Journal of Machine Learning Research 26 (2025) 1-99 Submitted 10/23; Revised 11/24; Published 1/25

Estimating Network-Mediated Causal Effects via Principal Components Network Regression

Alex Hayes ALEX.HAYES@WISC.EDU
Department of Statistics
University of Wisconsin–Madison
Madison, WI, USA

Mark M. Fredrickson MFREDRIC@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI, USA

Keith Levin KDLEVIN@WISC.EDU
Department of Statistics
University of Wisconsin–Madison
Madison, WI, USA

Editor: Ilya Shpitser

Abstract

We develop a method to decompose causal effects on a social network into an indirect effect mediated by the network, and a direct effect independent of the social network. To handle the complexity of network structures, we assume that latent social groups act as causal mediators. We develop principal components network regression models to differentiate the social effect from the non-social effect. Fitting the regression models is as simple as principal components analysis followed by ordinary least squares estimation. We prove asymptotic theory for regression coefficients from this procedure and show that it is widely applicable, allowing for a variety of distributions on the regression errors and network edges. We carefully characterize the counterfactual assumptions necessary to use the regression models for causal inference, and show that current approaches to causal network regression may result in over-control bias. The method is very general, so that it is applicable to many types of structured data beyond social networks, such as text, areal data, psychometrics, images and omics.

Keywords: causal mediation, latent mediators, network regression, principal components regression, random dot product graph, spectral embedding

1. Introduction

Recent years have seen a concerted effort to study causal effects on networks, motivated by striking claims about contagions in social networks (Christakis and Fowler, 2007). One of the key ideas to emerge from this push is the need to account for clustering in networks (Shalizi and Thomas, 2011). Sociologists have long known that people in social networks are mostly connected to other people like themselves, which is often expressed informally as "birds of a feather flock together," and more formally called "homophily". To identify and estimate causal effects in social settings, it is thus fundamental to model how social

©2025 Alex Hayes, Mark M. Fredrickson, and Keith Levin.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/23-1317.html.

groups form in networks, as well any downstream effects of social group membership. This is challenging, as social groups in a network are typically unobserved.

To account for unobserved social structure in networks, Shalizi and Thomas (2011) proposed using latent space models, where each node has an embedding that determines its propensity to connect with other nodes. More recently, McFowland and Shalizi (2021) showed that certain types of causal effects on networks can be estimated by controlling for latent node embeddings in linear network regression models. There is now a rapidly growing literature investigating how embeddings can be used for causal inference on networks (Paul et al., 2022a; Veitch et al., 2019, 2020; Louizos et al., 2017; Guo et al., 2020; Chu et al., 2021; Guo et al., 2020; Cristali and Veitch, 2022; Chen et al., 2022; Ogburn et al., 2022; Liu and Tchetgen Tchetgen, 2022; O'Malley et al., 2014; Leung, 2019; Egami and Tchetgen Tchetgen, 2021; Ogburn et al., 2022, 2020).

We make two contributions to this literature. First, we develop broad semi-parametric theory for network regression, showing how to incorporate node embeddings into linear regression. These network regression models are useful for causal inference on networks, as they allow practitioners to account for homophily, but they are also of substantial independent interest as an observation model. Similar models have previously appeared in Li et al. (2019); Le and Li (2022); Fosdick and Hoff (2015); He and Hoff (2019) and in concurrent work Nath et al. (2023); Chang and Paul (2024). Compared to these approaches, we require substantially weaker assumptions on the edge distribution in the network and the error distribution in the regression models. We allow weighted (i.e., real-valued) edges and allow edges to be observed with noise. Edges in the network and regression errors can be sampled from any distribution that satisfies a sub-gamma tail bound. Further, our regression models allow for heteroscedastic errors. Altogether, our results show estimating network embeddings with principal components analysis, together with ordinary least squares to estimate regression coefficients, is applicable under broad semi-parametric conditions. This contrasts with previous approaches that have largely focused on specific parametric models.

Our second contribution is to characterize how latent node-level variables (i.e., embeddings) can act as causal mediators (Imai et al., 2010). To date, causal methods for network data have mainly focused on homophily as a confounding factor. In contrast, we articulate the causal mechanisms involved when treatments influence latent social group membership. This is perhaps best illustrated by an example. Suppose we are interested in understanding and reducing adolescent smoking (Di Maria et al., 2022; Michell and Amos, 1997). Understanding why adolescents smoke is challenging, as smoking is both a sexually differentiated behavior and a social behavior. To study the effect of sex on adolescent smoking, we decompose the effect of sex into a direct effect independent of the network and an indirect effect that operates through the network. For instance, sex may directly cause higher cigarette consumption through sexed expectations about smoking. Sex may also have an indirect effect mediated by the network: adolescents may prefer to form sex-homophilous friendships, and social norms about the acceptability of smoking may vary across friend groups. That is, sex may influence an adolescent's social circumstances, which in turns causes a change in smoking behavior. These effects correspond to natural direct and indirect effects, specialized to the network setting by treating latent community memberships as mediators.

To estimate network-mediated effects, we propose an approach based on a generalization of the random dot product graph (Athreya et al., 2018; Levin et al., 2022) and

principal components network regression (Le and Li, 2022; Cai et al., 2021; Paul et al., 2022a; McFowland and Shalizi, 2021; Upton and Carvalho, 2017). To represent the mediating effect of the network, we embed the network into a low-dimensional Euclidean space via a singular value decomposition (Sussman et al., 2014). We then use the embeddings to develop two network regression models: (1) an outcome model that characterizes how nodal outcomes vary with nodal treatment, controls, and position in latent space; and (2) a mediator model that characterizes how latent positions vary with nodal treatment and controls. We estimate the regressions using ordinary least squares and Huber-White robust standard errors, and prove that the resulting coefficients are asymptotically normal under general semi-parametric conditions. Once estimation is complete, the coefficients from the outcome model can be used to estimate the direct effect of treatment, and the coefficients from the outcome model and the mediator model can be combined to estimate the indirect effect of treatment (VanderWeele and Vansteelandt, 2014). These estimators are essentially the well-known product-of-coefficients estimators (VanderWeele and Vansteelandt, 2014; Nguyen et al., 2021), but using the network embedding as a mediator. We anticipate that our causal estimators will be accessible to many social and natural scientists, as they are based on familiar product-of-coefficients mediation techniques. Importantly, the mediation estimands we consider in this work are distinct from peer effects such as contagion and interference, explored elsewhere in the literature (Ogburn et al., 2020, 2022). Our mediational estimands measure the effect of meso-level social structures, in contrast to more micro-level peer influence. In Section 2.5 we explore the relationship between these mechanisms.

While the product-of-coefficient estimator is familiar to many, it relies on strong functional form assumptions. We want to emphasize that our characterization of latent mediation is valuable even to readers who suspect these assumptions are too strong to hold in practice. Following McFowland and Shalizi (2021), practitioners and theorists are increasingly using linear regressions for causal inference on networks, and careful characterizations of the causal structure of these models is increasingly important (Paul et al., 2022a; McFowland and Shalizi, 2021; Veitch et al., 2019, 2020; Louizos et al., 2017; Guo et al., 2020; Chu et al., 2021; Guo et al., 2020; Cristali and Veitch, 2022; Chen et al., 2022; Ogburn et al., 2022; Liu and Tchetgen Tchetgen, 2022; O'Malley et al., 2014; Leung, 2019; Egami and Tchetgen Tchetgen, 2021; Ogburn et al., 2022, 2020). In our development of the product-of-coefficients estimator, we correct the misconception that homophily can only have a confounding effect, we describe the over-control bias that can be induced by causal misspecification, and we document a network data set where causal misspecification leads to substantially misleading results. Indeed, this work was originally motivated by a misinterpretation of coefficients in a principal components network regression model.

We expect our estimators to have significant causal and non-causal applications to outside of the network setting. Similar methods for low-rank data have been employed, for example, on spatial networks (Gilbert et al., 2021; Tiefelsdorf and Griffith, 2007; Doreian, 1981; Ord, 1975), text data (Keith et al., 2021; Veitch et al., 2020; Gerlach et al., 2018), psychometric surveys (Freier et al., 2022; Thurstone, 1947), imaging data (Zhao et al., 2020; Levin et al., 2022), and omics panels (Listgarten et al., 2010; Alter et al., 2000). Indeed, we apply our method to psychometric data, in addition to our running social network example. In data applications, we find that adolescent girls end up smoking more than adolescent boys primarily due to an indirect network effect. This suggests that public health interventions

might want to disrupt or alter social group formation. In an application to psychometrics data, we find that a meditation app reduces anxiety primarily by helping study participants defuse (i.e., step back and examine) from their emotions and by helping them feel less lonely. Our findings suggest that the meditation program might improve mental health outcomes by replacing meditation modules aimed at increasing a sense of meaning in life with additional modules targeting defusion or loneliness.

Our work is related to several extant lines of research. Past authors have heuristically proposed regression estimators for latent mediation in networks under the Hoff model (Hoff et al., 2002), albeit without asymptotic theory or much elaboration of causal mechanisms (Di Maria et al., 2022; Liu et al., 2021; Che et al., 2021). These ideas are related to work on non-parametric mediation analysis (Tchetgen Tchetgen and Shpitser, 2012; Farbmacher et al., 2022; Fulcher et al., 2020; Zheng and van der Laan, 2012) and semi-parametric methods for hidden mediators such as Cheng et al. (2022), as well as methods for proximal mediation such as Dukes et al. (2021) and Ghassami et al. (2021). Similarly, Sweet (2019); Sweet and Adhikari (2022); Guha and Rodriguez (2021); Zhao et al. (2022) consider mediation in networks, albeit treating entire networks as mediators, an approach that requires observing an entire network per unit of analysis.

Beyond causal considerations, this paper proposes semi-parametric methods for principal components network regression. Similar results have previously been established in more restrictive parametric settings. Le and Li (2022) and Paul et al. (2022a) show a related result in binary networks, under the assumption that regression errors are Gaussian, and Cai et al. (2021) considers Gaussian networks with a one-dimensional latent space. We substantially relax these assumptions to allow for latent spaces of arbitrary dimension and to permit far more general edge and regression error distributions. The regression models that we propose are related to several other forms of regression studied within the network literature, such as classic spatial and econometric interference models (Land and Deane, 1992; Manski, 1993) and regressions with network-coherence penalties (Li et al., 2019). For a review, we refer the interested reader to Le and Li (2022). Also related is a large body of work on network association testing, such as Ehrhardt and Wolfe (2019); Fredrickson and Chen (2019); Lee et al. (2019); Gao et al. (2022); Su et al. (2020).

Notation

For a matrix A, let $\|A\|, \|A\|_F$ and $\|A\|_{2,\infty}$ denote the spectral, Frobenius, and two-to-infinity norms, respectively. We write $A^\dagger$ for the Moore-Penrose pseudoinverse, $A_{i\cdot}$ for the i-th row, $A_{\cdot j}$ for the j-th column, and $\mathrm{vec}(A)$ for the column-wise vectorization of A, i.e., $\mathrm{vec}(A) = (A_{\cdot 1}^T, A_{\cdot 2}^T, \ldots, A_{\cdot n}^T)^T$ for a matrix with n columns. We use $\otimes$ to denote the Kronecker product. We write $[n]$ to denote the set $\{1, 2, \ldots, n\}$. $\mathbb{O}_d$ denotes the set of $d \times d$ orthogonal matrices. When we define a new symbol inline, we use $\equiv$. We use standard Landau notation, e.g., $\mathcal{O}(a_n)$ and $o(a_n)$ to denote growth rates, as well as the probabilistic variants $\mathcal{O}_p(a_n)$ and $o_p(a_n)$. For example, $g(n) = \mathcal{O}(f(n))$ means that for some constant $C > 0$, $|g(n)| < Cf(n)$ for all suitably large n. In proofs, C denotes a constant not depending on the number of vertices n, whose precise value may change from line to line, and occasionally within the same line.

2. The Causal Structure of Mediation in Latent Spaces

A central goal of this paper is to formalize and estimate natural direct and indirect effects as mediated by latent positions in a network. Briefly, the idea is that node-level treatments can affect formation of social groups, and that membership in social groups can further influence nodal outcomes.

2.1 Motivating Example: Social and Non-Social Elements of Teenage Smoking

Let us return to the adolescent smoking example mentioned previously, which is based on the *Teenage Friends and Lifestyle Study*, reported in Michell and West (1996), Michell and Amos (1997), Michell (1997), and Michell (2000). The *Teenage Friends and Lifestyle Study* collected three waves of survey data in a secondary school in Glasgow, beginning in January 1995. Students in the study filled out a questionnaire about their lifestyle and risk-taking behaviors, including alcohol, tobacco and drug use, and additionally were asked to list six of their friends. Beyond this quantitative data, researchers also conducted in-depth qualitative work, investigating the social dynamics of the school through focus groups, classroom observations, and interviews.

In the Glasgow data, as in prior investigations, researchers found that adolescent girls smoked more than adolescent boys. Michell and West (1996) and Michell (2000) proposed two distinct effects on smoking: one about athletic expectations, and the other about social influence. Michell (2000) proposed that athletic expectations affected smoking as follows: both boys and girls were subject to general societal pressure to smoke, but adolescent boys were subject to additional pressure to be good at sports (Michell and West, 1996). Since smoking was generally accepted as reducing athletic performance, this influenced adolescent boys to abstain from smoking in favor of pursuing athletic social capital (Michell, 2000).

The second proposed effect was a social effect. Smoking is known to be a collective behavior, where group members typically all partake or all abstain (Michell, 2000). Researchers found evidence of this form of group decision-making for various risk-taking behaviors in the Glasgow study, including tobacco consumption. This group-level decision-making led to differentiated behavior amongst adolescent boys and girls because adolescent friendships are highly sex-homophilous until the onset of puberty (Mehta and Strough, 2009). Taken all together, this meant that the effect of sex on smoking was mediated by friend group membership: adolescent friend group membership was heavily determined by sex, and smoking was heavily determined by friend group. More precisely, Michell (2000) found that smoking was mostly concentrated in friend groups composed of popular girls, unpopular students, and trouble-makers: "risk taking behaviour was heavily polarized within social categories so that, for instance, groups of individuals (and their peripherals) were in general either risk-taking or non-risk-taking [...] generally groups (and their peripherals) were either all boys or all girls."

The goal of this paper is to formalize a causal model for network-linked data that allows us to quantify and estimate causal effects like those proposed in the example above, a setting that we refer to as *homophilous mediation*. The first mechanism corresponds to a direct effect, independent of social considerations, and the second mechanism corresponds to an indirect effect, with friend group membership mediating the causal effect of sex on smoking. In the Glasgow study, it seems intuitive that sex exerted both a direct and indirect effect

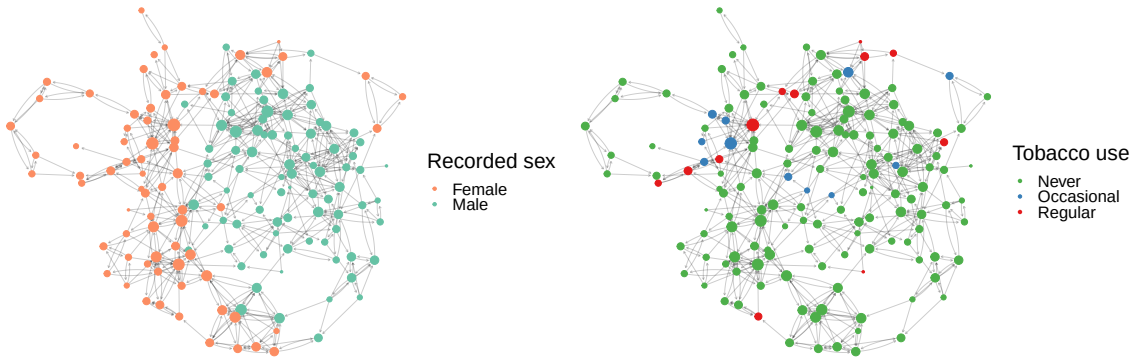


Figure 1: Directed friendships in a secondary school in Glasgow, reported in the Teenage Friends and Lifestyle Study (wave 1). Each node represents one student. An arrow from node i to node j indicates student i claimed student j as a friend. Node size is proportional to in-degree.

on tobacco consumption, driving adolescent girls to smoke more than adolescent boys. Our formalization, developed in Sections 2.2 and 2.3, also accommodates observed confounders influencing both the direct and the indirect pathways. For example, some students in the Glasgow study reported attending church, which has the potential to influence both their social group and attitudes about tobacco.

2.2 Causal Mediation in Latent Social Spaces

To formalize a model for mediated effects on a network, we begin by introducing some notation. We assume that we have a network with n nodes (corresponding to our experimental units), which we label according to the integers $[n] = \{1, 2, \ldots, n\}$. Let $T \in \{0, 1\}^n$ be a vector of observed binary treatment indicators for nodes $i \in [n]$, and let $Y \in \mathbb{R}^n$ be a vector of observed node-level outcomes. Let $X \in \mathbb{R}^{n \times d}$ be a matrix describing the unobserved "friend groups" of each node (to be formalized below), with the row vector $X_{i\cdot} \in \mathbb{R}^{1 \times d}$ encoding the friend group of node i for each $i \in [n]$. Similarly, let $C \in \mathbb{R}^{n \times p}$ be a matrix of observed confounders, with the row vector $C_{i\cdot} \in \mathbb{R}^{1 \times p}$ denoting the confounders associated with node i. Lastly, let $A_{ij} \in \mathbb{R}$ denote the strength of the friendship between node i and node j. For simplicity, our model takes friendships to be symmetric, such that $A_{ij} = A_{ji}$ for all $i, j \in [n]$, but this condition can be relaxed to allow for directed networks (see Section 5).

To define causal estimands, we introduce notation for the necessary counterfactual quantities, defined in the sense of structural causal models (Pearl, 2009).

Definition 1 *Let $Y_i(t)$ be the counterfactual value of the outcome measured for the ith node when T_i is set to t. Similarly, let $Y_i(t, x)$ be the counterfactual value of the ith outcome when T_i is set to t and $X_{i\cdot}$ is set to x, and let $X_{i\cdot}(t)$ be the counterfactual value of the mediator $X_{i\cdot}$ when T_i is set to t. The* average treatment effect*,* natural direct effect *and* natural

indirect effect *are defined, respectively, as*

$$\begin{aligned}\Psi_{\text{ate}}(t,t^*) &= \mathbb{E}[Y_i(t) - Y_i(t^*)],\\ \Psi_{\text{nde}}(t,t^*) &= \mathbb{E}[Y_i(t, X_{i\cdot}(t^*)) - Y_i(t^*, X_{i\cdot}(t^*))], \text{ and}\\ \Psi_{\text{nie}}(t,t^*) &= \mathbb{E}[Y_i(t, X_{i\cdot}(t)) - Y_i(t, X_{i\cdot}(t^*))].\end{aligned}$$

Note that the average treatment effect decomposes into the sum of the natural direct effect and the natural indirect effect: $\Psi_{\text{ate}}(t,t^*) = \Psi_{\text{nde}}(t,t^*) + \Psi_{\text{nie}}(t,t^*)$. We interpret all three estimands following Chapter 2 of VanderWeele (2015): the average treatment effect Ψ_{ate} describes how much the outcome Y_i would change on average over all nodes $i \in [n]$ if the treatment T_i were changed from $T_i = t$ to $T_i = t^*$. The natural direct effect describes how much the outcome Y_i would change on average over all nodes $i \in [n]$ if the exposure T_i were set at level $T_i = t^*$ versus $T_i = t$ but for each individual the mediator $X_{i\cdot}$ were kept at the level it would have taken for that individual, had T_i been set to t^*. The natural indirect effect describes how much the outcome Y_i would change on average over all nodes $i = [n]$ if the exposure were fixed at level $T_i = t^*$ but the mediator $X_{i\cdot}$ were changed from the level it would take under $T_i = t$ to the level it would take under $T_i = t^*$.

In slightly more plain language, we can interpret the natural direct effect as capturing the effect of the exposure on the outcome when the mediating pathway is disabled. In the smoking example discussed in Section 2.1, this would correspond to the effect of sexed expectations alone. Similarly, we can interpret the natural indirect effect as capturing the effect of the exposure on the outcome that operates by changing the mediator while keeping treatment fixed (VanderWeele, 2015). In the smoking example, this corresponds to the effect of sex on friend group, and then friend group on smoking behavior. The total effect of sex on smoking is the sum of the effects from sexed athletic expectations and friend group pressures.

In order to estimate counterfactual quantities, we must make identifying assumptions to relate unobserved counterfactual quantities to the observable data. Identification is implied by the properties of consistency, sequential ignorability, and positivity (Imai et al., 2010), which are standard sufficient conditions within the causal inference literature.

Assumption 1 (Non-parametric Identification of Natural Direct and Indirect Effects) *The random variables $(Y_i, Y_i(t,x), X_{i\cdot}, X_{i\cdot}(t), C_{i\cdot}, T_i)$ are independent over $i \in [n]$ and obey the following three properties.*

1. *Consistency:*
$$\begin{aligned}&\text{if } T_i = t, \text{ then } X_{i\cdot}(t) = X_{i\cdot} \text{ with probability 1, and}\\ &\text{if } T_i = t \text{ and } X_{i\cdot} = x, \text{ then } Y_i(t,x) = Y_i \text{ with probability 1}\end{aligned}$$

2. *Sequential ignorability:*
$$\{Y_i(t^*,x), X_{i\cdot}(t)\} \perp\!\!\!\perp T_i \mid C_{i\cdot} \quad \text{and} \quad \{Y_i(t^*,x)\} \perp\!\!\!\perp X_{i\cdot} \mid T_i = t, C_{i\cdot}$$

3. *Positivity:*
$$\begin{aligned}\mathbb{P}(x \mid T_i, C_{i\cdot}) > 0 \text{ for each } x \in \text{supp}(X_{i\cdot})\\ \mathbb{P}(t \mid C_{i\cdot}) > 0 \text{ for each } t \in \text{supp}(T_i)\end{aligned}$$

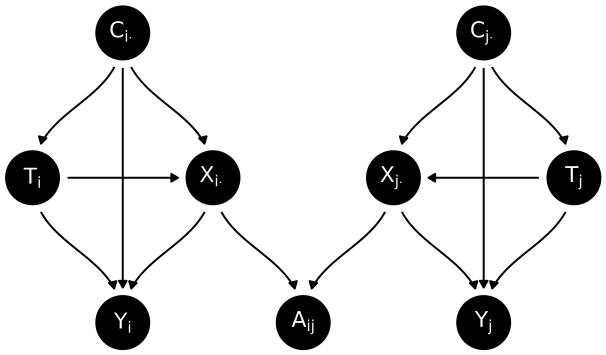


Figure 2: A directed acyclic graph (DAG) representing the causal pathways of latent mediation in a network with two nodes called i and j. Each node in the figure corresponds to a random variable, and edges indicate which random variables may cause which other random variables. We are interested in the causal effect of T_i on Y_i, as mediated by the latent position $X_{i\cdot}$.

Roughly speaking, a sufficient condition for natural direct and indirect effects to be non-parametrically identified is that the observed controls $C_{i\cdot}$ contain all confounders of the exposure-outcome ($T_i \to Y_i$), the exposure-mediator ($T_i \to X_{i\cdot}$) and the mediator-outcome ($X_{i\cdot} \to Y_i$) relationships. One structural causal model satisfying these requirements is given in Figure 2. Note that, crucially, our model differs from traditional mediation models because the mediators (i.e., the group memberships $X_{i\cdot}$) are unobserved.

Three assumptions are particularly important from a counterfactual perspective. As in tabular settings, the sequential ignorability assumption is strong and may not hold due to mediator-outcome confounding. Two other concerns are more specific to the network setting. In particular, positivity may be a problem if friend groups are highly homophilous. That is, conditional on treatment and controls, some regions of the latent space might have zero probability mass. In the context of the adolescent smoking example, this is potentially an issue, as friend groups are highly sexually homophilous. Since empirical networks often exhibit high degrees of homophily, positivity violations may present a larger challenge in network settings than in non-network settings. A third crucial assumption is that there are not peer effects such as contagion or interference, a topic we discuss in detail in Section 2.5.

2.3 Semi-Parametric Latent Space Structure

The counterfactual model we have presented thus far is a standard causal model for mediation, with the exception that we have not clarified the role of $X_{i\cdot}$, which we claimed should correspond to a latent measure of social group membership. We now fix ideas about $X_{i\cdot}$ and clarify the details of the network model by developing a statistical model for low-rank matrices. This low-rank model can be thought of as a generalization of the random dot product model (Athreya et al., 2018; Nickel, 2006; Bonato and Chung, 2007) to networks with weighted (i.e., non-binary) edges.

Describing (possibly) weighted edges requires a brief technical pre-requisite.

Definition 2 *Let Z be a mean-zero random variable with cumulant generating function $\psi_Z(t) = \log \mathbb{E}\left[e^{tZ}\right]$.*

1. *Z is ν-sub-Gaussian for $\nu > 0$ if $\psi_Z(t) \leq t^2\nu/2$ for all $t \in \mathbb{R}$.*
2. *Z is (ν, b)-sub-gamma for $\nu, b \geq 0$ if $\psi_Z(t) \leq \frac{t^2\nu}{2(1-bt)}$ and $\psi_{-Z}(t) \leq \frac{t^2\nu}{2(1-bt)}$ for all $t < 1/b$.*

The class of sub-gamma distributions is broad, and includes as special cases the Bernoulli, Poisson, Exponential, Gamma, and Gaussian distributions, as well as any sub-Gaussian or squared sub-Gaussian distribution, and all bounded distributions (see Boucheron et al., 2013, for a detailed treatment). Our primary assumption on the network structure is that the edges are sampled according to sub-gamma distributions and the network is low-rank in expectation (Boucheron et al., 2013; Tropp, 2015).

Assumption 2 (Sub-gamma network) *Let $A \in \mathbb{R}^{n\times n}$ be a random symmetric matrix, such as the adjacency matrix of an undirected graph. Let $P = \mathbb{E}[A \mid X] = XX^T$ be the expectation of A conditional on $X \in \mathbb{R}^{n\times d}$, which has independent and identically distributed rows $X_{1\cdot}, \ldots, X_{n\cdot}$. The matrix P has $\operatorname{rank}(P) = d$ and is positive semi-definite with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_n$. Conditional on X, the upper-triangular elements of $A - P$ are independent (ν_n, b_n)-sub-gamma random variables.*

This characterization of the network structure is very general and encompasses a range of popular network models as special cases. The most important example of the sub-gamma model for our purposes is the random dot product graph.

Example 1 (Random Dot Product Graph; Athreya et al. (2018)) *Let F be a distribution on $\mathbb{R}^d$ such that $x^T y \in [0, 1]$ whenever $x, y \in \operatorname{supp} F$. Draw $X_{1\cdot}, X_{2\cdot}, \ldots, X_{n\cdot}$ i.i.d. according to F, and collect these n points in the rows of $X \in \mathbb{R}^{n\times d}$. Conditional on X, the edges of graph G are generated independently, with probability of an edge $i \sim j$ given by $X_{i\cdot}X_{j\cdot}^T$. That is, conditional on X, the entries of the symmetric adjacency matrix A above the diagonal are independent with $A_{ij} \sim \operatorname{Bernoulli}(X_{i\cdot}X_{j\cdot}^T)$. Then we say that A is distributed according to a random dot product graph with latent position distribution F and write $(A, X) \sim \operatorname{RDPG}(F, n)$.*

Under the random dot product graph, each node in a network is associated with a latent vector, and these latent vectors characterize propensities to form edges with other nodes. Specifically, nodes close to each other in latent space are more likely to form connections, and nodes far apart are unlikely to form connections. When nodes cluster in the latent space, the result is that edges in the network are more likely to form between nodes with similar latent characteristics. This manifests as homophily in the resulting network.

Another particularly important sub-gamma model is the stochastic blockmodel, which is in fact a sub-model of the random dot product graph.

Example 2 (Stochastic Blockmodel) *The stochastic blockmodel (SBM; Holland et al., 1983) is a model of community membership, in which each vertex is assigned to a community (sometimes called a "block"). Conditional on assignments to communities, edges are*

generated independently between every pair of vertices in the network, and the probability of forming an edge between vertices i and j depends only on the community memberships of nodes i and j.

Let $B \in [0,1]^{d\times d}$ denote a fixed, positive semi-definite matrix of inter-block edge probabilities and let $Z_{i\cdot} \in \mathbb{R}^d$ be a vector encoding the block membership of node i. Conditional on the matrix $B \in [0,1]^{d\times d}$ and on the block memberships encoded by the rows of $Z \in \{0,1\}^{n\times d}$, the behavior of the stochastic blockmodel is characterized by

$$\mathbb{P}(A_{ij} = 1 \,|\, Z, B) = Z_{i\cdot} B Z_{j\cdot}^T.$$

The basic stochastic blockmodel, as introduced in Holland et al. (1983), forces each node to belong to exactly one block. That is, it requires $Z_{i\cdot} \in \{0,1\}^d$ to have exactly one entry equal to one. The degree-corrected stochastic blockmodel (Karrer and Newman, 2011) relaxes this restriction by giving each node a "degree-heterogeneity" parameter that encodes a node's propensity to form edges. This is equivalent to requiring $Z_{i\cdot} \in \mathbb{R}_+^d$, where $d-1$ entries of $Z_{i\cdot}$ are still zero. The overlapping stochastic blockmodel (Latouche et al., 2011) allows $Z_{i\cdot} \in \{0,1\}^d$ with no additional restrictions, such that nodes can belong to multiple blocks. The mixed membership stochastic blockmodel (Airoldi et al., 2008) restricts $Z_{i\cdot}$ to lie on the $d-1$ dimensional simplex, such that nodes can have partial membership in multiple blocks, but their total block participation must sum to one. The overlapping and mixed-membership variants may also be extended to include degree-correction, in which case $Z_{i\cdot}$ can be a nearly arbitrary vector in $\mathbb{R}_+^d$ (Jin et al., 2024; Zhang et al., 2020).

To express the stochastic blockmodel in more general sub-gamma form, take $X = ZB^{1/2}$. Thus, the latent positions X encode (1) block participation as characterized by Z, (2) degree-adjustment (i.e., popularity) captured by the row scales of Z, and (3) intra-block edge formation propensities contained in B.

Before moving on to describe how the latent positions X are related to other network covariates, some remarks about the general sub-gamma model are warranted. First, the sub-gamma model allows for edges to be observed with noise.

Example 3 (Noisily Observed Random Dot Product Graph) *Let $(\mathscr{A}, X) \sim \mathrm{RDPG}(F, n)$ and let $\{E_{ij} : 1 \le i < j \le n\}$ be independent, mean-zero sub-gamma random variables for $1 \le i \le j \le n$. For example, E_{ij} might correspond to (centered) Gaussian or Bernoulli noise. Then the network $A_{ij} = \mathscr{A}_{ij} + E_{ij}$ satisfies Assumption 2, as the sum of sub-gamma random variables remains sub-gamma.*

We note that all our results presented below can be extended to asymmetric P, rectangular P, and P with negative eigenvalues. The assumption that P is symmetric and positive semi-definitive is primarily to simplify notation, and our proofs can be extended to the general case using the techniques of Rubin-Delanchy et al. (2022) and Rohe and Zeng (2023). Thus, the sub-gamma model can be extended to handle bipartite and directed graphs (Qing and Wang, 2021; Rohe et al., 2016). Other natural extensions include models such as Gaussian mixtures with identity covariance, latent Dirichlet allocation (Rohe and Zeng, 2023; Blei et al., 2003), topic models (Gerlach et al., 2018), and psychometric factor models (Thurstone, 1947, 1934).

It is also important to note that, like the random dot product graph, the sub-gamma model considered here is subject to orthogonal non-identifiability. Since $P = XX^T = (XQ)(XQ)^T$ for any $d \times d$ orthogonal matrix Q, the latent positions X are only identifiable up to an orthogonal transformation (see Athreya et al. 2018 for further discussion). Luckily, this non-identifiability of latent positions does not influence identifiability of the natural direct and indirect effects.

Lastly, under the sub-gamma model, the latent positions X are sufficient for A. That is, the nodal covariates $T_i, C_{i\cdot}$ and Y_i do not directly influence the formation of edges of A_{ij}. The covariates T_i and $C_{i\cdot}$ *can* influence edge formation, but only via the intermediary X. Relaxing this constraint is an interesting topic for future work, but requires different estimators for the latent positions X than those we consider here (Mele et al., 2023; Binkiewicz et al., 2017).

With the network structure established, we now relate the latent positions $X_{i\cdot}$ to the node-level observations $(T_i, C_{i\cdot}, Y_i)$.

Assumption 3 (Linear Conditional Expectations) *The outcome regression functional is linear in $T_i, C_{i\cdot}$, and $X_{i\cdot}$ and the mediator regression functional is linear in T_i and $C_{i\cdot}$:*

$$\underbrace{\mathbb{E}[Y_i \,|\, T_i, C_{i\cdot}, X_{i\cdot}]}_{\mathbb{R}} = \underbrace{\beta_0}_{\mathbb{R}} + \underbrace{T_i}_{\{0,1\}} \underbrace{\beta_t}_{\mathbb{R}} + \underbrace{C_{i\cdot}}_{\mathbb{R}^{1\times p}} \underbrace{\beta_c}_{\mathbb{R}^p} + \underbrace{X_{i\cdot}}_{\mathbb{R}^{1\times d}} \underbrace{\beta_x}_{\mathbb{R}^d}, \quad \textit{(outcome model)}$$

$$\underbrace{\mathbb{E}[X_{i\cdot} \,|\, T_i, C_{i\cdot}]}_{\mathbb{R}^{1\times d}} = \underbrace{\theta_0}_{\mathbb{R}^{1\times d}} + \underbrace{T_i}_{\{0,1\}} \underbrace{\theta_t}_{\mathbb{R}^{1\times d}} + \underbrace{C_{i\cdot}}_{\mathbb{R}^{1\times p}} \underbrace{\Theta_c}_{\mathbb{R}^{p\times d}} \quad \textit{(mediator model)}$$

The columns of T, C and X must be linearly independent for regression coefficients to be identifiable. The latent dimension d and the number of nodal controls p are constants that do not vary with sample size.

In this model, β_0 and θ_0 play the role of intercept terms, while β_{t} and β_{c} encode the average associations between nodal covariates T, C and nodal outcomes Y, conditional on the effect of the latent positions X. θ_{t} and Θ_{c} describe how latent positions in the network vary with nodal covariates. In the general sub-gamma edge model that we consider here, it is difficult to provide an interpretation of β_{x} and the mediator coefficients because the latent positions X can take on several roles depending on the precise parametric sub-model under consideration (e.g., under the SBM as compared to the more general RDPG). Nonetheless, β_{x} represents the average association between latent network structure and nodal outcomes, conditional on nodal covariates. Since X is only identified up to orthogonal rotation, β_{x} and Θ are similarly only identified up to orthogonal rotation.

When we combine the statistical model entailed by Assumption 3 with the previous counterfactual assumptions entailed by Assumption 1, the regression coefficients β and Θ have known causal interpretations.

Proposition 3 (VanderWeele and Vansteelandt 2014) *Under Assumptions 1 and 3*

$$\Psi_{\mathrm{nde}}(t, t^*) = (t - t^*)\, \beta_t, \quad \textit{and} \tag{1}$$

$$\Psi_{\mathrm{nie}}(t, t^*) = (t - t^*)\, \theta_t\, \beta_x \tag{2}$$

Remark 4 *An immediate question is whether Proposition 3 is useful, given that θ_t and β_x are subject to an unknown orthogonal transformation. Luckily, non-identifiability of the regression coefficients does not impact identifiability of causal estimands. Ψ_{nde} depends only on β_t, which is fully identified. On the other hand, Ψ_{nie} is a function of θ_t and β_x, and the unknown orthogonal transformations for these estimates cancel (a result that follows immediately from Theorem 9 below). This is because Ψ_{nie} depends fundamentally on the projection of Y onto the span of $\{T, C, X\}$ and the projection of X onto the span of $\{T, C\}$, rather than the precise bases of those subspaces.*

Remark 5 *Proposition 3 identifies the natural direct and indirect effects based solely on outcome and mediator regression models, and does not require a propensity score model. It is possible to construct more complicated estimators that additionally leverage the propensity score, and these estimators can be more robust and efficient than the product-of-coefficients estimator considered here (Nguyen et al., 2021; Tchetgen Tchetgen and Shpitser, 2012). We are primarily interested in network regression in this work, and leave exploration of other estimators to future work.*

2.4 Causal Mechanisms in Latent Space

A first challenge when considering mediation in a latent space is to understand how latent mediation works in the abstract. To characterize one possible form of latent mediation, we develop some intuition in the context of degree-corrected mixed-membership stochastic blockmodels, as in Example 2. Recall that, under this model, the latent position $X_{i\cdot}$ encodes the group memberships of node i, and the "popularity" of node i (i.e., the propensity of node i to form connections with other nodes)[1]. The latent position $X_{i\cdot}$ is involved in several causal pathways.

1. The homophily-inducing pathway $T_i \to X_{i\cdot}$. Since $X_{i\cdot}$ encodes the group membership and popularity of node i, intervening on $X_{i\cdot}$ can cause node i to participant in different communities, or change its popularity, or simultaneously translate node i to new communities while also scaling its popularity. Assumption 3 implies that intervention must, on average, cause a translation in the latent space.

2. The network-formation pathway $X_{i\cdot} \to A_{i\cdot}$. Intervening on $X_{i\cdot}$ simultaneously modifies $\mathbb{P}(A_{ij} = 1 \mid X)$ for all $j \in [n]$. That is, intervening on node i's community membership changes node i's probability of connecting to every other node.

3. The social outcome effect pathway $X_{i\cdot} \to Y_i$. This encodes the idea that community membership and popularity influence outcomes.

1. More precisely, under the degree-corrected mixed-membership stochastic blockmodel, $\mathbb{P}(A_{ij} = 1 \mid Z, B) = Z_{i\cdot} B Z_{j\cdot}^T$, where $Z \in \mathbb{R}_+^{n \times d}$ is a matrix of popularity-scaled group membership weights, and B is a positive semi-definite mixing matrix. In particular, under this model, $X = ZB^{1/2}$. The $B^{1/2}$ factor in X can be slightly counter-intuitive at first. For the sake of intuition, one can conceptualize interventions as applying to the block memberships Z, while holding the mixing matrix B constant. In Proposition 19, we show that any intervention on Z satisfying the parametric constraints of Assumption 3 is equivalent to an intervention on X, which also satisfies the parametric constraints of Assumption 3. Thus, for concreteness, one can think about interventions applying directly to the block memberships Z. Under Assumption 3, interventions can also simultaneously influence B, but these interventions may feel less intuitive.

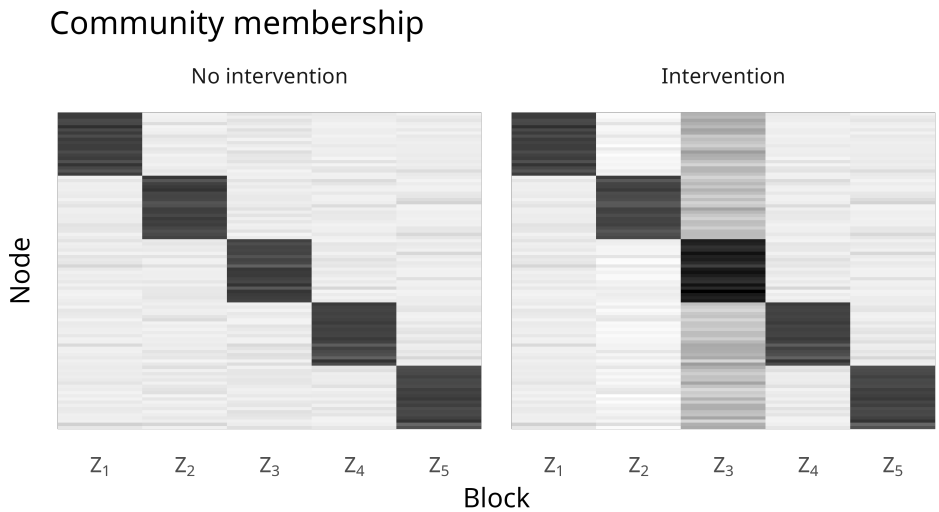
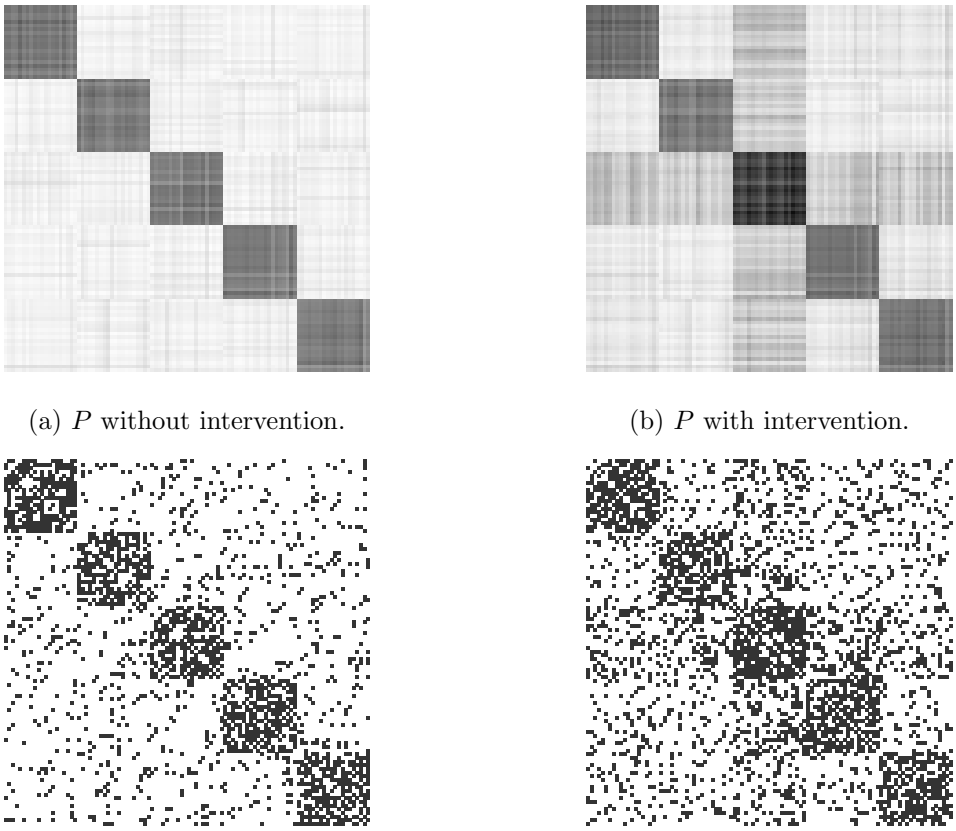


Figure 3: Block memberships in a degree-corrected mixed-membership stochastic blockmodel before and after intervention. Each row corresponds to a node in the network and each column corresponds to a block in the blockmodel. Darker colors indicated increased participation in a given block. The intervention decreases participation in the second community slightly and increases participation in the third community more dramatically. Nodes are sorted by community membership.

To develop intuition for interventions on the latent positions, consider a degree-corrected mixed membership stochastic blockmodel with five blocks, where nodes are primarily members of a single block and there is a small amount of degree heterogeneity. Interventions can increase or decrease participation in particular communities. For instance, in Figure 3, we visualize an intervention that decreases participation in the second community while increasing participation in the third community, in the latent space.

Intervention in turn alters the probability of friendship between pairs of nodes, and thus leads to different counterfactual networks via the pathway $T \to X \to A$. To understand the impact of the intervention on the network itself, Figure 4 visualizes the difference between counterfactual networks where intervention does and does not occur. In the top left panel, we see friendship probabilities when T does not impact X, and in the top right panel, we see friendship probabilities when T causes a change in X. Intervention causes all nodes to participate in the third community. The mixing matrix B is a diagonal matrix in this example, so nodes participating in the third community have some probability to connect with other nodes also participating in the third community. This means that, under intervention, all nodes in the network are more likely to connect with one another, due to mutual participation in the third block. The effect is especially pronounced for nodes that began in the third block. The bottom panels of Figure 4 show corresponding random samples from the networks conditional on latent positions X.

This synthetic example is primarily to develop intuition about the latent positions, and downstream consequences of an intervention. Even with this intuition, it may be difficult to understand what it means to intervene on a latent position. In practice, we believe that



(a) P without intervention.

(b) P with intervention.

(c) Realized network without intervention

(d) Realized network with intervention

Figure 4: A visualization of how intervention impacts the probability of edge formation P, and ultimately the realized network A. In all cases matrices as visualized as heatmaps. Each element of the heatmap corresponds to one edge in the network. Sub-figure a shows edge formation probabilities pre-intervention, and Sub-figure b shows edge formation probabilities post-intervention. Sub-figure c shows a network realized when the intervention doesn't occur, and Sub-figure d shows a network realized when the intervention does occur. Nodes are sorted by community membership.

the best way to resolve this ambiguity is by interpreting the latent positions in the context of the relevant data application. In practice, we compute estimates of the latent positions using the adjacency spectral embedding (see Section 3; one can think of A as a collection of proxy measurements that we use to infer X). With estimates of the latent positions in hand, it is often much easier to reason about causal mechanisms, because the estimated positions can be given a concrete interpretation via inspection, auxiliary data and domain knowledge. In the psychometric data application of Section 5.2, for instance, the latent positions have a very clear interpretation. To interpret the latent positions, we find that Gaussian mixture models and varimax rotation are often valuable tools (Rohe and Zeng, 2023; Rubin-Delanchy et al., 2022; Zhang et al., 2021; Priebe et al., 2019).

To summarize: we believe that the degree-corrected mixed-membership stochastic blockmodel offers compelling intuition about the structures that latent positions can represent. This intuition is helpful to determine when a latent mediation model may be applicable: for instance, in settings where one suspects that social groups act as mediators. However, there is an additional confirmatory step necessary for empirical work. It is critical to confirm that the estimated latent positions do in fact capture the hypothesized mediating constructs. Estimates of latent positions might reflect other network structure unrelated to hypothesized mediators. This is a weakness of our smoking example: we know from extensive sociology research that the adolescent social network is composed of numerous small social groups. However, we cannot confirm that the social structure observed by sociologists is the same as the social structure captured by the network principal components, an issue that we discuss in more detail in Section 5.1. In contrast, in the psychometric example of Section 5.2, the estimated latent positions are clearly measures of latent constructs of interest.

2.5 Relation to Peer Effects

The mediation mechanism that we have proposed considers how network effects manifest through group-level dynamics rather than individual-level interactions. In network settings, it is commonly assumed that there may be peer effects operating at the level of individual interactions. Two especially pertinent peer effects are contagion and interference (Hu et al., 2022; Ogburn et al., 2020). Contagion occurs when the outcome of neighbor j impacts the outcome of node i (i.e., there is a path $Y_j \to Y_i$). Interference occurs when the treatment applied to neighbor j impacts the outcome of node i (i.e., there is a path $T_j \to Y_i$). In the context of our smoking example, eliding some simultaneity issues, contagion would occur if a student's tobacco usage were influenced by their friends' tobacco usage. Interference would occur if a student's tobacco usage were influenced by their friends' sex.

Both our counterfactual assumptions and the parametric form of our outcome model assume that there are no peer effects. This restriction clearly limits the applicability of our method, but there are some caveats. First, in ongoing work, we are extending the network mediation model to allow for these peer effects. The extension is quite involved, but preliminary results and related work suggest that, at least some of the time, it is possible to estimate both mediated effects and peer effects using ordinary least squares and estimated latent positions $\widehat{X}$, as we propose here (Chang and Paul, 2024; Lee, 2002; Paul et al., 2022b; Trane, 2023). That is, one can view the network mediation model, and our characterization of mediation in a latent space, as descriptions of an important subset of causal mechanisms in networks. In many cases, these mechanisms should be jointly modelled, but developing methods to do so is a complicated technical endeavor that is outside the scope of this paper.

Part of the challenge in jointly modelling latent mediation and peer effects is that the mechanisms can be difficult or impossible to distinguish. Shalizi and Thomas (2011) showed this in a non-parametric longitudinal setting, but a very similar consideration arises in the parametric cross-sectional setting. To see this, consider the "linear-in-sums" model, one popular way to model contagion and interference (Paul et al., 2022a; McFowland and Shalizi, 2021; Egami and Tchetgen Tchetgen, 2021; Hu et al., 2022; Bramoullé et al., 2020):

$$\mathbb{E}[Y_i \,|\, T_i, C_{i\cdot}, X, A, Y_{-i}] = \beta_0 + T_i\beta_{\mathrm{t}} + C_{i\cdot}\beta_{\mathrm{c}} + X_{i\cdot}\beta_{\mathrm{x}} + \beta_{\mathrm{Ay}} \sum_{j \neq i}^{n} A_{ij} Y_j + \beta_{\mathrm{At}} \sum_{j \neq i}^{n} A_{ij} T_j.$$

The coefficient β_{Ay} encodes a contagion effect, and the coefficient β_{At} encodes an interference effect. In forthcoming work, we show that columns of the design matrix corresponding to β_{x}, β_{Ay} and β_{At} are collinear in the asymptotic limit. That is, the interference column AT and the contagion column AY of the design matrix are both contained in the column space of X in large samples. This occurs because the column spaces of A and X get closer with increasing sample size under the sub-gamma model, and thus peer effects that are linear functions of A can be equivalently represented as linear functions of X. From a parametric perspective, the coefficient β_{x} generalizes "linear-in-sums" style peer effects. In the context of our smoking example, the causal effect of belonging to a particular friend group is asymptotically indistinguishable from a contagion effect that depends on the number of friends who are smokers. Thus, the causal effects along the pathway $X_{i\cdot} \to Y_i$ can be thought of as effects of group memberships, or alternatively, as consequences of diffusions over the network. The fact that this equivalence holds only in the asymptotic limit, however, introduces a number of subtle caveats, which we describe in a forthcoming manuscript.

In addition to "linear-in-sums" style peer effect, it is also possible to consider "linear-in-means" style peer effects,

$$\mathbb{E}[Y_i \,|\, T_i, C_{i\cdot}, X, A, Y_{-i}] = \beta_0 + T_i\beta_{\text{t}} + C_{i\cdot}\beta_{\text{c}} + X_{i\cdot}\beta_{\text{x}} + \beta_{\text{Ay}} \sum_{j \in \mathcal{N}(i)} \frac{A_{ij}}{d_i} Y_j + \beta_{\text{At}} \sum_{j \in \mathcal{N}(i)} \frac{A_{ij}}{d_i} T_j$$

where $\mathcal{N}(i) = \{j \in [n] : A_{ij} = 1\}$ denotes the neighbors of node i and $d_i = \sum_j A_{ij}$ is the degree of node i (Bramoullé et al., 2009). In the linear-in-means approach, β_{x}, β_{Ay} and β_{At} are identified, even in the asymptotic limit, under some non-trivial assumptions about the network structure. Recent work has considered these models from a statistical point of view (Chang and Paul, 2024; Lee, 2002; Paul et al., 2022b; Trane, 2023), although we believe there are still substantial identifications concerns to address, from both a causal and a statistical perspective.

3. Estimation Theory for Principal Components Network Regression

Having established a model, we have several estimation targets. First, there are the network regression coefficients β and Θ from Assumption 3. Once we estimate the coefficients β and Θ, we can plug them into Equations (1) and (2) to obtain estimates of the natural direct and indirect effects, respectively.

3.1 Principal Components Network Regression

Before discussing estimation, we simplify notation by collecting the node-level covariates into $W = \begin{bmatrix} 1 & T & C \end{bmatrix} \in \mathbb{R}^{n \times (p+2)}$. Assumption 3 can then be re-written as

$$\begin{aligned} \mathbb{E}[Y_i \,|\, W_{i\cdot}, X_{i\cdot}] &= W_{i\cdot}\beta_{\text{w}} + X_{i\cdot}\beta_{\text{x}}, && \text{(outcome model)} \\ \mathbb{E}[X_{i\cdot} \,|\, W_{i\cdot}] &= W_{i\cdot}\Theta. && \text{(mediator model)} \end{aligned} \tag{3}$$

We would like to estimate $\beta_{\text{w}}, \beta_{\text{x}}$ and Θ by applying ordinary least squares regression to the vertex-level latent positions X. Unfortunately, the latent positions X are unobserved. To contend with this, we will estimate X from the observed network, then plug in $\widehat{X}$ for X

in subsequent regressions. We use the adjacency spectral embedding (ASE; Sussman et al., 2012) to achieve this estimation, but we note that several other methods are available (Xie and Xu, 2020, 2021; Wu and Xie, 2022).

Definition 6 (ASE; Sussman et al. (2014)) *Given a network with adjacency matrix A, the d-dimensional* adjacency spectral embedding *(ASE) of A is defined as*

$$\widehat{X} = \widehat{U}\widehat{S}^{1/2} \in \mathbb{R}^{n\times d},$$

where $\widehat{U}\widehat{S}\widehat{V}^T$ is the rank-d truncated singular value decomposition of A. That is, $\widehat{S} \in \mathbb{R}^{d\times d}$ is diagonal, with entries given by the d leading singular values of A, and $\widehat{U}, \widehat{V} \in \mathbb{R}^{n\times d}$ have the corresponding d orthonormal singular vectors as their columns.

The spectral embeddings $\widehat{X}$ converge to the true X uniformly over the rows of X, up to orthogonal non-identifiability (Levin et al., 2022; Lyzinski et al., 2014). We thus construct least squares estimators for the regression coefficients based on $\widehat{X}$ rather than X, as follows.

Definition 7 (Regression Point Estimators) *Define $\widehat{D} = \begin{bmatrix} W & \widehat{X} \end{bmatrix} \in \mathbb{R}^{n\times(2+p+d)}$. We estimate β_w and β_x via ordinary least squares as follows*

$$\begin{bmatrix} \widehat{\beta}_w \\ \widehat{\beta}_x \end{bmatrix} = \left(\widehat{D}^T\widehat{D}\right)^{-1}\widehat{D}^T Y.$$

Similarly, we estimate Θ via ordinary least squares as

$$\widehat{\Theta} = \left(W^T W\right)^{-1} W^T \widehat{X}.$$

Definition 8 (Regression variance estimators) *With notation as above, we define covariance estimators*

$$\widehat{\Sigma}_\beta = \widehat{A}_\beta^{-1} \cdot \widehat{B}_\beta \cdot \left(\widehat{A}_\beta^{-1}\right)^T \quad \text{and}$$

$$\widehat{\Sigma}_{\mathrm{vec}(\Theta)} = \widehat{A}_{\mathrm{vec}(\Theta)}^{-1} \cdot \widehat{B}_{\mathrm{vec}(\Theta)} \cdot \left(\widehat{A}_{\mathrm{vec}(\Theta)}^{-1}\right)^T,$$

where I_d is a $d \times d$ identity matrix, and letting $\widehat{\xi}_{i\cdot} = \widehat{X}_{i\cdot} - W_{i\cdot}\widehat{\Theta}$, we define

$$\widehat{A}_\beta = \frac{\widehat{D}^T\widehat{D}}{n}, \qquad \widehat{B}_\beta = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \widehat{D}_{i\cdot}\widehat{\beta}\right)^2 \widehat{D}_{i\cdot}^T \widehat{D}_{i\cdot},$$

$$\widehat{A}_{\mathrm{vec}(\Theta)} = \frac{I_d \otimes W^T W}{n}, \quad \text{and} \quad \widehat{B}_{\mathrm{vec}(\Theta)} = \frac{1}{n}\sum_{i=1}^n \widehat{\xi}_{i\cdot}^T \widehat{\xi}_{i\cdot} \otimes W_{i\cdot}^T W_{i\cdot}.$$

Our main technical results state that the ordinary least squares estimates based on $\widehat{X}$ converge to the same asymptotic distribution as the ordinary least squares estimates based on X. That is, the estimated regression coefficients asymptotically behavior as if we had access to the true latent positions, even though we do not. In order for plug-in estimation to be useful, the estimates based on the true latent positions X must themselves be well-behaved. Under fairly weak conditions, estimates based on the true latent positions are asymptotically normal.

Assumption 4 (Regularity Conditions for Ordinary Least Squares M-estimation) *We require standard regularity conditions for ordinary least squares estimates for both the outcome model and the mediator model. See Chapter 7 of Boos and Stefanski (2013) and Chapter 5 of van der Vaart (1998) for additional discussion.*

1. *Define $\xi = X - \mathbb{E}[X \mid W] \in \mathbb{R}^{n\times d}$ to be the matrix of errors in the mediator regression. ξ_{ij} and $\xi_{i'j}$ are independent for $i \neq i'$ and all $j \in [d]$, and that $\mathbb{E}[\xi_{ij} \mid W] = 0$ for all $i \in [n], j \in [d]$. Further,*

$$A_{\mathrm{vec}(\Theta)} \equiv \mathbb{E}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} W_{i\cdot}^T W_{i\cdot}\right] \in \mathbb{R}^{p\times p}$$

exists and is non-singular, and the matrix $B_{\mathrm{vec}(\Theta)} \in \mathbb{R}^{d\times d}$ defined according to

$$[B_{\mathrm{vec}(\Theta)}]_{j,j'} \equiv \mathbb{E}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \xi_{ij}\, W_{i\cdot}^T W_{i\cdot}\, \xi_{ij'}\right] \quad j, j' \in [d]$$

exists and has all entries finite.

2. *Define $\widehat{D} = \begin{bmatrix} W & \widehat{X} \end{bmatrix} \in \mathbb{R}^{n\times(2+p+d)}$, let $\varepsilon = Y - \mathbb{E}[Y \mid D] \in \mathbb{R}^n$ be the vector of errors in the outcome regression. The ε_i are independent, and obey $\mathbb{E}[\varepsilon_i \mid D] = 0$ for all $i \in [n]$. Further, $A_\beta \equiv \mathbb{E}\big[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} D_{i\cdot}^T D_{i\cdot}\big] \in \mathbb{R}^{(p+d)\times(p+d)}$ exists and is non-singular, and*

$$B_\beta \equiv \mathbb{E}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2\, D_{i\cdot}^T D_{i\cdot}\right]$$

exists and is finite.

For consistent estimation of β and Θ based on *known* X, the key requirement is that the residuals ε_i and ξ_{ij} have (conditional) mean zero, given the covariates D and W, respectively. Fundamentally, this means that $\mathbb{E}[Y_i \mid W_{i\cdot}, X_{i\cdot}]$ must be linear in both $W_{i\cdot}$ and $X_{i\cdot}$ and that $\mathbb{E}[X_{i\cdot} \mid W_{i\cdot}]$ must be linear in $W_{i\cdot}$. We note that the errors ε and ξ need not come from a particular distribution: any distribution satisfying the assumptions on $A_{\mathrm{vec}(\Theta)}, B_{\mathrm{vec}(\Theta)}, A_\beta$, and B_β will do. The rows of ξ must be independent of one another, but arbitrary dependence is acceptable within the rows of ξ. Once the regularity conditions in Assumption 4 are satisfied, these least-squares estimates for β and Θ are asymptotically normal about their estimands. Further, the well-known "robust" or "sandwich" covariance estimator is a consistent estimator for the covariance structures of these asymptotic distributions and the ordinary least squares estimates and covariance estimators can be used to obtain asymptotically valid confidence intervals for β and Θ.

Since we do not have access to the latent positions X, we must estimate them using the adjacency matrix, and as a result we need additional structure beyond Assumption 4.

Assumption 5 (Conditions for Two-stage Estimability) *With notation as above, we assume that*

1. *The smallest non-zero eigenvalue λ_d of P grows at a sufficiently fast rate as a function of n relative to the noise parameters ν_n and b_n. In particular,*
$$\left(\nu_n + b_n^2\right) = o\left(\frac{\lambda_d^2}{n \log^2 n}\right).$$
2. *The entries of the observation error vector ε have bounded second moments, i.e., $\max_{i\in[n]} \mathbb{E}\left[\varepsilon_i^2\right] < B$ for $B > 0$ not depending on n. Further, we assume that the error vector is such that $n^{-1}\sum_i \mathbb{E}\left[\varepsilon_i^4\right]$ is bounded as a function of n.*
3. *The rows $W_{1\cdot}, W_{2\cdot}, \ldots, W_{n\cdot}$ of W are independent. Within each row $W_{i,\cdot}$, the elements may be dependent, but are marginally sub-Gaussian with fixed, shared parameter $\sigma > 0$.*
4. *The latent positions $X_1, X_2, \ldots, X_n$ are such that $n^{-1}\sum_i \mathbb{E}\left[\|\xi_{i\cdot}\|^2\right]$ and $n^{-1}\sum_i \mathbb{E}\left[\|X_{i\cdot}\|^4\right]$ are bounded as functions of n.*

The first condition is a sufficient condition for $\widehat{X}$ to concentrate around X (see Levin et al., 2022, for further discussion). This condition primarily places requirements on the density of edges in the network, and in particular requires that the expected average degree is $\omega(\sqrt{n}\log n)$. In a random dot product graph, all degrees are asymptotically of the same order as the average degree, so our results are for dense networks. See Remark 12 for some additional comments on sparsity.

The second condition puts some additional (weak) conditions on the error in the outcome regression, beyond those already required for M-estimation. While the second condition strengthens Assumption 4, it still makes no distributional assumptions. For example, if the ε_i are independent and identically distributed sub-gamma random variables, our bounded second moment condition is satisfied. The third condition is necessary for control over $\|W\|$ in our proofs. The fourth-moment assumptions in the second and fourth conditions are needed to ensure convergence of our covariance estimator based on the spectral estimates. They ensure that error terms between this spectral-based covariance estimate and the covariance estimate based on the true (but unobserved) latent positions are suitably close. We anticipate that the moment bounds in the second, third and fourth conditions can be relaxed, but at the expense of additional proof complexity.

Our theoretical results, in their most general form, require certain growth rates on the sub-gamma parameters ν_n and b_n, and the largest and smallest non-zero eigenvalues λ_1 and λ_d of P, described in Assumption 6. For ease of presentation, we note that these rates are satisfied by random dot product graphs (and therefore stochastic blockmodels, since stochastic blockmodels are a submodel of random dot product graphs) and present Theorem 9 in the setting of a random dot product graph. Fully general bounds in terms of n, ν_n, b_n, λ_1 and λ_d along with proof details may be found in the Appendix.

Theorem 9 *If Example 1, and Assumptions 3, 4, and 5 hold, then there exists a sequence of orthogonal matrices $\{Q_n\}_{n=1}^{\infty}$ such that*
$$\sqrt{n}\,\widehat{\Sigma}_{\mathrm{vec}(\Theta)}^{-1/2}\left(\mathrm{vec}\left(\widehat{\Theta}\,Q_n^T\right) - \mathrm{vec}(\Theta)\right) \to \mathcal{N}(0, I_{pd}), \text{ and}$$
$$\sqrt{n}\,\widehat{\Sigma}_{\beta}^{-1/2}\begin{pmatrix}\widehat{\beta}_w - \beta_w \\ Q_n\,\widehat{\beta}_x - \beta_x\end{pmatrix} \to \mathcal{N}(0, I_d).$$

Theorem 9 shows that the ordinary least squares estimates based on $\widehat{X}$ are asymptotically normal about their estimands, up to orthogonal non-identifiability. Since $\widehat{X}$ only recovers X up to some unknown orthogonal transformation, $\widehat{\Theta}$ and $\widehat{\beta}_{\mathrm{x}}$ are only recovered up to this transformation, but β_{w} can be fully recovered.

Remark 10 (Specifying the latent dimension d) *Theorem 9 assumes, implicitly, that the latent dimension d is known or consistently estimated. In general, estimating the dimension of the latent space d is a challenging problem, but it can be addressed independently of network regression. Indeed, there are a variety of techniques for estimating the rank of random dot product graphs, such as those in Chen et al. (2021); Han et al. (2015); Fishkind et al. (2013); Landa et al. (2021); Li et al. (2020); Han et al. (2020), among others. Theoretically, any consistent estimator of the rank of a network suffices for Theorem 9 to hold.*

Practically, we propose that analysts use a consistent estimator of d, but also that they conduct a sensitivity analysis to investigate how much results vary with the embedding dimension d. When results are indeed sensitive to d, we recommend erring on the side of over-estimating d. It is well-known in the random dot product graph literature that over-estimating the rank of X can lead to estimates $\widehat{X}$ that are still useful for downstream tasks (Fishkind et al., 2013). We show via simulations that over-estimating d still leads to interval estimates of Ψ_{nde} and Ψ_{nie} with correct coverage in the network mediation setting. We also note that these estimates tend to vary with embedding dimension d when d is under-estimated, but they stabilize once the embedding dimension has been reached. This suggests that a reasonable way to choose the embedding dimension d is to look for a plateau in $\widehat{\Psi}_{\mathrm{nde}}$ and $\widehat{\Psi}_{\mathrm{nie}}$ as a function of d (see Section 4 and Figure 8 for details).

Remark 11 (Finite sample bias) *The noise in $\widehat{X}$ around X induces bias in the estimates of $\widehat{\beta}$ at finite sample sizes. In random dot product graphs, the coefficients $\widehat{\beta}$ will be shrunken towards zero, as the rows of $\widehat{X} - XQ$ are approximately normally distributed with mean zero (Athreya et al., 2015), such that standard results on noisily observed regressors hold. Asymptotically, however, the noise in $\widehat{X}$ around X vanishes, such that bias induced by measurement error disappears.*

Remark 12 (Sparsity) *Under Assumption 5, the average degree of a binary network must be $\omega(\sqrt{n}\log n)$, rather than the more typical $\omega(\log^c n)$. In turn, under Assumption 2, all degrees in the network are the same order, and so all degrees must be $\omega(\sqrt{n}\log n)$. This restriction to dense networks is a consequence of the proof technique, rather than a fundamental limit of the estimator: the sub-gamma bounds used in the present work are not tight when applied to networks with sparse Bernoulli edges. It is straightforward, however, to replace the sub-gamma bounds in our proofs with tighter bounds specialized to Bernoulli random variables (Lei and Rinaldo, 2015; Boucheron et al., 2013) and to thereby relax assumptions on average degree in binary networks. Given our focus on general semi-parametric results, we do not perform these calculations here. See Remark 15 of Levin et al. (2022) for further discussion.*

Remark 13 (Generalized linear models) *We expect that results similar to Theorem 9 hold for generalized linear models, and indeed for general regression M-estimators. Past work on random dot product graphs has established a functional central limit theorem for*

the latent positions $\widehat{X}$ (Tang et al., 2017, Theorem 4) and convergence of M-estimates that are functions of $\widehat{X}$ only (Athreya et al., 2021, Theorem 4). These results should extend to the sub-gamma model and regression M-estimators, and we anticipate this to be a fruitful avenue for future work. In particular, these results would enable practitioners to use many of the estimators considered in Nguyen et al. (2021) to estimate network-mediated causal effects, or alternatively may allow incorporation of spatial or network error structures (LeSage and Pace, 2009; Bramoullé et al., 2009). Similarly, such results would imply that $\widehat{X}$ could be used in place of X in the regression-based estimators proposed in VanderWeele (2015). Indeed, VanderWeele (2015) can be seen as a template for how our model could be extended to included non-linear terms or link functions.

3.2 Network Regression for Causal Estimation

The semi-parametric identification results of Proposition 3 suggest a regression estimator for the natural direct and indirect effects.

Definition 14 (Causal Point Estimators) *To estimate Ψ_{nde} and Ψ_{nie}, we combine regression coefficients from the network regression models*

$$\widehat{\Psi}_{\text{cde}} = \widehat{\Psi}_{\text{nde}} = (t - t^*)\,\widehat{\beta}_t \quad \textit{and}$$
$$\widehat{\Psi}_{\text{nie}} = (t - t^*)\,\widehat{\theta}_t\,\widehat{\beta}_x.$$

Remark 15 *When $\widehat{X} = X$, $\widehat{\Psi}_{\text{nie}}$ reduces to the multivariate product-of-coefficients estimator introduced in VanderWeele and Vansteelandt (2014). As such, there are numerous methods for sensitivity analysis that can be immediately applied to $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$ (VanderWeele, 2015, Chapter 3).*

Definition 16 (Causal Variance Estimators) *To estimate the variances of Ψ_{nde} and Ψ_{nie} in our semi-parametric setting, we combine coefficients from the network regression models:*

$$\widehat{\sigma}^2_{\text{nde}} = (t - t^*)^T \cdot \widehat{\Sigma}_{\beta_t} \cdot (t - t^*)$$

where $\widehat{\Sigma}_{\beta_t}$ denotes the element of $\widehat{\Sigma}_\beta$ corresponding to β_t. Using analogous notation, let

$$\widehat{\sigma}^2_{\text{nie}} = (t - t^*)^T \begin{bmatrix} \widehat{\beta}_x \\ \widehat{\theta}_t \end{bmatrix}^T \begin{bmatrix} \widehat{\Sigma}_{\theta_t} & 0 \\ 0 & \widehat{\Sigma}_{\beta_x} \end{bmatrix} \begin{bmatrix} \widehat{\beta}_x \\ \widehat{\theta}_t \end{bmatrix} (t - t^*).$$

As with Theorem 9, we present Theorems 17 and 18 in the setting of a random dot product graph. Fully general versions may be found in the Appendix.

Theorem 17 *In the setting of Example 1 under Assumptions 1, 3, 4 and 5,*

$$\sqrt{n\,\widehat{\sigma}^2_{\text{nde}}}\Big(\widehat{\Psi}_{\text{nde}} - \Psi_{\text{nde}}\Big) \to \mathcal{N}(0, 1).$$

The theorem follows from Definition 14 and Theorem 9 together with Slutsky's theorem and an application of the delta method. A similar distributional result holds for the natural indirect effect. Proofs for both results are given in the Appendix.

Theorem 18 *In the setting of Example 1 under Assumptions 1, 3, 4 and 5,*

$$\sqrt{n\,\widehat{\sigma}^2_{\text{nie}}}\Big(\widehat{\Psi}_{\text{nie}} - \Psi_{\text{nie}}\Big) \to \mathcal{N}(0,1).$$

The form of the variance estimator $\widehat{\sigma}^2_{\text{nie}}$ follows from an application of the delta method to the regression estimators $\widehat{\beta}$ and $\widehat{\Theta}$, which can be shown to have an asymptotically normal joint distribution via a stacked M-estimator argument (Boos and Stefanski, 2013; Nguyen et al., 2021; VanderWeele and Vansteelandt, 2014; He et al., 2024). Note that rotational non-identifiability of the regression coefficients does not impact our ability to recover Ψ_{nie}, as the unknown matrix Q cancels with a corresponding Q^T in the product of regression coefficients.

4. Simulations

We now turn to a brief exploration of our estimators' performance when applied to simulated data. In our results below, we find that our two-stage regression estimators are able to reliably recover regression coefficients and mediated effects, up to orthogonal non-identifiability where appropriate. We conduct simulations using two separate models to generate network structure, both based on the degree-corrected stochastic blockmodel.

We consider a degree-corrected SBM with d blocks, n nodes, and degree heterogeneity parameters γ sampled from a continuous uniform distribution on the interval $[1, 3]$. Block assignment is random and nodes have equal probability of assignment to all blocks. The mixing matrix B is set to 0.8 on the diagonal, and 0.03 off the diagonal, corresponding to strong assortative structure. Once the block memberships Z, degree heterogeneity parameters γ and mixing matrix B are known, we compute the latent positions X numerically based on the singular value decomposition of $\mathbb{E}[A \mid Z, \gamma, B]$.

To generate data for our simulations, we first sample a network A and latent positions X according to a degree-corrected stochastic blockmodel. Then we sample the nodal covariates W, according to one of two different models:

1. In the "uninformative" model, the nodal covariates are three-dimensional samples from a standard multivariate normal distribution, independent of all other parameters in the model. These are combined with an intercept column. One of the Gaussian columns is taken to be the treatment and the others are taken to be controls.

2. In the "informative" model, the nodal covariates are dummy-coded block membership indicators, using treatment coding and including an intercept column. The treatment T is taken to be the column corresponding to the indicator for the second block, and the controls are taken to be all other block membership indicators.

Then, we infer the implied mediator coefficients Θ via a linear regression of nodal covariates on the latent positions[2]. In the uninformative model, there is no association between W and

2. This process may seem counter-intuitive, since it gives up precise control over the mediator coefficients Θ. The upside is that we do not need to specify a generative model for the mediator regression errors ξ. Specifying ξ is challenging in binary networks where X must follow an inner product distribution to maintain $P_{ij} \in [0, 1]$.

X on average, so W is only idiosyncratically associated with X. In the informative model, W is a coarsened version of X where degree-correction information has been omitted, and so we expect a strong association between nodal covariates and latent positions.

Next we sample β from a multivariate Gaussian distribution with mean equal to the vector of all ones, and covariance equal to a diagonal matrix with $1/4$ on the diagonal. Finally, to generate the nodal outcomes, we generate errors ε from a t_5 distribution and use W, X, β and ε to produce Y satisfying the regression condition in Assumption 3. At this point, we also determine the induced direct and indirect effects based on β and Θ via Equations (1) and (2).

For each model, we sample (A, Y, W) for varying number of nodes n and latent dimensions d, and compute point estimates and confidence intervals for $\widehat{\Theta}, \widehat{\beta}, \widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$. We repeat this procedure 100 times for each combination of parameters. We focus here on the causal estimators $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$. Refer to the Appendix for further results on the consistency and finite sample bias of the regression coefficients.

In Figure 5, we consider the mean squared error of $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$. We observe that the point estimates $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$ converge to Ψ_{nde} and Ψ_{nie}, as expected per Theorem 9. In Figure 6 we observe that the proposed asymptotic confidence intervals achieve close to their nominal coverage rates in finite samples. This verifies that variance estimators accurately quantify the uncertainty in $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$, also as expected given Theorems 17 and 18. In the uninformative setting, coverage for the indirect effect is higher than the nominal rate, which is unsurprising, given that confidence intervals for the indirect effect based on the delta method can be overly conservative (He et al., 2024).

In Figure 7, we investigate the coverage of our asymptotic confidence intervals when the rank d of the network is misspecified. We see that underestimating the latent dimension dramatically degrades coverage of confidence intervals of Ψ_{nde} and Ψ_{nie}. However, when d is overestimated, confidence intervals for Ψ_{nde} and Ψ_{nie} obtain nominal coverage rates. The negative effect of underestimating d is more pronounced in the informative model, where treatment is strongly associated with latent position in the network, and weaker in the uninformative model, where treatment is weakly associated with latent position in the network. On the basis of these results, we suggest that practitioners err on the side of overestimating, rather than under-estimating, the rank d of the latent positions X. Intuitively, as d increases, $\widehat{X}$ captures more and more of the latent community structure in the network, until eventually $\widehat{X}$ captures all the latent structure in the network and the effects stabilize (see Figure 27 in the Appendix for additional simulation results in this vein).

Lastly, in Figure 8, we investigate the bias of $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$ as a function of the embedding dimension d. We observe that estimates of $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$ vary with the embedding dimension d when d is under-estimated. However, once the embedding dimension d is correctly specified, the estimates $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$ stabilize. This suggests that practitioners can use sensitivity curves (such as those in Figures 9 and 18) to estimate that embedding dimension d: in particular, they should look for the embedding dimension d that stabilizes the estimates $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$; any estimate using this embedding dimension or higher is likely to have good coverage properties.

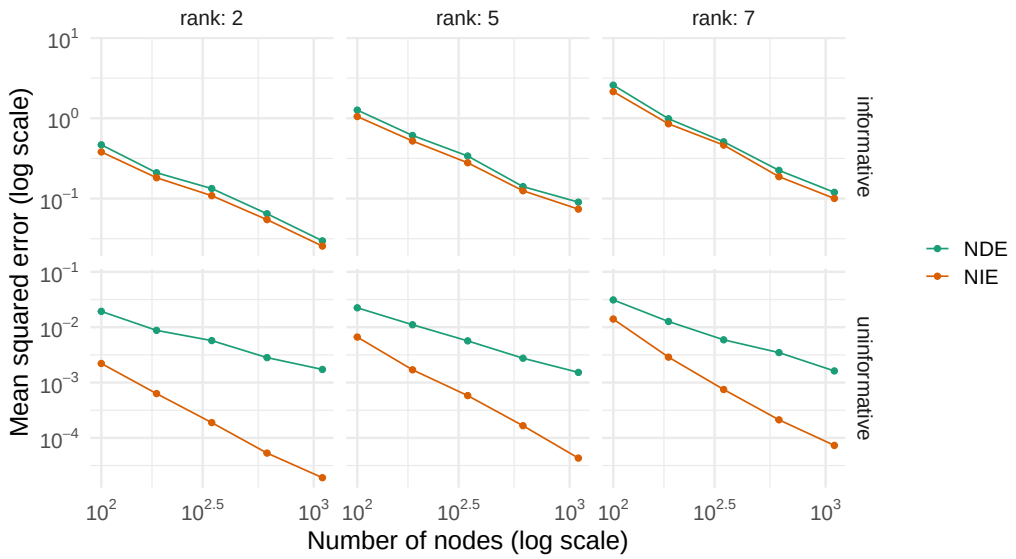


Figure 5: Convergence of $\widehat{\Psi}_{\text{nde}}$ to Ψ_{nde} and $\widehat{\Psi}_{\text{nie}}$ to Ψ_{nie}. Each panel shows the mean squared error (vertical axis, log scale) of $\widehat{\Psi}_{\text{nde}}$ (teal) and $\widehat{\Psi}_{\text{nie}}$ (orange) as a function of the number of nodes in the network (horizontal axis, log scale). Panels vary horizontally by number of latent communities (left: two blocks, middle: five block, right: seven blocks) and vertically by the simulation model (top: informative, bottom: uninformative).

5. Data Applications

We now illustrate our method by applying it to two data sets, one previously considered by Di Maria et al. (2022), and the other previously considered by Hirshberg et al. (2022, 2024).

5.1 Smoking in an Adolescent Social Network

We first revisit the Teenage Friends and Lifestyle Study described in Section 2.1, focusing on the causal effect of sex on smoking during the first wave of the study. Recall that the social network was collected by asking students "who are your best friends", and allowing them to list up to six responses. Sex and tobacco use were self-reported as nominal features with levels "Male" and "Female"; and "Never", "Occasional," and "Regular," respectively. To match the analysis of Di Maria et al. (2022), for the tobacco use measure we combined "Occasional" and "Regular" into a single level, and compared smokers with non-smokers. We treated age (continuous) and church attendance (nominal) as possible confounders.

We began by computing the adjacency spectral embedding of the social network A. In the Glasgow data, the social network is directed: an edge $i \to j$ indicates that student i listed student j as friend. This directedness means that students have two distinct co-embeddings corresponding to their propensity to send out-edges and receive in-edges. Letting $\widehat{A} \approx \widehat{U}\widehat{S}\widehat{V}^T$ be the truncated singular value decomposition of A, the left co-embedding $\widehat{X} = \widehat{U}\widehat{S}^{1/2}$ describes how students in the network send edges (i.e., claim friends), and the right

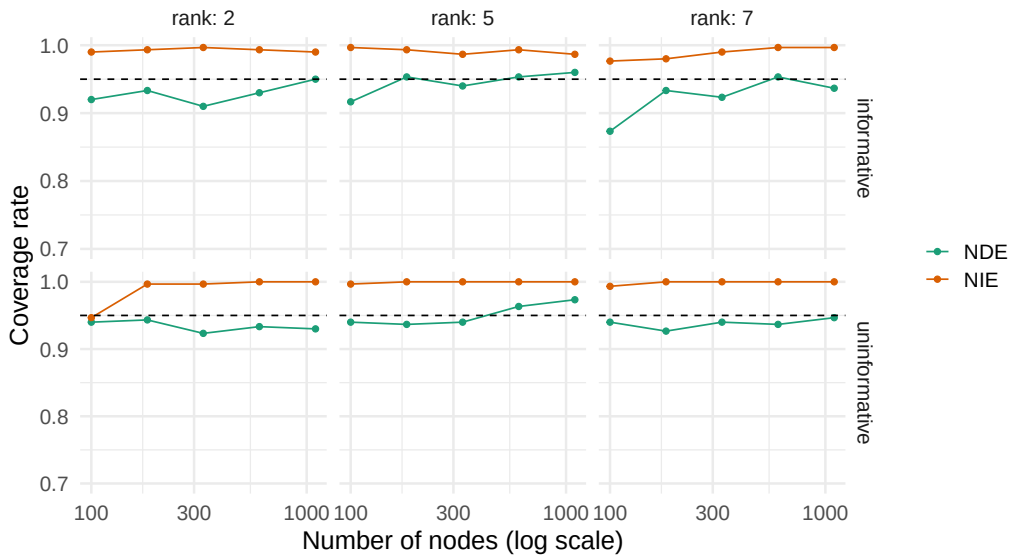


Figure 6: Finite sample coverage of asymptotic confidence intervals for Ψ_{nde} and Ψ_{nie}. Each panel shows coverage (vertical axis) of Ψ_{nde} (teal) and Ψ_{nie} (orange) as a function of the number of nodes in the network (horizontal axis, log scale). The dashed horizontal line denotes the nominal coverage rate of 95%. Panels vary horizontally by number of latent communities (left: two blocks, middle: five block, right: seven blocks) and vertically by the simulation model (top: informative, bottom: uninformative).

co-embedding $\widehat{F} \equiv \widehat{V}\widehat{S}^{1/2}$ describes how students receive edges (i.e., are claimed as friends). Here we report results using the right co-embeddings $\widehat{F}$. We did not select any particular dimension d for the latent space. Instead, we repeated our analysis for many values of d, to investigate the sensitivity of our results to the dimension of the latent space (see Remark 10 and the simulation study in Section 4 for additional commentary).

Once we obtained embeddings $\widehat{F}$ via the singular value decomposition, we performed two ordinary least squares regressions. Using the formula notation of Wilkinson and Rogers (1973) to specify the design and outcome matrices, we obtained least squares estimates for the following specifications:

$$\texttt{smoking} \sim \texttt{sex + age + church + Fhat}$$
$$\texttt{Fhat} \sim \texttt{sex + age + church}.$$

We then combined the regression coefficients (and covariances) per Definitions 14 and 16 to obtain point and interval estimates for the natural direct and indirect effects of sex on tobacco use. We visualized these results as a function of the embedding dimension d in Figure 9.

The estimates of $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$ stabilized as a function of the embedding dimension around $d = 12$, and the qualitative interpretation of the results is effectively the same for all $d \geq 12$. Since over-estimating d results in less estimation error than under-estimating d (see Remark 10 and simulation results in Section 4), we first interpreted interval estimates

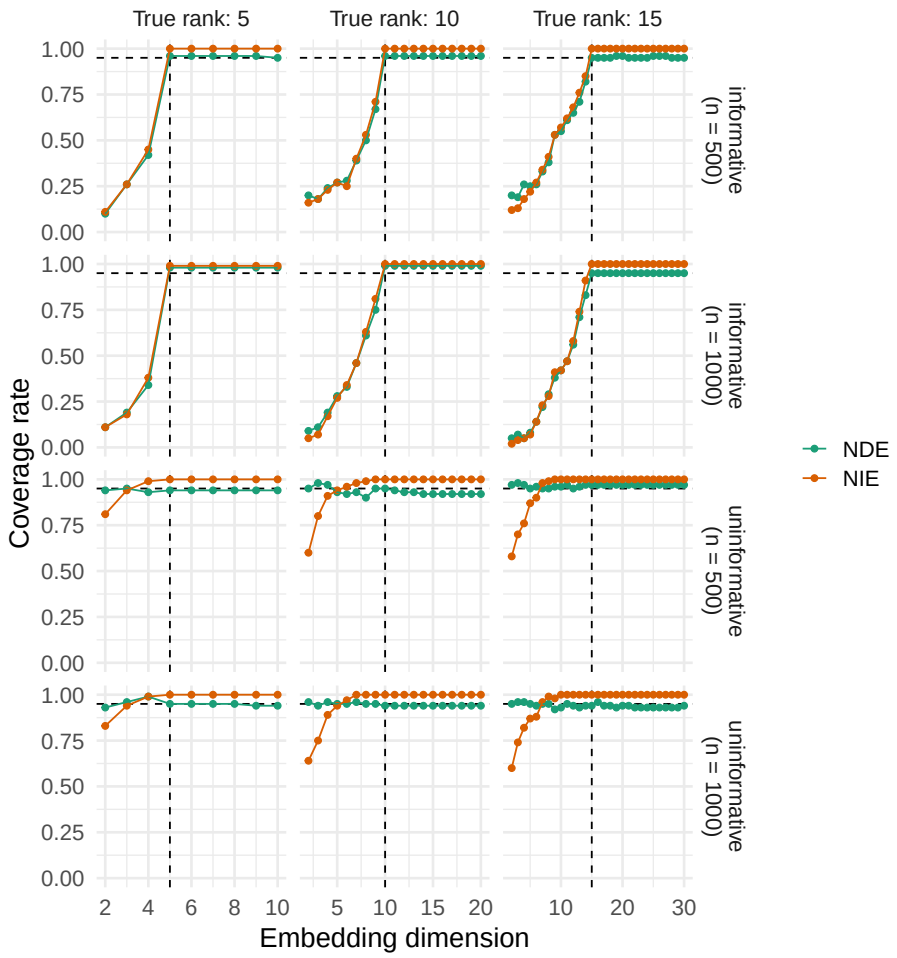


Figure 7: Coverage of confidence intervals for Ψ_{nde} and Ψ_{nie} when the dimension d is misspecified. Each panel shows coverage (vertical axis) of Ψ_{nde} (teal) and Ψ_{nie} (orange) as a function of the embedding dimension d (horizontal axis). The dashed horizontal line denotes the nominal coverage rate of 95% and the dashed vertical line denotes the true latent dimension. Panels vary horizontally by number of latent communities (left: five, middle: ten, right: fifteen) and vertically by the simulation model and number of nodes in the network.

when the latent dimension of the network is $d = 15$, under the assumption that $d = 15$ is correctly specified. For this latent dimension, we estimated a 95% CI for the direct effect of male sex, relative to female sex, to be $(-0.14, 0.17)$ and a 95% CI for the indirect effect to be $(-0.28, -0.04)$. That is, the estimates are consistent with no direct effect of sex on probability of tobacco usage. Nonetheless, there is substantial uncertainty in the estimate of the direct effect: the data is consistent with a direct effect in either direction of up to roughly 0.15. In contrast, the estimates are consistent with a negative indirect effect, though the scale of this effect is fairly uncertain.

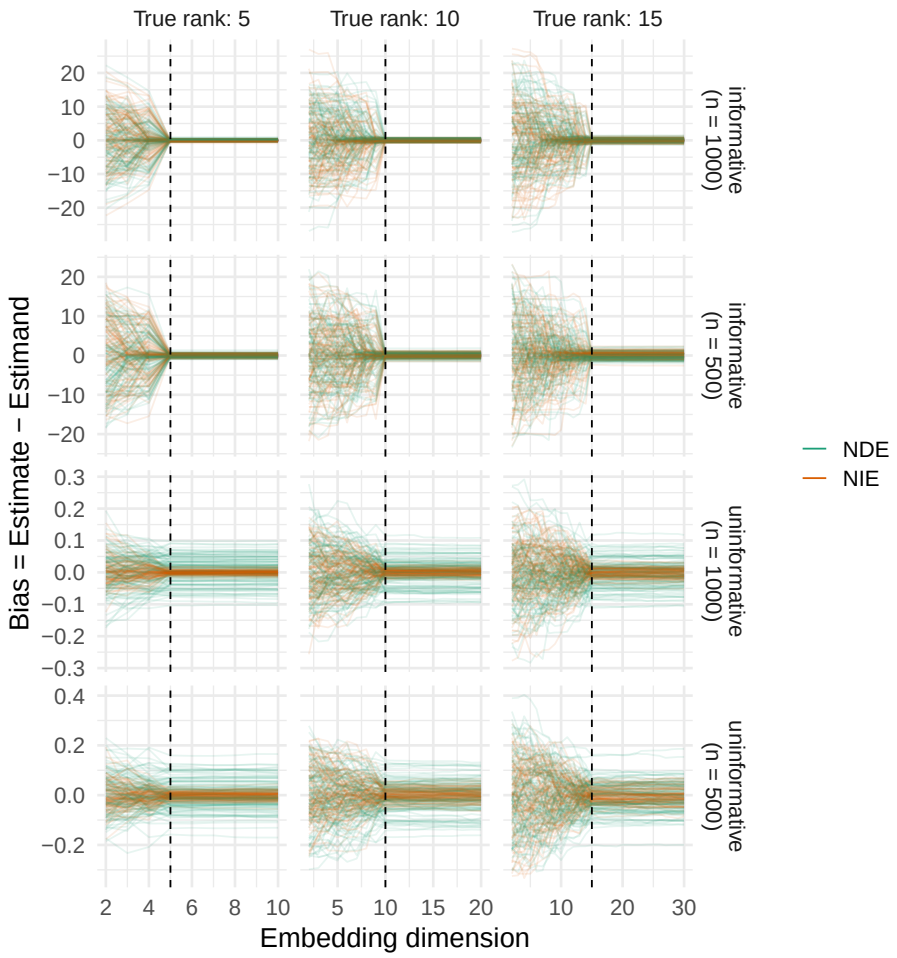


Figure 8: Stability of point estimates for Ψ_{nde} and Ψ_{nie} when the dimension d is misspecified. Each panel shows bias (vertical axis) of $\widehat{\Psi}_{\text{nde}}$ (teal) and $\widehat{\Psi}_{\text{nie}}$ (orange) as a function of the embedding dimension d (horizontal axis). The dashed vertical line denotes the true latent dimension. Panels vary horizontally by number of latent communities (left: five, middle: ten, right: fifteen) and vertically by the simulation model and number of nodes in the network.

There are, however, several reasons to be cautious about these estimates. First and foremost, we did not have auxiliary information about the social network that allowed us to directly interpret the embeddings $\widehat{F}$. This potentially leads to an issue with ill-defined interventions on F, as we discussed in Section 2.4. While substantial sociological research confirms the presence of meaningful social groups in the social network (Michell, 2000), as well as the fact that smoking predominantly occurs in majority female social groups (Michell, 2000, 1997), we could not verify that the social groups observed by sociologists match the social groups encoded by $\widehat{F}$. We must hope that low-rank structure is an appropriate way to capture these social groups. There is a second issue, namely that we have no particularly

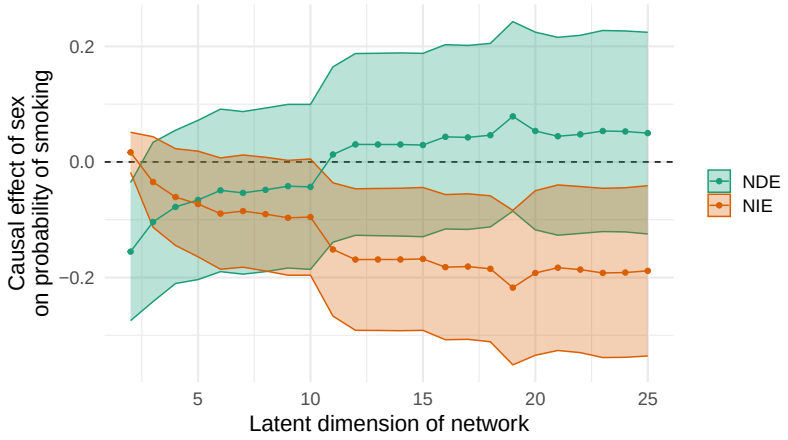


Figure 9: Estimated direct (teal) and indirect (orange) effects of sex on tobacco usage in the Glasgow social network. The estimated effects (vertical axis) vary with the dimension d of the latent space (horizontal axis), and are adjusted for possible confounding by age and church attendance. Positive values indicate a greater propensity for adolescent boys to smoke, while negative values indicate a greater propensity for adolescent girls to smoke.

compelling way to adjudicate between the sending co-embeddings $\widehat{X}$ and the receiving co-embeddings $\widehat{F}$. Our choice of $\widehat{F}$ is based primarily on folklore and personal experience that, in social networks, the receiving co-embeddings $\widehat{F}$ are more informative than the sending co-embeddings $\widehat{X}$. A sensitivity analysis using $\widehat{X}$ in place of $\widehat{F}$ yields smaller estimates of the indirect effect, which are not statistical distinguishable for zero (Figure 16 in the Appendix). A third and final reason to be cautious about this analysis is that the positivity assumption may be violated (Figure 15, also in the Appendix).

In light of these considerations, we consider this analysis to be illustrative. That said, the results in Figure 9 do align with previous sociological analyses, as well as the results obtained in Di Maria et al. (2022), who used a related estimator to conclude that "the probability of smoking tobacco regularly is higher for 13-year-old girls than for boys. In contrast, the total indirect effect for girls is negative, [such that] the effect of gender on the chance of smoking reduces through friendship relationships." Collectively, these analyses are suggestive of potential public health interventions, which could be further investigated. To reduce smoking, a public health intervention could focus on the indirect causal pathway, and could intervene on either the friend group formation process, or localized smoking within friend groups. For instance, students who smoke could be encouraged to connect socially with students who do not smoke. Alternatively, public health campaigns could focus on locating social groups where smoking is prevalent, and then performing more resource intensive interventions on those social groups.

5.2 Psychological Mediators of Anxiety in a Randomized Controlled Trial on Meditation

We next use our method to re-analyze data from a randomized controlled trial of a smartphone-based well-being training called the Healthy Minds Program, originally reported in Hirsh-

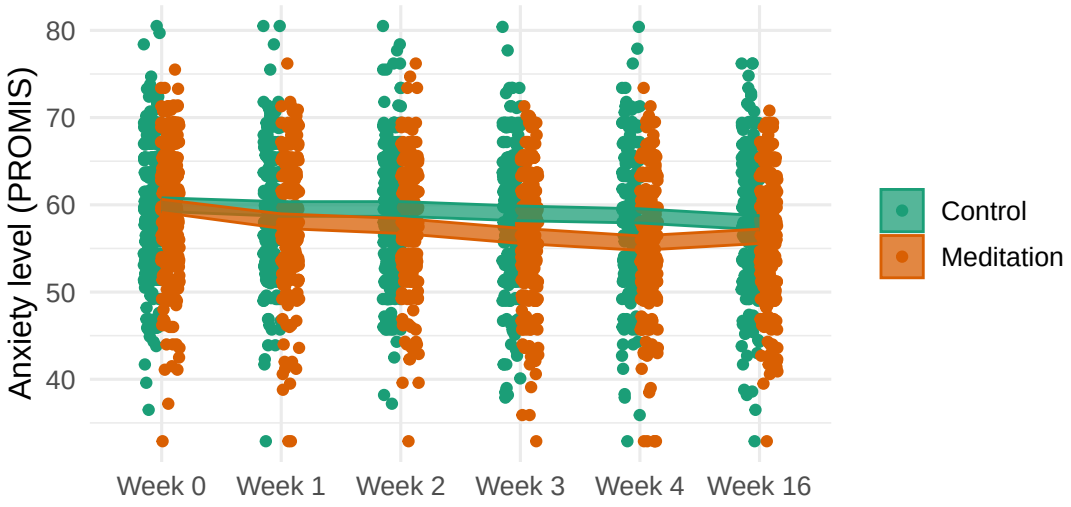


Figure 10: Anxiety levels (measured by the computer adaptive PROMIS test) for intervention and control groups in the Healthy Minds Study. Ribbons represent 95% confidence intervals for mean anxiety level in each group at each time point.

berg et al. (2022). We call this trial the Healthy Minds trial for convenience, but note that numerous trials have been conducted based on the Healthy Minds smartphone app and methodology.

In the trial, 662 adults were randomly assigned to a four-week meditation program or a control condition. During the intervention, participants were surveyed on a weekly basis to assess their psychological well-being (psychological distress, anxiety and depression) and four anticipated psychological mediators of well-being (mindful action, loneliness, cognitive defusion and purpose). Participants were also surveyed three months after the end of the intervention period.

Meditation based interventions are known to improve anxiety and depression, an observation empirically verifiable in the Healthy Minds data itself (see Figure 10). Further, there are theoretical reasons to believe that meditation can improve mindful action, loneliness, cognitive defusion and purpose, and that improvements in these dimensions can reduce depression and anxiety. In the four week long intervention program, one week is devoted to improving each of the hypothesized psychological mediators. The main goal of the Healthy Minds study was to investigate these mechanisms, and to improve knowledge of psychological mechanisms in order to design more effective interventions. We re-analyzed the data with this same goal in mind, focusing on the anxiety outcome at the end of the four week intervention. Following the original analysis, we control age and sex as potential confounders.

While the Healthy Minds data at first glance seems unrelated to social networks, there is a close connection to network data, as psychological constructs are typically considered latent constructs that must be measured via surveys followed by factor analysis (Rohe and Zeng, 2023). In the Healthy Minds study, mindful action was measured via the Five Facet Mindfulness Questionnaire Act with Awareness subscale (8 questions), loneliness via the

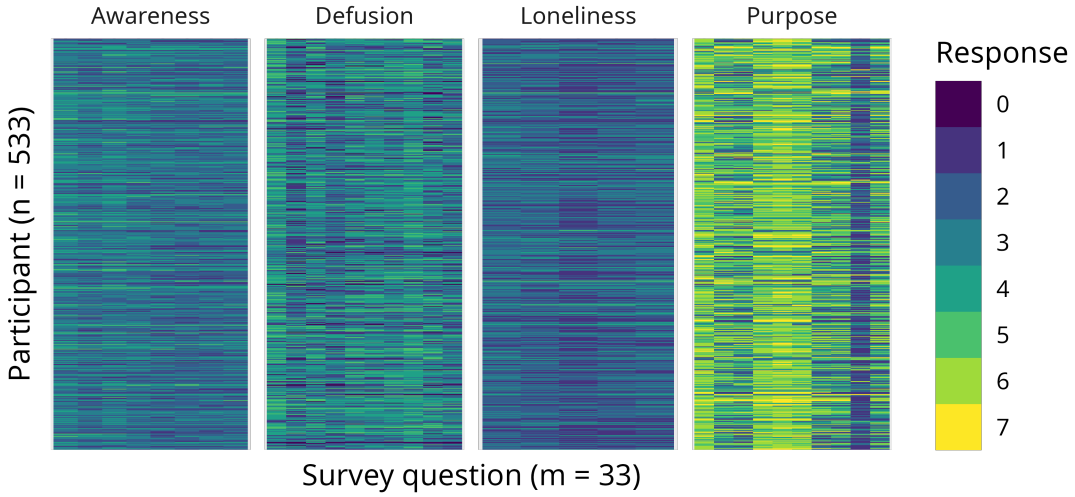


Figure 11: Survey responses for mediator measures at the end of four weeks.

NIH Toolbox Loneliness Questionnaire (5 questions), defusion via the Drexel Defusion Scale (10 questions), and purpose via the Life Questionnaire Presence subscale (10 questions). We can represent the survey responses as a bipartite network with adjacency matrix $A \in \mathbb{R}^{533\times 33}$, where A_{ij} denotes participant i's response to survey question j (for convenience, we only consider the 533 study participants who responded to all survey questions at the end of the intervention period). The survey questions were validated as measures of the corresponding latent constructs at the time that the surveys were developed.

Since A is rectangular, we compute a decomposition $A = \widehat{X}\widehat{F}^T$, where $\widehat{X} = \widehat{U}\widehat{S}^{1/2}$ and $\widehat{F} = \widehat{V}\widehat{S}^{1/2}$. Since the rows of A correspond to participants and the columns of A correspond to survey items, the left co-embeddings $\widehat{X}$ describe participants, and the right co-embeddings $\widehat{F}$ describe survey items. When the dot product of $\widehat{X}_{i\cdot}\widehat{F}_{j\cdot}^T$ is large, that indicates that participant i is expected to give a large response (e.g., "absolutely agree" rather than "neither agree nor disagree") to survey item j. When participants i and i' have embeddings $\widehat{X}_{i\cdot}$ and $\widehat{X}_{i'\cdot}$ that are close to each other, this indicates that they tended to respond to survey items in a similar manner. When survey items j and j' have embeddings $\widehat{F}_{j\cdot}$ and $\widehat{F}_{j'\cdot}$ that are close to each other, this means that participants responded to questions j and j' in a similar manner.

Our hope is that $\widehat{F}$ captures the hypothesized mediating constructs. Investigating if this is the case complicated by the fact that $\widehat{X}$ and $\widehat{F}$ are both subject to orthogonal non-identifiability, since $A = XQQ^TF^T$ for any orthogonal Q. To interpret the question embeddings, we varimax rotate the right co-embeddings, as described in Rohe and Zeng (2023). That is, we compute a varimax rotation R based on the unscaled right singular vectors V and then take $\widehat{X} = \widehat{U}\widehat{S}R$ and $\widehat{F} = R^T\widehat{V}^T$. Under the assumption that the embeddings are leptokurtic (i.e., more skewed than a Gaussian), the varimax rotated embeddings are identified up to column permutations and sign flips, making them much easier to interpret. We visualize the varimax-rotated results in Figure 12, where we see that a five dimensional embedding yields highly interpretable latent factors, which explain 70% of the

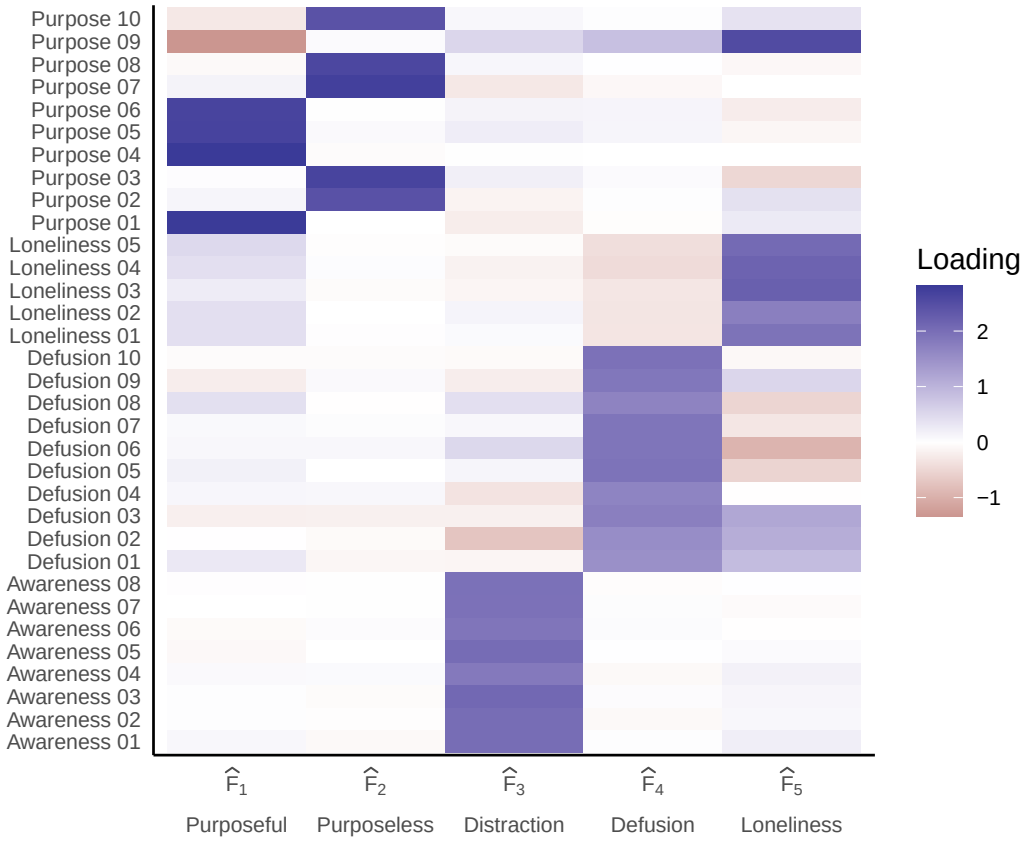


Figure 12: Survey item embeddings $\widehat{F}$ based on the survey responses at the end of four weeks.

variance in the survey responses. The factors $\widehat{F}_{\cdot 3}, \widehat{F}_{\cdot 4}, \widehat{F}_{\cdot 5}$ load primarily on the awareness, defusion and loneliness survey questions. The factors $\widehat{F}_{\cdot 1}$ and $\widehat{F}_{\cdot 2}$ load primarily on the purpose survey questions. However, two factors are need to capture the purpose latent construct because the survey questions are written with alternating valences:

1. I understand my life's meaning.
2. I am looking for something that makes my life feel meaningful.
3. I am always looking to find my life's purpose.
4. My life has a clear sense of purpose.
5. I have a good sense of what makes my life meaningful.
6. I have discovered a satisfying life purpose.
7. I am always searching for something that makes my life feel significant.
8. I am seeking a purpose or mission for my life
9. My life has no clear purpose.
10. I am searching for meaning in my life.

Items 1 and 2, for instance, are coded with opposite valences, but responses for all questions were on an integer scale from 1 (absolutely untrue) to 7 (absolutely true). Larger numerical

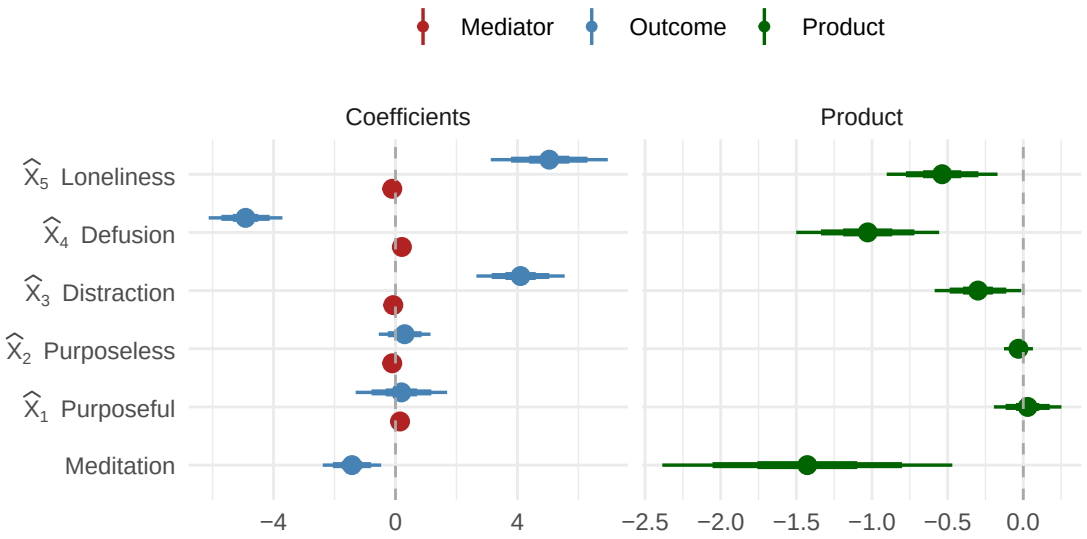


Figure 13: Left: Coefficients from the Healthy Minds mediator and outcome regression models, visualized as point intervals with interval widths 0.5, 0.8, 0.95, respectively. Coefficients for control variables not visualized. Intervals in the mediator model are very close to the point estimates, and thus not visible. All mediator coefficients are statistically distinguishable from zero. Right: confidence intervals for the corresponding coefficient products. The treatment coefficient, labelled "Meditation," is simply repeated from the left panel for a comparative sense of scale.

responses to item 1 indicate a greater sense of purpose, so we call $\widehat{F}_{\cdot 1}$ the "purposeful" factor. Larger numerical responses to item 2 indicate an absence of purpose, so we call $\widehat{F}_{\cdot 2}$ the "purposeless" factor. Interestingly, the 9-th item "My life has no clear purpose' is not particularly associated with the purposeful and purposeless factors $\widehat{F}_{\cdot 1}$ and $\widehat{F}_{\cdot 2}$, but rather the loneliness factors $\widehat{F}_{\cdot 5}$ (survey items for other scales are available in Appendix C). Using similar reasoning, we name $\widehat{F}_{\cdot 3}$ the "distraction" factor rather than the awareness factor, because larger numerical responses indicate lower levels of awareness (see Appendix C for the survey items and response scale).

In order to conduct the mediation analysis, we fit the following two regression models on the study participants:

$$\texttt{anxiety} \sim \texttt{meditation + sex + age + Xhat}$$
$$\texttt{Xhat} \sim \texttt{meditation + sex + age}.$$

We visualize the coefficients for both regressions in Figure 13. First, consider the mediator model coefficients. The intervention has a small but statistically significant effect on each of the latent mediators: it reduces loneliness, increases cognitive defusion, decreases distraction, decreases purposelessness and increases purposefulness. This suggests that the Healthy Minds Program is effectively improving the theorized psychological mediators. However, not all of the psychological mediators cause anxiety. Both coefficients for purpose-related factors are consistent with zero causal effects. Loneliness and distraction

(i.e., low awareness) both increased anxiety, and cognitive defusion (i.e., the ability to step back from thoughts and feelings and reflect on them) reduced anxiety. Meditation also appears to have a direct anxiety reducing effect, independent of its impact on the psychological mediators.

Considering the causal pathways, and the outcome regression and the mediation regression in aggregate, we estimate 95% confidence intervals for the natural direct effect as $[-2.38, -0.45]$ and for the natural indirect effect as $[-2.69, -1.05]$. That is, about half the anxiety reducing effect of the intervention was due to indirect effect, and about half was due to the direct effect. A natural followup question is if products of factor-specific regression coefficients can be interpreted as factor-specific indirect effects. This is the case, provided that we make the additional assumption that the latent psychometric factors are independent of one another conditional on intervention and controls. We visualize these products and their associated confidence intervals in the right panel of Figure 13.

If conditional independence of the factors is plausible, the defusion and loneliness pathways seem most important for reducing anxiety. The distraction estimate is smaller and nearly compatible with a null effect. The purpose effects are both fairly precisely measured zeroes. This factor-specific effects might motivate more effective meditation interventions. Each of the four weeks of the Healthy Minds intervention is devoted to improving one of the four hypothesized mediators, and our results suggest that it could be beneficial to replace the purpose module of the Healthy Minds program with an alternative, possibly doubling the time devoted to defusion or social connectedness skills. Alternatively, it may be worthwhile to investigate whether or not the factors are indeed independent; we find it particularly interesting that the Purpose 09 item ("My life has no clear purpose") was so strongly associated with the loneliness factor, possibly suggesting a connection between sense of meaning and social connectedness that could be leveraged in future interventions.

The analysis thus far has exclusively considered the survey responses and anxiety levels at the end of the four week intervention period. Participants were surveyed before the intervention period, then weekly for four weeks during the intervention period, and then three months after the end of the intervention. As a final step, we repeated our mediation analysis, considering each time point in the study independently. We compute direct and indirect effects at each stage of the study, and then plotted them in Figure 14. We see that the direct effect appears at week 1 of the intervention, and persists at the same level throughout the program, before fading back down three months post intervention. This leads us to believe that the direct effect is capturing the beneficial effect of calming breathing exercises, which have immediate and short term benefits. In contrast, the indirect effect grows slowly over time, exactly as one would expect if participants were slowly learning new and healthier habits of mind. These effects also seem to persist more strongly after the end of the intervention period. This suggests that a longer program might have more beneficial effect, and perhaps that the larger indirect effects may persist longer.

6. Discussion

In this paper, we have explored the use of principal components network regression for analyzing structured data and its potential applications in causal inference. We highlight four main takeaways from our research.

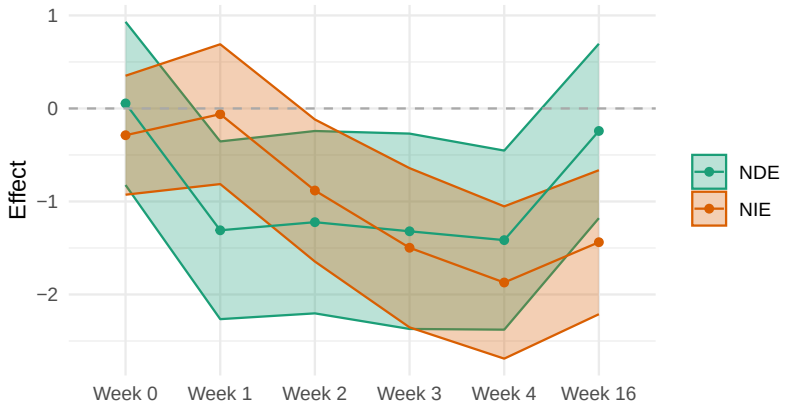


Figure 14: Natural direct and indirect effects of meditation on anxiety level, analyzed independently at each week in the Healthy Minds Trial, controlling for age and sex as confounders, and using a five dimensional embedding of survey responses at each time point.

Use principal components network regression. Principal components network regression is distributionally agnostic, robust to noisily observed networks, and computationally straightforward. The general low-rank sub-gamma model that we consider accommodates a wide variety of parametric submodels, as well as noisily observed networks, and is appropriate for many kinds of structured data. We recommend including spectral network embeddings in ordinary least squares regressions. We have shown that including spectral network embeddings in ordinary least squares estimators only requires semi-parametric assumptions. Asymptotically, it is equivalent to observing latent low-rank structure in the network and including population structure in regressions, although in finite samples there may be some bias induced by estimation error. Low-rank network regressions are consistent and asymptotically normal under weak and distribution-free assumptions. Although some regression coefficients are subject to an unknown orthogonal transformation, in practical applications this may not matter, or may be resolvable with varimax rotation (Rohe and Zeng, 2023).

Principal components network regression can be used for causal inference. Principal components network regression is useful in observational settings, and additionally for causal inference. We have carefully detailed the counterfactual and statistical assumptions required for regression coefficients to have causal interpretations. When latent network positions are mediators, coefficients from principal components network regression can be used in the product-of-coefficients method for estimating natural direct and indirect effects. Much of the causal intuition for tabular regressions also applies to our network regressions (VanderWeele and Vansteelandt, 2014), but there some network-specific concerns such as homophily causing positivity violations, and the presence or absence of peer effects.

Social groups can act as mediators, not just confounders. In other words, latent homophily can have a mediating effect rather than a confounder effect. As a result, it is critical to carefully consider the role of latent positions in causal settings, because causal analysis should adjust for confounders, but should not adjust for mediators. Mistaking a

mediator for a confounder and then adjusting for the mediator induces over-control bias. We empirically demonstrate how incorrect assumptions about the direction of causation can induce dramatic overcontrol bias into causal estimates. We suspect overcontrol bias is most likely to be an issue when considering causal effects of demographic features, which are likely to induce homophily in networks.

Network embeddings need to be interpreted. Network embeddings are not black boxes, magical controls, or inherent mediators. Network embeddings are estimates of latent structure in networks, and compelling inference using network embeddings requires an interpretation of that latent structure. It is not enough to claim that network embeddings capture the homophily relevant to a particular causal pathway. In applied projects, one must confirm that the network embeddings capture the important latent constructs, rather than noise, or latent constructs other than those originally hypothesized. As an important corollary, practitioners should not use network embeddings for causal inference unless they have sufficient domain knowledge or auxiliary data to give the embeddings a concrete and substantive interpretation. Pragmatically, interpreting network embeddings is complicated by orthogonal non-identifiability, but this identification challenge can often be resolved with varimax rotation (Rohe and Zeng, 2023), or mixture modeling in the latent space (Rubin-Delanchy et al., 2022).

Reproducibility

A replication package for our simulations and data analysis is available at `https://github.com/alexpghayes/network-mediation-replication`.

Acknowledgments

We thank Felix Elwert, Hyunseung Kang, Karl Rohe, Steven Wright, Sébastien Roch, Sameer Deshpande, Adeline Lo, Yehzee Ryoo, Tianxi Li, Can Le, Elizabeth Ogburn, Edward McFowland III, Simon Goldberg, Matthew Hirshberg, Isabel Fulcher, Nina Varsava, Ralph Trane, Bennett Zhu, Ben Lystig, Dan Bolt, Katharine Scott, and Emily Case for helpful discussions and feedback. Several anonymous reviewers provided feedback that greatly improved the manuscript. Support for this research was provided by the University of Wisconsin–Madison, Office of the Vice Chancellor for Research and Graduate Education with funding, as well as NSF grants DMS 2052918, DMS 1646108 and DMS 2023239. Support for this research was also provided by American Family Insurance through a research partnership with the University of Wisconsin-Madison's American Family Insurance Data Science Institute.

Appendix A. Alternative Blockmodel Parameterization

We develop our theory of network regression around the latent positions X. While this is mathematically convenient and allows us to relate our work to previous work on random dot product graphs, it is often difficult to develop good intuition for the latent positions X, even in well-known settings such as the stochastic blockmodel, as discussed in Section 2.4. Instead of parameterizing a network in terms of latent positions X, there is often a natural decomposition of $P = ZBZ^T$, which we explore here.

Proposition 19 (Equivalent Parameterizations for Network Regression) *Suppose that $Z \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{d\times d}$ are arbitrary full-rank matrices such that $\mathbb{E}[A \mid Z, B] = ZBZ^T$ and let*

$$Z = W\Theta' + \xi', \tag{4}$$

$$Y = W\beta_w + Z\beta_z + \varepsilon' \tag{5}$$

where $\xi' = Z - \mathbb{E}[Z \mid W] \in \mathbb{R}^{n\times d}$ and each row $\xi'_{i\cdot}$ is mean-zero and uncorrelated with the corresponding row $W_{i\cdot}$, and the ε_i are independent with bounded second moments. Then there exist $\Theta \in \mathbb{R}^{p\times d}, \xi \in \mathbb{R}^{n\times d}, \beta_x \in \mathbb{R}^d$ and $\varepsilon \in \mathbb{R}^n$ such that

$$X = W\Theta + \xi, \qquad \text{and}$$
$$Y = W\beta_w + X\beta_x + \varepsilon,$$

where ξ satisfies $\mathbb{E}[\xi_{ij} \mid W_{i\cdot}] = 0$ for $i \in [n], j \in [d]$, and the elements of ε are independent with bounded second moments.

This proposition has several implications. First, if there is a linear regression (4) to establish how Z varies with nodal covariates W, then there is another equivalent regression, also linear, to establish how X varies with nodal covariates W. That is, every "conveniently" parameterized mediator model implies an "inconveniently" parameterized mediator model (our proofs are developed in the "inconvenient" setting). Further, provided that the errors in the Z-regression are uncorrelated with W, then the errors in the X regression are also uncorrelated with W. Thus, if the coefficients in the convenient mediator regression are estimable, so are the coefficients in the inconvenient mediator regression. There is an analogous story for the outcome regression (5).

Remark 20 *The reparameterization above comes at the cost of a potential loss of identifiability. For example, there are some law-rank models where X is identifiable up to multiplication by a signed permutation matrix (e.g., Rohe and Zeng, 2022, Proposition 3.2), and much of the rotational ambiguity induced can be resolved via a varimax rotation. This in turn implies β_x is identifiable up to sign flips in the regression coefficients β_x. Since our network regression model does not leverage any additional identifying information about latent positions X, even when X is narrowly identified, β_x is always subject to non-identifiability up to a full orthogonal rotation.*

We conjecture that plugging varimax rotated singular vectors into a nodal regression results in a consistent estimator of regression coefficients. We further anticipate that these estimates are normally distributed in the large-n limit, effectively resolving the issue of rotational non-identifiability in network regression settings.

Proof [Proof of Proposition 19] First we consider the mediator model. We explicitly construct Θ' and ξ' and then verify the dependence properties of ξ'. Let $\Sigma_Z = Z^T Z$ and let $R_U^T D R_U$ be the singular value decomposition of $\Sigma_Z^{1/2} B \Sigma_Z^{1/2}$. By Proposition 3.2 of Rohe and Zeng (2022), D is a diagonal matrix containing the singular values of $\mathbb{E}[A \mid Z, B]$, and the singular vectors of $\mathbb{E}[A \mid Z, B]$ are given by

$$U = Z \Sigma_Z^{-1/2} R_U^T.$$

Thus, Z and B together imply that $X = Z \Sigma_Z^{-1/2} R_U^T D^{1/2}$. Then, by hypothesis,

$$X = Z \Sigma_Z^{-1/2} R_U^T D^{1/2} = (W\Theta + \xi) \Sigma_Z^{-1/2} R_U^T D^{1/2} = W\Theta' + \xi'.$$

where $\Theta' = \Theta\, T_Z$ and $\xi' = \xi\, T_Z$ and $T_Z = \Sigma_Z^{-1/2} R_U^T D^{1/2}$ is an invertible matrix depending only on Z and B. The tower law then yields

$$\mathbb{E}\left[\xi'_{i\cdot} \,\middle|\, W_{i\cdot}\right] = \mathbb{E}[\xi_{i\cdot}\, T_Z \mid W_{i\cdot}] = \mathbb{E}[\mathbb{E}[\xi_{i\cdot} \mid W_{i\cdot}]\, T_Z \mid Z, B] = \mathbb{E}[0 \cdot T_Z \mid Z, B] = 0.$$

Now, turning to the outcome model, we have $Z = X D^{-1/2} R_U \Sigma_Z^{1/2}$. Let $\beta_{\mathrm{x}} = D^{-1/2} R_U \Sigma_Z^{1/2} \beta_{\mathrm{z}}$. Then $Z\beta_{\mathrm{z}} = X\beta_{\mathrm{x}}$ and we can take $\varepsilon = \varepsilon'$. Since $X_{i\cdot}$ is a function of $Z_{i\cdot}$ and B, and B is fixed, independence of $Z_{i\cdot}$ and ε_i implies independence of $X_{i\cdot}$ and ε_i, completing the proof. ■

Appendix B. Additional Details about the Glasgow Data Example

In the Glasgow example, recall that the adolescent social network was highly sexually homophilous (see Figure 1). This high level of homophily suggests that there may be positivity violations (Sections 2.3 and 2.4), and so we investigate positivity empirically by plotting $\widehat{F}$. In Figure 15, we see that the latent embeddings in the Glasgow data likely violate the positivity assumption, as some regions of the latent space are only occupied by male or female students. This implies that causal identification may not hold in the Glasgow data set.

Also recall that that the Glasgow social network is a directed network, and as a result each node has a set of left co-embeddings $\widehat{X}$ and a set of right co-embeddings $\widehat{F}$. In Section 5.1, we used the right co-embeddings $\widehat{F}$, but we were unable to confirm that $\widehat{F}$ characterized the relevant social structure in the network. In Figure 16, we see that the results based on left co-embeddings $\widehat{X}$ are estimated to be smaller in magnitude than those based on the right co-embeddings $\widehat{F}$. Since we are unable to definitely adjudicate between $\widehat{X}$ and $\widehat{F}$, the difference in these analyses introduces ambiguity into the data analysis.

Appendix C. Additional Details about the Healthy Minds Data Example

Figure 17 characterizes the causal structure of mediation in a bipartite network, and Figure 18 shows that $\widehat{\Psi}_{\mathrm{nde}}$ and $\widehat{\Psi}_{\mathrm{nde}}$ are insensitive to the embedding dimension in the Healthy Minds data application.

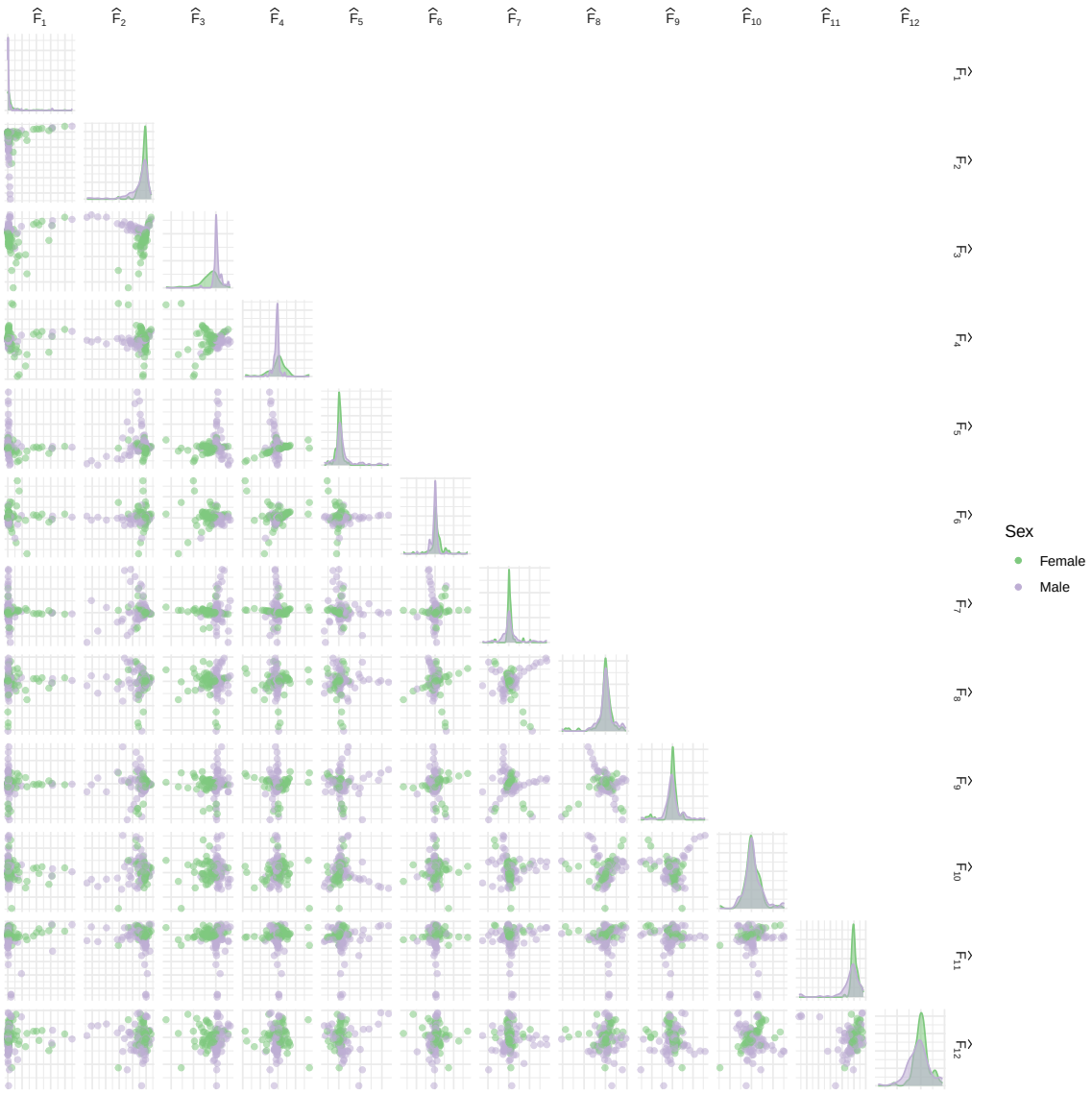


Figure 15: Right latent co-embeddings in the Glasgow data set, colored by reported sex.

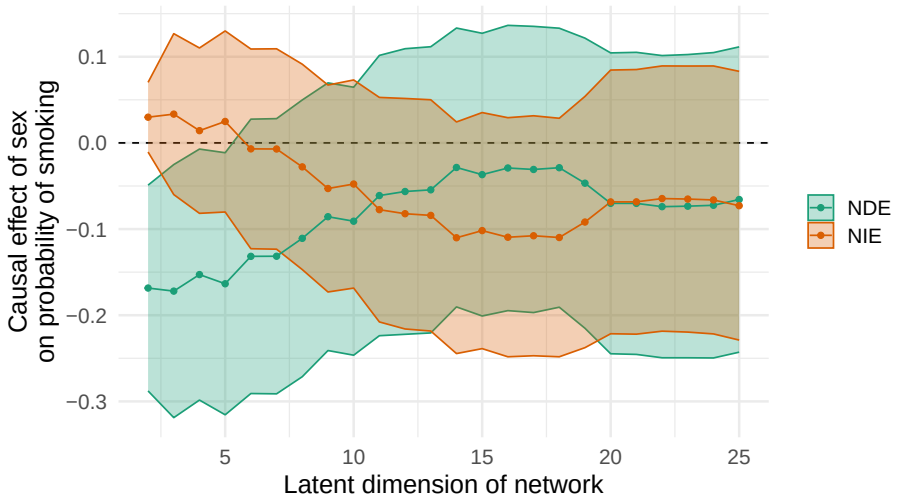


Figure 16: Estimated direct (teal) and indirect (orange) effects of sex on tobacco usage in the Glasgow social network, using the left co-embeddings $\widehat{X}$ rather than the right co-embeddings $\widehat{F}$ (compare with Figure 9). The estimated effects (vertical axis) vary with the dimension d of the latent space (horizontal axis), and are adjusted for possible confounding by age and church attendance. Positive values indicate a greater propensity for adolescent boys to smoke, while negative values indicate a greater propensity for adolescent girls to smoke.

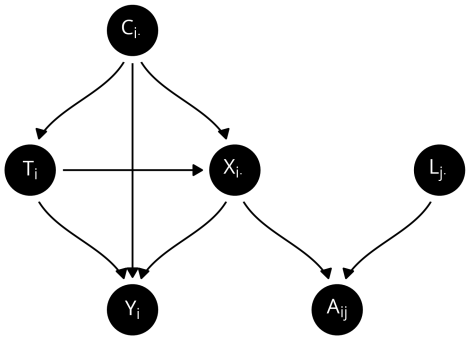


Figure 17: A directed acyclic graph (DAG) representing the causal pathways of latent mediation in a bipartite network. The network has two nodes called i and j and the node i is the unit of interest. Each node in the figure corresponds to a random variable, and edges indicate which random variables may cause which other random variables. We are interested in the causal effect of T_i on Y_i, as mediated by the latent position $X_{i\cdot}$.

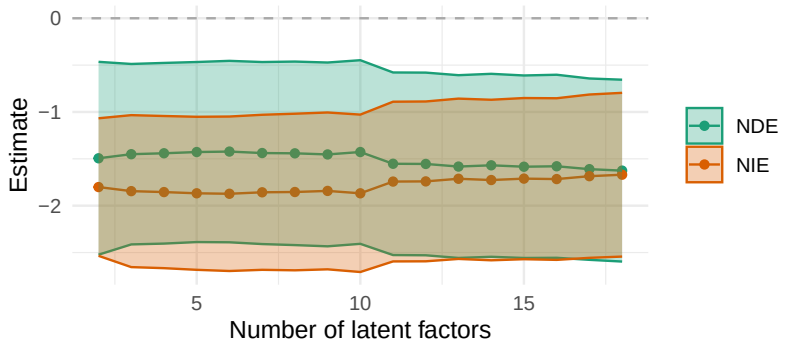


Figure 18: Estimated direct (teal) and indirect (orange) effects of the Healthy Minds program on anxiety levels at the end of the four week intervention period. The estimated effects (vertical axis) vary with the dimension d of the latent space (horizontal axis), and are adjusted for possible confounding by age and sex attendance. Positive values indicate that the intervention increased anxiety, while negative values indicate that intervention decreases anxiety.

C.1 Five Facet Mindfulness Questionnaire Act With Awareness subscale

1. When I do things, my mind wanders off and I'm easily distracted.
2. I don't pay attention to what I'm doing because I'm daydreaming, worrying, or otherwise distracted.
3. I am easily distracted.
4. I find it difficult to stay focused on what's happening in the present.
5. It seems I am 'running on automatic' without much awareness of what I'm doing.
6. I rush through activities without being really attentive to them.
7. I do jobs or tasks automatically without being aware of what I'm doing.
8. I find myself doing things without paying attention.

Score	Description
5	Very often or always true
4	Often true
3	Sometimes true
2	Rarely true
1	Never or very rarely true

C.2 Drexel Defusion Scale

1. Feelings of anger. You become angry when someone takes your place in a long line. To what extent would you normally be able to defuse from feelings of anger?
2. Cravings for food. You see your favorite food and have the urge to eat it. To what extent would you normally be able to defuse from cravings for food?
3. Physical pain. Imagine that you bang your knee on a table leg. To what extent would you normally be able to defuse from physical pain?

4. Anxious thoughts. Things have not been going well at school or your job, and work just keeps piling up. To what extent would you normally be able to defuse from anxious thoughts like "I'll never get this done."?
5. Thoughts of self. Imagine you are having a thought such as "no one likes me." To what extent would you normally be able to defuse from negative thoughts about yourself?
6. Thoughts of hopelessness. You are feeling sad and stuck in a difficult situation that has no obvious end in sight. You experience thoughts such as "Things will never get any better." To what extent would you normally be able to defuse from thoughts of hopelessness?
7. Thoughts about motivation or ability. Imagine you are having a thought such as "I can't do this" or "I just can't get started." To what extent would you normally be able to defuse from thoughts about motivation or ability?
8. Thoughts about your future. Imagine you are having thoughts like, "I'll never make it" or "I have no future." To what extent would you normally be able to defuse from thoughts about your future?
9. Sensations of fear. You are about to give a presentation to a large group. As you sit waiting for your turn, you start to notice your heart racing, butterflies in your stomach, and your hands trembling. To what extent would you normally be able to defuse from sensations of fear?
10. Feelings of sadness. Imagine that you lose out on something you really wanted. You have feelings of sadness. To what extent would you normally be able to defuse from feelings of sadness?

Score	**Description**
5	Very much
4	Quite a lot
3	Moderately
2	Somewhat
1	A little
0	Not at all

C.3 NIH Toolbox Loneliness Questionnaire

1. I feel alone and apart from others
2. I feel left out
3. I feel that I am no longer close to anyone
4. I feel alone
5. I feel lonely

Score	**Description**
5	Always
4	Usually
3	Sometimes
2	Rarely
1	Never

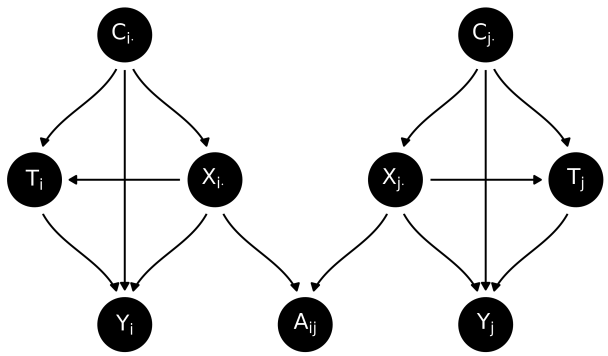


Figure 19: A DAG representing the causal pathways in a network with homophilous *confounding*, for a network with two nodes called i and j. Compare to Figure 2, where the direction of the $X_{i\cdot} \to T_i$ and $X_{j\cdot} \to T_j$ arrows are reversed.

Appendix D. Case Study: Over-Control Bias

Treating latent positions as confounders when they are in fact mediators leads to biased causal estimates. This bias can be substantial in empirical networks.

When the latent positions are confounders (see the structural causal model in Figure 19), β_{t} is the average treatment effect:

$$\mathbb{E}[Y_i \,|\, T_i, C_{i\cdot}, X_{i\cdot}] = \beta_0 + T_i \underbrace{\beta_{\mathrm{t}}}_{\substack{\text{average}\\ \text{treatment}\\ \text{effect}}} + C_{i\cdot}\beta_{\mathrm{c}} + X_{i\cdot}\beta_{\mathrm{x}}.$$

In contrast, when the latent positions are mediators (see the structural causal model in Figure 2), β_{t} is the natural direct effect:

$$\mathbb{E}[Y_i \,|\, T_i, C_{i\cdot}, X_{i\cdot}] = \beta_0 + T_i \underbrace{\beta_{\mathrm{t}}}_{\substack{\text{natural}\\ \text{direct}\\ \text{effect}}} + C_{i\cdot}\beta_{\mathrm{c}} + X_{i\cdot} \underbrace{\beta_{\mathrm{x}}}_{\substack{\text{effect of}\\ X \text{ on } Y}}, \quad \text{and}$$

$$\mathbb{E}[X_{i\cdot} \,|\, T_i, C_{i\cdot}] = \theta_0 + T_i \underbrace{\theta_{\mathrm{t}}}_{\substack{\text{effect of}\\ T \text{ on } X}} + C_{i\cdot}\Theta_{\mathrm{c}}.$$

Treating the latent positions as confounders when they are mediators implies the mistaken identification result $\beta_{\mathrm{t}} = \Psi_{\mathrm{ate}}$. However, in truth, $\beta_{\mathrm{t}} = \Psi_{\mathrm{nde}} = \Psi_{\mathrm{ate}} - \Psi_{\mathrm{nie}}$, and using β_{t} as an estimate of Ψ_{ate} induces a bias of Ψ_{nie} into the estimate of the average treatment effect. This bias is well-known as "over-control bias" (Cinelli et al., 2022).

Over-control bias can be large in network data. To demonstrate this, we re-analyzed the AddHealth data set investigated in the initial pre-print of Le and Li (2022). The AddHealth data consists of a self-reported social network of 2,152 high school students, along with grade level, sex, race, and a proxy measure of mental health for each student.

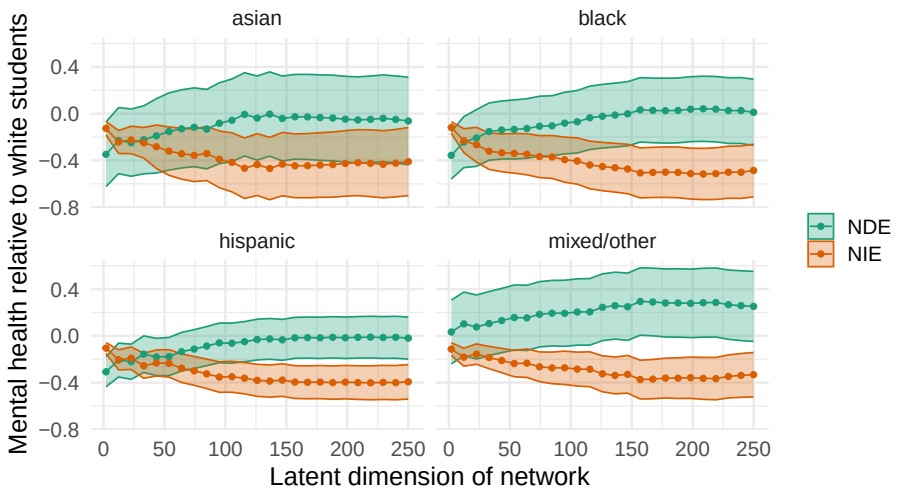


Figure 20: Confidence intervals (95%) for natural direct (teal) and indirect (orange) effects in the AddHealth data example. The vertical axis corresponds to the value of the causal effect, and the horizontal axis encodes the embedding dimension d. All contrasts are relative to white students.

Our goal is to investigate how mental health varies with race, controlling for grade level and sex.

In the original analysis, Le and Li (2022) used a procedure that is equivalent to the ordinary least squares regression

$$\texttt{mental_health} \sim \texttt{grade + race + sex + Xhat}$$

and found that race did not have a statistically significant effect on mental health. The original analysis did not interpret the regression coefficients causally, but it did suggest that the effect of race was plausibly zero.

Our mediation framework allows us to clarify the role of race. Race (as well as grade and sex) causes community structure, rather than the other way around, such that latent social groups are clearly mediators in the AddHealth network. Indeed, race precedes the network in time, and it is impossible for causal arrows to point backwards in time.

Using the framework developed in this paper, we fit models

$$\texttt{mental_health} \sim \texttt{grade + race + sex + Xhat}$$
$$\texttt{Xhat} \sim \texttt{grade + race + sex}$$

and then computed $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$, which we plotted with confidence intervals in Figure 20. Using the cross-validated eigenvalue method proposed in Chen et al. (2021), we determined that a reasonable choice of latent dimension was $d \approx 120$. Our estimates were stable in a neighborhood around this value of d, suggesting that they were reliable so long d was not badly misspecified (see Remark 10 and Section 4).

At $d = 120$, the estimated direct effect of non-white race was zero, and the estimated indirect effects of non-white race was large and statistically significant. This in turn implied

that there was a large average treatment effect of race on mental health. The effect simply operates along the indirect pathway, as race causes group membership.

The takeaways from this case study are two-fold. First, when latent positions are included in network regression models, they are likely to be interpreted, either implicitly or explicitly, as causal confounders. It's important to assess whether or not this is the case. Second, it's entirely plausible for causal effects in social networks to occur entirely via the indirect pathway. In this setting, mistakenly using the latent positions as confounders rather than mediators can result in over-control bias and very misleading estimates of causal effects.

Appendix E. Proof of Theorem 9

In order to prove Theorem 9, we will first prove a more general statement holds under the sub-gamma network model (Assumption 2), which generalizes the random dot product graph considered in Example 1. To work with the general sub-gamma model, we require the following assumptions about the sub-gamma parameters:

Assumption 6 (Growth rates) *Under the model in Assumption 2, the eigenvalues λ_1 and λ_d and the sub-gamma parameters ν_n and b_n grow with n in such a way that*

$$\lambda_1 = \mathcal{O}(n), \tag{6}$$

$$\lambda_1 = \Omega(1), \tag{7}$$

$$\frac{\sqrt{\nu_n + b_n^2}\, n \log^{3/2} n}{\lambda_d^{3/2}} = o(1), \tag{8}$$

$$\frac{\lambda_1(\nu_n + b_n^2) n^2 \log^2 n}{\lambda_d^{7/2}} = o(1), \tag{9}$$

and

$$\frac{\lambda_d}{\lambda_1^{1/2}} = \mathcal{O}(1). \tag{10}$$

We additionally take d, the rank of the latent positions X, and p, and the number of nodal controls, to be fixed asymptotically.

Note that Assumption 6 holds for the random dot product graph (Example 1).

Proposition 21 *Suppose that $(A, X) \sim \mathrm{RDPG}(F, n)$, as described in Example 1. Then (A, X) are generated according to a process that satisfies Assumption 2, and further $\lambda_1 = \Theta(n), \lambda_d = \Theta(n), \nu_n = c$, and $b_n = 1$ for some $c > 0$.*

Proof See Athreya et al. (2018, Remark 24) or Sussman et al. (2014, Prop 4.3). ■

A straightforward application of Proposition 21 shows that random dot product graphs satisfy Assumption 6. Thus, in order to prove Theorem 9, it is sufficient to replace Example 1 with Assumptions 2 and 6, as in the following. Recalling the W notation of Equation (3), we have the following theorem.

Theorem 22 *If Assumptions 2, 3, 4, 5, and 6 hold, then there exists a sequence of orthogonal matrices $\{Q_n\}_{n=1}^{\infty}$ such that*

$$\sqrt{n}\,\widehat{\Sigma}_{\mathrm{vec}(\Theta)}^{-1/2}\left(\mathrm{vec}\left(\widehat{\Theta}\,Q_n^T\right)-\mathrm{vec}(\Theta)\right)\to\mathcal{N}(0,I_{pd}),\ \textit{and}$$
$$\sqrt{n}\,\widehat{\Sigma}_{\beta}^{-1/2}\begin{pmatrix}\widehat{\beta}_w-\beta_w\\ Q_n\,\widehat{\beta}_x-\beta_x\end{pmatrix}\to\mathcal{N}(0,I_d).$$

The proof of Theorem 22 reduces to three key lemmas. Lemma 23 shows that least squares estimates are asymptotically normally distributed when the true latent positions are known. This is a standard result of M-estimation theory for regression.

Lemma 23 (Boos and Stefanski (2013), Theorems 7.2) *Define $\widetilde{\beta}_w, \widetilde{\beta}_x$ and $\widetilde{\Theta}$ analogously to the estimators in Definition 7, but using the true latent positions X rather than the spectral embedding $\widehat{X}$. Under Assumptions 2, 3, and 4,*

$$\sqrt{n}\left(\mathrm{vec}\left(\widetilde{\Theta}\right)-\mathrm{vec}(\Theta)\right)\to\mathcal{N}\big(0,\Sigma_{\mathrm{vec}(\Theta)}\big),\ \textit{and}$$
$$\sqrt{n}\begin{pmatrix}\widetilde{\beta}_w-\beta_w\\ \widetilde{\beta}_x-\beta_x\end{pmatrix}\to\mathcal{N}(0,\Sigma_\beta).$$

Further, when the true latent positions are known, the covariance of the least squares coefficients is also estimable using the robust covariance estimator.

Lemma 24 (Boos and Stefanski (2013), Theorems 7.3, 7.4) *Define $\widetilde{\Sigma}_\beta$ and $\widetilde{\Sigma}_{\mathrm{vec}(\Theta)}$ analogously to the estimators in Definition 8, but using the true latent positions X rather than the spectral embedding $\widehat{X}$. Under Assumptions 2, 3, and 4,*

$$\widetilde{\Sigma}_\beta\to\Sigma_\beta\quad\textit{in probability, and}$$
$$\widetilde{\Sigma}_{\mathrm{vec}(\Theta)}\to\Sigma_{\mathrm{vec}(\Theta)}\quad\textit{in probability.}$$

Since the true latent positions X are unobserved, we must use estimates $\widehat{X}$ in place of X in least squares estimators. This does not change the asymptotic distribution of the least squares coefficients.

Lemma 25 *Under Assumptions 2, 3, 4, 5, and 6,*

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_w-\widetilde{\beta}_w\\ Q_n\,\widehat{\beta}_x-\widetilde{\beta}_x\end{pmatrix}=o_p(1)\quad\textit{and}\quad\sqrt{n}\left(\widehat{\Theta}\,Q_n^T-\widetilde{\Theta}\right)=o_p(1).$$

We must show a similar result for the covariance estimators in order to obtain a consistent estimator based on $\widehat{X}$ rather than X.

Lemma 26 *Under Assumptions 2, 4, 5, and 6, we have*

$$\widehat{\Sigma}_\beta \to \widetilde{\Sigma}_\beta \qquad \textit{and} \qquad \widehat{\Sigma}_{\mathrm{vec}(\Theta)} \to \widetilde{\Sigma}_{\mathrm{vec}(\Theta)} \qquad \textit{both in probability.}$$

Proofs of Lemmas 25 and 26 are provided in Sections E.1 and E.4, respectively. Additional technical results are also given in Section E.3.

With these preliminaries in place, we are ready to prove Theorem 22. Note that in this proof, and subsequently, we often drop the subscript n from Q_n for convenience and write Q instead.

Proof [Proof of Theorem 22]

First, we show that $\widehat{\beta}$ and $\widehat{\Theta}$ are asymptotically normal. By Lemmas 25 and 23, asymptotically we have

$$\sqrt{n}\Big(\widehat{\Theta}\, Q^T - \Theta\Big) = \underbrace{\sqrt{n}\Big(\widehat{\Theta}\, Q^T - \widetilde{\Theta}\Big)}_{o_p(1)} + \underbrace{\sqrt{n}\Big(\widetilde{\Theta} - \Theta\Big)}_{\mathcal{N}(0,\Sigma_\Theta)}.$$

Slutsky's theorem is then sufficient to establish asymptotic normality. An analogous argument holds for $\widehat{\beta}$.

We use a similar argument to establish consistency of covariance estimation. By Lemmas 26 and 24

$$\widehat{\Sigma}_{\mathrm{vec}(\Theta)} - \Sigma_{\mathrm{vec}(\Theta)} = \underbrace{\widehat{\Sigma}_{\mathrm{vec}(\Theta)} - \widetilde{\Sigma}_{\mathrm{vec}(\Theta)}}_{o_p(1)} + \underbrace{\widetilde{\Sigma}_{\mathrm{vec}(\Theta)} - \Sigma_{\mathrm{vec}(\Theta)}}_{o_p(1)},$$

such that $\widehat{\Sigma}_{\mathrm{vec}(\Theta)} \to \Sigma_{\mathrm{vec}(\Theta)}$ in probability. Again, an analogous argument holds for the covariance of $\widehat{\beta}$. A final application of Slutsky's theorem to combine the above two results completes the proof. ■

E.1 Proof of Lemma 25

To prove Lemma 25, we will first consider the mediator regression coefficients, and then the outcome regression coefficients. We partition the outcome regression coefficients $\beta = (\beta_{\mathrm{w}}, \beta_{\mathrm{x}})$ using the Frisch-Waugh-Lowell theorem to deal with identified and unidentified coefficients separately. Here we present some important supporting lemmas that outline the proof; some tedious and less illuminating supporting lemmas are relegated to a later portion of the Appendix.

Lemma 27 (Sub-gamma mediator coefficient bound) *Suppose Assumptions 2, 4, 5 and 6 hold. Let $\{Q_n\}_{n=1}^\infty$ be the sequence of orthogonal matrices guaranteed by Lemma 31. Then*

$$\Big\|\sqrt{n}\Big(\widehat{\Theta}\, Q_n^T - \widetilde{\Theta}\Big)\Big\| = o_p(1).$$

Proof Using basic properties of the spectral norm,

$$\begin{aligned}\Big\|\sqrt{n}\Big(\widehat{\Theta}\, Q^T - \widetilde{\Theta}\Big)\Big\| &= \sqrt{n}\,\Big\|\big(W^TW\big)^{-1}W^T\Big(\widehat{X} - XQ\Big)\Big\| \\ &\le \sqrt{n}\,\frac{1}{n}\Big\|\Big(\frac{1}{n}W^TW\Big)^{-1}\Big\|\Big\|W^T\Big(\widehat{X} - XQ\Big)\Big\|.\end{aligned} \tag{11}$$

By the independence and moment conditions of Assumption 4, $\left(\frac{1}{n}W^TW\right)^{-1}$ converges to the inverse covariance matrix of W, so that

$$\left\|\left(\frac{1}{n}W^TW\right)^{-1}\right\| = \mathcal{O}(1).$$

By Lemma 49,

$$\left\|W^T\left(\widehat{X} - XQ\right)\right\| = o_p(\sqrt{n}).$$

Applying the above two displays to Equation (11),

$$\left\|\sqrt{n}\left(\widehat{\Theta}\,Q^T - \widetilde{\Theta}\right)\right\| = o_p(1),$$

completing the proof. ■

Theorem 28 (Lovell 1963; Frisch and Waugh 1933) *Let $\widehat{\beta}$ be as in Definition 7. Then*

$$\begin{bmatrix}\widehat{\beta}_w\\ \widehat{\beta}_x\end{bmatrix} = \begin{bmatrix}\left(W^TW\right)^{-1}W^T\left(Y - \widehat{X}\,\widehat{\beta}_x\right)\\ \left(\widehat{X}^TM\widehat{X}\right)^{-1}\widehat{X}^TMY\end{bmatrix} \tag{12}$$

where $M = I - W^T\left(W^TW\right)^{-1}W$.

We note that M projects vectors onto the orthogonal complement of the column space of W. Note that $\|M\| = 1$ since M is a projection matrix.

Lemma 29 (Sub-gamma outcome coefficient bound) *Suppose Assumptions 2, 4, 5 and 6 hold, and let $\{Q_n\}_{n=1}^{\infty}$ be the sequence of orthogonal matrices guaranteed by Lemma 31. Then*

$$\sqrt{n}\left(Q_n\,\widehat{\beta}_x - \widetilde{\beta}_x\right) = o_p(1).$$

Proof Applying the definition of $\widehat{\beta}_{\mathrm{x}}$ and $\widetilde{\beta}_{\mathrm{x}}$ from Theorem 28 and adding and subtracting appropriate quantities, we have

$$\begin{aligned}\sqrt{n}\left(Q\,\widehat{\beta}_{\mathrm{x}} - \widetilde{\beta}_{\mathrm{x}}\right) &= \sqrt{n}\left[Q\left(\widehat{X}^TM\widehat{X}\right)^{-1}\widehat{X}^T - \left(X^TMX\right)^{-1}X^T\right]MY\\ &= \sqrt{n}\left[Q\left(\widehat{X}^TM\widehat{X}\right)^{-1} - \left(X^TMX\right)^{-1}Q\right]\widehat{X}^TMY \qquad (13)\\ &\quad + \sqrt{n}\left(X^TMX\right)^{-1}\left(Q\widehat{X}^T - X^T\right)MY. \qquad (14)\end{aligned}$$

We bound the quantities (13) and (14) separately, starting with (13). Recalling the definition of M in Theorem 28 and expanding Y,

$$\left\|\widehat{X}^TMY\right\| = \left\|\widehat{X}^T\left(MX\beta + MW\xi + M\varepsilon\right)\right\| = \left\|\widehat{X}^T\left(MX\beta + M\varepsilon\right)\right\|.$$

Applying the triangle inequality, submultiplicativity and the fact that M is a projection matrix,

$$\left\|\widehat{X}^T MY\right\| \leq \left\|\widehat{X}\right\|\|\beta\| + \left\|\widehat{X}^T M\varepsilon\right\|.$$

Lemma 33 and the fact that β is a constant control the first term, while Lemma 43 with $H = \widehat{X}^T M$ controls the second term, and we have

$$\left\|\widehat{X}^T MY\right\| \leq C\lambda_1^{1/2} + \sqrt{B \operatorname{trace} \widehat{X}^T MM^T \widehat{X}}. \tag{15}$$

By cyclicity of the trace and Von Neumann's trace inequality,

$$\operatorname{trace} \widehat{X}^T MM^T \widehat{X} = \operatorname{trace} \widehat{X}\widehat{X}^T MM^T \leq \operatorname{trace} \widehat{X}\widehat{X}^T = \operatorname{trace} \widehat{S},$$

where we have used the fact that M is a projection matrix and the definition of $\widehat{X} = \widehat{U}\widehat{S}^{1/2}$. Bounding $\operatorname{trace} \widehat{S} \leq d\|\widehat{S}\|$, Lemma 33 implies

$$\operatorname{trace} \widehat{X}^T MM^T \widehat{X} \leq Cd\lambda_1.$$

Plugging this bound back into Equation (15) and using the fact that B and d are assumed constant, we conclude that

$$\left\|\widehat{X}^T MY\right\| = \mathcal{O}_p\left(\lambda_1^{1/2}\right). \tag{16}$$

Noting that for conformable matrices A and B, $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, submultiplicativity of the spectral norm implies

$$\begin{aligned}
&\left\|Q\left(\widehat{X}^T M\widehat{X}\right)^{-1} - \left(X^T MX\right)^{-1} Q\right\| \\
&\quad = \left\|Q\left(\widehat{X}^T M\widehat{X}\right)^{-1}\left[Q^T X^T MX - \widehat{X}^T M\widehat{X}\, Q^T\right]\left(X^T MX\right)^{-1} Q\right\| \\
&\quad \leq \left\|\left(\widehat{X}^T M\widehat{X}\right)^{-1}\right\|\left\|Q^T X^T MX - \widehat{X}^T M\widehat{X}\, Q^T\right\|\left\|\left(X^T MX\right)^{-1}\right\|.
\end{aligned}$$

Applying Lemmas 42 and 46, it follows that

$$\left\|Q\left(\widehat{X}^T M\widehat{X}\right)^{-1} - \left(X^T MX\right)^{-1} Q\right\| = o_p\left(\frac{1}{\sqrt{n}\lambda_1^{1/2}}\right). \tag{17}$$

Applying submultiplicativity, we can bound (13) as

$$\begin{aligned}
&\left\|\sqrt{n}\left[Q\left(\widehat{X}^T M\widehat{X}\right)^{-1} - \left(X^T MX\right)^{-1} Q\right]\widehat{X}^T MY\right\| \\
&\quad \leq \sqrt{n}\left\|Q\left(\widehat{X}^T M\widehat{X}\right)^{-1} - \left(X^T MX\right)^{-1} Q\right\|\left\|\widehat{X}^T MY\right\|.
\end{aligned}$$

Applying Equations (16) and (17), the quantity in Equation (13) is bounded as

$$\left\|\sqrt{n}\left[Q\left(\widehat{X}^T M\widehat{X}\right)^{-1} - \left(X^T MX\right)^{-1} Q\right]\widehat{X}^T MY\right\| = o_p(1). \tag{18}$$

Turning our attention to the quantity on line (14), by applying submultiplicativity and the triangle inequality, along with the fact that M is a projection matrix,

$$\begin{aligned}
&\left\|\sqrt{n}\left(X^T M X\right)^{-1}\left(Q\,\widehat{X}^T - X^T\right)MY\right\| \\
&\quad\leq \sqrt{n}\left\|\left(X^T M X\right)^{-1}\right\|\left\|\left(Q\,\widehat{X}^T - X^T\right)MY\right\| \\
&\quad= \sqrt{n}\left\|\left(X^T M X\right)^{-1}\right\|\left\|\left(Q\,\widehat{X}^T - X^T\right)MX\beta\left(Q\,\widehat{X}^T - X^T\right)M\varepsilon\right\| \\
&\quad\leq \sqrt{n}\left\|\left(X^T M X\right)^{-1}\right\|\left\|\left(Q\,\widehat{X}^T - X^T\right)MX\right\|\|\beta\| \\
&\qquad+ \sqrt{n}\left\|\left(X^T M X\right)^{-1}\right\|\left\|\left(Q\,\widehat{X}^T - X^T\right)M\varepsilon\right\|.
\end{aligned}$$

Applying Lemmas 40, 45 and 46 and using the fact that $\|\beta\|$ is a constant,

$$\left\|\sqrt{n}\left(X^T M X\right)^{-1}\left(Q\,\widehat{X}^T - X^T\right)MY\right\| = o_p(1).$$

Using the above display and Equation (18), respectively, to bound the terms on lines (13) and (14), we conclude that

$$\sqrt{n}\left(Q\,\widehat{\beta}_{\mathrm{x}} - \widetilde{\beta}_{\mathrm{x}}\right) = o_p(1),$$

completing the proof. ■

Lemma 30 *Suppose that Assumptions 2, 3, 4, 5 and 6 hold. Then, letting $\widetilde{\beta}_w$ and $\widetilde{\beta}_x$ be as specified in Definition 7,*

$$\sqrt{n}\left(\widehat{\beta}_w - \widetilde{\beta}_w\right) = o_p(1).$$

Proof Applying Theorem 28 and using basic properties of norms, we have

$$\begin{aligned}
\left\|\sqrt{n}\left(\widehat{\beta}_{\mathrm{w}} - \widetilde{\beta}_{\mathrm{w}}\right)\right\| &\leq \sqrt{n}\left\|\left(W^T W\right)^{-1} W^T\left(X - \widehat{X}Q^T\right)\right\|\left\|\widetilde{\beta}_{\mathrm{x}}\right\| \\
&\quad+ \sqrt{n}\left\|\left(W^T W\right)^{-1} W^T \widehat{X}\right\|\left\|Q^T\widetilde{\beta}_{\mathrm{x}} - \widehat{\beta}_{\mathrm{x}}\right\|.
\end{aligned} \tag{19}$$

By Lemmas 23 and 27,

$$\left\|\widetilde{\beta}_{\mathrm{x}}\right\| = \mathcal{O}_p(1) \quad \text{and} \quad \sqrt{n}\left\|\left(W^T W\right)^{-1} W^T\left(X - \widehat{X}\,Q^T\right)\right\| = o_p(1),$$

and it follows that

$$\sqrt{n}\left\|\left(W^T W\right)^{-1} W^T\left(X - \widehat{X}Q^T\right)\right\|\left\|\widetilde{\beta}_{\mathrm{x}}\right\| = o_p(1). \tag{20}$$

By Lemma 29,

$$\sqrt{n}\left\|Q^T\,\widetilde{\beta}_{\mathrm{x}} - \widehat{\beta}_{\mathrm{x}}\right\| = o_p(1),$$

and by Lemmas 23 and 27,

$$\left\|\left(W^T W\right)^{-1} W^T \widehat{X}\right\| = \left\|\widehat{\Theta}\,Q^T\right\| \leq \left\|\widehat{\Theta}\,Q^T - \widetilde{\Theta}\right\| + \left\|\widetilde{\Theta}\right\| = o_p\left(n^{-1/2}\right) + \mathcal{O}_p(1) = \mathcal{O}_p(1).$$

Combining the above two displays,

$$\sqrt{n}\left\|\left(W^TW\right)^{-1}W^T\widehat{X}\right\|\left\|Q^T\widetilde{\beta}_{\mathrm{x}}-\widehat{\beta}_{\mathrm{x}}\right\| = o_p(1).$$

Using this and Equation (20) to bound Equation (19) completes the proof. ■

E.2 Technical Preliminaries for Supporting Lemmas

The main technical components of our proofs are a series of concentration bounds similar to those in Levin et al. (2022). See Athreya et al. (2018) for an overview of proof techniques specialized to the RDPG setting.

Many of our results rely on the concentration $\widehat{X}$ around X and A around P.

Lemma 31 (Levin et al. (2022), Theorem 6) *Under Assumptions 2 and 5, with probability at least $1-\mathcal{O}\big(n^{-2}\big)$, there exists an orthogonal matrix $Q\in\mathbb{R}^{d\times d}$ such that*

$$\left\|\widehat{X}-XQ\right\|_{2,\infty}\leq\eta_n \tag{21}$$

where η_n is defined to be

$$\eta_n=\frac{C\,d}{\lambda_d^{1/2}}(\nu_n+b_n^2)^{1/2}\log n+\frac{C\,d\,n\,\lambda_1}{\lambda_d^{5/2}}\big(\nu_n+b_n^2\big)\log^2 n. \tag{22}$$

While Lemma 31 holds for any model satisfying Assumptions 2 and Assumption 5, the result is not very interesting unless η_n is $o(1)$. η_n must be $o(1)$ for convergence of $\widehat{\beta}$ and $\widehat{\Theta}$ in the general sub-gamma case. In the special case of a random dot product graph (Example 1), one can show that $\eta_n=\mathcal{O}_p\big(n^{-1/2}\log n\big)$. Under our growth assumptions outlined in Assumption 6, $\eta_n=o(1)$, as the next lemma shows.

Lemma 32 *Letting η_n be as defined in Lemma 31, under the growth conditions of Assumption 6, we have $\eta_n=o(1)$. Further, under the additional Assumptions 2 and 5,*

$$\left\|\widehat{X}-XQ\right\|_{2,\infty}=o_p(1).$$

Proof By Lemma 31 and the Borel-Cantelli lemma, there exists a sequence of orthogonal matrices $Q\in\mathbb{R}^{d\times d}$ such that, eventually,

$$\left\|\widehat{X}-XQ\right\|_{2,\infty}\leq\eta_n. \tag{23}$$

Applying the definition of η_n given in Lemma 31,

$$\eta_n=\frac{C\sqrt{\nu_n+b_n^2}\log n}{\lambda_d^{1/2}}+\frac{C\lambda_1(\nu_n+b_n^2)n\log^2 n}{\lambda_d^{5/2}}. \tag{24}$$

Using the fact that $\lambda_d \leq \lambda_1$ by definition and applying our growth assumptions in Equations (6)(8),

$$\frac{C\sqrt{\nu_n + b_n^2}\log n}{\lambda_d^{1/2}} = o(1).$$

Similarly, using $\lambda_d \leq \lambda_1 = \mathcal{O}(n)$ and Equation (9),

$$\frac{C\lambda_1(\nu_n + b_n^2)n\log^2 n}{\lambda_d^{5/2}} = o(1).$$

Applying the above two displays to Equation (24),

$$\eta_n = o(1),$$

establishing our desired growth rate on η_n. Applying this to Equation (23),

$$\left\|\widehat{X} - XQ\right\|_{2,\infty} = o_p(1),$$

as we set out to show. ■

Several other technical results will also prove useful. We collect them below.

Lemma 33 *Under Assumptions 2 and 5, with probability at least* $1 - \mathcal{O}\big(n^{-2}\big)$ *we have*

$$\left\|\widehat{S}^{-1/2}\right\| \leq C\lambda_d^{-1/2} \quad \textit{and} \quad \left\|\widehat{S}^{1/2}\right\| \leq C\lambda_1^{1/2}$$

for some universal constant $C > 0$.

Proof Both of these facts are shown in the course of proving Lemma 4 of Levin et al. (2022), in particular see Equations (28) and (32) in that work. ■

The following two lemmas are fundamental for determining rates of concentration throughout our proofs. Our goal is to produce bounds under very general assumptions on A, and as a result, under additional assumptions, it will often be possible to improve rates of convergence under specialized assumptions. For example, under the additional assumption that A is binary, (Lei and Rinaldo, 2015, Theorem 5.2) produces a notable improvement over a generic sub-gamma bound. We do not pursue specialized bounds here.

Lemma 34 (Levin et al. (2022), Lemma 5, taking $N = 1$) *Under Assumption 2 and 5, with probability at least* $1 - \mathcal{O}\big(n^{-2}\big)$,

$$\|A - P\| \leq C\sqrt{\nu_n + b_n^2}\sqrt{n}\log n.$$

Lemma 35 *Suppose that Assumptions 2 and 5 hold and let* $H \in \mathbb{R}^{n\times n}$ *be a fixed matrix satisfying*

$$\max_{i\in[n]} \sum_{j=1}^{n} H_{ij}^2 \leq C_H \tag{25}$$

for some constant $C_H \geq 0$. Then, with notation as above, with probability at least $1 - \mathcal{O}(n^{-2})$,

$$\left\|U^T(A-P)HU\right\|_F \leq Cd\sqrt{\nu_n + b_n^2}\log n$$

Proof We will show that $\left\|U^T(A-P)HU\right\|_F^2 \geq Cd^2(\nu_n + b_n^2)\log^2 n$ with probability no larger than $\mathcal{O}(n^{-2})$, whence taking square roots will yield the result.

For each $k, \ell \in [d]$, define

$$S_{k,\ell} = \left[U^T(A-P)HU\right]_{k,\ell} = \sum_{i=1}^n \sum_{j=1}^n (A-P)_{ij} U_{ik} (HU)_{j\ell}$$

and note that

$$\left\|U^T(A-P)HU\right\|_F^2 = \sum_{k=1}^d \sum_{\ell=1}^d S_{k,\ell}^2. \tag{26}$$

Since $(A-P)_{ij}$ are i.i.d. (ν_n, b_n)-sub-gamma, we have

$$\sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\Big[[(A-P)_{ij}U_{ik}(HU)_{j\ell}]^2\Big] < \nu_n \sum_{i=1}^n \sum_{j=1}^n U_{ik}^2 (HU)_{j\ell}^2$$

and thus by Corollary 2.11 in Boucheron et al. (2013), for any $t > 0$,

$$\mathbb{P}(|S_{k,\ell}| \geq t) \leq 2\exp\left\{\frac{-t^2}{2\Big(\nu_n \sum_{i=1}^n \sum_{j=1}^n U_{ik}^2 (HU)_{j\ell}^2 + b_n t\Big)}\right\}.$$

By Cauchy-Schwarz and our assumption in Equation (25),

$$(HU)_{j\ell}^2 = \left(\sum_{t=1}^n H_{jt}U_{t\ell}\right)^2 \leq \left(\sum_{t=1}^n H_{jt}^2\right)\left(\sum_{t=1}^n U_{t\ell}^2\right) \leq C_H,$$

and it follows that

$$\mathbb{P}(|S_{k,\ell}| \geq t) \leq 2\exp\left\{\frac{-t^2}{2(C_H\nu_n + b_n t)}\right\}.$$

Taking $t = C(\nu_n + b_n^2)^{1/2}\log n$ for $C > 0$ suitably large, it follows that

$$\mathbb{P}\Big(|S_{k,\ell}| \geq C(\nu_n + b_n^2)^{1/2}\log n\Big) \leq 2n^{-4}.$$

A union bound over all $k, \ell \in [d]$ implies that

$$\mathbb{P}\Big(\exists\ k, \ell \in [d] : |S_{k,\ell}| \geq C(\nu_n + b_n^2)^{1/2}\log n\Big) \leq \frac{2d^2}{n^4} \leq 2n^{-2},$$

and it follows from Equation (26) that

$$\mathbb{P}\Big(\left\|U^T(A-P)HU\right\|_F^2 \geq Cd^2(\nu_n + b_n^2)\log^2 n\Big) \leq 2n^{-2},$$

completing the proof. ■

We now define of a convenient decomposition of $\widehat{X} - XQ$.

Lemma 36 (Levin et al. 2022, Lemma 4) *Define the following three matrices:*

$$\begin{aligned} R_1 &= UU^T\widehat{U} - UQ \\ R_2 &= Q\widehat{S}^{1/2} - S^{1/2}Q \\ R_3 &= \widehat{U} - UU^T\widehat{U} + R_1 = \widehat{U} - UQ. \end{aligned}$$

Then

$$\begin{aligned} \widehat{X} - XQ &= \widehat{U}\widehat{S}^{1/2} - US^{1/2}Q \\ &= (A-P)US^{-1/2}Q + (A-P)U(Q\widehat{S}^{-1/2} - S^{-1/2}Q) \\ &\quad + UU^T(A-P)UQ\widehat{S}^{-1/2} + R_1\widehat{S}^{1/2} + UR_2 \\ &\quad + (I - UU^T)(A-P)R_3\widehat{S}^{-1/2}. \end{aligned}$$

Our proofs will rely on bounding each of the terms in the decomposition given in Lemma 36. The next few technical results will be used to ensure these bounds.

Proposition 37 (Levin et al. 2022, Proposition 19) *With notation as above, under Assumptions 2 and 5, it holds with probability at least* $1 - \mathcal{O}\big(n^{-2}\big)$ *that*

$$\|R_1\|_F = \left\|U\Big(U^T\widehat{U} - Q\Big)\right\|_F = \left\|U^T\widehat{U} - Q\right\|_F \leq \frac{d\|A-P\|^2}{\lambda_d^2} \leq \frac{Cd(\nu_n + b_n^2)\, n\log^2 n}{\lambda_d^2}.$$

Proof By Proposition 19 of Levin et al. (2022) and Lemma 34. ■

Lemma 38 *Under Assumptions 2 and 5, with probability at least* $1 - \mathcal{O}\big(n^{-2}\big)$,

$$\left\|\widehat{U} - UU^T\widehat{U}\right\|_F \leq \frac{C\sqrt{d}\ \|A-P\|}{\lambda_d} \leq \frac{C\sqrt{d}\sqrt{\nu_n + b_n^2}\sqrt{n}\log n}{\lambda_d}. \tag{27}$$

Furthermore,

$$\left\|Q\widehat{S} - SQ\right\|_F \leq \frac{C\lambda_1(\nu_n + b_n^2)n\log^2 n}{\lambda_d^2} + Cd\sqrt{\nu_n + b_n^2}\log n \tag{28}$$

$$\|R_2\|_F = \left\|Q\widehat{S}^{1/2} - S^{1/2}Q\right\|_F \leq \frac{C\lambda_1(\nu_n + b_n^2)n\log^2 n}{\lambda_d^{5/2}} + \frac{Cd\sqrt{\nu_n + b_n^2}\log n}{\lambda_d^{1/2}} \quad \textit{and,} \tag{29}$$

$$\left\|Q\widehat{S}^{-1/2} - S^{-1/2}Q\right\|_F \leq \frac{C\lambda_1(\nu_n + b_n^2)n\log^2 n}{\lambda_d^{7/2}} + \frac{Cd\sqrt{\nu_n + b_n^2}\log n}{\lambda_d^{3/2}} \tag{30}$$

Proof By Proposition 20 of Levin et al. (2022) and an application of Lemma 34 we obtain (27). Further, by Proposition 20 of Levin et al. (2022), we have

$$\begin{aligned}
\left\|Q\widehat{S} - SQ\right\|_F &\leq \frac{C\|A-P\|^2\lambda_1}{\lambda_d^2} + \left\|U^T(A-P)U\right\|_F \\
\left\|Q\widehat{S}^{1/2} - S^{1/2}Q\right\|_F &\leq \frac{\left\|Q\widehat{S} - SQ\right\|_F}{\lambda_d^{1/2}}, \\
\text{and } \left\|Q\widehat{S}^{-1/2} - S^{-1/2}Q\right\|_F &\leq \frac{\left\|Q\widehat{S} - SQ\right\|_F}{\lambda_d^{3/2}}.
\end{aligned}$$

First we apply Lemma 34 and Lemma 35 to bound the top term

$$\begin{aligned}
\left\|Q\widehat{S} - SQ\right\|_F &\leq \frac{C\|A-P\|^2\lambda_1}{\lambda_d^2} + \left\|U^T(A-P)U\right\|_F \\
&\leq \frac{C(\nu_n + b_n^2)n\log^2 n\lambda_1}{\lambda_d^2} + Cd\sqrt{\nu_n + b_n^2}\log n
\end{aligned}$$

and Equations (29) and (30) follow immediately. ■

Lemma 39 *Under Assumptions 2 and 5, it holds with probability* $1 - \mathcal{O}\big(n^{-2}\big)$ *that*

$$\left\|\widehat{U} - UQ\right\| = \|R_3\| \leq \frac{C\sqrt{d}\sqrt{\nu_n + b_n^2}\sqrt{n}\log n}{\lambda_d} + \frac{Cd(\nu_n + b_n^2)\, n\log^2 n}{\lambda_d^2}$$

Proof Adding and subtracting appropriate quantities, applying the triangle inequality and using basic properties of the Frobenius norm,

$$\begin{aligned}
\left\|\widehat{U} - UQ\right\| &\leq \left\|\widehat{U} - UU^T\widehat{U}\right\| + \left\|UU^T\widehat{U} - UQ\right\|_F \\
&\leq \left\|\widehat{U} - UU^T\widehat{U}\right\| + \left\|U^T\widehat{U} - Q\right\|_F.
\end{aligned}$$

Applying Lemmas 37 and 38, it follows that with probability at least $1 - \mathcal{O}\big(n^{-2}\big)$,

$$\left\|\widehat{U} - UQ\right\| = \|R_3\| \leq \frac{C\sqrt{d}\sqrt{\nu_n + b_n^2}\sqrt{n}\log n}{\lambda_d} + \frac{Cd(\nu_n + b_n^2)\, n\log^2 n}{\lambda_d^2},$$

completing the proof. ■

E.3 Supporting Results for Lemma 25 and Lemma 26

When we introduced Lemma 25 and Lemma 26 earlier, we presented a broad proof sketch and deferred the technical details to supporting lemmas, which we now present.

Lemma 40 *Suppose that Assumptions 2, 5 and 6 hold. Then*

$$\left\|(\widehat{X}Q^T - X)^T MX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right).$$

Proof Applying Lemma 36,

$$\begin{aligned}
&(\widehat{X}Q^T - X)^T MX \\
&\quad = Q\left(\widehat{U}\widehat{S}^{1/2} - US^{1/2}Q\right)^T MX \\
&\quad = QS^{-1/2}U^T(A-P)MX + Q(Q\widehat{S}^{-1/2} - S^{-1/2}Q)^T U^T(A-P)MX \\
&\qquad + Q\widehat{S}^{-1/2}Q^T U^T(A-P)UU^T MX + Q\widehat{S}^{1/2}R_1^T MX + QR_2^T U^T MX \\
&\qquad + Q\widehat{S}^{-1/2}R_3^T(I - UU^T)(A-P)MX.
\end{aligned} \tag{31}$$

We will bound each of the six terms on the right-hand side in turn.

Considering the first term, expanding the definition of X and using submultiplicativity of the norm, with probability $1 - \mathcal{O}(n^{-2})$,

$$\begin{aligned}
\left\|QS^{-1/2}U^T(A-P)MX\right\| &\leq \left\|S^{-1/2}\right\|\left\|U^T(A-P)MU\right\|\left\|S^{1/2}\right\| \\
&\leq \frac{Cd\lambda_1^{1/2}\sqrt{\nu_n + b_n^2}\log n}{\lambda_d^{1/2}},
\end{aligned}$$

where the second bound follows from Lemma 35. Applying Equations (6) and (8),

$$\left\|QS^{-1/2}U^T(A-P)MX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{32}$$

For the second term in Equation (31), we again use submultiplicativity of the spectral norm, along with Equation (30) from Lemma 38 and Lemma 35, to show that with probability $1 - \mathcal{O}(n^{-2})$,

$$\begin{aligned}
&\left\|Q(Q\widehat{S}^{-1/2} - S^{-1/2}Q)^T U^T(A-P)MX\right\| \\
&\quad \leq \left\|Q\widehat{S}^{-1/2} - S^{-1/2}Q\right\|\left\|U^T(A-P)MU\right\|\left\|S^{1/2}\right\| \\
&\quad \leq C\lambda_1^{1/2}\left(\frac{\lambda_1(\nu_n + b_n^2)n\log^2 n}{\lambda_d^{7/2}} + \frac{d\sqrt{\nu_n + b_n^2}\log n}{\lambda_d^{3/2}}\right)d\sqrt{\nu_n + b_n^2}\log n \\
&\quad \leq \frac{Cd\lambda_1^{3/2}(\nu_n + b_n^2)^{3/2}n\log^3 n}{\lambda_d^{7/2}} + \frac{Cd^2\lambda_1^{1/2}(\nu_n + b_n^2)\log^2 n}{\lambda_d^{3/2}}.
\end{aligned}$$

Applying Equation (6) and cubing the quantity in Equation (8) implies that the first of these two right-hand side quantities is $o(\lambda_d/\sqrt{n})$. Similarly, Equations (6) and (9) imply that the second right-hand term is $o(\lambda_d/\sqrt{n})$, whence

$$\left\|Q(Q\widehat{S}^{-1/2} - S^{-1/2}Q)^T U^T(A-P)MX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{33}$$

For the third term in Equation (31), by Lemmas 33 and 35, we have

$$\begin{aligned}\left\|Q\widehat{S}^{-1/2}Q^TU^T(A-P)UU^TMX\right\| &\le \left\|\widehat{S}^{-1/2}\right\|\left\|U^T(A-P)U\right\|\left\|S^{1/2}\right\| \\ &\le \frac{Cd\lambda_1^{1/2}\sqrt{\nu_n+b_n^2}\log n}{\lambda_d^{1/2}}.\end{aligned}$$

Applying Equations (6) and (8) along with the trivial $\log n = \Omega(1)$, we have

$$\left\|Q\widehat{S}^{-1/2}Q^TU^T(A-P)UU^TMX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{34}$$

For the fourth term in Equation (31), recalling $R_1 = UU^T\widehat{U} - UQ = U(U^T\widehat{U} - Q)$, observe that

$$Q\widehat{S}^{1/2}R_1^TMX = Q\widehat{S}^{1/2}(U^T\widehat{U}-Q)^TU^TMUS^{1/2},$$

whence Lemma 33 and Proposition 37 imply that

$$\begin{aligned}\left\|Q\widehat{S}^{1/2}R_1^TMX\right\| &\le \left\|\widehat{S}^{1/2}\right\|\left\|U^T\widehat{U}-Q\right\|_F\|MU\|\left\|S^{1/2}\right\| \\ &\le \frac{Cd\lambda_1(\nu_n+b_n^2)n\log^2 n}{\lambda_d^2}.\end{aligned}$$

Using Equation (6) and the trivial $\lambda_1 \ge \lambda_d$, Equation (9) then implies

$$\left\|Q\widehat{S}^{1/2}R_1^TMX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{35}$$

Similarly, for the fifth term in Equation (31), applying submultiplicativity followed by Equation (29) and Lemma 33,

$$\left\|QR_2^TU^TMX\right\| \le \|R_2\|\left\|S^{1/2}\right\| \le \frac{C\lambda_1^{3/2}(\nu_n+b_n^2)n\log^2 n}{\lambda_d^{5/2}} + \frac{Cd\lambda_1^{1/2}\sqrt{\nu_n+b_n^2}\log n}{\lambda_d^{1/2}}.$$

Equations (6) and (9) control the first of these terms, while Equations (6) and (8) control the second, and we conclude that

$$\left\|QR_2^TU^TMX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{36}$$

For the sixth term in Equation (31), we expand R_3 to write

$$\begin{aligned}&Q\widehat{S}^{-1/2}R_3^T(I-UU^T)(A-P)MX \\ &\quad= Q\widehat{S}^{-1/2}(\widehat{U}-UU^T\widehat{U})^T(I-UU^T)(A-P)MX \qquad (37)\\ &\qquad+ Q\widehat{S}^{-1/2}(UU^T\widehat{U}-UQ)^T(I-UU^T)(A-P)MX. \qquad (38)\end{aligned}$$

To bound (37), we use submultiplicativity, Lemma 33, (27) of Lemma 38, and Lemma 34 to write

$$\begin{aligned}
&\left\|Q\widehat{S}^{-1/2}(\widehat{U}-UU^T\widehat{U})^T(I-UU^T)(A-P)MX\right\| \\
&\quad\le C\left\|\widehat{S}^{-1/2}\right\|\left\|\widehat{U}-UU^T\widehat{U}\right\|\|A-P\|\left\|S^{1/2}\right\| \\
&\quad\le C\lambda_d^{-1/2}\frac{C\sqrt{d}\sqrt{\nu_n+b_n^2}\sqrt{n}\log n}{\lambda_d}C\lambda_1^{1/2}\sqrt{\nu_n+b_n^2}\sqrt{n}\log n \\
&\quad\le \frac{C\sqrt{d}\lambda_1^{1/2}(\nu_n+b_n^2)n\log^2 n}{\lambda_d^{3/2}},
\end{aligned}$$

Applying Equations (6) and (9),

$$\left\|Q\widehat{S}^{-1/2}(\widehat{U}-UU^T\widehat{U})^T(I-UU^T)(A-P)MX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{39}$$

To bound (38), we apply Lemma 33, Proposition 37 and Lemma 34 to see that

$$\begin{aligned}
&\left\|Q\widehat{S}^{-1/2}(UU^T\widehat{U}-UQ)^T(I-UU^T)(A-P)MX\right\| \\
&\quad\le \left\|\widehat{S}^{-1/2}\right\|\left\|U^T\widehat{U}-Q\right\|\|A-P\|\|MU\|\left\|S^{1/2}\right\| \\
&\quad\le C\lambda_d^{-1/2}\frac{Cd(\nu_n+b_n^2)\,n\log^2 n}{\lambda_d^2}\lambda_1^{1/2}\sqrt{\nu_n+b_n^2}\sqrt{n}\log n \\
&\quad\le \frac{C\lambda_1^{1/2}d(\nu_n+b_n^2)^{3/2}\,n^{3/2}\log^3 n}{\lambda_d^{5/2}}.
\end{aligned}$$

Applying Equations (6), (8) and (9),

$$\left\|Q\widehat{S}^{-1/2}(UU^T\widehat{U}-UQ)^T(I-UU^T)(A-P)MX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{40}$$

Using Equations (39) and (40), respectively, to bound (37) and (38), we conclude that

$$\left\|Q\widehat{S}^{-1/2}R_3^T(I-UU^T)(A-P)MX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{41}$$

Applying Equations (32), (33), (34), (35), (36) and (41) to the right-hand side of Equation (31), we conclude that

$$\left\|(\widehat{X}Q^T-X)^TMX\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right),$$

completing the proof. ■

Lemma 41 *Under Assumptions 2, 5 and 6, it holds with high probability that*

$$\left\|XQ-\widehat{X}\right\| = \mathcal{O}_p\left(\frac{\lambda_1(\nu_n+b_n^2)n\log^2 n}{\lambda_d^{5/2}}\right)+\mathcal{O}_p\left(\frac{\lambda_1^{1/2}\sqrt{\nu_n+b_n^2}\sqrt{n}\log n}{\lambda_d}\right).$$

Proof Using basic properties of the norm and the triangle inequality,

$$\begin{aligned}\left\|XQ-\widehat{X}\right\| &\leq \left\|U(S^{1/2}Q-Q\widehat{S}^{1/2})\right\| + \left\|UQ\widehat{S}^{1/2}-\widehat{U}\widehat{S}^{1/2}\right\| \\ &\leq \left\|S^{1/2}Q-Q\widehat{S}^{1/2}\right\| + \left\|UQ-\widehat{U}\right\|\left\|\widehat{S}^{1/2}\right\|.\end{aligned} \tag{42}$$

By Lemma 38, it holds with high probability that

$$\left\|S^{1/2}Q-Q\widehat{S}^{1/2}\right\| \leq \frac{C\lambda_1(\nu_n+b_n^2)n\log^2 n}{\lambda_d^{5/2}} + \frac{Cd\sqrt{\nu_n+b_n^2}\log n}{\lambda_d^{1/2}}$$

Applying Lemmas 33 and 39, it holds with high probability that

$$\left\|UQ-\widehat{U}\right\|\left\|\widehat{S}^{1/2}\right\| \leq \frac{C\sqrt{d}\lambda_1^{1/2}\sqrt{\nu_n+b_n^2}\sqrt{n}\log n}{\lambda_d} + \frac{Cd\lambda_1^{1/2}(\nu_n+b_n^2)\,n\log^2 n}{\lambda_d^2}.$$

Applying the above two displays to Equation (42),

$$\begin{aligned}\left\|XQ-\widehat{X}\right\| = \mathcal{O}_p\left(\frac{\lambda_1(\nu_n+b_n^2)n\log^2 n}{\lambda_d^{5/2}}\right) + \mathcal{O}_p\left(\frac{\sqrt{\nu_n+b_n^2}\log n}{\lambda_d^{1/2}}\right) \\ + \mathcal{O}_p\left(\frac{\lambda_1^{1/2}\sqrt{\nu_n+b_n^2}\sqrt{n}\log n}{\lambda_d}\right) + \mathcal{O}_p\left(\frac{\lambda_1^{1/2}(\nu_n+b_n^2)n\log^2 n}{\lambda_d^2}\right).\end{aligned}$$

Collecting terms and using the fact that $\lambda_d \leq \lambda_1$,

$$\left\|XQ-\widehat{X}\right\| = \mathcal{O}_p\left(\frac{\lambda_1(\nu_n+b_n^2)n\log^2 n}{\lambda_d^{5/2}}\right) + \mathcal{O}_p\left(\frac{\lambda_1^{1/2}\sqrt{\nu_n+b_n^2}\sqrt{n}\log n}{\lambda_d}\right),$$

completing the proof.

■

Lemma 42 *Under Assumptions 2, 5 and 6, for any orthogonal projection matrix M,*

$$\left\|Q^TX^TMX-\widehat{X}^TM\widehat{X}\,Q^T\right\| = o_p\left(\frac{\lambda_d^2}{\sqrt{n}\lambda_1^{1/2}}\right).$$

Proof Adding and subtracting appropriate quantities, applying the triangle inequality and using basic properties of the spectral norm,

$$\begin{aligned}\left\|Q^TX^TMX-\widehat{X}^TM\widehat{X}\,Q^T\right\| &\leq \left\|(XQ)^TM(X-\widehat{X}Q^T)\right\| + \left\|(XQ-\widehat{X})^TM(\widehat{X}Q^T-X)\right\| \\ &\quad + \left\|(XQ-\widehat{X})^TMX\right\| \\ &\leq 2\left\|(\widehat{X}Q^T-X)^TMX\right\| + \left\|XQ-\widehat{X}\right\|^2\|M\|.\end{aligned}$$

Using the fact that $\|M\| \leq 1$ and applying Lemma 40,

$$\left\|Q^T X^T M X - \widehat{X}^T M \widehat{X}\, Q^T\right\| \leq \left\|XQ - \widehat{X}\right\|^2 + o_p\left(\frac{\lambda_d}{\sqrt{n}}\right) \tag{43}$$

By Lemma 41,

$$\left\|XQ - \widehat{X}\right\|^2 = \mathcal{O}_p\left(\frac{\lambda_1^2(\nu_n + b_n^2)^2 n 2\log^4 n}{\lambda_d^5}\right) + \mathcal{O}_p\left(\frac{\lambda_1(\nu_n + b_n^2) n \log^2 n}{\lambda_d^2}\right).$$

Applying our growth assumption in Equation (9) controls the first of these terms as $o(\lambda_d/\sqrt{n})$, while Equations (7) and (9) control the second of these terms as $o(\lambda_d/\sqrt{n})$, so that

$$\left\|XQ - \widehat{X}\right\|^2 = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right).$$

Applying this bound to Equation (43),

$$\left\|Q^T X^T M X - \widehat{X}^T M \widehat{X}\, Q^T\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right).$$

Applying our assumption in Equation (10) completes the proof. ■

Lemma 43 *Under Assumption 5, for any (possibly random) matrix H independent of ε,*

$$\|H\varepsilon\| = \mathcal{O}_p\left(\sqrt{B \operatorname{trace} H^T H}\right).$$

In particular, taking $H = I$, $\|\varepsilon\| = \mathcal{O}_p\left(\sqrt{Bn}\right)$.

Proof We begin by noting that since ε is a vector of independent mean-zero random variables,

$$\mathbb{E}\|H\varepsilon\|^2 = \mathbb{E}\varepsilon^T H^T H \varepsilon \leq B \operatorname{trace} H^T H,$$

where $B > 0$ is the bound on the variance guaranteed by Assumption 5. Applying Markov's inequality, for any $t > 0$ and $\delta > 0$,

$$\mathbb{P}\left(\frac{\|H\varepsilon\|^2}{t} > \delta\right) \leq \frac{\mathbb{E}\|H\varepsilon\|^2}{t\delta} \leq \frac{B \operatorname{trace} H^T H}{t\delta}.$$

Let r_n be any function of n growing such that $r_n = \omega(B \operatorname{trace} H^T H)$. Then taking $t = r_n$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\|H\varepsilon\|^2}{r_n} > \delta\right) = 0.$$

Thus, $\|H\varepsilon\|^2 = o_p(r_n)$ for any $r_n = \omega(B \operatorname{trace} H^T H)$, and it follows that

$$\|H\varepsilon\|^2 = \mathcal{O}_p\big(B \operatorname{trace} H^T H\big).$$

Taking square roots completes the proof. ■

Lemma 44 *Under Assumptions 2 and 5, let $M \in \mathbb{R}^{n\times n}$ satisfy $\|M\| = 1$. Then*

$$\left\|U^T(A-P)M\varepsilon\right\| = \mathcal{O}_p\Big(\sqrt{d}\log n\sqrt{\nu_n Bn\log n + b_n^2}\Big).$$

Proof For each $k \in [d]$, define

$$S_k = \left[U^T(A-P)M\varepsilon\right]_k = \sum_{i=1}^n\sum_{j=1}^n (A-P)_{ij}U_{ik}(M\varepsilon)_j$$

and note that

$$\left\|U^T(A-P)M\varepsilon\right\|_2^2 = \sum_{k=1}^d S_k^2.$$

By Corollary 2.11 in Boucheron et al. (2013), for any $t > 0$,

$$\begin{aligned}
\mathbb{P}(|S_k| \geq t \,|\, X, \varepsilon) &\leq 2\exp\left\{\frac{-t^2}{2\Big(\nu_n \sum_{i=1}^n\sum_{j=1}^n U_{ik}^2(M\varepsilon)_j^2 + b_n t\Big)}\right\} \\
&= 2\exp\left\{\frac{-t^2}{2\Big(\nu_n \sum_{j=1}^n (M\varepsilon)_j^2 + b_n t\Big)}\right\} \\
&= 2\exp\left\{\frac{-t^2}{2(\nu_n\|M\varepsilon\|^2 + b_n t)}\right\} \\
&\leq 2\exp\left\{\frac{-t^2}{2(\nu_n\|\varepsilon\|^2 + b_n t)}\right\}.
\end{aligned}$$

Note that we can drop the conditioning on X in the above since the bound is free of terms that depend on X. We now need to drop the conditioning on ε. Let G_n denote the event $\left\{\|\varepsilon\|^2 < nB\log n\right\}$ and G_n^c denote the complement of G_n. By a slight modification of the proof of Lemma 43 with $H = I$, G_n occurs with probability at least $1 - \frac{1}{\log n}$. Thus

$$\begin{aligned}
\mathbb{P}(|S_k| \geq t) &= \mathbb{P}(|S_k| \geq t \,|\, G_n)\cdot\mathbb{P}(G_n) + \mathbb{P}(|S_k| \geq t \,|\, G_n^c)\cdot\mathbb{P}(G_n^c) \\
&\leq \mathbb{P}(|S_k| \geq t \,|\, G_n) + \mathbb{P}(G_n^c) \\
&\leq \mathbb{P}(|S_k| \geq t \,|\, G_n) + \frac{1}{\log n}
\end{aligned}$$

using our previous bound on $\mathbb{P}(|S_k| \geq t \,|\, X, \varepsilon)$. Let $\delta > 0$ be arbitrary. Taking $t = C\log n\big(\sqrt{\nu_n Bn\log n} + b_n\big)\delta$ for $C > 0$ suitably large, it follows that

$$\mathbb{P}\Big(|S_k| \geq C\log n\Big(\sqrt{\nu_n Bn\log n} + b_n\Big)\delta \,\Big|\, G_n\Big) \leq 2n^{-3}.$$

A union bound over all $k \in [d]$ implies that

$$\mathbb{P}\Big(\Big\{\exists k \in [d] : |S_k| \geq C\log n\Big(\sqrt{\nu_n Bn\log n} + b_n\Big)\delta\Big\} \,\Big|\, G_n\Big) \leq \frac{2d}{n^3} \leq 2n^{-2}$$

from we which we see that

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{\exists k \in [d] : |S_k| \geq C \log n \Big(\sqrt{\nu_n B n \log n} + b_n\Big)\delta\Big\}\Big) \\
&\quad = \mathbb{P}\Big(\Big\{\exists k \in [d] : |S_k| \geq C \log n \Big(\sqrt{\nu_n B n \log n} + b_n\Big)\delta\Big\} \,\Big|\, G_n\Big) \cdot \mathbb{P}(G_n) \\
&\qquad + \mathbb{P}\Big(\Big\{\exists k \in [d] : |S_k| \geq C \log n \Big(\sqrt{\nu_n B n \log n} + b_n\Big)\delta\Big\} \,\Big|\, G_n^c\Big) \cdot \mathbb{P}(G_n^c) \\
&\quad \leq \mathbb{P}\Big(\Big\{\exists k \in [d] : |S_k| \geq C \log n \Big(\sqrt{\nu_n B n \log n} + b_n\Big)\delta\Big\} \,\Big|\, G_n\Big) + \mathbb{P}(G_n^c) \\
&\quad \leq \frac{2}{n^2} + \frac{1}{\log n}.
\end{aligned}$$

Thus, $\big\|U^T(A-P)M\varepsilon\big\|_2^2 = \sum_{k=1}^d S_k^2 \geq Cd\big(\nu_n B n \log n + b_n^2\big)\delta^2 \log^2 n$ with probability $2n^{-2} + 1/\log n$; that is,

$$\big\|U^T(A-P)M\varepsilon\big\| = o_p\Big(\sqrt{d}\sqrt{\nu_n B n \log n + b_n^2}\log n\Big),$$

completing the proof. ■

Lemma 45 *Suppose that Assumptions 2, 5 and 6 hold. Then*

$$\Big\|(\widehat{X}Q^T - X)^T M\varepsilon\Big\| = o_p\Big(\frac{\lambda_d}{\sqrt{n}}\Big).$$

Proof Applying Lemma 36 to expand $\widehat{X} - XQ$,

$$\begin{aligned}
&(\widehat{X}Q^T - X)^T M\varepsilon \\
&\quad = Q\Big(\widehat{U}\widehat{S}^{1/2} - US^{1/2}Q\Big)^T M\varepsilon \\
&\quad = QS^{-1/2}U^T(A-P)M\varepsilon + Q(Q\widehat{S}^{-1/2} - S^{-1/2}Q)^T U^T(A-P)M\varepsilon \\
&\qquad + Q\widehat{S}^{-1/2}Q^T U^T(A-P)UU^T M\varepsilon + Q\widehat{S}^{1/2}R_1^T M\varepsilon + QR_2^T U^T M\varepsilon \\
&\qquad + Q\widehat{S}^{-1/2}R_3^T(I - UU^T)(A-P)M\varepsilon.
\end{aligned} \tag{44}$$

We will bound each of the six terms on the right-hand side in turn. In the first term, expanding the definition of X and using submultiplicativity of the norm,

$$\Big\|QS^{-1/2}U^T(A-P)M\varepsilon\Big\| \leq \Big\|S^{-1/2}\Big\|\big\|U^T(A-P)M\varepsilon\big\|.$$

Applying Lemmas 33 and 44,

$$\begin{aligned}
\Big\|QS^{-1/2}U^T(A-P)M\varepsilon\Big\| &= \mathcal{O}_p\left(\frac{(\nu_n B n \log n + b_n^2)^{1/2}\log n}{\lambda_d^{1/2}}\right) \\
&= \mathcal{O}_p\left(\frac{\sqrt{\nu_n + b_n^2}\sqrt{n}\log^{3/2} n}{\lambda_d^{1/2}}\right),
\end{aligned}$$

where we have used the trivial upper bound

$$\nu_n n \log n + b_n^2 \leq n(\nu_n + b_n^2) \log n \tag{45}$$

along with the assumption that d and B are constant in n. Equation (8) then implies that

$$\left\| Q S^{-1/2} U^T (A - P) M \varepsilon \right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}} \right). \tag{46}$$

For the second term in Equation (44), we use submultiplicativity of the spectral norm again to write

$$\left\| Q(Q\widehat{S}^{-1/2} - S^{-1/2}Q)^T U^T (A - P) M \varepsilon \right\| \leq \left\| Q\widehat{S}^{-1/2} - S^{-1/2}Q \right\| \left\| U^T (A - P) M \varepsilon \right\|.$$

Equation (30) from Lemma 38 bounds the first multiplicand on the right-hand side, while Lemma 44 bounds the second, and we have

$$\begin{aligned}
&\left\| Q(Q\widehat{S}^{-1/2} - S^{-1/2}Q)^T U^T (A - P) M \varepsilon \right\| \\
&\quad \leq C \left(\frac{\lambda_1 (\nu_n + b_n^2) n \log^2 n}{\lambda_d^{7/2}} + \frac{\sqrt{\nu_n + b_n^2} \log n}{\lambda_d^{3/2}} \right) o_p\left(\sqrt{\nu_n B n \log n + b_n^2} \log n \right) \\
&\quad = o_p\left(\frac{\lambda_1 (\nu_n + b_n^2)^{3/2} n^{3/2} \log^{7/2} n}{\lambda_d^{7/2}} \right) + o_p\left(\frac{(\nu_n + b_n^2) n^{1/2} \log^{5/2} n}{\lambda_d^{3/2}} \right)
\end{aligned}$$

where we have again used the bound in Equation (45) and our assumption that B and d are constants. Raising the quantity in Equation (8) to the third power and applying Equation (6), the first of these as $o_p(\lambda_d/\sqrt{n})$. Trivially using $\lambda_1 \geq \lambda_d$ and $\log n = o(n)$, Equation (9) bounds the second term by the same rate, and it follows that

$$\left\| Q(Q\widehat{S}^{-1/2} - S^{-1/2}Q)^T U^T (A - P) M \varepsilon \right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}} \right). \tag{47}$$

For the third term in Equation (44), submultiplicativity followed by Lemmas 33, 35 and 43 yields

$$\begin{aligned}
\left\| Q\widehat{S}^{-1/2} Q^T U^T (A - P) U U^T M \varepsilon \right\| &\leq \left\| \widehat{S}^{-1/2} \right\| \left\| U^T (A - P) U \right\| \|\varepsilon\| \\
&\leq C \lambda_d^{-1/2} d \sqrt{\nu_n + b_n^2} \|\varepsilon\| \log n \\
&= \mathcal{O}_p\left(\frac{\sqrt{\nu_n + b_n^2} \sqrt{n} \log n}{\lambda_d^{1/2}} \right).
\end{aligned}$$

The trivial bound $\log^{1/2} n = \Omega(1)$ and our growth bound in Equation (8) imply

$$\left\| Q\widehat{S}^{-1/2} Q^T U^T (A - P) U U^T M \varepsilon \right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}} \right). \tag{48}$$

For the fourth term in Equation (44), recalling the definition of $R_1 = UU^T\widehat{U} - UQ = U(U^T\widehat{U} - Q)$, observe that

$$Q\widehat{S}^{1/2}R_1^T M\varepsilon = Q\widehat{S}^{1/2}(U^T\widehat{U} - Q)^T U^T M\varepsilon.$$

Applying submultiplicativity followed by Lemma 33, Proposition 37 and Lemma 43, and using our assumption that B is constant in n,

$$\begin{aligned}\left\|Q\widehat{S}^{1/2}R_1^T M\varepsilon\right\| &\leq \left\|\widehat{S}^{1/2}\right\|\left\|U^T\widehat{U} - Q\right\|_F \|\varepsilon\| \\ &= o_p\left(\frac{\lambda_1^{1/2}(\nu_n + b_n^2)n^{3/2}\log^2 n}{\lambda_d^2}\right).\end{aligned}$$

Using the trivial upper bound $\lambda_1 \geq \lambda_d$ and our growth assumption in Equation (9), it follows that

$$\left\|Q\widehat{S}^{1/2}R_1^T M\varepsilon\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{49}$$

Similarly, for the fifth term in Equation (44), applying submultiplicativity followed by Equation (29) and Lemma 43,

$$\begin{aligned}\left\|QR_2^T U^T M\varepsilon\right\| &\leq \|R_2\|\|\varepsilon\| \\ &\leq C\left[\frac{\lambda_1(\nu_n + b_n^2)n\log^2 n}{\lambda_d^{5/2}} + \frac{(\nu_n + b_n^2)^{1/2}\log n}{\lambda_d^{1/2}}\right]\mathcal{O}_p(\sqrt{n}) \\ &= \mathcal{O}_p\left(\frac{\lambda_1(\nu_n + b_n^2)n^{3/2}\log^2 n}{\lambda_d^{5/2}}\right) + \mathcal{O}_p\left(\frac{\sqrt{\nu_n + b_n^2}\sqrt{n}\log n}{\lambda_d^{1/2}}\right).\end{aligned} \tag{50}$$

Our growth assumption in Equation (9) states that the first of these two rates is $o_p(\lambda_d/\sqrt{n})$. Our bound in Equation (8) along with the trivial bound $\log^{1/2} n = \Omega(1)$ implies that the second is $o_p(\lambda_d/\sqrt{n})$, whence

$$\left\|QR_2^T U^T M\varepsilon\right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{51}$$

For the sixth term in Equation (44), we expand R_3 to write

$$\begin{aligned}&Q\widehat{S}^{-1/2}R_3^T(I - UU^T)(A - P)M\varepsilon \\ &\quad = Q\widehat{S}^{-1/2}(\widehat{U} - UU^T\widehat{U})^T(I - UU^T)(A - P)M\varepsilon \qquad (52) \\ &\qquad + Q\widehat{S}^{-1/2}(UU^T\widehat{U} - UQ)^T(I - UU^T)(A - P)M\varepsilon. \qquad (53)\end{aligned}$$

To bound (52), we use submultiplicativity followed by Lemma 33, Equation (27) of Lemma 38, and Lemma 34 to see

$$\begin{aligned}&\left\|Q\widehat{S}^{-1/2}(\widehat{U} - UU^T\widehat{U})^T(I - UU^T)(A - P)M\varepsilon\right\| \\ &\quad \leq C\left\|\widehat{S}^{-1/2}\right\|\left\|\widehat{U} - UU^T\widehat{U}\right\|\|A - P\|\|\varepsilon\| \\ &\quad \leq \frac{C(\nu_n + b_n^2)n^{3/2}\log^2 n}{\lambda_d^{3/2}}.\end{aligned}$$

Applying the trivial upper bound $\lambda_1 \geq \lambda_d$ and our growth bound in Equation (9),

$$\left\| Q\widehat{S}^{-1/2}(\widehat{U} - UU^T\widehat{U})^T(I - UU^T)(A - P)M\varepsilon \right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{54}$$

To bound (53), we apply submultiplicativity followed by Lemma 33, Proposition 37, Lemma 34 and Lemma 43 to see

$$\begin{aligned}&\left\| Q\widehat{S}^{-1/2}(UU^T\widehat{U} - UQ)^T(I - UU^T)(A - P)M\varepsilon \right\| \\ &\quad \leq C\left\|\widehat{S}^{-1/2}\right\|\left\|U^T\widehat{U} - Q\right\|\|A - P\|\|M\varepsilon\| \leq \frac{C(\nu_n + b_n^2)^{3/2}n^2\log^3 n}{\lambda_d^{5/2}}.\end{aligned}$$

Applying Equations (8) and (9) along with the trivial bound $\lambda_d \leq \lambda_1 = \mathcal{O}(n)$ from Equation (6),

$$\left\| Q\widehat{S}^{-1/2}(UU^T\widehat{U} - UQ)^T(I - UU^T)(A - P)M\varepsilon \right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{55}$$

Applying Equations (54) and (55) to bound the respective quantities on lines (52) and (53), we conclude that

$$\left\| Q\widehat{S}^{-1/2}R_3^T(I - UU^T)(A - P)M\varepsilon \right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right). \tag{56}$$

Applying Equations (46), (47), (48), (49), (51) and (56) to control the terms of Equation (44), we conclude that

$$\left\| (\widehat{X}Q^T - X)^T M\varepsilon \right\| = o_p\left(\frac{\lambda_d}{\sqrt{n}}\right),$$

completing the proof. ■

Lemma 46 *Under Assumptions 2, 3, and 5*

$$\left\| \left(X^T M X\right)^{-1} \right\| = \mathcal{O}_p(\lambda_d^{-1}) \quad \textit{and} \quad \left\| \left(\widehat{X}^T M \widehat{X}\right)^{-1} \right\| = \mathcal{O}_p(\lambda_d^{-1}).$$

Proof Recall that $M = I - W\left(W^T W\right)^{-1} W^T$. Consider the full singular value decomposition

$$W = \begin{bmatrix} U_\parallel & U_\perp \end{bmatrix} \begin{bmatrix} S_\parallel & 0 \\ 0 & S_\perp \end{bmatrix} \begin{bmatrix} V_\parallel^T \\ V_\perp^T \end{bmatrix},$$

where $U_\parallel, V_\parallel \in \mathbb{O}_p$ and $U_\perp, V_\perp \in \mathbb{O}_{n-p}$. It can be shown that $M = U_\perp U_\perp^T$.

Recall that, for conformable real-valued matrices A and B, $(AB)^\dagger = B^\dagger A^\dagger$ when $B = A^T$, or when either A or B is an orthogonal matrix (Greville, 1966). Using this

fact repeatedly, together with submultiplicativity and orthogonal invariance of the spectral norm, and Lemma 33, we obtain:

$$\begin{aligned}\left\|\left(X^T M X\right)^{-1}\right\| &= \left\|\left(X^T U_\perp U_\perp^T X\right)^{-1}\right\| = \left\|\left(X^T U_\perp U_\perp^T X\right)^\dagger\right\| = \left\|\left(U_\perp^T X\right)^\dagger \left(X^T U_\perp\right)^\dagger\right\| \\ &\leq \left\|\left(U_\perp^T X\right)^\dagger\right\| \left\|\left(X^T U_\perp\right)^\dagger\right\| = \left\|X^\dagger \left(U_\perp^T\right)^\dagger\right\| \left\|U_\perp^\dagger \left(X^T\right)^\dagger\right\| \\ &\leq \left\|X^\dagger\right\| \left\|U_\perp^\dagger\right\| \left\|U_\perp^\dagger\right\| \left\|X^\dagger\right\| = \left\|\left(S^{1/2}\right)^\dagger\right\| \left\|\left(S^{1/2}\right)^\dagger\right\| \\ &= C\lambda_d^{-1/2} \cdot \lambda_d^{-1/2} = C\lambda_d^{-1}.\end{aligned}$$

The proof for the $\widehat{X}$ case is analogous, and uses the fact that $\widehat{S}$ concentrates around S, as characterized by Lemma 33. The fact that the inverses exist asymptotically follows from Assumption 4, which takes the regression coefficients β to be identified. ■

Lemma 47 (Concentration of Sub-Gaussian Norms) *Let $W \in \mathbb{R}^{n\times p}$ obey Assumption 5, so that W has independent rows, with the entries of each row being possibly dependent, but each marginally sub-Gaussian with parameter $\sigma > 0$. Then there exists a constant $C > 0$ such that with probability $1 - \mathcal{O}\left(n^{-2}\right)$,*

$$\|W\| \leq C\sqrt{pn\sigma^2}.$$

and, also with probability $1 - \mathcal{O}\left(n^{-2}\right)$, it holds for all $j \in [p]$ that

$$\|W_{\cdot j}\|^2 \leq Cn\sigma^2.$$

Proof To prove the spectral norm bound on W, we adapt the argument given in Theorem 4.6.1 in Vershynin (2020), for which we must first establish the Orlicz norm of each row $W_{i\cdot}$ (Vershynin, 2020, Definition 3.4.1). We begin by noting that for any unit vector $u \in \mathbb{R}^p$, any integer $q \geq 1$ and $i \in [n]$,

$$\mathbb{E}\left[\left(u^T W_{i\cdot}\right)^{2q}\right] \leq \mathbb{E}\left[p^{2q-1}\sum_{j=1}^{p}\left(u_j W_{ij}\right)^{2q}\right] = p^{2q-1}\sum_{j=1}^{p} u_j^{2q}\mathbb{E}\left[W_{ij}^{2q}\right],$$

where the inequality follows from the convexity of $x \mapsto x^{2q}$. Since W_{ij} is sub-Gaussian with parameter σ, using basic properties of sub-Gaussian random variables (see, e.g., Boucheron et al., 2013, Theorem 2.1),

$$\mathbb{E}\left[W_{ij}^{2q}\right] \leq q!(4\sigma)^{2q},$$

and it follows that, trivially upper bounding $q! \leq q^q$ and using $\|u\| = 1$,

$$\mathbb{E}\left[\left(u^T W_{i\cdot}\right)^{2q}\right] \leq q!(4p\sigma)^{2q}\frac{1}{p}\sum_{j=1}^{p} u_j^{2q} \leq q^q(4p\sigma)^{2q}.$$

Thus, we find that for any unit $u \in \mathbb{R}^p$, the random variable $u^T W_{i\cdot}$ satisfies

$$\left(\mathbb{E}\left[\left(u^T W_{i\cdot}\right)^{2q}\right]\right)^{1/2q} \leq \sqrt{2q}\left(2\sqrt{2}p\sigma^2\right). \tag{57}$$

That is, the random variable $u^T W_{i\cdot}$ has Orlicz norm (Vershynin, 2020, Proposition 2.5.2)

$$\|u^T W_{i\cdot}\|_{\Psi_2} \leq Cp\sigma^2.$$

Taking the supremum over all unit $u \in \mathbb{R}^p$, the random vector $W_{i\cdot}$ has Orlicz norm (see, e.g., Vershynin, 2020, Definition 3.4.1)

$$\|W_{i\cdot}\|_{\Psi_2} \leq Cp\sigma^2. \tag{58}$$

Following the argument of Theorem 4.6.1 in Vershynin (2020), let $\mathcal{N}$ be a (1/4)-net for the unit sphere in $\mathbb{R}^p$, which can be constructed with cardinality at most 9^p. It follows that

$$\left\|\frac{1}{n}W^T W\right\| \leq 2\max_{u\in\mathcal{N}}\left|\frac{1}{n}u^T W^T W u\right| = 2\max_{u\in\mathcal{N}}\frac{1}{n}\|Wu\|^2.$$

Fixing $u \in \mathcal{N}$, note that

$$\frac{1}{n}\|Wu\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(u^T W_{i\cdot}\right)^2.$$

Since the random vector $W_{i\cdot}$ has Orlicz norm as given in Equation (58), $u^T W_{i\cdot}$ is subgaussian with parameter $Cp\sigma^2$ and it follows that

$$\frac{1}{n}\left(\|Wu\|^2 - \mathbb{E}\left[\|Wu\|^2\right]\right) = \frac{1}{n}\sum_{i=1}^{n}\left[(u^T W_{i\cdot})^2 - \mathbb{E}\left[(u^T W_{i\cdot})^2\right]\right]$$

is the sample mean of n independent sub-exponential random variables, each with parameter $Cp\sigma^2$ (adjusting C by a suitable constant multiple). Define $\delta = C(\sqrt{p}+t)/\sqrt{n}$ for $t \geq 0$ and set $\epsilon = Cp\sigma^2\max\{\delta, \delta^2\}$. Applying Bernstein's inequality (Vershynin, 2020, Corollary 2.8.3), an argument essentially identical to that in Step 2 of Theorem 4.6.1 in Vershynin (2020), yields that

$$\mathbb{P}\left(\left|\frac{1}{n}\left(\|Wu\|^2 - \mathbb{E}\left[\|Wu\|^2\right]\right)\right| \geq \frac{\epsilon}{2}\right) \leq 2\exp\left\{-C(p+t^2)\right\}.$$

Setting $t = C\log^{1/2} n$ for suitably large $C > 0$ and taking a union bound over all at most 9^p vectors $u \in \mathcal{N}$, it follows that with probability at least $1 - \mathcal{O}\left(n^{-2}\right)$, it holds for all $u \in \mathcal{N}$ that

$$\left|\frac{1}{n}\left(\|Wu\|^2 - \mathbb{E}\left[\|Wu\|^2\right]\right)\right| \leq \frac{C\sigma^2 p(\sqrt{p}+\log^{1/2} n)^2}{\sqrt{n}} \leq \frac{Cp\sigma^2(p+\log n)}{\sqrt{n}}.$$

Thus, it follows that with probability at least $1 - \mathcal{O}\left(n^{-2}\right)$, for all $u \in \mathcal{N}$

$$\|Wu\|^2 \leq \mathbb{E}\left[\|Wu\|^2\right] + Cpn^{1/2}\sigma^2(p+\log n). \tag{59}$$

Expanding Wu and setting $q = 1$ in Equation (57),

$$\mathbb{E}\big[\|Wu\|^2\big] = \sum_{i=1}^{n} \mathbb{E}\big[(u^T W_{i\cdot})^2\big] \leq Cpn\sigma^2.$$

Applying this bound to Equation (59), with probability at least $1 - \mathcal{O}\big(n^{-2}\big)$, it holds for all $u \in \mathcal{N}$ that

$$\|Wu\|^2 \leq Cp\sigma^2(n + n^{1/2}\log n).$$

Thus, with probability at least $1 - \mathcal{O}\big(n^{-2}\big)$,

$$\|W\|^2 \leq 2\max_{u\in\mathcal{N}} \|Wu\|^2 \leq Cp\sigma^2 n.$$

Taking square roots yields our desired bound on the spectral norm. To prove the column-wise bound on W, observe that for any $j \in [p]$, we use a straight-forward adaptation of the proof of Theorem 3.1.1 in Vershynin (2020). We observe that

$$\frac{1}{n}\|W_{\cdot j}\|^2 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[W_{ij}^2\big] = \frac{1}{n}\sum_{i=1}^{n} \big(W_{ij}^2 - \mathbb{E}\big[W_{ij}^2\big]\big)$$

is the sample mean of independent mean-zero random variables, each of which is sub-exponential with parameter $C\sigma$ for suitably chosen constant $C > 0$. An application of Bernstein's inequality (Boucheron et al., 2013) then yields that with probability $1-\mathcal{O}\big(n^{-2}\big)$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} \big(W_{ij}^2 - \mathbb{E}\big[W_{ij}^2\big]\big)\right| \leq \frac{C\sqrt{\sigma^2 \log n}}{n^{1/2}}.$$

Thus, with probability at least $1 - \mathcal{O}\big(n^{-2}\big)$,

$$\|W_{\cdot j}\|^2 = \mathbb{E}\big[\|W_{\cdot j}\|^2\big] + \frac{C\sqrt{\sigma^2 \log n}}{n^{1/2}} \leq Cn\sigma^2$$

for $C > 0$ chosen suitably large. ■

Lemma 48 *Under Assumptions 2 and 5, with notation as above, it holds with probability at least* $1 - \mathcal{O}\big(n^{-2}\big)$ *that*

$$\big\|U^T(A-P)W\big\|_F \leq C\sqrt{dp(\nu_n + b_n^2)n\log n}.$$

Proof We will show that $\big\|U^T(A-P)W\big\|_F^2 \geq Cdp(\nu_n + b_n^2)n\log n$ with probability no larger than $\mathcal{O}(n^{-2})$, whence taking square roots will yield the result.

For each $k \in [d], \ell \in [p+2]$, define

$$S_{k,\ell} = \big[U^T(A-P)W\big]_{k,\ell} = \sum_{i=1}^{n}\sum_{j=1}^{n} (A-P)_{ij} U_{ik} W_{j\ell}$$

and note that

$$\left\|U^T(A-P)W\right\|_F^2 = \sum_{k=1}^{d}\sum_{\ell=1}^{p+2} S_{k,\ell}^2.$$

By Corollary 2.11 in Boucheron et al. (2013), for any $t > 0$,

$$\mathbb{P}(|S_{k,\ell}| \geq t \,|\, X, W) \leq 2\exp\left\{\frac{-t^2}{2\Big(\nu_n \sum_{i=1}^n \sum_{j=1}^n U_{ik}^2 W_{j\ell}^2 + b_n t\Big)}\right\}.$$

Let G_n denote the event $\left\{\|W_{\cdot\ell}\|^2 \leq C_W\, n\right\}$ for some constant C_W, and G_n^c denote the complement of G_n. By Lemma 47, G_n occurs with probability at least $1 - \mathcal{O}\big(n^{-2}\big)$. Thus

$$\begin{aligned}\mathbb{P}(|S_{k,\ell}| \geq t) &= \mathbb{P}(|S_{k,\ell}| \geq t \,|\, G_n)\cdot\mathbb{P}(G_n) + \mathbb{P}(|S_{k,\ell}| \geq t \,|\, G_n^c)\cdot\mathbb{P}(G_n^c)\\ &\leq \mathbb{P}(|S_{k,\ell}| \geq t \,|\, G_n) + \mathbb{P}(G_n^c)\\ &\leq \mathbb{P}(|S_{k,\ell}| \geq t \,|\, G_n) + \mathcal{O}\big(n^{-2}\big).\end{aligned}$$

Now observe that $\sum_{j=1}^n W_{j\ell}^2$ is the squared ℓ_2 norm of a column of W. By the definition of G_n,

$$\mathbb{P}(|S_{k,\ell}| \geq t \,|\, G_n) \leq 2\exp\left\{\frac{-t^2}{2(C_W n\nu_n + bt)}\right\}.$$

Thus

$$\mathbb{P}(|S_{k,\ell}| \geq t) \leq 2\exp\left\{\frac{-t^2}{2(C_W n\nu_n + bt)}\right\} + \mathcal{O}\big(n^{-2}\big)$$

Taking $t = C(\nu_n + b_n^2)^{1/2}\sqrt{n\log n}$ for $C > 0$ suitably large, it follows that

$$\mathbb{P}\Big(|S_{k,\ell}| \geq C(\nu_n + b_n^2)^{1/2}\sqrt{n\log n}\Big) \leq 2n^{-4} + \mathcal{O}\big(n^{-2}\big).$$

A union bound over all $k \in [d], \ell \in [p+2]$ implies that

$$\mathbb{P}\Big(\exists k \in [d], \ell \in [p] : |S_{k,\ell}| \geq C(\nu_n + b_n^2)^{1/2}\sqrt{n\log n}\Big) \leq \frac{2d(p+2)}{n^4} + \mathcal{O}\left(\frac{d(p+2)}{n^2}\right) = \mathcal{O}\big(n^{-2}\big),$$

as d and p are fixed as a function of n. It follows that

$$\mathbb{P}\Big(\left\|U^T(A-P)W\right\|_F^2 \geq Cdp(\nu_n + b_n^2)n\log^2 n\Big) \leq 2n^{-2},$$

completing the proof. ■

Lemma 49 *Suppose that Assumptions 2, 5 and 6 hold. Then*

$$\left\|(\widehat{X}Q^T - X)^T W\right\| = o_p\big(\sqrt{n}\big).$$

Proof Applying Lemma 36,

$$\begin{aligned}
&(\widehat{X}Q^T - X)^T W \\
&\quad = Q\Big(\widehat{U}\widehat{S}^{1/2} - US^{1/2}Q\Big)^T W \\
&\quad = QS^{-1/2}U^T(A-P)W + Q(Q\widehat{S}^{-1/2} - S^{-1/2}Q)^T U^T(A-P)W \\
&\qquad + Q\widehat{S}^{-1/2}Q^T U^T(A-P)UU^T W + Q\widehat{S}^{1/2}R_1^T W + QR_2^T U^T W \\
&\qquad + Q\widehat{S}^{-1/2}R_3^T(I-UU^T)(A-P)W.
\end{aligned} \tag{60}$$

We observe that since $\lambda_d/\sqrt{n} \leq \lambda_1\sqrt{n}$ trivially, our growth assumption in Equation (6) implies that it is sufficient to show that the above terms are all bounded as $o(\lambda_d/\sqrt{n})$. This is precisely the bound obtained in Lemma 45, and the proof of this lemma proceeds identically, but using Lemma 48 instead of Lemma 44 and using the bound in Lemma 47 instead of directly bounding $\|S^{1/2}\|$. The remaining details of the proof are omitted. ■

E.4 Proof of Lemma 26

The following fact will be useful in the subsequent proofs.

Proposition 50 *For $u, v \in \mathbb{R}^k$, $\|uu^T - vv^T\| \leq 2\|u-v\|\|v\| + \|u-v\|^2$.*

Proof Follows from adding and subtracting appropriate quantities and repeatedly applying the triangle inequality. ■

Proposition 51 *Under Assumption 5,*

$$\max_{i\in[n]} \|W_{i\cdot}\| = \mathcal{O}_p\Big(\sqrt{\log n}\Big).$$

Proof Since each W_{ij} is sub-Gaussian with variance parameter σ^2, by Theorem 2.1 of Boucheron et al. (2013),

$$\mathbb{P}(|W_{ij}| \geq t) \leq 2\exp\big(-t^2/2\sigma^2\big).$$

By a union bound,

$$\mathbb{P}\bigg(\max_{ij} |W_{ij}| \geq t\bigg) \leq 2n(p+2)\exp\big(-t^2/2\sigma^2\big).$$

Taking $t = \sqrt{C\sigma^2 \log n}$ and $C > 0$ sufficiently large,

$$\mathbb{P}\bigg(\max_{i\in[n], j\in[p+2]} |W_{ij}| \geq \sqrt{C\sigma^2\log n}\bigg) \leq 2n(p+2)\exp\big(-C\sigma^2\log n/2\sigma^2\big) = \frac{(4p+4)}{n^2}.$$

Observing that $\max_{i\in[n]} \|W_{i\cdot}\| \leq \sqrt{p+2}\,\max_{ij} W_{ij}$, we obtain the desired result. ■

Proposition 52 *Suppose Assumptions 2, 4, 5 and 6 hold. Letting $Q \in \mathbb{O}_d$ be as in Lemma 31, define*

$$\widetilde{Q} = \begin{bmatrix} I_{p+2} & 0 \\ 0 & Q \end{bmatrix}.$$

Then $\widehat{\Sigma}_\beta \to \widetilde{Q}\widetilde{\Sigma}_\beta\widetilde{Q}^T$ in probability, where $\widehat{\Sigma}_\beta$ and $\widetilde{\Sigma}_\beta$ are as defined in Definition 8.

Proof By definition,

$$\left\|\widehat{\Sigma}_\beta - \widetilde{Q}\widetilde{\Sigma}_\beta\widetilde{Q}^T\right\| = \left\|\widehat{A}_\beta^{-1} \cdot \widehat{B}_\beta \cdot \left(\widehat{A}_\beta^{-1}\right)^T - \widetilde{Q}^T\widetilde{A}_\beta^{-1}\widetilde{Q} \cdot \widetilde{Q}^T\widetilde{B}_\beta\widetilde{Q} \cdot \widetilde{Q}^T\left(\widetilde{A}_\beta^{-1}\right)^T\widetilde{Q}\right\|.$$

By the continuous mapping theorem, it is sufficient to show $\widehat{A}_\beta \to \widetilde{Q}^T\widetilde{A}_\beta\widetilde{Q}$ and $\widehat{B}_\beta \to \widetilde{Q}^T\widetilde{B}_\beta\widetilde{Q}$, with both convergences holding in probability.

We begin by observing that

$$\widehat{A}_\beta - \widetilde{Q}^T\widetilde{A}_\beta\widetilde{Q} = \frac{1}{n}\begin{bmatrix} 0 & W^T(\widehat{X} - XQ) \\ (\widehat{X} - XQ)^T W & \widehat{X}^T\widehat{X} - Q^T X^T X Q \end{bmatrix},$$

and so, applying Lemma 49 to bound $\|(\widehat{X} - XQ)^T W\| = o_p(\sqrt{n})$,

$$\begin{aligned} \left\|\widehat{A}_\beta - \widetilde{Q}^T\widetilde{A}_\beta\widetilde{Q}\right\| &\le \frac{2}{n}\left\|(\widehat{X} - XQ)^T W\right\| + \frac{1}{n}\left\|\widehat{X}^T\widehat{X} - Q^T X^T X Q\right\| \\ &= \frac{1}{n}\left\|\widehat{X}^T\widehat{X} - Q^T X^T X Q\right\| + o_p\left(n^{-1/2}\right). \end{aligned} \tag{61}$$

Applying Lemma 42 with $M = I$,

$$\frac{1}{n}\left\|\widehat{X}^T\widehat{X} - Q^T X^T X Q\right\| = o_p\left(\frac{\lambda_d^2}{n^{3/2}\lambda_1^{1/2}}\right).$$

Using the fact that $\lambda_d \le \lambda_1$ by definition and applying Equation (6),

$$\frac{1}{n}\left\|\widehat{X}^T\widehat{X} - Q^T X^T X Q\right\| = o_p(1).$$

Applying this to Equation (61),

$$\left\|\widehat{A}_\beta - \widetilde{Q}^T\widetilde{A}_\beta\widetilde{Q}\right\| = o_p(1) + o_p\left(n^{-1/2}\right) = o_p(1).$$

The continuous mapping theorem then implies that

$$\widehat{A}_\beta^{-1} \to \widetilde{Q}^T\widetilde{A}_\beta^{-1}\widetilde{Q} \quad \text{in probability.}$$

It remains to show that $\widehat{B}_\beta$ converges to $\widetilde{Q}^T\widetilde{B}_\beta\widetilde{Q}$ in probability. Toward this end, recall the definitions $\widehat{\varepsilon}_i = Y_i - \widehat{D}_{i\cdot}\widehat{\beta}$ and $\widetilde{\varepsilon}_i = Y_i - D_{i\cdot}\widetilde{\beta}$. Adding and subtracting appropriate

quantities,

$$\begin{aligned}\left\|\widehat{B}_\beta - \widetilde{Q}^T\widetilde{B}_\beta\widetilde{Q}\right\| &= \left\|\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot} - \widetilde{\varepsilon}_i^2\widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| \\ &= \left\|\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2\left[\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot} - \widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right] + \left[\widehat{\varepsilon}_i^2 - \widetilde{\varepsilon}_i^2\right]\widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| \\ &\leq \frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot} - \widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| + \left|\widehat{\varepsilon}_i^2 - \widetilde{\varepsilon}_i^2\right|\left\|D_{i\cdot}^T D_{i\cdot}\right\|.\end{aligned} \tag{62}$$

Adding and subtracting appropriate quantities and applying the triangle inequality,

$$\begin{aligned}\max_{i\in[n]}|\widehat{\varepsilon}_i - \widetilde{\varepsilon}_i| &= \max_{i\in[n]}\left|\widehat{D}_{i\cdot}\widetilde{Q}^T\widetilde{Q}\widehat{\beta} - D_{i\cdot}\widetilde{\beta}\right| \\ &\leq \max_{i\in[n]}\left\|\widehat{D}_{i\cdot}\widetilde{Q}^T\right\|\left\|\widetilde{Q}\widehat{\beta} - \widetilde{\beta}\right\| + \left\|\widehat{D}_{i\cdot}\widetilde{Q}^T - D_{i\cdot}\right\|\left\|\widetilde{\beta}\right\|.\end{aligned}$$

By definition, $\left\|\widehat{D}_{i\cdot} - D_{i\cdot}\widetilde{Q}\right\| = \left\|\widehat{X}_{i\cdot} - X_{i\cdot}Q\right\|$. Thus, applying Lemma 32 and using Lemma 33 to ensure that $\left\|\widehat{D}_{i\cdot}\right\| = \mathcal{O}_p\left(\lambda_1^{1/2}\right)$,

$$\max_{i\in[n]}|\widehat{\varepsilon}_i - \widetilde{\varepsilon}_i| \leq \left\|\widetilde{Q}\widehat{\beta} - \widetilde{\beta}\right\|\mathcal{O}_p\left(\lambda_1^{1/2}\right) + \eta_n\left\|\widetilde{\beta}\right\|.$$

Applying Assumption 6 and Lemma 25,

$$\max_{i\in[n]}|\widehat{\varepsilon}_i - \widetilde{\varepsilon}_i| \leq \eta_n + o_p(1).$$

Applying the bound on η_n from Lemma 32,

$$\max_{i\in[n]}|\widehat{\varepsilon}_i - \widetilde{\varepsilon}_i| = o_p(1).$$

This in turn yields, by Proposition 50,

$$\left|\widehat{\varepsilon}_i^2 - \widetilde{\varepsilon}_i^2\right| \leq |\widehat{\varepsilon}_i - \widetilde{\varepsilon}_i||\widetilde{\varepsilon}_i| + |\widehat{\varepsilon}_i - \widetilde{\varepsilon}_i|^2 = (1 + |\widetilde{\varepsilon}_i|)o_p(1). \tag{63}$$

By Proposition 50 and Lemma 31,

$$\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot} - \widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| \leq \left\|\widehat{D}_{i\cdot} - D_{i\cdot}\widetilde{Q}\right\|\left\|D_{i\cdot}\widetilde{Q}\right\| + \left\|\widehat{D}_{i\cdot} - D_{i\cdot}\widetilde{Q}\right\|^2 = \eta_n\|D_{i\cdot}\| + \eta_n^2. \tag{64}$$

Adding and subtracting appropriate quantities,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot} - \widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| \leq \frac{1}{n}\sum_{i=1}^n \widetilde{\varepsilon}_i^2\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot} - \widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| \\ + \frac{1}{n}\sum_{i=1}^n\left[\widehat{\varepsilon}_i^2 - \widetilde{\varepsilon}_i^2\right]\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot} - \widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\|\end{aligned} \tag{65}$$

By Equation (64),

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{\varepsilon}_i^2\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot}-\widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| \leq \frac{\eta_n}{n}\sum_{i=1}^{n}\widetilde{\varepsilon}_i^2\|D_{i\cdot}\|+\frac{\eta_n^2}{n}\sum_{i=1}^{n}\widetilde{\varepsilon}_i^2$$

By the regularity conditions in Assumptions 4 and 5, both averages on the right-hand side converge to constants. Thus, since $\eta_n = o(1)$ by Lemma 32,

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{\varepsilon}_i^2\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot}-\widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| \leq C\,\eta_n = o_p(1). \tag{66}$$

Appealing to Equations (63) and (64),

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\varepsilon}_i^2-\widetilde{\varepsilon}_i^2\right]\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot}-\widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| &= \frac{o_p(1)}{n}\sum_{i=1}^{n}(1+|\widetilde{\varepsilon}_i|)\eta_n\left(\|D_{i\cdot}\|+\eta_n\right)\\ &= \frac{o_p(\eta_n)}{n}\sum_{i=1}^{n}(1+|\widetilde{\varepsilon}_i|)\left(\|D_{i\cdot}\|+\eta_n\right).\end{aligned}$$

By Assumptions 4 and 5, the mean on the right-hand side converges to a constant, and again recalling that $\eta_n = o(1)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\varepsilon}_i^2-\widetilde{\varepsilon}_i^2\right]\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot}-\widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| = o_p(1).$$

Applying this and Equation (66) to Equation (65),

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{\varepsilon}_i^2\left\|\widehat{D}_{i\cdot}^T\widehat{D}_{i\cdot}-\widetilde{Q}^T D_{i\cdot}^T D_{i\cdot}\widetilde{Q}\right\| = o_p(1). \tag{67}$$

By Equation (63) and Proposition 50,

$$\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{\varepsilon}_i^2-\widetilde{\varepsilon}_i^2\right|\left\|D_{i\cdot}^T D_{i\cdot}\right\| \leq \frac{o_p(1)}{n}\sum_{i=1}^{n}(1+|\widetilde{\varepsilon}_i|)\|D_{i\cdot}\|^2.$$

Assumptions 4 and 5 ensure that the mean on the right-hand side converges to a constant, and we have

$$\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{\varepsilon}_i^2-\widetilde{\varepsilon}_i^2\right|\left\|D_{i\cdot}^T D_{i\cdot}\right\| = o_p(1).$$

Applying this and Equation (67) to Equation (62), we conclude that

$$\left\|\widehat{B}_\beta-\widetilde{Q}^T\widetilde{B}_\beta\widetilde{Q}\right\| = o_p(1),$$

completing the proof. ■

Remark 53 *Proposition 54 below states that the robust covariance estimator for $\widehat{\Theta}$ based on $\widehat{X}$ converges to the robust covariance estimator based on X, but subject to some orthogonal non-identifiability. The orthogonal non-identifiability has a somewhat nasty form because we have vectorized $\widehat{\Theta}$ in for the stake of M-estimation. To understand the result more intuitively, suppose that $d = 1$. Then Theorem 9 states*

$$\sqrt{n}\,\widehat{\Sigma}_{\mathrm{vec}(\Theta)}^{-1/2}\left(\widehat{\Theta}\,Q^T - \Theta\right) \to \mathcal{N}(0, I_p),$$

such that $\widehat{\Theta} \in \mathbb{R}^{d\times 1}$. Proposition 54 then states that $\widehat{\Sigma}_{\mathrm{vec}(\Theta)} \to \widetilde{Q}\widetilde{\Sigma}_{\mathrm{vec}(\Theta)}\widetilde{Q}^T$, analogously to Proposition 52.

Proposition 54 *Suppose Assumptions 2, 4, 5 and 6 hold. Let*

$$\widetilde{\Sigma}_{\mathrm{vec}(\Theta Q)} = \widetilde{A}_{\mathrm{vec}(\Theta)}^{-1} \cdot \widetilde{B}_{\mathrm{vec}(\Theta Q)} \cdot \left(\widetilde{A}_{\mathrm{vec}(\Theta)}^{-1}\right)^T,$$

where $\widetilde{A}_{\mathrm{vec}(\Theta)}$ is as defined in Definition 8, and

$$\widetilde{B}_{\mathrm{vec}(\Theta Q)} = \frac{1}{n}\sum_{i=1}^{n} Q^T\widetilde{\xi}_{i\cdot}^T\,\widetilde{\xi}_{i\cdot}Q \otimes W_{i\cdot}^T W_{i\cdot}.$$

Under Assumptions 2, 4, and 5, and Definition 8, $\widehat{\Sigma}_{\mathrm{vec}(\Theta)} \to \widetilde{\Sigma}_{\mathrm{vec}(\Theta Q)}$ in probability.

Proof By definition, $\widehat{A}_{\mathrm{vec}(\Theta)} = \widetilde{A}_{\mathrm{vec}(\Theta)}$. Thus, by the continuous mapping theorem, it will suffice to show that $\widehat{B}_{\mathrm{vec}(\Theta)} \to \widetilde{B}_{\mathrm{vec}(\Theta)}$.

Let $\widetilde{\xi}_{i\cdot} = X_{i\cdot} - W_{i\cdot}\widetilde{\Theta}$. Then, by the triangle inequality and properties of the Kronecker product,

$$\begin{aligned}
\left\|\widehat{B}_{\mathrm{vec}(\Theta)} - \widetilde{B}_{\mathrm{vec}(\Theta)}\right\| &= \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\xi}_{i\cdot}^T\widehat{\xi}_{i\cdot} - Q^T\widetilde{\xi}_{i\cdot}^T\widetilde{\xi}_{i\cdot}Q\right)\otimes W_{i\cdot}^T W_{i\cdot}\right\| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left\|\left(\widehat{\xi}_{i\cdot}^T\widehat{\xi}_{i\cdot} - Q^T\widetilde{\xi}_{i\cdot}^T\widetilde{\xi}_{i\cdot}Q\right)\otimes W_{i\cdot}^T W_{i\cdot}\right\| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\xi}_{i\cdot}^T\widehat{\xi}_{i\cdot} - Q^T\widetilde{\xi}_{i\cdot}^T\widetilde{\xi}_{i\cdot}Q\right\|\left\|W_{i\cdot}^T W_{i\cdot}\right\|.
\end{aligned}$$

Applying Cauchy-Schwarz, we obtain

$$\left\|\widehat{B}_{\mathrm{vec}(\Theta)} - \widetilde{B}_{\mathrm{vec}(\Theta)}\right\| \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\xi}_{i\cdot}^T\widehat{\xi}_{i\cdot} - Q^T\widetilde{\xi}_{i\cdot}^T\widetilde{\xi}_{i\cdot}Q\right\|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|W_{i\cdot}^T W_{i\cdot}\right\|^2}. \tag{68}$$

We will show that this product is $o_p(1)$, which will complete the proof. Using basic properties of the norm,

$$\mathbb{E}\left[\left\|W_{i\cdot}^T W_{i\cdot}\right\|^2\right] \le \mathbb{E}\left[\left\|W_{i\cdot}\right\|^4\right] = \mathbb{E}\left[\left(\sum_{j=1}^{p+2} W_{ij}^2\right)^2\right] \le C\sum_{j=1}^{p+2}\mathbb{E}\left[W_{ij}^4\right]$$

Our sub-Gaussian assumption on the entries of $W_{i\cdot}$ imply that $\mathbb{E}\Big[W_{ij}^4\Big] \leq C\sigma^4$ (Boucheron et al., 2013, Theorem 2.1), whence

$$\mathbb{E}\Big[\big\|W_{i\cdot}^T W_{i\cdot}\big\|^2\Big] \leq C(p+2)\sigma^4 = \mathcal{O}(1).$$

It follows that, by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n \big\|W_{i\cdot}^T W_{i\cdot}\big\|^2 = \mathcal{O}_p(1).$$

Applying this fact to Equation (68), our proof will be complete if we can show that

$$\frac{1}{n}\sum_{i=1}^n \Big\|\widehat{\xi}_{i\cdot}^T\widehat{\xi}_{i\cdot} - Q^T\widetilde{\xi}_{i\cdot}^T\widetilde{\xi}_{i\cdot}Q\Big\|^2 = o_p(1). \tag{69}$$

Recalling the definition of $\widehat{\xi}_{ij}$ from Definition 8, we observe that

$$\widehat{\xi}_{i\cdot} - \widetilde{\xi}_{i\cdot}Q = \widehat{X}_{i\cdot} - W_{i\cdot}\widehat{\Theta} - X_{i\cdot}Q + W_{i\cdot}\widetilde{\Theta}Q = \widehat{X}_{i\cdot} - X_{i\cdot}Q + W_{i\cdot}\Big(\widetilde{\Theta}Q - \widehat{\Theta}\Big).$$

Using Lemma 31 to bound $\max_i\Big\|\widehat{X}_{i\cdot} - X_{i\cdot}Q\Big\|$, Lemma 25 to bound $\Big\|\widetilde{\Theta}Q - \widehat{\Theta}\Big\|$ and Proposition 51 to bound $\|W_{i\cdot}\|$, it follows that

$$\max_{i\in[n]}\Big\|\widehat{\xi}_{i\cdot} - \widetilde{\xi}_{i\cdot}Q\Big\| \leq \eta_n + \max_{i\in[n]}\|W_{i\cdot}\|\, o_p\Big(n^{-1/2}\Big) = \eta_n + o_p\Big(n^{-1/2}\log^{1/2} n\Big).$$

Applying our growth assumptions to ensure that $\eta_n = o(1)$,

$$\max_{i\in[n]}\Big\|\widehat{\xi}_{i\cdot} - \widetilde{\xi}_{i\cdot}Q\Big\| = o_p(1). \tag{70}$$

Adding and subtracting appropriate quantities,

$$\widetilde{\xi}_{i\cdot} = \widetilde{\xi}_{i\cdot} - \xi_{i\cdot} + \xi_{i\cdot} = X_{i\cdot} - W_{i\cdot}\widetilde{\Theta} - X_{i\cdot} + W_{i\cdot}\Theta + \xi_{i\cdot} = W_{i\cdot}\Big(\Theta - \widetilde{\Theta}\Big) + \xi_{i\cdot},$$

so that $\Big\|\widetilde{\xi}_{i\cdot}\Big\| = \|\xi_{i\cdot}\| + o_p\Big(n^{-1/2}\log^{1/2} n\Big)$, and it follows that

$$\begin{aligned}
\Big\|\widehat{\xi}_{i\cdot}^T\widehat{\xi}_{i\cdot} - Q^T\widetilde{\xi}_{i\cdot}^T\widetilde{\xi}_{i\cdot}Q\Big\|^2 &= \Big\|\Big(\widehat{\xi}_{i\cdot}^T - Q^T\widetilde{\xi}_{i\cdot}^T\Big)\Big(\widehat{\xi}_{i\cdot} - \widetilde{\xi}_{i\cdot}Q\Big) + \Big(\widehat{\xi}_{i\cdot}^T - Q^T\widetilde{\xi}_{i\cdot}^T\Big)\widetilde{\xi}_{i\cdot} + \widetilde{\xi}_{i\cdot}^T\Big(\widehat{\xi}_{i\cdot} - \widetilde{\xi}_{i\cdot}Q\Big)\Big\|^2 \\
&\leq \Big\|\widehat{\xi}_{i\cdot} - \widetilde{\xi}_{i\cdot}Q\Big\|^4 + 2\,\|\xi_{i\cdot}\|\Big\|\widehat{\xi}_{i\cdot} - \widetilde{\xi}_{i\cdot}Q\Big\|^3 + \Big\|\widetilde{\xi}_{i\cdot}\Big\|^2\Big\|\widehat{\xi}_{i\cdot} - \widetilde{\xi}_{i\cdot}Q\Big\|^2 \\
&= o_p(1)\cdot\Big[1 + \|\xi_{i\cdot}\| + \|\xi_{i\cdot}\|^2\Big],
\end{aligned}$$

where we have made several applications of Equation (70). Thus,

$$\frac{1}{n}\sum_{i=1}^n \Big\|\widehat{\xi}_{i\cdot}^T\widehat{\xi}_{i\cdot} - Q^T\widetilde{\xi}_{i\cdot}^T\widetilde{\xi}_{i\cdot}Q\Big\|^2 \leq o_p(1)\cdot\frac{1}{n}\sum_{i=1}^n\Big(1 + \|\xi_{i\cdot}\| + \|\xi_{i\cdot}\|^2\Big).$$

By the law of large numbers and Assumption 5, the mean on the right-hand side converges in probability to a finite constant. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\xi}_{i\cdot}^{T}\widehat{\xi}_{i\cdot} - Q^{T}\widetilde{\xi}_{i\cdot}^{T}\widetilde{\xi}_{i\cdot}Q\right\|^{2} = o_p(1),$$

which establishes Equation (69) and completes the proof. ■

Appendix F. Proofs for Causal Estimators

Here we collect the proofs of Theorems 17 and 18.

F.1 Proof of Theorem 17

Proof Follows immediately from Theorem 9, Proposition 3 and the delta method. ■

F.2 Proof of Theorem 18

Before giving a formal proof of Theorem 18, we give a high-level sketch of the argument. As above, let $\widetilde{\Psi}_{\text{nie}}$ denote the analogous estimator to $\widehat{\Psi}_{\text{nie}}$, but based on X rather than $\widehat{X}$.

First, we show that $\widetilde{\Psi}_{\text{nie}}$ converges to a normal distribution centered on Ψ_{nie} by a stacked M-estimation argument. This requires an extension of Theorem 22 to show that $(\widetilde{\beta}, \widetilde{\Theta})$ are jointly asymptotically normal rather than marginally asymptotically normal. Then, using the delta method, we show

$$\sqrt{n}\Big(\widetilde{\Psi}_{\text{nie}} - \Psi_{\text{nie}}\Big) \to \mathcal{N}\big(0, \sigma^2_{\text{nie}}\big).$$

Next we show that we can replace the true latent positions X with the estimates $\widehat{X}$ without changing the asymptotic distribution of our estimates. By Slutsky's theorem it is sufficient show that

$$\left|\widehat{\Psi}_{\text{nie}} - \widetilde{\Psi}_{\text{nie}}\right| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

Finally, we argue that $\widehat{\sigma}^2_{\text{nie}}$ is a consistent estimator for $\widetilde{\sigma}^2_{\text{nie}}$, which is itself a consistent estimator for σ^2_{nie} under some mild conditions (Boos and Stefanski, 2013, Theorem 7.3, Theorem 7.4).

Proof Define $\boldsymbol{\theta} = \big(\Theta_{\cdot 1}^{T}, ..., \Theta_{\cdot d}^{T}, \beta^{T}\big)^{T}$. Observe that $\Big(\widehat{\Theta}_{\cdot 1}^{T}, ..., \widehat{\Theta}_{\cdot d}^{T}, \widehat{\beta}^{T}\Big)^{T}$ is an M-estimator with estimating function

$$\psi((Y_i, W_{i\cdot}, D_{i\cdot}), \boldsymbol{\theta}^*) = \begin{bmatrix} (X_{i1} - W_{i\cdot}\Theta_{\cdot 1})W_{i\cdot} \\ \vdots \\ (X_{id} - W_{i\cdot}\Theta_{\cdot d})W_{i\cdot} \\ (Y_i - D_{i\cdot}\beta)D_{i\cdot} \end{bmatrix}.$$

Recall the prior definitions of D and W as the design matrices for the mediator and outcome regressions. Then, by straightforward calculation, or Boos and Stefanski (2013, Chapter 7), and recalling the definitions of $A_{\text{vec}(\Theta)} \in \mathbb{R}^{pd \times pd}$ and $A_\beta \in \mathbb{R}^{(d+p)\times(d+p)}$ from Assumption 4, we have, under sufficient regularity conditions for limits and expectations to be interchangeable,

$$\begin{aligned}
A_{\boldsymbol{\theta}} &= \lim_{n\to\infty} \mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n} \psi'((Y_i, W_{i\cdot}, D_{i\cdot}), \boldsymbol{\theta})\right] \\
&= \lim_{n\to\infty} \mathbb{E}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} W_{i\cdot}W_{i\cdot}^T & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 \\ \vdots & 0 & \frac{1}{n}\sum_{i=1}^{n} W_{i\cdot}W_{i\cdot}^T & 0 \\ 0 & 0 & 0 & \frac{1}{n}\sum_{i=1}^{n} D_{i\cdot}D_{i\cdot}^T \end{bmatrix} \\
&= \begin{bmatrix} A_{\text{vec}(\Theta)} & 0 \\ 0 & A_\beta \end{bmatrix}.
\end{aligned}$$

Further, again recalling the definitions of $B_{\text{vec}(\Theta)} \in \mathbb{R}^{pd\times pd}$ and $B_\beta \in \mathbb{R}^{(d+p)\times(d+p)}$ from Assumption 4, we have

$$\begin{aligned}
B_{\boldsymbol{\theta}} &= \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \psi((Y_i, W_{i\cdot}, D_{i\cdot}), \boldsymbol{\theta})\, \psi((Y_i, W_{i\cdot}, D_{i\cdot}), \boldsymbol{\theta})^T\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \xi_{i1}W_{i\cdot}W_{i\cdot}^T\xi_{i1}^T & \dots & \xi_{i1}W_{i\cdot}W_{i\cdot}^T\xi_{id}^T & 0 \\ \vdots & \ddots & \vdots & 0 \\ \xi_{id}W_{i\cdot}W_{i\cdot}^T\xi_{i1}^T & \dots & \xi_{id}W_{i\cdot}W_{i\cdot}^T\xi_{id}^T & 0 \\ 0 & 0 & 0 & \varepsilon_i^2 D_{i\cdot}D_{i\cdot}^T \end{bmatrix}\right] \\
&= \begin{bmatrix} B_{\text{vec}(\Theta)} & 0 \\ 0 & B_\beta \end{bmatrix}.
\end{aligned}$$

While the diagonal blocks of $B_{\boldsymbol{\theta}}$ correspond to $B_{\text{vec}(\Theta)}$, and B_β, we show the derivation of the off-diagonal blocks. By the tower law, definition of D and ξ, and then Assumption 4, we have

$$\begin{aligned}
B_{14} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\mathbb{E}\big[\xi_{i1}W_{i\cdot}D_{i\cdot}^T\varepsilon_i^T \,\big|\, D_{i\cdot}\big]\big] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[(X_{i1} - W_{i\cdot}\Theta_{\cdot 1})W_{i\cdot}D_{i\cdot}^T\mathbb{E}\big[\varepsilon_i^T \,\big|\, D_{i\cdot}\big]\big] \\
&= 0
\end{aligned}$$

where the derivations to show that $B_{23} = B_{34} = 0$ are analogous to the B_{14} case. Under Assumption 4, Theorem 7.2 of Boos and Stefanski (2013) implies that

$$\begin{bmatrix} \text{vec}(\widehat{\Theta}) \\ \widehat{\beta} \end{bmatrix} \to \mathcal{N}\left(\begin{bmatrix} \text{vec}(\Theta) \\ \beta \end{bmatrix}, \begin{bmatrix} \Sigma_{\text{vec}(\Theta)} & 0 \\ 0 & \Sigma_\beta \end{bmatrix}\right),$$

where $\Sigma_{\mathrm{vec}(\Theta)}$ and Σ_β are the same marginal covariances for $\mathrm{vec}\left(\widehat{\Theta}\right)$ and $\widehat{\beta}$ from Theorem 22. This is an extension of Theorem 22 in that it shows that $\widehat{\Theta}$ and $\widehat{\beta}$ are jointly, rather than marginally, asymptotically normal. Also note that the asymptotic covariances between $\widehat{\Theta}$ and $\widehat{\beta}$ are zero, such that we can concatenate the previous marginal covariance estimators to obtain a joint covariance estimate. Next we show that

$$\left|\widehat{\Psi}_{\mathrm{nie}} - \widetilde{\Psi}_{\mathrm{nie}}\right| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

Recall that

$$\widehat{\Psi}_{\mathrm{nie}} = (t - t^*)\,\widehat{\theta}_{\mathrm{t}}\,\widehat{\beta}_{\mathrm{x}}$$

We apply the submultiplicativity to obtain

$$\frac{\left|\widehat{\Psi}_{\mathrm{nie}} - \widetilde{\Psi}_{\mathrm{nie}}\right|}{(t - t^*)} \leq \left\|\widehat{\theta}_{\mathrm{t}}\widehat{\beta}_{\mathrm{x}} - \widetilde{\theta}_{\mathrm{t}}\widetilde{\beta}_{\mathrm{x}}\right\|$$

We now bound the first term on the right-hand side via

$$\begin{aligned}\left|\widehat{\theta}_{\mathrm{t}}\widehat{\beta}_{\mathrm{x}} - \widetilde{\theta}_{\mathrm{t}}\widetilde{\beta}_{\mathrm{x}}\right| &\leq \left\|\widehat{\theta}_{\mathrm{t}}\left(\widehat{\beta}_{\mathrm{x}} - \widetilde{\beta}_{\mathrm{x}}\right)\right\| + \left\|\left(\widehat{\theta}_{\mathrm{t}} - \widetilde{\theta}_{\mathrm{t}}\right)\widetilde{\beta}_{\mathrm{x}}\right\| \\ &\leq \left\|\widehat{\theta}_{\mathrm{t}} - \widetilde{\theta}_{\mathrm{t}}\right\|\left\|\widehat{\beta}_{\mathrm{x}} - \widetilde{\beta}_{\mathrm{x}}\right\| + \left\|\widetilde{\theta}_{\mathrm{t}}\right\|\left\|\widehat{\beta}_{\mathrm{x}} - \widetilde{\beta}_{\mathrm{x}}\right\| + \left\|\widehat{\theta}_{\mathrm{t}} - \widetilde{\theta}_{\mathrm{t}}\right\|\left\|\widetilde{\beta}_{\mathrm{x}}\right\|.\end{aligned}$$

By Lemma 23, $\left\|\widetilde{\theta}_{\mathrm{t}}\right\|$ and $\left\|\widetilde{\beta}_{\mathrm{x}}\right\|$ are both $\mathcal{O}_p(1)$. By Theorem 9, $\left\|\widehat{\theta}_{\mathrm{t}} - \widetilde{\theta}_{\mathrm{t}}\right\|$ and $\left\|\widehat{\beta}_{\mathrm{x}} - \widetilde{\beta}_{\mathrm{x}}\right\|$ are both $o_p\left(n^{-1/2}\right)$. Thus, we obtain that the upper display is $o_p\left(n^{-1/2}\right)$, as desired.

Finally, we show that $\widehat{\sigma}^2_{\mathrm{nie}}$ is a consistent estimator for σ^2_{nie}. By Theorem 7.2 of Boos and Stefanski (2013) and the delta method, the asymptotic variance of $\widetilde{\Psi}_{\mathrm{nie}}$ is given by

$$\sigma^2_{\mathrm{nie}} = (t - t^*)^T \begin{bmatrix}\beta_{\mathrm{x}} \\ \theta_{\mathrm{t}}\end{bmatrix}^T \begin{bmatrix}\Sigma_{\theta_{\mathrm{t}}} & 0 \\ 0 & \Sigma_{\beta_{\mathrm{x}}}\end{bmatrix} \begin{bmatrix}\beta_{\mathrm{x}} \\ \theta_{\mathrm{t}}\end{bmatrix} (t - t^*).$$

By Theorem 9, Proposition 52, Proposition 54, and the continuous mapping theorem, $\widehat{\sigma}^2_{\mathrm{nie}}$ converges to $\widetilde{\sigma}^2_{\mathrm{nie}}$ in probability. ■

Appendix G. Additional Simulation Results

In this section we report additional results from our simulation study.

G.1 Convergence Rates for $\widehat{\beta}$ and $\widehat{\Theta}$

In Section 4, we showed that $\widehat{\Psi}_{\mathrm{nde}}$ converged to Ψ_{nde} and $\widehat{\Psi}_{\mathrm{nie}}$ converged to Ψ_{nie}. We now show similar convergence results for the regression estimators $\widehat{\beta}$ and $\widehat{\Theta}$. Recall that $\widehat{\beta}_{\mathrm{x}}$ and $\widehat{\Theta}$ only recover β_{x} and Θ up to an unknown orthogonal transformation Q. Luckily, since the true latent positions are known in our simulations, we can align $\widehat{X}$ with the latent X by solving a Procrustes alignment problem (Gower and Dijksterhuis, 2004). As a result,

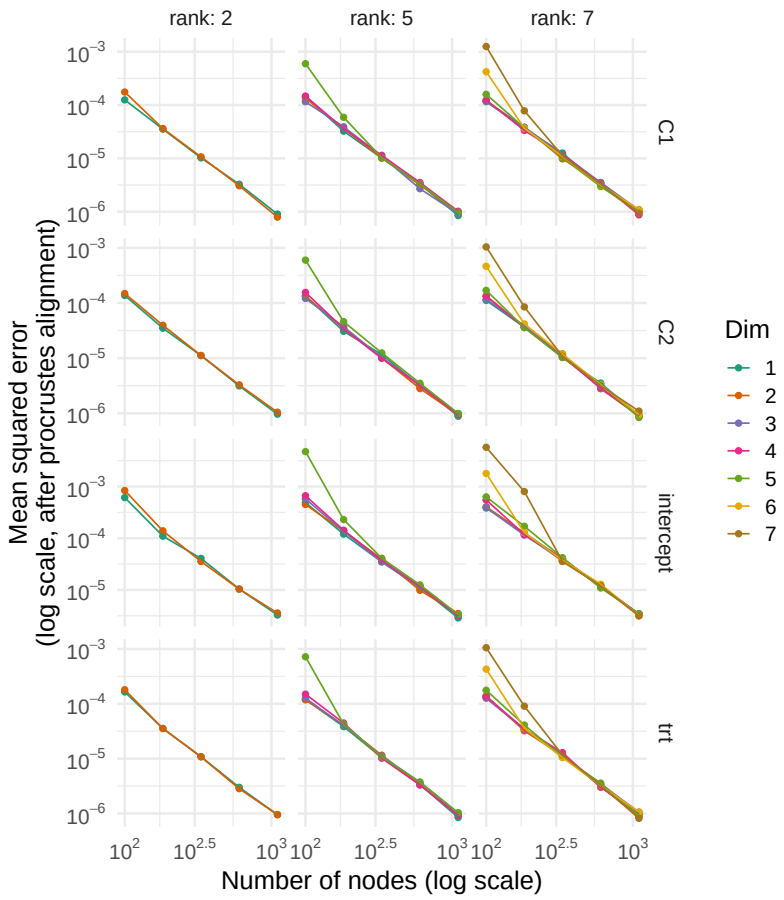


Figure 21: Elementwise ℓ_1 convergence of $\widehat{\Theta}\, Q^T$ to Θ under the uninformative model. Recall that $\widehat{\Theta}$ is a matrix-valued estimator. Each panel shows the ℓ_1 error (vertical axis, log scale) of a portion of $\widehat{\Theta}$ as a function of the number of nodes in the network (horizontal axis, log scale). Within each panel, each line represents the error for a single coefficient corresponding to a particular dimension of the latent space. Panels vary horizontally by number of latent communities (left: two blocks, middle: five block, right: seven blocks) and vertically by column of the design matrix W.

we can investigate element-wise parameter recovery even in the presence of orthogonal non-identifiability. Figures 21 and 22 visualize results for $\widehat{\Theta}$. Figures 23 and 24 visualize results $\widehat{\beta}$. These figures show convergence of each element of $\widehat{\Theta}$ and $\widehat{\beta}$ to the corresponding elements of Θ and β. This convergence occurs at the same $\sqrt{n}$-rate across several simulation settings, as expected under Theorem 9.

G.2 Finite Sample Bias in $\widehat{\beta}$

As explained in Remark 11, it is well-known that ordinary least squares estimates are biased when regression covariates are measured with error. Asymptotically, this is not an issue for $\widehat{\beta}$, as the deviance of $\widehat{X}$ around X tends goes to zero in two-to-infinity norm, such

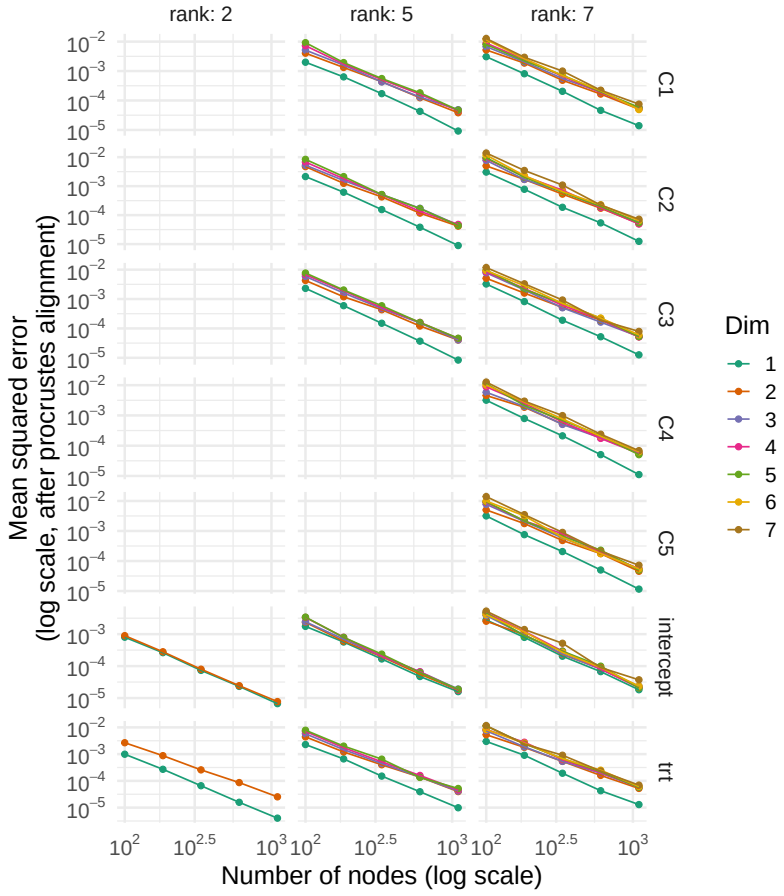


Figure 22: Elementwise ℓ_1 convergence of $\widehat{\Theta}\, Q^T$ to Θ under the informative model. Recall that $\widehat{\Theta}$ is a matrix-valued estimator. Each panel shows the ℓ_1 error (vertical axis, log scale) of a portion of $\widehat{\Theta}$ as a function of the number of nodes in the network (horizontal axis, log scale). Within each panel, each line represents the error for a single coefficient corresponding to a particular dimension of the latent space. Panels vary horizontally by number of latent communities (left: two blocks, middle: five block, right: seven blocks) and vertically by column of the design matrix W.

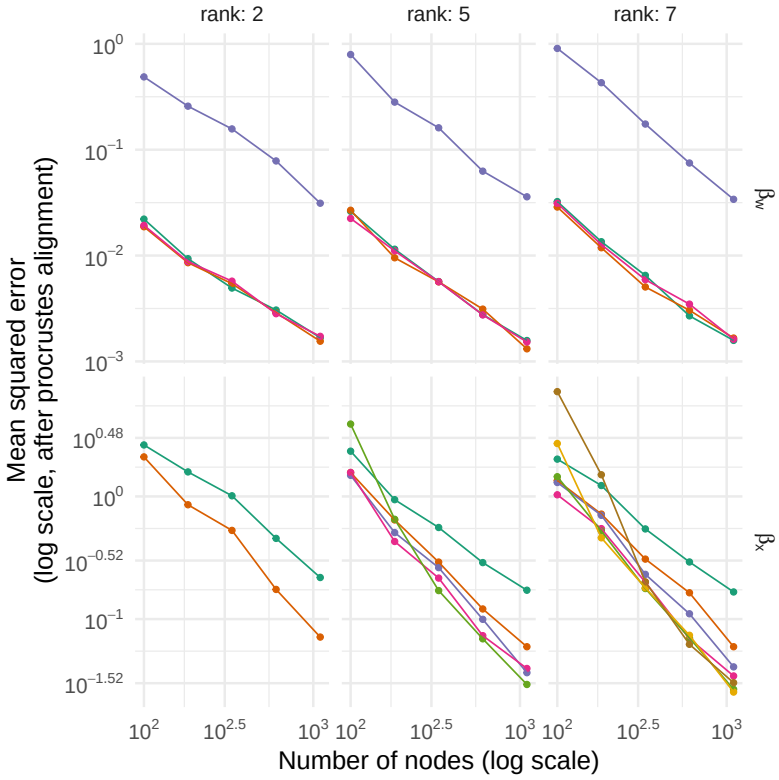


Figure 23: Convergence of $\widehat{\beta}_{\mathrm{w}}$ to β_{w} and $Q\widehat{\beta}_{\mathrm{x}}$ to β_{x} under the uninformative model. Each panel shows the ℓ_1 error (vertical axis, log scale) of a portion of $\widehat{\beta}$ as a function of the number of nodes in the network (horizontal axis, log scale). Within each panel, each line represents the error for a single coefficient. We visualize results for $\widehat{\beta}_{\mathrm{w}}$ and $\widehat{\beta}_{\mathrm{x}}$ in separate rows of panels, since only $\widehat{\beta}_{\mathrm{x}}$ is subject to rotational non-identifiability. Panels vary horizontally by number of latent communities (left: two blocks, middle: five block, right: seven blocks).

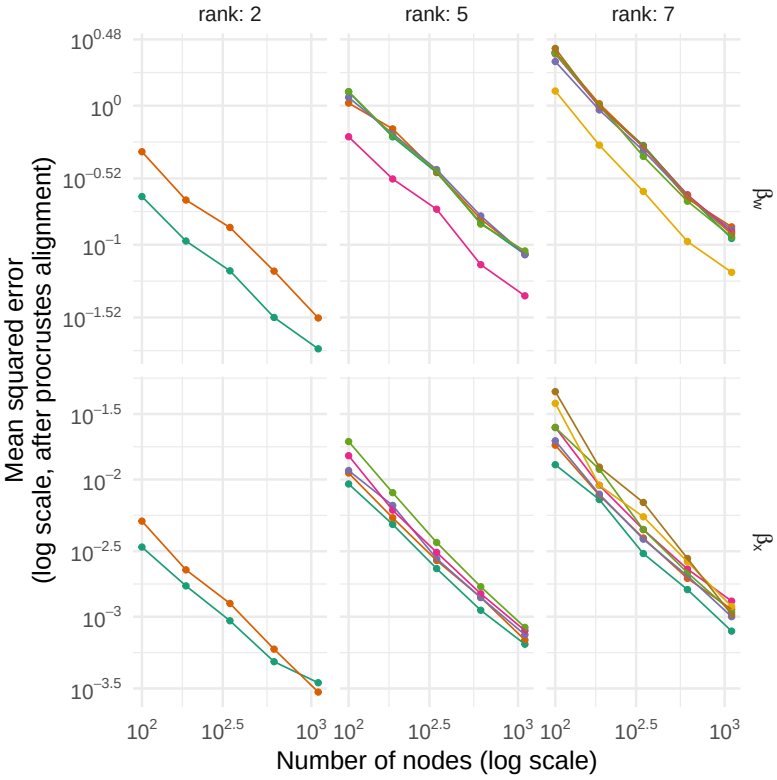


Figure 24: Convergence of $\widehat{\beta}_{\mathrm{w}}$ to β_{w} and $Q\widehat{\beta}_{\mathrm{x}}$ to β_{x} under the informative model. Each panel shows the ℓ_1 error (vertical axis, log scale) of a portion of $\widehat{\beta}$ as a function of the number of nodes in the network (horizontal axis, log scale). Within each panel, each line represents the error for a single coefficient. We visualize results for $\widehat{\beta}_{\mathrm{w}}$ and $\widehat{\beta}_{\mathrm{x}}$ in separate rows of panels, since only $\widehat{\beta}_{\mathrm{x}}$ is subject to rotational non-identifiability. Panels vary horizontally by number of latent communities (left: two blocks, middle: five block, right: seven blocks).

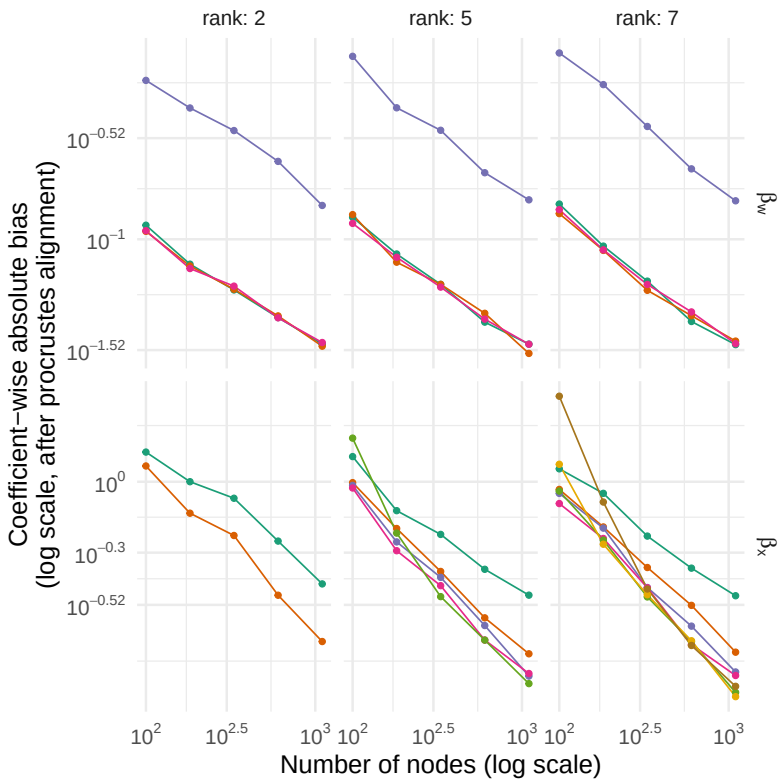


Figure 25: Elementwise bias of $\widehat{\beta}$ for β under the uninformative model. This can be thought of as measurement error bias induced by using $\widehat{X}$ in place of X. Each column of panels corresponds to a distinct model, where models have varying numbers of latent communities. We visualize results for $\widehat{\beta}_{\mathrm{w}}$ and $\widehat{\beta}_{\mathrm{x}}$ in separate rows of plots. Within each panel, each line represents results for a single coefficient. Note that the bias disappears asymptotically, as $\widehat{X}$ converges to X.

that "measurement error" shrinks to zero as the number of nodes in the network grows. Nonetheless, we visualize the finite sample bias of $\widehat{\beta}$ in Figures 25 and 26. Unsurprisingly, $\widehat{\beta}$ is biased, as expected due to the noise in $\widehat{X}$ around X. As n increases and $\widehat{X}$ converges to X, the bias rapidly shrinks.

G.3 Robustness of Causal Point Estimates to Rank Misspecification

In Theorems 17 and 18, the dimension d of the latent space is taken to be known or otherwise correct specified. In Figure 27, we investigate the estimation error of $\widehat{\Psi}_{\mathrm{nde}}$ and $\widehat{\Psi}_{\mathrm{nie}}$ when the dimension of the latent space is misspecified. As in Figure 6 (which investigates coverage rates when d is specified), we find that it is dramatically better, in terms of estimation error, to overestimate d than it is to underestimate d. This aligns with previous results that suggest overestimating the dimension of the embedding space in stochastic blockmodels incurs a performance penalty but otherwise retains nice properties of estimators like consistency (Fishkind et al., 2013).

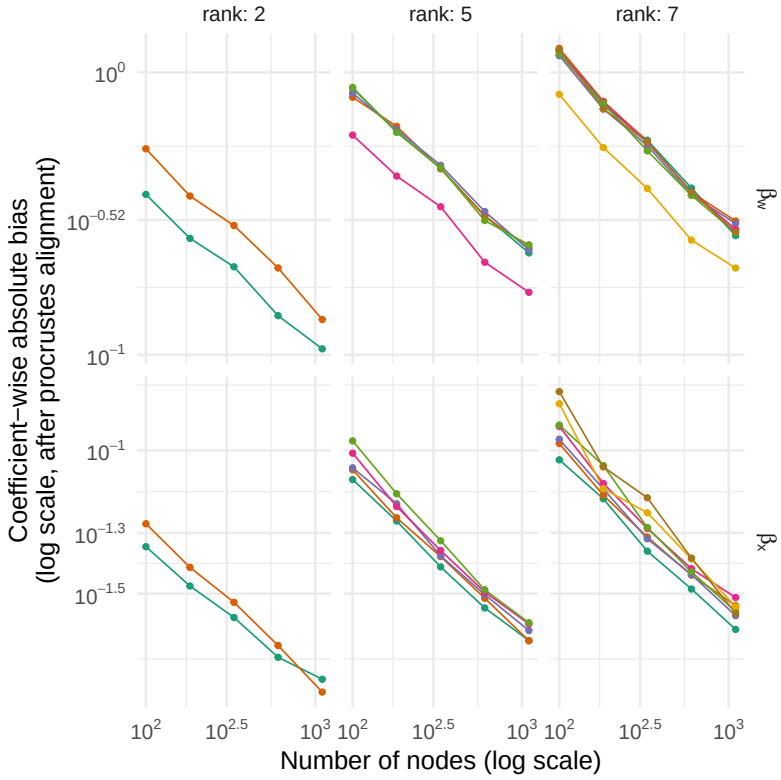


Figure 26: Elementwise bias of $\widehat{\beta}$ for β under the informative model. This can be thought of as measurement error bias induced by using $\widehat{X}$ in place of X. Each column of panels corresponds to a distinct model, where models have varying numbers of latent communities. We visualize results for $\widehat{\beta}_{\mathrm{w}}$ and $\widehat{\beta}_{\mathrm{x}}$ in separate rows of plots. Within each panel, each line represents results for a single coefficient. Note that the bias disappears asymptotically, as $\widehat{X}$ converges to X.

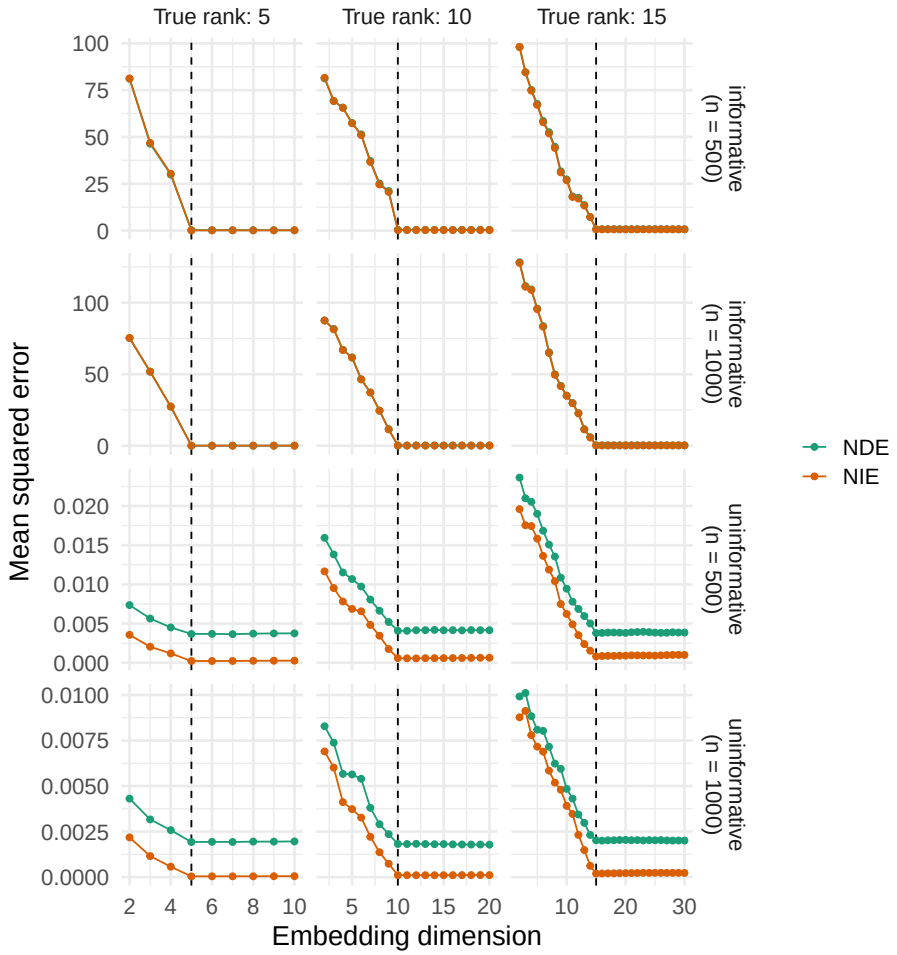


Figure 27: Mean squared error of $\widehat{\Psi}_{\text{nde}}$ and $\widehat{\Psi}_{\text{nie}}$ when the dimension d is misspecified. Each panel shows mean squared error (vertical axis) of Ψ_{nde} (teal) and Ψ_{nie} (orange) as a function of the embedding dimension d (horizontal axis). The dashed vertical line denotes the true latent dimension. Panels vary horizontally by number of latent communities (left: five, middle: ten, right: fifteen) and vertically by the simulation model and number of nodes in the network.

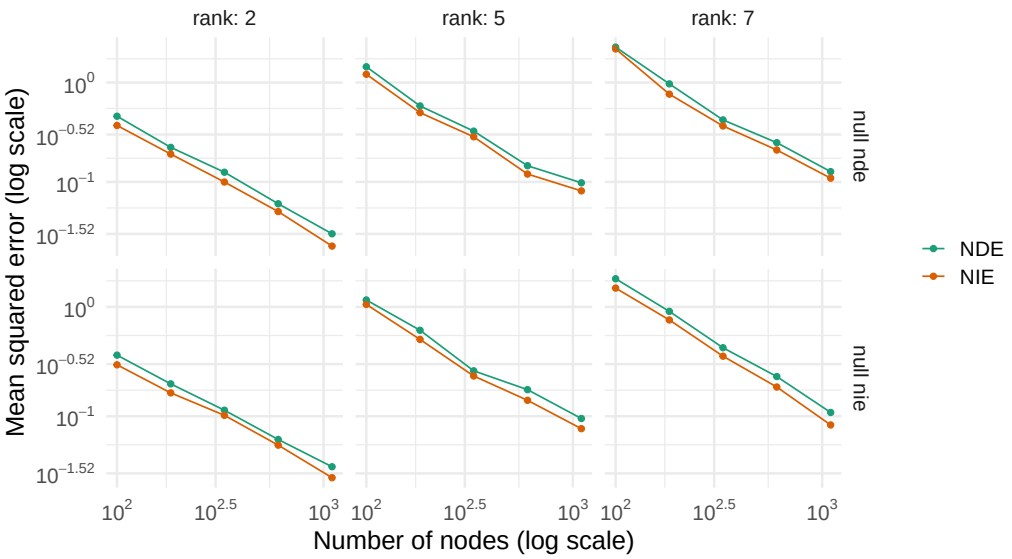


Figure 28: Convergence of $\widehat{\Psi}_{\text{nde}}$ to Ψ_{nde} and $\widehat{\Psi}_{\text{nie}}$ to Ψ_{nie}. Each panel shows the mean squared error (vertical axis, log scale) of $\widehat{\Psi}_{\text{nde}}$ (teal) and $\widehat{\Psi}_{\text{nie}}$ (orange) as a function of the number of nodes in the network (horizontal axis, log scale). Panels vary horizontally by number of latent communities (left: two blocks, middle: five block, right: seven blocks) and vertically by the simulation model (top: informative, bottom: uninformative).

G.4 Causal Estimation Error When Either Ψ_{nde} or Ψ_{nie} is Zero

We additionally investigate if estimation error and converge rates behave as expected when either $\Psi_{\text{nde}} = 0$ or $\Psi_{\text{nie}} = 0$. To generate data where $\Psi_{\text{nde}} = 0$, we simulate from the informative model with $\beta_{\text{t}} = 0$, and so to generate data where $\Psi_{\text{nie}} = 0$, we simulate from the informative model with $\beta_{\text{x}} = 0$. The results, in Figures 28 and 29, show that estimator error and coverage rates do not behave any differently in the setting where Ψ_{nde} or Ψ_{nie} is zero.

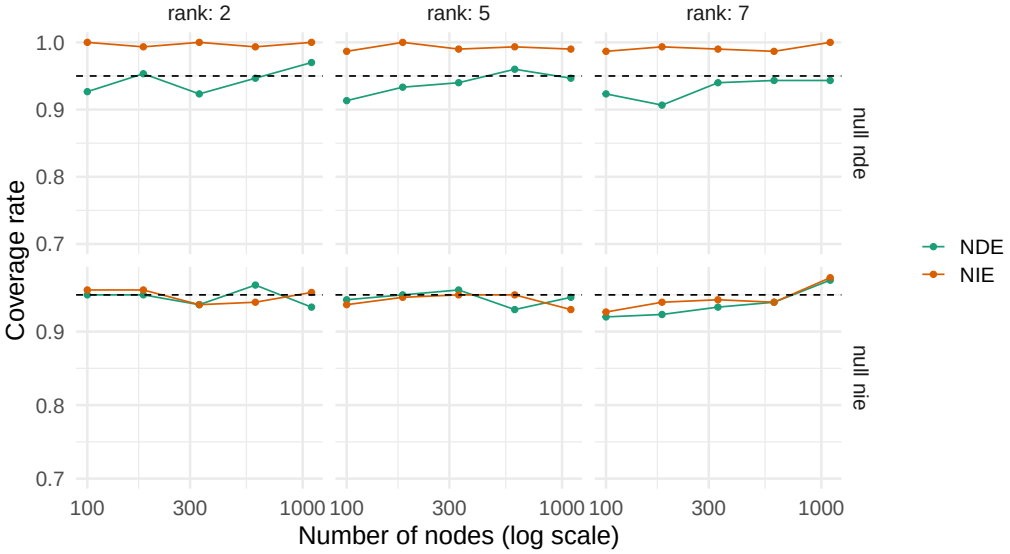


Figure 29: Finite sample coverage of asymptotic confidence intervals for Ψ_{nde} and Ψ_{nie}. Each panel shows coverage (vertical axis) of Ψ_{nde} (teal) and Ψ_{nie} (orange) as a function of the number of nodes in the network (horizontal axis, log scale). The dashed horizontal line denotes the nominal coverage rate of 95%. Panels vary horizontally by number of latent communities (left: two blocks, middle: five block, right: seven blocks) and vertically by the simulation model (top: informative, bottom: uninformative).

References

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008. URL http://jmlr.csail.mit.edu/papers/volume9/airoldi08a/airoldi08a.pdf.

Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, August 2000. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.97.18.10101. URL https://pnas.org/doi/full/10.1073/pnas.97.18.10101.

Avanti Athreya, Carey E. Priebe, Minh Tang, Vince Lyzinski, David J. Marchette, and Daniel L. Sussman. A Limit Theorem for Scaled Eigenvectors of Random Dot Product Graphs. *Sankhya A: The Indian Journal of Statistics*, 78(1):1–18, 2015. ISSN 0976-836X, 0976-8378. doi: 10.1007/s13171-015-0071-x. URL http://link.springer.com/10.1007/s13171-015-0071-x.

Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical Inference on Random Dot Product Graphs: A Survey. *Journal of Machine Learning Research*, 18:1–92, 2018. URL https://jmlr.org/papers/v18/17-448.html.

Avanti Athreya, Minh Tang, Youngser Park, and Carey E Priebe. On Estimation and Inference in Latent Structure Random Graphs. *Statistical Science*, 36(No. 1):68–88, 2021.

N. Binkiewicz, J. T. Vogelstein, and K. Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, June 2017. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asx008. URL https://academic.oup.com/biomet/article/104/2/361/3074977.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Anthony Bonato and Fan R. K. Chung, editors. *Algorithms and Models for the Web-graph: 5th International Workshop, WAW 2007, San Diego, CA, USA, December 11-12, 2007 ; Proceedings*. Number 4863 in Lecture Notes in Computer Science. Springer, Berlin Heidelberg, 2007. ISBN 978-3-540-77003-9.

Dennis D Boos and L. A Stefanski. *Essential Statistical Inference*, volume 120 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013. ISBN 978-1-4614-4817-4 978-1-4614-4818-1. doi: 10.1007/978-1-4614-4818-1. URL http://link.springer.com/10.1007/978-1-4614-4818-1.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, February 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199535255.001.0001/acprof-9780199535255.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, May 2009. ISSN 03044076. doi: 10.1016/j.jeconom.2008.12.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0304407609000335.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Peer Effects in Networks: A Survey. *Annual Review of Economics*, 12(1):603–629, 2020.

Junhui Cai, Dan Yang, Wu Zhu, Haipeng Shen, and Linda Zhao. Network regression and supervised centrality estimation. *arXiv:2111.12921 [cs, econ, stat]*, November 2021. URL http://arxiv.org/abs/2111.12921.

Jae Ho Chang and Subhadeep Paul. Embedding Network Autoregression for time series analysis and causal peer effect inference, June 2024. URL http://arxiv.org/abs/2406.05944.

Chang Che, Ick Hoon Jin, and Zhiyong Zhang. Network Mediation Analysis Using Model-Based Eigenvalue Decomposition. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(1):148–161, January 2021. ISSN 1070-5511, 1532-8007. doi: 10.1080/10705511.2020.1721292. URL https://www.tandfonline.com/doi/full/10.1080/10705511.2020.1721292.

Fan Chen, Sebastien Roch, Karl Rohe, and Shuqi Yu. Estimating Graph Dimension with Cross-validated Eigenvalues. *arXiv:2108.03336 [cs, math, stat]*, August 2021. URL http://arxiv.org/abs/2108.03336.

Xi Chen, Yan Liu, and Cheng Zhang. Distinguishing Homophily from Peer Influence Through Network Representation Learning. *INFORMS Journal on Computing*, 34(4):1958–1969, March 2022. ISSN 1091-9856, 1526-5528. doi: 10.1287/ijoc.2022.1171. URL http://pubsonline.informs.org/doi/10.1287/ijoc.2022.1171.

Lu Cheng, Ruocheng Guo, and Huan Liu. Causal Mediation Analysis with Hidden Confounders. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pages 113–122, New York, NY, USA, February 2022. Association for Computing Machinery. ISBN 978-1-4503-9132-0. doi: 10.1145/3488560.3498407. URL https://doi.org/10.1145/3488560.3498407.

Nicholas A. Christakis and James H. Fowler. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357(4):370–379, July 2007. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMsa066082. URL http://www.nejm.org/doi/abs/10.1056/NEJMsa066082.

Zhixuan Chu, Stephen L. Rathbun, and Sheng Li. Graph Infomax Adversarial Learning for Treatment Effect Estimation with Networked Observational Data. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 176–184, August 2021. doi: 10.1145/3447548.3467302. URL http://arxiv.org/abs/2106.02881.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A Crash Course in Good and Bad Controls. *Sociological Methods & Research*, pages 1–34, 2022. ISSN 1556-5068. doi: 10.1177/00491241221099552. URL https://www.ssrn.com/abstract=3689437.

Irina Cristali and Victor Veitch. Using Embeddings for Causal Estimation of Peer Influence in Social Networks. *arXiv:2205.08033 [cs, stat]*, May 2022. doi: 10.48550/arXiv.2205.08033. URL http://arxiv.org/abs/2205.08033.

Chiara Di Maria, Antonino Abbruzzo, and Gianfranco Lovison. Networks as mediating variables: A Bayesian latent space approach. *Statistical Methods & Applications*, 31:1015–1035, February 2022. ISSN 1618-2510, 1613-981X. doi: 10.1007/s10260-022-00621-w. URL https://link.springer.com/10.1007/s10260-022-00621-w.

Patrick Doreian. Estimating Linear Models with Spatially Distributed Data. *Sociological Methodology*, 12:359–388, 1981. ISSN 00811750. doi: 10.2307/270747. URL https://www.jstor.org/stable/270747?origin=crossref.

Oliver Dukes, Ilya Shpitser, and Eric J. Tchetgen Tchetgen. Proximal mediation analysis. *arXiv:2109.11904 [stat]*, September 2021. URL http://arxiv.org/abs/2109.11904.

Naoki Egami and Eric J. Tchetgen Tchetgen. Identification and Estimation of Causal Peer Effects Using Double Negative Controls for Unmeasured Network Confounding. *arXiv:2109.01933 [stat]*, September 2021. URL http://arxiv.org/abs/2109.01933.

Beate Ehrhardt and Patrick J. Wolfe. Network Modularity in the Presence of Covariates. *SIAM Review*, 61(2):261–276, January 2019. ISSN 0036-1445, 1095-7200. doi: 10.1137/17M1111528. URL https://epubs.siam.org/doi/10.1137/17M1111528.

Helmut Farbmacher, Martin Huber, Lukáš Lafférs, Henrika Langen, and Martin Spindler. Causal mediation analysis with double machine learning. *The Econometrics Journal*, 25(2):277–300, May 2022. ISSN 1368-4221. doi: 10.1093/ectj/utac003. URL https://doi.org/10.1093/ectj/utac003.

Donniell E. Fishkind, Daniel L. Sussman, Minh Tang, Joshua T. Vogelstein, and Carey E. Priebe. Consistent Adjacency-Spectral Partitioning for the Stochastic Block Model When the Model Parameters Are Unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, January 2013. ISSN 0895-4798, 1095-7162. doi: 10.1137/120875600. URL http://epubs.siam.org/doi/10.1137/120875600.

Bailey K. Fosdick and Peter D. Hoff. Testing and Modeling Dependencies Between a Network and Nodal Attributes. *Journal of the American Statistical Association*, 110(511):1047–1056, July 2015. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2015.1008697. URL http://www.tandfonline.com/doi/full/10.1080/01621459.2015.1008697.

Mark M. Fredrickson and Yuguo Chen. Permutation and randomization tests for network analysis. *Social Networks*, 59:171–183, October 2019. ISSN 03788733. doi: 10.1016/j.socnet.2019.08.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0378873318303381.

Anna Freier, Johannes Kruse, Bjarne Schmalbach, Sandra Zara, Samuel Werner, Elmar Brähler, Jörg M. Fegert, and Hanna Kampling. Supplementary data for the mediation effect of personality functioning – Gender differences, separate analyses of depression and anxiety symptoms and inferential statistics of the relationship between personality

functioning and different types of child maltreatment. *Data in Brief*, 42:108272, June 2022. ISSN 23523409. doi: 10.1016/j.dib.2022.108272. URL https://linkinghub.elsevier.com/retrieve/pii/S2352340922004747.

Ragnar Frisch and Frederick V. Waugh. Partial Time Regressions as Compared with Individual Trends. *Econometrica*, 1(4):387, October 1933. ISSN 00129682. doi: 10.2307/1907330. URL https://www.jstor.org/stable/1907330?origin=crossref.

Isabel R. Fulcher, Ilya Shpitser, Stella Marealle, and Eric J. Tchetgen Tchetgen. Robust inference on population indirect causal effects: The generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82 (1):199–214, February 2020. ISSN 13697412. doi: 10.1111/rssb.12345. URL https://onlinelibrary.wiley.com/doi/10.1111/rssb.12345.

Lucy L. Gao, Daniela Witten, and Jacob Bien. Testing for association in multiview network data. *Biometrics*, 78(3):1018–1030, September 2022. ISSN 0006-341X, 1541-0420. doi: 10.1111/biom.13464. URL https://onlinelibrary.wiley.com/doi/10.1111/biom.13464.

Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. A network approach to topic models. *Science Advances*, 4(7):1–11, July 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aaq1360. URL http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aaq1360.

AmirEmad Ghassami, Ilya Shpitser, and Eric J. Tchetgen Tchetgen. Proximal Causal Inference with Hidden Mediators: Front-Door and Related Mediation Problems. *arXiv:2111.02927 [math, stat]*, November 2021. URL http://arxiv.org/abs/2111.02927.

Brian Gilbert, Abhirup Datta, and Elizabeth Ogburn. Approaches to spatial confounding in geostatistics. *arXiv:2112.14946 [stat]*, December 2021. URL http://arxiv.org/abs/2112.14946.

John C. Gower and Garmt B. Dijksterhuis. *Procrustes Problems*. Number 30 in Oxford Statistical Science Series. Oxford University Press, Oxford ; New York, 2004. ISBN 978-0-19-851058-1.

T. N. E. Greville. Note on the Generalized Inverse of a Matrix Product. *SIAM Review*, 8(4):518–521, October 1966. ISSN 0036-1445, 1095-7200. doi: 10.1137/1008107. URL http://epubs.siam.org/doi/10.1137/1008107.

Sharmistha Guha and Abel Rodriguez. Bayesian Regression With Undirected Network Predictors With an Application to Brain Connectome Data. *Journal of the American Statistical Association*, 116(534):581–593, April 2021. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2020.1772079. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1772079.

Ruocheng Guo, Jundong Li, and Huan Liu. Counterfactual Evaluation of Treatment Assignment Functions with Networked Observational Data. In *Proceedings of the 2020 SIAM*

International Conference on Data Mining (SDM), Proceedings, pages 271–279. Society for Industrial and Applied Mathematics, January 2020. doi: 10.1137/1.9781611976236.31. URL https://epubs.siam.org/doi/10.1137/1.9781611976236.31.

Qiuyi Han, Kevin S Xu, and Edoardo M Airoldi. Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32 Nd International Conference on Machine Learning*, volume Volume 37, Lille, France, 2015. URL http://proceedings.mlr.press/v37/hanb15.pdf.

Xiao Han, Qing Yang, and Yingying Fan. Universal Rank Inference via Residual Subsampling with Application to Large Networks. *arXiv:1912.11583 [math, stat]*, July 2020. URL http://arxiv.org/abs/1912.11583.

Yanjun He and Peter D. Hoff. Multiplicative coevolution regression models for longitudinal networks and nodal attributes. *Social Networks*, 57:54–62, May 2019. ISSN 03788733. doi: 10.1016/j.socnet.2018.12.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0378873318305008.

Yinqiu He, Peter X K Song, and Gongjun Xu. Adaptive bootstrap tests for composite null hypotheses in the mediation pathway analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):411–434, April 2024. ISSN 1369-7412, 1467-9868. doi: 10.1093/jrsssb/qkad129. URL https://academic.oup.com/jrsssb/article/86/2/411/7420463.

Matthew J. Hirshberg, Corrina Frye, Cortland J. Dahl, Kevin M. Riordan, Nathan J. Vack, Jane Sachs, Robin Goldman, Richard J. Davidson, and Simon B. Goldberg. A randomized controlled trial of a smartphone-based well-being training in public school system employees during the COVID-19 pandemic. *Journal of Educational Psychology*, 114(8): 1895–1911, November 2022. ISSN 0022-0663. doi: 10.1037/edu0000739. URL https://ezproxy.library.wisc.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2022-42990-001&site=ehost-live&scope=site.

Matthew J. Hirshberg, Cortland J. Dahl, Daniel Bolt, Richard J. Davidson, and Simon B. Goldberg. Psychological Mediators of Reduced Distress: Preregistered Analyses From a Randomized Controlled Trial of a Smartphone-Based Well-Being Training. *Clinical Psychological Science*, page 21677026241233262, March 2024. ISSN 2167-7026. doi: 10.1177/21677026241233262. URL https://doi.org/10.1177/21677026241233262.

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, December 2002. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214502388618906. URL http://www.tandfonline.com/doi/abs/10.1198/016214502388618906.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, June 1983. ISSN 03788733. doi: 10.1016/0378-8733(83)90021-7. URL https://linkinghub.elsevier.com/retrieve/pii/0378873383900217.

Yuchen Hu, Shuangning Li, and Stefan Wager. Average direct and indirect causal effects under interference. *Biometrika*, 109(4):1165–1172, November 2022. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asac008. URL https://academic.oup.com/biomet/article/109/4/1165/6524620.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1), February 2010. ISSN 0883-4237. doi: 10.1214/10-STS321. URL https://projecteuclid.org/journals/statistical-science/volume-25/issue-1/Identification-Inference-and-Sensitivity-Analysis-for-Causal-Mediation-Effects/10.1214/10-STS321.full.

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Mixed membership estimation for social networks. *Journal of Econometrics*, 239(2):105369, February 2024. ISSN 0304-4076. doi: 10.1016/j.jeconom.2022.12.003. URL https://www.sciencedirect.com/science/article/pii/S0304407622002081.

Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, January 2011. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.83.016107. URL http://arxiv.org/abs/1008.3926.

Katherine Keith, Douglas Rice, and Brendan O'Connor. Text as Causal Mediators: Research Design for Causal Estimates of Differential Treatment of Social Groups via Language Aspects. In *Proceedings of CI+NLP: First Workshop on Causal Inference and NLP*, pages 21–32, November 2021. URL https://aclanthology.org/2021.cinlp-1.2.pdf.

Kenneth C. Land and Glenn Deane. On the Large-Sample Estimation of Regression Models with Spatial- Or Network-Effects Terms: A Two-Stage Least Squares Approach. *Sociological Methodology*, 22:221, 1992. ISSN 00811750. doi: 10.2307/270997. URL https://www.jstor.org/stable/270997?origin=crossref.

Boris Landa, Thomas T. C. K. Zhang, and Yuval Kluger. Biwhitening Reveals the Rank of a Count Matrix. *arXiv:2103.13840 [cs, math, stat]*, March 2021. URL http://arxiv.org/abs/2103.13840.

Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, March 2011. ISSN 1932-6157. doi: 10.1214/10-AOAS382. URL http://arxiv.org/abs/0910.2098.

Can M. Le and Tianxi Li. Linear regression and its inference on noisy network-linked data. *arXiv:2007.00803 [stat]*, August 2022. URL http://arxiv.org/abs/2007.00803.

Lung-Fei Lee. Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory*, 18(2):252–277, April 2002. ISSN 0266-4666, 1469-4360. doi: 10.1017/S0266466602182028. URL https://www.cambridge.org/core/product/identifier/S0266466602182028/type/journal_article.

Youjin Lee, Cencheng Shen, Carey E Priebe, and Joshua T Vogelstein. Network dependence testing via diffusion maps and distance-based correlations. *Biometrika*, 106(4):857–873, December 2019. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asz045. URL https://academic.oup.com/biomet/article/106/4/857/5578430.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, February 2015. ISSN 0090-5364. doi: 10.1214/14-AOS1274. URL http://arxiv.org/abs/1312.2050.

James P. LeSage and R. Kelley Pace. *Introduction to Spatial Econometrics*. Statistics, Textbooks and Monographs. CRC Press, Boca Raton, 2009. ISBN 978-1-4200-6424-7.

Michael Leung. Causal Inference Under Approximate Neighborhood Interference. *SSRN Electronic Journal*, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3479902. URL https://www.ssrn.com/abstract=3479902.

Keith Levin, Asad Lodhia, and Elizaveta Levina. Recovering shared structure from multiple networks with unknown edge distributions. *Journal of Machine Learning Research*, 23: 1–48, 2022.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132–164, March 2019. ISSN 1932-6157. doi: 10.1214/18-AOAS1205. URL https://projecteuclid.org/euclid.aoas/1554861644.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, June 2020. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asaa006. URL https://academic.oup.com/biomet/article/107/2/257/5816043.

Jennifer Listgarten, Carl Kadie, Eric E. Schadt, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465–16470, September 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1002425107. URL https://pnas.org/doi/full/10.1073/pnas.1002425107.

Haiyan Liu, Ick Hoon Jin, Zhiyong Zhang, and Ying Yuan. Social Network Mediation Analysis: A Latent Space Approach. *Psychometrika*, 86(1):272–298, March 2021. ISSN 0033-3123, 1860-0980. doi: 10.1007/s11336-020-09736-z. URL https://link.springer.com/10.1007/s11336-020-09736-z.

Lan Liu and Eric Tchetgen Tchetgen. Regression-based negative control of homophily in dyadic peer effect analysis. *Biometrics*, 78(2):668–678, June 2022. ISSN 0006-341X, 1541-0420. doi: 10.1111/biom.13483. URL https://onlinelibrary.wiley.com/doi/10.1111/biom.13483.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal Effect Inference with Deep Latent-Variable Models. In *31st Conference on Neural Information Processing Systems*, page 11, Long Beach, CA, USA., 2017.

Michael C Lovell. Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963. URL https://www.jstor.org/stable/2283327.

Vince Lyzinski, Daniel L. Sussman, Minh Tang, Avanti Athreya, and Carey E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922, January 2014. ISSN 1935-7524, 1935-7524. doi: 10.1214/14-EJS978. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-8/issue-2/Perfect-clustering-for-stochastic-blockmodel-graphs-via-adjacency-spectral-embedding/10.1214/14-EJS978.full.

Charles F. Manski. Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, 60(3):531, July 1993. ISSN 00346527. doi: 10.2307/2298123. URL https://academic.oup.com/restud/article-lookup/doi/10.2307/2298123.

Edward McFowland and Cosma Rohilla Shalizi. Estimating Causal Peer Influence in Homophilous Social Networks by Inferring Latent Locations. *Journal of the American Statistical Association*, 0(0):1–12, July 2021. ISSN 0162-1459. doi: 10.1080/01621459.2021.1953506. URL https://doi.org/10.1080/01621459.2021.1953506.

Clare M. Mehta and JoNell Strough. Sex segregation in friendships and normative contexts across the life span. *Developmental Review*, 29(3):201–220, September 2009. ISSN 02732297. doi: 10.1016/j.dr.2009.06.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0273229709000227.

Angelo Mele, Lingxin Hao, Joshua Cape, and Carey E. Priebe. Spectral Estimation of Large Stochastic Blockmodels with Discrete Nodal Covariates. *Journal of Business & Economic Statistics*, 41(4):1364–1376, October 2023. ISSN 0735-0015, 1537-2707. doi: 10.1080/07350015.2022.2139709. URL https://www.tandfonline.com/doi/full/10.1080/07350015.2022.2139709.

L. Michell and P. West. Peer pressure to smoke: The meaning depends on the method. *Health Education Research*, 11(1):39–49, 1996. ISSN 0268-1153, 1465-3648. doi: 10.1093/her/11.1.39. URL https://academic.oup.com/her/article-lookup/doi/10.1093/her/11.1.39.

Lynn Michell. Loud, sad or bad: Young people's perceptions of peer groups and smoking. *Health Education Research*, 12(1):1–14, 1997. ISSN 0268-1153, 1465-3648. doi: 10.1093/her/12.1.1-a. URL https://academic.oup.com/her/article-lookup/doi/10.1093/her/12.1.1-a.

Lynn Michell and Amanda Amos. Girls, pecking order and smoking. *Social Science & Medicine*, 44(12):1861–1869, June 1997. ISSN 02779536. doi: 10.1016/S0277-9536(96)00295-X. URL https://linkinghub.elsevier.com/retrieve/pii/S027795369600295X.

Michael Pearson Michell, Lynn. Smoke Rings: Social network analysis of friendship groups, smoking and drug-taking. *Drugs: Education, Prevention and Policy*, 7(1): 21–37, January 2000. ISSN 0968-7637, 1465-3370. doi: 10.1080/dep.7.1.21.37. URL http://www.tandfonline.com/doi/full/10.1080/dep.7.1.21.37.

Shanjukta Nath, Keith Warren, and Subhadeep Paul. Identifying Peer Influence in Therapeutic Communities, October 2023. URL http://arxiv.org/abs/2203.14223.

Trang Quynh Nguyen, Elizabeth L. Ogburn, Elizabeth B. Sarker, Noah Greifer, Ian Schmid, Ina M. Koning, and Elizabeth A. Stuart. Causal mediation analysis: From simple to more robust strategies for estimation of marginal natural (in)direct effects. *arXiv:2102.06048 [stat]*, May 2021. URL http://arxiv.org/abs/2102.06048.

Christine Leigh Myers Nickel. *Random Dot Product Graphs a Model for Social Networks*. PhD thesis, John Hopkins University, 2006.

Elizabeth L. Ogburn, Ilya Shpitser, and Youjin Lee. Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1659–1676, October 2020. ISSN 0964-1998, 1467-985X. doi: 10.1111/rssa.12594. URL https://onlinelibrary.wiley.com/doi/10.1111/rssa.12594.

Elizabeth L. Ogburn, Oleg Sofrygin, Iván Díaz, and Mark J. van der Laan. Causal Inference for Social Network Data. *Journal of the American Statistical Association*, 0(ja):1–46, October 2022. ISSN 0162-1459. doi: 10.1080/01621459.2022.2131557. URL https://doi.org/10.1080/01621459.2022.2131557.

A. James O'Malley, Felix Elwert, J. Niels Rosenquist, Alan M. Zaslavsky, and Nicholas A. Christakis. Estimating peer effects in longitudinal dyadic data using instrumental variables. *Biometrics*, 70(3):506–515, 2014. ISSN 1541-0420. doi: 10.1111/biom.12172. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12172.

Keith Ord. Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, 70(349):120–126, March 1975. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1975.10480272. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1975.10480272.

Subhadeep Paul, Shanjukta Nath, and Keith Warren. Causal Network Influence with Latent Homophily and Measurement Error: An Application to Therapeutic Community. *arXiv:2203.14223 [stat]*, March 2022a. URL http://arxiv.org/abs/2203.14223.

Subhadeep Paul, Shanjukta Nath, and Keith Warren. Network Influence with Latent Homophily and Measurement Error, August 2022b. URL http://arxiv.org/abs/2203.14223.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Minh Tang, Avanti Athreya, Joshua Cape, and Eric Bridgeford. On a two-truths

phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, 116(13):5995–6000, March 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1814462116. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1814462116.

Huan Qing and Jingli Wang. Directed mixed membership stochastic blockmodel. *arXiv:2101.02307 [cs, stat]*, October 2021. URL http://arxiv.org/abs/2101.02307.

Karl Rohe and Muzhe Zeng. Vintage Factor Analysis with Varimax Performs Statistical Inference. *arXiv:2004.05387 [math, stat]*, 2022. URL http://arxiv.org/abs/2004.05387.

Karl Rohe and Muzhe Zeng. Vintage factor analysis with Varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4): 1037–1060, September 2023. ISSN 1369-7412, 1467-9868. doi: 10.1093/jrsssb/qkad029. URL https://academic.oup.com/jrsssb/article/85/4/1037/7221295.

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45):12679–12684, November 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1525793113. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1525793113.

Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E. Priebe. A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4):1446–1473, 2022. ISSN 1467-9868. doi: 10.1111/rssb.12509. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12509.

Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research*, 40(2):211–239, May 2011. ISSN 0049-1241, 1552-8294. doi: 10.1177/0049124111404820. URL http://journals.sagepub.com/doi/10.1177/0049124111404820.

Lin Su, Wenbin Lu, Rui Song, and Danyang Huang. Testing and Estimation of Social Network Dependence With Time to Event Data. *Journal of the American Statistical Association*, 115(530):570–582, April 2020. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2019.1617153. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2019.1617153.

Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, and Carey E. Priebe. A Consistent Adjacency Spectral Embedding for Stochastic Blockmodel Graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, September 2012. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2012.699795. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2012.699795.

Daniel L. Sussman, Minh Tang, and Carey E. Priebe. Consistent Latent Position Estimation and Vertex Classification for Random Dot Product Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):48–57, January 2014. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2013.135. URL http://ieeexplore.ieee.org/document/6565321/.

Tracy M. Sweet. Modeling Social Networks as Mediators: A Mixed Membership Stochastic Blockmodel for Mediation. *Journal of Educational and Behavioral Statistics*, 44(2):210–240, April 2019. ISSN 1076-9986, 1935-1054. doi: 10.3102/1076998618814255. URL http://journals.sagepub.com/doi/10.3102/1076998618814255.

Tracy M. Sweet and Samrachana Adhikari. A hierarchical latent space network model for mediation. *Network Science*, pages 1–18, May 2022. ISSN 2050-1242, 2050-1250. doi: 10.1017/nws.2022.12. URL https://www.cambridge.org/core/product/identifier/S2050124222000121/type/journal_article.

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, and Carey E. Priebe. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23(3), August 2017. ISSN 1350-7265. doi: 10.3150/15-BEJ789. URL https://projecteuclid.org/journals/bernoulli/volume-23/issue-3/A-nonparametric-two-sample-hypothesis-testing-problem-for-random-graphs/10.3150/15-BEJ789.full.

Eric J. Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics*, 40(3), June 2012. ISSN 0090-5364. doi: 10.1214/12-AOS990. URL https://projecteuclid.org/journals/annals-of-statistics/volume-40/issue-3/Semiparametric-theory-for-causal-mediation-analysis--Efficiency-bounds-multiple/10.1214/12-AOS990.full.

L. L. Thurstone. The vectors of mind. *Psychological Review*, 41(1):1–32, 1934. ISSN 0033-295X. doi: 10.1037/h0075959. URL http://content.apa.org/journals/rev/41/1/1.

L. L. Thurstone. Factorial analysis of body measurements. *American Journal of Physical Anthropology*, 5(1):15–28, March 1947. ISSN 0002-9483, 1096-8644. doi: 10.1002/ajpa.1330050103. URL http://doi.wiley.com/10.1002/ajpa.1330050103.

Michael Tiefelsdorf and Daniel A Griffith. Semiparametric Filtering of Spatial Autocorrelation: The Eigenvector Approach. *Environment and Planning A: Economy and Space*, 39(5):1193–1221, May 2007. ISSN 0308-518X, 1472-3409. doi: 10.1068/a37378. URL http://journals.sagepub.com/doi/10.1068/a37378.

Ralph Møller Trane. *Practical Insights into Causal Methods: Nonparametric IV Bounds and ANCOVA under General Interference*. PhD thesis, University of Wisconsin-Madison, Madison, WI, December 2023. URL https://www.proquest.com/pqdtglobal/docview/2903265703/2C1FAB9582FE4AFBPQ/1?accountid=465&sourcetype=Dissertations%20&%20Theses.

Joel A. Tropp. *An Introduction to Matrix Concentration Inequalities*. Now Publishers, 2015. ISBN 978-1-60198-839-3. URL https://ieeexplore.ieee.org/document/8186879.

Elizabeth Upton and Luis Carvalho. Bayesian Network Regularized Regression for Modeling Urban Crime Occurrences. *arXiv:1708.05047 [stat]*, August 2017. URL http://arxiv.org/abs/1708.05047.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK ; New York, NY, USA, 1998. ISBN 978-0-521-49603-2.

Tyler J. VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York, 2015. ISBN 978-0-19-932587-0.

Tyler J. VanderWeele and Stijn Vansteelandt. Mediation Analysis with Multiple Mediators. *Epidemiologic methods*, 2(1):95–115, January 2014. ISSN 2194-9263. doi: 10.1515/em-2012-0010. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287269/.

Victor Veitch, Yixin Wang, and David Blei. Using Embeddings to Correct for Unobserved Confounding in Networks. In *33rd Conference on Neural Information Processing Systems*, page 11, Vancouver, Canada, 2019. URL https://papers.nips.cc/paper/2019/file/af1c25e88a9e818f809f6b5d18ca02e2-Paper.pdf.

Victor Veitch, Dhanya Sridhar, and David M Blei. Adapting Text Embeddings for Causal Inference. In *Proceedings of the 36 Th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, page 10, 2020. URL http://proceedings.mlr.press/v124/veitch20a/veitch20a.pdf.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 1 edition, June 2020. ISBN 978-1-108-23159-6 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://www.cambridge.org/core/product/identifier/9781108231596/type/book.

G. N. Wilkinson and C. E. Rogers. Symbolic Description of Factorial Models for Analysis of Variance. *Applied Statistics*, 22(3):392, 1973. ISSN 00359254. doi: 10.2307/2346786. URL https://www.jstor.org/stable/10.2307/2346786?origin=crossref.

Dingbo Wu and Fangzheng Xie. Statistical inference of random graphs with a surrogate likelihood function. *arXiv:2207.01702 [math, stat]*, July 2022. URL http://arxiv.org/abs/2207.01702.

Fangzheng Xie and Yanxun Xu. Optimal Bayesian estimation for random dot product graphs. *Biometrika*, 107(4):875–889, December 2020. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asaa031. URL https://academic.oup.com/biomet/article/107/4/875/5867841.

Fangzheng Xie and Yanxun Xu. Efficient Estimation for Random Dot Product Graphs via a One-Step Procedure. *Journal of the American Statistical Association*, pages 1–14, August 2021. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2021.1948419. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2021.1948419.

Yini Zhang, Fan Chen, and Karl Rohe. Social Media Public Opinion as Flocks in a Murmuration: Conceptualizing and Measuring Opinion Expression on Social Media. *Journal of Computer-Mediated Communication*, 27(1):zmab021, November 2021. ISSN 1083-6101. doi: 10.1093/jcmc/zmab021. URL https://academic.oup.com/jcmc/article/doi/10.1093/jcmc/zmab021/6472790.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting Overlapping Communities in Networks Using Spectral Methods. *SIAM Journal on Mathematics of Data Science*, 2(2):265–283, January 2020. ISSN 2577-0187. doi: 10.1137/19M1272238. URL https://epubs.siam.org/doi/10.1137/19M1272238.

Yi Zhao, Martin A. Lindquist, and Brian S. Caffo. Sparse Principal Component based High-Dimensional Mediation Analysis. *Computational Statistics & Data Analysis*, 142: 106835, February 2020. ISSN 01679473. doi: 10.1016/j.csda.2019.106835. URL http://arxiv.org/abs/1806.06118.

Yize Zhao, Tianqi Chen, Jiachen Cai, Sarah Lichenstein, Marc N. Potenza, and Sarah W. Yip. Bayesian network mediation analysis with application to the brain functional connectome. *Statistics in Medicine*, 41(20):3991–4005, 2022. ISSN 1097-0258. doi: 10.1002/sim.9488. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9488.

Wenjing Zheng and Mark J. van der Laan. Targeted Maximum Likelihood Estimation of Natural Direct Effects. *The International Journal of Biostatistics*, 8(1):1–40, January 2012. ISSN 1557-4679. doi: 10.2202/1557-4679.1361. URL https://www.degruyter.com/document/doi/10.2202/1557-4679.1361/html.