estimation of the random effects.

Lemma C.1.5. *Let $W \subset \{1, 2, \ldots, n\}$ denote a subset of indices which form a clique (C_ℓ). Let d_i denote the degree of node $i \in W$ where $|W| = \ell$ and assume that the points in the clique have a common latent position. Denote the estimator of $\gamma_{\mathcal{W}}$. Let $\mu_\nu := E[\nu]$. Then for any $\mu_\nu < \widetilde{\mu}_\nu < 0$*

$$\widehat{\gamma}_{\mathcal{W}} = \log\left(\frac{1}{\ell}\sum_{i\in W}\frac{d_i}{\max_{j\in W} d_j}\right).$$

Then if (E1) in Lemma 4.2.8 holds, then define $\widetilde{\mu}_\nu$ as in Lemma 4.2.8

$$\widehat{\gamma}_{\mathcal{W}} - \gamma_{\mathcal{W}} = \mathcal{O}_P\left(\max\left\{\exp(\widetilde{\mu}_\nu \ell), \frac{1}{\sqrt{n}}\right\}\right) \tag{C.2}$$

Therefore, estimation of $\gamma_{\mathcal{W}}$ within a clique can occur at an exponential rate, meaning that this estimation will be negligible compared to the average cross-clique probabilities.

Proof. The proof here is a straightforward application of the plug in estimator of $\widehat{p}_{\mathcal{X}\mathcal{Y}}$ and the set of random effects $\widehat{\gamma}_{X/Y}$. Then by the Lindeberg-Feller CLT:

$$\sqrt{\ell^2\sigma_\ell}(\widehat{p}_{\mathcal{X}\mathcal{Y}} - p_{\mathcal{X}\mathcal{Y}}) \to_d N(0, 1)$$

We first consider the localization of nodes within a clique, specifically

$$\begin{aligned} p_{\mathcal{X}\mathcal{Y}} &= \frac{1}{\ell^2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\exp(\nu_x + \nu_y - d_{xy}) \\ &= \frac{1}{\ell^2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\exp(\nu_x + \nu_y - d_{\mathcal{X}\mathcal{Y}} + o_P(-\widetilde{\mu}_d\ell)) \text{ by Lemma 4.2.7} \\ \implies d_{\mathcal{X}\mathcal{Y}} &= -\log(p_{\mathcal{X}\mathcal{Y}}) + \gamma_X + \gamma_Y + o_P(-\widetilde{\mu}_d\ell) \end{aligned}$$

We now study our distance estimator as per equation (4.12)

$$\begin{aligned}
\widehat{d}_{\mathcal{XY}} &= -\log(\widehat{p}_{\mathcal{XY}}) + \widehat{\gamma}_X + \widehat{\gamma}_Y + o_P(\widetilde{\mu}_d \ell) + o_P(-\widetilde{\mu}_d \ell) \\
\sqrt{\ell^2}\widehat{d}_{\mathcal{XY}} &= -\sqrt{\ell^2}\log(\widehat{p}_{\mathcal{XY}}) + \sqrt{\ell^2}\widehat{\gamma}_X + \sqrt{\ell^2}\widehat{\gamma}_Y \\
&= -\sqrt{\ell^2}\log(\widehat{p}_{\mathcal{XY}}) + \gamma_X + \gamma_Y + o_P(\ell n^{-1/2}) + o_P((\ell+1)\exp(-\widetilde{\mu}_d \ell)) + o_P(\ell(-\widetilde{\mu}_\nu \ell)) \text{ By Lemma 4.2.8} \\
&= -\sqrt{\ell^2}\log(\widehat{p}_{\mathcal{XY}}) + \gamma_X + \gamma_Y + o_P(1).
\end{aligned}$$

And hence the higher order terms are asymptotically negligible compared to the estimation of $p_{\mathcal{XY}}$. Therefore by the delta method we can derive the asymptotic distribution of the distance estimator

$$\sqrt{\ell^2 \frac{\sigma_\ell}{p_{\mathcal{XY}}}}\left(\widehat{d}_{\mathcal{XY}} - d_{\mathcal{XY}}\right) \to_d N(0,1)$$

□

C.1.9 Proof of Lemma C.1.5

Proof. Consider a clique $\mathcal{X}$ of size ℓ. We note that by Lemma 4.2.7 that the nodes within the clique are exponentially close to one another, and $d_{xx'} = o_P(\exp(-\widetilde{\mu}_d \ell))$. Therefore we consider the ratio of the connection probability to any node in the network $P(A_{xi} = 1|\mathcal{X})/P(A_{x'i} = 1|\mathcal{X})$, where

$$\begin{aligned}
P(A_{xi} = 1|\mathcal{X}) &= \int \exp(\nu_x + \nu_i - d(Z_x, Z_i)) dF_\nu(\nu_i) dF_Z(Z_i) \\
&= \exp(\nu_x) \int \exp(\nu_i - d(Z_x, Z_i)) dF_\nu(\nu_i) dF_Z(Z_i) \\
&= \exp(\nu_x) \int \exp(\nu_i - d(Z_{x'}, Z_i) + o_P(\exp(-\widetilde{\mu}_d \ell))) dF_\nu(\nu_i) dF_Z(Z_i) \\
\implies P(A_{xi} = 1|\mathcal{X})/P(A_{x'i} = 1|\mathcal{X}) &= \exp(\nu_x - \nu_{x'}) \exp(o_P(\exp(-\widetilde{\mu}_d \ell)))) \\
&= \exp(\nu_x - \nu_{x'}) + o_P(\exp(-\widetilde{\mu}_d \ell)))
\end{aligned}$$

Next we note that the empirical ratio of the connection probabilities can be computed using a ratio of degrees

$$\frac{\widehat{P}(A_{xi}=1|\mathcal{X})}{\widehat{P}(A_{x'i}=1|\mathcal{X})}=\frac{d_x/n}{d_{x'}/n}$$

By a simple application of Hoeffding's inequality, we see that

$$|\widehat{P}(A_{xi}=1|\mathcal{X})-P(A_{xi}=1|\mathcal{X})|=O_P(1/\sqrt{n})$$

And therefore

$$\frac{\widehat{P}(A_{xi}=1|\mathcal{X})}{\widehat{P}(A_{x'i}=1|\mathcal{X})}=\frac{P(A_{xi}=1|\mathcal{X})}{P(A_{x'i}=1|\mathcal{X})}+O_P(1/\sqrt{n})$$

Lastly we apply this to the estimation of $\gamma_{\mathcal{X}}$ where.

$$\gamma_{\mathcal{X}}=\log\left(\frac{1}{\ell}\sum_{x\in\mathcal{X}}\exp(\nu_i)\right)$$

By Lemma C.1.5 we note that the random effects converge to 0 within a clique, therefore setting the smallest magnitude estimate to 0 we finally observe that the final rate for estimating the random effects is as follows:

$$\widehat{\gamma}_{\mathcal{X}}-\gamma_{\mathcal{X}}=O_P(1/\sqrt{n})+o_P(\exp(-\widetilde{\mu}_d\ell)))+o_P(\exp(-|\widetilde{\mu}_\nu|\ell))).$$

□

C.1.10 Proof of Theorem C.1.4

In order to prove Theorem C.1.4 we first introduce a series of lemmas that will be useful for proof. The proofs of these are contained in the subsequent subsections of the appendix.

Let X_i be a random variable with marginal distribution function F, and let $X=(X_1,X_2,\ldots,X_n)$ denote a random vector for which all marginal distributions are F, that may

in general be correlated. Let H denote the joint distribution. By Sklar's Theorem [SKLAR and M., 1959, Durante et al., 2013] we can express the joint distribution as

$$\begin{aligned} H(\boldsymbol{x}) &= P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \\ &= \mathbb{C}(F(x_1), F(x_2), \ldots, F(x_n)) \end{aligned}$$

where $\mathbb{C}$ is a copula for the joint distribution. A copula is simply a multivariate distribution function for a set of uniform random variables which explains their dependence.

Lemma C.1.6. *For a set of random variables X_i with equal marginal distribution, with arbitrary correlation structure denoted by a copula $\mathbb{C}$, then $\widehat{\mathbb{M}}(X) = F^{-1}(\widehat{\mathbb{M}}(U))$ where $U \sim \mathbb{C}$ is a multivariate uniform random variable $U = (U_1, U_2, \ldots, U_n)$.*

This simply relates the distribution of the median of correlated random variables to the median in the copula representation.

Lemma C.1.7. *Consider for a fixed z, a set of intervals $A_I = \cap_{i \in I}\{U_i \leq z\}$ where I is a subset of $\{1, 2, \ldots, n\}$. Consider a sequence of sets A_J $J \in \mathcal{J}$ such that any pair of J, J' contain at least 1 overlapping index. Then*

$$\mathbb{C}(\cup_{J \in \mathcal{J}} A_J) \leq \mathbb{W}(\cup_{J \in \mathcal{J}} A_J) \tag{C.3}$$

where $\mathbb{W}$ is the perfectly collinear measure corresponding to the upper Fréchet Hoeffding Bound [Frechet, 1957, Hoeffding, 1940].

This lemma is useful since it can describe an upper bound on the union of sets of intervals A_J. It is used in an intermediary step in the derivation of Lemma C.1.8.

Lemma C.1.8. *For any copula, $\mathbb{C}$*

$$\mathbb{C}(\widehat{\mathbb{M}}(U) \leq z) \leq 2z \tag{C.4}$$

This lemma is useful in that it provides a worst-case bound on the median of uniform random variables. From here the proof is short.

Proof. The proof immediately is a result of Lemma C.1.6 and C.1.8.

$$\begin{aligned} P(\widehat{\mathbb{M}}(X) \leq t) &= \mathbb{C}(F^{-1}(\widehat{\mathbb{M}}(U)) \leq t) \text{ Lemma C.1.6} \\ &= \mathbb{C}(\widehat{\mathbb{M}}(U) \leq F(t)) \\ &\leq 2F(t) \text{ Lemma C.1.8} \\ &= 2P(X_1 \leq t) \end{aligned}$$

□

This concludes the proof of Theorem C.1.4. We subsequently prove the results of Lemmas C.1.6, C.1.7, C.1.8.

C.1.11 Proof of Lemma C.1.6

Proof. We can express $X = F^{-1}(U)$ where the vector U is sampled from the copula $\mathbb{C}$. Then if n is odd, with the subscript (i) denoting the i^{th} order statistic, $X_{((n+1)/2)}$ is the median. In this case, clearly $X_{(i)} = F^{-1}(U_{(i)})$. Now consider the case if n is even. If z is any median of U then $z \in [U_{(n/2)}, U_{(n/2+1)}]$ then

$$\begin{aligned} & z \in [U_{(n/2)}, U_{(n/2+1)}] \\ \iff & F^{-1}(z) \in [F^{-1}(U_{(n/2)}), F^{-1}(U_{(n/2+1)})] \\ \iff & F^{-1}(z) \in [X_{(n/2)}, X_{(n/2+1)}] \end{aligned}$$

Hence medians of the uniform distribution generated by the corresponding copula are mapped to the median of the observed variables. □

C.1.12 Proof of Lemma C.1.7

Proof. We prove this by induction. Consider the base case where there are two sets of intervals A_{J_1}, A_{J_2}. Suppose there exists some copula $\mathbb{Q}$ such that $\mathbb{Q}(A_{J_1} \cup A_{J_2}) > \mathbb{W}(A_{J_1} \cup$

A_{J_2}). Firstly under the perfectly correlated copula $\mathbb{W}$

$$\begin{aligned}\mathbb{W}(A_{J_1} \cup A_{J_2}) &= \mathbb{W}(A_{J_1}) + \mathbb{W}(A_{J_2}) - \mathbb{W}(A_{J_1} \cap A_{J_2}) \\ &= \min\{1, z\} + \min\{1, z\} - \min\{1, z\} \\ &= z + z - z \\ &= z\end{aligned}$$

Then since $\{A_{J_1} \cup A_{J_2}\} \subset \{U_k \leq z\}$ for some index k. Then

$$\mathbb{Q}(A_{J_1} \cup A_{J_2}) \leq \mathbb{Q}(U_k \leq z)$$

However, then this implies that the marginal distribution of U_k under $\mathbb{Q}$ is not uniform, i.e. $\mathbb{Q}(U_k \leq z) > z$ and thus $\mathbb{Q}$ is not a copula. Therefore it must be the case that $\mathbb{Q}(A_1 \cup A_{J_2}) \leq \mathbb{W}(A_{J_1} \cup A_{J_2})$ holds in the base case.

We next prove the induction step. Suppose the following holds for any copula $\mathbb{Q}$ and a sequence of n sets of intervals $\mathcal{J}_n$.

$$\mathbb{Q}(\cup_{J \in \mathcal{J}_n} A_j) \leq \mathbb{W}(\cup_{J \in \mathcal{J}_n} A_j)$$

for some n. Denote $B = \cup_{J \in \mathcal{J}_n} A_J$ then we must show for a new set of intervals A_{n+1}

$$\mathbb{Q}(B \cup A_{n+1}) \leq \mathbb{W}(B \cup A_{n+1}).$$

Again, we prove this by contradiction. Suppose that there exists $\mathbb{Q}$ such that $\mathbb{Q}(B \cup A_{n+1}) > \mathbb{W}(B \cup A_{n+1})$

$$\mathbb{W}(B \cup A_{n+1}) = \mathbb{W}(B) + \mathbb{W}(A_{n+1}) - \mathbb{W}(B \cap A_{n+1})$$

Clearly by definition of $\mathbb{W}$: $\mathbb{W}(B) = z$, $\mathbb{W}(A_{n+1}) = z$ and $\mathbb{W}(B \cap A_{n+1}) = z$. Then

$$\begin{aligned}\mathbb{Q}(B \cup A_{n+1}) &= \mathbb{Q}(B) + \mathbb{Q}(A_{n+1}) - \mathbb{Q}(B \cap A_{n+1}) \\ &\leq \mathbb{W}(B) + \mathbb{Q}(A_{n+1}) - \mathbb{Q}(B \cap A_{n+1}) \\ &\leq z + \underbrace{\mathbb{Q}(A_{n+1}) - \mathbb{Q}(B \cap A_{n+1})}_{\geq 0}\end{aligned}$$

which generates a contradiction hence

$$\mathbb{Q}(B \cup A_{n+1}) \leq \mathbb{W}(B \cup A_{n+1})$$

□

C.1.13 Proof of Lemma C.1.8

Proof. Define the events A_J as in Lemma C.1.7 of size $\lceil n/2 \rceil$. Then the event $\{\widehat{\mathbb{M}}(U) \leq z\}$ is equal to the union of all such A_J as the median will be equivalent to the case when at least half of all uniforms are below z. We can partition $\{A_J\}_{J \in \mathcal{J}}$ into two sets S_1, S_2 for which any two pairs of A_J in a set have at least one overlapping index. This can be done by taking all the sets $\{A_J\}_{J \in \mathcal{J}}$ which suggest $\{U_1 \leq z\}$ and placing them into set S_1. Then all other sets must be placed in set S_2. Since there are $(n-1)$ possible remaining indices $\{2, 3, \ldots, n\}$ available for S_2 then any two events must have an overlapping index by the pigeonhole principle. Therefore

$$\begin{aligned}
\mathbb{C}(\widehat{\mathbb{M}}(U) \leq z) &= \mathbb{C}(\cup A_i) \\
&= \mathbb{C}(S_1 \cup S_2) \\
&\leq \mathbb{C}(S_1) + \mathbb{C}(S_2) \text{ Union Bound} \\
&\leq \mathbb{W}(S_1) + \mathbb{W}(S_2) \text{ Lemma C.1.7} \\
&= 2z.
\end{aligned}$$

□

C.1.14 A corollary of Lemma C.1.8

A corollary immediately follows from Lemma C.1.8. This bounds the deviation of an arbitrary set of correlated uniform random variables by the distribution of its marginal.

Corollary C.1.9. *For any copula* $\mathbb{C}$*.*

$$\mathbb{C}(|\widehat{\mathbb{M}}(U) - 1/2| > \epsilon) \leq 2P(|U - 1/2| > \epsilon) \tag{C.5}$$

Proof. $P(|U - 1/2| > \epsilon)$ is simply the marginal distribution of a uniform distribution.

$$P(|U - 1/2| > \epsilon) = \max\{1 - 2\epsilon, 0\}$$

Next we note that

$$\{|\widehat{\mathbb{M}}(U) - 1/2| > \epsilon\} = \{\widehat{\mathbb{M}}(U) < 1/2 - \epsilon\} \cap \{\widehat{\mathbb{M}}(U) > 1/2 + \epsilon\}$$

Note that we can define $1 - V = U$ where the measure of V, $\widetilde{\mathbb{C}}$ is also a copula, since this is a joint distribution of marginally uniform variables.

$$\{\widehat{\mathbb{M}}(U) > 1/2 + \epsilon\} = \{\widehat{\mathbb{M}}(V) < 1/2 - \epsilon\}$$

hence by a union bound.

$$\begin{aligned}\mathbb{C}(|\widehat{\mathbb{M}}(U) - 1/2| > \epsilon) &\leq \mathbb{C}(\widehat{\mathbb{M}}(U) < 1/2 - \epsilon) + \widetilde{\mathbb{C}}(\widehat{\mathbb{M}}(V) < 1/2 - \epsilon) \\ &= 2(1/2 - \epsilon) + 2(1/2 - \epsilon)\end{aligned}$$

then since the copula is non-negative

$$\begin{aligned}\mathbb{C}(|\widehat{\mathbb{M}}(U) - 1/2| > \epsilon) &\leq 2\max\{1 - 2\epsilon, 0\} \\ &= 2P(|U - 1/2| > \epsilon)\end{aligned}$$

□

C.2 Additional Computational Details

C.2.1 Newton Method for $\widehat{\kappa}$

Given a set of distances $\widehat{d}$ we can estimate the curvature using a newton method. Firstly, we compute the derivative of $g(\kappa, d)$ with respect to κ.

$$\begin{aligned}\frac{\partial}{\partial\kappa}g(\kappa,\boldsymbol{d}^{\triangle}) = (1/(4\kappa^2))\Bigg(&8\cos(d_{xm}\sqrt{\kappa}) - 4\cos(d_{xz}\sqrt{\kappa})\sec(d_{yz}\frac{\sqrt{\kappa}}{2}) \\ &+ 4d_{xm}\sqrt{\kappa}\sin(d_{xm}\sqrt{\kappa}) \\ &- 2d_{xy}\sqrt{\kappa}\sec(d_{yz}\frac{\sqrt{\kappa}}{2})\sin(d_{xy}\sqrt{\kappa}) \\ &- 2d_{xz}\sqrt{\kappa}\sec(d_{yz}\frac{\sqrt{\kappa}}{2})\sin(d_{xz}\sqrt{\kappa}) \\ &+ d_{yz}\sqrt{\kappa}\cos(d_{xz}\sqrt{\kappa})\sec(d_{yz}\frac{\sqrt{\kappa}}{2})\tan(d_{yz}\frac{\sqrt{\kappa}}{2}) \\ &+ \cos(d_{xy}\sqrt{\kappa})\sec((d_{yz}\sqrt{\kappa})/2)(-4 + d_{yz}\sqrt{\kappa}\tan(d_{yz}\frac{\sqrt{\kappa}}{2})\Bigg)\end{aligned}$$

This allows us to construct a newton method for estimating the root $\widehat{\kappa}$

$$\widehat{\kappa}_{(m+1)} = \widehat{\kappa}_{(m)} - \frac{g(\widehat{\kappa}_{(m)},\widehat{d})}{\frac{\partial}{\partial\kappa}g(\widehat{\kappa}_{(m)},\widehat{d})}$$

C.2.2 Distance Estimation

We recall the problem of estimating a distance matrix from a set of cliques $\mathcal{C}$. Though this problem is convex, due to the $O(K^3)$ restrictions in the problem, it the problem is often slow to reach a solution in `CVXR`. Instead, we solve this problem using a successive second order approximation.

$$
\begin{aligned}
f(D) &:= \sum_{x,y\in\mathcal{C},i\in x,j\in Y} \Big(A_{ij}\left(\nu_i+\nu_j-d_{xy}\right) \\
&+ (1-A_{ij})\log\left(1-\exp\left(\nu_i+\nu_j-d_{xy}\right)\right)\Big) \\
&\approx \sum_{x,y\in\mathcal{C},i\in x,j\in Y} \Bigg(A_{ij}\left(\nu_i+\nu_j-D_{0,x,y}\right) \\
&+ (1-A_{ij})\log\left(1-\exp\left(\nu_i+\nu_j-D_{0,x,y}\right)\right) \\
&+ \left((A_{ij}-1)\frac{\exp\left(\nu_i+\nu_j-D_{0,x,y}\right)}{1-\exp\left(\nu_i+\nu_j-D_{0,x,y}\right)} - A_{ij}\right)\left(d_{xy}-D_{0,x,y}\right) \\
&+ (A_{ij}-1)\frac{\exp\left(\nu_i+\nu_j-D_{0,x,y}\right)}{\left(1-\exp\left(\nu_i+\nu_j-D_{0,x,y}\right)\right)^2}\frac{\left(d_{xy}-D_{0,x,y}\right)^2}{2}\Bigg) \\
&:= \widetilde{g}(D,D_0)
\end{aligned}
$$

Hence to compute the global solution $\widehat{\boldsymbol{D}}$, we can iteratively solve the following optimization problem

$$
\begin{aligned}
\widehat{\boldsymbol{D}}_{t+1} &= \operatorname{argsup}_{D\in\in\mathbb{R}^{K\times K}}\widetilde{g}(D,\widehat{\boldsymbol{D}}_t) \\
D_{ij} &\geq 0 \quad \text{for all } i,j \\
\operatorname{Diag}(D) &= 0 \\
\operatorname{tr}(E_s^\top D) &\geq 0 \forall s\in\mathcal{S}
\end{aligned}
$$

This process can be further sped up by choosing a good initialization matrix. We can do this by using the unconstrained maximum likelihood estimate $\widehat{\boldsymbol{D}}_U$, which is very fast to compute but does not enforce triangle inequality restrictions. This can be computed analogously as in Theorem 4.2.9. Though many of the distances $\widehat{\boldsymbol{D}}_U$ may not satisfy the triangle inequality, we can trim the distances so that $\widehat{\boldsymbol{D}}_U$ form a distance matrix, and use this as the starting point. The Floyd-Warshall Algorithm is a possible option for constructing a distance matrix from a noisy matrix which might not have a distance structure. A natural extension to this in our context is seen in Algorithm 15.

Algorithm 15 Adapted Floyd-Warshall Algorithm

Require: $D \in \mathbb{R}_{\geq 0}^{K \times K}$

Ensure: $D = D^\top$

Trim entries below 0: $D[D < 0] \leftarrow 0$
for $k \in \{1, 2, \ldots, n\}$ **do**
    for $j \in \{1, 2, \ldots, n\}$ **do**
        for $i \in \{1, 2, \ldots, n\}$ **do**
            if $D_{ij} > (D_{ik} + D_{kj})$ **then**
                $D_{ij} \leftarrow (D_{ik} + D_{kj})$
            end if
        end for
    end for
end for

One can draw similarities here to the problem of sparse metric repair. Metric repair seeks to adjust the fewest entries in a noisy distance matrix so that it still preserves the properties of being a metric (positivity, triangle inequality). Gilbert and Jain [2017] illustrated this Floyd-Warshall algorithm to be a solution to a special type known as decrease only metric repair.

C.3 Additional Discussion on the Latent Distance Model

C.3.1 Other Link Functions

Another common link function is the logistic link where the generative model for the network is as follows:

$$\nu_i \sim F_\nu, \quad \nu_i \leq 0$$
$$Z_i \sim F_Z, \quad Z_i \in \mathcal{M}^p$$
$$P(A_{ij} = 1) = \text{logit}(\nu_i + \nu_j + \varphi - d(Z_i, Z_j)).$$

We can consistently estimate the node level parameters up to a constant shift using conditional maximum likelihood as in the semiparametric Rasch model [Andersen, 1970]. The parameter φ controls the global sparsity. Similar to before, we note that ν_i terms are likely to be very large in a cliques then we can set the largest parameter in each group to be nearly zero.

Other link functions may be used but will likely all need specific methods to estimate ν_i parameters within each clique. However, we have shown how it can be developed in these two canonical cases.

C.3.2 Alternative Estimators of the Distance Matrix

Additional methods for estimating the distance matrix can be developed. A promising direction is to utilize structured sparsity in the distance matrix.

Our model exhibits numerous similarities to the β-model of network formation. In this framework, each node in the network possesses a gregariousness parameter β. An extension of this is the sparse beta model Chen et al. [2021]. In our context, ν functions analogously to β:

$$P(A_{ij} = 1|\nu) = \Lambda(\mu + \nu_i + \nu_j)$$

where $\|\nu\|_0 \leq s$ for some $s \ll n$. This parameterization facilitates a diminishing value of μ and is suitable for sparsely growing networks.

Furthermore, a distance matrix can be incorporated as follows:

$$P(A_{ij} = 1|\nu) = \Lambda(\mu + \nu_i + \nu_j - d_{ij})$$

where $\|D\|_0 \leq s_D \ll n$, $D \in \mathcal{D}$. Here, $\mathcal{D}$ represents the convex region of matrices constrained to be distances. If sparsity is not maintained, the interior dimension of $\mathcal{D}$ for a set of n points is $\binom{n}{2}$. Consequently, restricting the analysis solely to the distance matrix still

results in $\binom{n}{2}$ observations, making estimation infeasible without additional shared structure, i.e. through sparsity of the distance matrix.

Additionally, this formulation would be able to adjust to the sparsity level of the network data by asymptotically letting $\mu \to -\infty$. While a lasso-type procedure could be considered for estimating this distance matrix, a complete discussion of such methodologies is beyond the scope of this paper and is designated as future work.

C.3.3 Rates of Clique Formation

Here we clarify the notion of the likelihood of forming cliques of a given size. A famous result by Grimmett and McDiarmid [1975] states that the largest clique within an Erdos-Reyni random graph, $CN(n)$

$$\frac{CN(n)}{\log(n)} \to_{a.s.} \frac{2}{\log(1/p)}$$

where $\to_{a.s.}$ indicates almost sure convergence.

The behaviour governing the formation of the cliques is determined by the concentration of the latent positions in the latent space. Since this requires points that are exceedingly close together, the curvature of the space does not come into play here, but rather in the tendency of connections across cliques. In the case of a hyperbolic space, this tends to generate tree-like structures between clusters of cliques, as seen in Figure 4.11(b). Hence a hyperbolic space itself does not prevent the formation of cliques, so long as there as there is a reasonable concentration of the positions in the space.

However, in our case, due to the fact that we assume that there is some continuous distribution of latent positions and gregariousness parameters, then we can illustrate a polynomial growth in the size of cliques. This results in probabilities of connection approaching arbitrarily close to 1 rather than being bounded away from 1 by a fixed probability. Let $\mathfrak{S}(\ell)$ denote the combinations of nodes of size ℓ and $S \in \mathfrak{S}(\ell)$.

Let W_S denote the event that the nodes in S form a clique. Our goal is to express the expected number of cliques of size ℓ generated from the model $\mathfrak{W} = \sum_{S \in \mathfrak{S}(\ell)} W_S$.

Theorem C.3.1. *Let $\mathfrak{W}_{n,\ell}$ be the number of cliques formed of size ℓ formed from a network of size n. If assumptions (D1) in 4.2.7 and (E1) in 4.2.8 hold, and $\ell = \mathcal{O}(n^{1/(p+2)-\epsilon'})$ for any $\epsilon' > 0$. Then the expected number of cliques of size ℓ diverges i.e.*

$$\lim_{n\to\infty} \mathbb{E}[\mathfrak{W}_{n,\ell}] \to \infty \tag{C.6}$$

Proof. We first lower bound $\mathbb{E}[W_S]$ then utilize the linearity of expectation in order to compute this lower bound.

$$\begin{aligned}\mathbb{E}[W_S] &= \int \prod_{i<j} \exp(\nu_i + \nu_j - d(Z_i, Z_j)) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}) dF_{\boldsymbol{Z}}(\boldsymbol{Z}) \\ &= \int \exp((\ell-1)\sum_{i=1}^{\ell} \nu_i - \sum_{i<j}^{\ell} d(Z_i, Z_j)) dF_{\boldsymbol{\nu}}(\boldsymbol{\nu}) dF_{\boldsymbol{Z}}(\boldsymbol{Z})\end{aligned}$$

Next, using assumptions (D1) in 4.2.7 and (E1) in 4.2.8 we focus on the concentration of the latent positions.

Suppose that $\epsilon_\delta = \{\max_{i=1}^{\ell} d(Z_i, Z_j) \leq \delta\}$ and $\varepsilon_\delta = \{\min_{i=1}^{\ell} \nu_i > -\delta\}$ then $P(\epsilon_\delta) \geq (c_2\delta/2)^{p\ell}$ and $P(\varepsilon_\delta) \geq (c_3\delta)^{\ell}$.

Therefore:

$$\mathbb{E}[W_S] \geq \exp(-2\ell(\ell-1)\delta)(c_2\delta/2)^{p\ell}(c_3\delta)^{\ell}$$

and we let $\delta = \frac{\delta'}{\ell}$ for some $\delta' > 0$

$$\begin{aligned}\mathbb{E}[W_S] &\geq \exp(-2(\ell-1)\delta')(c_2\delta'/(2\ell))^{p\ell}(c_3\delta'/\ell)^{\ell} \\ &= \exp(-2(\ell-1)\delta' - (p+1)\ell\log(\ell))(c_2\delta'/(2))^{p\ell}(c_3\delta')^{\ell}\end{aligned}$$

The dominant term here is the $\exp(-(p+1)\ell\log(\ell))$. Next summing over $\mathfrak{S}(\ell)$.

$$\begin{aligned}\mathbb{E}[\mathfrak{W}] &= \sum_{S\in\mathfrak{S}(\ell)} \mathbb{E}[W_S] \\ &\geq \binom{n}{\ell} \exp(-2(\ell-1)\delta' - (p+1)\ell\log(\ell))(c_2\delta'/(2))^{p\ell}(c_3\delta')^{\ell}.\end{aligned}$$

We can lower bound the binomial coefficient $\binom{n}{\ell} \geq \left(\frac{n}{\ell}\right)^{\ell}$. Therefore, we consider the relationship between n and ℓ such that the expected number of cliques of size ℓ grows to infinity. We see that letting $\ell = n^{1/(p+2)+\epsilon'}$ for any $\epsilon' > 0$ ensures that this lower bound diverges.

Then for large n

$$\begin{aligned}
\mathbb{E}[\mathfrak{W}] &\gtrsim \left(\frac{n}{\ell}\right)^{\ell} \exp(-(p+1)\ell \log(\ell)) \\
&= \left(\frac{\ell^{(p+2)+\epsilon'}}{\ell}\right)^{\ell} \exp(-(p+1)\ell \log(\ell)) \\
&= \exp\left((p+1+\epsilon')\ell \log(\ell) - (p+1)\ell \log(\ell)\right) \\
&\to_{\ell \to \infty} \infty
\end{aligned}$$

and hence we expect the number of cliques of this size to diverge to infinity. □

C.4 Riemannian Geometry Definitions

In this section, we review some definitions of the sectional and scalar curvature as well as the volume elements. A Riemannian manifold $\mathcal{M} = (M, g)$ is a smooth manifold M equipped with a Riemannian inner product g_q on the tangent space $T_q(\mathcal{M})$ at any point $q \in \mathcal{M}$, $g_q(u, v) : T_q(\mathcal{M}) \times T_q(\mathcal{M}) \mapsto \mathbb{R}$.

This inner product can be used to define the Riemann curvature tensor at a point $q \in \mathcal{M}$ $R_q(u, v)w$, which takes 3 vectors u, v, w and returns an element of the tangent space

$$\begin{aligned}
R_q(u, v)w : &\quad T_q(\mathcal{M}) \times T_q(\mathcal{M}) \times T_q(\mathcal{M}) \mapsto T_q(\mathcal{M}) \\
R_q(u, v)w &:= [\nabla_u, \nabla_v]w - \nabla_{[u,v]}w
\end{aligned}$$

where $[u, v]$ is the lie bracket of vector fields and $[\nabla_u, \nabla_v]$ is the commuter of differential operators. The Riemann curvature tensor can be used to define our main quantity of interest, the sectional curvature at a point $\kappa_q(u, v) : \times T_q(\mathcal{M}) \times T_q(\mathcal{M}) \mapsto \mathbb{R}$. The sectional curvature takes two linearly independent elements of the tangent space and maps them to the real line.

$$\kappa_q(u, v) := \frac{g_q(R_q(u, v)v, u)}{g_q(u, u)g_q(v, v) - g_q(u, v)^2}.$$

The sectional curvature is independent of the coordinate system used, but depends only on the linear subspace spanned by u, v. Furthermore, in the canonical manifolds $\kappa_q(u, v) = \kappa$ by construction.

From the sectional curvature, we can define the scalar curvature $S(m)$,

$$S(q) := \sum_{i \neq j} \kappa(e_i, e_j)$$

where $\{e_i\}_{i=1}^p$ form an orthonormal frame for $T_q(\mathcal{M})$. We can think of the scalar curvature as an "average of sectional curvatures" across the manifold.

Next we consider the distance induced by the metric tensor. Given a smooth curve $\gamma : [a, b] \to \mathcal{M}^p$ with $\gamma(a) = q_1$ and $\gamma(b) = q_2$, the **length** $L(\gamma)$ of γ is defined by:

$$L(\gamma) = \int_a^b \sqrt{g_{\gamma(t)}\left(\dot{\gamma}(t), \dot{\gamma}(t)\right)}\, dt,$$

where $\dot{\gamma}(t)$ is the tangent vector to the curve γ at time t.

The Riemannian distance $d(q_1, q_2)$ between two points $q_1, q_2 \in \mathcal{M}^p$ is defined as:

$$d(q_1, q_2) = \inf_{\gamma} L(\gamma),$$

where the infimum is taken over all smooth curves $\gamma : [a, b] \to M$ such that $\gamma(a) = q_1$ and $\gamma(b) = q_2$.

We lastly define a volume form (also known as the Levi-Civita Tensor) via the Riemannian inner product g. If ω is a local oriented coordinate system near a point q then

$$dV := \sqrt{|\det(g)|} d\omega.$$

where g_q is the metric tensor evaluated on the basis coordinate system ω. For further details on these quantities, see Klingenberg [1995].

From this definition of a volume, we can define probability density functions on the manifold. A density function f corresponding to a measure F with support on the manifold can be defined as follows. For a set $\mathcal{X} \subset \mathcal{M}$.

$$P(X \in \mathcal{X}) = \int_{x \in \mathcal{X}} f(x) dV_x$$

See Pennec [1999] for further introduction for defining probabilities on the manifold.

C.5 Assumptions on $\mathcal{M}$

Here we verify that the Algebraic Midpoint properties, as well as locally Euclidean properties are satisfied for a complete simply connected smooth Riemannian manifolds.

If the algebraic midpoint property is satisfied for any complete metric space $\mathfrak{M}$ then by Theorem 1.8 of Gromov [2007], then $\mathfrak{M}$ is a **path metric space**. The authors follow up in discussion a list of examples of path metric spaces, which include Riemannian manifolds with boundary.

Secondly, if $\mathcal{M}^p$ has scalar curvature at point q, $S(q)$ then

$$\frac{\text{Vol}(B_{\mathcal{M}^p}(\epsilon, q))}{\text{Vol}(B_{\mathbb{E}^p}(\epsilon, q))} = 1 - \frac{S(q)}{6(p+2)}\epsilon^2 + o(\epsilon^3)$$

by Theorem 3.98 of Gallot et al. [2004]. Since this holds, then for a latent metric which is generated by distances on a Riemannian manifold, the locally Euclidean volume property will hold.

C.6 Graph Statistics From Simulations

Columns denote the scale factor used in the simulations.

C.7 Values of Tuning Parameter $C_\triangle$ Used in Simulations And Applications

Here we provide details on the choice of $C_\triangle$ used in various simulations.

Table C.1: $\kappa = -2$ graph statistics summary.

Scale (ρ)	0.7	1	2
Edge.fraction	0.016 (0.001)	0.012 (0.001)	0.007 (0.001)
Max.Degree	341.3 (23.596)	412.29 (26.364)	656.63 (33.954)
Mean.Degree	56.726 (5.279)	58.538 (5.351)	73.975 (5.448)
Distinct.Cliques$\geq \ell$	78.665 (6.955)	53.675 (5.427)	43.515 (4.41)
Max.Clique.Size	24.215 (4.985)	28.725 (5.494)	NaN (NA)
Mean.Degree.Centrality	0.155 (0.009)	0.134 (0.007)	0.107 (0.005)

Table C.2: $\kappa = -1$ graph statistics summary.

Scale (ρ)	0.7	1	2
Edge.fraction	0.017 (0.002)	0.012 (0.001)	0.008 (0.001)
Max.Degree	343.69 (25.21)	416.825 (28.796)	661.12 (32.257)
Mean.Degree	58.943 (5.354)	61.168 (5.636)	78.239 (5.6)
Distinct.Cliques$\geq \ell$	80.14 (6.488)	54.965 (5.404)	42.32 (5.027)
Max.Clique.Size	24.005 (5.456)	28.735 (6)	NaN (NA)
Mean.Degree.Centrality	0.159 (0.008)	0.137 (0.007)	0.112 (0.005)

Section (Figure)	$C_\triangle$
4.3.1 (Figure 4.7)	1.5
4.4.1 (Figure 4.8)	1.2
4.4.2 (Figure 4.9)	1.2
4.4.3 (Figure 4.10)	1.3
4.5.1 (Figure 4.12)	1.4
4.5.1 (Figure 4.13)	1.4

Table C.3: $\kappa = -0.5$ graph statistics summary.

Scale (ρ)	0.7	1	2
Edge.fraction	0.017 (0.001)	0.013 (0.001)	0.008 (0.001)
Max.Degree	340.785 (21.648)	413.115 (24.818)	656.07 (39.433)
Mean.Degree	59.658 (4.776)	62.67 (4.989)	79.873 (6.566)
Distinct.Cliques$\geq \ell$	80.205 (6.493)	55.155 (5.288)	39.95 (5.289)
Max.Clique.Size	23.62 (5.216)	29.52 (6.073)	NaN (NA)
Mean.Degree.Centrality	0.161 (0.009)	0.141 (0.007)	0.114 (0.005)

Table C.4: $\kappa = 0$ graph statistics summary.

Scale (ρ)	0.7	1	2
Edge.fraction	0.012 (0.001)	0.008 (0)	0.005 (0)
Max.Degree	205.38 (16.619)	242.88 (18.178)	371.805 (24.336)
Mean.Degree	40.815 (2.626)	41.759 (2.415)	52.465 (2.813)
Distinct.Cliques$\geq \ell$	70.2 (13.19)	46.84 (6.057)	41.54 (4.576)
Max.Clique.Size	23.735 (5.244)	28.49 (5.421)	41.63 (7.913)
Mean.Degree.Centrality	0.169 (0.017)	0.149 (0.013)	0.125 (0.009)

Table C.5: $\kappa = 0.5$ graph statistics summary.

Scale (ρ)	0.7	1	2
Edge.fraction	0.017 (0.001)	0.013 (0.001)	0.008 (0)
Max.Degree	267.08 (17.432)	318.615 (18.062)	488.545 (25.046)
Mean.Degree	61.311 (2.803)	63.539 (2.96)	80.373 (3.454)
Distinct.Cliques$\geq \ell$	87.945 (12.778)	55.385 (7.343)	39.535 (5.591)
Max.Clique.Size	23.595 (4.844)	28.45 (6.158)	39.925 (7.861)
Mean.Degree.Centrality	0.21 (0.016)	0.183 (0.012)	0.153 (0.008)

Table C.6: $\kappa = 1$ graph statistics summary.

Scale (ρ)	0.7	1	2
Edge.fraction	0.025 (0.001)	0.018 (0.001)	0.012 (0)
Max.Degree	351.89 (16.965)	421.36 (18.32)	652.265 (26.131)
Mean.Degree	88.331 (3.89)	91.901 (3.805)	117.569 (4.089)
Distinct.Cliques$\geq \ell$	91.71 (30.245)	64.23 (17.006)	41.765 (4.761)
Max.Clique.Size	24.07 (5.334)	29.29 (6.259)	40.52 (7.969)
Mean.Degree.Centrality	0.241 (0.012)	0.21 (0.01)	0.174 (0.007)

C.8 Additional Miscellanea

C.8.1 Embeddings and Graph Distances

Our focus on this paper is the estimation of curvature of latent spaces, however, one may consider a similar problem of embedding the graph distances (D_G), (i.e. shortest path

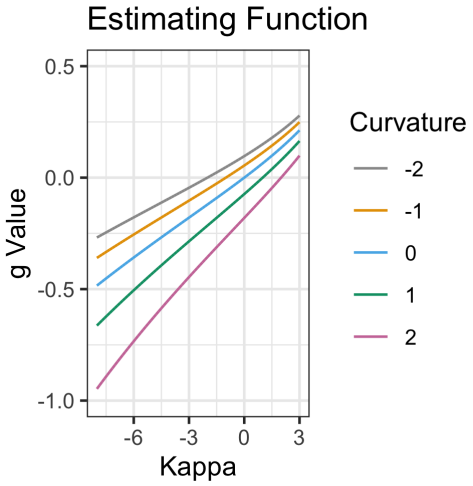


Figure C.1: Example of estimating functions g as a function of κ.

distances) as in Gu et al. [2018]. This is a distinct, but related problem that we can study using our estimating equation for curvature. Firstly, we note that using a set of graph distances will generally allow for the formation of many good quality midpoints, as any chain of 3 node forms a midpoint set. Our method may be useful here for identifying nodes who's distances may not be preserved well using a standard embedding in a space of constant curvature. We leave this possible extension of our method as future work.

C.8.2 Smoothness of The Estimating Equation

Note: We moved this section to the main text. To highlight the smoothness of our estimating function we plot a set of examples. For a unit equilateral triangle, we compute the corresponding midpoint distance d_{xm} for each curvature space with $\kappa \in \{-2, -1, 0, 1, 2\}$. We see in Figure C.1 that our proposed estimating equation is differentiable around the solution with non-zero derivative, allowing one to identify the curvature from the $\kappa : g(\kappa, d) = 0$.