Leveraging Graph Structure for Optimal Design, Estimation, and Inference under Interference

Vydhourie R T Thiyageswaran

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Maryam Fazel, Chair

Tyler H. McCormick, Chair

Marina Meilă

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Leveraging Graph Structure for Optimal Design, Estimation, and Inference under Interference

Vydhourie R T Thiyageswaran

Co-Chairs of the Supervisory Committee:
Professor Maryam Fazel
Department of Electrical and Computer Engineering

Professor Tyler H. McCormick
Department of Statistics

Many systems of scientific interest are composed of interacting and interdependent units. In such settings, observations cannot generally be treated as independent, and in experiments, one unit's treatment assignment may affect the outcomes of other units through interference. This dissertation studies inference, testing, estimation, and experimental design in these dependent-data settings, with a particular focus on leveraging network structure.

First, I develop general covariance-based sufficient conditions for multivariate central limit theorems with dependent triangular arrays. The framework allows nonzero dependence between all observations and I show how it encompasses several commonly studied dependence structures. Second, I study conditional randomization tests for detecting network interference by conditioning on focal units, deriving power characterizations that motivate optimization-based focal-unit selection. Third, I consider treatment-effect estimation when the correct exposure threshold is unknown. I propose a data-adaptive procedure that estimates the bias–variance trade-off across candidate exposure thresholds and selects the threshold minimizing estimated mean squared error. Finally, I study optimal experimental design under network interference, homophily, and heterogeneous variation. I derive worst-case mean-squared-error bounds that lead to an optimization over the covariance matrix of the treatment assignment, and develop design procedures based on semidefinite programming with Gaussian rounding

and vector balancing via the Gram–Schmidt Walk.

Together, these results study how dependence structure can be used at different stages of statistical analysis: to justify inference, improve detection of interference, adapt treatment-effect estimators, and construct optimal experimental designs.

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors and committee members for the mentorship, support, brilliance, and inspiration they shared with me during my PhD. From each of them, I learned to approach problems with creativity, joy, and resilience. To feel both intellectually stimulated and emotionally supported is a rare privilege for a PhD student, and I am deeply grateful to have experienced it. Thank you to Maryam Fazel and Tyler McCormick for giving me so much freedom in the research I chose to pursue. Maryam, thank you for your unwavering support, for sharing your wealth of knowledge, and your immense generosity in providing me with guidance even during your busiest times. I will always carry your encouragement to think deeply about problems. Tyler, your kindness, patience, and gentle prodding of my ideas created a wonderful space for open discussion. I will cherish our whiteboard sessions and your reminders to always look at the big picture rather than getting lost in the weeds, and to celebrate even the small wins. To Marina Meilă, thank you so very much for your constant support all the way from admissions till graduation, your creative and exciting research ideas, your encouragement to join the NSA problem sessions, which brought me new excitement, and countless many other things. I will miss dropping by your office just to say hello, and the patio parties. To Stefan Steinerberger, thank you for having always been so kind to me and for generously listening to my half-baked ideas, especially early on. You have shaped how I want to approach problems – fearlessly and creatively.

I am also deeply grateful to Jennifer Brennan, who kindly took me under her wing to explore network interference. Your patience in helping me navigate a new field, and your friendship has meant so much to me. To Christina Lee Yu, for your wonderful mentorship and guidance on both research and on navigating my career. To Arun Chandrasekhar, for your incredible support during my first few years working on the CLT project; you showed me how fun, quick, excited, and creative thinking can be. To Matt Jackson and Thomas Rothvoss, thank you for generously sharing ideas and listening to my research problems. I

have also benefited immensely from the insights of many other professors, including Chris Harshaw, Daniela Witten, Kevin Jamieson, Rene Carmona, and Thomas Richardson. A special thank you to Miklos Racz and Michael Guerzhoy for allowing me the freedom to think creatively during my independent research at Princeton, which gave me the confidence to pursue a PhD. Miki, your continued mentorship through advisor selections, collaborations, and career moves has been invaluable.

I am grateful for the UW Statistics administrative staff, especially Ellen Reynolds, Kristine Chan, Tracy Pham, and Veronica Bei. They were constants without which the whole system would probably fall apart.

To Laura Leal, thank you for your constant generosity. From sharing my first LaTeX template to offering advice on courses, advisors, and jobs, your friendship has meant so much to me. To Fernando Taboada and the late Jorge Sarmiento, thank you for taking a chance on me and encouraging my first taste of research during my freshman summer at Princeton. To Pn. Renuka, thank you for seeing my potential and giving me the space to express myself in high school. To Mr. Alderson who encouraged me to pursue maths and acting at university, thank you.

During my PhD, Dominique Perrault-Joncas had been an incredible source of honest advice and support, from internships to securing a job. My industry internships have taught me invaluable lessons about research and life; thank you especially to Ayal Chen-Zion, Aurelien Bibaut, Lorenzo Masoero, Nathan Kallus, and Stefan Hut, for the intellectually stimulating conversations. To Lorenzo and Aurelien, thank you also for your generous advice navigating post-graduation plans.

To Lilly Meng and Sean-Wyn Ng, thank you for making my Seattle adventures so meaningful. I will miss watching tennis grand-slams with Lilly, Tommy, and Sung. I am grateful that I got to share parts of my PhD experience with Anand Hemmady, Anupreet Porwal, Antonio Olivas-Martinez, Apara Venkat, Erin Lipman, Elizabeth Maher, Garrett Mulcahy, James Buenfil, Jess Kunke, Lars van der Laan, Nina Galanter, Sara LaPlante, Shreya Prakash, and Yikun Zhang. I will fondly remember Momi dinners with Anand, and Sunday rundays with Jess and gang. Thank you to Lars and Seth Temple for generously reading my work and providing feedback.

To Annie Song, Carmen Ng, James Lee, Janie Kim, Jean Luo, Livia Qoshe, and Qi Zhuang Siah, thank you for your support, for bringing so much joy into my life, and for reminding me that I have great friends, by construction. To Nardeen Khella, you have been the sister I never knew I wanted, checking in during tough times while being both the silliest and strongest person in the room.

To Alex, thank you for always being there for me. I deeply appreciate all the excited times we shared research ideas/thoughts, your encouragement to think outside the box when I wasn't, and your constant, unconditional support and kindness, especially on days when I struggled to be kind to myself.

Finally, to my parents: thank you for being my pillars of strength and for always believing in me. Thank you for sacrificing so much to provide me with the resources and emotional support to explore my passions—from math and academics to music, dance, and acting. When research felt overwhelming, I often coped through the other creative outlets you gave me. To my father, thank you for encouraging me to always be courageous even in math; "just start writing" is always at the back of my mind on days when I struggle to sit and think about some problem, 20 years later. To my mother, thank you for recognizing that I could and wanted to do the "serious" things, as well as the "non-serious" things, to be a whole human being. To my brother, thank you for having always had your unique ability to listen to what I had to say, and to bring the best out of every moment. And to my grandparents and great-grandparents, who sacrificed so much so that we could live happier, freer lives, thank you.

DEDICATION

To my parents

Chapter 1
INTRODUCTION

Many systems of scientific interest are comprised of interacting and interdependent units. Examples include social systems, in which individuals are connected through social relationships, marketplaces composed of sellers and buyers interacting through demand and supply, transportation systems, biological systems, and geological systems. Observations and processes from such settings cannot, in general, be treated as independent. Moreover, understanding the effects of interventions on units in these systems is complicated by the possibility that interactions between units modify these effects. This dissertation studies statistical inference, testing, estimation, and experimental design in such dependent-data settings, with a particular focus on network interference.

A large portion of the study of statistics is dedicated to settings where data are independent, or where dependence is sufficiently restricted that methods developed for independent data can still be used. In many applications, however, this is not a reasonable approximation. An individual's behavior may depend on the behavior of their friends, demand in one part of a marketplace may affect outcomes elsewhere in the market, or information introduced at one location in a network may diffuse throughout the population. In experimental settings, there is an additional complication: the treatment assigned to one unit may affect the outcomes of other units. Thus, the dependence of interest may arise both from the underlying system and from the intervention itself.

Throughout this dissertation, networks provide a useful representation of this structure. I use the term network to refer to a finite graph, possibly with weighted edges and possibly augmented by node-level covariates. The nodes represent the units of study, such as individuals in a social system or buyers and sellers in a marketplace. An edge represents a relationship between two units that is relevant to the scientific problem, while a weighted edge allows the strength of this relationship to vary. Covariates may include demographic or

behavioral characteristics, capacity or budget constraints in a marketplace, or other observed characteristics of the units.

The network provides information about the structure relating the observations. A central perspective of this dissertation is that such structure is useful not only for describing why standard statistical assumptions fail, but also for determining how statistical procedures should be constructed. In particular, information about dependence can be used to determine when asymptotic inference remains valid, which observations are most informative for detecting interference, how observations should be used when estimating treatment effects, and how treatments should be assigned before an experiment is conducted.

1.1 Dependence and Network Interference

Suppose that $X_1, \ldots, X_n$ are the observations available to a researcher. Even if there is no treatment or intervention, these observations may be correlated. This affects, among other things, the applicability of central limit theorems and the construction of inferential procedures. Classical approaches typically make assumptions on how this dependence is structured. For example, observations may be assumed to become approximately independent as their temporal or spatial distance grows, or the dependence may be summarized by a sparse dependency graph. Such assumptions can be appropriate in particular applications, but they can be restrictive when dependencies propagate through a network.

A problem building on this is causal interference in the presence of network interference. Let $z \in \{0,1\}^n$ denote a treatment assignment and $Y_i(z)$ the potential outcome of unit i. Under the usual no-interference component of the Stable Unit Treatment Value Assumption (SUTVA) [Cox, 1958, Rubin, 1978, Manski, 1990], we require that

$$Y_i(z) = Y_i(z') \qquad \text{whenever} \qquad z_i = z'_i. \tag{1.1.1}$$

That is, holding the treatment of unit i fixed, changing the assignments of other units does not change the potential outcome of unit i. Under network interference this is no longer true. If i is connected to j, then changing z_j may change Y_i. More generally, an intervention may propagate through multiple paths in the network, so that the treatment assignment of

a distant unit can also have some effect on i.

For example, consider an experiment on a social network where the treatment is access to a new product or feature. If a treated user begins interacting with the product more frequently, this may change the behavior of the user's friends, including friends assigned to control. Consequently, an outcome observed under a mixture of treated and control users need not correspond to the outcome that the same user would have under global treatment or global control. Similarly, when the goal is simply to detect whether such interference exists, the treatments received by a unit's neighbors determine how much information about interference is contained in that unit's outcome.

A common simplification is to summarize the assignments of other units through an exposure mapping Aronow and Samii [2017]. For instance, if W is a weighted adjacency matrix and D the corresponding degree matrix, a simple exposure mapping is

$$e_i = \left(D^{-1}Wz\right)_i, \tag{1.1.2}$$

the weighted fraction of unit i's neighbors assigned to treatment. The potential outcome may then be modeled as a function $Y_i(z_i, e_i)$. This reduction is useful, but it creates another modeling problem: the correct exposure mapping is typically not known. Whether 20%, 80%, or all of one's neighbors must receive treatment before interference becomes negligible is a scientific question that often cannot be answered from domain knowledge alone.

These issues affect different statistical tasks in different ways. For inference, the relevant question is how much dependence can be allowed before asymptotic normality fails. For testing interference, the relevant question is which units experience enough treatment exposure variation among their neighbors to make interference detectable. For estimation, the experimenter must trade off protection against interference-induced bias against the variance created by using only units satisfying restrictive exposure conditions. Finally, at the design stage, the treatment assignments themselves can be chosen to control interference under this exposure, but designs that perform well for interference may perform poorly in the presence of other forms of structure, such as homophily. At this design stage also, treatment assignments can be chosen to control misspecified networks, and therefore misspecified

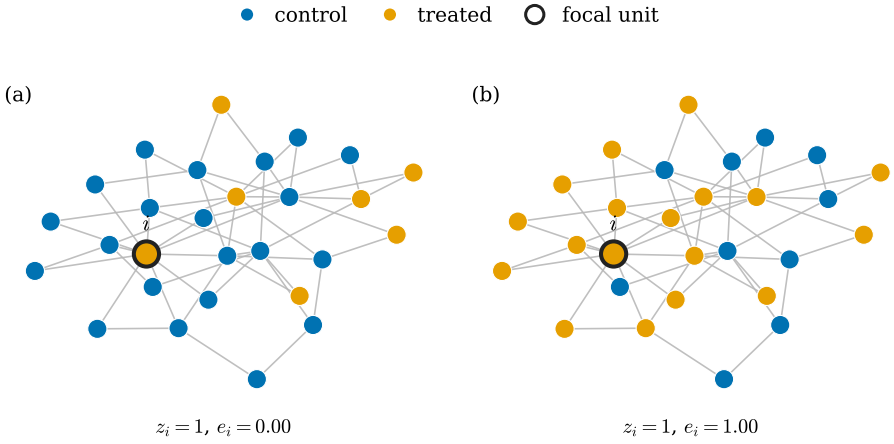


Figure 1.1.1: Illustration of network interference. The focal unit i has the same individual treatment assignment in the two panels, but the assignments of its neighbors change. Under interference, the resulting change in exposure can change the potential outcome of unit i even though z_i is held fixed.

exposures.

The chapters of this dissertation study these problems at different points of the statistical pipeline.

1.2 Leveraging Structure Under Dependence

A recurring idea throughout this dissertation is that dependence need not only be treated as an obstacle. If something is known about the structure of the dependence, that information can be used when constructing the statistical procedure.

This perspective first appears in Chapter 2 at the level of asymptotic inference. Rather than requiring observations outside of a prescribed neighborhood to be independent, I study conditions under which the total dependence can be decomposed into relatively stronger and weaker components, across many different dependency settings. In particular, the chapter introduces *affinity sets*, which collect the observations having relatively stronger dependence with a given observation. Correlation outside of these sets is allowed to be nonzero. What matters is whether the different covariance contributions remain sufficiently controlled

relative to the variance of the overall sum. Thus, even when every pair of observations is correlated at every finite n, asymptotic normality may still follow.

In the later chapters the network becomes something that can be directly acted upon. In Chapter 3, the problem is not merely to account for the dependence induced by the network, but to select units according to the graph so that a test has more power. In Chapter 4, the network determines the exposure distribution, and the observed exposure–response relationship is used to decide how restrictive an exposure condition should be. In Chapter 5, the optimization is moved one step earlier: instead of adapting the analysis to an observed treatment assignment, the assignment distribution itself is chosen using the network.

This progression also reflects different amounts of information about the underlying problem. If little is known beyond the fact that observations are dependent, one may seek high-level conditions under which standard inference remains possible. If the network is known, it can additionally be used to construct a more informative test. If outcomes and exposures have already been observed, the data can inform the choice of estimator. Finally, if the experiment has not yet been conducted, assumptions about the interference and potential outcome structure can be incorporated directly into the experimental design.

Throughout, the models used to improve statistical performance need not be treated as exactly true. In Chapter 3, a model is used to understand and optimize power, while the conditional randomization test itself remains valid under model misspecification. In Chapter 4, the linear exposure model is used as a first-order approximation for estimating the bias rather than as an assumption that the true potential outcome function is linear. In Chapter 5, uncertainty not explained by the interference or homophily models is represented explicitly through heterogeneous variation, and the experimental design trades off exploitation of network structure against additional randomization. Thus, across the chapters, while the network structure is leveraged as much as is scientifically reasonable, the statistical consequences of the structure that remains unknown is accounted for.

1.3 Statistical Questions Considered in this Dissertation

The dissertation specifically studies the following four questions that arise naturally in dependent data and in the experimentation pipeline.

The first question is a foundational inferential one:

> *When can sums or averages of dependent observations still be treated as asymptotically normal?*

This is the focus of Chapter 2. Many estimators can be reduced, exactly or asymptotically, to sums of random variables. Central limit theorems therefore provide the basis for a large range of inferential procedures. Existing central limit theorems for dependent observations often require dependence to have a particular form, such as mixing in a metric space or a sparse dependency graph. These conditions can be difficult to justify directly from the scientific problem. Chapter 2 instead develops sufficient conditions stated through covariance quantities.

The second question arises once an experiment has been conducted:

> *Is there evidence that interference is present at all?*

This question is important because designing and analyzing experiments under interference can be substantially more complicated than under no interference. For example, cluster randomization may reduce interference-induced bias, but it also reduces the effective sample size, i.e., the number of independent randomization units and can increase variance. Before using more complicated procedures, it is therefore useful to know whether there is evidence of interference. Chapter 3 studies conditional randomization tests for this purpose and, in particular, how the selection of focal units affects statistical power.

The third question also assumes that the experiment has already been conducted and that interference must be addressed:

> *How should treatment effects be estimated when the correct exposure mapping is unknown?*

One possible approach is to impose a conservative exposure condition, such as requiring all of a unit's neighbors to share the unit's treatment assignment. This can reduce bias, but the resulting exposure probabilities may be extremely small. A more permissive condition uses more observations but incurs more interference-induced bias. Chapter 4 treats the exposure

threshold as a statistical decision and selects it by estimating this bias–variance trade-off from the observed data.

Finally, before an experiment is run, there is a corresponding design question:

> *Can the treatment assignments themselves be selected so that treatment effects are easier to estimate?*

This is the focus of Chapter 5. A natural approach under interference is cluster randomization, because assigning neighboring units together reduces the number of edges connecting treatment and control. However, networks often exhibit homophily: connected units tend to have similar characteristics and potentially similar outcomes or treatment effects. Cluster randomization can then place entire groups of similar units in only one treatment arm, which can substantially increase estimation error as the treatment and control samples are no longer representative of the population. At the same time, outcomes may exhibit heterogeneous variation that is not well described by the observed network. Chapter 5 develops a framework that treats these considerations jointly.

These four questions move from general inference to testing, estimation, and finally design. The later chapters therefore progressively use more of the available structure of the problem.

1.4 Contributions

1.4.1 General Covariance-Based Conditions for Central Limit Theorems with Dependent Data

Chapter 2 develops a general central limit theorem for dependent triangular arrays of random vectors. The central goal is to replace model-specific dependence assumptions with a small number of covariance-based sufficient conditions that are easier to interpret in terms of the scientific problem.

The framework associates each observation with an affinity set containing the variables with which it may have relatively high dependence. Importantly, observations outside the affinity set are not assumed to be independent. Instead, three conditions control the

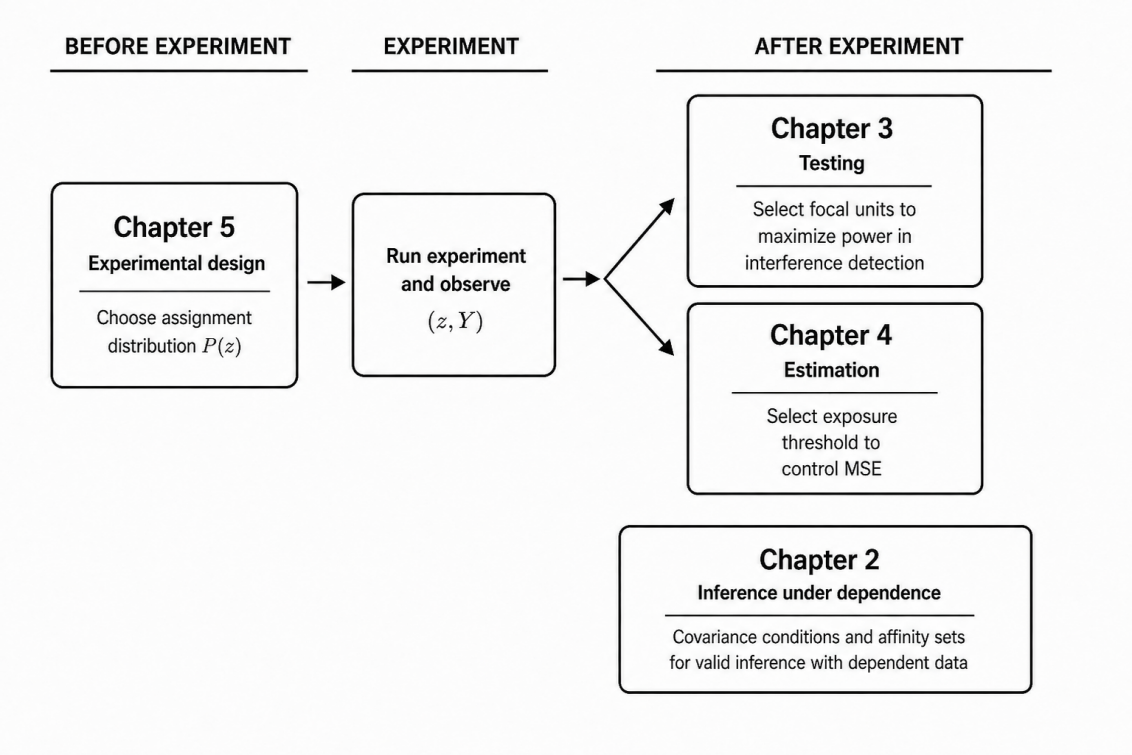


Figure 1.3.1: Overview of the dissertation. Chapter 2 provides a general inferential foundation for dependent data. Chapters 3–5 use the observed network at different points of the experimental pipeline: Chapter 5 designs treatment assignments before the experiment, while Chapters 3 and 4 use the resulting data to test for interference and estimate treatment effects.

covariance within affinity sets, covariance across affinity sets, and the aggregate influence of observations outside of the affinity sets. When these quantities are sufficiently small relative to the variance contributed by the affinity sets, the appropriately normalized sum converges to a multivariate normal distribution.

This differs from dependency graph approaches, where observations outside a specified neighborhood are typically required to be independent. It also differs from mixing approaches, where dependence must decay with respect to a prescribed temporal or spatial metric. In the framework of Chapter 2, there may be nonzero dependence between every pair of observations. The main requirement is instead that the stronger dependence be sufficiently concentrated relative to the total variance.

I show how the conditions encompass several commonly studied dependence settings that do not themselves necessarily nest each other. These include M-dependent processes, mixing random fields, non-mixing autoregressive processes, and dependency graphs. The chapter then applies the framework to settings where the ability to allow weak global dependence is particularly useful. These include treatment effects with spillovers, covariance estimation based on socio-economic rather than purely spatial distance, and diffusion processes.

For example, in a sub-critical diffusion process on a connected network, infection originating from one seed may in principle reach every node. Thus, infection indicators for every pair of units can be correlated. Nevertheless, most of the dependence may be generated by relatively small local outbreaks. The affinity sets can collect the regions in which infection from a given seed occurs with appreciable probability, while allowing weaker dependence outside of these regions. This gives a central limit theorem even though a literal dependency graph representation would be complete.

The chapter provides a statistical foundation for the remainder of the dissertation in another sense as well. In Chapter 3, the asymptotic analysis of the conditional randomization test involves sums of dependent exposure-based quantities. The affinity-set framework provides a direct route for establishing their asymptotic normality. Thus, the asymptotic analysis of dependent data in Chapter 2 reappears when analyzing the power of procedures specifically designed for network interference.

1.4.2 An Optimization Framework for Focal Node Selection in Interference Detection

Chapter 3 turns to the problem of testing for interference. I consider conditional randomization tests of the sharp null that the outcome of a focal unit does not depend on the treatment assignments of other units. Following the conditional testing framework of Aronow [2012] and Athey et al. [2018], the treatment assignments of a selected set of *focal units* are fixed, while assignments of the remaining *auxiliary units* are rerandomized. The test statistic evaluated on the focal units under the observed assignment is then compared to its rerandomization distribution.

The validity of this procedure follows from the randomization mechanism under the

appropriate conditioning. Its power, however, can change substantially with the focal set. If a focal unit has little variation in its exposure under rerandomization, its outcome provides little information about interference. At the opposite extreme, imposing overly strong separation requirements between focal units can drastically reduce the number of focal observations available to the test.

I develop an asymptotic power characterization under a linear neighborhood-interference model and show that power is governed by a residualized exposure-variation quantity. In the notation of Chapter 3, this quantity is

$$C_M(F) = \mathbb{E}\left[\tilde{z}_R^{\mathsf{T}} B_{RF} M_F B_{FR} \tilde{z}_R\right], \tag{1.4.1}$$

where the focal set is indexed by F, and the auxiliary set is indexed by $R = V \setminus F$. This provides an explicit objective for focal-unit selection rather than relying only on heuristic graph constructions.

The characterization also explains several procedures previously proposed in the literature. For example, maximizing the number of focal–auxiliary edges in Athey et al. [2018] appears as an approximation to the power objective if M_F is taken to be I. Independent-set constructions in Athey et al. [2018] can also be understood through their effect on the covariance and residualization of focal exposures, leading to centered null rerandomization distributions. I show, however, that making focal units too separated can be counterproductive: on cycle graphs, a less restrictive net can have greater power because the gain in the number of focal units outweighs the reduction in exposure dependence obtained by additional separation.

The exact power objective leads naturally to optimization formulations based on mixed-integer quadratic programs, second-order cone programs and semidefinite program relaxations together with simpler graph-based approximations. In synthetic experiments, the resulting optimization-based focal selections lead to substantial power improvements over random focal selection and common graph heuristics.

An important feature of this chapter is the separation between validity and power. The conditional randomization test remains valid even when the linear potential outcome model used for the power analysis is misspecified. The model determines which focal units are

expected to be informative, while the randomization argument protects the Type-I error. This is one example of the broader robustness perspective used throughout the dissertation.

1.4.3 Data-Adaptive Exposure Thresholds under Network Interference

Chapter 4 focuses on treatment-effect estimation after an experiment has been conducted. A common approach to estimation under interference is to specify an exposure mapping and use a Horvitz–Thompson estimator based on units satisfying the corresponding exposure condition [Aronow and Samii, 2017, Ugander et al., 2013]. For example, under an h-fractional neighborhood exposure condition, a treated unit is considered treatment-exposed when at least an h-fraction of its neighbors are also treated.

The choice of h creates a direct bias–variance trade-off. When h is close to one, the observed outcomes better approximate the potential outcomes under global treatment and control, reducing interference-induced bias. At the same time, the probability that a unit satisfies such a restrictive exposure condition can become very small, particularly for high-degree units, and the inverse-probability weights in the Horvitz–Thompson estimator can become large. When h is smaller, more observations are used and the variance decreases, but units whose neighbors largely receive the opposite treatment are included, increasing bias.

Without sufficient domain expertise, there may be no principled way to specify this threshold in advance. Chapter 4 therefore treats the threshold as a data-adaptive tuning parameter and chooses

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{\mathrm{MSE}}(\widehat{\tau}_h). \tag{1.4.2}$$

The method, which I call AdaThresh, estimates the bias of each candidate threshold by first fitting a linear approximation of outcomes as a function of treatment and exposure. The coefficient on exposure provides a *bias slope*, measuring the average sensitivity of outcomes to changes in exposure. This is then combined with a Horvitz–Thompson-style calculation that accounts for the empirical exposure distribution. The variance is estimated using a design-based variance estimator.

The linear regression is not intended to assert that the potential outcome function is globally linear. Rather, it gives a first-order approximation to the rate at which outcomes

change as exposure moves away from the global-treatment or global-control boundary. This distinction allows the method to remain useful when the true exposure–response function is nonlinear. The chapter also studies local linear versions of the bias approximation.

The theoretical results characterize the probability that the estimated threshold agrees with the MSE-optimal threshold under the best linear approximation. The quality of this selection depends naturally on the sample size, the separation between candidate MSEs, and the dependence induced by the graph. Under unit-level Bernoulli randomization this dependence is controlled by the maximum node degree; an analogous cross-cluster degree appears under cluster randomization. The results are also extended to quantify approximation error when the true exposure function deviates from its best linear approximation.

Experiments on simulated graphs, a village network, and an Amazon product-similarity network illustrate that the adaptive procedure interpolates between permissive and conservative fixed exposure thresholds. The chapter therefore addresses uncertainty about interference at the estimation stage: rather than requiring the experimenter to know the exact threshold defining effective treatment, the observed data are used to determine an appropriate point on the bias–variance trade-off.

1.4.4 Optimal Design under Interference, Homophily, and Robustness Trade-offs

Chapter 5 moves from optimizing the estimator after observing an experiment to optimizing the experimental design itself. The goal is estimation of the global average treatment effect,

$$\tau = \frac{1}{n} \langle \mathbf{1}, Y(\mathbf{1}) - Y(\mathbf{0}) \rangle . \tag{1.4.3}$$

A common strategy under network interference is cluster randomization. Assigning strongly connected units to treatment together reduces the number of edges connecting units with conflicting treatment assignments, thereby reducing interference-induced bias. However, this strategy can work against another common feature of networks: homophily.

Under homophily, connected units tend to be similar along socioeconomic, demographic, behavioral, or other dimensions that may also determine potential outcomes. If treatment is randomized at the cluster level, an entire group of similar units may be assigned to one

treatment arm. The resulting treated and control groups may then be poorly balanced. Thus, interference encourages assigning neighboring units together, while homophily can encourage spreading treatment assignments throughout the graph.

I formalize this tension through a potential outcome model that separately represents homophily, network interference, and heterogeneous variation. Let L denote the graph Laplacian. Homophilous components of the baseline outcomes and treatment effects are constrained through quadratic forms involving L, so that neighboring units have similar potential-outcome components. Interference is constrained through the pseudoinverse $L^\dagger$, and heterogeneous variation collects variation not explained by these structural components.

Writing $x = 2z - 1$ and $X = \mathbb{E}[xx^\mathsf{T}]$ for the second-moment matrix of the design, the resulting worst-case MSE bound has the schematic form

$$\mathrm{MSE} \lesssim \frac{1}{n^2}\left\{\gamma \operatorname{Tr}(LX) + \eta \operatorname{Tr}(L^\dagger X) + \kappa\|X\|_q + \Delta \operatorname{Tr}(\mathbf{1}\mathbf{1}^\mathsf{T} X)\right\}. \tag{1.4.4}$$

The individual terms have direct design interpretations. The interference term $\operatorname{Tr}(LX)$ is related to a graph cut and encourages neighboring units to receive positively correlated assignments. The homophily term $\operatorname{Tr}(L^\dagger X)$ has an effective-resistance interpretation and encourages assignments that are spread throughout the network. The Schatten-norm term penalizes excessive correlation in the assignment covariance and produces robustness to heterogeneous variation not captured by the structural model. The final term prevents degenerate assignments.

Consequently, familiar experimental designs arise as limiting cases of the same optimization problem. When interference dominates, the design moves toward clustered assignments. When homophily dominates, treatment is dispersed through the network. When heterogeneous variation dominates, additional randomization becomes valuable and, for the relevant Schatten norms, the design approaches unit-level Bernoulli randomization. This characterization also provides a generalization to randomized saturation designs. At the covariance level, randomized saturation designs interpolate between cluster and unit-level Bernoulli randomization by varying the within-cluster assignment correlation. The covariance optimization in Chapter 5 contains this path as a restricted class, while allowing more general correlation

structures. At the same time, I discuss the limitation of a second-moment formulation: different randomized saturation designs may share the same marginal treatment probabilities and covariance matrices while differing in higher moments.

I propose two complementary computational approaches for constructing designs. The first formulates a semidefinite relaxation over X, followed by Gaussian rounding to generate binary treatment assignments. The approximation analysis draws on ideas underlying the Goemans–Williamson approximation for MAXCUT. The second adapts the Gram–Schmidt Walk, a vector-balancing algorithm, by interpreting graph-induced quantities as covariates that should be balanced against robustness. In particular, this uncovers a natural interpretation of interference and homophily by viewing eigenvalue-scaled Laplacian eigenvector embeddings as unit-level covariates, connecting the covariate-balancing and interference literatures. The SDP provides finer direct control over the covariance objective, while the adapted Gram–Schmidt Walk is more computationally attractive on larger problems.

Thus, Chapter 5 moves the same basic principle used in Chapters 3 and 4 to the design stage. In Chapter 3, network structure determines which units should be used to maximize power. In Chapter 4, it determines how the observations should be weighted and thresholded to minimize MSE. In Chapter 5, it determines the distribution from which the treatment assignment itself should be drawn.

1.5 Network Structure and Optimization

Several ideas recur throughout the dissertation.

First, the chapters trade off the statistical information gained from dependence against the difficulties that dependence creates. In Chapter 2, stronger dependence increases the covariance of the observations and can prevent normal approximation unless it is appropriately controlled. In Chapter 3, dependence among focal exposures can similarly reduce useful variation, while separating the focal units too aggressively can leave too little data for the test. In Chapter 4, using observations with more heterogeneous neighborhood assignments increases the sample available for estimation but also increases interference-induced bias. In Chapter 5, assigning neighboring units together protects against interference but may reduce representativeness in a homophilous network.

Second, optimization appears at many stages of the analysis. Chapter 3 optimizes the subset of observations used for a test. Chapter 4 optimizes a parameter of the estimator after observing the data. Chapter 5 optimizes the experimental design before observing the outcomes. In each case, the objective is statistical—power or mean squared error—while the graph determines the structure of the optimization problem.

Third, the dissertation distinguishes between structural assumptions that improve performance and guarantees that remain valid under weaker assumptions. This is clearest in Chapter 3, where model-based focal selection affects power but not the randomization validity of the test. Chapter 4 uses a simple exposure–response approximation while explicitly studying deviations from that approximation. Chapter 5 places the unknown component of the potential outcomes directly into a robustness class and increases randomization when less of the outcome structure can be safely attributed to the observed graph.

Finally, the role of the graph itself changes across the dissertation. In the general central limit theorem, a graph is only one possible source of dependence, and the key objects are covariance-based affinity sets. In the remaining chapters, the graph is observed and can be used directly in the procedure. Nevertheless, even there, the graph should not necessarily be treated as an exact description of the underlying system. Exposure mappings may be misspecified, outcomes may vary for reasons not explained by the network, and the observed network itself may be noisy. The methods developed here aim to exploit the information contained in network structure without requiring that it explain all aspects of the data.

Taken together, the chapters study a sequence of statistical decisions for interacting populations: when dependence still permits conventional inference, how interference can be detected, how its effects can be estimated when its precise form is uncertain, and how experiments can be designed to account for it before the data are collected. The central perspective throughout is that the structure producing dependence can itself be useful statistical information. Rather than treating a network only as a violation of independence, this dissertation uses it to guide inference, testing, estimation, and experimental design.

1.6 Organization of the Dissertation

The remainder of the dissertation is organized as follows. Chapter 2 develops general covariance-based sufficient conditions for central limit theorems under dependent triangular arrays and studies several dependence models and applications. Chapter 3 develops an optimization framework for focal-unit selection in conditional randomization tests of network interference, deriving asymptotic power characterizations and optimization-based selection procedures. Chapter 4 introduces AdaThresh, a data-adaptive procedure for selecting exposure thresholds in Horvitz–Thompson estimation under network interference. Chapter 5 develops an experimental-design framework that jointly accounts for network interference, homophily, and robustness to heterogeneous variation, together with semidefinite-programming and vector-balancing approaches for constructing treatment assignments.

Chapter 2

GENERAL COVARIANCE-BASED CONDITIONS FOR CENTRAL LIMIT THEOREMS WITH DEPENDENT TRIANGULAR ARRAYS

This chapter is based on the paper "A General Central Limit Theorem for Dependent Data," coauthored with Arun G. Chandrasekhar[‡,§], Matthew O. Jackson[‡,⋆], and Tyler H. McCormick[◇,¶]. It is currently under review.

[‡] Stanford, Department of Economics, USA

[§] J-PAL, USA

[⋆] Santa Fe Institute, USA

[◇] University of Washington, Department of Statistics, USA

[¶] University of Washington, Department of Sociology, USA

We present a general central limit theorem with simple, easy-to-check covariance-based sufficient conditions for triangular arrays of random vectors when all variables could be interdependent. The result is constructed from Stein's method, but the conditions are distinct from those of related work. Existing approaches require checking bespoke conditions that in many contexts are difficult to verify scientifically and either (i) impose rigid structure, often difficult to interpret and lacking microfoundations (e.g., strong mixing in random fields) or (ii) allow more flexibility but limit the extent of correlations (e.g., sparsity restrictions in dependency graphs). Our approach, in contrast, permits researchers to work with high-level but intuitive conditions based on overall correlation. We show that these covariance conditions nest standard assumptions studied in the literature such as M-dependence, mixing random fields, non-mixing autoregressive processes, and dependency graphs, which themselves need not imply each other. We apply our result to practical settings previously not covered in the literature, such as treatment effects with spillovers in more settings than previously admitted, covariance matrices, and processes with global dependencies such as epidemic spread and information diffusion.

Keywords: Central limit theorem, dependent data, Stein's method

2.1 Introduction

In many contexts researchers use an interdependent set of random vectors to develop estimators and need to establish whether the estimators are asymptotically normal. In existing results, dependency is modeled in idiosyncratic ways, with perhaps unintuitive if not unappealing conditions to describe assumptions on correlation. Further, in many settings—such as treatment effects with spillovers, epidemic spread, information diffusion, general equilibrium effects, and so on,—the correlation is non-zero for any finite set of observations across all random vectors, which is not allowed in the literature.

We present simple covariance-based sufficient conditions for a central limit theorem to be applied to a triangular array of dependent random vectors. We use Stein's method [Stein, 1986] to derive three high-level, easy-to-interpret conditions. Stein's method is widely used in theoretical arguments (in fact, a special case of the argument here first appeared in [Chandrasekhar and Jackson, 2024]) and our goal is not to advance a new technique for proving limit theorems. Instead, we present a new approach to proving asymptotic normality that is extremely general, nesting many existing approaches. More importantly, it consists of easily-interpretable conditions that are also easy to check. Researchers can check our conditions by thinking directly about correlations using calculations that we demonstrate are straightforward in several consequential examples. This feature makes the result accessible to a wider range of researchers without having to derive bespoke limit theorems in every setting. Our approach also eases restrictions required by existing methods, again doing so in a way that is not idiosyncratically imposed by a specific setting.

We follow a literature that uses Stein's method to prove asymptotic normality. Our goal is not to derive a new proof, but rather to provide much more general conditions that allow for some amount of dependence between all observations but at the same time are minimal in terms of parametric/shape restrictions, making them flexible and easy to check.

To review, Stein's method observes that

$$\mathbb{E}\left[Y f(Y)\right] = \mathbb{E}\left[f'(Y)\right]$$

for all continuously differentiable functions if and only if Y has a standard normal distribution. So, when considering normalized sums that take the role of Y, it is enough to show that this equality holds for all such functions asymptotically. Rinott and Rotar [2000] and Ross [2011], among others, provide a detailed view of the method.Proving normality using Stein's method typically amounts to checking dependence conditions. Bolthausen [1982], for example, establishes how the probability of a joint set of events differs from independence decays in temporal distance. So, as the mixing coefficients decay fast enough in distance, the proof proceeds by checking that the Stein argument follows under the mixing structure. Distance in time can be generalized to space and, further, to random fields. The existing literature is organized around a number of such conditions that apply in particular data settings.

A peculiar consequence of this organization, however, is that these conditions generally lack a scientific basis for the assumed dependence structure. For example, spatial standard errors are often used—, e.g., in Froot [1989], Conley [1999], Driscoll and Kraay [1998]—when performing Z-estimation (or GMM). However, in actual applications, for instance agricultural shocks such as rainfall or pests or soil, it is not clear that they follow a specific form of interdependence satisfying ϕ-mixing with a certain decay rate as invoked in Conley [1999] (cross-sectionally) or Driscoll and Kraay [1998] (temporally and cross-sectionally). Surely shocks correlate over space, but it is hard to argue empirically that their correlation matches the required mixing conditions.

To take a different example, some models of social network formation orient themselves by embedding individuals in a random field to deliver central limit theorems. Distance in the metric space is, then, inversely proportional to the likelihood of a connection. Given the structure of the field, researchers can toggle dependence as a function of distance, reducing

the size of the covariance sum. However, the structure of the field has implications for consequential properties of the graph, such as clustering patterns [Hoff et al., 2002, Lubold et al., 2023], that then often fail to match the empirical observations. As in the previous example, the researcher probably has some intuition as to whether graph properties (e.g., clustering) are likely or not, but assessing the reasonableness of specific mixing random field assumptions is all but impossible.

As an alternative to embedding variables in a metric space and toggling dependency by distance, a literature on dependency graphs emerged [Baldi and Rinott, 1989, Goldstein and Rinott, 1996, Ross, 2011] in part to be more flexible on the correlation structure. There, the cost is extreme sparsity: observations have indices in a graph, where those that are not edge-adjacent are independent. This provides a different strategy to apply Stein's method, by creating for each observation a dependency neighborhood. Sufficient sparsity in the graph structure allows a central limit theorem to apply, despite not forcing a time- or space-like structure. Examples include Ross [2011], Goldstein and Rinott [1996] and a more general treatment in Chen and Shao [2004].

Both embedding indices in a metric space or using a more unstructured dependency structure are similar in the sense that they constrain the total amount of correlation between the n random variables. In principle there are $\binom{n}{2}$-order components to this sum, but, via mixing conditions or sparsity conditions, this sum is assumed to be of order n. In many settings—such as treatment effects with spillovers, epidemic spread, information diffusion, or general equilibrium effects—we see dependency that are neither random fields nor are there any conditional independencies. The correlation is non-zero for any finite set of observations across all random vectors of interest. Our approach allows for this general dependence.

In the remainder of this section, we provide a brief overview of our approach, which we formalize in the next section. We consider a triangular array of n random vectors $X_{1:n}^n \in \mathbb{R}^p$, which are neither necessarily independent nor identically distributed. We study conditions under which their appropriately normalized sample mean is asymptotically

normally distributed. In principle, at any n, all X_i^n and X_j^n can be correlated. The proof follows the well-known Stein's method, though we develop and apply specific bounds for our purpose. To apply Stein's method we first associate each random variable with a set of other random variables with which it could have a higher level of correlation, keeping in mind that it could in principle have some non-zero correlation with all variables.

We call these *affinity sets*, denoted $\mathcal{A}_{(i,d)}^n$, to capture the other random variables with which i may have high correlations in the d-th dimension. We use the term *affinity set* rather than "dependency neighborhood" to emphasize the possibility of high and low non-zero covariance structures with arbitrary groupings that satisfy summation conditions. We provide sufficient conditions for asymptotic normality in terms of the total amount of covariance within an affinity set and the total amount of covariance across affinity sets. As long as, in the limit, the amount of interdependence in the overall mean comes from the covariances within affinity sets, asymptotic normality follows.

This yields substantially weaker conditions than in the previous literature. In Arratia et al. [1989], the authors present Chen's method [Chen, 1975] for Poisson approximation rather than normality, which has a similar approach to ours in collecting random variables into dependencies. While this results in nice finite sample bounds, these bounds consist of three almost separate pieces, making it less friendly to understanding these bounds together with growing sums of covariance of the samples. In fact, all five of the examples studied in Arratia et al. [1989] are limited to cases where at least one of these pieces is identically zero in the cases where Chen's method succeeds. Many empirically relevant examples do not have such zeros, and so our approach substantially expands the relevant applications.

To preview our conditions, presented formally below, let us take $p = 1$ so X_i^n is scalar, and let Z_i^n be centered, $Z_i^n := X_i^n - \mathbb{E}(X_i^n)$. Let

$$\Omega_n := \sum_{i=1}^{n} \sum_{j \in \mathcal{A}_i^n} \mathrm{Cov}\left(Z_i, Z_j\right),$$

denote the total covariance within the affinity sets. Let also $\mathbf{Z}_{-i} := \sum_{(j) \notin \mathcal{A}_i^n} Z_j$, the sum of

the random variables outside of i's affinity set. Informally, our conditions are the following.

1. Within affinity set covariance control:

$$\sum_i \sum_{j,k \in \mathcal{A}_i^n} \mathbb{E}(Z_j Z_k \cdot |Z_i|) \text{ is small relative to } (\Omega_n)^{3/2}.$$

   The average covariance between two random variables in an affinity set, when weighted by the realized magnitude of the reference variable, and the covariance between the averages given the magnitude of the reference variable, is small relative to the comparably adjusted covariance between the reference variables and their affinity sets.

2. Cross affinity set covariance control:

$$\sum_{i,j} \sum_{k \in \mathcal{A}_i^n, l \in \mathcal{A}_j^n} \text{Cov}(Z_i Z_k, Z_j Z_l) \text{ is small relative to } (\Omega_n)^2.$$

   The average covariance across members of two different affinity sets (weighted by the reference variables) is sufficiently small compared to the squared covariance within affinity sets.

3. Outside affinity set covariance control:

$$\sum_i \mathbb{E}(|\mathbf{Z}_{-i} \mathbb{E}(Z_i \mid \mathbf{Z}_{-i})|) \text{ is small relative to } \Omega_n.$$

   The average absolute conditional covariance outside of affinity sets is small compared to the covariance within affinity sets.

These conditions are general and easy to check, as we show. They also further simplify in a number of cases such as with positive correlations (as in diffusion models and auto-regressive processes) and with binary variables.

The advantages of this approach, which we see as most salient for empirical scientists, are several-fold. First, we do not require a sparse dependency structure at any n. That is,

there can be non-zero correlation between any pair X_i, X_j. Much of the dependency graph literature leverages an independence structure in constructing their bounds and, therefore, the bounds we build are different.

Second, because of this possibility of non-zero covariance across all random vectors, we organize our bounds through covariance conditions. We are reminded of a discussion in Chen [1975] in the context of Poisson approximation. Covariance conditions are easy-to-interpret and check, and from an applied perspective often easier to justify from a microfoundation.

Third, our result is for random vectors, and while the application of the Cramér-Wold device is simple in our setting—by the nature of how indexing works—it is useful to have and instructive for a practitioner.

Fourth, our setup nests many of the previous literature's examples, most of which do not nest each other. We illustrate the utility of our central limit theorem through several distinct applications. We begin with an example of M-dependence in a stochastic process, noting that this implies many other types of mixing, such as α-, ϕ-, or ρ- mixing [Bradley, 2005]. We then move to random fields, where we show an example with α- and ϕ- dependence. These mixing approaches require constructing idiosyncratic notions of dependence based on the underlying probability distribution that happen to imply bounds on the covariance function (see, for example, Rio [1993] or Rio [2017]), which, as noted above, is not derived from, and may not match, micro-economic foundations or scientific principles. Our covariance-based arguments are compact and direct, placing restrictions on the covariance explicitly and, thus, in a matter that is salient in the scientific context. They are also general rather than based on a specific model or type of dependence. We also show that our framework is applicable outside the context of mixing, giving examples with non-mixing autoregressive processes, dependency graphs, among other examples.

Fifth, we show that our generalizations permit a wider and more practical set of analyses that were otherwise ruled out or limited in the literature. This includes treatment effects with spillover models, covariance matrices, and things like epidemic and diffusion models.

Specifically, we extend the treatment effects with spillovers analysis, as in Aronow and Samii [2017], to allow every individual's exposure to treatment to possibly be increasing in every other node's treatment assignment, and nonetheless, the relevant estimator is still asymptotically normally distributed. This case, which is ubiquitous in practice: e.g., in diffusion, epidemic, and financial flow models in principle a shock anywhere could theoretical impact any other node's outcomes (albeit with potentially small odds). But this is assumed away in applied work because conventional central limit theorems do not cover such a case. We also show how a researcher can model covariance matrices without forcing a random field structure as in Conley [1999] or Driscoll and Kraay [1998]. This allows applied researchers to proceed with greater generality, and permits structure across units that do not have a natural ordering such as race, ethnicity, caste, and occupation.

The next two examples concern diffusion. First, we look at a sub-critical epidemic process with a number of periods longer than the graph's diameter. So, whether an individual is infected is correlated with the infection status of any other individual (assuming a connected, unweighted graph). Again, this practical situation is excluded by the previous central limit theorems in the literature. Second, we look at diffusion in stochastic block models to show that our conditions characterize when asymptotic normality holds and when it does not.

The remainder of the paper is organized as follows, Section 2.2 proves our main result. Section 2.3 shows that the conditions we provide nest several commonly used characterizations of dependence: M-dependence, non-mixing autoregressive processes, random fields, and dependency graphs. Section 2.4 discusses the process for checking our conditions in several applications (peer effects models, socio-demographic distances, sub-critical diffusion models, stochastic block models, and irregularly observed spatial processes), while also yielding new results. Section 2.5 provides a discussion.

2.2 The Theorem

We consider a triangular array of n random variables $X^n_{1:n} \in \mathbb{R}^p$, with entries $X^n_{i,d}$ and $d \in \{1, \ldots, p\}$, each of which has finite variance (possibly varying with n). We let $Z^n_i \in \mathbb{R}^p$

denote the corresponding de-meaned variables, $Z_i^n = X_i^n - \mathrm{E}\left[X_i^n\right]$. The sum, $S^n \in \mathbb{R}^p$, is given by $S^n := \sum_{i=1}^n Z_i^n$. We suppress the dependency on n for clarity; writing $X_{i,d}$ unless otherwise needed.

2.2.1 Affinity Sets

Each real-valued random variable $X_{i,d}$ has an *affinity set*, denoted $\mathcal{A}_{(i,d)}^n$, which can depend on n. We require $(i,d) \in \mathcal{A}_{(i,d)}^n$. Heuristically, $\mathcal{A}_{(i,d)}^n$ includes the indices j, d' for which the covariance between $X_{j,d'}$ and $X_{i,d}$ is relatively high in magnitude, but not those for which the covariance is low.

There is no independence requirement at any n and, in fact, our sufficient conditions for the central limit theorem bound the total sums of covariances within and across affinity sets. The precise construction of affinity sets is flexible, as long as these bounds on the respective total sums are respected.

In the multivariate case, we aggregate affinity sets across dimensions

$$\widetilde{\mathcal{A}}_i^n = \bigcup_{d=1}^{p} \mathcal{A}_{i,d}^n.$$

2.2.2 The Central Limit Theorem

Let Ω_n be a $p \times p$ matrix that captures the bulk of the covariance across observations and dimensions, obtained by summing the covariances between each observation and the observations in its unified affinity set:

$$\Omega_n := \sum_{i=1}^{n} \sum_{j \in \widetilde{\mathcal{A}}_i^n} \mathrm{Cov}(Z_i, Z_j).$$

This is distinct from the total variance-covariance matrix Σ_n, whose (d, d') entry is

$$\Sigma_{n,dd'} := \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(Z_{i,d}, Z_{j,d'}),$$

and which therefore also includes covariance terms with $j \notin \widetilde{\mathcal{A}}_i^n$. Also define

$$\mathbf{Z}_{-i} := \sum_{j \notin \widetilde{\mathcal{A}}_i^n} Z_j,$$

the vector sum of the random variables outside of i's unified affinity set.

In what follows, we maintain the assumption that $\liminf_{n\to\infty} \lambda_{\min}(\Omega_n) > 0$, where $\lambda_{\min}(\Omega_n)$ is the smallest eigenvalue of Ω_n. For simplicity, we suppose Ω_n is symmetric, which follows if the affinity sets are symmetric, in the sense that j being in the affinity set of i implies that i is in the affinity set for j. If Ω_n is not symmetric, the following theorem will still hold, substituting $\tilde{\Omega}_n = (\Omega_n + \Omega_n^\top)/2$ for Ω. We also presume that for all i, n, the individual moment ratios are bounded: there exists $K < \infty$ such that for every non-zero $u \in \mathbb{R}^p$, $\mathbb{E}[|u^\top Z_i|^3]/(u^\top \operatorname{Cov}(Z_i)u)^{3/2} \leq K$.

Our first assumption is that the total mass of the variance-covariance is not driven by the covariance between members of a given affinity set neither of which are the reference random variables themselves. That is, given reference variable $X_{i,d}$, the covariance of some $X_{j,d'}$ and $X_{k,d''}$ where both are in the reference variable's affinity set is relatively small in total across all such triples of variables compared to the variance coming from the reference variable and its affinity sets.

Assumption 2.2.1 (Bound on total weighted-covariance within affinity sets)**.**

$$\sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\| \cdot \left\| \sum_{j \in \tilde{\mathcal{A}}_i} Z_j \right\|^2 \right] = o\left(\lambda_{\min}(\Omega_n)^{3/2} \right).$$

The second assumption is that the total mass of the variance-covariance is not driven by random variables across affinity sets relative to two distinct reference variables. That is, given two random variables $X_{i,d}$ and $X_{j,d'}$, the aggregate amount of weighted covariance between two other random variables—each within one of the reference variables' affinity sets—is small compared to the (squared) variance coming from the reference variable and

its affinity sets.

Assumption 2.2.2 (Bound on total weighted-covariance across affinity sets)**.**

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in\tilde{\mathcal{A}}_i}\sum_{l\in\tilde{\mathcal{A}}_j}\sum_{d_1,d_2,d_3,d_4=1}^{p}\left|\mathrm{Cov}(Z_{i,d_1}Z_{k,d_2},Z_{j,d_3}Z_{l,d_4})\right| = o\left(\lambda_{\min}(\Omega_n)^2\right).$$

The third assumption is that the total mass of variance-covariance is not driven by reference random variables and the variables outside of their affinity sets, again compared to the variance coming from the reference variable and its affinity sets.

Assumption 2.2.3 (Bound on total weighted-covariance from outside of affinity sets)**.**

$$\sum_{i=1}^{n}\mathbb{E}\left[\|Z_i\|\,\left\|\mathbb{E}[Z_i\mid\mathbf{Z}_{-i}]\right\|\right] = o(\lambda_{\min}(\Omega_n))\,.$$

These three assumptions imply a central limit theorem.

Theorem 2.2.4. *Suppose that, for all i, n and for every non-zero $u\in\mathbb{R}^p$, there exists K such that $\mathbb{E}[|u^\top Z_i|^3]/(u^\top\mathrm{Cov}(Z_i)u)^{3/2}\le K$. If Assumptions 2.2.1, 2.2.2, and 2.2.3 are satisfied and $\liminf_{n\to\infty}\lambda_{\min}(\Omega_n)>0$,*

$$\Omega_n^{-1/2}S^n\xrightarrow{d}\mathcal{N}(0,I_p)$$

Here $\lambda_{\min}(X)$ is used to denote the minimal eigenvalue of the matrix X. The proof is provided in the Appendix. The argument follows by applying the Cramér-Wold device to the arguments following Stein's method, as Chandrasekhar and Jackson [2024] argued for the univariate case. Since the Cramér-Wold device requires for all $c\in\mathbb{R}^p$ fixed in n that the c-weighted sum satisfies a central limit theorem [Biscio et al., 2018]—that is, $(c'\Omega_n c)^{-1/2}c'S^n\rightsquigarrow\mathcal{N}(0,1)$—we can consider a problem of np random variables with affinity sets.

An important special case is where the affinity sets are the variables themselves. In

that case, the conditions simplify to a total bound on the overall sum of covariances across variables (the univariate case is in Chandrasekhar and Jackson [2024]). It nests many cases in practice, and we provide an illustration in our second application.

Corollary 2.2.5 (Singleton Affinity Sets)**.** *Suppose the unified affinity sets are singletons such that* $\tilde{\mathcal{A}}_i = \{i\}$ *for all observations* i*, that* $\liminf_{n\to\infty} \lambda_{\min}(\Omega_n) > 0$*, and*

$$\sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\|^3\right] = o\left(\lambda_{\min}(\Omega_n)^{3/2}\right).$$

If $\mathbb{E}\left[Z_{i,d_1}\mathbb{E}[Z_{i,d_2} \mid Z_{-i}]\right] \geq 0$ *for every* i*, and*

1. $\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{d_1,d_2,d_3,d_4=1}^{p} |\mathrm{Cov}(Z_{i,d_1}Z_{i,d_2}, Z_{j,d_3}Z_{j,d_4})| = o\left(\lambda_{\min}(\Omega_n)^2\right)$

2. $\displaystyle\sum_{i=1}^{n}\sum_{j\neq i}\sum_{d_1,d_2=1}^{p} |\,\mathrm{Cov}(Z_{i,d_1}, Z_{j,d_2})| = o\left(\lambda_{\min}(\Omega_n)\right)$

Then $\Omega_n^{-1/2} S^n \xrightarrow{d} \mathcal{N}(0, I_p)$*. If the non-negativity condition does not hold, Condition 2 can be replaced by the outside-affinity influence bound:*

$$\sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\| \cdot \|\mathbb{E}\left[Z_i \mid Z_{-i}\right]\|\right] = o\left(\lambda_{\min}(\Omega_n)\right).$$

Also, it is useful to note that, for instance, if $p = 1$ and the X_i's are Bernoulli random variables with $\mathbb{E}[X_i] \to 0$ (uniformly), then condition (ii) implies condition (i) [Chandrasekhar and Jackson, 2024].

Corollary 2.2.6 (Total dependence)**.** *Suppose the unified affinity sets encompass the entire sample, such that* $\tilde{\mathcal{A}}_i = \{1, \ldots, n\}$ *for all observations* $i \in \{1, \ldots, n\}$*. Assume that for all* i, n *and for every non-zero* $u \in \mathbb{R}^p$*, there exists a constant* $K > 0$ *such that* $\mathbb{E}[|u^\top Z_i|^3]/(u^\top cov(Z_i)u)^{3/2} \leq K$*.*

Furthermore, suppose $\liminf_{n\to\infty} \lambda_{\min}(\Omega_n) > 0$*, and that the following global bounds hold:*

$$\sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\| \cdot \|S^n\|^2\right] = o\left(\lambda_{\min}(\Omega_n)^{3/2}\right) \tag{2.2.1}$$

$$\sum_{i,j,k,l=1}^{n} \sum_{d_1,d_2,d_3,d_4=1}^{p} |cov(Z_{i,d_1} Z_{k,d_2}, Z_{j,d_3} Z_{l,d_4})| = o\left(\lambda_{\min}(\Omega_n)^2\right) \tag{2.2.2}$$

Then it follows that $\Omega_n^{-1/2} S^n \xrightarrow{d} \mathcal{N}(0, I_p)$ *as* $n \to \infty$.

2.3 Models of Dependence

We first present four applications from the literature that prove asymptotic normality: (i) M-dependence, (ii) non-mixing autoregressive processes, (iii) mixing random fields, and (iv) dependency graphs. We provide a sketch of the core assumptions made in the relevant papers and show how these assumptions imply our covariance assumptions. We maintain consistent use of notation defined in the previous sections. The remaining notation, however, is kept consistent only within each subsection to facilitate comparison with the original work.

2.3.1 M-dependence

2.3.1.1 Environment

We consider Theorem 2.1 of Romano and Wolf [2000]. In this application there are real-valued time series data, so $p = 1$ (and we drop the index d) and Ω_n is a scalar. Under Romano and Wolf's setup, $Z_{n,i}$ and $Z_{n,j}$ are independent if $|i - j| > M$. Here, $\{Z_{n,i}\}$ are mean zero random variables. For convenience of the reader, we include the assumptions made in their paper: Suppose $Z_{n,1}, Z_{n,2}, ..., Z_{n,r}$ is an M-dependent sequence of random variables for some $\delta > 0$ and $-1 \leq \gamma < 1$,

1. $\mathbb{E}|Z_{n,i}|^{2+\delta} \leq \Delta_n$ for all i
2. $\text{Var}\left(\sum_{i=a}^{a+k-1} Z_{n,i}\right) k^{-1-\gamma} \leq K_n$ for all a and $k \geq M$
3. $\text{Var}\left(\sum_{i=1}^{r} Z_{n,i}\right) r^{-1} M^{-\gamma} \geq L_n$

4. $K_n = O(L_n)$

5. $\Delta_n = O\left({L_n}^{1+\delta/2}\right)$

6. $M^{1+(1-\gamma)(1+2/\delta)} = o(r)$

2.3.1.2 Application of Theorem 2.2.4

We consider the M-ball, $\mathcal{A}_i^n = \{j : |j - i| \leq M\}$. We drop the subscript n in $Z_{n,i}$ for convenience. In this case, $\text{Cov}(Z_i, Z_j) = 0$ by independence for all j with $|i - j| > M$, so Assumption 2.2.3 is satisfied. By their Assumptions (3) and (6), we have that $\frac{\sum_{j\in\mathcal{A}_i}\text{Cov}(Z_i,Z_j)}{M} \geq L_n(r/M)^{1+2/\delta}$. This, together with their Assumptions (2) and (4), results in $\sum_i \sum_{j\in\mathcal{A}_i} \text{Cov}(Z_i, Z_j) = \Omega(nM)$ (with $\Omega(\cdot)$ being the big Omega notation). Under bounded third and fourth moments, we check the remaining assumptions. Assumption 2.2.1 is easily verified:

$$\sum_{i;j,k\in\mathcal{A}_i^n} \mathbb{E}[|Z_i|Z_jZ_k] = O(M^2 \sum_i \mathbb{E}[|Z_i|^3]) = o\left(n^{3/2}M^{3/2}\right) = o\left(\Omega_n^{3/2}\right),$$

following their Assumption 6. Our Assumption 2.2.2 is satisfied similarly following their Assumption 6:

$$\sum_{i,j;k\in\mathcal{A}_i^n,l\in\mathcal{A}_j^n} \text{Cov}(Z_iZ_k, Z_jZ_l) = O(\sum_{\substack{i,j:|i-j|\leq M,\\ k:|k-i|\leq M,\\ l:|l-i|\leq 2M}} \text{Var}(Z_i^2)) = O(n \cdot M^3 \cdot \text{Var}(Z_i^2)) = o(n^2M^2) = o(\Omega_n^2).$$

This is due to the fact that if they are not within that distance, then automatically it is impossible for the Z_k and Z_l to induce any correlation as well.

Following the hierarchy established in Bradley [2007], M-dependence implies several of these commonly used forms of mixing, such as ρ-mixing [Bradley, 2005]. It also implies ϕ-mixing and α-mixing in the time series context. We also given a example using α- and ϕ- mixing in the context of random fields below. Characterizing mixing in the context of

random fields requires imposing restrictions on the dependence between σ-algebras as the number of points in those sets increases, see Jenish and Prucha [2009] or Bradley [2005]. Our conditions generate bounded fourth moment requirements, which is not necessarily a condition invoked in every analysis of M-dependent processes in the literature, which sometimes have slightly lower moment requirements.

2.3.2 Andrews' Non-Mixing Autoregressive Processes

2.3.2.1 Environment

This application is from Andrews [1984], which allows interdependence in a time series, but one that does not satisfy strong (α-) mixing, in order to clarify the distinction between dependence and mixing. Again, $p = 1$. We take $X_t = \sum_{l=0}^{\infty} \rho^l \epsilon_{t-l}$, where $\rho \in (0, 1/2]$. The ϵ_t's are independent, and come from a Bernoulli distribution with success probability q. We define Z_t as normalized, mean zero, X_t. Assume, without loss of generality, that $s > t$, so

$$\operatorname{Cov}(Z_t, Z_s) = \operatorname{Cov}(\sum_{l=0}^{\infty} \rho^l \epsilon_{t-l}, \sum_{k=0}^{\infty} \rho^k \epsilon_{s-k}) = \rho^M \sum_{l=0}^{\infty} \rho^{2l} \operatorname{Var}(\epsilon_{t-l}) = \frac{\rho^M}{1-\rho^2} \cdot q(1-q)$$

where $M = s - t$. For a constant C depending only on ρ, we show asymptotic normality of Z_t: $\frac{1}{\sqrt{Cnq(1-q)}} \sum_t Z_t \rightsquigarrow \mathcal{N}(0, 1)$.

2.3.2.2 Application of Theorem 2.2.4

We begin by verifying our conditions for truncated versions of our random variables and then show that the full result applies. To define what we mean by "truncation", for each $i \in [n]$, let

$$X_i^{(D)} = \sum_{j=0}^{D} \rho^j \epsilon_{i-j}$$

for some D. Let also $S_n^{(D)} := \sum_{i=1}^n Z_i^{(D)} \equiv \sum_{i=1}^n (X_i^{(D)} - \mathbb{E}[X_i^{(D)}])$, and $\Omega_n^{(D)} = \sum_i \sum_{j \in \mathcal{A}_i^n} \operatorname{Cov}(Z_i^{(D)}, Z_j^{(D)})$. We then define the affinity sets to be $\mathcal{A}_i = \{k : |i - k| \leq D\}$.

Note that, since $\rho > 0$, $\Omega_n^{(D)} = \Omega(nD)$, where $\Omega(\cdot)$ is the big Omega notation.

We begin by verifying Assumption 2.2.1:

$$\sum_{i;j,k\in\mathcal{A}_i^n} \mathbb{E}[|Z_i^{(D)}|Z_j^{(D)}Z_k^{(D)}] = O(D^2 \sum_i \mathbb{E}[|Z_i|^3]) = o\left(n^{3/2}D^{3/2}\right) = o\left(\Omega_n^{(D)^{3/2}}\right),$$

Next, we verify Assumption 2.2.2:

$$\begin{aligned}\sum_{i,j;k\in\mathcal{A}_i^n, l\in\mathcal{A}_j^n} \mathrm{Cov}(Z_i^{(D)}Z_k^{(D)}, Z_j^{(D)}Z_l^{(D)}) &= O(\sum_{\substack{i,j:|i-j|\le D,\\ k:|k-i|\le D,\\ l:|l-i|\le 2D}} \mathrm{Var}({Z_i^{(D)}}^2)) \\ &= O(n\cdot D^3\cdot \mathrm{Var}({Z_i^{(D)}}^2) = o(n^2D^2) = o({\Omega_n^{(D)}}^2).\end{aligned}$$

To verify Assumption 2.2.3 is trivial; we have $\sum_i \mathbb{E}[|\mathbf{Z}_{-i}^{(D)}\mathbb{E}[Z_i^{(D)}|\mathbf{Z}_{-i}^{(D)}]|] = \sum_i \mathbb{E}[|\mathbf{Z}_{-i}^{(D)}\mathbb{E}[Z_i^{(D)}]|] = 0$. Therefore, we have that the truncated variables $\left(\Omega_n^{(D)}\right)^{-1/2} S_n^{(D)} \rightsquigarrow \mathcal{N}(0,1)$, and we would like to show the result in full generality, $\Omega_n^{-1/2}S_n \rightsquigarrow \mathcal{N}(0,1)$. To do that, we begin by writing

$$\frac{S_n}{\Omega_n^{1/2}} = \frac{S_n^{(D)}}{{\Omega_n^{(D)}}^{1/2}} \frac{{\Omega_n^{(D)}}^{1/2}}{\Omega_n^{1/2}} + \frac{S_n^{(\bar{D})}}{{\Omega_n}^{1/2}} \tag{2.3.1}$$

where $S_n^{(\bar{D})} = \sum_i^n (Z_i - Z_i^{(D)}) \equiv \sum_i^n Z_i^{(\bar{D})}$. We will show that $\lim_{D\to\infty}\lim_{n\to\infty} \frac{\Omega_n^{(\bar{D})}}{\Omega_n} = 0$, where $\Omega_n^{(\bar{D})} = \sum_{i,j=1} \mathrm{Cov}(Z_i^{(\bar{D})}, Z_j^{(\bar{D})})$, and use this to control the second term in the right-hand-side of the decomposition given in expression (2.3.1) above, via Chebyshev's inequality. Then, we will also show that $\lim_{D\to\infty}\lim_{n\to\infty} \frac{{\Omega_n^{(D)}}^{1/2}}{\Omega_n^{1/2}} = 1$, which will result in the weak convergence of the first term on the right-hand-side of (2.3.1) to $\mathcal{N}(0,1)$. We start with

$$\begin{aligned}\lim_{D\to\infty}\lim_{n\to\infty} \frac{\Omega_n^{(\bar{D})}}{\Omega_n} &= \lim_{D\to\infty}\lim_{n\to\infty} \frac{\sum_{i,j} \mathrm{Cov}(Z_i^{(\bar{D})}, Z_j^{(\bar{D})})}{\sum_{i,j}\mathrm{Cov}(Z_i, Z_j)} \\ &= \lim_{D\to\infty}\lim_{n\to\infty} \frac{\sum_{i,j} \rho^{2D+2+|i-j|}\sum_{k=0}^{\infty}\rho^{2k}q(1-q)}{\sum_{i,j}\rho^{|i-j|}\sum_{k=0}^{\infty}\rho^{2k}q(1-q)}\end{aligned}$$

$$= \lim_{D\to\infty} \lim_{n\to\infty} \frac{\sum_{i,j} \rho^{2D+2+|i-j|}}{\sum_{i,j} \rho^{|i-j|}} = 0.$$

Together with Chebyshev's inequality, we have that

$$\lim_{D\to\infty} \lim_{n\to\infty} \mathbb{P}\left(\left| \frac{S_n^{(\bar{D})}}{{\Omega_n}^{1/2}} \right| > \epsilon \right) \leq \lim_{D\to\infty} \lim_{n\to\infty} \frac{\Omega_n^{(\bar{D})}}{\epsilon^2 \Omega_n} = 0, \tag{2.3.2}$$

and this gives us convergence in probability of the second term in 2.3.1 to zero.

Next, we show that $\lim_{D\to\infty} \lim_{n\to\infty} \frac{{\Omega_n^{(D)}}^{1/2}}{\Omega_n^{1/2}} = 1$. We begin by considering

$$\begin{aligned}
\frac{\Omega_n^{(D)}}{\Omega_n} &= \frac{\sum_{i,j} \mathrm{Cov}(Z_i^{(D)}, Z_j^{(D)})}{\sum_{i,j} \mathrm{Cov}(Z_i, Z_j)} = \frac{\sum_{i,j} \rho^{|i-j|} \sum_{k=0}^{D-|i-j|} \rho^{2k} q(1-q)}{\sum_{i,j} \rho^{|i-j|} \sum_{k=0}^{\infty} \rho^{2k} q(1-q)} \\
&= 1 - \frac{\sum_{i,j} \rho^{|i-j|} \sum_{k=D-|i-j|+1\vee 0}^{\infty} \rho^{2k}}{\sum_{i,j} \rho^{|i-j|} \sum_{k=0}^{\infty} \rho^{2k}} \\
&= 1 - \frac{\sum_{i,j:|i-j|\leq D} \rho^{2D-|i-j|+2}}{\sum_{i,j} \rho^{|i-j|}} - \frac{\sum_{i,j:|i-j|>D} \rho^{|i-j|}}{\sum_{i,j} \rho^{|i-j|}}
\end{aligned} \tag{2.3.3}$$

We now consider the second term in the above expression, (2.3.3), $\sum_{i,j:|i-j|\leq D} \rho^{2D-|i-j|+2} / \sum_{i,j} \rho^{|i-j|}$. The lower bound $\sum_{i,j} \rho^{|i-j|} \sum_{k=0}^{\infty} \rho^{2k} = \Omega(n)$ (i.e. there exists some constant c such that $\sum_{i,j} \rho^{|i-j|} \geq cn$). For a fixed $\epsilon > 0$, there exists a $D > 0$ such that $\rho^d < \epsilon$ for all $d \geq D$. Therefore, we have that $\sum_{i,j} \rho^{|i-j|} \sum_{k=D-|i-j|+1}^{\infty} \rho^{2k} = O(n\epsilon)$. Hence,

$$\frac{\sum_{i,j:|i-j|\leq D} \rho^{2D-|i-j|+2}}{\sum_{i,j} \rho^{|i-j|}} = O(\epsilon).$$

Next, we consider the third term in 2.3.3

$$\frac{\sum_{i,j:|i-j|>D} \rho^{|i-j|}}{\sum_{i,j} \rho^{|i-j|}} = \frac{\sum_{i,j} \rho^{D+1} \rho^{j}}{\sum_{i,j} \rho^{|i-j|}}.$$

Once again, we have the lower bound $\sum_{i,j} \rho^{|i-j|} = \Omega(n)$ (i.e. there exists some constant c such that $\sum_{i,j} \rho^{|i-j|} \geq cn$), and for a fixed $\epsilon > 0$, there exists a $D > 0$ such that $\rho^d < \epsilon$ for

all $d \geq D$. Therefore, we have that $\sum_{i,j} \rho^{D+1}\rho^j = O(n\epsilon)$, and hence,

$$\frac{\sum_{i,j:|i-j|>D} \rho^{|i-j|}}{\sum_{i,j} \rho^{|i-j|}} = O(\epsilon).$$

Putting all of this together gives us $\lim_{D\to\infty} \lim_{n\to\infty} \frac{{\Omega_n^{(D)}}^{1/2}}{{\Omega_n}^{1/2}} = 1$. Finally, we have that the first term in 2.3.1, $\frac{S_n^{(D)}}{{\Omega_n^{(D)}}^{1/2}} \frac{{\Omega_n^{(D)}}^{1/2}}{\Omega_n^{1/2}} \rightsquigarrow N(0,1)$ since $\lim_{D\to\infty} \lim_{n\to\infty} \frac{{\Omega_n^{(D)}}^{1/2}}{\Omega_n^{1/2}} = 1$. Therefore, together with 2.3.2, we have that $\frac{S_n}{\Omega_n^{1/2}} \rightsquigarrow \mathcal{N}(0,1)$, and the proof is complete.

2.3.3 Random Fields

2.3.3.1 Environment

This example nests many time series and spatial mixing models. Take the setting of Jenish and Prucha [2009], Theorem 1. Their setting has either ϕ- or α- mixing in random fields, allowing for non-stationarity and asymptotically unbounded second moments. Throughout this section, we make frequent use of the notation introduced in their paper. For the reader's convenience, we briefly reintroduce the relevant definitions below and refer to Jenish and Prucha [2009] for further details and context. Let $\alpha_{k,l,n}(r) := \sup(\alpha(\sigma_n(X_{i,n}; i \in U), \sigma_n(X_{i,n}; i \in V)), |U| \leq k, |V| \leq l, \rho(U,V) \geq r)$, $\bar{\alpha}_{k,l}(r) = \sup_n \alpha_{k,l,n}(r)$, $\alpha_{inv}(u) := \max\{m \geq 0 : \bar{\alpha}_{1,1}(m) > u\}$, $Q_X(u) = F_X^{-1}(1-u)$, $Q^{(k)}_{|Z_i|} := Q_{|Z_i|\mathbf{1}(|Z_i|>k)}$, and $\bar{Q}^{(k)}_{|Z_i|} := Q_{|Z_i/c_{i,n}|\mathbf{1}(|Z_i/c_{i,n}|>k)}$.

We restate the relevant assumptions made in their paper for completeness:

- Assumption 2: $\lim_{k\to\infty} \sup_n \sup_{i\in D_n} \mathbb{E}[|Z_{i,n}/c_{i,n}|^2 \mathbf{1}\{|Z_{i,n}/c_{i,n}| > k\}] = 0$ for $c_{i,n} \in \mathbb{R}^+$

- Assumption 3: The following conditions must be satisfied by the α-mixing coefficients:

  1. $\lim_{k\to\infty} \sup_n \sup_{i\in D_n} \int_0^1 \alpha_{inv}^d(u) \left(Q_{|Z_{i,n}/c_{i,n}|\mathbf{1}\{Z_{i,n}/c_{i,n}>k\}}\right)^2 du = 0$
  2. $\sum_{m=1}^{\infty} m^{d-1} \sup_n \alpha_{k,l,n}(m) < \infty$ for $k + l \leq 4$ where

$$\alpha_{k,l,n}(r) = \sup(\alpha_n(U,V), |U| \leq k, |V| \leq l, \rho(U,V) \geq r)$$

3. $\sup_n \alpha_{1,\infty,n}(m) = O\left(m^{-d-\delta}\right)$

- Assumption 4: The following conditions must be satisfied by the ϕ-mixing coefficients:

  1. $\sum_{m=1}^{\infty} m^{d-1}\bar{\phi}_{1,1}^{1/2}(m) < \infty$
  2. $\sum_{m=1}^{\infty} m^{d-1}\bar{\phi}_{k,l}(m) < \infty$ for $k+1 \leq 4$
  3. $\bar{\phi}_{1,\infty}(m) = \mathcal{O}(m^{-d-\epsilon})$ for some $\epsilon > 0$

- Assumption 5: $\liminf_{n\to\infty} |D_n|^{-1} {M_n}^{-2}\sigma_n^2 > 0$, where $M_n = \max_i c_{i,n}$ for $c_{i,n}$ in Assumption 3: (1), and $\sigma_n^2 = \mathrm{Var}(S_n)$

2.3.3.2 Application of Theorem 2.2.4

In the following, we assume that the $Z_{i,n}$s have bounded second (i.e. $c_{i,n} = 1$ in the above), third and fourth, moments. Here, for any $\epsilon > 0$, we take $\mathcal{A}_i^n = \{j : \rho(i,j) \leq K^i(\epsilon_n)\}$, where K^i is a non-increasing function. That is, we pick $K^i(\epsilon_n)$ to be large enough, and this can be decided by understanding the cumulative distribution function of the random variables.

By Assumption 3 (and Proposition B.10), and Lemma B.1 (a) in Jenish and Prucha [2009], we know that for any $k \neq i$ such that $\rho(i,k) \geq K^i(\epsilon_n)$, we have that for some constant C,

$$C\int_0^{\bar{\alpha}_{1,1}(\rho(i,k))} Q_{i,n}^2(u)du \leq \epsilon_n. \tag{2.3.4}$$

In particular, we note that we can pick $K^i(\epsilon_n)$ by observing that ϵ_n above satisfies $O(1/\rho(i,k)^r)$ for $r > 1$.

Taking $\epsilon_n = \omega(\frac{1}{\rho_0^d n^\gamma})$ allows control of the size of the affinity sets. Indeed, via a packing number calculation, we see that while this allows $K^i(\epsilon_n)$ to grow with n, it grows more slowly than n. Specifically, taking $\epsilon_n = \omega\left(\frac{1}{\rho_0^d n^\gamma}\right)$ for any $0 < \gamma < 1$, and since $\delta > 0$ together with

their Assumption 3, we have,

$$\left(\frac{K^i(\epsilon_n)}{\rho_0}\right)^d = \left(\frac{(1/\epsilon_n)^{-(d+\delta)}}{\rho_0}\right)^d < \left(\frac{1}{\epsilon_n \rho_0^d}\right) = o(n)$$

We write $K := \max_i K^i(\epsilon_n)$. By their Assumption 5, $\Omega_n = \Omega(nK)$, where $\Omega(\cdot)$ is the big Omega notation.

Now, we verify that our key conditions are satisfied in this setting. First, we check Assumption 2.2.1:

$$\sum_{i;j,k\in\mathcal{A}_i^n} \mathbb{E}[|Z_{i,n}|Z_{j,n}Z_{k,n}] = O(K^2 \sum_i \mathbb{E}[|Z_{i,n}|^3]) = O(K^2 n) = o(n^{3/2}K^{3/2}) = o\left(\Omega_n^{3/2}\right)$$

since $\Omega_n^{3/2} = (\sum_i \sum_{j\in\mathcal{A}_i^n} \mathbb{E}[Z_i Z_j])^{3/2}$. The second inequality holds by the arithmetic-geometric mean inequality. The remaining argument relies on rearranging the summations and using the growth rate of K, i.e. $K = o(n)$ in the third equality, as defined above.

Next, we check that Assumption 2.2.2 is satisfied using similar arguments and assuming bounded fourth moments:

$$\sum_{i,j;k\in\mathcal{A}_i^n, l\in\mathcal{A}_j^n} \mathrm{Cov}(Z_{i,n}Z_{k,n}, Z_{j,n}Z_{l,n}) = O(\sum_{\substack{i,j:|i-j|\le K,\\ k:|k-i|\le K,\\ l:|l-i|\le 2K}} \mathbb{E}(Z_{i,n}^4)) = O(n \cdot K^3 \cdot \mathbb{E}(Z_{i,n}^4)) = o(\Omega_n^2)$$

where the first equality comes from the covariance terms within the affinity sets dominating those with outside the affinity sets (see the following verification of Assumption 2.2.3).

Finally, we check Assumption 2.2.3. Together with 2.3.4, taking $\epsilon_n = \omega(\frac{1}{\rho_0^d n^\gamma})$ where $\gamma = 1 - \beta$ for arbitrarily small $\beta > 0$, we have, $\frac{n}{K^{1+1/(d+\delta)}} = o(1)$. Thus, defining the random variable $\xi_{i,-i}$ such that $\xi_{i,-i} = 1$ if $\mathbb{E}(Z_{i,n}|\mathbf{Z}_{-i,n}) > 0$, and $\xi_{i,-i} = -1$ otherwise just as in Rio [2013], we have

$$\sum_i \mathbb{E}(|\mathbb{E}(Z_{i,n}\mathbf{Z}_{-i,n}|\mathbf{Z}_{-i,n})|) = \sum_i \sum_{j \notin \mathcal{A}_i^n} \text{Cov}(\xi_{i,-i} Z_{j,n}, Z_{i,n})$$
$$\leq 2 \sum_i \sum_{j \notin \mathcal{A}_i^n} \int_0^{\bar{\alpha}(\rho(i,j))} Q_{i,n}(u) Q_{j,n}(u) du$$
$$= O(n(n-K)\epsilon_n) = O(n(n-K)K^{-1/(2+\delta)}) = o(nK) = o\left(\Omega_n\right),$$

by Rio's inequality [Rio, 1993].

2.3.4 Dependency Graphs and Chen and Shao [2004]

Next, we consider dependency graphs. There is an undirected, unweighted graph G with dependency neighborhoods $N_i := \{j : G_{ij} = 1\}$ such that Z_i is independent of all Z_j for $j \notin N_i$ [Baldi and Rinott, 1989, Chen and Shao, 2004, Ross, 2011]. Let $\mathcal{A}_i^n = \{i\} \cup \{j : G_{ij} = 1\}$. Denote the maximum cardinality of these to be D. In Ross [2011] (see Theorem 3.6), together with a bounded fourth moment assumption, we see that the conditions there imply the conditions here. Indeed, we see that

$$\sum_{i;j,k \in \mathcal{A}_i^n} \mathbb{E}\left[|Z_i| Z_j Z_k\right] \leq D^2 \sum_{i=1}^n \mathbb{E}\left[Z_i^3\right],$$

and for $\Omega_n^{-1/2} S_n \rightsquigarrow N(0,1)$, we need $D^2 \sum_{i=1}^n \mathbb{E}\left[Z_i^3\right] = o\left(\Omega_n^{3/2}\right)$ in Ross [2011] (Theorem 3.6), and hence Assumption 2.2.1 is satisfied. Similarly,

$$\sum_{i,i';j \in \mathcal{A}(i,n), j' \in \mathcal{A}_i^n} \text{Cov}\left(Z_i Z_j, Z_{i'} Z_{j'}\right) = O(D^3 \sum_{i=1}^n \mathbb{E}[Z_i^4]),$$

and for $\Omega_n^{-1/2} S_n \rightsquigarrow N(0,1)$, we need $D^3 \sum_{i=1}^n \mathbb{E}[Z_i^4] = o\left((\Omega_n)^2\right)$ in Ross [2011] (Theorem 3.6), so Assumption 2.2.2 is satisfied. Assumption 2.2.3 holds by definition of the dependency neighborhoods.

Now, we consider Chen and Shao [2004]. In particular, we consider their weakest

assumption, LD1: Given index set $\mathcal{I}$, for any $i \in \mathcal{I}$, there exists an $A_i \in \mathcal{I}$ such that X_i and $X_{A_i^C}$ are independent. The affinity sets can be defined by the complement of the independence sets. So $\mathcal{A}_i^n = \{j : Z_j \text{ is not independent of } Z_i\}$, which is similar to the dependency graphs setting. The goals of their paper are different. They develop finite-sample Berry-Esseen bounds, with bounded p-th moments, where $2 < p \leq 3$. This is different from our approach. In our paper, we focus on covariance conditions in the asymptotics, and collect relatively more dependent sets along the triangular array.

2.4 Applications

2.4.1 Peer Effects Models

2.4.1.1 Environment

We now turn to an example of treatment effects with spillovers. Consider a setting in which units in a network are assigned treatment status $T_i \in \{0, 1\}$, as in Aronow and Samii [2017]. The network is a graph, $\mathcal{G}$, consisting of individuals (nodes) and connections (edges). For now, we consider the case in which treatment assignments are independent across nodes. However, there are spillovers in treatment effects determined by the topology of the network where treatment status within one's network neighborhood may influence one's own outcome - for instance whether a friend is vaccinated affects a person's chance of being exposed to a disease.

Rather than being arbitrary, Aronow and Samii [2017] consider an exposure function f that takes on one of K finite values; i.e., $f(T_i; T_{1:n}, \mathcal{G}) \in \{d_1, \ldots, d_K\}$. An estimand of the average causal effect is of the form

$$\tau(d_k, d_l) = \frac{1}{n} \sum_i y_i(d_k) - \frac{1}{n} \sum y_i(d_l)$$

where d_k and d_l are the induced exposures under the treatment vectors. The Horvitz and

Thompson estimator from Horvitz and Thompson [1952] provides:

$$\hat{\tau}_{HT}(d_k, d_l) = \frac{1}{n}\sum_i \mathbf{1}\{D_i = d_k\}\frac{y_i(d_k)}{\pi_i(d_k)} - \frac{1}{n}\sum \mathbf{1}\{D_i = d_l\}\frac{y_i(d_l)}{\pi_i(d_l)}$$

where $\pi_i(d_k)$ is the probability that that node i receives exposure d_k over all treatments.

The challenge is that the treatment effects are not independent across subjects. Let $N_{i,:}$ denote a dummy vector of whether j is in i's neighborhood: let $N_{ij} = 1$ when $G_{ij} = 1$ with the convention $N_{ii} = 0$. Aronow and Samii [2017] consider an empirical study with $K = 4$ that are: (i) only i is treated in their neighborhood ($d_1 = T_i \cdot 1\{T'_{1:n}N_{i,:} = 0\}$), (ii) at least one is treated in i's neighborhood and i is treated ($d_2 = T_i \cdot 1\{T'_{1:n}N_{i,:} > 0\}$), (iii) i is not treated but some member of the neighborhood is ($d_3 = (1 - T_i) \cdot 1\{T'_{1:n}N_{i,:} > 0\}$), and (iv) neither i nor any neighbor is treated ($d_4 = (1 - T_i) \cdot \prod_{j:\ N_{ij}=1}(1 - T_j)$). We show that our result allows for a more generalized setting.

2.4.1.2 Application of Theorem 2.2.4

To obtain consistency and asymptotic normality Aronow and Samii [2017] assume a covariance restriction of local dependence (Condition 5) and apply Chen and Shao [2004] to prove the result. Namely, their restriction is that there is a dependency graph H (with entries $H_{ij} \in \{0, 1\}$) with degree that is uniformly bounded by some integer m independent of n. That is, $\sum_j H_{ij} \leq m$ for every i. This setting is much more restrictive than our conditions, especially as there can exist indirect correlation in choices as effects propagate or diffuse through the graph. We can work with larger real exposure values, and in settings concentrating the mass of influence in a neighborhood while allowing from spillovers from everywhere. This is important to allow for in centrality-based diffusion models, SIR models, and financial flow networks, since the spillovers in these settings are less restricted than the sparse dependency graph in their Condition 5.

Indeed, we can even allow the dependency graph to be a complete graph, as long as the correlations between the nodes in this dependency graph satisfy our Assumptions 2.2.1-2.2.3.

That is, we can handle cases where, for a given treatment assignment, each node has n real exposure conditions for which the exposure conditions of the whole graph can be well approximated by simple functions, where small perturbations to any node in large regions of small correlations do not substantially perturb the outcomes in these regions (i.e., across affinity sets) while perturbations of the same size in any region of larger correlations (i.e., within affinity sets) can cause significant changes in the outcomes in that region. One can think of the “shorter" constant intervals in the simple function to be over affinity sets, and the “longer" constant intervals to be across different affinity sets. One can think about monotonically non-decreasing functions, for instance, in epidemic spread settings where any increase in the “treatment" cannot decrease the number of infected nodes.

To take an example, let the true exposure of i be given by $e_i(T_{1:n})$. Then consider a case where $e_i(T^+_{1:n}) - e_i(T_{1:n}) \geq 0$, where $T^+_{1:n}$ indicates an increase in any element $j \in [n]$ from $T_{1:n}$. This is a structure that would happen naturally in a setting with diffusion. The potential outcome for i given treatment assignment is assumed to be $y_i(e_i)$.

In practice, for parsimony and ease exposures are often binned. So consider the problem where the 2^n possible exposures $e_i(T_{1:n})$ can be approximated by K well-separated “effective" exposures $\{d_1, d_2, \ldots, d_K\}$ where $|d_k - d_l| > \delta$ for any $k, l \in \{1, 2, ..., K\}$, $k \neq l$, and some $\delta > 0$. For any $r \in \{1, 2, \ldots, K\}$, and any exposure realizations $e_i(T_{1:n}), e_j(T_{1:n})$, we can take $e_i(T_{1:n}), e_j(T_{1:n})$ to be in bin b_r with effective exposure d_r if and only if $|e_i(T_{1:n}) - e_j(T_{1:n})| < \delta$ and we have $y_i(e_i)$ smooth in its argument for every i.

Then, following the above, the researcher's target estimand is the average causal effect switching between two exposure bins,

$$\tau(d_k, d_l) = \frac{1}{n} \sum_i 1\{e_i(T_{1:n}) \in b_k\} y_i(e_i) - \frac{1}{n} \sum_i 1\{e_i(T_{1:n}) \in b_l\} y_i(e_i).$$

The estimator for this estimand cannot directly be shown to be asymptotically normally distributed using the prior literature. It is ruled out by Condition 5 in Aronow and Samii [2017] which uses Chen and Shao [2004]. However, it is straightforward to apply our result.

An example of this is a sub-critical diffusion process with randomly selected set of nodes M_n being assigned some treatment, and every other node is subsequently infected with some probability. We provide two examples which speak to this in Subsections 2.4.3 and 2.4.4.

2.4.2 Covariance Estimation using Socio-economic Distances

2.4.2.1 Environment

One application of mixing random fields is to use them to develop covariance matrices for estimators (e.g., Driscoll and Kraay [1998], Bester et al. [2011], Barrios et al. [2012], Cressie [2015]). Here we consider the example of Conley and Topa [2002], which builds on Conley [1999]. Essentially, their approach is to parameterize the characteristics (observable or unobservable) of units that drive correlation in shocks by the Euclidean metric, as we further describe below. This, however, rules out examples that are common in practice that include (discrete) characteristics with no intrinsic ordering driving degrees of correlation. For instance, correlational structures across race, ethnicity, caste, occupation, and so on, are not readily accommodated in the framework. For a concrete example, correlations between ethnicities e_i and e_j for units i, j that are parametrized by $p_{e_i e_j}$ in an unstructured manner are ruled out. Like this, many of these examples only admit partial orderings, if that. Yet these are important, practical considerations in applied work. The results in our Theorem 2.2.4 allows an intuitively nice treatment of such cases. Our discussion below also applies to combinations of temporal (and possibly cross-sectional) dependence as in Driscoll and Kraay [1998]. Our conditions also provide consistent estimators for covariance matrices of moment conditions for parameters of interest in the GMM setting, under full-rank conditions of expected derivatives Conley [1999], since the author uses the CLT from Bolthausen [1982] under stationary random fields which is generalized above.

In Conley [1999], the model is that the population lives in a Euclidean space (taken to be $\mathbb{R}^2$ for the purposes of exposition), with each individual i at location s_i. Each location has an associated random field X_{s_i}. Conley [1999] obtains the limiting distribution of parameter

estimates b of $\beta \in B$, where $B \subset \mathbb{R}^d$ is a compact subset and β is the unique solution to $\mathbb{E}[g(X_{s_i}; \beta)]$ for a moment function g. Conley [1999] lists sufficient conditions on the moment function to imply consistent estimation of the expected derivatives and having full rank:

a) for all $b \in B$, the derivative $Dg(X_{s_i}; b)$ with respect to b is measurable, continuous on B for all $X \in \mathbb{R}^k$, and first-moment continuous.

b) $\mathbb{E}[Dg(X_{s_i}; b)] < \infty$ and is of full-rank.

c) $\sum_{s \in \mathbb{Z}^2} \text{Cov}(g(X_0; \beta), g(X_s; \beta))$ corresponding to sampled locations X_s is a non-singular matrix.

In addition to the sufficient conditions on the expected derivatives above, we list the remaining sufficient conditions on the random field X_s itself used by Conley [1999] to obtain the limiting distribution of parameter estimates through the GMM; and these are nested in the conditions from Jenish and Prucha [2009], with the addition of bounded $2 + \delta$-moment of $||g(X_s; \beta)||$:

1. $\sum_{m=1}^{\infty} m\alpha_{k,l}(m) < \infty$ for $k + l \leq 4$

2. $\alpha_{1,\infty}(m) = o(m^{-2})$

3. for some $\delta > 0$, $\mathbb{E}[||g(X_s; \beta)||]^{2+\delta} < \infty$ and $\sum_{m=1}^{\infty} m(\alpha_{k,l}(m))^{\delta/(2+\delta)} < \infty$.

In Conley and Topa [2002], the authors develop consistent covariance estimators, using these conditions, combining different distance metrics including physical distance as well as ethnicity (or occupation, for another example) distance in L_2 at an aggregate level (using census tracts data). In particular, the authors use indicator vectors to encode ethnicities (or occupations), and take the Euclidean distance from aggregated (at the census tract level) indicator vectors, as a measure of these ethnic/occupational distances. They use the

Euclidean metric to write a “race and ethnicity" distance between census tracts i and j,

$$D_{ij} = \sqrt{\sum_{k=1}^{9}(e_{ik} - e_{jk})^2},$$

where the sum is taken over nine ethnicities/races, indexed by k, defined by the authors.

2.4.2.2 Application of Theorem 2.2.4

Consider random variables Z_i and Z_j, whose correlation we can decompose into components of physical distance and racial/ethnic distance (just as in Conley and Topa [2002]). It is direct to see how our work above takes care of the physical distance component, and so we turn to the remaining distance component. For this, for instance, one could consider the pairwise interaction probabilities, p_{e_i,e_j} characterizing the correlation between ethnicity e_i of i, and ethnicity e_j of j. Our affinity set structure then allows us to incorporate this correlation structure. That is, one can construct an affinity set $\mathcal{A}_i^n = \{j : \rho(i,j) \leq K^i(\epsilon/2), \text{ or } p_{e_i,e_j} \geq 1 - \epsilon/2\}$ with K^i defined just as in Subsection 2.3.3. Following our previous section, our generalization holds. Attempts to (non-parametrically) estimate covariance in the cross-section often leverage a time or distance structure. For example, Driscoll and Kraay [1998] assume a mixing condition on a random field such that the correlation between shocks ϵ_{it} and $\epsilon_{j,t-s}$ tends to zero as $s \to \infty$. This allows for reasonably agnostic cross sectional correlational structures but requires it to be temporally invariant and studies $T^{1/2}$ asymptotics.

Although such an assumption applies in certain contexts, there are many socio-economic contexts in which it does not apply and yet our theorem can be applied. For instance, in simple models of migration with migration cost, there is often persistence in how shocks to incentives to migrate in some areas affect populations in other areas. Nonetheless, there are very particular correlation patterns because, as an example, ethnic groups migrate to specific places based on existing populations, and so affinity sets are driven by the places to which a given ethnic group might consider moving and our central limit theorem can then be applied, provided our conditions on affinity sets apply. Another example comes

from social interaction. Individuals interact with others in small groups that experience correlated shocks which correlate grouped individuals' behaviors and beliefs. Each group involves only a tiny portion of the population, and any given person interacts in a series of groups over time. Thus individuals' behaviors or beliefs are correlated with others with whom they have interacted; but without any natural temporal or spatial structure. Each person has an affinity set composed of the groups (classes, teams, and so on) that they have been part of. People may also have their own idiosyncratic shocks to behaviors and beliefs. In this example again, our central limit theorem applies despite the lack of any spatial or temporal structure. One does not need to know the affinity sets, just to know that each person's affinity group is appropriately small relative to the population.

2.4.3 Sub-Critical Diffusion Models

2.4.3.1 Environment

A finite-time SIR diffusion process occurs on a sequence of unweighted and undirected graphs G_n. A first-infected set, or seed set M_n, of size m_n, of random nodes with treatment indicated $W_i \in \{0, 1\}$, are seeded (set to have $W_i = 1$) at $t = 0$ and in period $t = 1$ each infects each of its network neighbors, $\{j : G_{ij,n} = 1\}$ i.i.d. with probability q_n. The seeds then are no longer active in infecting others. In period $t = 2$ each of the nodes infected at period $t = 1$ infects each of its network neighbors who were never previously infected i.i.d. with probability q_n. The process continues for T_n periods. Let $X_i^n \in \{0, 1\}$ be a binary indicator of whether i was ever infected throughout the process.

Assume that the sequence of SIR models under study, (G_n, q_n, T_n, W_n) have $m_n \to \infty$ (with $\alpha_n := m_n/n = o(1)$), $q_n \to 0$, and $T_n \to \infty$ (with $T_n \geq \text{diam}(G_n)$ at each n), and are such that the process is sub-critical. Since the number of periods is at least as large as the diameter, it guarantees that for a connected G_n, $\text{Cov}(X_i^n, X_j^n) > 0$ for each i, j.

The statistician may be interested in a number of quantities. For instance, the unknown parameter q_n may be of interest. Suppose $\mathbb{E}[\Psi_i(X_i; q_n, W_{1:n})] = 0$ is a (scalar) moment

condition satisfied only at the true parameter q_n given known seeding $W_{1:n}$. The Z-estimator (or GMM) derives from the empirical analog, setting $\sum_i \Psi_i(X_i; \hat{q}, W_{1:n}) = 0$.

By a standard expansion argument

$$(\hat{q} - q_n) = -\left\{\sum_i \nabla_q \Psi_i(X_i; \tilde{q}, W_{1:n})\right\}^{-1} \times \sum_i \Psi_i(X_i; q_n, W_{1:n}) + o_p(n^{-1/2}).$$

To study the asymptotic normality of the estimator, we need to study

$$\frac{1}{\sqrt{\text{Var}\left(\sum_i \Psi_i\right)}} \sum_i \Psi_i(X_i; q_n, W_{1:n}),$$

which involves developing affinity sets for each Ψ_i. For simplicity we consider the case where the estimator may directly work with $X^n_{1:n}$. Letting $Z_i = X^n_i - \mathbb{E}(X^n_i)$ be the de-meaned outcome and we want to show that

$$\frac{1}{\sqrt{\sum_i \sum_{j \in \mathcal{A}^n_i} \text{Cov}(Z^n_i, Z^n_j)}} \sum_i Z^n_i \rightsquigarrow \mathcal{N}(0, 1).$$

Under sub-criticality, a vanishing share of nodes are infected from a single seed. Without sub-criticality, most of the graph can have a nontrivial probability of being infected and accurate inference cannot be made with a single network. Let us define

$$\mathcal{B}^n_j := \{i : \mathbb{P}(X^n_i = 1 \mid j \in M_n, m_n = 1) > \epsilon_n\}.$$

Then $\mathcal{B}^n_j$ is the set of nodes for which, if j is the only seed, the probability of being infected in the process is at least ϵ_n. As noted above, in a sub-critical process $\left|\mathcal{B}^n_j\right| = o(n)$ for every j, not necessarily uniformly in j. For simplicity assume that there is a sequence $\epsilon_n \to 0$ such that this holds uniformly (otherwise, one can simply consider sums). Let $\beta_n := \sup_j \left|\mathcal{B}^n_j\right| / n$ which tends to zero.

Next, we assume that the rate at which infections happen within the affinity set is higher than outside of it, and the share of seeds is sufficiently high and affinity sets are large enough

to lead to many small infection outbursts but not so large as to infect the whole network. That is, there exists some $\mathcal{B}_j^{n\prime} \subset \mathcal{B}_j^n$ such that $|\mathcal{B}_j^{n\prime}| = \Theta(|\mathcal{B}_j^n|)$ and $\mathbb{P}(X_i = 1 | j \in M_n, m_n = 1) \geq \gamma_n$ with $\gamma_n/\epsilon_n \to \infty$, such that $m_n = o(\gamma_n^{-1})$, $m_n = o(\beta_n^{-1/2})$, and $\gamma_n = o(\beta_n)$. This would apply to, for instance, targeted advertisements or promotions that lead to local spread of information about a product, but that does not go viral. None of the prior examples such as random fields and dependency graphs cover this case since all X_i^n are correlated. We now show that Theorem 2.2.4 applies to this case.

2.4.3.2 Application of Theorem 2.2.4

Let us define the affinity sets $\mathcal{A}_i^n = \mathcal{B}_i^n$. Consider the event $\mathcal{F}_S := \{|M_n \cap (\mathcal{B}_{i_1}^n \cup \mathcal{B}_{i_2}^n \cup \ldots \cup \mathcal{B}_{i_k}^n)| \leq 1, \{i_1, i_2, ..., i_k \in S \subseteq [n]\}\}$, where there is at most one seed in the affinity sets considered, and $|S| = O(1)$. Then, $\mathbb{P}(\mathcal{F}_S) \approx 1 - m_n^2\beta_n \to 1$, since the seeds are distributed uniformly at random across the nodes, and $m_n = o(\beta_n^{-1/2})$. Therefore, by the law of total probability, it is sufficient to consider the dominating terms resulting from conditioning on this event in the following. Let $\sigma^2 = \mathrm{Var}(Z_i)$. We begin by noting that $\Omega_n = \Theta(n^2\beta_n\sigma^2)$. Now, we check the first assumption:

$$\sum_{i;j,k \in \mathcal{A}_i^n} \mathbb{E}[|Z_i^n | Z_j^n Z_k^n] = O(n(n\beta_n)^2 \mathbb{E}[|Z_i^n|^3]) = o(n^3\beta_n^{3/2}\sigma^3) = o(\Omega_n^{3/2}).$$

The second condition is satisfied noting that

$$\sum_{i;j,k \in \mathcal{A}_i^n, r \in \mathcal{A}_k^n} \mathrm{Cov}(Z_i^n Z_j^n, Z_k^n Z_r^n) = O(n(n\beta_n)^3 \mathbb{E}[(Z_i^n)^4]) = O(n^4\beta_n^3) = o(n^4\beta_n^2) = o(\Omega_n^2).$$

Finally, we check the third condition. For fixed nodes $i, j \in [n]$, define the event $\mathcal{E}_{i,j,k} := \bigcap_{\substack{a,b \in \{i,j,k\} \\ a \neq b}} \{a \notin \mathcal{B}_b^n\}$ for each $k \in M_n$, where none of the nodes i, j, k are in each other's affinity sets, with k being a seed. Then, the probability that no seed is in the affinity sets of i, j, and i, j are not in the affinity sets of the seeds and of each other is $\mathbb{P}(\cap_{k \in M_n} \mathcal{E}_{i,j,k}) \approx (1 - \beta_n)^{2m_n} \approx 1 - 2m_n\beta_n \to 1$, by the uniform random distribution of

the seeds across the nodes, and $m_n = o(\beta_n^{-1/2})$. Therefore, it is sufficient to consider the dominating terms resulting from conditioning on this event in what follows. Then, notice that $\text{Cov}(Z_i, Z_j | \cap_{k \in M_n} \mathcal{E}_{i,j,k}) = O(\epsilon_n)$, for $j \notin \mathcal{A}_i$, for every i. Thus,

$$\sum_i \mathbb{E}[|Z^n_{-i}\mathbb{E}[Z^n_i|Z^n_{-i}]|] = O(n(n\beta_n)(\beta_n^{-1} - 1)\epsilon_n) = O(n^2(1-\beta_n)\epsilon_n) = o(n^2\beta_n) = o(\Omega_n).$$

2.4.4 Diffusion in Stochastic Block Models

2.4.4.1 Environment

A SIR diffusion process occurs on a sequence of unweighted and undirected networks, as in the previous example, except that the network has a block structure as generated by a standard stochastic block model (Holland et al. [1983]). The n nodes are partitioned into k_n blocks, where block sizes are equal, or within one of each other. With probability $p_n^{in} \in (0,1)$ links are formed inside blocks, and with probability $p_n^{ac} \in (0,1)$ they are formed across blocks, independently. Let q_n be the probability of infection, as in the last example. Inside link probabilities are large enough for percolation within blocks: $p_n^{in}\frac{n}{k_n}q_n >> \log\left(\frac{n}{k_n}\right)$. Across link probabilities are small enough for vanishing probabilities of contagion across blocks, even if all other blocks are infected: $p_n^{ac}n^2q_n << 1$. The infections are seeded with $k_n/2$ seeds. With probability going to 1, all nodes in the blocks with the seeds will be infected and no others. There is a correlation going to 1 of infection status of nodes within the blocks, which are the affinity sets; and there is correlation of infection status going to 0 across blocks, but it is always positive. If k_n is bounded, then a central limit theorem fails. If k_n grows without bound (while allowing $n/k_n \to \infty$ so that blocks are large), then the central limit theorem holds.

2.4.4.2 Application of Theorem 2.2.4

Given the previous examples, we simply sketch the application. The affinity set of node i, $\mathcal{A}_i^n$, is the block in which it resides. Letting σ_n^2 denote $\text{Var}(Z_i)$, it follows that $\Omega_n \approx n\frac{n}{k_n}\sigma_n^2$.

Then the first assumption is satisfied noting that if $k_n \to \infty$, then

$$\sum_{i;j,k\in\mathcal{A}_i^n} \mathbb{E}[|Z_i|Z_jZ_k] = O(\sum_{i;j,k\in\mathcal{A}_i^n} \mathbb{E}[|Z_i|^3]) = O\left(n\left(\frac{n}{k_n}\right)^2 \mathbb{E}[|Z_i|^3]\right) = o(\Omega_n^{3/2}).$$

Verification of the second assumption comes from noting that if $k_n \to \infty$ then

$$\begin{aligned}\sum_{i;j,k\in\mathcal{A}_i^n, r\in\mathcal{A}_k^n} \mathrm{Cov}(Z_i^n Z_j^n, Z_k^n Z_r^n) &= O\left(\sum_{i;j,k,r\in\mathcal{A}_i^n} \mathrm{Cov}(Z_i^n Z_j^n, Z_k^n Z_r^n)\right) \\ &= O\left(n\left(\frac{n}{k_n}\right)^3 \mathbb{E}[Z_i^4]\right) = o(\Omega_n^2).\end{aligned}$$

Next we check the third assumption. Let $\varepsilon_n = \mathrm{Cov}(Z_i^n, Z_j^n)$ for i, j in different blocks (ignoring the approximation due to the fact that blocks may be of slightly different sizes). Note that $\varepsilon_n = \mathrm{Cov}(Z_i^n, Z_j^n)$ is on the order of contagion across blocks, which is $p_n^{ac} q_n (n^2/k_n^2) = o(1/k_n^2)$. Both blocks could also be infected by some other nodes, which happens of order at most $n^2(p^{ac}q_n(n/k_n))^2$, which is also of order $o(1/k_n^2)$. If k_n grows without bound

$$\sum_i \mathbb{E}\left(|\mathbb{E}[Z_i \mathbf{Z}_{-i}|\mathbf{Z}_{-i}]|\right) = O\left(n^2 \frac{k_n - 1}{k_n}\varepsilon_n\right) = o\left(\frac{n^2}{k_n}\sigma_n^2\right).$$

In this example, not only do the assumptions fail if k_n is bounded, but also the conclusion of the theorem fails to hold as well; so the conditions are tight in that sense.

2.5 Discussion

We have provided an organizing principle for modeling dependency and obtaining a central limit theorem: affinity sets. It allows for non-zero correlation across all random vectors in the triangular array and places focus on correlations within and across sets. These conditions are intuitive and we illustrate their use through some practical applications for applied research. In some cases, as in several of our applied examples, our result is needed as previous conditions do not apply.

We now reflect on settings that our theorem does not cover. For example, the martingale

central limit theorem (e.g., Billingsley [1961], Ibragimov [1963], Hall and Heyde [2014] among others) is not covered by our theorem, without modification. It admits nontrivial unconditional correlation between all variables, but relies on other structural properties to deduce the result. In fact, proofs of the martingale central limit theorem did not appeal to Stein's method, until Röllin [2018]. By combining Stein and Lindeberg methods, Röllin [2018] develops a shorter proof but did not find a direct proof using the Stein technique alone. Some biased processes that do not fall under the martingale umbrella can still generate a central limit theorem if they satisfy the covariance structure that we have provided.

Online Appendix

Proof of Central Limit Theorem 2.2.4 and Corollary 2.2.5

2.A Preliminary Univariate CLT

To prove Theorem 2.2.4, we apply the Cramer-Wold device to a univariate affinity set CLT, previously appearing at Theorem 2 in Chandrasekhar and Jackson [2024]. Consider a triangular array of scalar real-valued random variables X_i^n with indices $i \in \{1, \ldots, n\}$. Let the mean-normalized variables be $Z_i^n = X_i^n - \mathbb{E}[X_i^n]$.

For each i, n, the index set $\{1, \ldots, n\}$ is partitioned into an *affinity set* $\mathcal{A}_i^n \subseteq \{1, \ldots, n\}$ (with $i \in \mathcal{A}_i^n$) and its complement. Define the variance normalizer

$$\Omega_n := \sum_{i,j:j\in\mathcal{A}_i^n} \mathrm{Cov}(Z_i, Z_j),$$

and let $\mathbf{Z}_{-i} := \sum_{j\notin\mathcal{A}_i^n} Z_j$ denote the sum over all variables outside the reference index's affinity set.

The following conditions on the covariance are required to control dependence amongst observations.

Assumption 2.A.1 (Within affinity set covariance control)**.**

$$\sum_{i,j,k:j,k\in\mathcal{A}_i^n} \mathbb{E}\left[|Z_i| \cdot Z_j Z_k\right] = o\left(\Omega_n^{3/2}\right).$$

Assumption 2.A.2 (Cross-affinity set covariance control)**.**

$$\sum_{i,i',j,j':j\in\mathcal{A}_i^n, j'\in\mathcal{A}_{i'}^n} \left|\mathrm{Cov}(Z_i Z_j, Z_{i'} Z_{j'})\right| = o\left(\Omega_n^2\right).$$

Assumption 2.A.3 (Outside affinity set covariance control)**.**

$$\sum_i \mathbb{E}\left[|\mathbf{Z}_{-i} \cdot \mathbb{E}[Z_i \mid \mathbf{Z}_{-i}]|\right] = o\left(\Omega_n\right).$$

Theorem 2.A.4 (Univariate CLT)**.** *Suppose that* $\mathbb{E}[|Z_i^n|^3]/\mathbb{E}[(Z_i^n)^2]^{3/2}$ *is bounded above for all* i, n *and assume that* $\liminf_{n\to\infty} \Omega_n > 0$*. If Assumptions 2.A.1, 2.A.2, and 2.A.3 are satisfied, then*

$$\sum_{i=1}^{n} Z_i^n / \sqrt{\Omega_n} \xrightarrow{d} \mathcal{N}(0,1).$$

In Chandrasekhar and Jackson [2024], Theorem 2.A.4 is stated as requiring $\Omega_n \to \infty$, but inspection of the proof reveals that $\liminf_{n\to\infty} \Omega > 0$ is sufficient for the theorem to hold. We re-produce a proof of the theorem here, including this slight relaxation on Ω.

The proof uses Stein's lemma from Stein [1986]. The key observation of Stein [1986] is that if a random variable satisfies

$$\mathbb{E}[f'(Y) - Yf(Y)] = 0$$

for every $f(\cdot)$ that is continuously differentiable, then it must have a standard normal distribution. This observation leads to a useful lemma, that allows one to characterize the Kolmogorov distance between a random variable Y and a standard normally distributed Z, denoted $d_K(Y, Z)$. We can bound this from above by (a constant times) the Wasserstein distance, $d_W(Y, Z)$, which itself is bounded by the below expression.

Lemma 2.A.5 (Stein [1986], Ross [2011]). If Y is a random variable and Z has the standard normal distribution, then

$$d_W(Y, Z) \leq \sup_{\{f:||f||,||f''||\leq 2,||f'||\leq\sqrt{2/\pi}\}} \left|\mathbb{E}[f'(Y) - Yf(Y)]\right|.$$

Further $d_K(Y, Z) \leq (2/\pi)^{1/4}(d_W(Y, Z))^{1/2}$.

By this Lemma 2.A.5, if we show that a normalized sum of random variables satisfies

$$\sup_{\{f:||f||,||f''||\leq 2,||f'||\leq\sqrt{2/\pi}\}} \left|\mathbb{E}[f'(\overline{S}^n) - \overline{S}^n f(\overline{S}^n)]\right| \to 0,$$

then $d_W(\overline{S}^n, Z) \to 0$, and so it must be asymptotically normally distributed.

The following lemmas are useful in the proof.

Lemma 2.A.6 (Chandrasekhar and Jackson [2024], Lemma B.2). A solution to $\max_h \mathbb{E}[Zh(Y)]$ s.t. $|h| \leq 1$ (where h is measurable) is $h(Y) = \text{sign}(\mathbb{E}[Z|Y])$, where we break ties, setting $\text{sign}(\mathbb{E}[Z|Y]) = 1$ when $\mathbb{E}[Z|Y] = 0$.

Lemma 2.A.7 (Chandrasekhar and Jackson [2024], Lemma B.3). $\mathbb{E}[XYh(Y)]$ when $h(\cdot)$ is measurable and bounded by $\sqrt{\frac{2}{\pi}}$ satisfies

$$\mathbb{E}[XYh(Y)] \leq \sqrt{\frac{2}{\pi}}\mathbb{E}\left[XY \cdot \text{sign}(\mathbb{E}[X|Y]Y)\right].$$

Proof of Theorem 2.A.4. By Lemma 2.A.5, it is sufficient to show that the normalized sequence of random variables $\overline{S}^n$ satisfies

$$\sup_{\{f:||f||,||f''||\leq 2,||f'||\leq\sqrt{2/\pi}\}} \left|\mathbb{E}[f'(\overline{S}^n) - \overline{S}^n f(\overline{S}^n)]\right| \to 0.$$

Let $\mathbf{Z}_{-i} := \sum_{j\notin\mathcal{A}_i^n} Z_j$ and let $\overline{S}_i := \Omega_n^{-1/2}\mathbf{Z}_{-i}$. Recall $\overline{S}^n = \Omega_n^{-1/2}S^n$. We evaluate the expectation in the Stein equation:

$$\begin{aligned}\mathbb{E}\left[\overline{S}^n f\left(\overline{S}^n\right)\right] &= \mathbb{E}\left[\frac{1}{\Omega_n^{1/2}}\sum_{i=1}^n Z_i \cdot f\left(\overline{S}^n\right)\right] \\ &= \mathbb{E}\left[\frac{1}{\Omega_n^{1/2}}\sum_{i=1}^n Z_i\left(f\left(\overline{S}^n\right) - f\left(\overline{S}_i\right)\right)\right] + \mathbb{E}\left[\frac{1}{\Omega_n^{1/2}}\sum_{i=1}^n Z_i \cdot f\left(\overline{S}_i\right)\right].\end{aligned}$$

We first show that the second term vanishes asymptotically. By a first-order Taylor

approximation of f around 0, and noting that $\mathbb{E}[Z_i] = 0$, we can bound this term:

$$\left|\mathbb{E}\left[\frac{1}{\Omega_n^{1/2}}\sum_{i=1}^n Z_i \cdot f\left(\overline{S}_i\right)\right]\right| \leq \left|\mathbb{E}\left[\frac{1}{\Omega_n^{1/2}}\sum_{i=1}^n Z_i \cdot f\left(0\right)\right]\right| + \left|\mathbb{E}\left[\frac{1}{\Omega_n^{1/2}}\sum_{i=1}^n Z_i \cdot \overline{S}_i f'\left(\tilde{S}_i\right)\right]\right|$$
$$= \left|\frac{1}{\Omega_n}\mathbb{E}\left[\sum_{i=1}^n Z_i \mathbf{Z}_{-i} f'\left(\tilde{S}_i\right)\right]\right|,$$

where $\tilde{S}_i$ lies between 0 and $\overline{S}_i$. Applying Lemma 2.A.7 and using the bound $||f'|| \leq \sqrt{2/\pi}$, we obtain:

$$\left|\frac{1}{\Omega_n}\mathbb{E}\left[\sum_{i=1}^n Z_i \mathbf{Z}_{-i} f'\left(\tilde{S}_i\right)\right]\right| \leq \frac{1}{\Omega_n}\sqrt{\frac{2}{\pi}}\left|\mathbb{E}\left[\sum_{i=1}^n Z_i \mathbf{Z}_{-i} \cdot \text{sign}(\mathbb{E}[Z_i \mathbf{Z}_{-i} \mid \mathbf{Z}_{-i}])\right]\right|$$
$$\leq \frac{1}{\Omega_n}\sqrt{\frac{2}{\pi}}\sum_{i=1}^n \mathbb{E}\left[|\mathbf{Z}_{-i}\mathbb{E}\left[Z_i \mid \mathbf{Z}_{-i}\right]|\right].$$

By Assumption 2.A.3, this sum is $o(\Omega_n)$. Because Ω_n is uniformly bounded away from zero (i.e., $\Omega_n \geq c > 0$ for some constant c for sufficiently large n), scaling by Ω_n^{-1} perfectly preserves the $o(1)$ convergence. Thus, the second term evaluates to $o(1)$.

We now bound the remaining elements of the Stein discrepancy. Adding and subtracting $(\overline{S}^n - \overline{S}_i)f'(\overline{S}^n)$ yields:

$$\left|\mathbb{E}[f'(\overline{S}^n) - \overline{S}^n f(\overline{S}^n)]\right|$$
$$\leq \left|\mathbb{E}\left[\frac{1}{\Omega_n^{1/2}}\sum_{i=1}^n Z_i\left(f(\overline{S}^n) - f(\overline{S}_i) - (\overline{S}^n - \overline{S}_i)f'(\overline{S}^n)\right)\right]\right|$$
$$+ \left|\mathbb{E}\left[f'(\overline{S}^n)\left(1 - \frac{1}{\Omega_n^{1/2}}\sum_{i=1}^n Z_i(\overline{S}^n - \overline{S}_i)\right)\right]\right| + o(1).$$

We denote the first two terms on the right-hand side as A_1 and A_2. We bound A_1 using a second-order Taylor expansion, $|f(x) - f(y) - (x-y)f'(x)| \leq \frac{||f''||}{2}(x-y)^2$:

$$A_1 \leq \frac{||f''||}{2\Omega_n^{1/2}}\sum_{i=1}^n \mathbb{E}\left[|Z_i|(\overline{S}^n - \overline{S}_i)^2\right]$$

$$= \frac{||f''||}{2\Omega_n^{3/2}} \sum_{i=1}^{n} \mathbb{E}\left[|Z_i| \left(\sum_{j \in \mathcal{A}_i^n} Z_j \right)^2 \right]$$
$$= \frac{||f''||}{2\Omega_n^{3/2}} \sum_{i=1}^{n} \sum_{j,k \in \mathcal{A}_i^n} \mathbb{E}\left[|Z_i| Z_j Z_k \right].$$

By Assumption 2.A.1, the covariance sum is $o(\Omega_n^{3/2})$. Because $\Omega_n \geq c > 0$ for sufficiently large n, we have $A_1 = o(1)$.

Finally, we bound A_2. Note that $\overline{S}^n - \overline{S}_i = \Omega_n^{-1/2} \sum_{j \in \mathcal{A}_i^n} Z_j$, and that $\Omega_n = \sum_{i=1}^n \sum_{j \in \mathcal{A}_i^n} \mathrm{Cov}(Z_i, Z_j) = \mathbb{E}\left[\sum_{i=1}^n \sum_{j \in \mathcal{A}_i^n} Z_i Z_j \right]$. We have:

$$A_2 = \left| \mathbb{E}\left[f'(\overline{S}^n) \left(1 - \frac{1}{\Omega_n} \sum_{i=1}^{n} \sum_{j \in \mathcal{A}_i^n} Z_i Z_j \right) \right] \right|$$
$$\leq \frac{||f'||}{\Omega_n} \mathbb{E}\left[\left| \Omega_n - \sum_{i=1}^{n} \sum_{j \in \mathcal{A}_i^n} Z_i Z_j \right| \right].$$

Applying the Cauchy-Schwarz inequality to the expectation yields:

$$A_2 \leq \frac{||f'||}{\Omega_n} \sqrt{ \mathrm{Var}\left(\sum_{i=1}^{n} \sum_{j \in \mathcal{A}_i^n} Z_i Z_j \right) }$$
$$\leq \frac{\sqrt{2/\pi}}{\Omega_n} \left(\sum_{i,j \in \mathcal{A}_i^n} \sum_{i',j' \in \mathcal{A}_{i'}^n} \mathrm{Cov}(Z_i Z_j, Z_{i'} Z_{j'}) \right)^{1/2}.$$

By Assumption 2.A.2, the variance sum is $o(\Omega_n^2)$. Thus, the square root evaluates to $o(\Omega_n)$. Scaled by Ω_n^{-1}, we obtain $A_2 = o(1)$.

Combining these bounds, the total Stein discrepancy evaluates to $o(1)$, establishing convergence to the standard normal distribution and completing the proof. ■

2.B Extension to multivariate CLT

To extend the univariate affinity set CLT to the multivariate case, we apply the Cramér-Wold device under appropriately updated conditions.

Lemma 2.B.1 (Cramér and Wold [1936]). The sequence $\{X_n\}_n$ of random vectors of dimension $p \in \mathbb{N}$ weakly converges to the random vector $X \in \mathbb{R}^p$, as $n \to \infty$, if and only if for any $c \in \mathbb{R}^p$, as $n \to \infty$,

$$c^\intercal X_n \rightsquigarrow c^\intercal X$$

With Lemma 2.B.1 in hand, we are ready to prove Theorem 2.2.4.

Proof of Theorem 2.2.4. By the Cramér-Wold device, $\Omega_n^{-1/2} S^n \xrightarrow{d} \mathcal{N}(0, I_p)$ if and only if for every fixed, non-zero projection vector $\lambda \in \mathbb{R}^p$,

$$\lambda^\top (\Omega_n^{-1/2} S^n) \xrightarrow{d} \mathcal{N}(0, \|\lambda\|^2).$$

Thus, we fix an arbitrary $\lambda \neq 0$ and define $c_n = \Omega_n^{-1/2}\lambda$. We will apply Theorem 2.A.4 to $Y_i = c_n^\top Z_i = \lambda^\top \Omega_n^{-1/2} Z_i$. The random variables Y_i are mean zero since, $\mathbb{E}[Y_i] = c_n^\top \mathbb{E}[Z_i] = 0$. Further, they satisfy the moment ratio condition, since, taking $u = c_n = \Omega_n^{-1/2}\lambda$ and applying the multivariate moment ratio hypothesis

$$\frac{\mathbb{E}\left[\left|c_n^\top Z_i\right|^3\right]}{\left(c_n^\top \mathbb{E}\left[Z_i Z_i^\top\right] c_n\right)^{3/2}} = \frac{\mathbb{E}\left[\left|c_n^\top Z_i\right|^3\right]}{\mathbb{E}\left[c_n^\top Z_i Z_i^\top c_n\right]^{3/2}} = \frac{\mathbb{E}\left[|Y_i|^3\right]}{\mathbb{E}\left[\left(c_n^\top Z_i\right)^2\right]^{3/2}} = \frac{\mathbb{E}\left[|Y_i|^3\right]}{\mathbb{E}\left[Y_i^2\right]^{3/2}} \leq K$$

The univariate variance stabilizer, which we denote as V_n, is the sum of the covariances strictly within the unified affinity sets:

$$V_n = \sum_{i=1}^n \sum_{j \in \tilde{\mathcal{A}}_i} \operatorname{cov}(Y_i, Y_j) = \sum_{i=1}^n \sum_{j \in \tilde{\mathcal{A}}_i} \operatorname{cov}(c_n^\top Z_i, c_n^\top Z_j) = \sum_{i=1}^n \sum_{j \in \tilde{\mathcal{A}}_i} c_n^\top \operatorname{cov}(Z_i, Z_j) c_n = c_n^\top \Omega_n c_n$$

Substituting $c_n = \Omega_n^{-1/2}\lambda$, and using the fact that $\liminf_{n\to\infty} \lambda_{\min}(\Omega_n) > 0$, such that Ω_n is non-singular for sufficiently large n,

$$\liminf_{n\to\infty} V_n = \liminf_{n\to\infty} c_n^\top \Omega_n c_n = \liminf_{n\to\infty} \left(\Omega_n^{-1/2}\lambda\right)^\top \Omega_n \left(\Omega_n^{-1/2}\lambda\right) = \lambda^\top \lambda = \|\lambda\|^2$$

Thus, the initial standardization by $\Omega_n^{-1/2}$ ensures that the univariate affinity-set covariance normalizer for the projected sequence evaluates to a strictly constant $V_n = \|\lambda\|^2$. Since $\liminf_{n\to\infty} V_n > 0$, the corresponding univariate condition of Theorem 2.A.4 is satisfied.

We also bound the squared norm of the weighting vector c_n, which will be used to control the bounds of the univariate assumptions:

$$\|c_n\|^2 = \lambda^\top \Omega_n^{-1} \lambda \le \|\lambda\|^2 \lambda_{\max}(\Omega_n^{-1}) = \frac{\|\lambda\|^2}{\lambda_{\min}(\Omega_n)} \tag{2.B.1}$$

It remains to verify that Assumptions 2.A.1, 2.A.2, and 2.A.3 hold when applied to Y_i. To verify Assumption 2.A.1, expand $Y_i = c_n^\top Z_i$, apply Cauchy-Schwartz, eq. (2.B.1) and Assumption 2.2.1:

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j,k\in\tilde{\mathcal{A}}_i} \mathbb{E}\left[|Y_i| \cdot Y_j Y_k\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\left|c_n^\top Z_i\right| \left(c_n^\top \sum_{j\in\tilde{\mathcal{A}}_i} Z_j\right)^2\right] \\
&\le \|c_n\|^3 \sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\| \cdot \left\|\sum_{j\in\tilde{\mathcal{A}}_i} Z_j\right\|^2\right] \\
&\le \frac{\|\lambda\|^3}{\lambda_{\min}(\Omega_n)^{3/2}} \sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\| \cdot \left\|\sum_{j\in\tilde{\mathcal{A}}_i} Z_j\right\|^2\right] \\
&= \frac{\|\lambda\|^3}{\lambda_{\min}(\Omega_n)^{3/2}} \cdot o(\lambda_{\min}(\Omega_n)^{3/2}) \\
&= o(1) = o\left(V_n^{3/2}\right) = o\left(\|\lambda\|^3\right).
\end{aligned}$$

To verify that Assumption 2.A.2 holds, using eq. (2.B.1) and Assumption 2.2.2:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in\tilde{\mathcal{A}}_i}\sum_{l\in\tilde{\mathcal{A}}_j} \left|\operatorname{Cov}\left((c_n^\top Z_i)(c_n^\top Z_k), (c_n^\top Z_j)(c_n^\top Z_l)\right)\right| \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in\tilde{\mathcal{A}}_i}\sum_{l\in\tilde{\mathcal{A}}_j} \left|\sum_{d_1,d_2,d_3,d_4=1}^{p} (c_n)_{d_1}(c_n)_{d_2}(c_n)_{d_3}(c_n)_{d_4} \operatorname{Cov}\left(Z_{i,d_1}Z_{k,d_2}, Z_{j,d_3}Z_{l,d_4}\right)\right| \\
&\le \|c_n\|^4 \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in\tilde{\mathcal{A}}_i}\sum_{l\in\tilde{\mathcal{A}}_j}\sum_{d_1,d_2,d_3,d_4=1}^{p} \left|\operatorname{Cov}(Z_{i,d_1}Z_{k,d_2}, Z_{j,d_3}Z_{l,d_4})\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\|\lambda\|^4}{\lambda_{\min}(\Omega_n)^2} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in\tilde{\mathcal{A}}_i}\sum_{l\in\tilde{\mathcal{A}}_j}\sum_{d_1,d_2,d_3,d_4=1}^{p} |\operatorname{Cov}(Z_{i,d_1}Z_{k,d_2}, Z_{j,d_3}Z_{l,d_4})| \\
&= \frac{\|\lambda\|^4}{\lambda_{\min}(\Omega_n)^2} \cdot o(\lambda_{\min}(\Omega_n)^2) \\
&= o(1) = o\left(V_n^2\right) = o\left(\|\lambda\|^4\right).
\end{aligned}
$$

Verifying that Assumption 2.A.3 holds completes the proof. Let $\mathbf{Y}_{-i} = c_n^\top \mathbf{Z}_{-i}$ be the scalar sum outside i's unified affinity set. Apply eq. (2.B.1) and Assumption 2.2.3:

$$
\begin{aligned}
\sum_{i=1}^{n}\mathbb{E}\left[|\mathbf{Y}_{-i}\cdot\mathbb{E}[Y_i \mid \mathbf{Y}_{-i}]|\right] &= \sum_{i=1}^{n}\mathbb{E}\left[|\mathbf{Y}_{-i}\cdot\mathbb{E}[\mathbb{E}[Y_i \mid \mathbf{Z}_{-i}] \mid \mathbf{Y}_{-i}]|\right] \\
&\leq \sum_{i=1}^{n}\mathbb{E}\left[|\mathbf{Y}_{-i}|\cdot\mathbb{E}[|\mathbb{E}[Y_i \mid \mathbf{Z}_{-i}]| \mid \mathbf{Y}_{-i}]\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[|\mathbf{Y}_{-i}|\cdot|\mathbb{E}[Y_i \mid \mathbf{Z}_{-i}]|\right] \\
&\leq \|c_n\|^2 \sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{Z}_{-i}\|\cdot\|\mathbb{E}[Z_i \mid \mathbf{Z}_{-i}]\|\right] \\
&\leq \frac{\|\lambda\|^2}{\lambda_{\min}(\Omega_n)} \sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{Z}_{-i}\|\cdot\|\mathbb{E}[Z_i \mid \mathbf{Z}_{-i}]\|\right] \\
&= \frac{\|\lambda\|^2}{\lambda_{\min}(\Omega_n)} \cdot o(\lambda_{\min}(\Omega_n)) \\
&= o(1) = o\left(V_n\right) = o\left(\|\lambda\|^2\right).
\end{aligned}
$$

■

Proof of Corollary 2.2.5. Under the singleton restriction $\tilde{\mathcal{A}}_i = \{i\}$, Assumption 2.2.1 reduces to the assumed bound $\sum_{i=1}^{n} E[\|Z_i\|^3] = o(\lambda_{\min}(\Omega_n)^{3/2})$. Assumption 2.2.2 simplifies to Condition (1). For Assumption 2.2.3, if $E[Z_{i,d_1}E[Z_{i,d_2}|Z_{-i}]] \geq 0$, iterated expectations allow bounding the outside-affinity influence by the cross-observation covariances, which is $o(\lambda_{\min}(\Omega_n))$ by Condition (2). If this non-negativity fails, the corollary's alternative condition directly imposes Assumption 2.2.3. ■

Proof of Corollary 2.2.6. Under the whole-sample restriction $\tilde{\mathcal{A}}_i = \{1, \ldots, n\}$, $\sum_{j\in\tilde{A}_i} Z_j =$

S^n, so Assumption 2.2.1 reduces to the assumed bound $\sum_{i=1}^n \mathbb{E}[\|Z_i\| \cdot \|S^n\|^2] = o(\lambda_{\min}(\Omega_n)^{3/2})$. Assumption 2.2.2 simplifies to first condition of the Corollary. For Assumption 2.2.3, the complement set is empty, rendering $Z_{-i} = 0$. Because $\mathbb{E}[Z_i] = 0$, the conditional expectation $\mathbb{E}[Z_i \mid Z_{-i}] = \mathbb{E}[Z_i] = 0$. Consequently, the outside-affinity influence evaluates to $\sum_{i=1}^n \mathbb{E}[\|Z_i\| \cdot 0] = 0$, which is $o(\lambda_{\min}(\Omega_n))$. ■

2.C Gap Between Marginal and Joint Conditions

Consider a sequence of random vectors $X_1, ..., X_n \in \mathbb{R}^p$. If the univariate conditions (Assumptions 2.A.1, 2.A.2, and 2.A.3) hold marginally for each dimension of X_i, the multivariate conditions (Assumptions 2.2.1, 2.2.2, and 2.2.3) may not hold. In the following, we provide two examples that demonstrate the difference between marginal and joint asymptotic normality. We then provide additional conditions that combine with Assumptions 2.A.1, 2.A.2, and 2.A.3 to satisfy the multivariate conditions of Assumption 2.2.1, 2.2.2, and 2.2.3.

Example 2.C.1 (Degenerate Joint Covariance)**.** Let $X_i \sim \mathcal{N}(0, 1)$ and $\epsilon_i \sim \mathcal{N}(0, n^{-2})$ be mutually independent sequences. Consider the bivariate sequence:

$$Z_i = \begin{pmatrix} X_i \\ -X_i + \epsilon_i \end{pmatrix}.$$

For dimension $d = 1$, $Z_{i,1} = X_i$ is an independent sequence. The marginal affinity set is minimal: $\mathcal{A}_{i,1}^n = \{i\}$. The variance normalizer is $a_{n,1} = n \to \infty$. Because $\mathcal{A}_{i,1}^n$ only contains i, within-affinity sums are $O(1)$, cross-affinity covariances are 0, and outside-affinity influence is 0. For dimension $d = 2$, $Z_{i,2} = -X_i + \epsilon_i$ is similarly independent with $\mathcal{A}_{i,2}^n = \{i\}$ and $a_{n,2} = n(1 + n^{-2}) \to \infty$. The univariate conditions of Theorem 2.A.4 hold trivially.

However, the multivariate conditions of Theorem 2.2.4 do not hold. The within-affinity set covariance matrix of the partial sum $S^n = \sum_{i=1}^n Z_i$ is:

$$\Omega_n = \sum_{i=1}^n \text{Cov}(Z_i) = n \begin{pmatrix} 1 & -1 \\ -1 & 1 + n^{-2} \end{pmatrix}$$

The eigenvalues λ_1, λ_2 must satisfy $\lambda_1\lambda_2 = 1$ and $\lambda_1 + \lambda_2 \approx 2n$. Consequently, the minimal eigenvalue is:

$$\lambda_{\min}(\Omega_n) \approx 1/(2n) \to 0.$$

The multivariate premise $\lambda_{\min}(\Omega_n) \to \infty$ fails.

Example 2.C.2 (Cross-Dimensional Contagion and Affinity Set Explosion)**.** Let $X_1, \ldots, X_n$ and $\nu_{1,1}, \ldots, \nu_{n,1}$ and $\nu_{1,2}, \ldots, \nu_{n,2}$ be jointly and mutually independent standard normal random variables. Define the bivariate sequence on a ring (with $X_{n+1} \equiv X_1$):

$$Z_{i,1} = X_i + \nu_{i,1}$$
$$Z_{i,2} = X_{i+1} + \nu_{i,2}$$

Both sequences $Z_{i,1}$ and $Z_{i,2}$ are collections of independent $\mathcal{N}(0, 2)$ variables. The marginal affinity sets are $\mathcal{A}^n_{i,1} = \{i\}$ and $\mathcal{A}^n_{i,2} = \{i\}$. The variance normalizers are $a_{n,1} = 2n$ and $a_{n,2} = 2n$. The conditions for the marginal univariate CLT hold.

However, there is too much dependence for Theorem 2.2.4 to imply a jointly normal limiting distribution. In particular, Assumption 2.2.3 does not hold.

The unified affinity set is $\tilde{\mathcal{A}}_i = \{i\}$ and the joint covariance matrix Ω_n captures within-observation variance. Because $X_i, X_{i+1}, \nu_{i,1}$, and $\nu_{i,2}$ are mutually independent standard normal variables, Ω_n is purely diagonal:

$$\Omega_n = \sum_{i=1}^{n} \begin{pmatrix} \mathrm{Var}(X_i + \nu_{i,1}) & 0 \\ 0 & \mathrm{Var}(X_{i+1} + \nu_{i,2}) \end{pmatrix} = \begin{pmatrix} 2n & 0 \\ 0 & 2n \end{pmatrix}.$$

Consequently, the minimal eigenvalue is $\lambda_{\min}(\Omega_n) = 2n \to \infty$.

However, we must evaluate the outside-affinity influence bound (Assumption 6), which for singletons requires the sum of cross-observation covariances to be $o(\lambda_{\min}(\Omega_n))$. Due to the ring structure, each dimension of Z_i is correlated with exactly one adjacent variable

outside its affinity set: $Z_{i,1}$ covaries exclusively with $Z_{i-1,2}$ (sharing X_i), and $Z_{i,2}$ covaries exclusively with $Z_{i+1,1}$ (sharing X_{i+1}). Both covariances evaluate to 1.

Aggregating this outside influence across all n observations yields:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{d_1,d_2=1}^{2}|\mathrm{Cov}\left(Z_{i,d_1},Z_{j,d_2}\right)| &= \sum_{i=1}^{n}\left(|\mathrm{Cov}\left(Z_{i,1},Z_{i-1,2}\right)|+|\mathrm{Cov}\left(Z_{i,2},Z_{i+1,1}\right)|\right)\\ &= \sum_{i=1}^{n}(1+1)=2n.\end{aligned}$$

Assumption 2.2.3 requires $2n = o(\lambda_{\min}(\Omega_n)) = o(2n)$, which does not hold.

2.D Additional conditions to bridge the gap between marginal and joint normality

To ensure the univariate conditions imply the multivariate assumptions, we define the maximal marginal variance normalizer $\Omega_{\max} = \max_d \Omega_{n,d}$ and introduce three structural constraints:

Assumption 2.D.1 (Strict Eigenvalue Lower-Bounding)**.** There exists $C_1 > 0$ such that $\lambda_{\min}(\Omega_n) \geq C_1\Omega_{\max}$.

Assumption 2.D.2 (Proportional Cross-Moments)**.** For all dimensions d_1, d_2, d_3, d_4, there exists $C > 0$ such that joint moments are dominated by their marginal counterparts:

1. $\sum_{i=1}^{n}\mathbb{E}\left[|Z_{i,d_1}|\left(\sum_{j\in\tilde{\mathcal{A}}_i}Z_{j,d_2}\right)^2\right] \leq C\max_d\sum_{i=1}^{n}\mathbb{E}\left[|Z_{i,d}|\left(\sum_{j\in\mathcal{A}^n_{i,d}}Z_{j,d}\right)^2\right]$
2. $\sum_{i,j=1}^{n}\sum_{k\in\tilde{\mathcal{A}}_i,l\in\tilde{\mathcal{A}}_j}|\mathrm{Cov}(Z_{i,d_1}Z_{k,d_2},Z_{j,d_3}Z_{l,d_4})| \leq C\max_d\sum_{i,j=1}^{n}\sum_{k\in\mathcal{A}^n_{i,d},l\in\mathcal{A}^n_{j,d}}|\mathrm{Cov}(Z_{i,d}Z_{k,d},Z_{j,d}Z_{l,d})$
3. $\sum_{i=1}^{n}\mathbb{E}\left[|Z_{i,d_1}|\cdot|\mathbb{E}[Z_{i,d_2}\mid Z_{-i}]|\right] \leq C\max_d\sum_{i=1}^{n}\mathbb{E}\left[|Z_{i,d}|\cdot|\mathbb{E}[Z_{i,d}\mid Z_{-i,d}]|\right]$

Assumption 2.D.3 (Uniform Moment Control)**.** $\mathbb{E}\left[|Z_{i,d_1}Z_{j,d_2}Z_{k,d_3}|\right] \leq C_3\max_d\mathbb{E}\left[|Z_{i,d}Z_{j,d}Z_{k,d}|\right]$.

Theorem 2.D.4. *Assumptions 2.A.1, 2.A.2, and 2.A.3 together with Assumptions 2.D.1, 2.D.2, and 2.D.3 imply Assumptions 2.2.1, 2.2.2, and 2.2.3.*

Proof. First we show that Assumption 2.2.1 holds.

$$\sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\| \left\|\sum_{j\in\tilde{\mathcal{A}}_i} Z_j\right\|^2\right] \leq p \sum_{d_1,d_2=1}^{p} \sum_{i=1}^{n} \mathbb{E}\left[|Z_{i,d_1}| \left(\sum_{j\in\tilde{\mathcal{A}}_i} Z_{j,d_2}\right)^2\right]$$

By Assumption 2.D.2.1, this is bounded by $p^3 C \max_d \sum_{i=1}^n \mathbb{E}\left[|Z_{i,d}| \left(\sum_{j\in\mathcal{A}^n_{i,d}} Z_{j,d}\right)^2\right]$. Assumption 2.A.1 ensures this marginal sum is $o((\Omega_{\max})^{3/2})$. Applying Assumption 2.D.1 evaluates the bound to $o(\lambda_{\min}(\Omega_n)^{3/2})$.

To see that Assumption 2.2.2 holds, expanding the quadruple sum over p^4 dimensional combinations and applying Assumption 2.D.2.2 bounds the expression by:

$$p^4 C \max_d \sum_{i,j=1}^{n} \sum_{k\in\mathcal{A}^n_{i,d}, l\in\mathcal{A}^n_{j,d}} |\mathrm{Cov}(Z_{i,d}Z_{k,d}, Z_{j,d}Z_{l,d})|$$

By Assumption 2.A.2 and 2.D.1, this is $o((\Omega_{\max})^2) = o(\lambda_{\min}(\Omega_n)^2)$.

Lastly, consider Assumption 2.2.3.

$$\sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\| \cdot \|\mathbb{E}[Z_i \mid Z_{-i}]\|\right] \leq \sum_{d_1,d_2=1}^{p} \sum_{i=1}^{n} \mathbb{E}\left[|Z_{i,d_1}| \cdot |\mathbb{E}[Z_{i,d_2} \mid Z_{-i}]|\right]$$

Applying Assumption 2.D.2.3 bounds this by $p^2 C \max_d \sum_{i=1}^n \mathbb{E}\left[|Z_{i,d}| \cdot |\mathbb{E}[Z_{i,d} \mid Z_{-i,d}]|\right]$. By Assumptions 2.A.3 and 2.D.1, this evaluates to $o(\Omega_{\max}) = o(\lambda_{\min}(\Omega_n))$. ■

Chapter 3

AN OPTIMIZATION FRAMEWORK FOR FOCAL NODE SELECTION IN INTERFERENCE DETECTION

Conditional randomization tests provide a finite-sample valid approach for detecting interference in network experiments, but their power can depend substantially on the choice of focal units whose treatment assignments are held fixed. We study the problem of selecting focal units to maximize power. For a regression-based test statistic under local network interference, we derive an asymptotic characterization of power for dependent network data and show that, under local alternatives, focal-set quality is governed by the residualized variation in exposure induced by rerandomizing the auxiliary units. This characterization yields a principled optimization objective; we develop mixed-integer quadratic programming formulations and semidefinite and second-order cone relaxations to select the focal set. The framework also clarifies the roles of previously proposed independent-set and focal–auxiliary edge heuristics, identifying the aspects of the power objective that each captures and omits. In experiments across several network models, direct optimization of the proposed criterion yields substantial power gains over random selection and commonly used greedy heuristics.

3.1 Introduction

Network interference is the setting where a unit's treatment assignment or intervention may affect the outcomes of its neighbors in a network. In settings where there is significant interference between experimental units, it is often necessary for experimenters to run an experimental design more sophisticated than a unit-level Bernoulli randomization, to mitigate the bias introduced by the interference. This requires a significant amount of effort, and incurs some cost in terms of variance. Thus, to decide if such a design is even necessary, the problem of detecting interference is an important one. This problem has been studied

before, for example, in Aronow [2012], Athey et al. [2018].

For a concrete instance of the problem, consider a friendship network with a presumed 1-neighborhood interference structure. That is, if there is interference, it is assumed to occur only through one's immediate friends in the network, such that one's outcome from a "treatment" roll-out is through their own treatment assignment, and their immediate friends'. In Aronow [2012], Athey et al. [2018], the authors propose a conditional treatment resampling framework to test the hypothesized interference for every unit in the conditioning set. In particular, they propose fixing the treatment assignments of select "focal" units, and resampling the treatment assignments of the others, called the "auxiliary" units. Then they look at a test statistic, such as a regression coefficient of the outcomes of the focal units on the treatment assignments of the auxiliary units. Repeating this procedure of resampling the treatment assignments of the auxiliary units many times would lead to a distribution of test statistics. From this, we can compute exact p-values of the originally observed test statistic, in comparison to the test-statistics from the resampled treatment assignments. Intuitively, if there were no interference, resampling the treatment assignments of the auxiliary units should not drastically change the test statistic, and thus the p-value calculated will be large. Conversely, if there is interference, resampling the treatment assignments of the auxiliary units would lead to a "scrambling" of the original signal behind the test-statistics, leading to a small p-value. Figure 3.1.1 illustrates this process. This testing procedure conditional on the focal units selected takes a particular form of conditional randomization tests.

In Athey et al. [2018], the authors briefly consider heuristics for focal units selection depending on the hypothesized interference structure. For example, they separately consider forming independent sets and maximizing focal-auxiliary edges to select the focal units. However, the motivation for these approaches have not been formally established. In fact, power characterizations in this literature have been scarce, with the closest being in Puelz et al. [2022], where the authors establish a lower bound for their biclique selection procedure under certain assumptions, and in Katsevich and Ramdas [2022], where the authors build

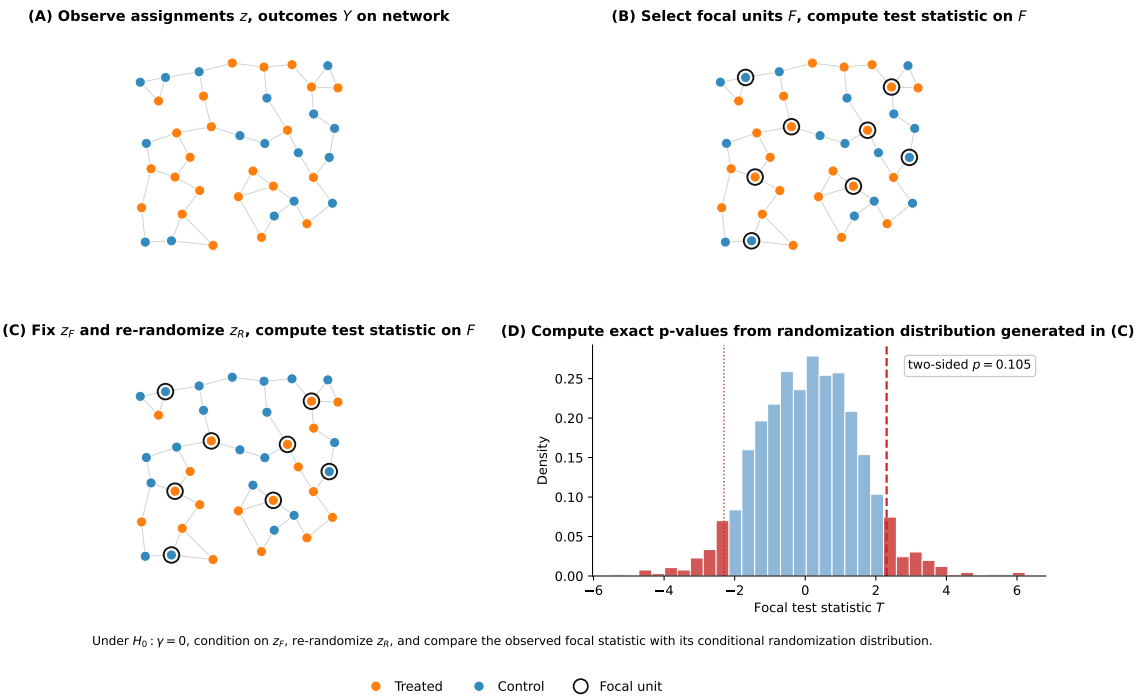


Figure 3.1.1: Conditional randomization test of network interference. The orange and blue colors of the network nodes depict different treatment assignments under i.i.d. Bernoulli(0.5).

upon the Neyman-Pearson lemma for the i.i.d. setting in conditional randomization testing against local semiperimetric alternatives. Even in the adjacent literature of knockoffs [Candes et al., 2018], power characterizations is limited. In this paper, we derive an asymptotic characterization of power for dependent network data and show that this characterization yields a principled optimization objective. We develop mixed-integer quadratic programming formulations and semidefinite and second-order cone relaxations as focal-unit selection methods with the goal of rigorously maximizing statistical power. In doing so, we establish precise connections to heuristics proposed in Athey et al. [2018].

Moreover, the optimization-based formulation is flexible and can be extended to accommodate additional features of networked settings, such as homophily and heterogeneous variation; related formulations are studied in Thiyageswaran et al. [2026a]

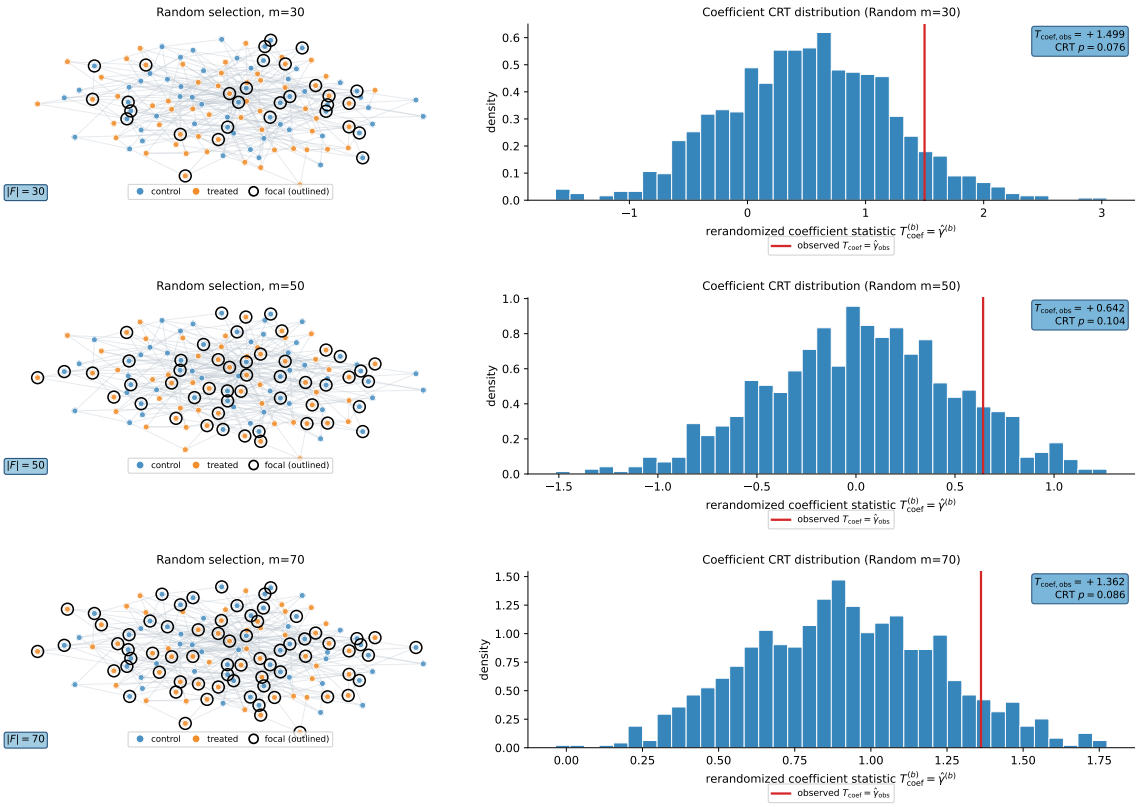


Figure 3.1.2: Left column: Focal units randomly selected from a graph with 120 nodes. Right column: null distribution generated from the resampling procedure. The rows (from top to bottom) correspond to focal units sizes of 30, 50, and 70, respectively. The red vertical line on the histogram represents the observed test statistic restricted to the focal set.

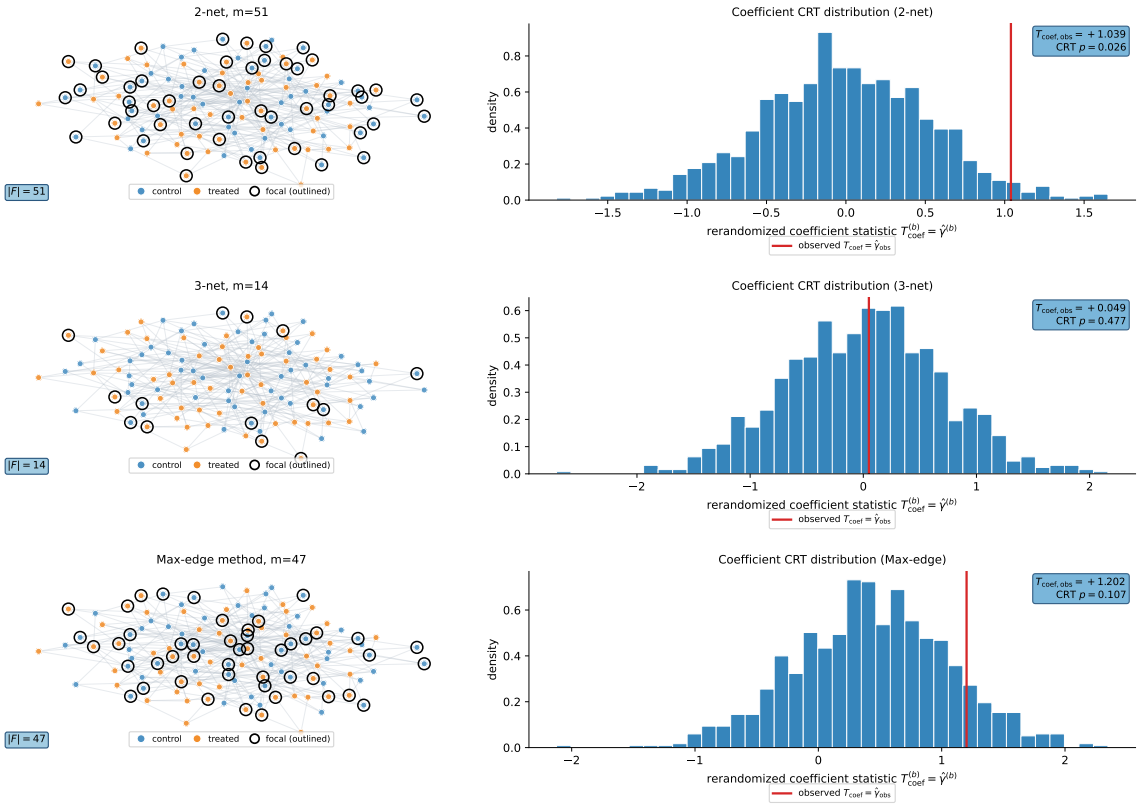


Figure 3.1.3: Left column: Focal units selected from a graph with 120 nodes. Right column: null distribution generated from the resampling procedure. The rows (from top to bottom) correspond to focal units selected according to greedy 2-net, 3-net, and max-edge methods, respectively. The 2-net and the max edge method correspond to heuristics proposed by Athey et al. [2018]. The red vertical line on the histogram represents the observed test statistic restricted to the focal set.

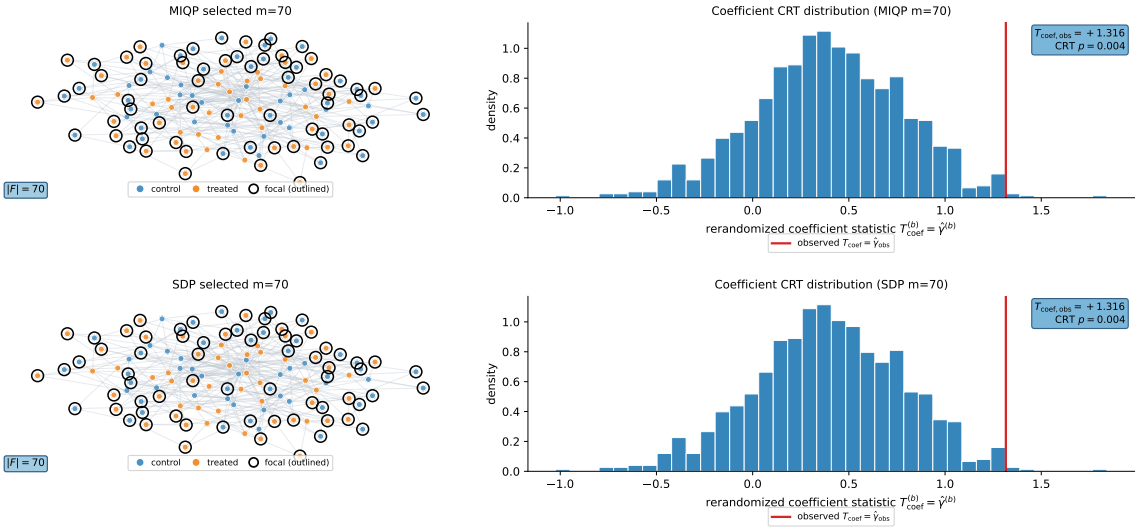


Figure 3.1.4: Left column: Focal units selected from a graph with 120 nodes. Right column: null distribution generated from the resampling procedure. The rows correspond to focal units selected by different optimization procedures discussed in Section 3.4. The red vertical line on the histogram represents the observed test statistic restricted to the focal set.

3.2 Conditional randomization tests and power

3.2.1 Null Hypothesis

Let Y denote the vector of outcomes, and $\mathbb{Z} = \{0,1\}^n$ be the set of possible treatment assignments. For each i,

$$Y_i(z) = Y_i(z'), \tag{3.2.1}$$

for all $z, z' \in \mathbb{Z}$, where $z_i = z'_i$.

We would like to find exact p-values, while maximizing statistical power. To find exact p-values with respect to the interference term, we run a conditional randomization test, conditioning on the focal units whose treatment assignments we fix. This allows us to test the sharp null hypothesis in Equation (3.2.1) restricted to focal units. See Figure 3.1.1 for an illustration of this idea. To maximize statistical power, we would like to optimally pick

the focal units. Figures 3.1.2 to 3.1.4 illustrate how selecting different focal units can affect the statistical power of the tests.

3.2.2 Potential outcome model

For the sake of simplicity, we focus on the setting of the following linear outcome model for most of this paper:

$$Y_i = \alpha + \beta z_i + \gamma_{n,i} e_i + \varepsilon_i, \tag{3.2.2}$$

where $z_i \in \{0, 1\}$ are the treatment assignments, e_i are the fraction of treated neighbors, and ε_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, for $i = 1, 2, ..., n$. We would like to test if there is interference:

$$H_0 : \gamma_{n,i} = 0 \quad \text{for all } i \tag{3.2.3}$$

3.2.3 Test statistic

Motivated by Athey et al. [2018], Wang and Janson [2022] and its practical simplicity, we focus on the regression coefficient statistic, $\hat{\gamma_n}$.

3.2.4 Robust Validity

Our procedure is conditionally valid. This simply means that under the appropriate conditioning, the test has the correct level, i.e., the Type-I error rate is equal to α under the null. In our case, the conditioning event is simply the focal selection procedure and the observed treatment assignments. See Aronow [2012], Athey et al. [2018] for more details.

We note that this conditional randomization testing procedure results in valid tests that are robust to model-misspecifications. Indeed, even if the true outcome model were not linear, the tests remain valid. What we may lose is power, since the focal selection procedure leverages the model structure.

3.2.5 Conditional sharpness

In this paper, we focus on focal selection procedures so that we can test the sharp null conditional on the focal units. The null in Equation (3.2.1) is not globally sharp. With direct effects, one cannot impute the outcomes of a unit under a global rerandomization. However, with a restricted rerandomization procedure so that the treatment assignments of the focal units are fixed, the test restricted to the focal units becomes sharp. We refer the reader to Aronow [2012], Athey et al. [2018] for more details. Then, we can infer the randomization distribution of the test-statistic for this new sharp null Fisher [1950], Rosenbaum [1984, 2002, 2007].

3.2.6 Comparisons to related work

In this work, we provide rigorous motivations for heuristics proposed in Athey et al. [2018]. In particular, our general framework provides an lens through which to understand why their proposed greedy implementations of the independent set and focal-auxiliary edge maximization heuristics may perform well in certain settings. Basse et al. [2019] formalize the conditioning mechanism and test validity based on flexible conditioning for conditional randomization testing frameworks. As testing for interference falls under this conditional randomization testing framework, their formalization allows focal set selection, for the problem of interference detection, to depend on the original treatment assignments, if properly accounted for. In Puelz et al. [2022], the authors propose generating an auxiliary null exposure bipartite graph as a function of the exposure model, the original interference network, the null of interest, and the treatment assignment mechanism. The bipartite graph consists of the units/nodes on one side and treatment assignment sets on the other. The focal set is then selected as the nodes of the largest biclique in this bipartite graph, with the aim of increasing statistical power. The authors demonstrate a lower bound of the power in terms of the magnitude of the size of the biclique under a certain set of assumptions on the distribution the design and null, and the outcome model. Additionally, this approach is

more suited for particular exposure models. In particular, for more general exposure models, the null exposure graph very quickly becomes a complete bipartite graph. For example, the fractional neighborhood exposure model we focus on paired with the null Equation (3.2.1) we care about we focus on would lead to a complete bipartite graph, unless we changed the exposure model to a thresholded one (see, for example, Thiyageswaran et al. [2026b]). Moreover, constructing the bipartite graph, as well as finding the largest biclique, with the largest possible number of edges in the remaining null exposure graph can be computationally expensive. In fact, the latter is NP-hard. The authors propose a binary inclusion-maximum biclustering approach as an approximation.

Unlike our work, the biclique approach in Puelz et al. [2022] also does not account for exposure variation that is uninformative to the test statistic. Yanagi and Sei [2022] proposed an improvement to the biclique approach for a special case of the difference-in-means estimator, and a uniform exposure variation parameter. Our approach is much more general than this. Yanagi and Sei [2022] build on the power analysis in the difference-in-means estimator in i.i.d. settings by Krieger et al. [2020].

Other related works include Hoshino and Yanagi [2023] where the authors consider the problem of testing correct specifications of exposure mappings. Basse et al. [2024] propose randomization-based stratified permutation tests to account for group formation designs. In Wang and Janson [2022], the authors derive asymptotic power for CRTs for various test statistics in a high-dimensional linear model in the i.i.d. setting. Guo et al. [2025] focuses on ML models to capture the outcome models with and without the effects of interest to improve power, and test-statistics based on the discrepancies between the cross-validation errors from the models above in CRTs. Our work focuses on the subsequent focal selection optimization problem for power maximization. Other model-based approaches include Gao et al. [2026], Bowers et al. [2013].

Another thread of the interference testing literature Pouget-Abadie et al. [2019], Baird et al. [2018], Toulis and Kao [2013] focuses on design-based frameworks to effectively test

for particular effects. Adjacent to the CRT literature is the model-x knockoffs one (see, for example, Candes et al. [2018]). However, it is much more tricky to work with dependent structures like networks in the knockoffs framework.

3.2.7 Notation

Denote the sequence of graphs by $G^{(n)} = (V^{(n)}, E^{(n)})$, where $V^{(n)}$ is the sequence of vertex sets, and $E^{(n)}$ is the sequence of edge sets. Let W be the edge weights matrix, and D denote the diagonal degree matrix of the network. That is, $D_{ii} = \sum_j W_{ij}$. Define $B = D^{-1}W$ and thus, $e_i = \sum_j B_{ij} z_j$. Decompose $z = \mu + \tilde{z}$, where $\mu := \mathbb{E}[z] = p\mathbf{1}$. Throughout, we use F and R to index the focal and auxiliary sets, respectively. $M_F = I - X_F(X_F^T X_F)^{-1} X_F^T$ is the projection matrix, partialling out the unit-level treatment assignments and the constant vector, with $X_F = [\mathbf{1} \quad z_F]$.

Throughout this paper, we focus on i.i.d. Bernoulli randomized experimental designs for simplicity. Our general framework should apply to any initial experimental design. We also focus on upper-tail tests to illustrate the key ideas. All arguments should extend to lower-tail and two-sided tests.

3.3 Power under asymptotic analysis

First we list a series of assumptions we may make use of in our analysis.

Assumption 3.3.1 (Bernoulli randomization)**.** The treatment assignments are i.i.d. Bernoulli(p), with $p \in (0, 1)$, bounded away from zero and one.

Assumption 3.3.2 (Bounded degree)**.** The graph sequence has uniformly bounded degree:

$$\max_{i \in V_n} d_i \leq d < \infty.$$

Assumption 3.3.3 (No isolated nodes)**.** There are no isolated nodes: $d_i > 0$ for all $i \in V^{(n)}$.

Assumption 3.3.4 (Bounded weights)**.** For every edge $(i, j) \in E^{(n)}$,

$$b_{\min} \leq B_{ij} \leq b_{\max}.$$

Assumption 3.3.5 (Local asymptotics)**.**

$$\frac{\gamma_n}{\sigma} = o(1),$$

and that $\gamma_n < \infty$.

Assumption 3.3.6 (Boundary mass)**.** Every focal unit has at least one auxiliary neighbour. That is, for all $i \in F$, $\sum_{j \in R} B_{ij} > 0$.

Assumption 3.3.7 (Variance order)**.** Define $C_M(s) = \mathbb{E}[\tilde{z}_R^T B_{RF} M_F B_{FR} \tilde{z}_R]$, $L_F(s) = M_F(B_{FF} z_F + B_{FR} \pi_R)$, and $\Sigma_R = M_F B_{FR} \tilde{z}_R \tilde{z}_R^T B_{RF}^T M_F$. Then,

$$L_F(s)^T \Sigma_R L_F(s) + \operatorname{Tr}(\Sigma_R^2) = O\left(C_M(s)\right). \tag{3.3.1}$$

Theorem 3.3.8 (Asymptotic power optimality of $C_M(F)$)**.** *Let $\mathcal{F}_n$ be a class of proper focal sets $F \subsetneq V^{(n)}$. Suppose Assumptions 3.3.1 to 3.3.5 hold. Assume $\gamma_n \geq 0$, corresponding to the upper-tail test. For each $F \in \mathcal{F}_n$, let*

$$C_M(F) := p(1-p) \operatorname{Tr}\left(M_F B_{FR} B_{FR}^\top M_F\right),$$

and let

$$F^\star \in \arg\max_{F \in \mathcal{F}_n} C_M(F).$$

Then $F^\star$ satisfies Assumption 3.3.6 with probability tending to one. Moreover, Assumption 3.3.7 holds for $F^\star$. If

$$\frac{\gamma_n \sqrt{C_M(F^\star)}}{\sigma} = O_P(1),$$

then

$$\text{Power}(F^\star) = 1 - \Phi\left(c_{1-\alpha} - \frac{\gamma_n\sqrt{C_M(F^\star)}}{\sigma}\right) + o_P(1), \tag{3.3.2}$$

and

$$\text{Power}(F^\star) = \max_{F\in\mathcal{F}_n} \text{Power}(F) + o_P(1). \tag{3.3.3}$$

If instead

$$\frac{\gamma_n\sqrt{C_M(F^\star)}}{\sigma} \to \infty,$$

then

$$\text{Power}(F^\star) \to 1,$$

and (3.A.2) *again follows. Thus, maximizing* $C_M(F)$ *is asymptotically sufficient for maximizing the power of the conditional randomization test.*

Thus, we maximize $C_M(s)$ over all $s \in \{0,1\}^n$. We find the best focal set over the grid of candidate focal size m.

Theorem 3.3.9 (Central Limit Theorem)**.** *Suppose Assumptions 3.3.1 to 3.3.4 and 3.3.6 hold, with* $\gamma_n = \gamma > 0$. *Define* $D_m := \|L_F(s)\|^2 + \mathbb{E}[\tilde{z}_R^T B_{RF} M_F B_{FR}\tilde{z}_R]$, *where* $L_F(s) = M_F(B_{FF}z_F + B_{FR}\pi_R)$. *Let* $r = M_F B_{FR}\tilde{z}_R$, *and* $\Delta_o = \frac{r^T Y}{D_m}$. *Then,*

$$\Delta_o \Rightarrow \mathcal{N}(\frac{\gamma C_M(s)}{D_m}, V_m),$$

where

$$V_m = \frac{1}{D_m^2}\left[\sigma^2 C_M(s) + \gamma^2 \,\text{Var}(r^T L_F + \|r\|^2)\right]. \tag{3.3.4}$$

This results in the following corollary.

Corollary 3.3.10. *Suppose Assumptions 3.3.1 to 3.3.6 hold. Then,*

$$\Delta_o \Rightarrow \mathcal{N}(\frac{\gamma_n C_M(s)}{D_m}, V_m),$$

where

$$V_{m,n} = \frac{1}{D_m^2}\left[\sigma^2 C_M(s) + \gamma_n^2 \operatorname{Var}(r^T L_F + \|r\|^2)\right]. \tag{3.3.5}$$

If, in addition, Assumption 3.3.7 holds,

$$V_{m,n} = \frac{1}{D_m^2}\left[\sigma^2 C_M(s) + \gamma_n^2 \operatorname{Var}(r^T L_F + \|r\|^2)\right] = \frac{\sigma^2 C_M(s)}{D_m^2}\{1 + o(1)\},$$

and therefore,

$$\frac{\Delta_o - \gamma_n C_M(s)/D_m}{\sigma\sqrt{C_M(s)}/D_m} \Rightarrow \mathcal{N}(0,1)$$

Lemma 3.3.11 (Asymptotic monotonicity in focal gains)**.** *Suppose Assumptions 3.3.1 to 3.3.3 holds. Let F contain a focal unit v that has no auxiliary neighbors. Let $F^- = F \setminus v$. Then,*

$$C_M(F^-) - C_M(F) = c - o_{\mathbb{P}}(1),$$

for some constant $c > 0$.

3.4 Optimal focal units selection

To make the problem tractable, we solve the optimization problems for fixed numbers of focal units, $|\mathcal{F}| = m$. Then, we choose the solutions across m that maximize our objective function. Let $\mathcal{M}$ denote our candidate numbers of focal units.

Thus, for $m \in \mathcal{M}$, the optimization problem we seek to solve is

$$\begin{aligned} \max_{s \in \mathbb{R}^n} \quad & C_M(s) \\ \text{s.t.} \quad & s_i \in \{0,1\}, \quad i = 1, 2, ..., n, \\ & \mathbf{1}^T s = m. \end{aligned} \tag{3.4.1}$$

Here, $C_M(F) = \mathbb{E}[\tilde{z}_R^T B_{RF} M_F B_{FR} \tilde{z}_R]$.

We solve these problems by formulating Mixed-integer quadratic programs (MIQP), Second-order conic program relaxations (SOCP), and semidefinite programs (SDP) relaxations. See, for example, Boyd and Vandenberghe [2004] for details on each of these.

3.4.1 Max focal-auxiliary edges

As an exercise in illustrating some ideas, we replace M_F with I. Thus, the optimization problem becomes the following, resembling MAXCUT.

$$\begin{aligned} \max_{s \in \mathbb{R}^n, t \in \mathbb{R}} \quad & s^T (B \circ B)(1 - s) \\ \text{s.t.} \quad & s_i \in \{0, 1\}, \quad i = 1, 2, ..., n, \\ & \mathbf{1}^T s = m. \end{aligned} \tag{3.4.2}$$

Lifting and SDP relaxation

$$\begin{aligned} \max_{s \in \mathbb{R}^n, t \in \mathbb{R}} \quad & s^T (B \circ B)(1 - s) \\ \text{s.t.} \quad & 0 \leq s_i \leq 1, \quad i = 1, 2, ..., n, \\ & \mathbf{1}^T s = m. \end{aligned}$$

In fact, one of the two main approaches proposed by Athey et al. [2018] is this, which they call edge comparison maximization. Thus, we formalize the approach from Athey et al. [2018] for selecting focal units. For computational reasons, they propose a greedy approximation; they begin with all units assigned to the auxiliary set. Then, at each iteration, they add a unit to the focal set if it results in the biggest gain in the number of edges between that unit and the remaining nodes, i.e., the remaining auxiliary units. They stop when there is no longer such a gain.

3.4.1.1 Comparison to the sensor selection problem

We note that this formulation also bears some resemblance to the sensor selection problem studied in Joshi and Boyd [2008]:

$$\max_{s} \quad \sum_{i} s_i e_i^2 \tag{3.4.3}$$

$$\text{s.t.} \quad s_i \in \{0,1\} \quad i = 1,2,...,n \tag{3.4.4}$$

$$\mathbf{1}^T s = m, \tag{3.4.5}$$

where s is the indicator over the sensors. Relaxing the binary constraints so that $w_i \in [0,1]$ for all $i = 1,2,...,n$, gives us a convex optimization problem. The largest m values of the solution $w^\star$ can be taken to correspond to the focal units. The difference here is the contrast vectors s and $1-s$ we would consider.

3.4.1.2 Independent sets

The other approach proposed by Athey et al. [2018] is the construction of maximal independent sets. They construct a 2-net greedily. They start with empty focal and auxiliary sets. At every iteration, they randomly select a node in neither sets and assign it as a focal unit. All of its neighbours are assigned as auxiliary units. This is continued until no nodes are left. Under Assumption 3.3.1, observe that the L_F term in Equation (3.3.5) vanishes for every n under an independent set because $B_{FF} = 0$ under the independent set. Asymptotically, the null re-randomization distribution becomes centered. For finite n, if $p = 1/2$, then, the null re-randomization distribution is also centered. See Figure 3.4.1 for an illustration of this phenomenon. This motivates the use of an independent set, or a 2-net.

A curious reader might ask if the 2-net is sufficient, when the 3-net maybe better suited to targeting covariance minimization of the exposures of the focal units, since under the 3-net, the covariance between the exposures of any pair of focal units is zero, under the 1-hop neighbourhood exposure model. It turns out that the 3-net can be too strong a restriction,

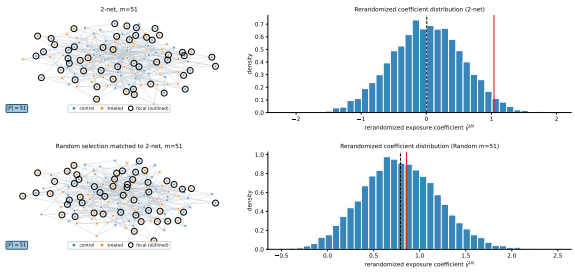


Figure 3.4.1: The null rerandomization distribution under the 2-net is centered around zero.

as the number of focal units under the 3-net can be significantly fewer than from the 2-net.

Moreover, it is not clear if a centered null is necessarily optimal. We should be able to trade-off this bias in the null distribution for some reduction in variance.

Remark 3.4.1 (Greedy selection and submodularity). The focal-selection criteria considered here do not, in general, fall into the class of monotone submodular maximization problems for which classical greedy approximation guarantees apply. For example, the focal–auxiliary edge objective in Equation (3.4.6) is a weighted cut function and is therefore submodular, but it is non-monotone: adding a unit to the focal set may remove more focal–auxiliary edges than it creates. Once the centering or residualization terms relevant for statistical power are incorporated, the resulting objectives $C_H(F)$ and $C_M(F)$ need not retain even this submodular structure, since the marginal contribution of a candidate focal unit depends jointly on the other selected focal units through the centering or projection term. Consequently, the greedy focal-selection procedures discussed above should be viewed as computational heuristics, and do not inherit the standard approximation guarantees for greedy maximization of a monotone submodular objective.

In the following subsection, we now retain the full residualization matrix M_F.

3.4.2 Fisher information and Schur complement

By the Frisch-Waugh-Lovell theorem Frisch and Waugh [1933], Lovell [1963] (see also, Yule [1907]), $e_F^T M_F e_F$ is the Fisher information for γ after partialling out the nuisance parameters

η. Notice that under an independent focal set, as $m \to \infty$, $e_F^T M_F e_F \approx C_M(F)$. For this setting, we can write our objective function in terms of the Schur complement of $I_{\eta\eta}(s)$, where $\eta = (\alpha, \beta)$. Indeed, notice that,

$$e_F^T M_F e_F = I_{\gamma|\eta}$$

The Schur complement $I_{\gamma|\eta}$ of $I_{\eta\eta}$ in the information matrix I is,

$$I_{\gamma|\eta} = I_{\gamma\gamma} - I_{\gamma\eta} I_{\eta\eta}^{-1} I_{\eta\gamma}.$$

The components are $I_{\gamma\gamma} = e^T e$, $I_{\gamma\eta} = e^T X$, $I_{\eta\gamma} = X^T e$, and $I_{\eta\eta} = X^T X$. With the indicator set s, the components are now $I_{\gamma\gamma}(s) = e^T \operatorname{diag}(s) e$, $I_{\gamma\eta} = e^T \operatorname{diag}(s) X$, $I_{\eta\gamma} = X^T \operatorname{diag}(s) e$, and $I_{\eta\eta} = X^T \operatorname{diag}(s) X$. Thus, for the indicator set s of the focal units F,

$$e_F^T M_F e_F = I_{\gamma|\eta}(s) = I_{\gamma\gamma}(s) - I_{\gamma\eta}(s) I_{\eta\eta}^{-1}(s) I_{\eta\gamma}(s).$$

Finding an approximately optimal focal set then amounts to solving the following SDP relaxation:

$$\begin{aligned} \max_{s \in \mathbb{R}^n, t \in \mathbb{R}} \quad & t \\ \text{s.t.} \quad & s_i \in \{0,1\}, \quad i = 1, 2, ..., n, \\ & \begin{bmatrix} I_{\eta\eta}(s) & I_{\eta\gamma}(s) \\ I_{\gamma\eta}(s) & I_{\gamma\gamma}(s) - t \end{bmatrix} \succeq 0, \\ & 0 \leq s_i + s_j \leq 1 \quad \forall (i,j) \in E^{\varepsilon}, \\ & \mathbf{1}^T s = m. \end{aligned} \tag{3.4.6}$$

The solutions to this approach is a function of both the graph as well as the treatment assignments of the focal units. Thus, for valid tests, when generating the treatment assign-

ments on the auxiliary units, we should condition on the selection procedure of the focal units [Puelz et al., 2022]. In practice, one would implement this via a rejection sampling procedure, which would be computationally inefficient. In the following subsection, we leverage the concentration of treatment assignments on the focal subset to run easier CRTs.

3.4.3 Constant Projections

Under Assumptions 3.3.1 and 3.3.2, $C_H(F) = C_M(F) + O(1)$, where the projection matrix M_F is replaced with $H_F := I - \frac{1}{m}\mathbf{1}\mathbf{1}^T$. Thus, under these assumptions, for large m, it suffices to solve the following optimization problem as an approximation. Then,

$$\begin{aligned} C_H(s) = \mathrm{Tr}(\Sigma_R) &= p(1-p)\,\mathrm{Tr}(H_F B_{FR} B_{FR}^T H_F) \\ &= p(1-p)\left[\mathrm{Tr}(B_{FR}B_{FR}^T) - \frac{1}{m}\mathbf{1}^T B_{FR}B_{FR}^T\mathbf{1}\right] \\ &= p(1-p)\left[\sum_{i\in F}\sum_{j\in R} B_{ij}^2 - \frac{1}{m}\sum_{j\in R}\left(\sum_{i\in F} B_{ij}\right)^2\right]. \end{aligned}$$

Under the constraint $0 \le s_i + s_j \le 1 \quad \forall (i,j) \in E^\varepsilon$, $B_{FF} = 0$, and this becomes

$$p(1-p)\left[s^T b - \frac{1}{m} s^T B B^T s\right],$$

where $b := \sum_j B_{ij}^2$.

Thus, we can formulate the following Mixed-Integer Quadratic Programming (MIQP):

$$\begin{aligned} \max_{s\in\mathbb{R}^n} \quad & s^T b - \frac{1}{m} s^T B B^T s \\ \text{s.t.} \quad & s_i \in \{0,1\}, \quad i = 1,2,...,n, \\ & 0 \le s_i + s_j \le 1 \quad \forall (i,j) \in E^\varepsilon, \\ & \mathbf{1}^T s = m. \end{aligned} \tag{3.4.7}$$

We can also formulate an SDP relaxation:

$$\begin{aligned} \max_{X \in \mathbb{R}^{n \times n}} \quad & s^T b - \frac{1}{m} \operatorname{Tr}(BB^T X) \\ \text{s.t.} \quad & 0 \leq s_i \leq 1, \quad i = 1, 2, ..., n, \\ & 0 \leq s_i + s_j \leq 1 \quad \forall (i,j) \in E^{\varepsilon}, \\ & X_{ii} = s_i, \\ & X_{ij} = 0 \quad \forall (i,j) \in E^{\varepsilon}, \\ & \begin{bmatrix} 1 & s \\ s & X \end{bmatrix} \succeq 0, \\ & \mathbf{1}^T s = m. \end{aligned} \tag{3.4.8}$$

In addition to that, we can carry out a greedy swapping algorithm, similar to Joshi and Boyd [2008].

3.4.3.1 Full Mixed-integer quadratic programming formulation

For $i, j \in [n]$, introduce the auxiliary variable $x_{ij} = s_i(1 - s_j)$, indicating that i is focal and j is auxiliary. Because s is binary, this product can be represented exactly using its envelopes. In addition, define

$$q_j = \sum_{i=1}^{n} B_{ij} x_{ij}, \qquad j \in [n].$$

The resulting mixed-integer quadratic program is

$$\begin{aligned}
\max_{s,x,q} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}^2 x_{ij} - \frac{1}{m}\sum_{j=1}^{n} q_j^2 \\
\text{s.t.} \quad & \mathbf{1}^\top s = m, \\
& q_j = \sum_{i=1}^{n} B_{ij} x_{ij}, && j \in [n], \\
& x_{ij} \le s_i, && i,j \in [n], \\
& x_{ij} \le 1 - s_j, && i,j \in [n], \\
& x_{ij} \ge s_i - s_j, && i,j \in [n], \\
& x_{ij} \ge 0, && i,j \in [n], \\
& s_i \in \{0,1\}, && i \in [n].
\end{aligned} \tag{3.4.9}$$

3.4.3.2 Full Second-order cone relaxation

We next relax the binary constraint to $s \in [0,1]^n$. Let X be the lifted representation of $ss^\top$, and define

$$r_{ij} = s_i - X_{ij},$$

so that $r_{ij} = s_i(1 - s_j)$ whenever $X = ss^\top$. Then, $X\mathbf{1} = ms$, which holds for $\mathbf{1}^\top s = m$.

Next, we introduce the variables t_j satisfying $t_j \ge q_j^2$. The second-order cone relaxation

is

$$
\begin{aligned}
\max_{s,X,r,q,t} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}^2 r_{ij} - \frac{1}{m}\sum_{j=1}^{n} t_j \\
\text{s.t.} \quad & 0 \le s_i \le 1, && i \in [n], \\
& \mathbf{1}^\top s = m, \\
& X = X^\top, \qquad \operatorname{diag}(X) = s, \qquad X\mathbf{1} = ms, \\
& X_{ij} \ge 0, && i,j \in [n], \\
& X_{ij} \le s_i, \qquad X_{ij} \le s_j, && i,j \in [n], \\
& X_{ij} \ge s_i + s_j - 1, && i,j \in [n], \\
& r_{ij} = s_i - X_{ij}, && i,j \in [n], \\
& q_j = \sum_{i=1}^{n} B_{ij} r_{ij}, && j \in [n], \\
& (t_j, 1/2, q_j) \in \mathcal{Q}_r^3, && j \in [n],
\end{aligned}
\tag{3.4.10}
$$

where

$$
\mathcal{Q}_r^3 = \left\{ (u,v,w) \in \mathbb{R}^3 : u \ge 0,\ v \ge 0,\ 2uv \ge w^2 \right\}
$$

denotes the second-order cone. In particular,

$$
(t_j, 1/2, q_j) \in \mathcal{Q}_r^3 \quad \Longleftrightarrow \quad t_j \ge q_j^2.
$$

3.4.3.3 Semidefinite programming relaxation

We can also formulate an SDP relaxation:

$$
\begin{aligned}
\max_{s,X,r,q,t} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}^2 r_{ij} - \frac{1}{m}\sum_{j=1}^{n} t_j \\
\text{s.t.} \quad & 0 \le s_i \le 1, && i \in [n], \\
& \mathbf{1}^\top s = m, \\
& X = X^\top, \qquad \operatorname{diag}(X) = s, \qquad X\mathbf{1} = ms, \\
& \begin{bmatrix} 1 & s^\top \\ s & X \end{bmatrix} \succeq 0, \\
& X_{ij} \ge 0, && i,j \in [n], \\
& X_{ij} \le s_i, \qquad X_{ij} \le s_j, && i,j \in [n], \\
& X_{ij} \ge s_i + s_j - 1, && i,j \in [n], \\
& r_{ij} = s_i - X_{ij}, && i,j \in [n], \\
& q_j = \sum_{i=1}^{n} B_{ij} r_{ij}, && j \in [n], \\
& \begin{bmatrix} t_j & q_j \\ q_j & 1 \end{bmatrix} \succeq 0, && j \in [n].
\end{aligned}
\tag{3.4.11}
$$

By the Schur complement, the final constraint is equivalent to $t_j \ge q_j^2$.

For $X = ss^\top$, and $r_{ij} = s_i(1 - s_j)$, the feasible regions satisfy

$$\mathcal{F}_{\text{integer}} \subseteq \mathcal{F}_{\text{SDP}} \subseteq \mathcal{F}_{\text{SOCP}}.$$

Therefore, the SDP relaxation provides an upper bound on the optimal value of the integer problem that is at least as tight as the upper bound provided by the SOCP relaxation.

3.5 Cycle graphs

In this section, we consider cycle graphs with equal edge weights. Recall that

$$C_H(F) = \sum_{i \in F, j \in R} B_{ij}^2 - \frac{1}{m} \sum_{j \in R} \left(\sum_{i \in F} B_{ij} \right)^2 .$$

Let t_j denote the number of focal units connected to auxiliary unit j. For graphs with equal edge weights w, this becomes,

$$C_H(F) = w^2 \sum_{j \in R} \left(t_j - \frac{t_j^2}{m} \right). \tag{3.5.1}$$

3.5.1 Minimum vertex covers are optimal when $n \geq 7$

A vertex cover is a set of vertices/nodes such that every edge has at least one endpoint in it. Minimum vertex covers are the smallest such sets. We formalize this below. For an even cycle graph, the minimum vertex cover is equal to the maximum independent set.

Proposition 3.5.1. *Let C_n be the equal edge-weighted cycle graph with $n \geq 7$. If n is even, C_H is maximized by the maximum focal independent set of cardinality $n/2$, i.e., the alternating focal sets. If n is odd, C_H is maximized by the focal sets with cardinality $(n+1)/2$ and are almost alternating: there is exactly one pair of adjacent focal nodes, and the remaining nodes alternate between focal and auxiliary units.*

In Figure 3.6.1, we see that the full optimization-based approaches perform as well as the optimization-based approaches with the independent set constraints, and the maximal independent set, i.e., alternating focal set, approach. We see that the greedy heuristics perform only slightly worse.

3.5.1.1 What happens when $n = 6$.

We first consider $m = 3$. Let's consider the alternating focal set, F_{alt}. Here, $c(F_{\text{alt}}) = 3$ and $a(F) = 3$. Thus, $C_H(F_{\text{alt}}) = \frac{2w^2}{3}(6 - 3) = 2w^2$.

Next, let us consider the focal set composed of two adjacent nodes, and one singleton, so that $c(F) = 2$, and $a(F) = 1$. In this case, $C_H(F_{\text{alt}}) = \frac{2w^2}{3}(2 \cdot 2 - 1) = 2w^2$, once again.

Let us now consider $m = 4$. In this case, we can have the following scenarios:

1. one contiguous focal block, with $c(F) = 1$, and $a(F) = 0$. Then, $C_H(F) = \frac{2w^2}{4}(3 \cdot 1) = \frac{3}{2}w^2$.

2. two contiguous focal blocks, with $c(F) = 2$, and $a(F) = 2$. Then, $C_H(F) = \frac{2w^2}{4}(3 \cdot 2 - 2) = 2w^2$.

Next, consider $m = 5$. In this case, we can have one contiguous focal block, and one singleton auxiliary unit. That is, $c(F) = 1$, and $a(F) = 1$. Then, $C_H(F) = \frac{2w^2}{5}(4 \cdot 1 - 1) = \frac{6}{5}w^2$.

Now we consider the cases where $m < 3$. We start with $m = 2$. We can have the following scenarios:

1. one contiguous focal block, so $c(F) = 1$, and $a(F) = 0$. Then, $C_H(F) = \frac{2w^2}{2}(1) = w^2$.

2. two non-adjacent focal units, so $c(F) = 2$, and $a(F) = 0$ (the case where $a(F) > 0$ would give a smaller C_H). Then, $C_H(F) = \frac{2w^2}{2}(2) = 2w^2$.

Finally, let us consider $m - 1$. Here, $c(F) = 1$, and $a(F) = 0$. Thus, $C_H(F) = 0$.

3.5.1.2 What happens when $n = 4$.

Proposition 3.5.2. *Let F_{alt} denote the maximal independent focal set, which is just the alternating focal set in the cycle graph. Let F_{adj} denote the focal set where $m = 2$, and the two focal units are adjacent to each other. Then,* $\max_F C_H(F) = C_H(F_{adj}) > C_H(F_{alt})$.

3.6 Experiments

In Figure 3.1.3, we see that the selection of focal units can significantly affect the null rerandomization distribution, and thus the power of our tests. As discussed in the previous

section, Figure 3.6.1 demonstrates how on a cycle graph, full optimization-based approaches perform as well as the optimization-based approaches with the independent set constraints, and the maximal independent set, i.e., alternating focal set, approach. In Figure 3.6.2, we see that the full optimization-based approaches perform much better than the heuristics, as well as the optimization-based approaches with the independent set constraints. Similar plots are observed for preferential-attachment graphs using the Barabasi-Albert model (see Figure 3.6.3), and clustered graphs generated from stochastic block models (see Figures 3.6.4 to 3.6.8).

The grid $\mathcal{M}$ used was 10 to 70, in increments of 10. To make the comparisons as fair as possible, we choose the m with the best (according to our C_H objective) for the greedy approaches as well as the random selection approach.

We generate treatments independently according to

$$z_i \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$$

and generate outcomes from

$$Y_i = \alpha + \beta z_i + \gamma(Bz)_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

We consider

$$\gamma \in \{0, 0.15, 0.30, 0.45, 0.60, 0.80, 1.00, 1.25, 1.50\}.$$

Table 3.1 provides the list of methods we compare in our experiments.

3.6.1 Application to a real social network

We additionally evaluate the focal-unit selection procedures on an observed social network from Cai et al. [2015]. Cai et al. [2015] study the diffusion of information about a weather-insurance product among rice farmers in rural China. Their experiment contains approximately 5,300 households across 185 villages. As part of the study, household heads

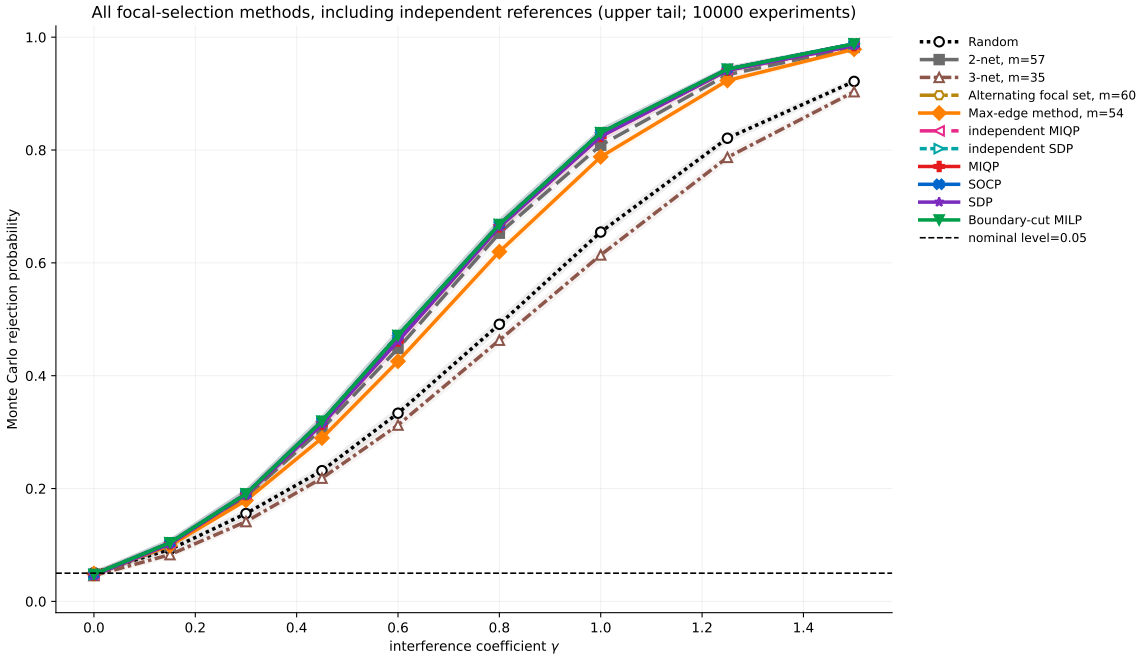


Figure 3.6.1: Power comparisons of all the methods across increasing γ for cycle graphs, with $n = 120$.

were asked to identify five close friends, within or outside their village, with whom they most frequently discussed rice production or financial matters. These nominations provide a natural empirical network on which to evaluate the focal-unit selection methods developed above.

Network construction. We use the natural village identified in the replication data as `dayudengxiongmin1`. We construct the network from the household survey file `0422survey.dta` and the corresponding network file `0422allinforawnet.dta`. We retain nominations for which both endpoints are surveyed households belonging to the selected natural village, remove self-nominations and duplicate directed nominations, and then form the undirected "general-link" network used in our analysis. In particular, an undirected edge $\{i, j\}$ is included whenever either household i nominated household j or household j nominated household i. We retain the observed network without artificially connecting

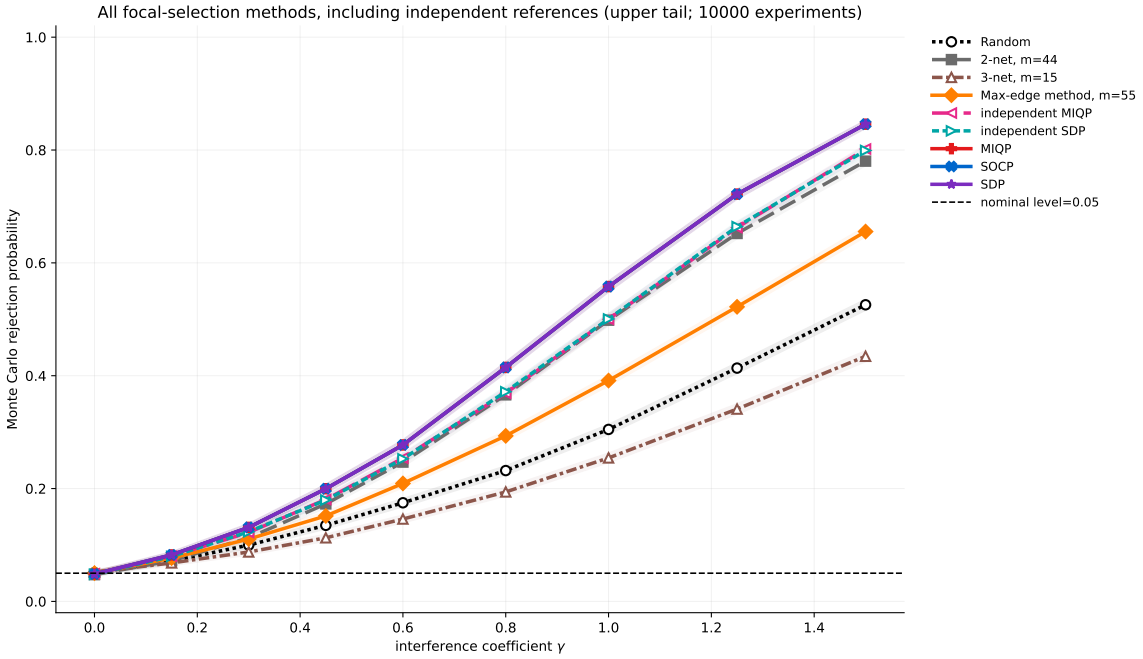


Figure 3.6.2: Power comparisons of all the methods across increasing γ for Erdös-Rényi graphs, with $n = 120$. Edge probability is 0.05.

separate components or removing isolated households. The resulting graph contains

$$n = 55 \qquad \text{and} \qquad |E| = 187$$

households and edges, respectively, with 1 connected component. The mean degree is 6.8, and the maximum degree is 14. See Figure 3.6.9, for a visualization of the underlying network.

Similar to the synthetic network settings, we compare the exact MIQP formulation, the SOCP and SDP relaxations followed by rounding, a uniformly selected random focal set, greedy 2- and 3-net constructions, and the max focal-auxiliary edge procedure of Athey et al. [2018]. For completeness, we also report the independent-set MIQP and SDP procedures, which impose the restriction that no two focal units are adjacent. For methods that optimize

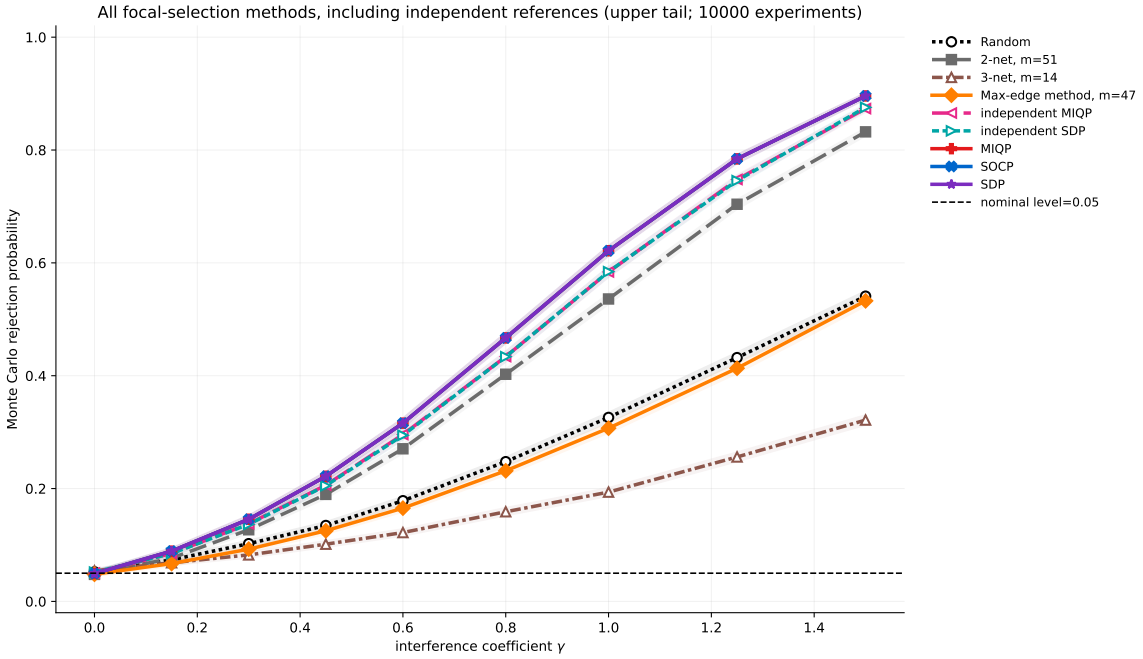


Figure 3.6.3: Power comparisons of all the methods across increasing γ for Barabasi-Albert graphs, with $n = 120$.

over focal-set cardinality, we search over

$$m \in \{10, 20, \ldots, 90\},$$

subject to feasibility, and retain the cardinality yielding the largest graph-only value of $C_H(F)$. Power is estimated using 10,000 independently simulated experiments and 1,000 auxiliary-treatment rerandomizations per experiment. The test is conducted at level $\alpha = 0.05$.

Figure 3.6.10 displays the average (over the 10000 experiments) rejection probability of the conditional randomization test as a function of the interference coefficient γ. At $\gamma = 0$, the rejection probabilities are close to the nominal 5% level across the methods. The optimization-based focal-selection procedures provide a clear improvement over all other focal selection procedures.

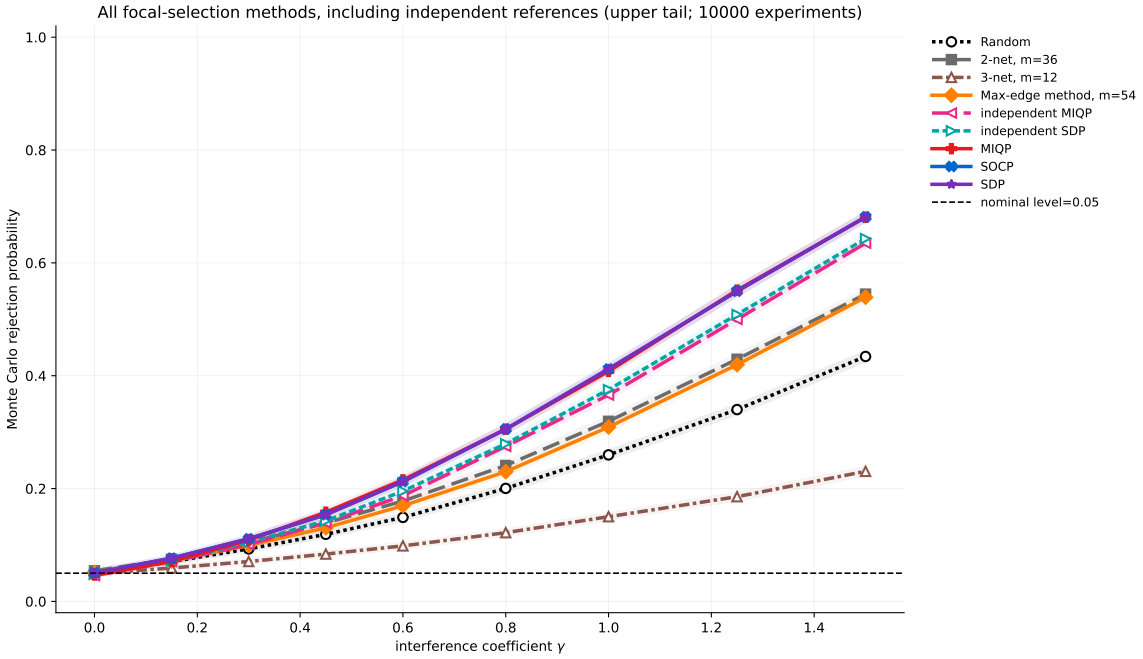


Figure 3.6.4: Power comparisons of all the methods across increasing γ for SBM graphs with two clusters, with $n = 120$. Within cluster connection probability is 0.1, and across cluster connection probability is 0.02.

3.7 Discussion and conclusion

This paper studies focal-unit selection for conditional randomization tests to detect network interference as an explicit statistical power optimization problem. Our asymptotic analysis shows that, under local alternatives, power is governed by the amount of residualized exposure variation induced on the focal units by rerandomizing the auxiliary units. Thus, a useful focal set must create sufficiently large exposure variation that remains informative after the nuisance components of the test statistic have been partialled out.

The constant-projection approximation $C_H(F)$ makes this trade-off particularly transparent. Up to the factor $p(1-p)$,

$$C_H(F) = \sum_{i \in F} \sum_{j \in R} B_{ij}^2 - \frac{1}{m} \sum_{j \in R} \left(\sum_{i \in F} B_{ij} \right)^2 .$$

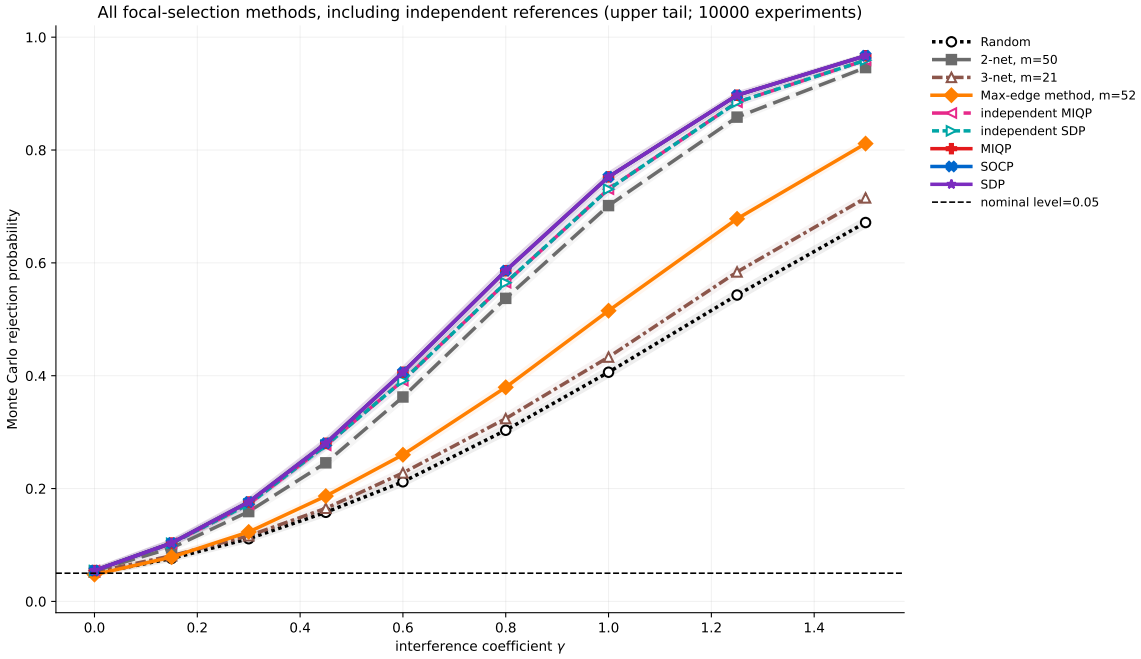


Figure 3.6.5: Power comparisons of all the methods across increasing γ for SBM graphs with four clusters of equal in size on average, with $n = 120$. Within cluster connection probability is 0.1, and across cluster connection probability is 0.02.

The first term rewards a large weighted boundary between focal and auxiliary units: rerandomization must be capable of changing the exposures of the focal units. The second term penalizes concentration of that variation through the same auxiliary units. This decomposition helps explain both the strengths and the limitations of existing heuristics. Focal–auxiliary edge maximization targets the first term but ignores the covariance correction. Independent-set constructions eliminate focal–focal adjacency and can center the rerandomization distribution in important settings, but independence can impose an unnecessarily strong combinatorial restriction.

We formulate the resulting selection problem as a mixed-integer quadratic program and develop second-order cone and semidefinite relaxations for settings in which exact optimization is impractical. The experiments support the theory established. Across Erdos–Renyi, preferential-attachment, and stochastic block model networks, the optimization-based

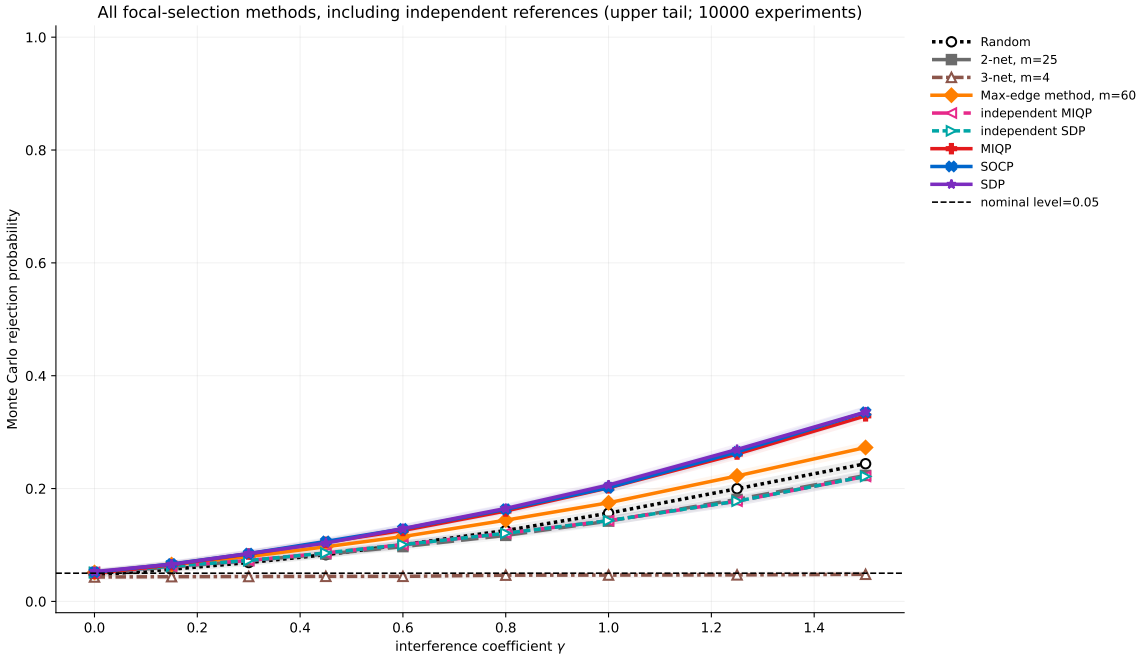


Figure 3.6.6: Power comparisons of all the methods across increasing γ for SBM graphs with four clusters of equal in size on average, with $n = 120$. Within cluster connection probability is 0.4, and across cluster connection probability is 0.02.

methods generally achieve higher power than random focal selection and the greedy graph-based heuristics. The gains can be substantial on heterogeneous network structures, where simple edge counts or separation constraints do not capture the full geometry of exposure variation. At the same time, the cycle-graph results illustrate settings in which simple graph-theoretic constructions can in fact be optimal, thereby identifying when the existing heuristics have a principled justification.

There are several directions in which the present analysis can be extended. Our theoretical results focus primarily on Bernoulli treatment assignment, a local neighborhood exposure model, bounded-degree graph sequences, and a linear working outcome model. More general treatment designs and exposure mappings may produce different forms of informative exposure variation, but the same principle, selecting focal units to maximize the component of rerandomized exposure variation that is informative for the test statistic, provides a

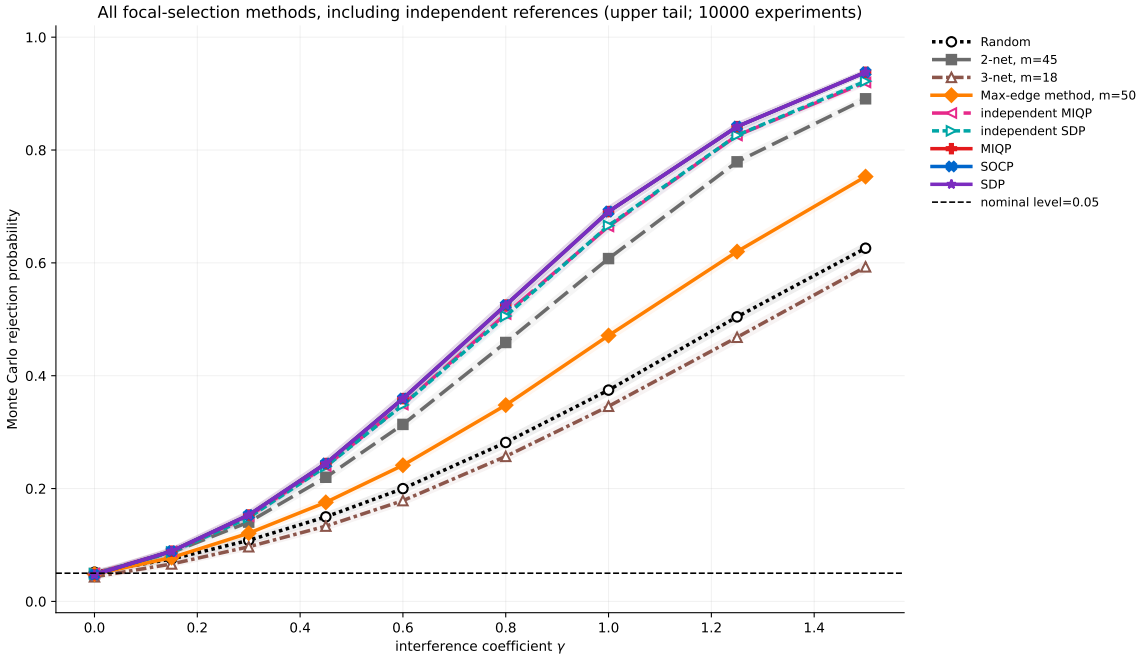


Figure 3.6.7: Power comparisons of all the methods across increasing γ for SBM graphs with four clusters of sizes 10%, 20%, 30%, and 40% on average, with $n = 120$.Within cluster connection probability is 0.1, and across cluster connection probability is 0.02.

natural starting point. Likewise, incorporating covariates, heterogeneous interference effects, and homophilous outcome variation should lead to objectives that distinguish variation aligned with the effect of interest from variation explained by observed or latent network structure. Finally, the choice of test statistic also affects power, and thus optimizing both the test statistic and the focal sets jointly would be an important direction for future work.

Appendix

3.A Proofs

We start by proving Theorem 3.3.8.

Theorem 3.A.1 (Asymptotic power optimality of $C_M(F)$)**.** *Let $\mathcal{F}_n$ be a class of proper focal sets $F \subsetneq V^{(n)}$. Suppose Assumptions 3.3.1 to 3.3.5 hold. Assume $\gamma_n \geq 0$, corresponding to*

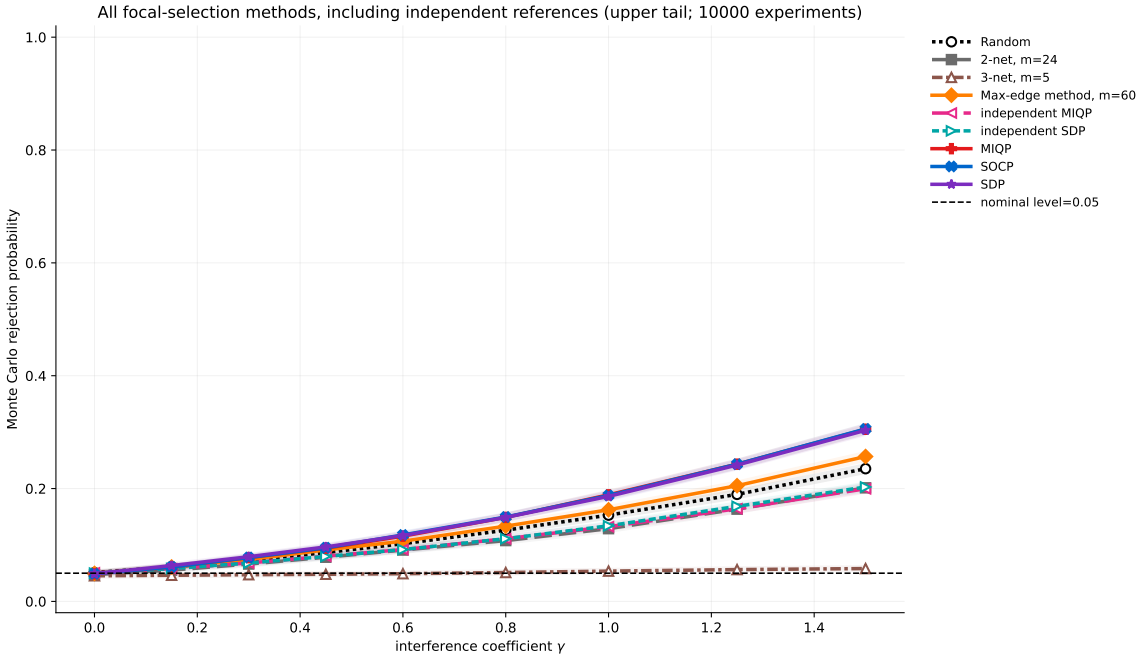


Figure 3.6.8: Power comparisons of all the methods across increasing γ for SBM graphs with four clusters of sizes 10%, 20%, 30%, and 40% on average, with $n = 120$.Within cluster connection probability is 0.4, and across cluster connection probability is 0.02.

the upper-tail test. For each $F \in \mathcal{F}_n$, let

$$C_M(F) := p(1-p)\,\mathrm{Tr}\left(M_F B_{FR} B_{FR}^\top M_F\right),$$

and let

$$F^\star \in \arg\max_{F \in \mathcal{F}_n} C_M(F).$$

Then $F^\star$ satisfies Assumption 3.3.6 with probability tending to one. Moreover, Assumption 3.3.7 holds for $F^\star$. If

$$\frac{\gamma_n \sqrt{C_M(F^\star)}}{\sigma} = O_P(1),$$

then

$$\mathrm{Power}(F^\star) = 1 - \Phi\left(c_{1-\alpha} - \frac{\gamma_n \sqrt{C_M(F^\star)}}{\sigma}\right) + o_P(1), \tag{3.A.1}$$

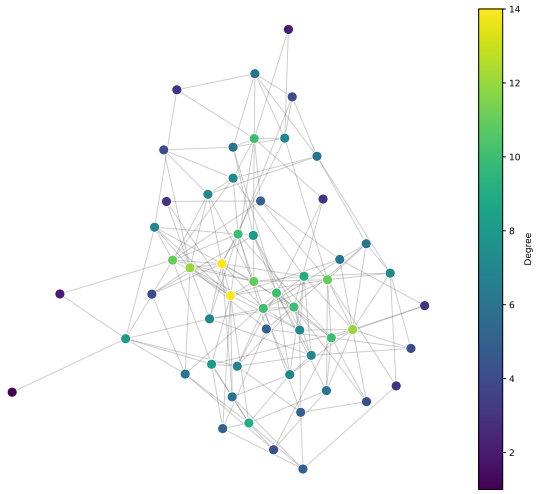


Figure 3.6.9: Village network from dayudengxiongmin1. An undirected edge $\{i, j\}$ is included whenever either household i nominated household j or household j nominated household i

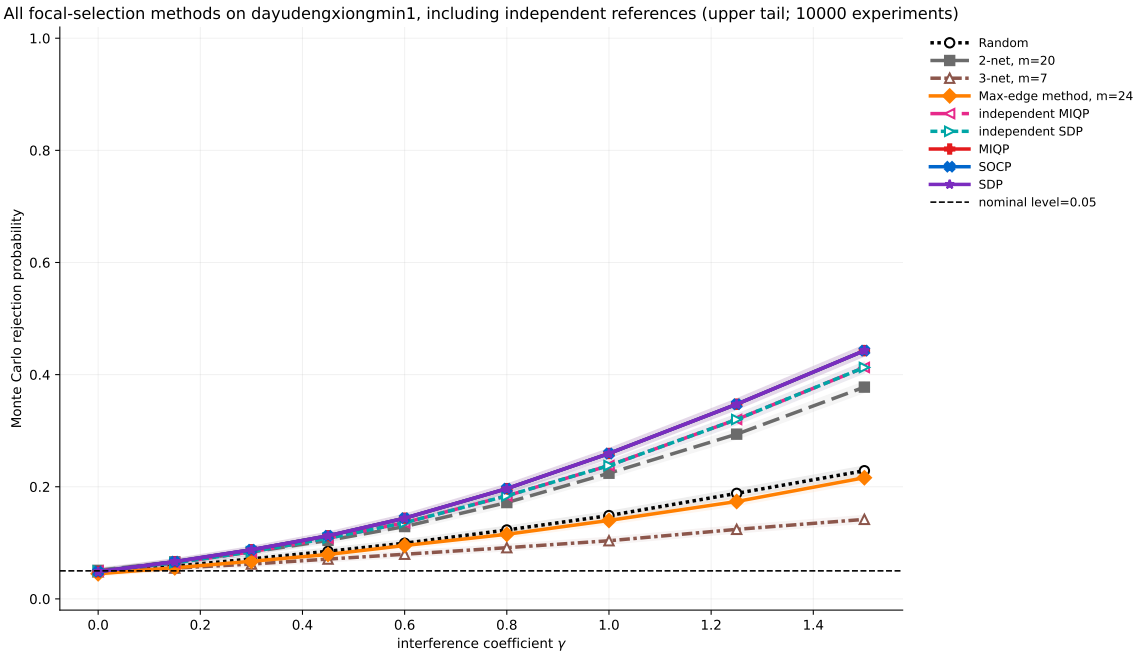


Figure 3.6.10: Power comparisons of all the methods across increasing γ on the village network.

and

$$\mathrm{Power}(F^\star) = \max_{F \in \mathcal{F}_n} \mathrm{Power}(F) + o_P(1). \tag{3.A.2}$$

If instead

$$\frac{\gamma_n \sqrt{C_M(F^\star)}}{\sigma} \to \infty,$$

then

$$\mathrm{Power}(F^\star) \to 1,$$

and (3.A.2) *again follows. Thus, maximizing $C_M(F)$ is asymptotically sufficient for maximizing the power of the conditional randomization test.*

Proof. We proceed in five steps. We first record several bounds that hold without imposing Assumption 3.3.6. We then verify Assumption 3.3.6 at the C_M-maximizer using the focal-gain lemma. Third, we derive the asymptotic power formula for any sequence satisfying $C_M(F) \to \infty$. Fourth, we show that focal sets with bounded $C_M(F)$ have asymptotic power at most α. Finally, we combine these two regimes to establish the optimality of $F^\star$.

Bounding each component. Fix $F \in \mathcal{F}_n$, write $m = |F|$, and let $R = V^{(n)} \setminus F$. Recall

$$P_F := I - M_F, \qquad \rho := B_{FR}\widetilde{z}_R, \qquad r := M_F \rho,$$

and define

$$C_\rho := \mathbb{E}\|\rho\|^2 = p(1-p)\|B_{FR}\|_F^2.$$

Also define

$$\Sigma_R := \mathbb{E}[rr^\top] = p(1-p) M_F B_{FR} B_{FR}^\top M_F,$$

so that

$$C_M(F) = \mathrm{Tr}(\Sigma_R) = \mathbb{E}\|r\|^2.$$

Because M_F and P_F are complementary orthogonal projections,

$$\|r\|^2 = \|\rho\|^2 - \rho^\top P_F \rho.$$

Taking expectations,

$$C_\rho - C_M(F) = p(1-p)\,\mathrm{Tr}\left(P_F B_{FR} B_{FR}^\top\right). \tag{3.A.3}$$

Since $\mathrm{rank}(P_F) \le 2$,

$$0 \le C_\rho - C_M(F) \le 2p(1-p)\|B_{FR}\|_{\mathrm{op}}^2.$$

By Assumptions 3.3.2 and 3.3.4,

$$\|B_{FR}\|_{\mathrm{op}}^2 \le \|B_{FR}\|_1 \|B_{FR}\|_\infty = O(1).$$

Hence

$$C_\rho = C_M(F) + O(1). \tag{3.A.4}$$

In particular,

$$C_M(F) \to \infty \quad \Longleftrightarrow \quad C_\rho \to \infty. \tag{3.A.5}$$

We will also use the uniform operator-norm bound

$$\|\Sigma_R\|_{\mathrm{op}} \le p(1-p)\|B_{FR}\|_{\mathrm{op}}^2 = O(1). \tag{3.A.6}$$

Consequently,

$$\begin{aligned} \mathrm{Tr}(\Sigma_R^2) &\le \|\Sigma_R\|_{\mathrm{op}}\,\mathrm{Tr}(\Sigma_R) \\ &= O(C_M(F))\,. \end{aligned} \tag{3.A.7}$$

Recall also that

$$L_F = M_F \left(B_{FF} z_F + B_{FR} \pi_R \right).$$

The vector inside the projection has uniformly bounded coordinates. Because projection on $\mathrm{span}\{\mathbf{1}_F, z_F\}$ amounts to subtracting the appropriate focal assignment means, the coordinates of L_F are uniformly bounded as well.

Since $M_F L_F = L_F$,

$$L_F^\top \Sigma_R L_F = p(1-p) \| B_{FR}^\top L_F \|^2.$$

For every $j \in R$, Assumption 3.3.2 and Cauchy–Schwarz give

$$\left(\sum_{i \in F} B_{ij} L_{F,i} \right)^2 \le C \sum_{i \in F} B_{ij}^2.$$

Summing over $j \in R$,

$$L_F^\top \Sigma_R L_F = O(C_\rho). \tag{3.A.8}$$

Therefore, whenever $C_M(F) \to \infty$, (3.A.4) implies

$$L_F^\top \Sigma_R L_F + \mathrm{Tr}(\Sigma_R^2) = O(C_M(F)). \tag{3.A.9}$$

Thus Assumption 3.3.7 holds automatically along every sequence for which $C_M(F) \to \infty$.

The C_M-maximizer satisfies boundary mass condition. We now consider $F^\star$. Suppose, to the contrary, that $F^\star$ contains a focal unit v having no auxiliary neighbors, so that

$$B_{vR} = 0.$$

Let

$$F^- = F^\star \setminus \{v\}.$$

By Lemma 3.3.11,

$$C_M(F^-) - C_M(F^\star) = p(1-p)\|B_{F^-v}\|^2 + o_P(1). \tag{3.A.10}$$

The $o_P(1)$ term is the residualization loss generated by deleting one observation from the regression on $\mathbf{1}_F$ and z_F; it tends to zero because the treated and control focal counts used in that regression both diverge.

By Assumption 3.3.3, v has at least one neighbor. Since v has no auxiliary neighbor, every such neighbor lies in F^-. Because the graph is undirected, at least one entry of B_{F^-v} is nonzero. Hence, by Assumption 3.3.4,

$$\|B_{F^-v}\|^2 \geq b_{\min}^2.$$

Substituting into (3.A.10),

$$C_M(F^-) - C_M(F^\star) \geq p(1-p)b_{\min}^2 + o_P(1).$$

Because p is bounded away from zero and one, the right-hand side is strictly positive with probability tending to one. This contradicts the definition of $F^\star$ as a maximizer of $C_M(F)$.

Therefore,

$$\mathbb{P}\left(F^\star \text{ satisfies Assumption 3.3.6}\right) \to 1. \tag{3.A.11}$$

By Assumption 3.3.6 and Assumption 3.3.4,

$$\|B_{F^\star R}\|_F^2 \geq |F^\star| b_{\min}^2.$$

Combining this with (3.A.4),

$$C_M(F^\star) = \Theta(|F^\star|).$$

Since

$$|F^\star| \to \infty,$$

we obtain

$$C_M(F^\star) \to \infty. \tag{3.A.12}$$

Equation (3.A.9) therefore also proves Assumption 3.3.7 along $F^\star$.

Power for sequences satisfying $C_M(F) \to \infty$. We now fix any sequence $F \in \mathcal{F}_n$ such that

$$C_M(F) \to \infty.$$

For an auxiliary rerandomization,

$$z_R^{(b)} = \pi_R + \widetilde{z}_R^{(b)},$$

and

$$\widetilde{e}_{(b)} = L_F + r_{(b)}, \qquad r_{(b)} := M_F B_{FR} \widetilde{z}_R^{(b)}.$$

Define

$$D_m := \|L_F\|^2 + C_M(F).$$

Then

$$D_m = \mathbb{E}_R \left[\widetilde{e}_{(b)}^\top \widetilde{e}_{(b)} \right].$$

Indeed,

$$\widetilde{e}_{(b)}^\top \widetilde{e}_{(b)} = D_m + \delta_{m,(b)},$$

where

$$\delta_{m,(b)} = 2L_F^\top r_{(b)} + \left\{ r_{(b)}^\top r_{(b)} - C_M(F) \right\}. \tag{3.A.13}$$

By (3.A.8) and (3.A.4),

$$\mathrm{Var}_R(L_F^\top r_{(b)}) = L_F^\top \Sigma_R L_F = O(C_M(F)).$$

Likewise, the standard quadratic-form variance bound for independent bounded Bernoulli variables gives

$$\mathrm{Var}_R\left(r_{(b)}^\top r_{(b)}\right) = O(C_M(F)).$$

Hence

$$\delta_{m,(b)} = O_{P_R}\left(\sqrt{C_M(F)}\right). \tag{3.A.14}$$

Since

$$D_m \geq C_M(F) \to \infty,$$

we have

$$\frac{\widetilde{e}_{(b)}^\top \widetilde{e}_{(b)}}{D_m} = 1 + O_{P_R}\left(C_M(F)^{-1/2}\right). \tag{3.A.15}$$

The same statement holds for the observed assignment.

Recall the observed and rerandomized regression coefficients

$$\gamma_{n,o} = \frac{\widetilde{e}_o^\top Y_F}{\widetilde{e}_o^\top \widetilde{e}_o}, \qquad \gamma_{n,(b)} = \frac{\widetilde{e}_{(b)}^\top Y_F}{\widetilde{e}_{(b)}^\top \widetilde{e}_{(b)}}.$$

Define

$$\mu_R = \frac{L_F^\top Y_F}{D_m}, \qquad \Delta_o = \frac{r^\top Y_F}{D_m}, \qquad \Delta_{(b)} = \frac{r_{(b)}^\top Y_F}{D_m}.$$

Using (3.A.15) and the linear outcome model,

$$\gamma_{n,(b)} = \mu_R + \Delta_{(b)} + o_{P_R}\left(\frac{\sigma\sqrt{C_M(F)}}{D_m}\right), \tag{3.A.16}$$

and

$$\gamma_{n,o} = \mu_R + \Delta_o + o_P\left(\frac{\sigma\sqrt{C_M(F)}}{D_m}\right). \tag{3.A.17}$$

To see the claimed rate explicitly, write

$$\overline{\gamma}_{n,(b)} := \frac{(L_F + r_{(b)})^\top Y_F}{D_m}.$$

Then

$$\gamma_{n,(b)} - \overline{\gamma}_{n,(b)} = -\overline{\gamma}_{n,(b)} \frac{\delta_{m,(b)}}{D_m + \delta_{m,(b)}}.$$

The numerator calculation under the linear model gives

$$\overline{\gamma}_{n,(b)} = O_{P_R}\left(\gamma_n + \frac{\sigma}{\sqrt{D_m}}\right),$$

while

$$\frac{\delta_{m,(b)}}{D_m + \delta_{m,(b)}} = O_{P_R}\left(\frac{\sqrt{C_M(F)}}{D_m}\right).$$

Therefore,

$$\frac{\gamma_{n,(b)} - \overline{\gamma}_{n,(b)}}{\sigma\sqrt{C_M(F)}/D_m} = O_{P_R}\left(\frac{\gamma_n}{\sigma} + \frac{1}{\sqrt{D_m}}\right) = o_{P_R}(1)$$

by Assumption 3.3.5. This proves (3.A.16); the observed version follows in the same way.

The common term μ_R is fixed across rerandomizations and therefore cancels from the CRT comparison.

We next consider the observed distribution. As in the central limit theorem, write

$$S_m := D_m\left(\Delta_o - \frac{\gamma_n C_M(F)}{D_m}\right).$$

Let

$$\rho = B_{FR}\widetilde{z}_R, \qquad C_\rho = \mathbb{E}\|\rho\|^2,$$

and define

$$T_m := \gamma_n\left\{\rho^\top L_F + \|\rho\|^2 - C_\rho\right\} + \rho^\top \varepsilon_F.$$

As in the proof of the central limit theorem,

$$T_m = \sum_{i \in F} \psi_{m,i},$$

where

$$\psi_{m,i} = \gamma_n \left\{ L_{F,i}\rho_i + \rho_i^2 - \mathbb{E}[\rho_i^2] \right\} + \rho_i \varepsilon_i.$$

Only focal units having at least one auxiliary neighbor contribute a nonzero summand. For every such focal unit,

$$\mathbb{E}[\rho_i^2] = p(1-p) \sum_{j \in R} B_{ij}^2 \geq p(1-p) b_{\min}^2.$$

Consequently, the number of nonzero summands is $O(C_\rho)$. By bounded degree, their affinity sets have uniformly bounded cardinality. Moreover, the fourth moments of the nonzero $\psi_{m,i}$ are uniformly bounded.

Let

$$\Omega_m := \mathrm{Var}(T_m).$$

The bounded affinity sets imply

$$\Omega_m = O(C_\rho).$$

For the lower bound, write

$$T_m = A_m^\rho + B_m^\rho,$$

where

$$B_m^\rho = \rho^\top \varepsilon_F.$$

Since

$$\mathbb{E}[B_m^\rho \mid \rho] = 0,$$

we have

$$\mathrm{Cov}(A_m^\rho, B_m^\rho) = 0,$$

and hence

$$\begin{aligned}\Omega_m &\geq \mathrm{Var}(B_m^\rho) \\ &= \sigma^2 \mathbb{E}\|\rho\|^2 \\ &= \sigma^2 C_\rho.\end{aligned}$$

Thus

$$\Omega_m = \Theta(C_\rho) = \Theta(C_M(F)). \tag{3.A.18}$$

The three affinity-set conditions now follow exactly as in the central limit theorem, but with C_ρ replacing m as the effective number of nonzero summands. Specifically,

$$\sum_i \sum_{j,k \in A_i} \mathbb{E}|\psi_i \psi_j \psi_k| = O(C_\rho) = o(\Omega_m^{3/2}),$$

and

$$\sum_{i,j} \sum_{k \in A_i, l \in A_j} \mathrm{Cov}(\psi_i \psi_k, \psi_j \psi_l) = O(C_\rho) = o(\Omega_m^2).$$

The third condition is zero because ψ_i is independent of the summands outside its affinity set. Therefore,

$$\frac{T_m}{\sqrt{\Omega_m}} \Rightarrow N(0,1). \tag{3.A.19}$$

It remains to transfer this result from T_m to S_m. Since $M_F = I - P_F$,

$$S_m = T_m + R_m,$$

where

$$R_m = -\gamma_n \left\{ \rho^\top P_F \rho - \mathbb{E}[\rho^\top P_F \rho] \right\} - \rho^\top P_F \varepsilon_F.$$

Because $\operatorname{rank}(P_F) \le 2$ and $\|B_{FR}\|_{\mathrm{op}} = O(1)$, standard quadratic-form moment bounds give

$$\mathbb{E}[R_m^2] = O(1).$$

Combining this with (3.A.18),

$$\frac{R_m}{\sqrt{\Omega_m}} \xrightarrow{P} 0.$$

Hence

$$\frac{S_m}{\sqrt{\operatorname{Var}(S_m)}} \Rightarrow N(0,1).$$

Therefore

$$\frac{\Delta_o - \gamma_n C_M(F)/D_m}{\sqrt{V_m}} \Rightarrow N(0,1), \tag{3.A.20}$$

where

$$V_m = \frac{1}{D_m^2}\left[\sigma^2 C_M(F) + \gamma_n^2 \operatorname{Var}\left(r^\top L_F + \|r\|^2\right)\right].$$

The same dependency bounds used above give

$$\operatorname{Var}\left(r^\top L_F + \|r\|^2\right) = O(C_M(F)).$$

Thus, by Assumption 3.3.5,

$$V_m = \frac{\sigma^2 C_M(F)}{D_m^2}\{1 + o(1)\}. \tag{3.A.21}$$

We now analyze the rerandomization distribution. Conditional on the observed data,

$$D_m \Delta_{(b)} = \tilde{z}_R^{(b)\top} B_{FR}^\top M_F Y_F.$$

Its conditional variance is

$$V_R = \frac{Y_F^\top \Sigma_R Y_F}{D_m^2}.$$

Expanding $M_F Y_F = \gamma_n(L_F + r) + M_F \varepsilon_F$ gives

$$Y_F^\top \Sigma_R Y_F = \sigma^2 C_M(F) + \gamma_n^2 L_F^\top \Sigma_R L_F + \gamma_n^2 \operatorname{Tr}(\Sigma_R^2) + o_P\Big(\sigma^2 C_M(F)\Big).$$

By (3.A.9) and Assumption 3.3.5,

$$\gamma_n^2 \left\{ L_F^\top \Sigma_R L_F + \operatorname{Tr}(\Sigma_R^2) \right\} = o\Big(\sigma^2 C_M(F)\Big).$$

Hence

$$V_R = \frac{\sigma^2 C_M(F)}{D_m^2}\{1 + o_P(1)\}. \tag{3.A.22}$$

Combining (3.A.21) and (3.A.22),

$$\frac{V_R}{V_m} \xrightarrow{P} 1. \tag{3.A.23}$$

Finally, the conditional Lindeberg condition for $\Delta_{(b)}$ follows because the number of nonzero columns of B_{FR} is $O(C_\rho)$, each column has uniformly bounded support, and $C_\rho \asymp C_M(F) \to \infty$. Thus,

$$\frac{\Delta_{(b)}}{\sqrt{V_R}} \Rightarrow N(0,1) \qquad \text{conditionally in probability.}$$

Therefore the conditional $(1-\alpha)$ rerandomization quantile is

$$c_{1-\alpha}\sqrt{V_R} + o_P(\sqrt{V_R}).$$

Using (3.A.16)–(3.A.17), the CRT therefore rejects asymptotically when

$$\Delta_o > c_{1-\alpha}\sqrt{V_R} + o_P(\sqrt{V_m}).$$

By (3.A.20),

$$\text{Power}(F) = 1 - \Phi\left(c_{1-\alpha}\sqrt{\frac{V_R}{V_m}} - \frac{\gamma_n C_M(F)/D_m}{\sqrt{V_m}}\right) + o_P(1).$$

Using (3.A.21) and (3.A.23),

$$\text{Power}(F) = 1 - \Phi\left(c_{1-\alpha} - \frac{\gamma_n\sqrt{C_M(F)}}{\sigma}\{1 + o_P(1)\}\right) + o_P(1). \tag{3.A.24}$$

In particular, if

$$\frac{\gamma_n\sqrt{C_M(F)}}{\sigma} = O_P(1),$$

then the Lipschitz continuity of Φ yields

$$\text{Power}(F) = 1 - \Phi\left(c_{1-\alpha} - \frac{\gamma_n\sqrt{C_M(F)}}{\sigma}\right) + o_P(1). \tag{3.A.25}$$

If instead

$$\frac{\gamma_n\sqrt{C_M(F)}}{\sigma} \to \infty,$$

then (3.A.24) gives

$$\text{Power}(F) \to 1. \tag{3.A.26}$$

Focal sets with bounded $C_M(F)$. We next show that a focal-set sequence with bounded $C_M(F)$ cannot have nontrivial asymptotic power under the local alternatives. Conditional on F and z_F, consider the reference distribution under which z_R has the same Bernoulli law as under the experiment, but

$$Y_F = \alpha\mathbf{1}_F + \beta z_F + \gamma_n\left(B_{FF}z_F + B_{FR}\pi_R\right) + \varepsilon_F.$$

Under this reference distribution, Y_F does not depend on the realized auxiliary assignments z_R. Therefore the conditional randomization test is level α.

Under the alternative model, the only additional mean term is

$$\gamma_n B_{FR} \widetilde{z}_R.$$

The KL-divergence from the alternative joint law of (z_R, Y_F) to the preceding reference law is

$$\begin{aligned} D_{\mathrm{KL}} &= \frac{\gamma_n^2}{2\sigma^2} \mathbb{E}\left[\| B_{FR} \widetilde{z}_R \|^2 \right] \\ &= \frac{\gamma_n^2}{2\sigma^2} C_\rho. \end{aligned}$$

By (3.A.4), if

$$C_M(F) = O(1),$$

then

$$C_\rho = O(1).$$

Hence, by Assumption 3.3.5,

$$D_{\mathrm{KL}} = o(1).$$

Pinsker's inequality gives

$$d_{\mathrm{TV}} \leq \sqrt{\frac{D_{\mathrm{KL}}}{2}} = o(1).$$

Since the test has level α under the reference distribution, its power under the alternative satisfies

$$\mathrm{Power}(F) \leq \alpha + o(1). \tag{3.A.27}$$

Thus bounded-C_M focal sets are asymptotically powerless against the local alternatives beyond the nominal level.

Maximizing $C_M(F)$ maximizes power. We now compare $F^\star$ with an arbitrary sequence of competing focal sets.

By (3.A.12), $C_M(F^\star) \to \infty$.

First suppose

$$\frac{\gamma_n \sqrt{C_M(F^\star)}}{\sigma} = O_P(1).$$

Then

$$\text{Power}(F^\star) = 1 - \Phi\left(c_{1-\alpha} - \frac{\gamma_n \sqrt{C_M(F^\star)}}{\sigma}\right) + o_P(1).$$

Consider any competing sequence $F \in \mathcal{F}_n$.

If, along a subsequence,

$$C_M(F) \to \infty,$$

then the third section in this proof gives

$$\text{Power}(F) = 1 - \Phi\left(c_{1-\alpha} - \frac{\gamma_n \sqrt{C_M(F)}}{\sigma}\right) + o_P(1).$$

Since $F^\star$ maximizes C_M,

$$C_M(F) \leq C_M(F^\star).$$

Because $\gamma_n \geq 0$ and Φ is increasing,

$$\text{Power}(F) \leq \text{Power}(F^\star) + o_P(1).$$

If instead $C_M(F)$ is bounded along a subsequence, then by (3.A.27),

$$\text{Power}(F) \leq \alpha + o(1).$$

On the other hand,

$$1 - \Phi\left(c_{1-\alpha} - \frac{\gamma_n \sqrt{C_M(F^\star)}}{\sigma}\right) \geq 1 - \Phi(c_{1-\alpha}) = \alpha,$$

so

$$\text{Power}(F) \leq \text{Power}(F^\star) + o_P(1)$$

again.

Hence the preceding two arguments, together with the subsequence principle, imply

$$\text{Power}(F) \leq \text{Power}(F^\star) + o_P(1)$$

for every competing sequence. Taking a power-maximizing sequence therefore gives

$$\max_{F \in \mathcal{F}_n} \text{Power}(F) \leq \text{Power}(F^\star) + o_P(1).$$

The reverse inequality holds trivially because $F^\star \in \mathcal{F}_n$. Thus,

$$\text{Power}(F^\star) = \max_{F \in \mathcal{F}_n} \text{Power}(F) + o_P(1).$$

Finally, if

$$\frac{\gamma_n \sqrt{C_M(F^\star)}}{\sigma} \to \infty,$$

then (3.A.26) gives

$$\text{Power}(F^\star) \to 1.$$

Since power is bounded above by one,

$$0 \leq \max_{F \in \mathcal{F}_n} \text{Power}(F) - \text{Power}(F^\star) \leq 1 - \text{Power}(F^\star) \to 0.$$

This establishes the result in the diverging-signal regime as well. ■

Next, we restate and prove Theorem 3.3.9.

Theorem 3.A.2 (Central Limit Theorem)**.** *Suppose Assumptions 3.3.1 to 3.3.4 and 3.3.6*

hold, with $\gamma_n = \gamma > 0$. *Then,*

$$\Delta_o \Rightarrow \mathcal{N}(\frac{\gamma C_M(s)}{D_m}, V_m),$$

where

$$V_m = \frac{1}{D_m^2}\left[\sigma^2 C_M(s) + \gamma^2 \operatorname{Var}(r^T L + \|r\|^2)\right]. \tag{3.A.28}$$

Proof. Because $r = M_F B_{FR} \widetilde{z}_R$ lies in the range of M_F, the nuisance terms in the outcome model vanish when premultiplied by $r^\top$. Moreover, $M_F e_F = L_F + r$. Hence

$$\begin{aligned} D_m \Delta_o &= r^\top Y_F \\ &= \gamma\, r^\top (L_F + r) + r^\top \varepsilon_F. \end{aligned}$$

Since $\mathbb{E}[r] = 0$ and $\mathbb{E}\|r\|^2 = C_M(s)$,

$$\mathbb{E}[\Delta_o] = \frac{\gamma C_M(s)}{D_m}.$$

Therefore,

$$S_m := D_m \left(\Delta_o - \frac{\gamma C_M(s)}{D_m}\right) = \gamma\{r^\top L_F + \|r\|^2 - C_M(s)\} + r^\top \varepsilon_F. \tag{3.A.29}$$

We prove a central limit theorem for S_m in two steps.

Removing the rank-two projection. Set $\rho = B_{FR}\widetilde{z}_R$, so that $r = M_F \rho$, and recall

$$C_\rho = \mathbb{E}\|\rho\|^2.$$

We start by considering the simple case. Consider

$$T_m = \gamma\{\rho^\top L_F + \|\rho\|^2 - C_\rho\} + \rho^\top \varepsilon_F.$$

This can be written as $T_m = \sum_{i \in F} \psi_{m,i}$, where

$$\psi_{m,i} = \gamma\{L_{F,i}\rho_i + \rho_i^2 - \mathbb{E}[\rho_i^2]\} + \rho_i \varepsilon_i.$$

For each $i \in F$, let

$$A_i := \{k \in F : (N_i \cap R) \cap (N_k \cap R) \neq \varnothing\} \cup \{i\}.$$

If $k \notin A_i$, then $\psi_{m,i}$ and $\psi_{m,k}$ depend on disjoint auxiliary assignments and independent outcome errors, and are therefore independent. By bounded degree,

$$|A_i| \leq 1 + d + d(d-1)$$

uniformly in i and m.

Assumptions 3.3.1, 3.3.2 and 3.3.4 imply $|\rho_i| = O(1)$ uniformly. Bernoulli randomization also gives $(X_F^\top X_F)^{-1} = O_P(m^{-1})$, and hence $|L_{F,i}| = O_P(1)$ uniformly. Because the errors are Gaussian,

$$\sup_{i \in F} \mathbb{E}|\psi_{m,i}|^4 < \infty.$$

It follows that

$$\Omega_m := \mathrm{Var}(T_m) = O(m),$$

because only $O(m)$ covariance terms are nonzero. For the matching lower bound, condition on ρ. Since $\mathbb{E}[\rho^\top \varepsilon_F \mid \rho] = 0$, the covariance between $\rho^\top \varepsilon_F$ and the remaining part of T_m is zero. Therefore

$$\Omega_m \geq \mathrm{Var}(\rho^\top \varepsilon_F) = \sigma^2 \mathbb{E}\|\rho\|^2 = \sigma^2 C_\rho.$$

Assumptions 3.3.4 and 3.3.6 imply $C_\rho = \Theta(m)$, while bounded degree gives $C_\rho = O(m)$. Consequently,

$$\Omega_m = \Theta(m).$$

We now verify the three affinity-set conditions. Bounded fourth moments and uniformly bounded $|A_i|$ give

$$\sum_{i\in F}\sum_{j,k\in A_i} \mathbb{E}|\psi_{m,i}\psi_{m,j}\psi_{m,k}| = O(m) = o(\Omega_m^{3/2}).$$

Likewise, for each fixed (i,k) with $k \in A_i$, there are only $O(1)$ pairs (j,ℓ) for which $\mathrm{Cov}(\psi_{m,i}\psi_{m,k}, \psi_{m,j}\psi_{m,\ell})$ can be nonzero. Hence

$$\sum_{i,j\in F}\sum_{k\in A_i,\,\ell\in A_j} \mathrm{Cov}(\psi_{m,i}\psi_{m,k}, \psi_{m,j}\psi_{m,\ell}) = O(m) = o(\Omega_m^2).$$

Finally, with $\psi_{-i} = \sum_{j\notin A_i}\psi_{m,j}$, independence gives

$$\mathbb{E}[\psi_{m,i} \mid \psi_{-i}] = \mathbb{E}[\psi_{m,i}] = 0,$$

so the third affinity-set term is identically zero. The central limit theorem of Chandrasekhar et al. therefore yields

$$\frac{T_m}{\sqrt{\Omega_m}} \Rightarrow N(0,1). \tag{3.A.30}$$

Restoring the projection. Write $M_F = I - P_F$. Since $L_F = M_F L_F$,

$$r^\top L_F = \rho^\top L_F, \qquad \|r\|^2 = \|\rho\|^2 - \rho^\top P_F \rho, \qquad r^\top \varepsilon_F = \rho^\top \varepsilon_F - \rho^\top P_F \varepsilon_F,$$

and

$$C_M(s) = C_\rho - \mathbb{E}[\rho^\top P_F \rho].$$

Thus $S_m = T_m + R_m$, where

$$R_m = -\gamma\{\rho^\top P_F \rho - \mathbb{E}[\rho^\top P_F \rho]\} - \rho^\top P_F \varepsilon_F.$$

Now

$$\rho^\top P_F \rho = (X_F^\top \rho)^\top (X_F^\top X_F)^{-1} (X_F^\top \rho),$$

and similarly

$$\rho^\top P_F \varepsilon_F = (X_F^\top \rho)^\top (X_F^\top X_F)^{-1} (X_F^\top \varepsilon_F).$$

Bounded degree gives $X_F^\top \rho = O_P(\sqrt{m})$; Gaussian errors give $X_F^\top \varepsilon_F = O_P(\sqrt{m})$; and Bernoulli randomization gives $(X_F^\top X_F)^{-1} = O_P(m^{-1})$. Therefore

$$\rho^\top P_F \rho = O_P(1), \qquad \rho^\top P_F \varepsilon_F = O_P(1), \qquad R_m = O_P(1).$$

The same fourth-moment bounds strengthen the last display to $\mathbb{E}[R_m^2] = O(1)$. Since $\Omega_m = \Theta(m)$,

$$\frac{R_m}{\sqrt{\Omega_m}} = o_P(1),$$

and (3.A.30) implies

$$\frac{S_m}{\sqrt{\Omega_m}} \Rightarrow N(0,1).$$

Moreover, if $W_m = \mathrm{Var}(S_m)$, then

$$|W_m - \Omega_m| \le \mathbb{E}[R_m^2] + 2\sqrt{\Omega_m \mathbb{E}[R_m^2]} = o(m),$$

so $W_m/\Omega_m \to 1$. Hence

$$\frac{S_m}{\sqrt{W_m}} \Rightarrow N(0,1).$$

Using (3.A.29) and

$$W_m = \sigma^2 C_M(s) + \gamma^2 \operatorname{Var}(r^\top L_F + \|r\|^2) = D_m^2 V_m$$

completes the proof:

$$\frac{\Delta_o - \gamma C_M(s)/D_m}{\sqrt{V_m}} \Rightarrow N(0,1).$$

■

Next, we present the proof to Corollary 3.3.10.

Corollary 3.A.3. *Suppose Assumptions 3.3.1 to 3.3.6 hold. Then,*

$$\Delta_o \Rightarrow \mathcal{N}(\frac{\gamma_n C_M(s)}{D_m}, V_m),$$

where

$$V_{m,n} = \frac{1}{D_m^2}\left[\sigma^2 C_M(s) + \gamma_n^2 \operatorname{Var}(r^T L + \|r\|^2)\right]. \tag{3.A.31}$$

If, in addition, Assumption 3.3.7 holds,

$$V_{m,n} = \frac{1}{D_m^2}\left[\sigma^2 C_M(s) + \gamma_n^2 \operatorname{Var}(r^T L + \|r\|^2)\right] = \frac{\sigma^2 C_M(s)}{D_m^2}\{1 + o(1)\},$$

and therefore,

$$\frac{\Delta_o - \gamma_n C_M(s)/D_m}{\sigma\sqrt{C_M(s)}/D_m} \Rightarrow \mathcal{N}(0,1)$$

Proof. The proof of the preceding theorem remains valid with γ replaced by γ_n. Indeed, the upper moment bounds are uniform for bounded γ_n, while the variance lower bound

$$\operatorname{Var}(T_m) \geq \sigma^2 C_\rho = \Theta(m)$$

does not depend on γ_n. Therefore

$$\frac{\Delta_o - \gamma_n C_M(s)/D_m}{\sqrt{V_{m,n}}} \Rightarrow N(0,1),$$

where

$$V_{m,n} = \frac{1}{D_m^2}\left\{\sigma^2 C_M(s) + \gamma_n^2 \operatorname{Var}(r^\top L_F + \|r\|^2)\right\}.$$

The same quadratic-form variance bound used above gives

$$\operatorname{Var}(r^\top L_F + \|r\|^2) = O\Big(L_F^\top \Sigma_R L_F + \operatorname{Tr}(\Sigma_R^2)\Big).$$

Under Assumptions 3.3.5 and 3.3.7,

$$\frac{\gamma_n^2 \operatorname{Var}(r^\top L_F + \|r\|^2)}{\sigma^2 C_M(s)} = o(1).$$

Consequently,

$$V_{m,n} = \frac{\sigma^2 C_M(s)}{D_m^2}\{1 + o(1)\},$$

and hence

$$\frac{\Delta_o - \gamma_n C_M(s)/D_m}{\sigma\sqrt{C_M(s)}/D_m} \Rightarrow N(0,1).$$

■

We now prove Lemma 3.3.11.

Lemma 3.A.4 (Asymptotic monotonicity in focal gains)**.** *Suppose Assumptions 3.3.1 to 3.3.3 holds. Let F contain a focal unit v that has no auxiliary neighbors. Let $F^- = F \setminus v$. Then,*

$$C_M(F^-) - C_M(F) = c - o_{\mathbb{P}}(1),$$

for some constant $c > 0$.

Proof. Let $F = F^- \cup \{v\}$ and $R = V \setminus F$. Since v has no auxiliary neighbor, $B_{vR} = 0$, and

after moving v to the auxiliary set, $R^- = R \cup \{v\}$. Therefore

$$B_{FR} = \begin{bmatrix} B_{F^-R} \\ 0 \end{bmatrix}, \qquad B_{F^-R^-} = \begin{bmatrix} B_{F^-R} & B_{F^-v} \end{bmatrix}.$$

It follows that

$$\frac{C_M(F^-) - C_M(F)}{p(1-p)} = \|M_{F^-} B_{F^-v}\|_2^2 \tag{3.A.32}$$

$$- \left\{ \left\| M_F \begin{bmatrix} B_{F^-R} \\ 0 \end{bmatrix} \right\|_F^2 - \|M_{F^-} B_{F^-R}\|_F^2 \right\}. \tag{3.A.33}$$

We bound the two terms on the right separately.

Write $A = B_{F^-R}$, $X_{F^-} = [\mathbf{1}_{F^-} \; z_{F^-}]$, and $x_v = (1, z_v)^\top$. The recursive least-squares identity gives

$$\left\| M_F \begin{bmatrix} A \\ 0 \end{bmatrix} \right\|_F^2 - \|M_{F^-} A\|_F^2 = \frac{\|x_v^\top (X_{F^-}^\top X_{F^-})^{-1} X_{F^-}^\top A\|_2^2}{1 + x_v^\top (X_{F^-}^\top X_{F^-})^{-1} x_v}.$$

Let

$$f^- = |F^-|, \qquad n_1 = \sum_{i \in F^-} z_i, \qquad n_0 = \sum_{i \in F^-} (1 - z_i), \qquad q = X_{F^-} (X_{F^-}^\top X_{F^-})^{-1} x_v.$$

Bernoulli randomization gives $n_1, n_0 = \Theta_P(f^-)$. Additionally,

$$X_{F^-}^T X_{F^-} = \begin{bmatrix} f^- & n_1 \\ n_1 & n_1 \end{bmatrix},$$

and

$$(X_{F^-}^T X_{F^-})^{-1} = \frac{1}{n_1 n_0} \begin{bmatrix} n_1 & -n_1 \\ -n_1 & f^- \end{bmatrix}.$$

Hence

$$\|q\|_2^2 = x_v^\top (X_{F^-}^\top X_{F^-})^{-1} x_v = O_P((f^-)^{-1}).$$

Since bounded degree and bounded weights imply $\|A\|_{\text{op}} = O(1)$,

$$\frac{\|A^\top q\|_2^2}{1 + \|q\|_2^2} \le \|A\|_{\text{op}}^2 \|q\|_2^2 = O_P((f^-)^{-1}).$$

For the first term in (3.A.33),

$$\begin{aligned} \|M_{F^-} B_{F^- v}\|_2^2 = \|B_{F^- v}\|_2^2 \\ - B_{F^- v}^\top X_{F^-} (X_{F^-}^\top X_{F^-})^{-1} X_{F^-}^\top B_{F^- v}. \end{aligned}$$

The column $B_{F^- v}$ has only $O(1)$ nonzero entries, each uniformly bounded. Thus $\|X_{F^-}^\top B_{F^- v}\| = O(1)$, while $(X_{F^-}^\top X_{F^-})^{-1} = O_P((f^-)^{-1})$. Consequently,

$$\|M_{F^-} B_{F^- v}\|_2^2 = \|B_{F^- v}\|_2^2 + O_P((f^-)^{-1}).$$

Substituting the last two bounds into (3.A.33) gives

$$C_M(F^-) - C_M(F) = p(1-p)\|B_{F^- v}\|_2^2 + o_P(1).$$

For an undirected graph, Assumptions 3.3.3 and 3.3.4 imply that v has at least one neighbor in F^- and hence $\|B_{F^- v}\|_2^2 \ge b_{\min}^2$. Therefore

$$C_M(F^-) - C_M(F) \ge p(1-p) b_{\min}^2 + o_P(1),$$

which proves the claim. ■

We next restate and prove Proposition 3.5.1.

Proposition 3.A.5. *Let C_n be the equal edge-weighted cycle graph with $n \geq 7$. If n is even, C_H is maximized by the maximum focal independent set of cardinality $n/2$, i.e., the alternating focal sets. If n is odd, C_H is maximized by the focal sets with cardinality $(n+1)/2$ and are almost alternating: there is exactly one pair of adjacent focal nodes, and the remaining nodes alternate between focal and auxiliary units.*

Proof. For every nonempty proper focal set, let $c(F)$ be the number of focal blocks, let $a(F)$ be the number of singleton auxiliary blocks, and let $q = n - m$. Every focal block contributes two focal–auxiliary edges, so

$$\sum_{j \in R} t_j = 2c(F).$$

Because $t_j \in \{0, 1, 2\}$ and $t_j = 2$ exactly when j is a singleton auxiliary block,

$$\sum_{j \in R} t_j^2 = 2c(F) + 2a(F).$$

Substitution into the objective gives the exact identity

$$C_H(F) = \frac{2w^2}{m}\{(m-1)c(F) - a(F)\}. \tag{3.A.34}$$

The block counts satisfy

$$c(F) \leq \min\{m, q\}, \qquad a(F) \geq \max\{0, 2c(F) - q\}.$$

The second inequality follows because the $c(F) - a(F)$ nonsingleton auxiliary blocks contain at least two vertices each:

$$q \geq a(F) + 2\{c(F) - a(F)\} = 2c(F) - a(F).$$

For fixed m, the right-hand side of (3.A.34) is maximized by taking $c(F)$ as large and $a(F)$ as small as the preceding constraints allow. To see this directly, substitute the lower bound on $a(F)$. The resulting expression is piecewise linear in $c(F)$, with slopes $m-1$ and $m-3$. For $m \geq 3$ both pieces are nondecreasing; the cases $m = 1, 2$ are immediate when $n \geq 7$. Hence, for $n \geq 7$,

$$\max_{|F|=m} C_H(F) = \begin{cases} \dfrac{2w^2}{m}\left\{m(m-1) - \max(0, 3m-n)\right\}, & m \leq n/2, \\ \dfrac{2w^2}{m}(m-2)(n-m), & m > n/2. \end{cases} \tag{3.A.35}$$

The bounds are attainable by arranging the focal and auxiliary blocks so that $c(F) = \min\{m, q\}$ and $a(F) = \max\{0, 2c(F) - q\}$.

Even cycles. Let $n = 2\ell$, with $\ell \geq 4$. At $m = \ell$, (3.A.35) gives

$$\max_{|F|=\ell} C_H(F) = 2w^2(\ell - 2).$$

Equality requires $c(F) = a(F) = \ell$, so every focal and auxiliary block is a singleton. Thus the only maximizers are the two alternating focal sets.

If $m < \ell$ and $3m \leq 2\ell$, then

$$\max_{|F|=m} C_H(F) = 2w^2(m-1) < 2w^2(\ell - 2).$$

If $m < \ell$ and $3m > 2\ell$, then

$$\begin{aligned} 2w^2(\ell-2) - \max_{|F|=m} C_H(F) &= 2w^2\left(\ell - m + 2 - \frac{2\ell}{m}\right) \\ &= \frac{2w^2}{m}(\ell - m)(m-2) > 0. \end{aligned}$$

Finally, if $m > \ell$, then

$$2w^2(\ell-2) - \max_{|F|=m} C_H(F) = 2w^2\left\{\ell - 2 - \frac{(m-2)(2\ell-m)}{m}\right\} > 0.$$

The last inequality follows directly after multiplying by m, because the resulting difference is $(m-\ell)(m-4) > 0$. Thus the alternating focal sets are the global maximizers when n is even.

Odd cycles. Let $n = 2\ell+1$, with $\ell \geq 3$. At $m = \ell+1$, (3.A.35) gives

$$\max_{|F|=\ell+1} C_H(F) = \frac{2w^2}{\ell+1}\ell(\ell-1).$$

Equality requires $c(F) = a(F) = \ell$. Hence every auxiliary block is a singleton, while the $\ell+1$ focal vertices form ℓ focal blocks. There is therefore exactly one adjacent focal pair, and all remaining focal vertices alternate with auxiliary vertices.

For $m \leq \ell$ with $3m \leq 2\ell+1$,

$$\max_{|F|=m} C_H(F) = 2w^2(m-1) < \frac{2w^2}{\ell+1}\ell(\ell-1).$$

For $m \leq \ell$ with $3m > 2\ell+1$, the first line of (3.A.35) is strictly increasing in m, because

$$\left\{m+1-4+\frac{2\ell+1}{m+1}\right\} - \left\{m-4+\frac{2\ell+1}{m}\right\} = 1 - \frac{2\ell+1}{m(m+1)} > 0.$$

It therefore suffices to compare $m = \ell$ and $m = \ell+1$:

$$\begin{aligned}\frac{2w^2}{\ell+1}\ell(\ell-1) - 2w^2\left(\ell-2+\frac{1}{\ell}\right)\\ = \frac{2w^2(\ell-1)}{\ell(\ell+1)} > 0.\end{aligned}$$

For $m > n/2$, the second line of (3.A.35) can be written as

$$2w^2 \left\{ n + 2 - m - \frac{2n}{m} \right\}.$$

Its increment from m to $m+1$ is

$$2w^2 \left\{ -1 + \frac{2n}{m(m+1)} \right\} < 0$$

for every $m \geq \ell+1$. Hence it is maximized at $m = \ell+1$. This proves the odd-cycle claim. ■

We restate and prove Proposition 3.5.2.

Proposition 3.A.6. *Let s_{alt} denote the maximal independent focal set, which is just the alternating focal set in the cycle graph. Let s_{adj} denote the focal set where $m = 2$, and the two focal units are adjacent to each other. Then,* $\max_s C_H(s) = C_H(s_{adj}) > C_H(s_{alt})$.

Proof. We consider first $m = 2$.

Then, $1 \leq c(F) \leq 2$, and $a(F) \geq \max(0, 2c(F) - q)$, and $q = 2$.

$$C_H(F) \leq \frac{2w^2}{m}\{(m-3)c(F) + 2\} = w^2(-c(F) + 2) \leq w^2.$$

Now, for F_{adj}, $c(F_{\text{adj}}) = 1$, and $a(F) = 0$.

$$C_H(F_{\text{adj}}) = w^2(c(F) - a(F)) = w^2,$$

and thus, the upper-bound is realized. For F_{alt}, $c(F_{\text{alt}}) = 2$, and $a(F_{\text{alt}}) = 2$.

$$C_H(F_{\text{alt}}) = w^2(c(F) - a(F)) = 0,$$

and thus the adjacent focal set is better than the independent focal set with cardinality $m = 2$. To see that F_{adj} is optimal over all possible F, we can simply calculate C_H for focal

sets of cardinality 1 and 3. We begin with $F = 1$. In this case $c(F) = 1$, and $a(F) = 0$, and therefore, $C_H(F) = 2w^2(-a(F)) = 0$. For $|F| = 3$, we have $c(F) = 1$, and $a(F) = 1$. Therefore, $C_H(F) = \frac{2w^2}{3}(2c(F) - a(F)) = \frac{2w^2}{3} < w^2$. Thus, F_{adj} is the optimal focal set. ■

Table 3.1: Summary of the focal-unit selection methods used in the numerical experiments. Here F and R denote the focal and auxiliary sets, respectively, $m = |F|$, and $\mathcal{M}$ denotes the candidate grid of focal-set cardinalities. Unless otherwise stated, candidates are compared using the graph-only criterion $C_H(F)$ before the power simulation.

Plot label	**Description**
Random	For each $m \in \mathcal{M}$, draws one uniformly random focal set of size m. Among these candidates, retains the set with the largest $C_H(F)$.
2-net	Constructs a maximal independent focal set greedily, so that no two focal units are adjacent. Multiple randomized greedy orderings are considered, and the resulting 2-net with the largest $C_H(F)$ is retained. This is the independent-set heuristic of Athey et al. (2018).
3-net	Imposes stronger separation between focal units, so that distinct focal units have no common one-hop auxiliary neighbor under the one-neighborhood exposure model. Multiple randomized greedy orderings are considered, and the candidate with the largest $C_H(F)$ is retained.
Alternating focal set	Cycle-only benchmark. For an even cycle, selects every other vertex as focal, giving $m = n/2$.
Max-edge method	Implements the greedy edge-comparison heuristic of Athey et al. (2018). Within each restart, focal units are added according to the increase in the number of focal–auxiliary graph edges, and the procedure stops when no positive gain remains. Multiple randomized tie-breaking restarts are run, and the resulting candidate with the largest $C_H(F)$ is retained.
Independent MIQP	For each feasible $m \in \mathcal{M}$, maximizes $s^\top b - \frac{1}{m} s^\top BB^\top s$ over binary s, subject to $s_i + s_j \leq 1$ on every graph edge. The candidate attaining the largest $C_H(F)$ across feasible m is used in the power comparison.
Independent SDP	Semidefinite relaxation of the independent-set optimization above, followed by rounding to a feasible independent focal set of size m. The procedure is evaluated over feasible $m \in \mathcal{M}$, and the candidate with the largest $C_H(F)$ is retained.
MIQP	Exactly optimizes the unrestricted constant-projection criterion $C_H(F)$ for each $m \in \mathcal{M}$ using a mixed-integer quadratic formulation. The best candidate across m according to $C_H(F)$ is retained.
SOCP	Second-order cone relaxation of the unrestricted $C_H(F)$ optimization. For each $m \in \mathcal{M}$, the relaxed solution is rounded to a focal set of size m and refined using the exact objective. The candidate with the largest $C_H(F)$ across m is retained.
SDP	Semidefinite relaxation of the unrestricted $C_H(F)$ optimization, followed by rounding and local refinement at each $m \in \mathcal{M}$. The candidate with the largest $C_H(F)$ across m is retained. The SDP relaxation is tighter than the corresponding SOCP relaxation.
Boundary-cut MILP	For each $m \in \mathcal{M}$, exactly optimizes the weighted focal–auxiliary boundary objective $\sum_{i \in F} \sum_{j \in R} B_{ij}^2$. The resulting focal sets are evaluated using $C_H(F)$, and the candidate with the largest $C_H(F)$ across m is retained for the power comparison.

Chapter 4

DATA-ADAPTIVE EXPOSURE THRESHOLDS UNDER NETWORK INTERFERENCE

This chapter is based on the paper "Data-Adaptive Exposure Thresholds under Network Interference," coauthored with Tyler H. McCormick[◇,¶], and Jennifer Brennan[⋆], published in Advances in Neural Information Processing Systems 38 (2026): 131360-131393.

[⋆] Google Research, USA

[◇] University of Washington, Department of Statistics, USA

[¶] University of Washington, Department of Sociology, USA

Randomized controlled trials often suffer from interference, a violation of the Stable Unit Treatment Value Assumption (SUTVA), where a unit's outcome is influenced by its neighbors' treatment assignments. This interference biases naive estimators of the average treatment effect (ATE). A popular method to achieve unbiasedness pairs the Horvitz-Thompson estimator of the ATE with a known exposure mapping, a function that identifies units in a given randomization unaffected by interference. For example, an exposure mapping may stipulate that a unit experiences no further interference if at least an h-fraction of its neighbors share its treatment status. However, selecting this threshold h is challenging, requiring domain expertise; in its absence, fixed thresholds such as $h = 1$ are often used. In this work, we propose a data-adaptive method to select the h-fractional threshold that minimizes the mean-squared-error (MSE) of the Horvitz-Thompson estimator. Our approach estimates the bias and variance of the Horvitz-Thompson estimator paired with candidate thresholds by leveraging a first-order approximation, specifically, linear regression of potential outcomes on exposures. We present simulations illustrating that our method improves upon non-adaptive threshold choices, and an adapted Lepski's method. We further illustrate the performance of our estimator by running experiments with synthetic outcomes on a

real village network dataset, and on a publicly-available Amazon product similarity graph. Furthermore, we demonstrate that our method remains robust to deviations from the linear potential outcomes model.

4.1 Introduction

Estimating the Average Treatment Effect (ATE), the difference in the average outcomes of units when all units are treated versus when none are, is challenging in the presence of network interference. Under interference, the Stable Unit Treatment Value Assumption (SUTVA) [Cox, 1958, Rubin, 1978, Manski, 1990] is violated, as a unit's outcome is influenced by its neighbors' treatment assignments. This problem is salient in randomized controlled trials, the setting we examine. For example, estimating the adoption of a new idea through random assignment of people to advertisements is complicated by the dissemination of information through social interactions. Consider a social media platform seeking to estimate the efficacy of a product innovation in increasing user engagement. Measuring outcomes for users in the control group is complicated by their interactions with treated users. If the innovation successfully increases engagement among treated users, their friends in the control group may also exhibit increased engagement through their interactions on the platform. As such discrete interactions can be modeled through network interference models, the problem of estimating average treatment effects under network interference becomes a ubiquitous one.

A fundamental challenge arises from the lack of knowledge about the exact interference pattern. *Exposure mappings*, as defined by Aronow and Samii [2017], are functions that partition the space of treatment assignments and individual-level features (e.g., social neighborhood structure) into distinct exposure values. These mappings encapsulate the concept of "effective treatments" introduced in [Manski, 2013]. For example, let $z \in \{0,1\}^n$ denote the treatment assignment vector, and $W \in [0,1]^{n \times n}$ represent the (weighted) adjacency matrix, where W_{ij} encodes the relationship between units i and j. A simple exposure mapping takes the form $f\colon (z, W) \mapsto Wz$, assigning each unit a weighted sum of its neighbors'

treatments. Aronow and Samii [2017], Ugander et al. [2013], Sussman and Airoldi [2017], Hardy et al. [2019], Auerbach and Tabord-Meehan [2021] expound upon different exposure mappings, and estimation under these settings. Following [Eckles et al., 2017, Toulis and Kao, 2013], we characterize exposure in terms of the fraction of treated neighbors. For instance, a unit may be considered as experiencing no additional interference if at least an h-fraction of its neighbors share its treatment assignment. This fractional neighborhood exposure mapping aligns with the fractional thresholds model [Watts, 1999], where treated and untreated peers exert opposing influences on a unit's outcome. Centola and Macy [2007] illustrate this with the example of refraining from littering in a neighborhood: an individual's decision to avoid littering depends on the relative number of neighbors who also refrain from littering.

Under a known h-fractional neighborhood exposure mapping, the Horvitz-Thompson estimator [Horvitz and Thompson, 1952] is a popular unbiased estimator of the average treatment effect. Without domain knowledge of the interference structure, however, such estimators based on non-adaptive treatment exposure conditions often suffer from high bias or variance. For instance, the Horvitz-Thompson estimator can be paired with an extreme threshold h to achieve unbiasedness, requiring that a unit and all its neighbors be treated (or in control). However, this can result in high variance, since the probability of such configurations is very low, especially for high-degree nodes. Conversely, setting a low threshold h introduces bias into the estimator by including units with limited treatment (resp. control) exposure in the treatment (resp. control) exposed group. We propose a simple data-dependent approach to find an MSE-optimal threshold for the Horvitz-Thompson estimator in the finite population setting. We compute the *bias slope*, an approximate rate of change of bias, which we use to estimate the bias of our estimator. We use this information together with variance estimates to select a threshold minimizing the Mean-Squared-Error (MSE) of the estimator.

In [Basse and Airoldi, 2018, Cai et al., 2015], a (known) generalized linear form of

network effects on neighborhood treatments was assumed. Related to our work, Zhu et al. [2024] fit functionals on the treatment assignment and exposure. In [Chin, 2019], the author proposes regression adjustment estimators that predict potential outcomes under global treatment and control conditions. Unlike traditional regression adjustments, their approach constructs adjustment variables from functions of the treatment assignment vector, framing the learning of a more flexible exposure mapping as a feature engineering problem. While Belloni et al. [2022] focus on estimating direct effects under network interference, their approach to determining the exposure radius, i.e. the m_i-hop influence, shares conceptual ground with our approach of selecting optimal exposure thresholds.

In this paper, we present our framework for the Horvitz-Thompson estimator of the ATE, motivated by its widespread adoption in practice. Nevertheless, our approach could be adapted for other estimators, like the Hájek, Difference-in-Means, and Augmented Inverse Propensity Weighted (AIPW) estimators. We formulate the adaptively-thresholded estimator for the Difference-in-Means estimator, which is a special case of the Hájek estimator, in the Appendix 4.A.7.5. We note however, that the Hájek estimator is only approximately unbiased at the true threshold. This makes formal characterization of its exact bias–variance tradeoff more challenging than for the Horvitz-Thompson estimator.

Within off-policy evaluation in reinforcement learning, Su et al. [2020] address the classical problem of adaptive bandwidth selection, well studied in non-parametric statistics [Fan and Gijbels, 1992, Ruppert, 1997, Kallus and Zhou, 2018, Mukherjee et al., 2015], by applying Lepski's method [Lepskii, 1992, 1993, Lepski and Spokoiny, 1997, Goldenshluger and Lepski, 2011]. Our work explores an extension of this framework to optimal bandwidth selection for the Horvitz-Thompson estimator in average treatment effect estimation under network interference (see Appendix 4.A.3). To our knowledge, this connection has not been explored previously. We compare our procedure to an adaptation of Lepski's method tailored to our problem setting. In our context, the network dependence structure can lead to violations of the monotonicity and decay rates conditions that Lepski's method relies on. Our approach

offers a simple and principled data-adaptive alternative designed to address these challenges in our context. Philosophically, these ideas are similar. This perspective on threshold, or equivalently bandwidth, selection at the estimation stage can loosely be viewed as the "dual" of choosing cluster sizes in cluster-randomization approaches such as Ugander et al. [2013], Eckles et al. [2017] at the design stage.

4.2 Setup

4.2.1 Notation.

We use h to denote the threshold for "treatment exposure", and $1-h$ for "control exposure". Therefore, for larger h, i.e. h closer to one, we have a more restrictive setting, where only the subset of treated (resp. control) units with most of their neighbors treated (resp. control) are counted as treatment (resp. control) exposed. For smaller h, i.e. h close to zero, we have a less restrictive setting, as a larger subset of treated (resp. control) units satisfy the threshold condition. Throughout the paper we use W to denote the adjacency matrix with $W_{ii} = 0$, $d_i = \sum_j W_{ij}$ for the degree of node i, and d when the graph is regular. We use D to denote the diagonal matrix of degrees d_i.

4.2.2 Problem Setup

We study the finite population setting. Additionally, we consider a fractional neighborhood exposure mapping. That is, for any treatment assignment vector $z \in \{0,1\}^n$, the effective influence of the graph's treatment assignment on the outcome of node i is equivalent to the influence of fractional treatment assignments in the neighborhood of node i. This reduces our potential outcomes model to the following form:

$$y_i(z) = y_i(z_i, e_i) = \alpha_i + \psi(z_i, e_i) + \epsilon_i, \tag{4.2.1}$$

where $z_i \in \{0,1\}$ is the treatment assignment of unit i, $e_i \in [0,1]$ is the fraction of treated neighbors of unit i (excluding i itself), and the ϵ_i are uniformly-bounded and non-random.

Throughout the paper, we focus on $\psi(z_i, e_i) = g(z_i) + f(e_i)$, but our framework generalizes. For instance, we will consider the linear potential-outcome model in some examples:

$$y_i(z) = \alpha_i + \psi(z_i, e_i) + \epsilon_i = \alpha_i + \beta_i z_i + \gamma_i e_i + \epsilon_i. \tag{4.2.2}$$

We are interested in estimating the average treatment effect (ATE),

$$\tau = \frac{1}{n}\sum_{i=1}^{n} y_i(1,1) - \frac{1}{n}\sum_{i=1}^{n} y_i(0,0) = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i + \beta_i + \gamma_i + \epsilon_i - \alpha_i - \epsilon_i) = \frac{1}{n}\sum_{i=1}^{n}(\beta_i + \gamma_i), \tag{4.2.3}$$

though we could, more generally, take any difference between exposure categories.

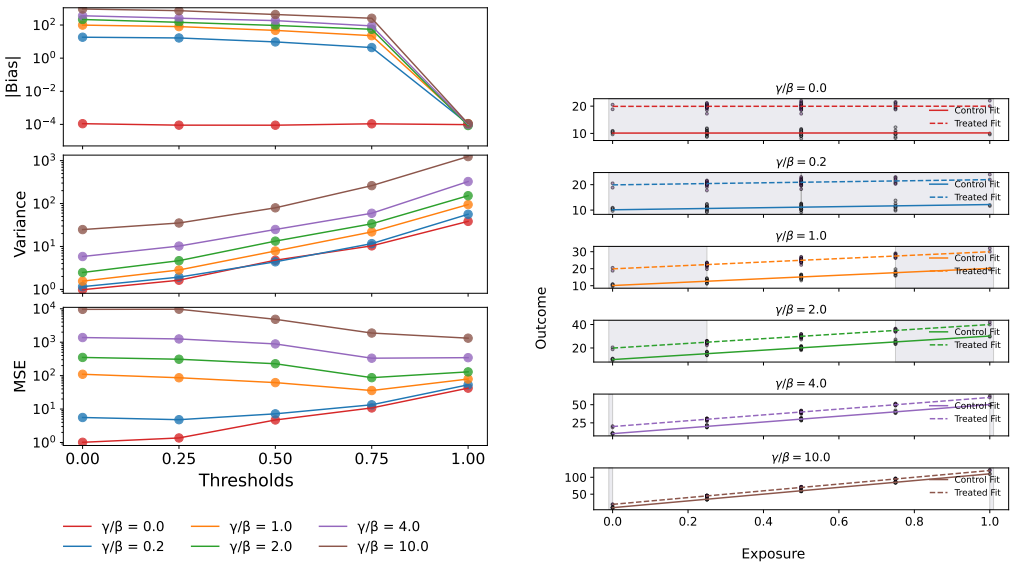
Let Y_i denote the observed outcome for unit i, $i \in [n]$. We consider the Horvitz-Thompson estimator [Horvitz and Thompson, 1952] for a given exposure threshold h:

$$\hat{\tau}_h \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 1, e_i \geq h\}}{\mathbb{P}\{z_i = 1, e_i \geq h\}} Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 0, e_i \leq 1-h\}}{\mathbb{P}\{z_i = 0, e_i \leq 1-h\}} Y_i \tag{4.2.4}$$

Our goal then is to select the threshold h in the Horvitz-Thompson estimator above that minimizes the MSE. More formally, given a candidate set of thresholds H, we select

$$h^* := \operatorname*{argmin}_{h \in H} \mathrm{MSE}(\hat{\tau}_h) = \operatorname*{argmin}_{h \in H} \mathrm{Bias}^2(\hat{\tau}_h) + \mathrm{Var}(\hat{\tau}_h). \tag{4.2.5}$$

Stronger interference, i.e. when $\psi(\cdot, e_i)$ is more sensitive to changes in exposure e_i, induces greater bias through edges that connect units with differing treatment assignments. To mitigate this bias, one might pick a higher exposure threshold h. However, increasing h reduces the exposure probabilities $\mathbb{P}\{z_i = 1, e_i \geq h\}$ and $\mathbb{P}\{z_i = 0, e_i \leq 1 - h\}$, thereby increasing variance. To illustrate this bias–variance trade-off, we display the biases and variances under the linear model $Y_i = \alpha + \beta z_i + \gamma e_i + \epsilon_i$ evaluated across varying thresholds in the Horvitz-Thompson estimator (4.2.4) above, in the left panel of Figure 4.2.1. This trade-off is illustrated for various ratios of γ/β. The right panel of Figure 4.2.1 illustrates



(a) Bias-variance tradeoff

(b) Linear regression across parameters

Figure 4.2.1: Left: bias, variance, and RMSE of AdaThresh across different thresholds for the 1000-node 2nd-power cycle graph (see Appendix 4.A.6) with unit-level Bernoulli randomization, and linear model outcome $Y_i = 10 + 10z_i + \gamma e_i + \epsilon_i$, for fixed ϵ_i generated from $\mathcal{N}(0, 1)$. Right: toy illustration of linear fits used to approximate exposure bias caused by including more data across different thresholds. The "jittered" datapoints signify the amount of variance reduction across thresholds. Rectangular blocks depict the data regions under the MSE-optimal thresholds.

the different fitted linear models for various γ/β ratios, whose slopes we call the bias slope. We then use these bias slopes to compute the bias estimates for $\{\hat{\tau}_h\}_h$ (4.2.4), as described in detail in the next section. Combining these bias estimates with the corresponding variance estimates allows us to select the MSE-optimal threshold, shown by the shaded regions in the right panel of Figure 4.2.1 (matching the left panel).

We adopt the MSE as our objective as we are interested in precisely and accurately estimating the ATE, trading off bias against variance in the estimator. See Deng et al. [2024, Section 2.2] for more discussions on this. We note that, while the overall trend of bias decreasing and variance increasing with h holds, neither quantity is necessarily monotonic due to the dependence between units and, consequently, their exposures. To further illustrate

this, in Appendix 4.A.6, we draw upon toy examples of circulant graphs from Ugander et al. [2013].

4.3 Estimating bias and variance

With the goal of obtaining an MSE-optimal threshold in mind, we find the *bias slope* $\hat{\gamma}_n$, which captures how the bias changes with the threshold h. More concretely, the bias slope quantifies the average change in bias as h is varied from 0 to 1, estimated by the slope of a linear fit. Generally, this linear regression coefficient $\hat{\gamma}_n$ is simply a first-order approximation of how $\psi(\cdot, e_i)$ changes with e_i, estimating the average change in the mean outcome per unit change in exposure. In our context, this also reflects the average change in bias as we move away from the boundary exposures, i.e., from 0 for control or 1 for treated, since in our setting bias arises when exposures deviate from these extremes. This interpretation motivates the use of the weights $e_i - 0 = e_i$ and $1 - e_i$ in Equation 4.3.1, corresponding to deviations from the control and treated boundaries, respectively.

Building on this interpretation, we propose the following Horvitz-Thompson-style estimator for the bias, which integrates the bias slope over the empirical exposure distribution to yield an average (absolute) bias estimate at a given threshold:

$$\hat{b}(\hat{\tau}_h) = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-e_i)\hat{\gamma}_n \mathbf{1}\{z_i = 1, e_i \geq h\}}{\mathbb{P}\{z_i = 1, e_i \geq h\}} + \frac{1}{n}\sum_{i=1}^{n} \frac{e_i \hat{\gamma}_n \mathbf{1}\{z_i = 0, e_i \leq 1-h\}}{\mathbb{P}\{z_i = 0, e_i \leq 1-h\}}, \tag{4.3.1}$$

where $\hat{\gamma}_n$, the bias slope, is the linear regression coefficient of the outcome on the exposure variable. If the potential outcomes depend strongly and positively on the treatment exposures, $\hat{\gamma}_n$ will be larger. Consequently, our estimator will capture a more pronounced increase in bias caused by including more data when using a smaller threshold h. We use a Horvitz-Thompson-style bias estimator to account for the exposure distribution, which may pose challenges for Lepski's method (see description in Appendix 4.A.6).

To estimate the variance associated with a threshold h, we use the variance estimator discussed in Aronow and Samii [2017], available in Appendix 4.A.1.

Together, they inform the choice of threshold, and our adaptive estimator, AdaThresh. We describe our procedure in Algorithm 1.

Algorithm 1 AdaThresh

Require: Graph adjacency matrix W, outcome vector Y, treatment vector z

Compute exposure: $e \leftarrow D^{-1}Wz$
Fit linear model: $Y = \beta z + \gamma e + c$
Let $\hat{\gamma}$ be the estimated coefficient for e
for each threshold $h \in \mathcal{H}$ **do**
    Estimate bias: $\widehat{\text{Bias}}(h)$ using $\hat{\gamma}, Y, z, W$, and h ▷ See Eq. (4.3.1)
    Estimate variance: $\widehat{\text{Var}}(h)$ using Y, z, W, and h ▷ See Appendix 4.A.1
    Compute $\widehat{\text{MSE}}(h) \leftarrow \widehat{\text{Bias}}^2(h) + \widehat{\text{Var}}(h)$
end for
$\hat{h} \leftarrow \arg\min_{h \in \mathcal{H}} \widehat{\text{MSE}}(h)$
return $\hat{\tau}_{\hat{h}}$ ▷ See (4.2.4)

The bias slope from the linear fit captures the average change in bias as h is varied from 0 to 1. We do not assume an underlying linear model $Y = \alpha + \beta z + \gamma e$, for which the sum of linear regression coefficients $\hat{\beta} + \hat{\gamma}$ would be the optimal estimate of the ATE. Rather, we use a linear regression only to obtain the bias slope, a first-order approximation of how $\psi(\cdot, e_i)$ changes with e_i. This idea is similar in spirit to the work of Baird et al. [2018], where the authors leverage the slope of spillovers with respect to treatment saturation levels as an approximation, without assuming linearity in the true data-generating process.

As we first estimate the bias slope to inform our overall estimator of the ATE, sample-splitting is one potential strategy for consistent estimation. However, since the units of interest are dependent (as modeled by the network interference), one would have to prove performance guarantees by leveraging results such as Hart and Vieu [1990]. In our setting, however, it is not necessary that we split the data as we leverage the simplicity of the model class [Van Der Vaart et al., 1996]. In Appendix 4.A.9, we discuss this using Donsker results [Van Der Vaart et al., 1996] to demonstrate that our approach does not lead to overfitting.

4.4 Theoretical Results

Denote by H the (finite) set of exposure thresholds. For example, in cycle graphs (see Appendix 4.A.6 for details), $H = \{\frac{u}{d} : u \in \{0, 1, ..., d\}\}$ where d is the degree of the graph. We assume the following:

Assumption 4.4.1 (Unconfoundedness, Positivity)**.**

1. (Unconfoundedness) For all $i \in [n]$,

$$\{y_i(\tilde{z}_i, \tilde{e}_i) : \tilde{z}_i \in \{0, 1\}, \tilde{e}_i \in [0, 1]\} \perp\!\!\!\perp (z_i, e_i)$$

2. (Positivity) For all $i \in [n]$, $h \in [0, 1]$,

$$\mathbb{P}\{z_i = 1,\ e_i \geq h\} > 0 \quad \text{and} \quad \mathbb{P}\{z_i = 0,\ e_i \leq 1 - h\} > 0.$$

The first part of Assumption 4.4.1 states that the individual and neighborhood treatment assignments are independent of the potential outcomes. This follows from the unit-level and cluster-level Bernoulli randomization designs we consider in this paper. The second part of the assumption is also a fundamental one in the causal inference literature [Rosenbaum and Rubin, 1983, Petersen et al., 2012], ensuring that we have sufficient data to estimate the ATE. It also relates to statistical leverage [Mahoney et al., 2011, Martinsson and Tropp, 2020, Young, 2019, Pilanci and Wainwright, 2015]. The idea, informally, is that under positivity, statistical leverage across the input data is more homogeneous. As a result of this, there are no observations that have too much control over the linear regression.

Assumption 4.4.2 (Bounded variation treatment-exposure function)**.** Let the potential outcome function be $y_i = \alpha_i + \psi(z_i, e_i) + \epsilon_i$ as in (4.2.1). The function $\psi(z_i, e_i)$ has bounded variation.

Assumption 4.4.3 (Bounded First-Order Interactions)**.** For unit-level randomization, we assume that all nodes have bounded degree. Specifically, there exists a finite constant $d_{\max} < \infty$ such that, for every unit $i \in [n]$, the degree d_i satisfies

$$d_i \leq d_{\max}.$$

For cluster-level randomization, we assume that all clusters have bounded degrees of cross-cluster interactions. Let s_i denote the number of cross-cluster connections involving the cluster to which unit i belongs. Then, there exists a finite constant $s_{\max} < \infty$ such that, for every cluster $i \in \mathcal{C}$, the cross-cluster degree s_i satisfies

$$s_i \leq s_{\max}.$$

These restrictions are consistent with the observation that networks in practice tend to be sparse and clustered [Barabási, 2013, Chandrasekhar, 2016, Chandrasekhar et al., 2020]. If this assumption is violated, exposure positivity may fail, and Proposition 4.A.5 will no longer hold. However, as discussed in Remark 4.4.4 below, this assumption can be weakened under certain conditions. Some examples of graph classes that fall under the categorization of Assumption 4.4.3 are expander graphs, and growth-restricted graphs [Alon, 1986, Arora et al., 2009, Gkantsidis et al., 2003, Kuhn et al., 2005, Krauthgamer and Lee, 2003, Kowalski, 2019]. This is related to requiring the graph interference structure to have small k_n-conductance (i.e. the k_n-partition generalization of the Cheeger constant) for k_n growing with n. In the causal inference literature, the growth-restricted graph setting was studied in [Ugander et al., 2013].

Remark 4.4.4. Assumption 4.4.3 is also employed by Aronow and Samii [2017]. As we further discuss in Appendix 4.A.8, it can be relaxed by "binning" the exposures and verifying whether the "affinity set" conditions of Chandrasekhar et al. [2023] are satisfied. Specifically, in more general weighted settings, one can partition exposures into bins indexed by $b = 1, 2, \ldots, K$,

each associated with an effective exposure level $\tilde{e}_b$. In this framework, node degrees need not be bounded, provided the affinity set conditions are met. These conditions subsume the approximate neighborhood interference (ANI) condition of Leung [2022a] when the maximum clique size in the network remains bounded. For cases with growing clique size, we refer the reader to Leung [2022a] for further discussion of the ANI framework.

In the following theorem, we characterize the probability of choosing the correct threshold under the best average linear fit, for which the exposure slope is γ^*. The correct threshold h^* under the best average linear fit is the threshold minimizing the sum of the true variance and the true squared-bias under $Y_i = \alpha + \beta z_i + \gamma^* e_i$. We write $M_n^*(h), \hat{M}_n(h)$ to denote the MSE under the true average best fit line and the MSE estimated by our methods, respectively, at threshold h and finite-population size n.

Theorem 4.4.5. *Suppose Assumptions 4.4.1, 4.4.2, and 4.4.3 hold. Further assume unit-level Bernoulli randomization with treatment assignment probability p. Let $\Delta_n := \min_{h\in H\setminus\{h^*\}} |M_n^*(h) - M_n^*(h^*)|$, and let $U_n \in \mathbb{R}$ satisfy $\max_h |b_n^*(\hat{\tau}_h)| \le U_n$, which exists by Assumptions 4.4.1(2) and 4.4.2. Denote by h_n^* the optimal threshold under the true average best-fit line, and by $\hat{h}_n$ the threshold chosen by our method. Finally, let H be the set of exposures. Then,*

$$\begin{aligned}\mathbb{P}\{\hat{h}_n \neq h_n^*\} &\le \sum_{h\in H} \mathbb{P}\{|\hat{M}_n(h) - M_n^*(h)| > \Delta_n/2\} \\ &\le 3|H| \max\Bigg\{ \exp\Bigg(-\frac{\Delta_n np(1-p)}{8cd_{\max}} + 1\Bigg), \quad \exp\Bigg(-\frac{\Delta_n^2 np(1-p)}{(16U_n)^2 cd_{\max}} + 1\Bigg), \\ &\qquad 6\exp\Bigg(-\frac{Cn(\frac{\Delta_n}{4} - \frac{cd_{\max}^2}{n})}{\sqrt{A_{n,2}} + \sqrt{(\frac{\Delta_n}{4} - \frac{cd_{\max}^2}{n})M_{n,2}}}\Bigg)\Bigg\},\end{aligned}$$

for some constants c, C. Here, $A_{n,p} \le 16\|\tilde{v}\|_{L_1}^2\{c_1 + \frac{(\log n)^4}{n}\}^2$, for some constant c_1 and $\tilde{v}$ is the Fourier transform of the summands of the variance components, and $M_{n,2} = 4\|\tilde{v}\|_{L_1}(\log n)^4$.

If, instead, the design is cluster-level Bernoulli randomization with probability p, denote

by $s_{\max}$ the maximum cross-cluster degree, i.e., the largest number of distinct clusters to which any given cluster is connected. Then, we can replace $d_{\max}$ by $s_{\max}$ in the bounds above.

In Theorem 4.4.5, we assume without loss of generality that the MSE gap is lower-bounded by a constant, i.e. $\Delta_n \geq \Delta$, as otherwise all candidate thresholds are optimal. The proof to the theorem is in Appendix 4.A.5.1. In this proof, we make use of the results from [Ziemann et al., 2024, Theorem 3.1], and [Shen et al., 2020, Theorem 2.1].

Our results tell us that generally, for fixed n, and under a unit-level Bernoulli randomization, our approach yields a higher probability of choosing the optimal threshold when the maximum degree is smaller. Under a Bernoulli cluster-level randomization, our approach yields a higher probability of choosing the optimal threshold when the maximum cross-cluster degree is smaller. Therefore, for a well-clusterable graph, our approach would yield a higher probability of choosing the optimal threshold under cluster-level randomization, as the clustering would give us a reduction in the cross-cluster degree. Not surprisingly, when n is larger, and the degree ranges are kept fixed, the probability of choosing the optimal threshold is higher. Additionally, as the variance of the variance estimator is larger, the probability of picking the right threshold is smaller. The error scales with the size of the bias, appearing in the second term in the bound. Finally, considering symmetric h for treated and control units, our approach yields a higher probability of choosing the optimal threshold when the treatment assignment probability $p = 0.5$. This can be seen from the $p(1-p)$ term in the bound which is maximized at $p = 0.5$.

This gives us the following corollary. Denote the true MSE by M_n^{**}, and the corresponding bias and optimal threshold by b_n^{**} and h_n^{**}, respectively.

Corollary 4.4.6. *Under the conditions of Theorem 4.4.5, further assume that $\psi(z_i, e_i) = g(z_i) + f(e_i)$, and that $\sup_{e_i} |f(e_i) - \gamma^* e_i| \leq \delta$. Let $U_n^* \in \mathbb{R}$ be such that $\max_h |b_n^{**}(\hat{\tau}_h)| \leq U_n^*$, which exists by Assumptions 4.4.1 (2) and 4.4.2. Let $\tilde{\delta} := 16\delta^2 + 8\delta U_n^*$. Then,*

$$\mathbb{P}\{\hat{h}_n \neq \hat{h}_n^{**}\} \leq \sum_{h \in H} \mathbb{P}\{|\hat{M}_n(h) - M_n^{**}(h)| > \Delta_n/2\}$$

$$\leq 3|H| \max \Bigg\{ \exp\Bigg(- \frac{(\Delta_n/8 - \tilde{\delta})np(1-p)}{cd_{\max}} + 1 \Bigg),$$
$$\exp\Bigg(- \frac{(\Delta_n^2/(16U_n)^2 - \tilde{\delta})np(1-p)}{cd_{\max}} + 1 \Bigg),$$
$$6\exp\Bigg(- \frac{Cn(\frac{\Delta_n}{4} - \frac{cd_{\max}^2}{n} - \tilde{\delta})}{\sqrt{A_{n,2}} + \sqrt{(\frac{\Delta_n}{4} - \frac{cd_{\max}^2}{n} - \tilde{\delta})M_{n,2}}} \Bigg) \Bigg\},$$

for some constants c, C*. Here,* $A_{n,p} \leq 16\|\tilde{v}\|_{L_1}^2\{c_1 + \frac{(\log n)^4}{n}\}^2$*, for some constant* c_1 *and* $\tilde{v}$ *is the Fourier transform of the summands of the variance components, and* $M_{n,2} = 4\|\tilde{v}\|_{L_1}(\log n)^4$.

If, instead, the design is cluster-level Bernoulli randomization with probability p*, denote by* $s_{\max}$ *the maximum cross-cluster degree, i.e., the largest number of distinct clusters to which any given cluster is connected. Then, we can replace* $d_{\max}$ *by* $s_{\max}$ *in the bounds above.*

Corollary 4.4.6 tells us that if the maximum linear approximation error between the best average linear fit and the true exposure function $f(\cdot)$ is small enough relative to the minimum MSE gap Δ_n, our estimator will be optimal with high probability for large n. Indeed, when the $n(\Delta_n - \tilde{\delta})$ terms are large, the $\exp(-n(\Delta_n - \tilde{\delta}))$ terms are small. If $f(x) = \gamma x$ for all $x \in [0,1]$ indeed, then with necessarily we have that with high probability for large n, our estimator is optimal.

4.5 Experiments

We now compare the performance of our approach with that of non-adaptive Horvitz-Thompson estimators $HT(1)$ and $HT(0)$, with $h = 1$, and $h = 0$, respectively. Additionally, since our objective is essentially an MSE-optimal bandwidth selection problem for the indicator kernel (see Appendices 4.A.3, 4.A.7), we also compare our estimator to a variant where the threshold is selected via Lepski's method [Goldenshluger and Lepski, 2011, Su et al., 2020]. We note that Lepski's method relies on the monotonicity of the bias and the variance to work well. In our setting, due to the implicit dependency structure from the graph appearing in the exposures e_i, the monotonicity assumption may be violated. We

write out these three estimators we compare against in Appendix 4.A.2.

In Figure 4.5.2, we display simulation results with outcomes generated from the linear model with $\psi(z_i, e_i) = g(z_i) + f(e_i)$, $\alpha_i = 10$, $g(z_i) = \beta z_i = 10 z_i$, $f(e_i) = \gamma e_i$, with fixed ϵ_i generated from $\mathcal{N}(0, 1)$, for a 200-node graph, generated from an SBM with 40 underlying clusters (see Figure 4.5.1 (left panel)). We focus on varying the ratio γ/β as we consider a fixed graph. We see that our adaptive thresholding estimator, AdaThresh, generally performs better than existing estimators, interpolating between the fixed 0/1-threshold Horvitz-Thompson estimators, and out-performing the Lepski-based Horvitz-Thompson estimators when the threshold γ/β is high. We also display simulation results for 1000-node 2nd-power cycle graphs in Figure 4.A.3.

Furthermore, cluster-randomized designs better control interference, reducing bias over wider ranges of spillover ratios. In such settings, the variance tends to dominate the bias. Since optimizing for bias leads to $HT(1)$, and optimizing for variance leads to $HT(0)$, we observe better intermediate thresholds for longer ranges of spillover ratios under cluster-randomized designs compared to unit-level designs. These patterns become even more pronounced in the cycle graphs presented in Figure 4.A.3, where the node degree is held constant. From the design perspective (see, for instance, [Viviano et al., 2023]), if spillover ratios are larger, then one might analogously perform more cluster-randomized designs to contain more of the interference and reduce bias.

4.5.1 Real Data

We evaluate the performance of our estimator on village (No.6) network data from [Banerjee et al., 2013a]. Here, $n = 110$, with nodes representing villagers and edges indicating that the adjacent villagers have visited each other's homes. We ran experiments with synthetic potential outcomes, averaging over 1000 trials, under unit-level and cluster-level Bernoulli randomizations, where clusters were formed using an $\epsilon(= 1)$-net clustering (see [Ugander et al., 2013] for details on ϵ-net clustering). We generate simulated data using the linear

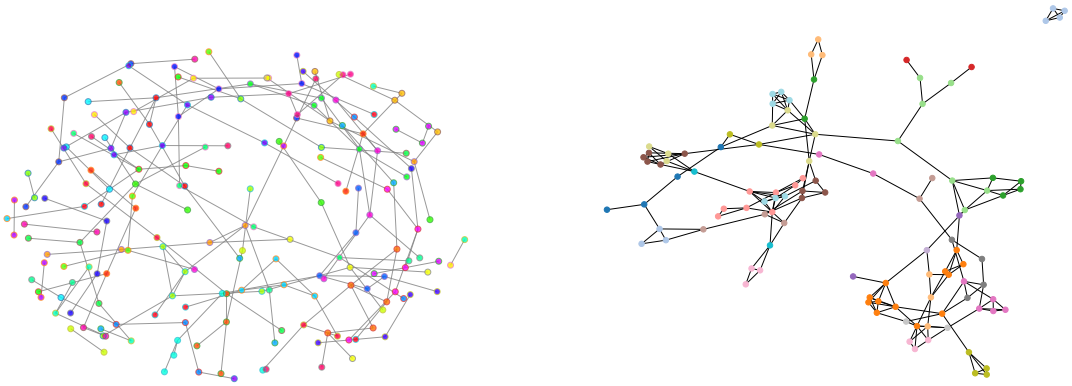


Figure 4.5.1: Left: A graph of size $n = 200$ generated from the Stochastic Block Model (SBM). Node colors reflect ground-truth cluster membership, corresponding to the underlying block membership. Right: Village network dataset, with node colors reflecting $\epsilon = 1$-net clustering (see [Ugander et al., 2013] for details on ϵ-net clustering).

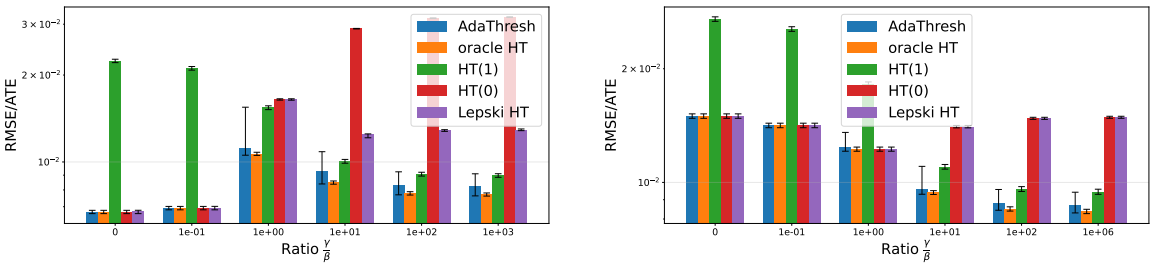


Figure 4.5.2: RMSE (normalized by the ATE) of different Horvitz-Thompson estimators for the SBM graph in Figure 4.5.1 (left panel). Left: unit-level Ber(0.5) randomization. Right: cluster-level Ber(0.5) randomization with ground truth clusters. The error bars are the empirical 95% confidence intervals around the mean.

model with $\psi(z_i, e_i) = g(z_i) + f(e_i)$, $\alpha_i = 10, g(z_i) = \beta z_i = 10 z_i$, $f(e_i) = \gamma e_i$, with fixed ϵ_i generated from $\mathcal{N}(0, 1)$. To compute the exposure probabilities, we used 2×10^4 Monte-Carlo trials. We focus on varying the ratio γ/β as we consider a fixed graph. Figure 4.5.3 demonstrates that our method generally interpolates between HT(1) and HT(0), achieving intermediate performance on smaller graphs (i.e., smaller n). While it does not consistently outperform the HT(0) baseline across all settings, the results suggest that our approach remains competitive across a large range of regimes and offers a reasonable bias–variance trade-off. We hypothesize that the observed imprecision and inaccuracy of our estimator stems from inaccuracies in estimating smaller exposure probabilities, particularly for higher-degree nodes. For larger n and smaller $d_{\max}$, performance improves further, supporting our theoretical findings, as demonstrated in Appendix 4.A.7 on the Amazon (DVD) products similarity network [Leskovec et al., 2007] (see Figure 4.A.2), and on various circulant graphs.

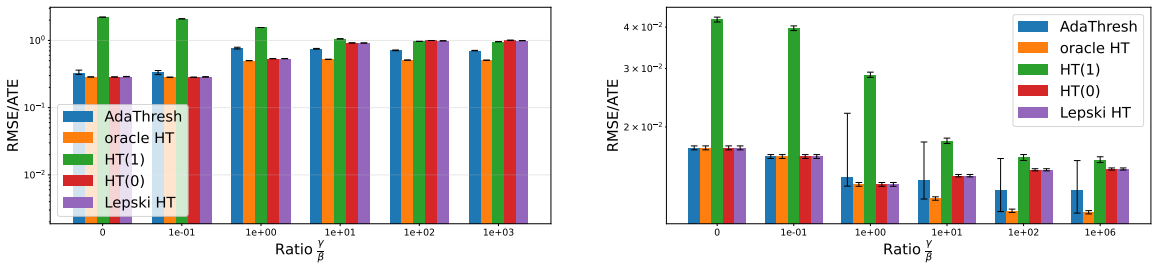


Figure 4.5.3: RMSE (normalized by the ATE) of different Horvitz-Thompson estimators for the village network in Figure 4.5.1 (right panel). Left: unit-level Ber(0.5) randomization. Right: cluster-level Ber(0.5) randomization with $\epsilon(= 1)$-net clustering. The error bars are the empirical 95% confidence intervals around the mean.

4.6 Discussion

In this paper, we focused on the additive model for ψ, but note that our framework is applicable more broadly. We investigated AdaThresh, an adaptive Horvitz-Thompson estimator for symmetric thresholds h and $1 - h$ for treatment and control, respectively. Our framework applies more generally with estimator thresholds h_1 and $1 - h_0$, respectively

with $h_1 \neq h_0$. Additionally, in Appendix 4.A.7.5, we demonstrate the performance of this approach using the Difference-in-Means estimator incorporating exposure thresholds. We leave it to future work to extend our results to more general exposure estimation procedures. In [Chin, 2019], the author proposes learning the feature variables that are most predictive of the potential outcomes. One could extend our work by first learning the feature variables, as proposed by [Chin, 2019], followed by then learning the appropriate optimal threshold associated with these features, using our approach.

In Appendix 4.A.7.4, we illustrate the robustness of our estimator to deviations from linearity in the potential outcomes model. We note that our framework would apply in broader settings, including a direct extension to friends-of-friends (or further connections) or to neighborhoods that are either observed or learned through a clustering algorithm. An additional interesting future direction could be to extend our work to the framework presented in Chandrasekhar et al. [2023], allowing for a complete graph underlying the interference structure.

Furthermore, to improve upon robustness to non-linear settings, we propose an extension using local regression to estimate the rate of change of bias within the $[0, 1-h], [h, 1]$ windows. As long as there is sufficient concentration of exposures in these windows, robustness is never worse. In Appendix 4.A.7.6, we display our simulation results in this setting using local linear regression to estimate the rate of change of bias within the $[0, 1-h], [h, 1]$ windows. One could also use other local regression approaches, such as kernel regression, etc., while maintaining sufficiently small model complexity to avoid needing to sample-split.

This paper prioritizes minimizing the mean squared error (MSE) to effectively balance the bias–variance trade-off in a Horvitz-Thompson-type estimator, rather than focusing on inference. However, we acknowledge the significance of characterizing inferential tools, such as confidence intervals, in specific contexts. Notably, [Aronow and Samii, 2017] and [Athey et al., 2018] offer valuable insights into inferential methods for estimating average treatment effects under network interference.

Acknowledgements

The authors would like to thank Alex Kokot and Lars van der Laan for helpful discussions; Dean Eckles for pointing us to a reference, and Matthew Eichhorn for suggesting the Amazon product similarity network dataset.

4.A Appendix / supplemental material

Notation

In this supplement, we write $b^*(\tau_h), \hat{\tau}_b(h), v^*(\tau_h), \hat{v}(\tau_h)$ to represent the true bias, estimated bias, true variance, and estimated variance, respectively, at threshold h. We also write $d_{\max}$ to represent the upper-bound (Assumption 4.4.3) on the degrees of the nodes in the network.

4.A.1 Variance estimator

We use the variance estimator, for the Horvitz-Thompson estimator, proposed by [Aronow and Samii, 2017, Eqn. 7,9].

$$
\begin{aligned}
\hat{v}(\hat{\tau}_h) = &\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 1, e_i \geq h\} Y_i^2}{n^2 \pi_i^1} \left(\frac{1}{\pi_i^1} - 1\right) \\
&+ \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\mathbf{1}\{z_i = 1, e_i \geq h\} \mathbf{1}\{z_j = 1, e_j \geq h\} Y_i Y_j}{n^2 \pi_{ij}^{11}} \left(\frac{\pi_{ij}^{11}}{\pi_i^1 \pi_j^1} - 1\right) \\
&+ \sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 0, e_i \leq 1 - h\} Y_i^2}{n^2 \pi_i^0} \left(\frac{1}{\pi_i^0} - 1\right) \\
&+ \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\mathbf{1}\{z_i = 0, e_i \leq 1 - h\} \mathbf{1}\{z_j = 0, e_j \leq 1 - h\} Y_i Y_j}{n^2 \pi_{ij}^{00}} \left(\frac{\pi_{ij}^{00}}{\pi_i^0 \pi_j^0} - 1\right) \\
&- \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j \in [n]; \pi_{ij}^{10} > 0} \left(\mathbf{1}\{z_i = 1, e_i \geq h\} \mathbf{1}\{z_j = 0, e_j \leq 1 - h\} Y_i Y_j\right) \left(\frac{1}{\pi_i^1 \pi_j^0} - \frac{1}{\pi_{ij}^{10}}\right) \\
&+ \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j \in [n]; \pi_{ij}^{10} = 0} \left(\frac{\mathbf{1}\{z_i = 1, e_i \geq h\} Y_i^2}{2\pi_i^1} + \frac{\mathbf{1}\{z_j = 0, e_j \leq 1 - h\} Y_j^2}{2\pi_j^0}\right)
\end{aligned}
\tag{4.A.1}
$$

Here, π_i^x, π_{ij}^{xy} are defined with respect to the threshold h. That is, $\pi_i^1 = \mathbb{P}\{z_i = 1, e_i \geq h\}$, $\pi_i^0 = \mathbb{P}\{z_i = 0, e_i \leq 1 - h\}$, $\pi_{ij}^{10} = \mathbb{P}\{z_i = 1, e_i \geq h, z_j = 0, e_j \leq 1 - h\}$, $\pi_{ij}^{11} = \mathbb{P}\{z_i = 1, e_i \geq h, z_j = 1, e_j \geq h\}$, $\pi_{ij}^{01} = \mathbb{P}\{z_i = 0, e_i \leq 1 - h, z_j = 1, e_j \geq h\}$, $\pi_{ij}^{00} = \mathbb{P}\{z_i = 0, e_i \leq 1 - h, z_j = 0, e_j \leq 1 - h\}$.

4.A.2 Horvitz-Thompson estimators with existing approaches

We write the vanilla Horvitz-Thompson estimator (at threshold one) 4.A.2, the Horvitz-Thompson estimator at threshold zero 4.A.3, and the Horvitz-Thompson estimator with the threshold selected via Lepski's method 4.A.6 below.

4.A.2.1 Vanilla Horvitz-Thompson estimator (at threshold one)

$$\hat{\tau}_{\mathrm{HT}_1} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 1, e_i = 1\}}{\mathbb{P}\{z_i = 1, e_i = 1\}} Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 0, e_i = 0\}}{\mathbb{P}\{z_i = 0, e_i = 0\}} Y_i \tag{4.A.2}$$

4.A.2.2 Horvitz-Thompson estimator at threshold zero

$$\hat{\tau}_{\mathrm{HT}_0} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 1\}}{\mathbb{P}\{z_i = 1\}} Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 0\}}{\mathbb{P}\{z_i = 0\}} Y_i \tag{4.A.3}$$

4.A.2.3 Lepski's method

As described in [Lepski and Spokoiny, 1997], we first take

$$I(h) := [\hat{\tau}_{\mathrm{HT}_h} - 2\widehat{\mathrm{SDEV}}(\hat{\tau}_{\mathrm{HT}_h}), \hat{\tau}_{\mathrm{HT}_h} + 2\widehat{\mathrm{SDEV}}(\hat{\tau}_{\mathrm{HT}_h})]. \tag{4.A.4}$$

Then, take

$$\hat{h}_{\mathrm{Lepski}} := \min\{h \in H : \cap_{h' \in H : h' \geq h} I(h') \neq \emptyset\}, \tag{4.A.5}$$

and

$$\hat{\tau}_{\mathrm{LepskiHT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 1, e_i \geq \hat{h}_{\mathrm{Lepski}}\}}{\mathbb{P}\{z_i = 1, e_i \geq \hat{h}_{\mathrm{Lepski}}\}} Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 0, e_i \leq 1 - \hat{h}_{\mathrm{Lepski}}\}}{\mathbb{P}\{z_i = 0, e_i \leq 1 - \hat{h}_{\mathrm{Lepski}}\}} Y_i \tag{4.A.6}$$

Lepski's method requires monotonicity and decay rate conditions to be satisfied. Our setting may lead to violations of these conditions.

4.A.3 An equivalent formulation of the Horvitz-Thompson estimator with the exposure threshold

We can rewrite the display in 4.2.4 in the following form:

$$\hat{\tau}_h = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{|z_i - e_i| \leq 1 - h\}}{\mathbb{P}\{|z_i - e_i| \leq 1 - h \mid z_i\}} \tilde{z}_i Y_i, \tag{4.A.7}$$

where $\tilde{z}_i = 2z_i - 1$, so that $\tilde{z}_i \in \{-1, 1\}$. It is then clear that the threshold h controls how much dissimilarity between the treatment status of units and their neighbors, we allow in our estimation. This allows us to reframe the problem as an optimal bandwidth selection one.

4.A.4 Bias and Variance estimation errors

In this subsection, we write down the estimation errors of the bias and the variance terms in the MSE estimation.

For the bias terms, we have,

$$\begin{aligned}\hat{b}(\hat{\tau}_h) - b^*(\hat{\tau}_h) = &\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_n(1-e_i)\mathbf{1}\{z_i=1, e_i \geq h\}}{\mathbb{P}(z_i=1, e_i \geq h)} - \frac{1}{n}\sum_{i=1}^{n}\sum_{x_i \in X_i}\frac{\mathbf{1}\{x_i \geq h\}}{|x_i : x_i \in X_i \cap x_i \geq h|}\gamma_n^*(1-x_i)\Big) \\ &+ \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_n e_i \mathbf{1}\{z_i=0, e_i \leq 1-h\}}{\mathbb{P}(z_i=0, e_i \leq 1-h)} - \frac{1}{n}\sum_{i=1}^{n}\sum_{x_i \in X_i}\frac{\mathbf{1}\{x_i \leq 1-h\}}{|x_i : x_i \in X_i \cap x_i \leq 1-h|}\gamma_n^*(x_i)\Big),\end{aligned}$$

where γ is the slope of the best average linear fit, and where x_i ranges over the set X_i of possible fractions of degree i.

By an abuse of notation, we use $y_i(h^+)$ to represent the average (across possible exposure fractions for node i) potential outcome for unit i with exposures that are at least h, while we use $y_i(h^-)$ to represent the average (across possible exposure fractions for node i) potential outcome for unit i with exposures that are at most $1-h$.

In the variance terms,

$$\begin{aligned}v^*(\hat{\tau}_h) = &\sum_{i=1}^{n}\frac{y_i(h^+)^2}{n^2}\left(\frac{1}{\pi_i^1} - 1\right) \\ &+ \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\frac{y_i(h^+)y_j(h^+)}{n^2}\left(\frac{\pi_{ij}^{11}}{\pi_i^1\pi_j^1} - 1\right) \\ &+ \sum_{i=1}^{n}\frac{y_i(h^-)^2}{n^2\pi_i^0}\left(\frac{1}{\pi_i^0} - 1\right) \\ &+ \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\frac{y_i(h^-)y_j(h^-)}{n^2}\left(\frac{\pi_{ij}^{00}}{\pi_i^0\pi_j^0} - 1\right) \\ &- \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\in[n];\pi_{ij}^{10}>0} y_i(h^+)y_j(h^-)\left(\frac{\pi_{ij}^{10}}{\pi_i^1\pi_j^0} - 1\right)\end{aligned}$$

$$+\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\in[n];\pi_{ij}^{10}=0} y_i(h^+)y_j(h^-).$$

Therefore, from the above and Section 4.A.1, we have that

$$\begin{aligned}
&\sup_h \hat{v}(\hat{\tau}_h) - v^*(\hat{\tau}_h) \\
&= \sup_h \Bigg[\left(\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i=1, e_i \geq h\} Y_i^2}{n^2 \pi_i^1} \left(\frac{1}{\pi_i^1} - 1 \right) - \sum_{i=1}^{n} \frac{y_i(h^+)^2}{n^2} \left(\frac{1}{\pi_i^1} - 1 \right) \right) \\
&\quad + \left(\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\mathbf{1}\{z_i=1, e_i \geq h\} \mathbf{1}\{z_j=1, e_j \geq h\} Y_i Y_j}{n^2 \pi_{ij}^{11}} \left(\frac{\pi_{ij}^{11}}{\pi_i^1 \pi_j^1} - 1 \right) \right. \\
&\qquad \left. - \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{y_i(h^+) y_j(h^+)}{n^2} \left(\frac{\pi_{ij}^{11}}{\pi_i^1 \pi_j^1} - 1 \right) \right) \\
&\quad + \left(\sum_{i=1}^{n} \frac{\mathbf{1}\{z_i=0, e_i \leq 1-h\} Y_i^2}{n^2 \pi_i^0} \left(\frac{1}{\pi_i^0} - 1 \right) - \sum_{i=1}^{n} \frac{y_i(h^-)^2}{n^2 \pi_i^0} \left(\frac{1}{\pi_i^0} - 1 \right) \right) \\
&\quad + \left(\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\mathbf{1}\{z_i=0, e_i \leq 1-h\} \mathbf{1}\{z_j=0, e_j \leq 1-h\} Y_i Y_j}{n^2 \pi_{ij}^{00}} \left(\frac{\pi_{ij}^{00}}{\pi_i^0 \pi_j^0} - 1 \right) \right. \\
&\qquad \left. - \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{y_i(h^-) y_j(h^-)}{n^2} \left(\frac{\pi_{ij}^{00}}{\pi_i^0 \pi_j^0} - 1 \right) \right) \\
&\quad - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{\substack{j\in[n] \\ \pi_{ij}^{10}>0}} \left(\mathbf{1}\{z_i=1, e_i \geq h\} \mathbf{1}\{z_j=0, e_j \leq 1-h\} Y_i Y_j \left(\frac{1}{\pi_i^1 \pi_j^0} - \frac{1}{\pi_{ij}^{10}} \right) \right. \\
&\qquad \left. - y_i(h^+) y_j(h^-) \left(\frac{\pi_{ij}^{10}}{\pi_i^1 \pi_j^0} - 1 \right) \right) \\
&\quad + \frac{2}{n^2} \sum_{i=1}^{n} \sum_{\substack{j\in[n] \\ \pi_{ij}^{10}=0}} \left(\frac{\mathbf{1}\{z_i=1, e_i \geq h\} Y_i^2}{2\pi_i^1} + \frac{\mathbf{1}\{z_j=0, e_j \leq 1-h\} Y_j^2}{2\pi_j^0} \right. \\
&\qquad \left. - y_i(h^+) y_j(h^-) \right) \Bigg].
\end{aligned}$$

4.A.5 Proofs to Theorem 4.4.5 and Corollary 4.4.6

4.A.5.1 Proof to Theorem 4.4.5

Proof to Theorem 4.4.5. We first consider the variance terms. Define $\bar{v} = \mathbb{E}[\hat{v}]$. We have that,

$$\begin{aligned}
\mathbb{P}(|\hat{v} - v^*| > \Delta_n/4) &= \mathbb{P}(|\hat{v} - \bar{v} + \bar{v} - v^*| > \Delta_n/4) \\
&\overset{(a)}{=} \mathbb{P}(|\hat{v} - \bar{v}| > \Delta_n/4 - cd_{\max}^2/n) \\
&\overset{(b)}{\leq} 6\exp\left(-\frac{Cn\left(\frac{\Delta_n}{4} - \frac{cd_{\max}^2}{n}\right)}{\sqrt{A_{n,2}} + \sqrt{\left(\frac{\Delta_n}{4} - \frac{cd_{\max}^2}{n}\right)M_{n,2}}}\right)
\end{aligned}$$

where $A_{n,p} \leq 16\|\tilde{v}\|_{L_1}^2\{c_1 + \frac{(\log n)^4}{n}\}^2$, for some constant c_1 and $\tilde{v}$ is the Fourier transform of the variance summands, and $M_{n,2} = 4\|\tilde{v}\|_{L_1}(\log n)^4$.

The equality (a) is obtained as a result of Assumption 4.4.3. Indeed, we know that $\bar{v} - v^*$ is non-zero only when there exists i, j such that $\pi_{ij}(h^+, h^-) = 0$, i.e. the joint exposure assignment probability of $e_i \geq h$, and $e_i \leq 1 - h$. Since the contribution of these to $\bar{v} - v^*$ are bounded above by $cd_{\max}^2/n$ for some constant c, we get (a). The inequality (b) is obtained from a direct application of [Shen et al., 2020, Theorem 2.1], as it easy to see that our exposure dependencies satisfy α-mixing conditions and that together with Assumption 2, the variance summands satisfy the Fourier transform conditions of [Shen et al., 2020, Theorem 2.1].

Next, we consider the bias terms, for any constant J,

$$\mathbb{P}((\hat{\gamma}_n - \gamma_n^*)^2 > \Delta_n/J) \leq \exp\left(-\frac{\Delta_n np(1-p)}{Jcd_{\max} + 1}\right),$$

for some constant c. We use the results from [Ziemann et al., 2024, Theorem 3.1]. Below, we verify that the conditions are met in our setting. We begin with the condition that for every $v \in \mathbb{R}$ such that $v = 1/(\mathbb{E}[e_i^2])^{1/2}$, there exists $h \in \mathbb{R}^+$ such that, $\mathbb{E}[(ve_i)^4] \leq h^2v^2\mathbb{E}[e_i^2]$. This is trivially satisfied in our setting since our exposures is bounded. Next, we consider the first part of [Ziemann et al., 2024, Condition (3.3)]. This is also trivially satisfied in our setting since our block sizes ($= d_{\max}^2$) are bounded by Assumption 4.4.3. The second part of [Ziemann et al., 2024, Condition (3.3)] is also satisfied since our de-meaned noise-class interaction variables (as defined in [Ziemann et al., 2024, Eq.(2.5)]) are sub-gaussian by assumption of independence between the potential outcomes' residuals

and the treatment design. Next, [Ziemann et al., 2024, Condition (3.4)] is also trivially satisfied since we take our blocks to be of size $d_{\max}^2$ almost, with the possible exception of the final remaining block, uniformly. Finally, it is easy to see that [Ziemann et al., 2024, Condition (3.5)] is also satisfied since the exposure variables are independent outside of the radius $d_{\max}^2$, since exposure variables e_i and e_j for any $i, j \in [n]$ are only dependent when nodes i and j share a neighbor. We quickly note that the monotone partitioning (as described in [Ziemann et al., 2024, Theorem 3.1] is not necessary in our setting, as the "blocking" procedure introduced by [Yu, 1994] holds more generally.

For convenience, in the following we write $b^*, \hat{b}$ to mean $b^*(\hat{\tau}_h), \hat{b}(\hat{\tau}_h)$, and similarly $v^*, \hat{v}$ to mean $v^*(\hat{\tau}_h), \hat{v}(\hat{\tau}_h)$. H is the set of exposures. Therefore, putting these together by a union bound,

$$
\begin{aligned}
\mathbb{P}\{\hat{h}_n \neq h_n^*\} &\leq \sum_h \mathbb{P}\{|\hat{M}_n(h) - M_n^*(h)| > \Delta_n/2\} \\
&\leq |H|\mathbb{P}\{|\hat{b}^2 + \hat{v} - {b^*}^2 - v^*| > \Delta_n/2\} \\
&\leq |H|\big(\mathbb{P}\{|\hat{b}^2 - {b^*}^2| > \Delta_n/4\} + \mathbb{P}\{|\hat{v} - v^*| > \Delta_n/4\}\big) \\
&\leq |H|\big(\mathbb{P}\{(\hat{b} - b^*)^2 > \Delta_n/8\} + \mathbb{P}\{(\hat{b} - b^*)^2 > \Delta_n^2/(16U_n)^2\} \\
&\quad + \mathbb{P}\{|\hat{v} - v^*| > \Delta_n/4\}\big) \\
&\leq |H|\big(\mathbb{P}\{(\hat{\gamma}_n - \gamma_n^*)^2 > \Delta_n/8\} + \mathbb{P}\{(\hat{\gamma}_n - \gamma_n^*)^2 > \Delta_n^2/(16U_n)^2\} \\
&\quad + \mathbb{P}\{|\hat{v} - v^*| > \Delta_n/4\}\big) \\
&\leq 3|H| \max\Bigg\{ \exp\bigg(- \frac{\Delta_n np(1-p)}{8cd_{\max}} + 1\bigg), \\
&\quad \exp\bigg(- \frac{\Delta_n^2 np(1-p)}{(16U_n)^2 cd_{\max}} + 1\bigg), \\
&\quad 6\exp\Bigg(- \frac{Cn\big(\frac{\Delta_n}{4} - \frac{cd_{\max}^2}{n}\big)}{\sqrt{A_{n,2}} + \sqrt{\big(\frac{\Delta_n}{4} - \frac{cd_{\max}^2}{n}\big)M_{n,2}}}\Bigg)\Bigg\}.
\end{aligned}
$$

The proof for the cluster-level Bernoulli randomization follows immediately.

■

4.A.5.2 Proof to Corollary 4.4.6

Proof of Corollary 4.4.6. Suppose $\sup_{e_i} |f(e_i) - \gamma^* e_i| \leq \delta$. Let $\tilde{\delta} := 16\delta^2 + 8\delta U_n^*$. We begin by considering the difference between the MSE under the true best average linear fit and the true MSE:

$$\begin{aligned}
|M_n^*(h) - M_n^{**}(h)| &= |{b_n^*}^2(h) - {b_n^{**}}^2(h)| \\
&\leq |(b_n^*(h) - b_n^{**}(h))(b_n^*(h) + b_n^{**}(h))| \\
&\leq |(b_n^*(h) - b_n^{**}(h))\big((b_n^*(h) - b_n^{**}(h)) + 2b_n^{**}(h)\big)| \\
&\leq \big((b_n^*(h) - b_n^{**}(h))\big)^2 + 2|b_n^*(h) - b_n^{**}(h)|U_n^* \\
&\leq (4\delta)^2 + 2U_n^*(4\delta) \equiv \tilde{\delta}.
\end{aligned}$$

This gives us,

$$\begin{aligned}
&\mathbb{P}\{\hat{h}_n \neq \hat{h}_n^{**}\} \\
&\leq \sum_{h \in H} \mathbb{P}\{|\hat{M}_n(h) - M_n^{**}(h)| > \Delta_n/2\} \\
&= \sum_{h \in H} \mathbb{P}\{|\hat{M}_n(h) - M_n^*(h) + M_n^*(h) - M_n^{**}(h)| > \Delta_n/2\} \\
&\leq \sum_{h \in H} \mathbb{P}\{|\hat{M}_n(h) - M_n^*(h)| + |M_n^*(h) - M_n^{**}(h)| > \Delta_n/2\} \\
&\leq \sum_{h \in H} \mathbb{P}\{|\hat{M}_n(h) - M_n^*(h)| > \Delta_n/2 - \tilde{\delta}\}
\end{aligned}$$

where the second inequality is obtained by applying a triangle inequality.

Finally, applying Theorem 4.4.5 gives us the result. ■

4.A.6 Toy examples: Tradeoffs in Circulant graphs

We consider unit-level and cluster-level Bernoulli(p) randomizations in the kth-power cycle graphs, see Figure 4.A.1. We say a graph is a kth-power cycle graph if there exists an edge between each node and $2k$ of its nearest neighbors [Ugander et al., 2013].

Proposition 4.A.1 (Absolute bias in k-th power cycle graphs under unit-randomization)**.** *When $p = 1/2$ and the potential outcome model is simply linear, i.e. $Y_i = \alpha + \beta z_i + \gamma e_i$, the absolute bias of the Horvitz-Thompson estimator for a given threshold $h \equiv l/2k$ for $l = 0, 2, ..., 2k$, in the kth-power*

cycle graph, under unit-level randomization, is equal to $\gamma \times \left[\frac{\sum_{r=l}^{2k}(r/k-1)\binom{2k}{r}}{\sum_{r=l}^{2k}\binom{2k}{r}} - 1\right]$.

Proposition 4.A.2 (Variance in k-th power cycle graphs under unit-randomization)**.** *Denote the degree of the nodes by* $d(=2k)$*. When* $p = 1/2$ *and the potential outcome model is simply linear, i.e.* $Y_i = \alpha + \beta z_i + \gamma e_i$*, the variance of the Horvitz-Thompson estimator for a given threshold* h*, in the* k*-th power cycle graph, under unit-level randomization is proportional to*

$$\mathcal{O}\left(\frac{1}{np^{dh}}\left[(\alpha+\beta+\gamma dh)^2 + (\alpha+\beta+\gamma d(1-h))^2 - 2\gamma h(1-h)d\right]\right).$$

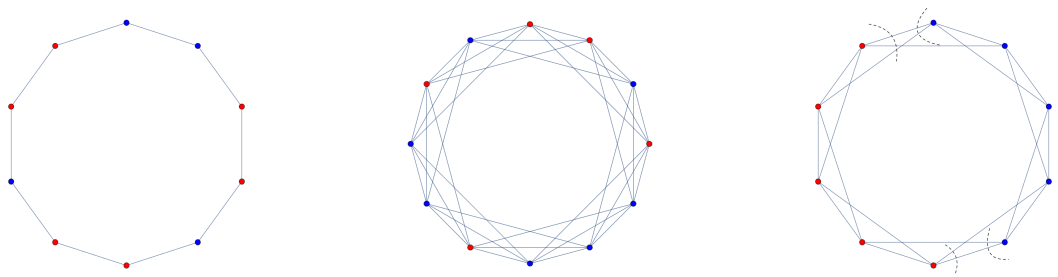


Figure 4.A.1: Circulant graphs with unit-level randomization and cluster-level randomization. Blue and red nodes represent treated and control units, respectively. Left: (1st-power) cycle graph with Ber(0.5) unit randomization. Center: 3rd-power cycle graph with Bernoulli(0.5) unit randomization. Right: 2nd-power cycle graph with Bernoulli(0.5) cluster randomization, with clusters of size 5 (=2k + 1).

Therefore, the optimal threshold h^* here depends on γ, β, n, p, d, the graph structure, and the number of other candidate thresholds.

From this toy example, it is not difficult to see that for unit-level Bernoulli randomization for more general graphs, the squared-bias would be $\Theta(\gamma^2 h^2)$, and the variance would be $\Theta(p^{-dh}\beta^2/n)$. Therefore, minimizing the MSE corresponds to balancing these quantities.

For kth-power cycle graphs, we also consider cluster-randomized design with cluster size $2k+1$. In [Ugander et al., 2013], the authors show that this clustering size minimizes variance under full neighborhood exposure. We state the approximate absolute bias and variance for this setting in the following propositions. Here, define the threshold $h = \frac{l}{d}$ with $l = 0, 1, ..., d$.

Proposition 4.A.3 (Absolute bias in k-th power cycle graph under cluster-randomization)**.** *When* $p = 1/2$*, and the potential outcome model is simply linear, i.e.* $Y_i = \alpha + \beta z_i + \gamma e_i$*, the bias of the*

Horvitz-Thompson estimator for a given threshold h in the k-th power cycle graph under cluster randomization, with cluster-sizes $2k+1$, is approximately $2\gamma(h-1)$ for $h \geq 1/2$.

Proposition 4.A.4 (Variance in the k-th power cycle graph under cluster-randomization)**.** *When $p = 1/2$ and the potential outcome model is simply linear, i.e. $Y_i = \alpha + \beta z_i + \gamma e_i$, the variance of the Horvitz-Thompson estimator for a given threshold h, in the k-th power cycle graph under cluster-level randomization, with cluster-sizes $2k+1$, is proportional to*

$$\frac{1}{np^2}(3d+1-2dh)[(\beta+\gamma h)^2+(\gamma(1-h))^2] = \Theta(\frac{\beta^2h^2}{np^{2h}}).$$

Under cluster randomization with cluster sizes $2k+1$ for the k-th power cycle graphs, the variance grows linearly in the degrees of the graph. Therefore, informally, compared to the unit-randomized design setting, higher node degrees lead to stronger bias than variance for a fixed exposure function $\psi(\cdot, e_i)$.

4.A.6.1 Proofs to Propositions 4.A.1, 4.A.2, 4.A.3, 4.A.4

Proof to Proposition 4.A.1. When $p = 1/2$ and the potential outcome model is simply linear, i.e. $Y_i = \alpha + \beta z_i + \gamma e_i$, the absolute bias of the Horvitz-Thompson estimator for a given threshold $h \equiv l/2k$ for $l = 0, 2, ..., 2k$, in the kth-power cycle graph under unit-randomization is equal to

$$\begin{aligned}
&\frac{1}{\mathbb{P}\{z_i=1, e_i \geq h\}} \sum_{x_i \in X} \frac{\mathbf{1}\{x_i \geq h\}}{|x_i : x_i \in X_i \cap x_i \geq h|} \mathbb{E}[\mathbf{1}\{z_i=1, e_i=x_i\}] y_i(x_i) \\
&- \frac{1}{\mathbb{P}\{z_i=0, e_i \leq 1-h\}} \sum_{x_i \in X} \frac{\mathbf{1}\{x_i \leq 1-h\}}{|\{x_i : x_i \in X_i \cap x_i \leq 1-h\}|} \mathbb{E}[\mathbf{1}\{z_i=0, e_i=x_i\}] y_i(x_i) \\
&- (\beta+\gamma) \\
&= \frac{1}{\sum_{r=l}^{2k} \binom{2k}{r} p^{2r}} \sum_{x_i \in X} \frac{\mathbf{1}\{x_i \geq h\}}{|\{x_i : x_i \in X_i \cap x_i \leq 1-h\}|} \mathbb{E}[\mathbf{1}\{z_i=1, e_i=x_i\}] y_i(x_i) \\
&- \frac{1}{\sum_{r=l}^{2k} \binom{2k}{r} p^{2r}} \sum_{x_i \in X} \frac{\mathbf{1}\{x_i \leq 1-h\}}{|\{x_i : x_i \in X_i \cap x_i \leq 1-h\}|} \mathbb{E}[\mathbf{1}\{z_i=0, e_i=x_i\}] y_i(x_i) \\
&- (\beta+\gamma) \\
&= \frac{1}{\sum_{r=l}^{2k} \binom{2k}{r}} \left(\sum_{r=l}^{2k} \binom{2k}{r} \beta + \binom{2k}{r} (r/k-1)\gamma \right) - (\beta+\gamma) \\
&= \frac{1}{\sum_{r=l}^{2k} \binom{2k}{r}} \left(\sum_{r=l}^{2k} \binom{2k}{r} (r/k-1)\gamma \right) - \gamma
\end{aligned}$$

where $X = \{0, 1/2k, 2/2k, ..., 1\}$. ■

Proof to Proposition 4.A.2. When $p = 1/2$ and the potential outcome model is simply linear, i.e. $Y_i = \alpha + \beta z_i + \gamma e_i$, it is not difficult to see that the variance of the Horvitz-Thompson estimator for a given threshold $h \equiv l/2k$, in the k-th power cycle graph under unit-level randomization is proportional to

$$\frac{1}{n}\sum_{l=0}^{2k}\binom{2k}{l}\mathbf{1}\{l/2k \geq h\}\Bigg[(\alpha+\beta+\gamma l/2k)^2\Big(\frac{1}{\sum_{l=0}^{2k}\binom{2k}{l}\mathbf{1}\{l/2k\geq h\}p^{2l}}-1\Big)$$
$$+\ 2(\beta+\gamma l/(2k))(\gamma(1-l/2k))\Big(2k+1-\frac{2k}{\sum_{l=0}^{2k}\binom{2k}{l}\mathbf{1}\{l/2k\geq h\}p^{2l}}\Big)$$
$$+\ (\alpha+\gamma(1-l/2k))^2\Big(\frac{1}{\sum_{l=0}^{2k}\binom{2k}{l}\mathbf{1}\{l/2k\geq h\}p^{2l}}-1\Big)\Bigg].$$

Considering the dominating terms gives us the result. ■

Proof to Proposition 4.A.3. When $p = 1/2$ and the potential outcome model is simply linear, i.e. $Y_i = \alpha + \beta z_i + \gamma e_i$, the absolute bias of the Horvitz-Thompson estimator for a given threshold $h \equiv l/2k$ for $l = k, k+1, ..., 2k$, in the kth-power cycle graph under cluster-randomization is equal to

$$\frac{1}{\mathbb{P}\{z_i=1, e_i\geq h\}}\sum_{x_i\in X}\frac{\mathbf{1}\{x_i\geq h\}}{|x_i : x_i\in X_i\cap x_i\geq h|}\mathbb{E}[\mathbf{1}\{z_i=1,e_i=x_i\}]y_i(x_i)$$
$$-\frac{1}{\mathbb{P}\{z_i=0, e_i\leq 1-h\}}\sum_{x_i\in X}\frac{\mathbf{1}\{x_i\leq 1-h\}}{|x_i : x_i\in X_i\cap x_i\leq 1-h|}\mathbb{E}[\mathbf{1}\{z_i=0,e_i=x_i\}]y_i(x_i)$$
$$-(\beta+\gamma)$$
$$=\frac{(p/(2k+1)+2k/(2k+1)\cdot p^2)(\beta+\gamma)+2p/(2k+1)\sum_{r=1}^{k}\mathbf{1}\{r\geq d-l\}(\beta+\gamma\cdot(2l-d)/d)}{p/(2k+1)+p^2\cdot 2k/(2k+1)+2p/(2k+1)\sum_{r=1}^{k}\mathbf{1}\{r\geq d-l\}}$$
$$-(\beta+\gamma)$$
$$=\mathcal{O}(2\gamma(h-2))$$

where $X = \{0, 1/2k, 2/2k, ..., 1\}$. When $l < k$, the bias scales is the same as when $l = k$. ■

Proof to Proposition 4.A.4. When $p = 1/2$ and the potential outcome model is simply linear, i.e. $Y_i = \alpha + \beta z_i + \gamma e_i$, it is not difficult to see that the variance of the Horvitz-Thompson estimator

for a given threshold $h \equiv l/2k$, in the k-th power cycle graph under cluster-randomization, with cluster-size $2k+1$, is proportional to

$$\sum_{u=0}^{d} \mathbf{1}\{u/d \geq h\} \binom{d}{u} \sum_{i=1}^{n} \left(\frac{(\beta + \gamma u/d)^2}{n^2}\right)\left[(3d+1-2l)\left(\frac{1}{p^2}-1\right) + (2l+1)\left(\frac{1}{p}-1\right)\right]$$
$$+ \sum_{u=0}^{d} \mathbf{1}\{u/d \geq h\} \binom{d}{u} \sum_{i=1}^{n} \left(\frac{(\gamma(\frac{d-u}{d}))^2}{n^2}\right)\left[(3d+1-2l)\left(\frac{1}{p^2}-1\right) + (2l+1)\left(\frac{1}{p}-1\right)\right]$$
$$- \sum_{u=0}^{d} \mathbf{1}\{u/d \geq h\} \binom{d}{u} \frac{2}{n}(\beta + \gamma \cdot u/d)\left(\gamma \cdot \frac{d-u}{d}\right).$$

Considering the dominating terms gives us the result. ■

4.A.7 Simulations

4.A.7.1 Experimental details and results on the Amazon product similarity graph data

We also evaluate the performance of our estimator on the Amazon (DVD) products similarity network [Leskovec et al., 2007]. We ran experiments with synthetic potential outcomes, averaging over 1000 trials, under unit-level randomization and spectral cluster-level randomizations. We generate simulated data using the linear model with $\psi(z_i, e_i) = g(z_i) + f(e_i)$, $\alpha_i = 10, g(z_i) = \beta z_i = 10 z_i$, $f(e_i) = \gamma e_i$, and ϵ_i is generated from $\mathcal{N}(0,1)$ under a unit-level Bernoulli(0.5) randomization design. We focus on varying the ratio γ/β as we consider a fixed graph. Figure 4.A.2 shows that we consistently perform better than existing fixed estimators.

The simulations were run on a CPU. Our experiments focus on a subset of (the first) 1000 nodes of the 19828-node DVD graph. In particular, we considered the 17924-node subgraph by removing all isolated nodes. To compute the exposure probabilities, we used 10^6 simulations. We then selected the first 1000 nodes of the 19828 to analyze. 200 replicates were run, generating random treatment assignments and the corresponding outcomes. For each of the replicates, 1000 separate simulation runs were generated to compute the oracle MSEs. In total, this took approximately 2 hours. The graph data is available at: https://snap.stanford.edu/data/amazon-meta.html

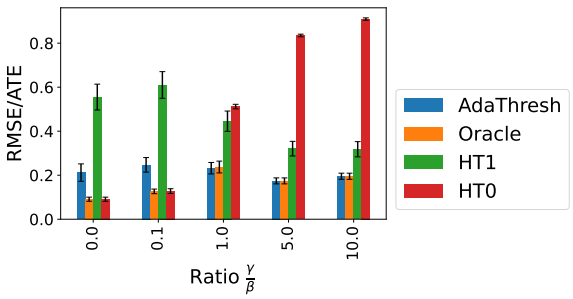


Figure 4.A.2: RMSE (normalized by the ATE) across different thresholds on the Amazon product network data. The error bars are two times the standard deviation.

4.A.7.2 Experimental details on simulated graphs

All synthetic graph simulations were run on a machine of Intel® Xeon® processors with 48 CPU cores, and 50GB of RAM. We simulated 1000 replicates, generating random treatment assignments and the corresponding outcomes, with each oracle MSE computed using 1000 separate simulation runs. The exposure probabilities under each threshold were computed as proposed in [Aronow and Samii, 2017] using 20000 simulation iterations. In total, this took approximately 30 minutes on average (across the different potential outcomes, and graph settings). The code is available at: https://github.com/Vydhourie/AdaThresh.git

In the following, we focus on key experimental results on circulant graphs (see Section 4.A.6 for discussions on circulant graphs).

4.A.7.3 More Simulations for Horvitz-Thompson estimator with adaptive exposure thresholds

In Figure 4.A.3, we simulate outcomes from the linear model with $\psi(z_i, e_i) = g(z_i) + f(e_i)$, $\alpha_i = 10$, $g(z_i) = \beta z_i = 10 z_i$, $f(e_i) = \gamma e_i$, and fixed ϵ_i generated from $\mathcal{N}(0, 1)$ for a 1000-node 2nd-power cycle graph.

4.A.7.4 Simulations for non-linear potential outcomes models

We investigate the robustness of our estimator to potential outcomes models that are non-linear in the exposure. In particular, we consider simulations from the sigmoid, and sine exposure functions.

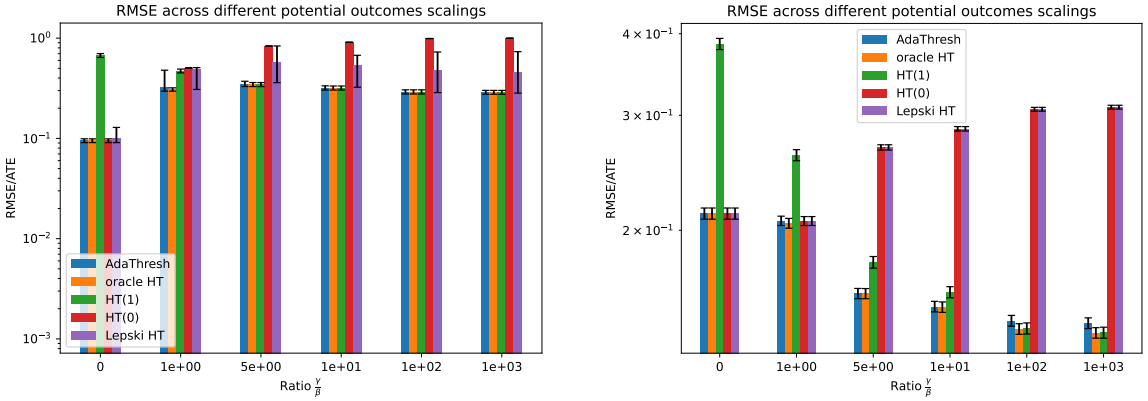


Figure 4.A.3: RMSE (normalized by the ATE) of different Horvitz-Thompson estimators. Left: 2nd-power cycle graph under unit-level Ber(0.5) randomization. Right: 2nd-power cycle graph under cluster-level Ber(0.5) randomization with cluster sizes 5 (=$2k+1$). The error bars are two times the standard deviation.

In Figure 4.A.4, we display the performance of our estimator under a sigmoid (left) and sine (right) interference function, respectively. Our adaptive threshold Horvitz-Thompson estimator (AdaThresh) generally improves upon other existing Horvitz-Thompson estimators.

4.A.7.5 Simulations for Difference-in-Means estimator with adaptive exposure thresholds

We consider our approach using the Difference-in-Means estimator incorporating exposure thresholds:

$$\hat{\tau}_{\mathrm{DiM}_h} = \frac{\sum_{i=1}^n \mathbf{1}\{z_i = 1, e_i \geq h\} Y_i}{\sum_{i=1}^n \mathbf{1}\{z_i = 1, e_i \geq h\}} - \frac{\sum_{i=1}^n \mathbf{1}\{z_i = 0, e_i \leq 1-h\} Y_i}{\sum_{i=1}^n \mathbf{1}\{z_i = 0, e_i \leq 1-h\}}. \tag{4.A.8}$$

We use the following bias estimator:

$$\hat{b}(\hat{\tau}_h) = \sum_{i=1}^n \frac{(1-e_i)\hat{\gamma}_n \mathbf{1}\{z_i = 1, e_i \geq h\}}{\sum_{i=1}^n \mathbf{1}\{z_i = 1, e_i \geq h\}} + \sum_{i=1}^n \frac{e_i \hat{\gamma}_n \mathbf{1}\{z_i = 0, e_i \leq 1-h\}}{\sum_{i=1}^n \mathbf{1}\{z_i = 0, e_i \leq 1-h\}},$$

where $\hat{\gamma}$ is the linear regression coefficient for the exposure variable.

We use the following variance estimator, decomposing it into its treatment and control parts:

$$\hat{v}(\hat{\tau}) = \frac{2}{n-1}\left(\hat{v}_T + \hat{v}_C\right)$$

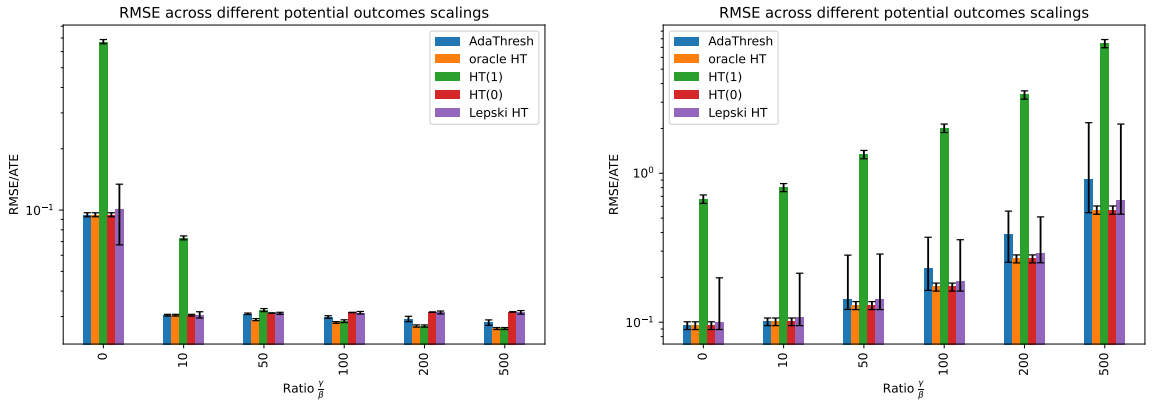


Figure 4.A.4: RMSE (normalized by the ATE) induced by the different Horvitz-Thompson estimators. We take $\psi(z_i, e_i) = g(z_i) + f(e_i)$. Left: 2nd-power cycle graph under unit-level Ber(p), $p = 0.5$, randomization under sigmoid $f(e_i) = \gamma/(1 + \exp(-10(e_i - p)))$. Right: 2nd-power cycle graph under unit-level Ber(p), $p = 0.5$, randomization with $f(e_i) = \gamma(1 - \sin(\pi \cdot e_i)))$ being a sine function. We focus on varying the ratio γ/β as we consider a fixed graph, with $n = 1000$. The error bars are two times the standard deviation.

where

$$\hat{v}_T = \frac{1}{n_1}\sum_{i=1}^{n}\left(Y_i\mathbf{1}\{z_i = 1, e_i \geq h\} - \frac{1}{n_1}\sum_{i=1}^{n} Y_i\mathbf{1}\{z_i = 1, e_i \geq h\}\right)^2$$

where $n_1 := \sum_{i=1}^{n}\mathbf{1}\{z_i = 1, e_i \geq h\}$, and

$$\hat{v}_C = \frac{1}{n_0}\sum_{i=1}^{n}\left(Y_i\mathbf{1}\{z_i = 0, e_i \leq 1 - h\} - \frac{1}{n_0}\sum_{i=1}^{n} Y_i\mathbf{1}\{z_i = 0, e_i \leq 1 - h\}\right)^2$$

where $n_0 := \sum_{i=1}^{n}\mathbf{1}\{z_i = 0, e_i \leq 1 - h\}$.

We compare the performance of our adaptive estimator to the vanilla difference-in-means estimator, the difference-in-means analogue of the vanilla Horvitz-Thompson estimator, and the difference-in-means estimator with a threshold plugin via Lepski's method. We write these out below:

- vanilla difference-in-means estimator

$$\hat{\tau}_{\mathrm{DiM}_0} = \frac{\sum_{i=1}^{n}\mathbf{1}\{z_i = 1\}Y_i}{\sum_{i=1}^{n}\mathbf{1}\{z_i = 1\}} - \frac{\sum_{i=1}^{n}\mathbf{1}\{z_i = 0\}Y_i}{\sum_{i=1}^{n}\mathbf{1}\{z_i = 0\}} \tag{4.A.9}$$

- difference-in-means estimator at threshold $h = 1$

$$\hat{\tau}_{\text{DiM}_1} = \frac{\sum_{i=1}^{n} \mathbf{1}\{z_i = 1, e_i = 1\} Y_i}{\sum_{i=1}^{n} \mathbf{1}\{z_i = 1, e_i = 1\}} - \frac{\sum_{i=1}^{n} \mathbf{1}\{z_i = 0, e_i = 0\} Y_i}{\sum_{i=1}^{n} \mathbf{1}\{z_i = 0, e_i = 0\}} \tag{4.A.10}$$

- difference-in-means estimator at threshold $\hat{h}_{\text{Lepski}}$ where

$$\hat{h}_{\text{Lepski}} := \min\{h \in H : \cap_{h' \in H : h' \geq h} I(h') \neq \emptyset\}, \tag{4.A.11}$$

with

$$I(h) := [\hat{\tau}_{\text{DIM}_h} - 2\widehat{\text{SDEV}}(\hat{\tau}_{\text{DIM}_h}), \hat{\tau}_{\text{DIM}_h} + 2\widehat{\text{SDEV}}(\hat{\tau}_{\text{DIM}_h})], \tag{4.A.12}$$

and

$$\hat{\tau}_{\text{LepskiDiM}} = \sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 1, e_i \geq \hat{h}_{\text{Lepski}}\}}{\sum_{i=1}^{n} \mathbf{1}\{z_i = 1, e_i \geq \hat{h}_{\text{Lepski}}\}} Y_i - \sum_{i=1}^{n} \frac{\mathbf{1}\{z_i = 0, e_i \leq 1 - \hat{h}_{\text{Lepski}}\}}{\sum_{i=1}^{n} \mathbf{1}\{z_i = 0, e_i \leq 1 - \hat{h}_{\text{Lepski}}\}} Y_i \tag{4.A.13}$$

In Figures 4.A.5 and 4.A.6, we display the performance of our adaptive threshold (AdaThresh) Difference-in-Means estimators in comparison to other estimators. AdaThresh improves upon fixed threshold Difference-in-Means estimators.

4.A.7.6 Simulations using local linear regression

We extend our global linear regression approach to a local linear regression one to estimate the rate of change of the bias. We write the new bias estimators for the Horvitz-Thompson and Difference-in-Means estimators. We illustrate the performance of the local linear regression in the settings above, as well as settings that significantly deviates from linearity. We note that local and global linear regression involve distinct bias-variance trade-offs, which we do not pursue here as they fall outside the scope of the paper.

We write the bias estimator for the Horvitz-Thompson estimator as:

$$\hat{b}(\hat{\tau}_h) = \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - e_i)\hat{\gamma}_n^{(h)} \mathbf{1}\{z_i = 1, e_i \geq h\}}{\mathbb{P}\{z_i = 1, e_i \geq h\}} + \frac{1}{n} \sum_{i=1}^{n} \frac{e_i \hat{\gamma}_n^{(1-h)} \mathbf{1}\{z_i = 0, e_i \leq 1 - h\}}{\mathbb{P}\{z_i = 0, e_i \leq 1 - h\}},$$

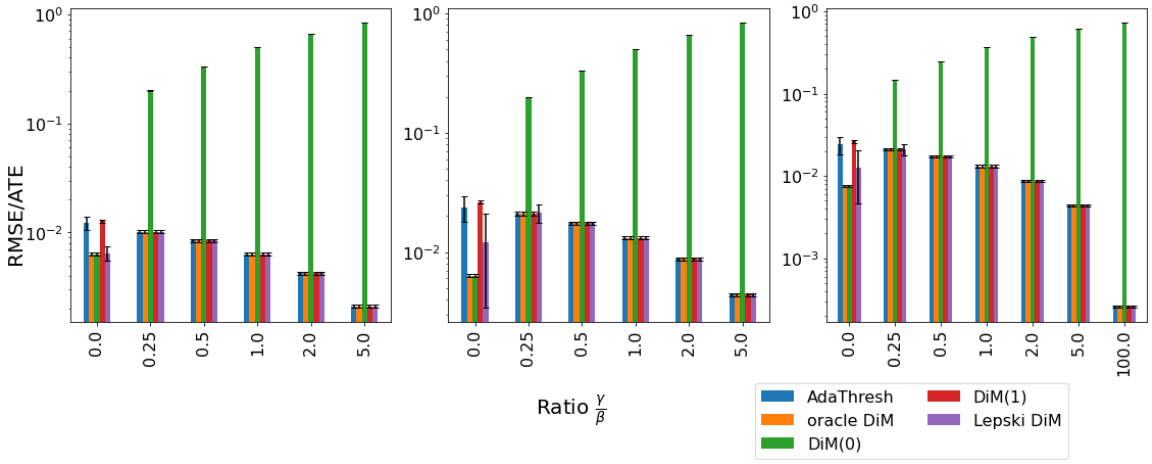


Figure 4.A.5: RMSE (normalized by the ATE) under the linear model with $\psi(z_i, e_i) = g(z_i) + f(e_i)$, $\alpha_i = 10$, $g(z_i) = \beta z_i = 10 z_i$, $f(e_i) = \gamma e_i$, and fixed ϵ_i generated from $\mathcal{N}(0, 1)$, induced by the different Difference-in-Means estimators. We focus on varying the ratio γ/β as we consider a fixed graph. Left: Cycle graph under unit-level Ber(0.5) randomization. Center: 2nd-power cycle graph under unit-level Ber(0.5) randomization. Right: 2nd-power cycle graph under cluster-level Ber(0.5) randomization with cluster sizes 5 (=2k + 1). The error bars are two times the standard deviation.

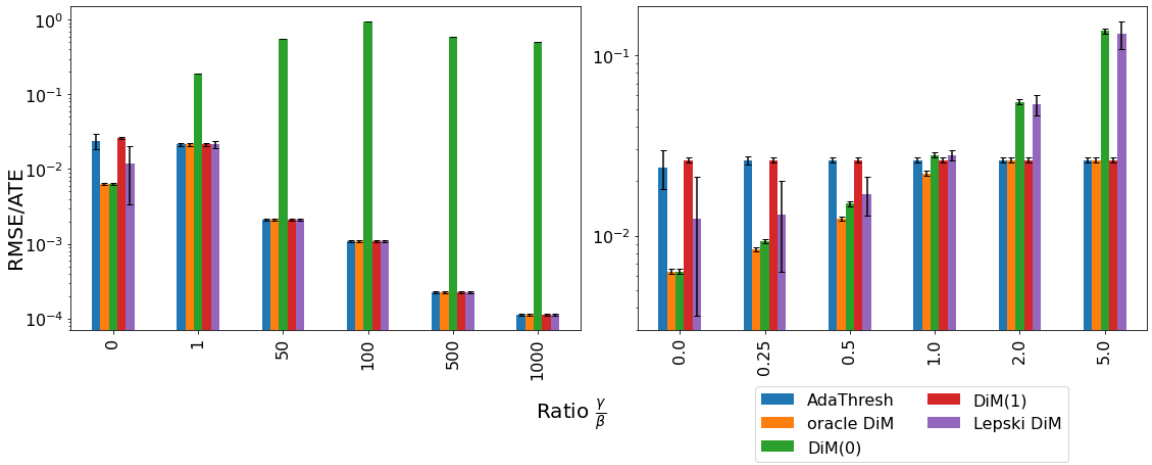


Figure 4.A.6: RMSE (normalized by the ATE) induced by the different Difference-in-Means estimators. We take $\psi(z_i, e_i) = g(z_i) + f(e_i)$. Left: 2nd-power cycle graph under unit-level Ber(p), $p = 0.5$, randomization under sigmoid $f(e_i) = \gamma/(1 + \exp(-e_i))$. Right: 2nd-power cycle graph under unit-level Ber(p), $p = 0.5$, randomization with $f(e_i) = \gamma(1 - \sin(\pi \cdot e_i)))$ being a sine function. We focus on varying the ratio γ/β as we consider a fixed graph, with $n = 1000$. The error bars are two times the standard deviation.

where $\hat{\gamma}_n^{(h)}$ and $\hat{\gamma}_n^{(1-h)}$ are the local linear regression coefficients for the exposure variable in the intervals $[h, 1]$, and $[0, 1-h]$, respectively.

Similarly, write the bias estimator for the Difference-in-Means estimator as:

$$\hat{b}(\hat{\tau}_h) = \sum_{i=1}^{n} \frac{(1-e_i)\hat{\gamma}_n^{(h)} \mathbf{1}\{z_i = 1, e_i \geq h\}}{\sum_{i=1}^{n} \mathbf{1}\{z_i = 1, e_i \geq h\}} + \sum_{i=1}^{n} \frac{e_i \hat{\gamma}_n^{(1-h)} \mathbf{1}\{z_i = 0, e_i \leq 1-h\}}{\sum_{i=1}^{n} \mathbf{1}\{z_i = 0, e_i \leq 1-h\}},$$

where $\hat{\gamma}_n^{(h)}$ and $\hat{\gamma}_n^{(1-h)}$ are the local linear regression coefficients for the exposure variable in the intervals $[h, 1]$, and $[0, 1-h]$, respectively.

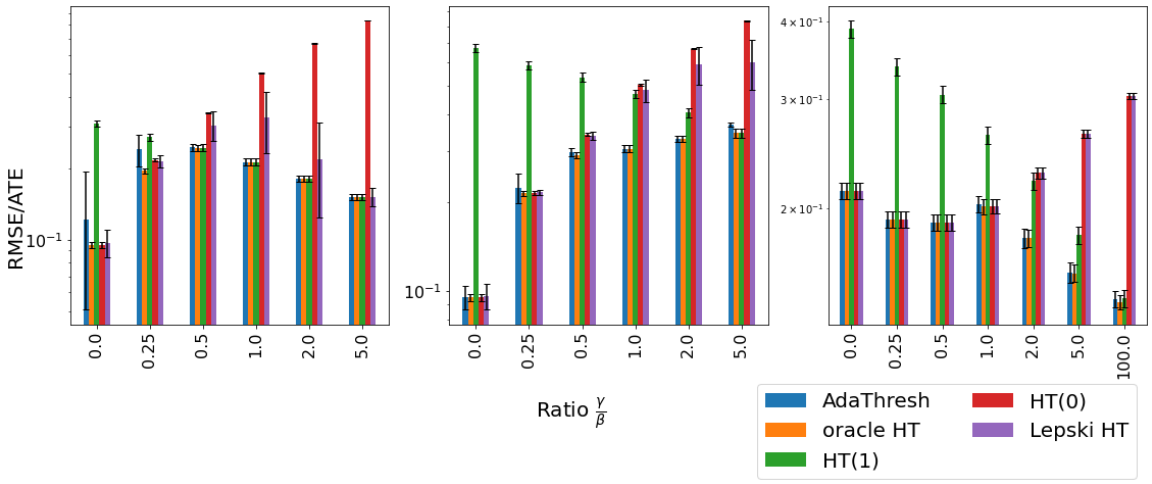


Figure 4.A.7: RMSE (normalized by the ATE) under the linear model with $\psi(z_i, e_i) = g(z_i) + f(e_i)$, $\alpha_i = 10$, $g(z_i) = \beta z_i = 10 z_i$, $f(e_i) = \gamma e_i$, and fixed ϵ_i generated from $\mathcal{N}(0, 1)$, induced by the different (local) Horvitz-Thompson estimators. We focus on varying the ratio γ/β as we consider a fixed graph, with $n = 1000$. Left: Cycle graph under unit-level Ber(0.5) randomization. Center: 2nd-power cycle graph under unit-level Ber(0.5) randomization. Right: 2nd-power cycle graph under cluster-level Ber(0.5) randomization with cluster sizes 5 (=2k + 1). The error bars are two times the standard deviation.

Figures 4.A.7, 4.A.9, 4.A.8, and 4.A.10, display how local linear regression bias estimates perform.

4.A.8 Discussion on relaxing the bounded degree assumption

In more general settings, we can relax Assumption 4.4.3 by partitioning the exposures into exposure bins E_b, $b = 1, 2, \ldots, K$, each associated with an "effective exposure" $\tilde{e}_b$. That is, $[0, 1] = E_1 \cup E_2 \cup \ldots \cup E_K$, with $E_i \cap E_j = \emptyset$, for all $i \neq j \in [K]$. In particular, in the weighted graph setting, we can

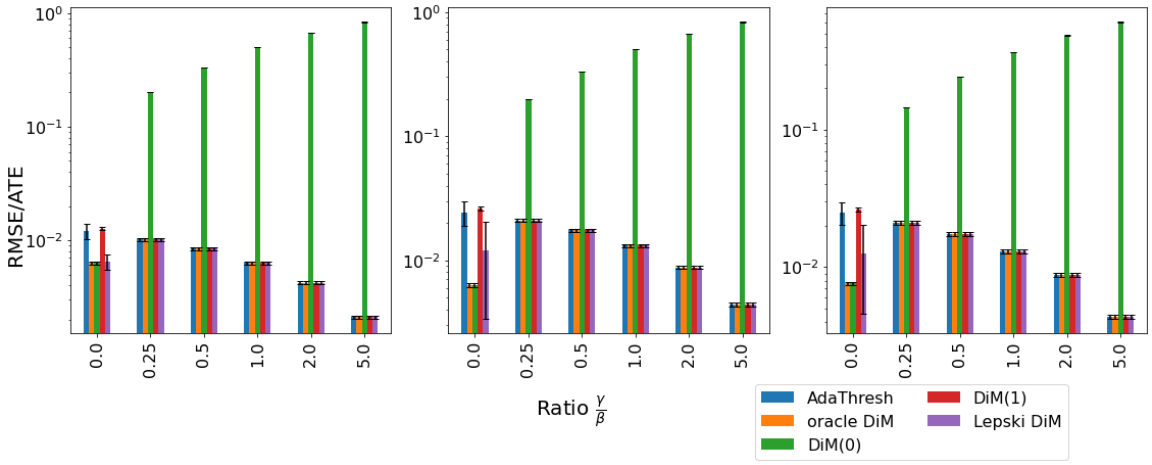


Figure 4.A.8: RMSE (normalized by the ATE) under the linear model with $\psi(z_i, e_i) = g(z_i) + f(e_i)$, $\alpha_i = 10$, $g(z_i) = \beta z_i = 10 z_i$, $f(e_i) = \gamma e_i$, and fixed ϵ_i generated from $\mathcal{N}(0, 1)$, induced by the different (local) Difference-in-Means estimators. We focus on varying the ratio γ/β as we consider a fixed graph, with $n = 1000$. Left: Cycle graph under unit-level Ber(0.5) randomization. Center: 2nd-power cycle graph under unit-level Ber(0.5) randomization. Right: 2nd-power cycle graph under cluster-level Ber(0.5) randomization with cluster sizes 5 (=2k + 1). The error bars are two times the standard deviation.

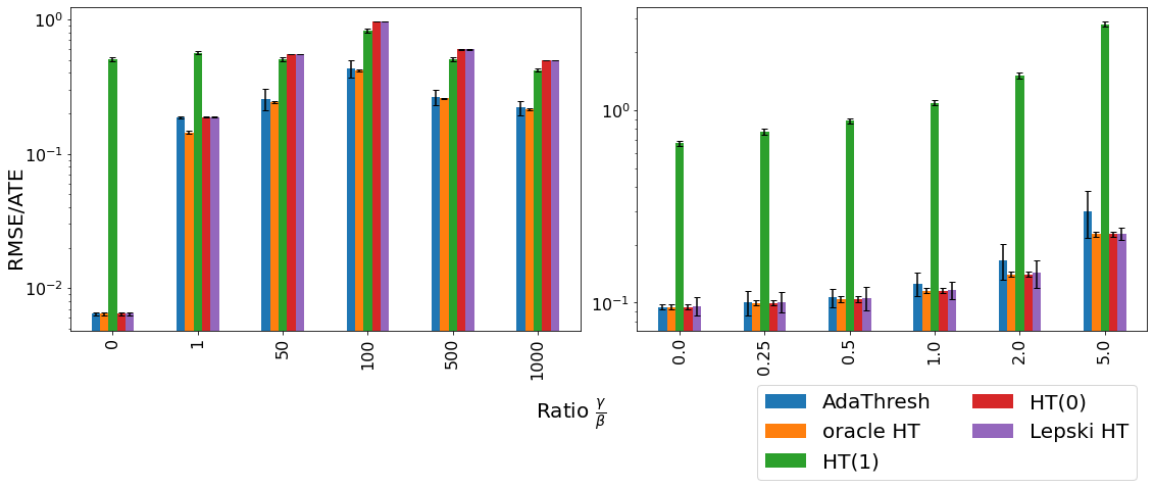


Figure 4.A.9: RMSE (normalized by the ATE) under the linear model with $\psi(z_i, e_i) = g(z_i) + f(e_i)$, $\alpha_i = 10$, $g(z_i) = \beta z_i = 10 z_i$, and fixed ϵ_i generated from $\mathcal{N}(0, 1)$, induced by the different (local) Horvitz-Thompson estimators. Left: 2nd-power cycle graph under unit-level Ber(0.5) randomization under sigmoid $f(e_i) = \gamma/(1 + \exp(-e_i)))$. Right: 2nd-power cycle graph under unit-level Ber(0.5) randomization with $f(e_i) = \gamma(1 - \sin(\pi \cdot e_i)))$ being a sine function. We focus on varying the ratio γ/β as we consider a fixed graph, with $n = 1000$. The error bars are two times the standard deviation.

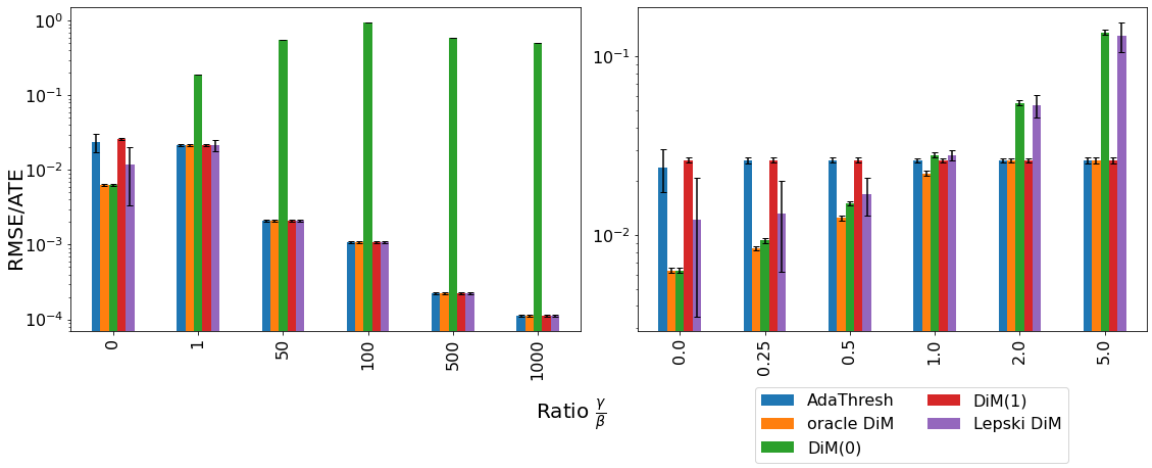


Figure 4.A.10: RMSE (normalized by the ATE) induced by the different (local) Difference-in-Means estimators. We take $\psi(z_i, e_i) = g(z_i) + f(e_i)$. Left: 2nd-power cycle graph under unit-level Ber(0.5) randomization under sigmoid $f(e_i) = \gamma/(1 + \exp(-e_i)))$. Right: 2nd-power cycle graph under unit-level Ber(0.5) randomization with $f(e_i) = \gamma(1 - \sin(\pi \cdot e_i))$ being a sine function. We focus on varying the ratio γ/β as we consider a fixed graph, with $n = 1000$. The error bars are two times the standard deviation.

take the effective exposure of unit i, $\tilde{e}_i = \tilde{e}_b$ if and only if $i \in E_b$, and such that exposure positivity, for all $i \in [n]$, $r \in \{0, 1\}$, and $s \in [0, 1]$, $\mathbb{P}(z_i = r, \tilde{e}_i = s) > 0$, is satisfied. For instance, for uniformly sized bins with associated effective exposures $\tilde{e}_b$ that well approximate e_i, $|E_b| = 1/K$. Let $B_r(i)$ denote the ball of radius r around unit i, such that it decomposes, $|B_1(i)| = |B_1^S(i)| + |B_1^W(i)|$ for all $i \in [n]$. $B_1^S(i)$ is the ball of radius 1 around unit i with strong connections η_i^s, and $|B_1^W(i)|$ be the ball of radius 1 around unit i with weak connections δ_i^w such that $\sup_i \frac{\eta_i^s}{\delta_i^w} \to \infty$. We choose our partition $B_1, \dots B_K$, for some K, as a function of δ_i^w, and $B_1^W(i)$, $i \in [n]$, such that the exposure bins well approximate the true exposures. Additionally, for all $i \in [n]$, we require that $|B_1^S(i)| = \mathcal{O}(1)$. With an abuse of notation, write $e_i(\mathcal{N})$ to mean the exposure of unit i as a function of the treated units in the subset $\mathcal{N} \subseteq B_1(i)$. Then, we require also that $e_i(B_1^S(i)) = \Omega(e_i(B_1(i)))$. Additionally, $B_1^S(i)$ must satisfy the three conditions around "affinity sets" in [Chandrasekhar et al., 2023] generalizing [Aronow and Samii, 2017], with an additional condition that the sum of the affinity set covariances is $\Omega(n)$, in the bias terms as well as the variance terms in the MSE. This gives us uniform control on the estimated MSE of the estimator with respect to the true MSE of the estimator and E_b, $b = 1, \dots, K$, and all the results follow through. These conditions subsume the approximate neighborhood interference

(ANI) condition in [Leung, 2022a] analogously to α-mixing conditions in spatial settings, such as in [Jenish and Prucha, 2009], when the maximum clique size in the network does not grow. Indeed, when the maximum clique size in the network does not grow, we can embed the ψ-dependent network in [Leung, 2022a] in a lattice structure of a random field in a fixed-dimensional Euclidean space with α-mixing just as in [Jenish and Prucha, 2009].

4.A.9 Discussion on "double-dipping"

Since our approach involves "double-dipping" into the data, we need to make sure that there is no overfitting that occurs. The authors of [Chernozhukov et al., 2018], [Belloni et al., 2015], and [Kennedy, 2022] describe this in more detail. While sample-splitting would simply take care of this in the case of independent data, it does not apply to our setting where the data are dependent, as modeled by the network. One could potentially leverage results under the dependent-data setting, such as Hart and Vieu [1990], but this is outside the scope of our paper. We propose to use our approach on the whole sample data and argue this via empirical process theory arguments. In particular, we can think of our threshold h as a nuisance parameter, and we show that our estimator for h has a simple enough associated MSE function class. We note that it is sufficient to show that we are not overfitting by showing that the rate of convergence of the estimated MSE under $\hat{h}$ converges to the true MSE under the optimal threshold, under the best possible linear fit, h^* at least at a $\mathcal{O}(n^{-1/2})$-rate.

Proposition 4.A.5. *Suppose that Assumptions 2, 4.4.2, and 4.4.3 are satisfied. The corresponding bias terms in the estimated and true MSE, $\hat{M}_n^b$, and M_n^b, under "double prediction", satisfy*

$$\mathbb{E}\left[\sup_{\delta/2<|h_n^*-\hat{h}_n|<\delta}\sqrt{n}\left(\hat{M}_n^b(\hat{h}_n)-M_n^b(\hat{h}_n)-\hat{M}_n^b(h^*)-M_n^b(h^*)\right)\right]\to 0,$$

as $\delta\to 0$.

Suppose that Assumptions 2, 4.4.2, and 4.4.3 are satisfied. We aim to show that the corresponding biases in the MSE terms under " double prediction" satisfy

$$\mathbb{E}\left[\sup_{\delta/2<|h^*-h_n|<\delta}\sqrt{n}\left(\hat{M}_n^b(h_n)-M_n^b(h_n)-\hat{M}_n^b(h^*)-M_n^b(h^*)\right)\right]\to 0,$$

as $\delta\to 0$, where x_i ranges over the set X_i of possible fractions of degree i.

Proof to Proposition 4.A.5. We begin by considering the empirical process in question. We write out

$$\begin{aligned}
\hat{M}_n(h) = &\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_h(1-e_i)\mathbf{1}\{z_i=1,e_i\geq h\}}{\mathbb{P}\{z_i=1,e_i\geq h\}}+\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_h(1-e_i)\mathbf{1}\{z_i=1,e_i\geq h\}}{\mathbb{P}\{z_i=1,e_i\geq h\}}\right)^2\\
&+\sum_{i=1}^{n}\frac{\mathbf{1}\{z_i=1,e_i\geq h\}Y_i^2}{n^2\pi_i^1}(\frac{1}{\pi_i^1}-1)\\
&+\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\frac{\mathbf{1}\{z_i=1,e_i\geq h\}\mathbf{1}\{z_j=1,e_j\geq h\}Y_iY_j}{n^2\pi_{ij}^{11}}(\frac{\pi_{ij}^{11}}{\pi_i^1\pi_j^1}-1)\\
&+\sum_{i=1}^{n}\frac{\mathbf{1}\{z_i=0,e_i\leq 1-h\}Y_i^2}{n^2\pi_i^0}(\frac{1}{\pi_i^0}-1)\\
&+\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\frac{\mathbf{1}\{z_i=0,e_i\leq 1-h\}\mathbf{1}\{z_j=0,e_j\leq 1-h\}Y_iY_j}{n^2\pi_{ij}^{00}}(\frac{\pi_{ij}^{00}}{\pi_i^0\pi_j^0}-1)\\
&-\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\in[n];\pi_{ij}^{10}>0}(\mathbf{1}\{z_i=1,e_i\geq h\}\mathbf{1}\{z_j=0,e_j\leq 1-h\}Y_iY_j)(\frac{1}{\pi_i^1\pi_j^0}-\frac{1}{\pi_{ij}^{10}})\\
&+\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\in[n];\pi_{ij}^{10}=0}(\frac{\mathbf{1}\{z_i=1,e_i\geq h\}Y_i^2}{2\pi_i^1}+\frac{\mathbf{1}\{z_j=0,e_j\leq 1-h\}Y_j^2}{2\pi_j^0}).
\end{aligned}$$

The absolute Horvitz-Thompson bias estimate at threshold h (and at $1-h$) is

$$\hat{b}(\hat{\tau}_h)=\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_n(1-e_i)\mathbf{1}\{z_i=1,e_i\geq h\}}{\mathbb{P}(z_i=1,e_i\geq h)}+\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_n(e_i)\mathbf{1}\{z_i=0,e_i\leq 1-h\}}{\mathbb{P}(z_i=0,e_i\leq 1-h)} \tag{4.A.14}$$

Then, for every h, using the identity $x^2-y^2=(x+y)(x-y)$, and defining $U_n\in\mathbb{R}$ such that $U_n\geq M_n^b(h)$ for all h (note that there exists such a $U_n<\infty$ by Assumptions 2, 4.4.2), we have

$$\hat{M}_n^b(\hat{\tau}_h)-M_n^b(\hat{\tau}_h)\leq 2U_n\left(\hat{b}(\hat{\tau}_h)-b^*(\hat{\tau}_h)\right)+\left(\hat{b}(\hat{\tau}_h)-b^*(\hat{\tau}_h)\right)^2,$$

where, from Section 4.A.4, the difference between the bias estimate and the true bias induced by the best average linear fit is

$$\begin{aligned}
\hat{b}(\hat{\tau}_h)-b^*(\hat{\tau}_h)=&\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_n(1-e_i)\mathbf{1}\{z_i=1,e_i\geq h\}}{\mathbb{P}(z_i=1,e_i\geq h)}-\frac{1}{n}\sum_{i=1}^{n}\sum_{x_i\in X_i}\frac{\mathbf{1}\{x_i\geq h\}}{|x_i:x_i\in X_i\cap x_i\geq h|}\gamma_n(1-x_i)\right)\\
&+\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_n e_i\mathbf{1}\{z_i=0,e_i\leq 1-h\}}{\mathbb{P}(z_i=0,e_i\leq 1-h)}-\frac{1}{n}\sum_{i=1}^{n}\sum_{x_i\in X_i}\frac{\mathbf{1}\{x_i\leq 1-h\}}{|x_i:x_i\in X_i\cap x_i\leq 1-h|}\gamma_n(x_i)\right),
\end{aligned}$$

where γ_n is the slope of the best average linear fit, and where x_i ranges over the set X_i of possible fractions of degree i.

We now proceed to consider each of the terms in $\hat{b}(\hat{\tau}_h) - b^*(\hat{\tau}_h)$ above. Each of the parentheses satisfies Donsker conditions. Indeed, under Assumptions 2, 4.4.2, and 4.4.3, they are sample-mean terms with uniformly bounded coefficients, and $\hat{\gamma}_n$ converges to γ_n^* at a rate of $\mathcal{O}(n^{-1/2})$, which we compose with $3b^*(h)$ (see Section 3.4.3.2. in [Van Der Vaart et al., 1996]). Thus, the terms in $\hat{M}_n^b(h) - M_n^b(h)$ have bounded entropy integral. Under our bounded-variation potential-outcome model (Assumption 4.4.2), we have that $\log N_{[]}(\epsilon, \Theta_n^b, L_2(\mathbb{P}) \leq K\left(\frac{1}{\epsilon}\right)$, which gives us $\tilde{J}(\delta, \Theta_n^b, L_2(\mathbb{P})) \leq \delta^{1/2}$ [Van Der Vaart et al., 1996] (see also Example 19.11 in [Van der Vaart, 2000]).

Therefore, putting these together, we have the following maximal inequality for the bias terms

$$\mathbb{E}\left[\sup_{\delta/2<|h^*-\hat{h}_n|<\delta} \sqrt{n}\left(\hat{M}_n^b(h_n) - M_n^b(h_n) - (\hat{M}_n^b(h^*) - M_n^b(h^*))\right)\right] \tag{4.A.15}$$

$$\leq \tilde{J}(\delta, \Theta_n^b, L_2(\mathbb{P}))\left(1 + \frac{\tilde{J}(\delta, \Theta_n^b, L_2(\mathbb{P}))}{\delta\sqrt{n}}\right), \tag{4.A.16}$$

with $\tilde{J}(\delta, \Theta_n^b, L_2(\mathbb{P})) \leq \delta^{1/2}$.

■

Chapter 5

OPTIMAL DESIGN UNDER INTERFERENCE, HOMOPHILY, AND ROBUSTNESS TRADE-OFFS

This chapter is based on the paper "Optimal Design under Interference, Homophily, and Robustness Trade-offs," coauthored with Alex Kokot[⋆], Jennifer Brennan[‡], Marina Meilă[◇], Christina Lee Yu[¶] and Maryam Fazel[⋆]. It is currently under review.

[⋆] University of Washington, USA

[‡] Google LLC, USA

[◇] University of Waterloo, Canada

[¶] Cornell University, USA

To minimize the mean squared error (MSE) in global average treatment effect (GATE) estimation under network interference, a popular approach is to use a cluster-randomized design. However, in the presence of homophily, which is common in social networks, cluster randomization can instead increase the MSE. We develop a novel potential outcomes model that accounts for interference, homophily, and heterogeneous variation. In this setting, we establish a framework for optimizing designs for worst-case MSE under the Horvitz-Thompson estimator. This leads to an optimization problem over the covariance matrices of the treatment assignment, trading off interference, homophily, and robustness. We frame and solve this problem using two complementary approaches. The first involves formulating a semidefinite program (SDP) and employing Gaussian rounding, in the spirit of the Goemans-Williamson approximation algorithm for MAXCUT. The second is an adaptation of the Gram-Schmidt Walk, a vector-balancing algorithm which has recently received much attention. Finally, we evaluate the performance of our designs through various experiments on simulated network data and a real village network dataset.

Keywords: network interference, causal inference, optimization

5.1 Introduction

Network interference models settings in which the outcome of one unit depends on the treatment assignment of its neighbors in a network. The presence of network interference complicates the estimation of global average treatment effects (GATE), the difference between the average outcome when all units are treated and the average outcome when all units are assigned to control, as it violates the Stable Unit Treatment Value Assumption (SUTVA). In particular, in an experiment with partial treatment assignments where there is a mixture of treated units and control units, the difference in treatment assignments between neighboring units can bias each unit's outcome relative to the all-treated and all-control outcomes. Therefore, basic experimental designs such as unit-level Bernoulli randomization lead to biased GATE estimation under standard estimators.

Cluster randomization approaches [Ugander et al., 2013, Ugander and Yin, 2023, Brennan et al., 2022, Viviano et al., 2023] are considered to be a useful way to reduce this bias. All of these methods stem from the same principle. Since bias is introduced by the conflicting treatment assignments between neighboring units, if we were to cluster together units that are strongly connected before assigning treatment at the cluster level, we would successfully reduce the bias. Consider a social network in which the outcome of interest is each individual's likelihood of joining a premium social media platform. If treatment consists of offering a free trial, an individual's outcome may depend not only on their own assignment, but also on whether their friends receive free trials and subsequently join the platform. Consequently, the outcomes from when all people in the social network are treated and from when all people are assigned to control will be different from when there is an experiment run where people are uniformly randomly assigned to treatment and control, due to the interference

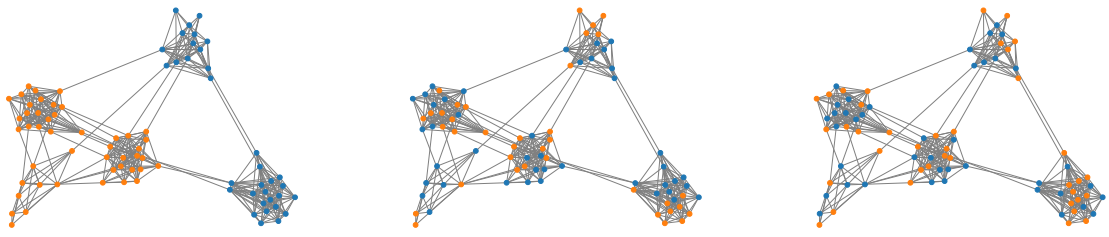


Figure 5.1.1: Optimized treatment assignments under interference-homophily tradeoffs on a synthetic network as homophily levels increase from left to right. See Appendix 5.B for more details.

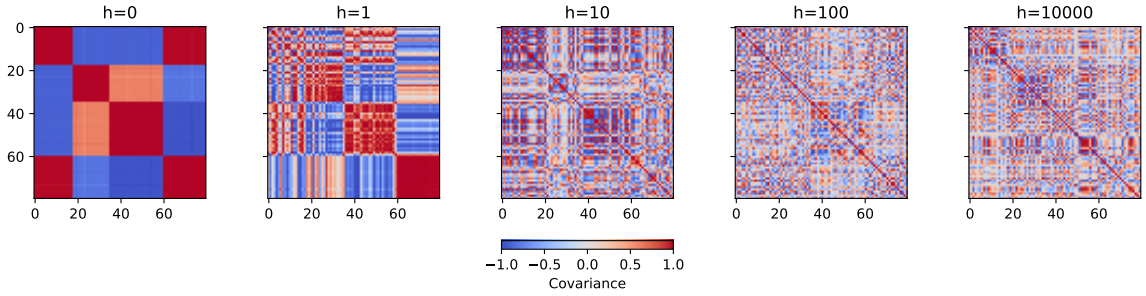


Figure 5.1.2: Optimal design covariance matrices under interference-homophily tradeoffs as homophily increases from left to right. See Appendix 5.B for more details.

between friends who have different treatment assignments. Assigning tightly connected friends to the same treatment group, whether treatment or control, reduces this discrepancy between the outcomes in the experimental setting and in the full treatment and control settings.

However, networks are often homophilous: individuals tend to form connections with others who are similar to themselves, a phenomenon first noted by Lazarsfeld et al. [1954] and commonly summarized as "birds of a feather flock together" [McPherson et al., 2001]. Similarities along socioeconomic, demographic, or behavioral dimensions [Jackson et al., 2023] increase the likelihood of social connection, and these same dimensions may also drive heterogeneous treatment responses. For example, younger and older individuals may respond differently to incentives for adopting new technologies. When cluster-randomized designs align with such homophilous groupings, covariate distributions can vary substantially across clusters, leading to unrepresentative treatment or control samples [Ugander and Yin, 2023, Cortez-Rodriguez et al., 2024, Fatemi and Zheleva, 2020]. Some covariate groups may be rarely or never exposed to treatment, inflating the imprecision of the estimator. In the presence of homophily, one would perhaps, now contradictorily, be incentivized to spread treatments across the network, to more effectively sample across different social groups. This reflects the typical goal of covariate balancing, where estimation accuracy is improved by enforcing similarity between treatment and control groups [Efron, 1971]. More broadly, causal effects estimation on network data must often account both for dependence transmitted along network ties and for dependence induced by latent similarities among connected nodes [Ogburn et al., 2024]. Finally, an effective design is one that is robust to heterogeneous variation, i.e., differences in potential outcomes poorly explained by the observed network structure. In Appendix 5.L.1, we simulate experiments

using network data from villages in Karnataka, India, collected by Banerjee et al. [2013a], which exhibit caste-based homophily. We observe that for large ranges of interference and heterogeneous variation levels, our designs, in the worst-case, perform better than existing unit-level and cluster-level randomization designs.

5.1.1 Our Contributions and Overview of Results

We formalize the trade-off between network interference and homophily, recovering well-studied designs in limiting regimes, and show how this trade-off interacts with randomization to yield a unifying three-way framework generalizing various results in the literature [Harshaw et al., 2024, Viviano et al., 2023]. We focus on the Horvitz-Thompson estimator which is commonly used in practice, and operate under the potential outcomes framework. We provide methods to solve for mean squared error (MSE)-optimal designs under the worst-case potential outcomes within the subclass $\chi(\eta, \gamma, \kappa)$ (made precise in Section 5.4.1), where homophily, interference, and heterogeneous variation conditions are separately satisfied with parameters η, γ, and κ, respectively.

We consider the GATE, $\tau = \frac{1}{n}\langle \mathbf{1}, Y(\mathbf{1}) - Y(\mathbf{0})\rangle = \frac{1}{n}\langle \mathbf{1}, \phi\rangle$, where $Y(z)$ is the outcome vector observed under the treatment assignment vector $z \in \{0,1\}^n$, and ϕ is the vector of global treatment effects. We focus on the vanilla Horvitz-Thompson estimator $\hat{\tau} = \left\langle Y, \frac{z}{pn} - \frac{\mathbf{1}-z}{(1-p)n} \right\rangle$ with uniform treatment probability p across nodes, $p = \mathbb{P}(z_i = 1) = \mathbb{P}(z_j = 1)$ for all $i, j \in [n]$. Let L be the graph Laplacian[1] underlying the interference structure, and $L^\dagger$ its pseudoinverse. We can then characterize the homophily, interference, and robustness trade-offs with parameters η, γ, and κ, respectively.

Theorem 5.1.1 (Informal version of Theorem 5.4.2)**.** *Let $x \in \{-1,1\}^n$ by defining $x := 2z - 1$. Define $X := \mathbb{E}[xx^T]$. Then, for $q \geq 1$,*

$$\sup_{\chi(\eta,\gamma,\kappa)} \mathbb{E}[(\hat{\tau} - \tau)^2] \leq \frac{C}{n^2}\left\{\operatorname{Tr}(\mathbf{1}\mathbf{1}^T X) + C_1\eta \operatorname{Tr}(L^\dagger X) + C_2\kappa\|X\|_q + C_3\gamma \operatorname{Tr}(LX)\right\},$$

for known constants C, C_1, C_2, C_3 that depend on p.

We propose two complementary approaches. The first approach involves formulating a semidefinite program (SDP) and employing Gaussian rounding, in the spirit of the Goemans-Williamson approximation algorithm [Goemans and Williamson, 1995] for MAXCUT. In particular, we cast the

[1]Here $L := D - W$, where D is the diagonal degree matrix and W is the edge-weight matrix; see Sections 5.1.3 and 5.2 for details.

optimal design problem as an optimization problem over X, the covariance matrix of the design. Reparameterizing with $\tilde{\gamma}$,$\tilde{\eta}$, and $\tilde{\kappa}$, so that $\tilde{\gamma} := C_3\gamma, \tilde{\eta} := C_1\eta,$ and $\tilde{\kappa} := C_2\kappa$, we can write an SDP more explicitly in terms of the new trade-off parameters:

$$\begin{aligned} \min_{X \in \mathbb{R}^{n\times n}} \quad & \tilde{\eta}\mathrm{Tr}(L^{\dagger}X) + \tilde{\gamma}\mathrm{Tr}(LX) + \tilde{\kappa}\|X\|_q + \mathrm{Tr}(\mathbf{1}\mathbf{1}^T X) \\ \text{s.t.} \quad & X \succeq 0, \quad \mathrm{diag}(X) = \mathbf{1}. \end{aligned} \tag{5.1.1}$$

To generate treatment assignments, we proceed with a Gaussian rounding procedure. As a second approach, we adapt the Gram-Schmidt Walk [Harshaw et al., 2024], a vector balancing algorithm. We uncover a natural interpretation of interference and homophily by viewing eigenvalue-scaled Laplacian eigenvector embeddings as unit-level covariates, connecting the covariate balancing and interference literatures.

Our tradeoff characterization recovers several familiar limiting regimes. When γ is large, our design reduces to a min-cut type clustering. When η is large, it encourages dispersed treatment assignments. Finally, when κ is large, it reduces to Bernoulli randomization. More broadly, our framework is robust to heterogeneous variation, including that induced by noisy network observations. We show that increasing randomization toward unit-level Bernoulli designs improves robustness to such misspecifications. We demonstrate improvements over existing designs through various experiments on simulated network data and a real village network dataset.

5.1.2 Related Work

Harshaw et al. [2024] propose the Gram-Schmidt Walk design that navigates trade-offs between covariate balancing and robustness under a potential outcomes model. Our paper is of a similar flavor as we study designs that trade off homophily, interference, and robustness to heterogeneous variation. We adapt the Gram-Schmidt Walk algorithm to our setting as one approach, and contrast its performance with an alternative method of solving an SDP followed by Gaussian rounding. We also note a connection to Bhat et al. [2020], who study treatment-effect estimation, under a different potential outcomes model, in a non-network setting with the goal of maximizing precision. They derive an SDP relaxation for this maximization problem and directly apply the Goemans–Williamson rounding scheme [Goemans and Williamson, 1995]. They further show that their precision maximization admits a covariate balancing interpretation. Our framework is formulated

instead through a quadratic-form minimization perspective, and our characterization of homophily and covariate balancing reveals a natural conceptual bridge between these approaches, with connections to Algorithms 2 and 4.

Optimization-based designs for network interference settings are well studied. Viviano et al. [2023] propose an SDP to construct clusters for cluster-randomized designs by minimizing the MSE of a difference-in-means estimator, but do not consider the effects of homophily. The work most closely related to ours is Chen et al. [2024], who also minimize an MSE objective over the treatment assignment covariance matrix. Their focus, however, is restricted to cluster randomization and does not address homophily. They consider a non-convex formulation arising from Grothendieck's identity and propose a projected gradient method to find stationary points. In contrast, we adopt a convex SDP formulation and apply a Gaussian rounding procedure that generalizes the rounding scheme for the MAXCUT approximation of Goemans and Williamson [1995].

Bhat et al. [2020] study treatment-effect estimation and show that maximizing precision is related to a covariate balancing objective. They derive an SDP relaxation for precision maximization and apply the Goemans–Williamson rounding scheme, as in the MAXCUT approximation algorithm [Goemans and Williamson, 1995]. Our quadratic-form characterization of homophily/covariate balancing terms provides a bridge between this and vector balancing approaches such as Harshaw et al. [2024], with natural connections to our Algorithms 2 and 4.

A common design approach with the goal of balancing covariates within a randomization scheme is re-randomization [Morgan and Rubin, 2012, Li et al., 2018]. While we do not directly compare our methods to re-randomization procedures, Harshaw et al. [2024] show superior trade-offs of their method, which ours resemble under reparameterization, in comparison to re-randomization. Basse and Airoldi [2018] study the re-randomization approach under homophily settings alone. Ugander and Yin [2023] introduce randomized graph cluster randomization (RGCR) to address exponentially small exposure probabilities arising in graph cluster randomization designs [Ugander et al., 2013]. They show that RGCR still exhibits large variance when the nodes in a ring graph are clustered into a small number of large randomization units under a potential outcomes model with a homophily drift, a phenomenon also observed more generally in simulations. Related observations on the tension between homophily and cluster randomization appear in Ugander et al. [2012], Cortez-Rodriguez et al. [2024].

Another strategy is a matched-pair design, where randomization is restricted to pre-specified

pairs of similar units. Fatemi and Zheleva [2020] propose CMatch, a two-stage procedure that clusters units via random walks on the graph and then matches clusters using a similarity measure to assign treatments. Our framework does not rely on local assumptions about inherent similarities between pairs of randomization units. We introduce a global optimization approach to address homophily, interference, and robustness simultaneously.

Jagadeesan et al. [2020] study estimation of direct effects by constructing treatment assignments that balance interference-related features between treated and control units. They introduce perfect quasi-colorings, which eliminate interference bias by matching units with identical treated and control neighborhood structure, and analyze the bias and variance of restricted randomization schemes that approximate such colorings when they do not exist. Their design can be viewed as stratified randomization based on node similarity, and they characterize the resulting MSE under various settings, including models with homophily and network interference, showing improvements over Bernoulli and complete randomization in some regimes.

Optimization-based designs in this setting include the approach of Fatemi and Zheleva [2023], where the authors formulate a mixed-integer program to identify maximally independent node sets for direct effect estimation. In a different application, Doudchenko et al. [2021] use a mixed-integer program to jointly optimize design and weights in synthetic control settings. Solving mixed-integer programs to optimality is not polynomial-time and does not scale; solving an SDP, as in our method, is polynomial-time and admits efficient solvers that can be sped up by exploiting problem structure.

Bipartite networks arising from preference-based matching processes, such as those studied by Gale and Shapley [1962], naturally induce homophily. Our framework therefore extends to bipartite settings, including marketplaces, generalizing existing bias–variance and graph-cut characterizations [Brennan et al., 2022, Holtz et al., 2025, Pouget-Abadie et al., 2018, Harshaw et al., 2023].

5.1.3 Finite-population setup and network structure

We consider a finite population of n units indexed by $V = [n]$. To represent possible pathways of interference, as well as patterns of homophily among the units, we equip this population with an undirected weighted network $G = (V, E, W)$, where $E \subseteq V \times V$ is the set of edges and $W \in [0,1]^{n \times n}$ is a symmetric weight matrix. We take $(i, j) \in E$ if and only if $W_{ij} > 0$, so that an edge between units i and j indicates that they are linked through a relationship relevant to interference or homophily, while the magnitude of W_{ij} reflects the strength of that relationship. We write $d_i := \sum_{j=1}^{n} W_{ij}$,

$D := \mathrm{Diag}(d)$, $L := D - W$, where d_i is the weighted degree of unit i, D is the diagonal degree matrix, and L is the graph Laplacian. We let $L^\dagger$ denote the Moore–Penrose pseudoinverse of L. Since our framework is formulated relative to an underlying edge-weight matrix W encoding the relevant interference and homophily structure, the main analysis does not require a stochastic model for how the network was generated. Throughout most of the paper, for clarity of exposition, we assume that the relevant interference/homophily network is observed without error. That is, the analyst observes a network $G_{\mathrm{obs}} = (V, E_{\mathrm{obs}}, W_{\mathrm{obs}})$ that coincides with the underlying network $G = (V, E, W)$ governing the dependence structure of interest.

In many applications, however, the relevant network is not directly observed and must instead be constructed from a proxy. Examples include self-reported ties, administrative records, spatial or infrastructural adjacency, digital interaction logs, and exposure records. Such proxies may be subject to missingness, measurement error, thresholding, temporal mismatch, or other distortions, so that the observed network need not coincide with the latent one. Within the network interference literature, a small body of work has begun to examine the consequences of network and exposure misspecification [Aronow and Samii, 2017, Sävje, 2024, Weinstein and Nevo, 2026]. In this spirit, Section 5.2.2 develops our robustness framework for settings in which the observed network is a noisy proxy for the underlying interference/homophily structure. In particular, we study a natural model in which the observed edge weights are unbiased for the latent edge weights, $\mathbb{E}[(W_{\mathrm{obs}})_{ij}] = W_{ij}$. We also discuss in Example 5.3.4 the setting in which the observed network is a first-order approximation to a richer interference structure, for example when interference extends beyond immediate neighbors and decays with distance. For ease of reference, a glossary containing a subset of key notation is provided in Table 5.1 in the Appendix.

In a similar spirit to Jagadeesan et al. [2020], we focus on the design problem for point estimation via worst-case MSE control in the finite population setting. Thus, we do not consider the problem of inference in detail, although our bounds could in principle be combined with generic concentration inequalities to yield conservative confidence intervals.

5.2 Homophily and Heterogeneous Variation

Network interference models how the outcome of units depend on the treatment assignments of neighboring units in the network. With regards to the GATE, the bias arising from network interference can be written as the component of the potential outcome model accounting for disparate

treatments among units. As such, the bias notably vanishes in a graph where all units are treated (respectively controlled). In this section, we focus on the terms that are left over, the baseline α, and treatment effect ϕ. We discuss the interference terms in full detail in Section 5.3. Denote the outcomes by $Y := Y(z)$, with treatment assignment $z \in \{0,1\}^n$. We consider outcomes associated with $\mathbf{0}$ and $\mathbf{1}$, the treatment assignment vectors of all 0s and 1s, respectively. We can express, for each instance of the sampled network,

$$\alpha := Y(\mathbf{0}), \quad \alpha + \phi := Y(\mathbf{0}) + [Y(\mathbf{1}) - Y(\mathbf{0})] = Y(\mathbf{1}). \tag{5.2.1}$$

We decompose these vectors $\alpha = h_\alpha + \varepsilon_\alpha$, and $\phi = h_\phi + \varepsilon_\phi$, where h_α, h_ϕ are homophilous components, capturing smooth variation of the potential outcome coefficients with the graph structure. We formalize this smoothness via a quadratic constraint, and define this formally in Definition 5.2.2. The $\varepsilon_\alpha, \varepsilon_\phi$ terms are heterogeneous variation, components not explained by homophily and interference, formalized by moment constraints. These are nothing but the residuals or model misspecifications, terms more commonly seen in other literature, if we model the true generative (potential) outcomes by just interference and homophily.

Remark 5.2.1. It is equivalent to write $\alpha = \bar{\alpha} + h_\alpha + \varepsilon_\alpha$, and $\phi = \bar{\phi} + h_\phi + \varepsilon_\phi$, where $h_\alpha, h_\phi, \varepsilon_\alpha, \varepsilon_\phi$ are mean-zero, and $\bar{\alpha}, \bar{\phi}$ are constant vectors. We note that the decompositions into $h_\alpha + \varepsilon_\alpha$ and $h_\phi + \varepsilon_\phi$ are generally not unique. Our results below hold for any admissible decomposition satisfying the stated homophily, interference, and heterogeneous-variation conditions. In Section 5.4.1, we distinguish a particular admissible decomposition, which we call the *min-min-max* decomposition.

5.2.1 Homophily

We formalize the definition of homophily we consider. To our knowledge, the only other work that has used a similar definition is Ugander and Yin [2023]. Informally, homophily captures the extent to which similar nodes are strongly connected and dissimilar nodes are weakly connected. The quantity $\sum_{(i,j)\in E} W_{ij}(h_i - h_j)^2$ encapsulates this idea by aggregating the dissimilarities in potential outcomes coefficients h_i across edges. In a connected graph that is well-clusterable, high levels of homophily could result in clusters of nodes that are very similar while being significantly dissimilar across clusters. Large dissimilarities across clusters would result in a large value of $\sum_{(i,j)\in E} W_{ij}(h_i - h_j)^2$.

Definition 5.2.2 (η-homophily)**.** We say that the vector h satisfies (η, L)-homophily, or is (η, L)-homophilous, if $h^T L h \leq \eta$, where $L := D - W$ is the graph Laplacian. That is, $\sum_{(i,j)\in E} W_{ij}(h_i - h_j)^2 \leq \eta$.

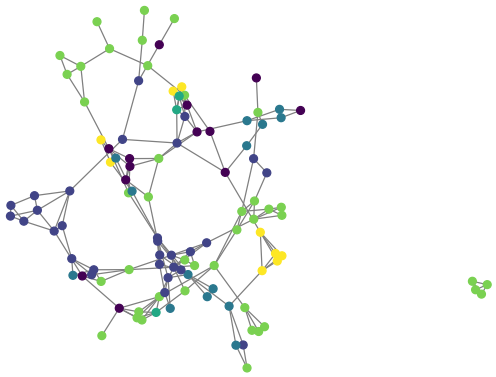


Figure 5.2.1: A real example of homophily: network of home visits between members of a village (no.6), using data from Banerjee et al. [2013a]. Node colors indicate the caste of each village member.

In such a setting, one natural design might be to have a spread out treatment assignment, to diversify treatment samples across the natural clusters. In Section 5.4, we demonstrate how this formalization leads to a design that distributes treatment assignments across the underlying graph clusters, aligning with intuition.

Remark 5.2.3. Unless otherwise specified, we use η-homophily to refer throughout to homophily defined with respect to the graph Laplacian L, and we suppress this dependence in the notation for brevity. Other kernels may be more appropriate in certain applications; see, for example, Remark 5.D.3. More generally, Appendix 5.F.1 treats $(\eta, \mathcal{K})$-homophily for an arbitrary similarity kernel $\mathcal{K}$, allowing the notion of homophily to capture, for example, spatial proximity or similarity in other covariates.

5.2.2 Robustness to heterogeneous variation

Our model accounts for heterogeneous variation via the ε components, which we only constrain by certain moment conditions. In its simplest form, this corresponds to an ℓ_2 constraint on ε which

leads to an operator norm as previously realized in Harshaw et al. [2024]. Indeed, let $X = \mathbb{E}[xx^T]$, $x = 2z - 1$, then $\sup_{\|\varepsilon_f\|_2^2 \leq \kappa} \mathrm{Tr}(\varepsilon_f^T X \varepsilon_f) = \kappa \|X\|_{\mathrm{op}}$, for $f = \alpha, \phi$.

Our framework encompasses heterogeneous variation arising both from unobserved subject covariates and from unobserved dependence between units independent of treatment assignment. In the latter case, κ bounds the second-moment magnitude of the dependence not captured by the observed network used in the homophily and interference models, which can be viewed as analogous to violations of conditional unconfoundedness. In works such as Li et al. [2021], the setting of interference under noisy network observations has previously been considered, resulting in stochasticity of the potential outcome model. The observed network is a sample from a super-population of possible dependencies between the finite population of experimental units. In Example 5.3.4, we discuss how this noisy interference is encapsulated by our generalized notion of network interference. Our notion of robustness in the current section addresses the resulting correlation in the baseline α and treatment effect ϕ in addition to that induced by unobserved subject covariates and homophily with respect to the observed network. In particular, we allow for ε_f, $f = \alpha, \phi$, to act as a random effect, drawing particular inspiration from the spatial statistics literature [Gaetan et al., 2010]. In our potential outcomes framework, we place constraints on $\Sigma_f := \mathbb{E}[\varepsilon_f \varepsilon_f^T]$, and we discuss how different penalties relate to inter-unit dependencies. This encompasses relationships between the outcomes that are not restricted to the observed network encoded by homophily. We employ duality to relate this back to the treatment design.

Example 5.2.4. A prominent example of stochastic heterogeneous variation is through the lens of noisy network observations. Consider the setting where α (and similarly ϕ) is a fixed function of the units in our finite population that is η-homophilous with respect to the graph G with Laplacian L. Thus, we can express $\alpha = L^{\dagger/2}\beta$ for a coefficient vector β with length $\sqrt{\eta}$. Now consider the observed network G_{obs} with corresponding Laplacian L_{obs} (see, for example, Figure 5.D.2). We can decompose $\alpha = L^{\dagger/2}\beta = h + \varepsilon, h := L_{\mathrm{obs}}^{\dagger/2}\beta$, and $\varepsilon := (L^{\dagger/2} - L_{\mathrm{obs}}^{\dagger/2})\beta$. In this decomposition, h is η-homophilous with respect to the observed graph, and ε is a function of the randomly observed graph structure. Imposing conditions on $\Sigma := \mathbb{E}[\varepsilon\varepsilon^T]$ imposes minimal moment constraints on the extent of heterogeneous variation.

We can characterize worst-case bounds in terms of Schatten norms $q^\star$, which constrains Σ, and q, which penalizes X. In particular, $\sup_{\|\Sigma_f\|_{q^\star} \leq \kappa} \mathrm{Tr}(X\Sigma_f) = \kappa \|X\|_q$. We write this out formally in Section 5.4. When $q^* = 1$, $q = \infty$, we recover the previous operator-norm bound. When $q^* = q = 2$,

we get a bound in terms of the Frobenius norm of the covariance matrix, $\kappa\|X\|_F$. This leads to a novel connection to Viviano et al. [2023].

Concretely, these are realized as penalizations that encourage randomization in our designs, with the extreme case being unit-level Bernoulli randomization, whenever $q > 1$. In this sense, randomization yields robust designs that ensure a controlled worst-case MSE under minimal moment assumptions. This underscores the interpolation of the two extremes of randomization and determinism. In principle, we uniformly independently randomize when we have no information and assumptions at all. Conversely, if we had all the necessary information, i.e., all potential outcomes were encoded through our notions of homophily and interference, then we can deterministically solve for an optimal design.

The importance of randomization for robustness has a long history, and here we mention only a few representative references most closely related to our perspective. Fisher [1925, 1935] first emphasized this importance through the role of randomization in yielding unbiased estimates of experimental error, which are vital for valid testing procedures. Wu [1981] further studied robustness from the perspective of the worst-case mean squared error; showing that the Bernoulli randomized design is minimax optimal over invariant sets of model violations, as is also reflected in our framework. The paper also describes how efficiency can be gained with structured assignments aligned with particular patterns of violations. More recently, Kallus [2018] showed that, without structural information on how the outcomes depend on the particular covariates, Bernoulli randomization is minimax optimal with respect to variance.

5.3 Potential Outcomes Model and Network Interference

Let $z \in \{0,1\}^n$ denote the binary treatment assignment vector. We consider the following potential outcome model

$$Y_i(z) = \alpha_i + \phi_i z_i + s_i(z), \tag{5.3.1}$$

with $z_i \in \{0,1\}$, for all $i \in [n]$. The term $s(z)$ denotes interference that vanishes under constant treatment, i.e., $z \in \{\mathbf{1}, \mathbf{0}\}$, with α denoting the baseline, and ϕ the treatment effect (see Equation (5.2.1)). Together with Equation (5.2.1), $s(z)$ is uniquely determined.

Definition 5.3.1 (γ-interference)**.** We say that the outcome component s satisfies (γ, L)-interference,

or is (γ, L)-interferent, if $\sup_{z\in\{0,1\}^n} s(z)^T L^\dagger s(z) \le \gamma$, where $L^\dagger$ is the pseudoinverse graph Laplacian.

Remark 5.3.2. Similar to our definition of homophily (see Remark 5.2.3), unless otherwise specified, we use γ-interference to refer to interference with respect to L, suppressing this dependence in the notation for brevity. For a similar level of interference with respect to a different kernel $\mathcal{K}$, we say that such a vector s is $(\gamma, \mathcal{K})$-interferent.

Example 5.3.3 (Neighborhood Interference)**.** A form of $s(z)$ that is commonly studied in the literature, such as in Chin [2019], Eckles et al. [2017], Harshaw et al. [2023], Yu et al. [2022], is the linear neighborhood interference term, i.e., $s(z) \propto D^{-1}Lz$. This corresponds to a neighborhood interference setting, where a unit's outcome depends only on its own and its neighbors' treatment assignments, not on assignments beyond its neighborhood. Let the normalized graph Laplacian $\tilde{L} := I - D^{-1}W$, i.e., $\tilde{L} = D^{-1}L$. A simple outcome model with such a $(\gamma, \tilde{L})$-interferent s component with a $\sqrt{n}$-scaling reminiscent of the local asymptotic framework [Viviano et al., 2023, Hirano and Porter, 2009] is $Y_i = \alpha_i + \beta_i z_i + \frac{\gamma}{\sqrt{n}} \sum_j \frac{W_{ij}}{\sum_j W_{ij}} z_j$. Here, $s(z) = -\frac{\gamma}{\sqrt{n}}\tilde{L}z$, i.e., s is $(\gamma, \tilde{L})$-interferent. Additionally, the normalized Laplacian term appropriately vanishes when the full population is treated or assigned to control. We describe this in more detail in Example 5.D.5.

To capture settings with stochastically misspecified interference, we cannot use the approach from Example 5.2.4, since the resulting misspecification would be correlated with treatment assignment, complicating our later analysis. Instead, we rely on the intrinsic robustness of our constraint, which allows generalization beyond neighborhood interference models.

Example 5.3.4 (Full Network Interference)**.** Our interference characterization allows for "perturbations" of the interference structure. Let $\tilde{L}$ be the graph Laplacian of the perturbed network, and let L be the graph Laplacian underlying a neighborhood interference structure satisfying γ-interference. If both the perturbed and unperturbed networks are connected, then the column and row spaces of L and $\tilde{L}$ are equal, and hence $\tilde{L}^{\dagger/2}L^{1/2}L^{\dagger/2} = \tilde{L}^{\dagger/2}$, and $L^{\dagger/2}L^{1/2}\tilde{L}^{\dagger/2} = \tilde{L}^{\dagger/2}$. Thus, if $\|\tilde{L}^\dagger L\|_{\text{op}} \le c$, and both the perturbed and unperturbed networks are connected, we can write $s(z)^T\tilde{L}^\dagger s(z) = s(z)^T(\tilde{L}^{\dagger/2}L^{1/2})L^\dagger(L^{1/2}\tilde{L}^{\dagger/2})s(z) \le cs(z)^T L^\dagger s(z) \le c\gamma$ for all $z \in \{0,1\}^n$. In other words, if L and $\tilde{L}$ are close in spectrum, i.e., $\|\tilde{L}^\dagger L\|_{\text{op}} \le c$, then s being (γ, L)-interferent implies that s is $(c\gamma, \tilde{L})$-interferent. For instance, if $\tilde{L}$ encapsulates a full-interference structure with decaying interference (for example, Leung [2022a]) across nodes with longer distances such that $\tilde{L}$ is well

approximated by the neighborhood interference Laplacian L, we can write $c = 1 + O(\psi)$ for some $\psi \ll 1$. We describe this in more detail in Example 5.D.6.

5.4 Estimation and Trade-offs

5.4.1 GATE estimation

Definition 5.4.1 (Global Average Treatment Effect (GATE))**.** The Global Average Treatment Effect (GATE) is defined to be

$$\tau = \frac{1}{n} \langle \mathbf{1}, Y(\mathbf{1}) - Y(\mathbf{0}) \rangle = \frac{1}{n} \langle \mathbf{1}, \phi \rangle .$$

Our primary focus is on the Horvitz-Thompson estimator with uniform treatment probability p across nodes, $p = \mathbb{P}(z_i = 1) = \mathbb{P}(z_j = 1)$ for all $i, j \in [n]$:

$$\hat{\tau} = \left\langle Y, \frac{z}{pn} - \frac{\mathbf{1} - z}{(1-p)n} \right\rangle .$$

For ease of exposition, we present our framework primarily in the symmetric setting, i.e., $p = 1/2$. The general case is given in Appendix 5.C.

We note here that the particular Horvitz-Thompson estimator we focus on throughout the paper is the vanilla one. In the literature on interference and cluster randomized designs, it is common to pair Horvitz-Thompson estimators with neighborhood exposure mappings [Aronow and Samii, 2017, Ugander et al., 2013, Leung, 2022b, Thiyageswaran et al., 2024, Eichhorn et al., 2024].

Under a correctly specified exposure mapping under network interference, the Horvitz-Thompson estimator remains unbiased, albeit with high variance. In contrast, the vanilla Horvitz-Thompson estimator is biased in the presence of interference, though it typically exhibits lower variance.

We focus on the vanilla estimator because it is much simpler, and our primary goal is to study the tradeoffs that occur under varying levels of interference, homophily and heterogeneous variation, for various designs. This includes settings without interference, where using a Horvitz-Thompson estimator paired with a neighborhood exposure mapping would introduce unnecessary variance. Moreover, applying exposure mappings requires specifying them correctly, and under more general forms of interference beyond neighborhood interference, characterizing the correct exposure mapping (and the imprecision that arises from a misspecified exposure mapping) becomes much more

complicated. In Appendix 5.I we consider our framework with respect to the difference-in-means estimator. In fact, our framework applies to more general estimators of the form $\hat{\tau} = \langle Y, w \rangle$, with general weights w.

This distinction is also important when interpreting MSE rates. In particular, under a correctly specified full-neighborhood exposure mapping, bounded outcomes, and unit-level Bernoulli randomization, uniformly bounded degree implies that the exposure probabilities are uniformly bounded away from zero. The corresponding exposure-based Horvitz–Thompson estimator is unbiased and attains an $(O(n^{-1}))$ variance. In what follows, we retain graph-dependent quantities appearing in our MSE bounds rather than using bounded degree as the primary point of comparison.

The measure of precision of the estimator we focus on is the worst-case mean squared error (MSE) $\sup_{h_f, \Sigma_f, s \in \chi(\eta,\gamma,\kappa)} \mathbb{E}[(\hat{\tau} - \tau)^2]$, with respect to the potential outcomes within the class $\chi(\eta, \gamma, \kappa) = \chi_{q^*}(\eta, \gamma, \kappa) = \{h_f, \varepsilon_f, s : \|L^{1/2} h_f\|^2 \leq \eta, \|\Sigma_f\|_{q^*} \leq \kappa, \|L^{\dagger/2} s(z)\|^2 \leq \gamma \text{ for all } z \in \{0,1\}^n, \text{ for } f = \alpha, \phi\}$. Here q^* denotes the Schatten exponent used to connstrain the exponent Σ_f, q denotes its Hölder conjugate exponent, defined by $1/q^* + 1/q = 1$. Consequently, the dual penalty appearing in the MSE bound is $\|X\|_q$.

When τ and $\hat{\tau}$ incorporate stochastic network misspecification, this expectation is understood to be the iterated expectation $\mathbb{E}[\mathbb{E}[(\hat{\tau} - \tau)^2 \mid \varepsilon_f]]$, $f = \alpha, \phi$, over the network misspecification residual and the design, respectively.

We call $\chi(\eta, \gamma, \kappa)$ the *min-min-max* model class to reflect (η, γ, κ) chosen to minimize the resulting worst-case MSE upper bound in Equation (5.4.1) among all admissible decompositions of α and ϕ. This parameterization, arising from the corresponding trade-off-admissible *min-min-max* decomposition of α and ϕ, yields the sharpest worst-case MSE control within our framework and is unique up to alternative optimal tuples. Specifically, the components h_α, h_ϕ and $\varepsilon_\alpha, \varepsilon_\phi$ optimally capture, relative to the underlying graph structure, the homophily and heterogeneous variation present in α and ϕ, by effectively separating graph-aligned low-Rayleigh-quotient variation from residual heterogeneous variation; see also Remark 5.D.4.

Theorem 5.4.2 (MSE bound)**.** *Let $p := \mathbb{P}(z_i = 1) = 1/2$ for all $i \in [n]$, $a = \langle \alpha, \mathbf{1}/n \rangle$, and $b = \langle \phi, \mathbf{1}/n \rangle$. Let $|\langle s(z), \mathbf{1}/n \rangle| \leq \sqrt{\delta}$ for all z. Let also $x \in \{-1, 1\}^n$ by defining $x := 2z - 1$. Define $X := \mathbb{E}[xx^T]$ and $\Delta := (a + b/2)^2 + \delta$. Then, for $q \geq 1$,*

$$\sup_{\chi(\eta,\gamma,\kappa)} \mathbb{E}[(\hat{\tau} - \tau)^2] \leq \frac{28}{n^2} \{\Delta \operatorname{Tr}(\mathbf{1}\mathbf{1}^T X) + \frac{5\eta}{4} \operatorname{Tr}(L^\dagger X) + \frac{5\kappa}{4} \|X\|_q + \gamma \operatorname{Tr}(LX)\}. \tag{5.4.1}$$

The bound in Equation (5.4.1) upper-bounds the worst-case MSE with terms linear in the covariance matrix X. We later show that this relaxation of the problem makes it tractable for simple optimization approaches. Thus, to control the worst-case MSE it suffices to minimize the right-hand side in Equation (5.4.1) over the covariance matrix X. The proof of Theorem 5.4.2 is in Appendix 5.J.2. We assume that the interference term is bounded in its center, i.e., $\sup_z \langle s(z), \mathbf{1}/n \rangle \leq \sqrt{\delta}$, and this is implicitly adopted in later MSE bounds.

When $\eta, \gamma, \kappa, q^\star$ are fixed, the scaling in Theorem 5.4.2 is governed jointly by the graph, the design covariance X, and the min-min-max decomposition through $\chi(\eta, \gamma, \kappa)$: the effective rate is $O(n^{-2} r_n)$, where $r_n = \max\{1^\top X 1, \mathrm{Tr}(LX), \mathrm{Tr}(L^\dagger X), \|X\|_q\}$. For the naive Bernoulli unit-level randomization where $X = I$, $r_n = O(n)$ for typical well-connected graph settings where $\mathrm{Tr}(L^\dagger) = O(n)$, like expander graphs, as well as bounded average degree settings.

In Appendix 5.G, we discuss an alternative where we constrain the quadratic form $s(z)^T(\delta \mathbf{1}\mathbf{1}^T + L)^{-1} s(z)$. If instead we were to constrain the quadratic form $s(z)^T(\lambda I + L)^{-1} s(z)$, we would arrive at the same bound up to a constant additive factor without necessitating this assumption. There are many alternatives that lead to fundamentally equivalent optimization problems, with another alternative provided in Example 5.F.7.

Comparison of graph-dependent rates. The bound in Theorem 5.4.2 can also be expressed in terms of standard graph quantities. Let $\lambda_2(L_n)$ denote the smallest nonzero eigenvalue of the graph Laplacian and let $\bar{d}_n := n^{-1} \mathrm{Tr}(L_n)$ denote the average weighted degree. Since a balanced design is feasible, the optimal worst-case objective is no larger than the value at a balanced feasible design. For any such design covariance X, we have $\mathrm{Tr}(X) = n$ and $\mathbf{1}^\top X \mathbf{1} = 0$. Moreover,

$$\mathrm{Tr}(L_n^\dagger X) \leq \|L_n^\dagger\|_{\mathrm{op}} \mathrm{Tr}(X) = \frac{n}{\lambda_2(L_n)},$$

and, for $q \geq 1$,

$$\|X\|_q \leq \|X\|_1 = n.$$

Finally,

$$\mathrm{Tr}(L_n X) = \mathbb{E}[x^\top L_n x] \leq 4 \sum_{\{i,j\} \in E_n} W_{ij} = 2n\bar{d}_n.$$

Consequently,

$$\sup_{\chi(\eta_n,\gamma_n,\kappa_n)} \mathbb{E}\left[(\widehat{\tau}-\tau)^2\right] \leq \frac{28}{n}\left\{\frac{5\eta_n}{4\lambda_2(L_n)} + \frac{5\kappa_n}{4} + 2\gamma_n\overline{d}_n\right\}. \tag{5.4.2}$$

Thus, the parametric $O(n^{-1})$ rate follows whenever

$$\frac{\eta_n}{\lambda_2(L_n)} + \kappa_n + \gamma_n\overline{d}_n = O(1),$$

while consistency only requires this quantity to be $o(n)$. In particular, bounded maximum degree is not required: for graph sequences with $\lambda_2(L_n)$ bounded away from zero and bounded average degree, the bound remains $O(n^{-1})$ for bounded model-class parameters even when $d_{\max,n}$ grows with n.

This gives a direct comparison with the graph-dependent guarantees of Ugander and Yin [2023]. To avoid confusion with our heterogeneous-variation parameter κ_n, let ρ_n denote their restricted-growth coefficient. Their Theorem 4.8 gives, under a correctly specified full-neighborhood exposure model and bounded outcomes,

$$\mathrm{Var}\left(\widehat{\tau}_{\mathrm{RGCR}}\right) \leq \frac{2\overline{Y}^2}{n}(d_{\max,n}+1)^2\rho_n^4\left\{p^{-1}+(1-p)^{-1}\right\}.$$

At $p = 1/2$, this becomes

$$\mathrm{Var}\left(\widehat{\tau}_{\mathrm{RGCR}}\right) \leq \frac{8\overline{Y}^2}{n}(d_{\max,n}+1)^2\rho_n^4.$$

Hence, the two bounds depend on different aspects of graph geometry: the guarantee from Ugander and Yin [2023] is governed by worst-case degree and local metric growth, whereas (5.4.2) is governed by average degree and algebraic connectivity. For example, if $\lambda_2(L_n)$ is bounded away from zero and $\overline{d}_n = O(1)$, our graph-dependent prefactor can remain bounded even when $d_{\max,n}$ increases. Conversely, when $\lambda_2(L_n)$ is small, (5.4.2) may be conservative, and the original design-specific quantities $\mathrm{Tr}(L_n^{\dagger}X)$ and $\mathrm{Tr}(L_nX)$ in Theorem 5.4.2 provide a sharper characterization.

We note that the exposure-based Horvitz–Thompson estimator in Ugander and Yin [2023] is unbiased under a correctly specified full-neighborhood exposure mapping, whereas our vanilla estimator controls total MSE over the broader class $\chi(\eta_n, \gamma_n, \kappa_n)$.We therefore view the comparison above as characterizing the different graph complexities governing the two approaches rather than as a uniform dominance result.

5.4.2 Trade-offs

Trade-off parameters

Our design flexibly navigates homophily, robustness, and interference trade-offs parameterized by $\eta, \kappa, \gamma \in \mathbb{R}$:

1. ${h_\alpha}^T L h_\alpha \leq \eta$, and ${h_\phi}^T L h_\phi \leq \eta$ (see also Remark 5.2.3 for a measure of homophily including degree similarity)
2. $\|\Sigma_\alpha\|_{q^*} \leq \kappa$, and $\|\Sigma_\phi\|_{q^*} \leq \kappa$
3. $s(z)^T L^\dagger s(z) \leq \gamma$ for all $z \in \{0,1\}^n$

As is typical in conducting power analyses, our method depends on baseline and treatment effect size, encapsulated by Δ. We dissect the three components in the trade-offs above.

5.4.2.1 Cut Minimization

If we fix the treatment vector x, then $X = xx^T$. The term $\mathrm{Tr}(LX) = x^T L x$ in the theorem statement above is the same as the objective function in the cut minimization problem. Indeed, for $L = D - W$, and $x \in \{-1,1\}^n$, $\frac{1}{4} x^T L x = \sum_{i \in S, j \in S^c} W_{ij}$, where S and S^c make up the cut partition. Thus, this $x^T L x$ term alone encourages a more "clustered" design. The term $\sum_{i \in S, j \in S^c} W_{ij}$ has previously appeared in Brennan et al. [2022], Viviano et al. [2023].

5.4.2.2 Effective Resistance

Once again, if we fix the treatment vector x, then $X = xx^T$. The term $\mathrm{Tr}(L^\dagger X) = x^T L^\dagger x$ can be expressed in terms of the effective resistances. Let R be the matrix of effective resistances across edges, so that the (i,j)-th entry is the effective resistance R_{ij} across edge (i,j). Then, for $x \in \{-1,1\}^n$, $x^T L^\dagger x = -\frac{1}{2} x^T \Pi^T R \Pi x$, where Π is the projection matrix to the subspace orthogonal to $\mathbf{1}$ (see Fontan and Altafini [2021]). In the graph sparsification literature, Spielman and Srivastava [2008] use effective resistances to sample network edges. Selecting a cut to minimize $x^T L^\dagger x$ corresponds to selecting a treatment design "spread-out" in the graph. We see in our MSE bounds that this provides a counterbalance to the $x^T L x$ term above that incentivizes clustered designs, and thus improves our estimation procedure in the presence of homophilous outcomes.

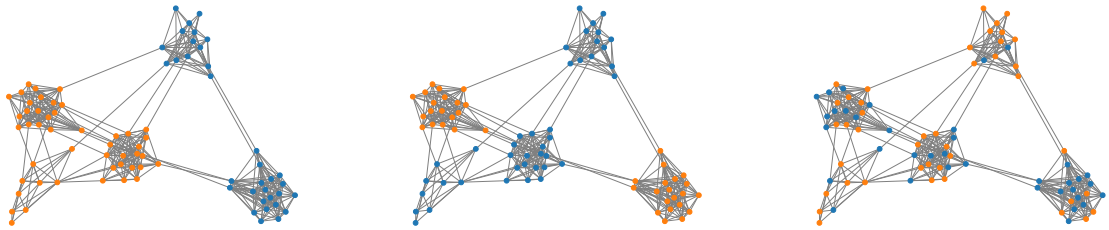


Figure 5.4.1: Optimized treatment assignments under interference-robustness tradeoffs on a synthetic network as heterogeneous variation levels increase from left to right. See Appendix 5.B for more details.

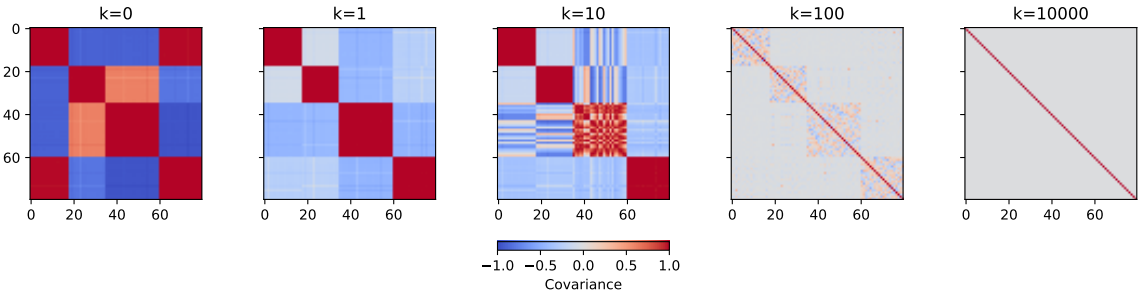


Figure 5.4.2: Optimal design covariance matrices under interference-robustness tradeoffs as heterogeneous variation levels increase from left to right. See Appendix 5.B for more details.

5.4.2.3 Randomization

We use Proposition 5.4.3 to bound the contribution of the heterogeneous variation ε_f to the MSE of our estimator. Note that for deterministic $\varepsilon_\alpha, \varepsilon_\phi$, $\Sigma_f = \varepsilon_f \varepsilon_f^T$, and $\|\Sigma_f\|_{q^*} = \|\varepsilon_f\|_{\ell_2}^2$, $f = \alpha, \phi$.

Proposition 5.4.3 (Deterministic heterogeneous variation bound)**.** *For PSD matrices $X \in \mathbb{R}^{n\times n}$, $\Sigma = \varepsilon_f \varepsilon_f^T \in \mathbb{R}^{n\times n}$, $f = \alpha, \phi$ and $\kappa \in \mathbb{R}$, $\sup_{\|\varepsilon_f\|_{\ell_2}^2 \le \kappa} \mathrm{Tr}(X\Sigma) = \sup_{\|\varepsilon_f\|_{\ell_2}^2 \le \kappa} \varepsilon_f^T X \varepsilon_f = \kappa \lambda_{max}(X)$, where $\lambda_{max}(X)$ is the maximum eigenvalue of X.*

This recovers the operator-norm bound that is used as a robustness measure in Harshaw et al. [2024]. We can generalize this further for the (uncentered) covariance matrix of the stochastic network misspecification.

Proposition 5.4.4 (Bhatia, 2013, Exercise IV.2.12)**.** *Let $\|\cdot\|_q$ denote the Schatten q-norm of a matrix. For symmetric positive semidefinite (PSD) matrices $X \in \mathbb{R}^{n\times n}$ and $\Sigma \in \mathbb{R}^{n\times n}$, and*

$1/q^* + 1/q = 1$, $\sup_{\|\Sigma\|_{q^\star} \leq \kappa} \operatorname{Tr}(X\Sigma) = \kappa \|X\|_q$.

Particular choices of q^* relate to existing trade-offs in the literature.

Corollary 5.4.5 (Worst-case bound over marginal variances)**.** *For PSD matrices $X \in \mathbb{R}^{n\times n}$ and $\Sigma \in \mathbb{R}^{n\times n}$, $q^\star = 1$, $q = \infty$, and $\kappa \in \mathbb{R}$,* $\sup_{\|\Sigma\|_1 \leq \kappa} \operatorname{Tr}(X\Sigma) = \kappa \lambda_{max}(X)$.

As a practical consideration, it is worth noting that the above supremization places little constraint on the off-diagonal of Σ. Indeed, the supremum is realized by a dense Σ, and as Σ is PSD, the Schatten-1 norm corresponds to the trace, hence the marginal variances of ε are directly constrained. However, in many circumstances, one may expect the coordinates of ε to be somewhat independent, making $q^* = 1$ an unrealistic modeling assumption.

In the other extreme Σ could be a diagonal matrix corresponding to independent heterogeneous variation, with the off-diagonal entries being completely nullified. Such a Σ is achieved if we were to instead supremize over the constraint $\|\Sigma\|_\infty \leq \kappa$, and the resulting error would be $\kappa \operatorname{Tr}(X) = \kappa n$, which is invariant with respect to the design. That is, if the (edge) misspecification was, for example, i.i.d. across the nodes, then additional randomization to account for network misspecification would be unnecessary, as this value κn is constant, independent of the chosen design. As a reasonable compromise between these two extremes, we highlight the Frobenius norm, $q^* = 2$.

Corollary 5.4.6 (Covariance structure worst-case bound)**.** *For PSD matrices $X \in \mathbb{R}^{n\times n}$ and $\Sigma \in \mathbb{R}^{n\times n}$, $q^* = q = 2$,* $\sup_{\|\Sigma\|_2 \leq \kappa} \operatorname{Tr}(X\Sigma) = \kappa \|X\|_2 = \kappa \sqrt{\operatorname{Tr}(X^2)} = \kappa \sqrt{\sum_{i=1}^n \lambda_i^2(X)}$. *Additionally,* $\sup_{\|\Sigma\|_2 \leq \kappa \|X\|_2} \operatorname{Tr}(X\Sigma) = \kappa \operatorname{Tr}(X^2) = \kappa \sum_{i=1}^n \lambda_i^2(X)$.

In the bounds in Corollary 5.4.6, the total covariance structure of $(\varepsilon_i)_{i=1}^n$ is constrained, rather than just the marginal variances. By constraining the Frobenius norm, we bound the sum of the squared entries of all of the elements of Σ, putting equal importance on on- and off-diagonal elements. The distinction between the two bounds in Corollary 5.4.6 is that the latter allows the norm of Σ to grow with X (and therefore n).

Under a cluster-level Bernoulli randomization design, the second part of Corollary 5.4.6 corresponds to the dominant term in the variance presented in Viviano et al. [2023]. Indeed, consider X such that $X_{ij} = 1$ if nodes i, j are in the same cluster, and 0 otherwise. This matrix X is the assignment covariance matrix for cluster-level Bernoulli randomization, where each cluster is independently assigned a treatment with probability p. The eigenvalues of this matrix are given by the sizes of the

clusters, $|\mathcal{C}_i|$, hence $\mathrm{Tr}(X^2) = \sum_i |\mathcal{C}_i|^2$. This is exactly the dominating term in the variance observed in Viviano et al. [2023, Theorem 3.2], hence the squared Frobenius norm generalizes this variance to generic designs. Further, as it appears as a convex penalty in our objective, it allows for easy computation. As unit-level clustering results in the smallest value of the Schatten q penalty, increases to the penalization parameter κ will force the solution towards this covariance structure.

Proposition 5.4.7. *Let $x \in \{-1, 1\}^n$ denote the treatment assignment under a Bernoulli randomization design. Then, for fixed $p = \mathbb{P}(x_i = 1)$, $i \in [n]$, and $q > 1$,* $\arg\min_{\substack{X \succeq 0; \\ X_{ii}=1}} \|X\|_q = \frac{1}{4p(1-p)} \mathrm{Cov}(x)$.

We note that the analogous result holds if $x \in \{-1, 1\}^n$ is the treatment assignment under a complete randomization design and $n_1 = n_0$, with the optimization constraint $X\mathbf{1} = \mathbf{0}$. We present this formally in Appendix 5.J.

Thus, large κ permits a more robust design. Indeed, κ penalizes the Schatten norm of X, driving the design toward a Bernoulli randomization as $\kappa \to \infty$ for $q > 1$, as formalized in Proposition 5.4.7. We illustrate this behavior in Figures 5.B.2 and 5.4.2, as the trade-off parameters in the SDP, described in Section 5.5.1, are varied.

Remark 5.4.8. We use the term randomization in a specific sense: by increased randomization, we refer to designs with smaller off-diagonal covariance elements, which lead to more robustness to heterogeneous variation. This stands in contrast to trivial randomizations, such as randomly flipping between fixed treatment and control assignment vectors, that preserve the covariance matrices and lie on level sets of the objective function.

To further demonstrate the trade-offs in our framework, we generate designs for a variety of trade-off parameters. Figure 5.1.2 illustrates the trade-offs of the minimum cut and minimum effective resistance components against each other. It depicts a "clustered" block structure in the covariance matrices on the left-end of the spectrum when $\gamma > \eta$, and a more "diffusive" covariance structure on the right-end of the spectrum when $\gamma < \eta$. Figures 5.1.1 and 5.4.1 depict the trade-offs on a graph generated from an SBM with equal cluster membership probability across five clusters. Figure 5.4.1 isolates the trade-offs between clustering and randomization. As the ratio κ/γ increases, the designs interpolate from a clustered design to a design that resembles unit-level Bernoulli randomization. Figure 5.1.1 isolates the trade-offs between clustering and effective resistance minimization. As the ratio η/γ increases, the designs interpolate from a clustered design to a design that is maximally spread-out.

Remark 5.4.9. The term $\mathrm{Tr}(\mathbf{1}\mathbf{1}^T X)$ in the bound encourages solutions that are orthogonal to $\mathbf{1}\mathbf{1}^T$ as $\Delta \to \infty$. Since the constant vector lies in the kernel of both the graph Laplacian and its pseudoinverse, this term encourages non-degenerate solutions.

5.4.3 Comparisons with Randomized Saturation Designs

Another popular approach in the literature is randomized saturation designs. In this class of designs, clusters are randomly assigned saturation levels, and each unit within a cluster is assigned to treatment independently with probability equal to the saturation level. We consider the subclass in which the saturation levels are independent across clusters and have a common mean and variance. As we show below, this class interpolates, at the covariance level, between unit-level Bernoulli randomization and cluster-level Bernoulli randomization.

Let $\pi_{\mathrm{RSD}}(c) \in [0,1]$ denote the random saturation assigned to cluster c, with

$$\mathbb{E}[\pi_{\mathrm{RSD}}(c)] = p.$$

Conditional on $\pi_{\mathrm{RSD}}(c)$, treatment assignments within the cluster are independent:

$$z_i \mid \pi_{\mathrm{RSD}}(c) \sim \mathrm{Bernoulli}(\pi_{\mathrm{RSD}}(c))\,, \qquad i \in c.$$

For two distinct units $i, j \in c$, the law of total covariance gives

$$\begin{aligned}\mathrm{Cov}(z_i, z_j) &= \mathbb{E}\big[\mathrm{Cov}\big(z_i, z_j \mid \pi_{\mathrm{RSD}}(c)\big)\big] \\ &\quad + \mathrm{Cov}(\mathbb{E}[z_i \mid \pi_{\mathrm{RSD}}(c)], \mathbb{E}[z_j \mid \pi_{\mathrm{RSD}}(c)]) \\ &= \mathrm{Var}(\pi_{\mathrm{RSD}}(c))\,,\end{aligned}$$

where the first term vanishes by conditional independence. Moreover,

$$\mathbb{P}(z_i = 1) = \mathbb{E}\big[\mathbb{P}\big(z_i = 1 \mid \pi_{\mathrm{RSD}}(c)\big)\big] = \mathbb{E}[\pi_{\mathrm{RSD}}(c)] = p,$$

so marginally $z_i \sim \mathrm{Bernoulli}(p)$ and therefore

$$\mathrm{Var}(z_i) = p(1-p).$$

Hence the induced correlation between two distinct treatment assignments in the same cluster is

$$\begin{aligned}\rho := \mathrm{Corr}(z_i, z_j) &= \frac{\mathrm{Cov}(z_i, z_j)}{\sqrt{\mathrm{Var}(z_i)\,\mathrm{Var}(z_j)}} \\ &= \frac{\mathrm{Var}(\pi_{\mathrm{RSD}}(c))}{p(1-p)}.\end{aligned}$$

Equivalently,

$$\mathrm{Var}(\pi_{\mathrm{RSD}}(c)) = \rho p(1-p). \tag{5.4.3}$$

Since $\pi_{\mathrm{RSD}}(c) \in [0,1]$ and $\mathbb{E}[\pi_{\mathrm{RSD}}(c)] = p$,

$$0 \le \mathrm{Var}(\pi_{\mathrm{RSD}}(c)) \le p(1-p),$$

and consequently $\rho \in [0,1]$.

We now express this interpolation in terms of the covariance matrix used in our design formulation. Let

$$x := 2z - \mathbf{1},$$

and denote by

$$X_{\mathrm{RSD}}(\rho) := \mathrm{Cov}(x)$$

the covariance matrix induced by the randomized saturation design with within-cluster treatment correlation ρ. Let X_{clust} denote the covariance matrix of x under independent cluster-level Bernoulli(p) randomization on the same collection of clusters. Define the corresponding normalized covariance matrices by

$$\bar{X}_{\mathrm{RSD}}(\rho) := \frac{X_{\mathrm{RSD}}(\rho)}{4p(1-p)}, \qquad \bar{X}_{\mathrm{clust}} := \frac{X_{\mathrm{clust}}}{4p(1-p)}.$$

For distinct units i, j in the same cluster,

$$\mathrm{Cov}(x_i, x_j) = 4\,\mathrm{Cov}(z_i, z_j) = 4\rho p(1-p),$$

whereas

$$\mathrm{Var}(x_i) = 4\,\mathrm{Var}(z_i) = 4p(1-p).$$

For units in different clusters, the covariance is zero by independence of the cluster saturation levels.

It follows that the normalized covariance matrix has diagonal entries equal to one, within-cluster off-diagonal entries equal to ρ, and between-cluster entries equal to zero. Therefore,

$$\bar{X}_{\mathrm{RSD}}(\rho) = \rho \bar{X}_{\mathrm{clust}} + (1-\rho) I. \tag{5.4.4}$$

The two extreme cases recover unit-level and cluster-level Bernoulli randomization. If $\rho = 0$, then

$$\mathrm{Var}(\pi_{\mathrm{RSD}}(c)) = 0,$$

so $\pi_{\mathrm{RSD}}(c) = p$ almost surely. Thus, conditional treatment assignment reduces to i.i.d. Bernoulli(p) randomization at the unit level, and

$$X_{\mathrm{RSD}}(0) = 4p(1-p)I, \qquad \bar{X}_{\mathrm{RSD}}(0) = I.$$

At the other extreme, if $\rho = 1$, then the saturation variance achieves its maximum possible value $p(1-p)$. This is attained by

$$\pi_{\mathrm{RSD}}(c) = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1-p, \end{cases}$$

independently across clusters. In this case all units within a cluster receive the same treatment assignment, which is exactly cluster-level Bernoulli(p) randomization. Hence

$$X_{\mathrm{RSD}}(1) = X_{\mathrm{clust}}, \qquad \bar{X}_{\mathrm{RSD}}(1) = \bar{X}_{\mathrm{clust}}.$$

Thus, as ρ varies from zero to one, randomized saturation designs trace the line segment between unit-level and cluster-level Bernoulli randomization in the space of normalized treatment-assignment covariance matrices.

Since $\min_{X \succeq 0, X_{ii}=1} J(X) \leq \min_{\rho \in [0,1]} J(X_{\mathrm{RSD}}(\rho))$, at the second-moment/covariance level, our design generalizes randomized saturation designs.

The simulations below demonstrate this behavior on a stochastic block model graph of four clusters of equal size. The probablity of edges connecting nodes within a cluster is 0.1 , while the probability of edges connecting nodes across a cluster is 0.0005. Figure 5.4.3 illustrates the interpolation behavior

formalized in Equation (5.4.4). Figures 5.4.4 and 5.4.5 illustrate $\operatorname{argmin}_{\rho\in[0,1]} J(X), \min_{\rho\in[0,1]} J(X)$ for different ratios of κ/γ. Figure 5.4.6 demonstrates that randomized saturation designs (with the optimal saturation levels) interpolate between cluster and unit-level Bernoulli randomizations, in their objective functions as well. Additionally, the SDP relaxation solution uniformly dominates the randomized saturation designs in performance, indicating the generalization above. The SDP together with the Gaussian rounding for $p = 1/2$ is shown to closely follow the optimized randomized saturation designs.

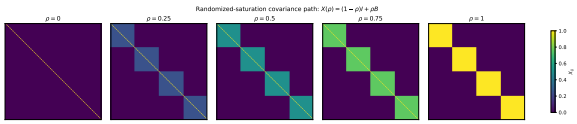


Figure 5.4.3: Covariance matrices of randomized saturation designs as saturation correlation levels ρ is varied.

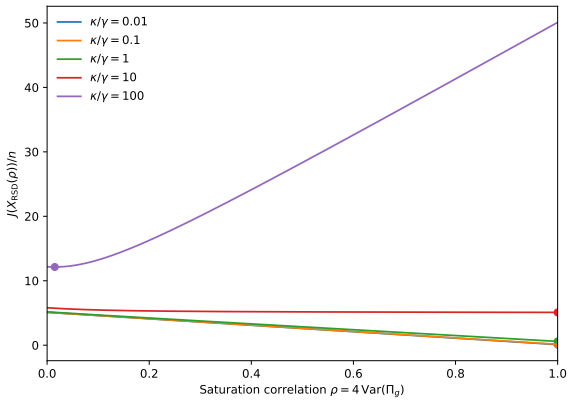


Figure 5.4.4: Normalized objective function across saturation correlation levels ρ together with the optimal saturation correlation levels for each κ/γ ratio.

5.4.4 Beyond second moments

Our approach is to relax the problem by upper-bounding the worst-case MSE with terms linear in the design covariance matrix X, so that it is amenable to optimization approaches. In reality, the true worst-case MSE may consist of terms with higher moments. To illustrate this concept, we can

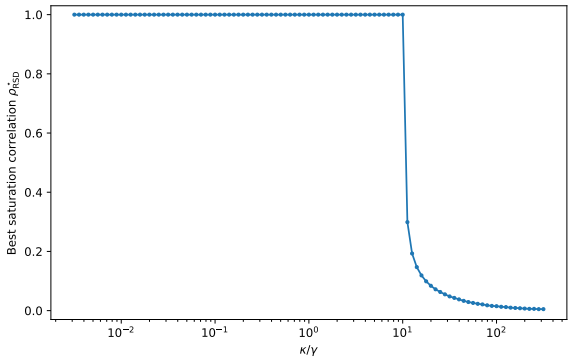


Figure 5.4.5: Optimal saturation correlation levels across varying κ/γ levels.

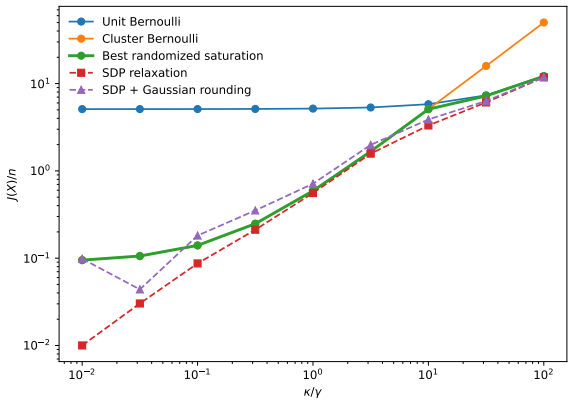


Figure 5.4.6: Normalized objective function evaluated across the different approaches under a SBM of 4 clusters of equal size. "Best randomized saturation" corresponds to the randomized saturation design with the optimal saturation correlation level ρ chosen to minimize the objective function (see Figures 5.4.4 and 5.4.5). "Cluster Bernoulli" corresponds to cluster randomization with clustering according to the generated clusters in the SBM.

consider as example the randomized saturation designs from the previous subsection. For this class of designs, while we can capture the covariance matrix behavior in Equation (5.4.4), different saturation designs with different higher order moments may share the same covariance matrix. Indeed, take for example, the setting where

$$\pi_{\mathrm{RSD},1}(c) = \begin{cases} 0.25 & \text{w.p. } 0.5, \\ 0.75 & \text{w.p. } 0.5. \end{cases}$$

and

$$\pi_{\mathrm{RSD},2}(c) = \begin{cases} 0 & \text{w.p. } 0.125, \\ 0.5 & \text{w.p. } 0.75, \\ 1 & \text{w.p. } 0.125. \end{cases}$$

Then, their marginal probabilities are equal $\mathbb{E}[\pi_{\mathrm{RSD},1}(c)] = 1/8 + 3/8 = 1/2$, and $\mathbb{E}[\pi_{\mathrm{RSD},2}(c)] = 0 \cdot 1/8 + 3/8 + 1/8 = 1/2$. Now, we check that the covariances are also equal. We begin by the variance calculation,

$$\mathrm{Var}(\pi_{\mathrm{RSD},1}(c)) = \frac{5}{16} - \frac{1}{8} = \frac{1}{16},$$

and similarly,

$$\mathrm{Var}(\pi_{\mathrm{RSD},2}(c)) = \frac{5}{16} - \frac{1}{8} = \frac{1}{16}.$$

Consider units i, j in the same cluster. Then, from the previous subsection, for both designs,

$$\mathrm{Cov}(z_i, z_j) = \mathrm{Var}(\pi_{\mathrm{RSD},1}(c)) = \frac{1}{16}.$$

Since the saturation probabilities are independent across clusters, $\mathrm{Cov}(z_i, z_j) = 0$, if i, j are in different clusters.

Thus, these different randomized saturation designs share the same marginal treatment probabilities, as well as the same covariance structure. They differ in their fourth moments; a simple

calculation shows that

$$\mathbb{E}[\pi_{\mathrm{RSD},1}(c)^4] = \frac{1}{2}\cdot\left(\frac{1}{4}\right)^4 + \frac{1}{2}\cdot\left(\frac{3}{4}\right)^4 = \frac{41}{256},$$

while

$$\mathbb{E}[\pi_{\mathrm{RSD},2}(c)^4] = \frac{1}{8}\cdot(0)^4 + \frac{3}{4}\cdot\left(\frac{1}{2}\right)^4 + \frac{1}{8}\cdot(1)^4 = \frac{44}{256}.$$

Thus, the covariance matrix may not necessarily characterize any particular design, and there may be multiple designs corresponding to the optimal covariance matrix we solve for.

5.5 Experimental Design

We describe the two design approaches we propose. To keep ideas simple and clear, in what follows we write the robustness trade-off terms only with general Schatten norms, rather than particular special cases.

5.5.1 MSE-minimizing SDP relaxation and Gaussian Rounding

Let $\Delta = (a + b/2)^2 + \delta$. Reparameterizing the right-hand side of Theorem 5.4.2 with $\tilde{\gamma}$,$\tilde{\eta}$, and $\tilde{\kappa}$, so that $\tilde{\gamma} := \frac{\gamma}{\Delta}, \tilde{\eta} := \frac{5\eta}{4\Delta}$, and $\tilde{\kappa} := \frac{5\kappa}{4\Delta}$, we can write an SDP more explicitly in terms of the new trade-off parameters:

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times n}} \quad & \tilde{\eta}\mathrm{Tr}(L^\dagger X) + \tilde{\gamma}\mathrm{Tr}(LX) + \tilde{\kappa}\|X\|_q + \mathrm{Tr}(\mathbf{1}\mathbf{1}^T X) \\ \text{s.t.} \quad & X \succeq 0 \\ & \mathrm{diag}(X) = \mathbf{1} \end{aligned} \tag{5.5.1}$$

The non-symmetric setting is presented in Appendix 5.C. We generate our treatment assignment vector ζ using Algorithm 2, drawing ideas from Goemans and Williamson [1995].

5.5.2 Adapted Gram-Schmidt Walk algorithm

As a second approach, we adapt the Gram-Schmidt Walk algorithm [Harshaw et al., 2024], a vector balancing procedure, to our problem. In particular, by specializing the robustness hyperparameter λ,

Algorithm 2 GaussianRounding at $p = 1/2$.

Require: Parameters: q, $\tilde{\gamma}$, $\tilde{\eta}$,$\tilde{\kappa}$, and marginal assignment probability p ▷ $\mathbb{P}(x_i = 1) = p$ for all $i \in [n]$

1: Solve the SDP in Equation (5.5.1) with parameters q, $\tilde{\gamma}$,$\tilde{\eta}$, and $\tilde{\kappa}$; let the solution be X_{SDP}
2: Sample $\xi \sim \mathcal{N}(0, X_{\text{SDP}})$ ▷ Multivariate Gaussian with covariance matrix X_{SDP}
3: **for** each $i \in [n]$ **do**
4: $\quad x'_i \leftarrow \text{sgn}(\xi_i)$
5: **end for**
6: **return** ζ

we provide an alternative algorithm that yields a guarantee for the worst-case MSE.

Focusing for simplicity on the homophily term $x^T L^\dagger x$ in our MSE expression, we can write $x^T L^\dagger x = x^T A^T A x$, with $A = \sqrt{\eta} L^{\dagger/2}$, or $A = \sqrt{\eta}\Lambda^{1/2} V^T$, where Λ is the matrix of eigenvalues of $L^\dagger$, V is the matrix of the corresponding eigenvectors. We then use the Gram-Schmidt Walk algorithm [Harshaw et al., 2024] on the augmented covariate matrix B with columns $B_i = [e_i \quad A_i]^T$ where e_i are standard basis vectors, and randomization parameter $\lambda = \frac{\sqrt{\kappa n^{1/q}}}{\sqrt{\omega_A^2 n + \sqrt{\kappa n^{1/q}}}}$ in Algorithm 4, where $\omega_A = \max_{i \in [n]} \|A_i\|$.

5.5.3 Comparing the two approaches

Both proposed methods extend to more general kernels, since the final optimization objective function retains the same form after replacing the Laplacian (see Appendix 5.F.1).

In the adapted Gram-Schmidt Walk algorithm, the bounds grow with n (see also discussion after Theorem 5.6.3), since this procedure is tailored to the minimization of the operator norm. By contrast, we can directly specify the Schatten-q norm penalties, via an SDP for common choices of q (e.g. $q \in \{1, 2, \infty\}$), or via convex optimization more broadly, leading to finer control. However, the computational efficiency of the SDP significantly drops for larger n, making vector balancing a practical alternative. We incorporate our trade-off formulation into the Gram-Schmidt Walk algorithm [Harshaw et al., 2024] in Algorithm 4, line 3.

Our application of the Gram-Schmidt Walk algorithm reduces to controlling $\text{Tr}(A^T A X)$ and $\tilde{\kappa}\|X\|_q$, where $A^T A := \tilde{\eta} L^\dagger + \tilde{\gamma} L + \mathbf{1}\mathbf{1}^T$. This casts worst-case MSE control as a covariate balancing problem, where the relevant covariates are induced by a choice of factor A of $A^T A$. One natural choice is the spectral factor arising from an eigendecomposition of $A^T A$. Other valid choices include

$A := (\tilde{\eta} L^{\dagger} + \tilde{\gamma} L + \mathbf{1}\mathbf{1}^T)^{1/2}$, or $A^T := [\tilde{\eta}^{1/2} L^{\dagger/2} \quad \tilde{\gamma}^{1/2} L^{1/2} \quad \mathbf{1}\mathbf{1}^{T\,1/2}]$. This framing would allow one to think about clustering as part of the covariate balancing problem, the original focus of Harshaw et al. [2024].

The SDP followed by Gaussian rounding described in Section 5.5.1 is specialized to uniform treatment probabilities, whereas the Gram-Schmidt Walk approach allows direct specification of non-uniform treatment probabilities: let $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ denote the vector of treatment probabilities, and set the initial vector $z_1 \leftarrow 2\mathbf{p} - 1$ in Algorithm 4, line 6. On the other hand, the SDP formulation makes it straightforward to add additional convex constraints (see also Appendix 5.I).

5.6 Theoretical guarantees

Theorem 5.6.2 provides worst-case MSE approximation bounds for the SDP solution paired with the Gaussian Rounding procedure. In Theorem 5.6.3, we provide worst-case MSE bounds for the Gram-Schmidt Walk algorithm by applying results from Harshaw et al. [2024]. The proofs are presented in Appendix 5.K.

5.6.1 SDP followed by Gaussian Rounding

Our proof of Theorem 5.6.2 utilizes the seminal ideas underlying the approximation bound for MAXCUT, as introduced in Goemans and Williamson [1995] (see Appendix 5.K.1 for more details). Our problem is similar in flavour, except we look at the minimization problem instead of maximization, and as such we would like an inequality in the opposite direction for a meaningful approximation bound.

We derive geometric identities associated with Algorithm 2. When $p := \mathbb{P}(\zeta_i = 1) = 1/2$, we have Grothendieck's identity, as derived in Goemans and Williamson [1995]. That is, $\mathbb{E}[\zeta_i \zeta_j] = \frac{2}{\pi} \arcsin \langle v_i, v_j \rangle$, where v_i is the i-th column vector of $X_{\mathrm{SDP}}^{1/2}$, and X_{SDP} is the solution to the SDP in Section 5.5.1. Indeed, when $p = 1/2$, it is equivalent to setting $\zeta_i = 1$ whenever $\langle \xi, v_i \rangle > 0$, $i \in [n]$. This is because $\mathbb{P}(\langle v_i, \xi \rangle > 0) = 1/2$. This identity allows us to compare the covariance matrix of the rounded assignment to the SDP solution, and hence to compare the resulting worst-case MSE bound after rounding.

Define $J(X) := \frac{28}{n^2} \{\Delta \operatorname{Tr}(\mathbf{1}\mathbf{1}^T X) + \frac{5\eta}{4} \operatorname{Tr}(L^{\dagger} X) + \frac{5\kappa}{4} \|X\|_q + \gamma \operatorname{Tr}(LX)\}$, the worst-case MSE upper bound given by the right-hand side of Equation (5.4.1). Let X^* denote the optimal achievable covariance matrix, i.e., one minimizing $J(X)$ over all covariance matrices achievable by a design. Let

$\Xi := \mathrm{Cov}(\zeta)$ be the covariance matrix of the rounded assignment returned by Algorithm 2.

Theorem 5.6.1 (Gaussian rounding MSE approximation bound)**.** *Let* $J(X) = \mathrm{Tr}(AX) + \tilde{\kappa}_n \|X\|_q$, *i.e.,* $A = \frac{28}{n^2}\{\Delta \mathbf{1}\mathbf{1}^T + \frac{5\eta}{4}L^\dagger + \gamma L\}$, *and* $\tilde{\kappa}_n = \frac{35\kappa}{n^2}$. *Let* q^* *be dual to* q, *so that* $1/q^* + 1/q = 1$. *Then, we get* $J(\Xi) \leq cJ(X^*)$, *where*

$$1 \leq c \leq \min\left(1 + \frac{\|A\|_{q^*}}{\tilde{\kappa}_n}, \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}\right),$$

whenever $A \succ 0$.

The result above bounds the approximation factor between the worst-case MSE upper-bound evaluated at the rounded design covariance and the minimum such upper-bound over all covariance matrices achievable by binary designs. We see that when there is more randomization, i.e., when $\tilde{\kappa}_n$ is larger relative to the interference, homophily, and non-degeneracy components, then the approximation factor bound is closer to one. Indeed, in this robustness-dominated regime, $\tilde{\kappa}_n \gg \|A\|_q$, and the "cost" of converting the SDP solution to a binary randomized design becomes small. On the other hand, if $\|A\|_{q^*} = \Theta(\tilde{\kappa})_n$, then $c = O(1)$. If the underlying network is highly clusterable, and $\lambda_2(L)$ is small, then unless $\tilde{\kappa}_n$ grows correspondingly, the upper-bound on the approximation factor might be large.

Theorem 5.6.2 (Gaussian hyperplane rounding MSE approximation bound)**.** *Assume the symmetric setting* $p = 1/2$. *Let* $\Xi := \mathrm{Cov}(\zeta)$ *be the covariance matrix of the rounded assignment returned by Algorithm 2, and let* X_{SDP} *be the solution to the SDP in 5.5.1. Assume that* X_{SDP} *is non-degenerate, i.e.,* $X_{SDP} \succeq \delta_\kappa I$, $\delta_\kappa > 0$. *Let* $c_\kappa := \left(\frac{2}{\pi} + \frac{2\psi_\kappa}{\pi\delta_\kappa} + \frac{1-2/\pi}{\delta_\kappa}\right)$, $\psi_\kappa = n\alpha_\kappa^3$, *where* α_κ *is the maximum off-diagonal entry of* X_{SDP} *in absolute value. Then, for the estimator* $\hat{\tau}_\zeta$ *under the design associated with* ζ, $\sup_{\chi(\eta,\gamma,\kappa)} \mathbb{E}[(\hat{\tau}_\zeta - \tau)^2] \leq J(\Xi) \leq c_\kappa J(X^*)$.

In the proof, we make use of the Taylor expansion of $\arcsin(X)$, following Nesterov [1997] in the MAXCUT problem. The resulting bounds depend on the tradeoff parameter κ: as $\kappa \to \infty$, the factor $\psi_\kappa \to 0$ while $\delta_\kappa \to 1$. To encourage non-degeneracy of X, the design parameters must be appropriately balanced. In particular, for clusterable graphs, the small Laplacian eigenvalues can lead to heavy penalization of the corresponding inverse Laplacian eigenvectors.

When κ is large, the rounding procedure closely approximates the optimal design covariance; in the limiting Bernoulli randomized design, $\alpha_\kappa = 0$. For a complete randomized design, $\alpha_\kappa = 1/(n-1)$.

Our results allow for $\alpha_\kappa = \Theta(1/n^{1/3})$, far more correlation than Bernoulli or complete randomizations, while maintaining a constant level of error. In contrast, when κ is very small, we can optimally solve the problem deterministically.

5.6.2 Adapted Gram-Schmidt Walk algorithm

We leverage Harshaw et al. [2024, Theorem 6.4] (see also Proposition 5.K.5) and its proof. In particular, we use the fact that the Gram-Schmidt Walk algorithm yields treatment assignments x that satisfy (1) $\mathrm{Cov}(x) \preceq \lambda^{-1} I$, and (2) $\mathrm{Cov}(Ax) \preceq \omega_A^2 (1-\lambda)^{-1} I$. We state the results when $p = 1/2$ formally below. The general version (incorporating non-symmetric designs) is presented in Appendix 5.K (Theorem 5.K.6).

Theorem 5.6.3 (Adapted Gram-Schmidt Walk error bound)**.** *Fix $q, p' \in [1, \infty]$, where $\|\cdot\|_{p'}$ denotes the Schatten-p' norm. Let $A := (\tilde{\eta} L^\dagger + \tilde{\gamma} L + \tilde{\Delta} \mathbf{1}\mathbf{1}^T)^{1/2}$, where $\tilde{\eta} = 5\eta/n^2$, $\tilde{\gamma} = 4\gamma/n^2$, $\tilde{\kappa} = 5\kappa/n^2$, and $\tilde{\Delta} = 4\{(a + b/2)^2 + \delta\}/n^2$, $a = \langle \alpha, \mathbf{1}/n \rangle$, $b = \langle \phi, \mathbf{1}/n \rangle$. The Gram-Schmidt Walk algorithm, with randomization parameter λ, returns an assignment vector x such that $\|\mathrm{Cov}(Ax)\|_{p'} \leq \omega_A^2/(1-\lambda)\|I\|_{p'}$, and $\|\mathrm{Cov}(x)\|_q \leq \|I\|_q/\lambda = n^{1/q}/\lambda$, for $\omega_A := \max_{i \in n} \|A_i\|$, A_i the ith column of A. Let $X := \mathrm{Cov}(x)$. Thus, for $q \geq 1$,*

$$\sup_{\chi(\eta,\gamma,\kappa)} \mathbb{E}[(\tau - \hat{\tau})^2] \leq 7 \left[\mathbb{E}[\|A(x-\mu)\|^2] + \tilde{\kappa} \|X\|_q \right] \leq 7 \left[\frac{\omega_A^2 n}{1-\lambda} + \tilde{\kappa} \frac{n^{1/q}}{\lambda} \right],$$

where $\mu_i = \mathbb{E}[x_i]$, for all $i \in [n]$. This is minimized for $\lambda = \frac{\sqrt{\tilde{\kappa} n^{1/q}}}{\sqrt{\omega_A^2 n} + \sqrt{\tilde{\kappa} n^{1/q}}}$, so that $\sup_{\chi(\eta,\gamma,\kappa)} \mathbb{E}[(\tau - \hat{\tau})^2] \leq 7(\omega_A n^{1/2} + \tilde{\kappa}^{1/2} n^{1/2q})^2$.

The asymptotic results in Harshaw et al. [2024] require the trade-off parameter $\lambda \to 1$. That is, their asymptotic results require that $\kappa/\eta, \kappa/\gamma, \kappa/\Delta \to \infty$ under $q = \infty$, mirroring the improvement in the approximation error bounds in Theorem 5.6.2.

For fixed γ, κ, η, appropriate choice of q, q^*, baselines, and treatment effects, the bound above is governed by the term $\omega_A n^{1/2}$. Indeed, the appropriate choice of the dual Schatten classes $(q^\star, q)$ depends on the form of the heterogeneous variation being modeled, as discussed in Section 5.4.2.3. The corresponding robustness parameter $\kappa = \kappa_n$ then has a natural scaling with n. In bounded degree, well-connected graphs such as expander graphs, for which $\max_i (L^\dagger)_{ii} = O(1)$ and $\max_i (L)_{ii} = O(1)$ one has $\omega_A = O(n^{-1})$, yielding an $O(n^{-1})$ worst-case MSE bound.

The results in Theorems 5.6.1 and 5.6.2 and Theorem 5.6.3 are notably different. In Theorems 5.6.1 and 5.6.2, we prove that rounding yields comparable MSE control to the optimal SDP solution. In contrast, in Theorem 5.6.3 we directly bound the worst-case MSE resulting from the adapted Gram-Schmidt Walk design. See also Section 5.5.3 where we discuss how the design from the Gram-Schmidt Walk algorithm compares to the design from the SDP followed by Gaussian rounding. In theory, we expect the bounds from the Gram-Schmidt Walk algorithm to be loose when n is larger, and q is smaller.

5.7 Discussion

We develop a novel framework to understand the three-way trade-off between homophily, network interference, and robustness to heterogeneous variation, recovering existing designs in extreme cases within the trade-off. Along the way, we generalize several existing results and designs in related literature. We demonstrated the complementary behaviors of our two proposed approaches, solving an SDP followed by Gaussian Rounding, and the adapted Gram-Schmidt Walk algorithm. In proving approximation guarantees via the SDP approach, we leveraged and generalized ideas utilized in proving approximation guarantees for the MAXCUT problem. This generalization may be of independent interest.

While we focus on the Horvitz–Thompson estimator for GATE estimation, our framework naturally extends to more general estimators and estimands, which we leave for future work. In this paper, we study optimal designs under known bounds η, κ, and γ. Such bounds may be informed by domain expertise or learned from prior experiments and historical network data, and our approach makes explicit how this information can be incorporated. An important direction for future work is extending the framework to online settings, where treatments are assigned sequentially and the parameters η, κ, and γ are learned adaptively. This paper focuses on the design problem for point estimation through worst-case MSE control in the finite-population setting. Developing accompanying inferential procedures, as well as identifying structural assumptions governing the sequence of networks as $n \to \infty$ under which consistency and sharp asymptotic guarantees hold, are natural directions for future work. Finally, for large-scale graphs, algorithms exploiting problem-specific structure could be used to accelerate the SDP, though this lies beyond the scope of the present work.

Code Availability

Code to replicate the experiments is available at this GitHub repository.

Acknowledgements

We thank Thomas Rothvoss for insightful discussions and for pointing us to Raghavendra and Tan [2012], Arun Chandrasekhar for pointing us to the village network dataset, Davide Viviano for sharing the causal clustering code for verification purposes, and Seth Temple and Lars van der Laan for feedback on an earlier draft.

Supplement to "Optimal Design under Interference, Homophily, and Robustness Trade-offs"

5.A Glossary of notation

Table 5.1: Selected glossary of notation used throughout the paper.

Term / symbol	Meaning
Population and network setup	
$L = D - W$	Graph Laplacian.
$L^{\dagger}$	Moore–Penrose pseudoinverse of the graph Laplacian.
$\widetilde{L} = I - D^{-1}W$	Normalized graph Laplacian used in neighborhood-interference examples.
Potential outcomes	
$x = 2z - \mathbf{1}$	Rademacher recoding of the treatment assignment, so $x \in \{-1, 1\}^n$.
α	Baseline vector, defined by $\alpha = Y(0)$.
φ	Treatment-effect vector, defined by $\alpha + \varphi = Y(1)$, equivalently $\varphi = Y(1) - Y(0)$.
$s(z)$	Interference component in the model $Y_i(z) = \alpha_i + \varphi_i z_i + s_i(z)$; it vanishes under constant treatment.
h_α, h_φ	Homophilous components of α and φ.
$\varepsilon_\alpha, \varepsilon_\varphi$	Heterogeneous-variation / residual components of α and φ.
Homophily, interference, robustness	
η	Homophily trade-off parameter.
γ	Interference trade-off parameter.
κ	Robustness trade-off parameter.
$\Sigma_f = \mathbb{E}[\varepsilon_f \varepsilon_f^{\top}]$	Covariance matrix of heterogeneous variation, for $f \in \{\alpha, \varphi\}$.
q^*	Schatten exponent used to constrain Σ_f.

Term / symbol	Meaning
q	Hölder-conjugate exponent to q^*, defined by $1/q^* + 1/q = 1$; it appears in the penalty $\|X\|_q$.
$\|\cdot\|_q$	Schatten-q norm of a matrix.
$\chi(\eta, \gamma, \kappa) = \chi_{q^*}(\eta, \gamma, \kappa)$	Model class of potential outcomes satisfying the homophily, interference, and heterogeneous-variation constraints.
$\widetilde{\gamma}$	Rescaling; γ/Δ in symmetric-case and $\gamma\beta_1^2$ in general
$\widetilde{\eta}$	Rescaling; $5\eta/(4\Delta)$ in symmetric-case and $\eta(\beta_1^2 + \beta_2^2)$ in general
$\widetilde{\kappa}$	Rescaling; $5\kappa/(4\Delta)$ in symmetric-case and $\kappa(\beta_1^2 + \beta_2^2)$ in general
Estimand, estimator, MSE quantities and design covariance	
p	Marginal treatment probability, $p = \mathbb{P}(z_i = 1)$, taken uniform across nodes in the main development.
a	Average baseline level, $a = \langle \alpha, \mathbf{1}/n \rangle$.
b	Average treatment-effect level, $b = \langle \varphi, \mathbf{1}/n \rangle$.
δ	Quantity controlling the average / centered interference contribution, via $\|\langle s(z), \mathbf{1}/n \rangle\| \leq \sqrt{\delta}$.
Δ	Combined scale parameter: in the symmetric case, $\Delta = (a + b/2)^2 + \delta$; in general, $\Delta = (a\beta_1 + b\beta_2)^2 + \beta_1^2\delta$
$X = \mathrm{Cov}(x)$	Design covariance matrix of the assignment.
X^*	Optimal achievable covariance matrix of a design.
X_{SDP}	SDP solution matrix used in the Gaussian-rounding algorithm.
β_1	Constant in the general MSE bound, $\beta_1 = 1/(2np(1-p))$.
β_2	Constant in the general MSE bound, $\beta_2 = 1/(2np)$.
Gaussian rounding and adapted Gram–Schmidt Walk notation	
ξ	Gaussian sample drawn as $\xi \sim \mathcal{N}(0, X^*)$.
ζ	Final rounded $\{\pm 1\}^n$ assignment returned by Gaussian rounding.

Term / symbol	Meaning
$\Xi = \mathrm{Cov}(\zeta)$	Covariance matrix of the rounded assignment.
λ	Randomization parameter in GSW balancing robustness against covariate balance.
ω_A	In the GSW analysis, $\omega_A = \max_i \|A_i\|$, i.e. the maximum column norm of A.
μ	Mean assignment vector, with entries $\mu_i = \mathbb{E}[x_i]$.

5.B More illustrations

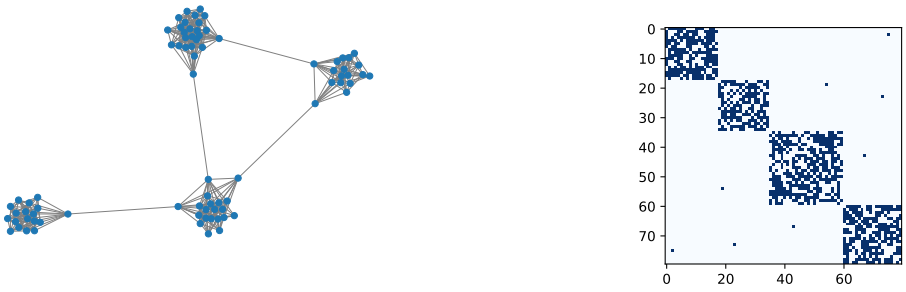


Figure 5.B.1: An SBM graph with four underlying clusters, and its adjacency matrix (with dark blue representing the presence of an edge, and light blue representing absence of an edge).

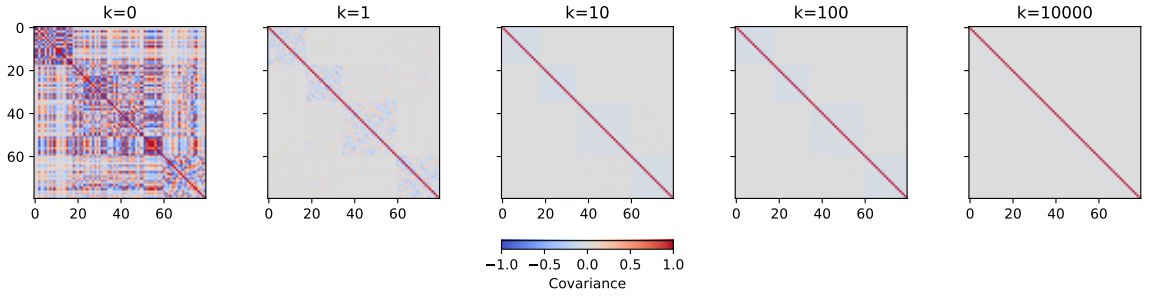


Figure 5.B.2: Solutions of the SDPs described in Section 5.5.1 on the graph in Figure 5.B.1 trading-off $10\,\mathrm{Tr}(L^\dagger X) + k\|X\|_F + 10\,\mathrm{Tr}(\mathbf{1}\mathbf{1}^T X)$. k ranges between $(0, 10^0, 10, 10^2, 10^4)$ from left to right.

In Figure 5.1.1, the treatment assignments are generated by Gaussian rounding. The panels vary the trade-off parameter ratio η/γ: Left: $\eta = 10, \gamma = 100, \kappa = 1$, $q = 2$; Center: $\eta = 1000, \gamma = 100, \kappa = 1$, $q = 2$; Right: $\eta = 10000, \gamma = 10, \kappa = 1$, $q = 2$.

Similarly, in Figure 5.4.1, the treatment assignments are generated by Gaussian rounding. The panels vary the trade-off parameter ratio κ/γ: Left: $\eta = 0, \gamma = 1000, \kappa = 1$, $q = 2$; Center: $\eta = 0, \gamma = 1, \kappa = 1$, $q = 2$; Right: $\eta = 0, \gamma = 1, \kappa = 1000$, $q = 2$.

In Figure 5.1.2, we display solutions of the SDPs described in Section 5.5.1, on the graph in Figure 5.B.1, trading-off $10\,\mathrm{Tr}(LX) + h\,\mathrm{Tr}(L^\dagger X) + 10\,\mathrm{Tr}(\mathbf{1}\mathbf{1}^T X)$. Here, h ranges between $(0, 10^0, 10, 10^2, 10^4)$ from left to right.

Similarly, in Figure 5.B.2, we display solutions of the SDPs described in Section 5.5.1, on the graph in Figure 5.B.1. This formulation explicitly trades off homophily and robustness by minimizing the objective $10\,\mathrm{Tr}(L^\dagger X) + k\|X\|_F + 10\,\mathrm{Tr}(\mathbf{1}\mathbf{1}^T X)$. Here, k ranges between $(0, 10^0, 10, 10^2, 10^4)$ from left to right.

In Figure 5.4.2, we display solutions of the SDPs described in Section 5.5.1, on the graph in Figure 5.B.1. This time, the formulation explicitly trades off interference and robustness by minimizing the objective $10\,\mathrm{Tr}(LX) + k\|X\|_F + 10\,\mathrm{Tr}(\mathbf{1}\mathbf{1}^T X)$. Here, k ranges between $(0, 10^0, 10, 10^2, 10^4)$ from left to right.

5.C Non-symmetric Case

We now consider the non-symmetric case, i.e., when $\mathbb{P}(z_i = 1) \neq \mathbb{P}(z_i = 0)$. We assume that the treatment probability p is uniform across nodes, i.e., $p = \mathbb{P}(z_i = 1) = \mathbb{P}(z_j = 1)$ for all $i, j \in [n]$.

As before, our goal is to relax this problem to the minimization of a quadratic form $x^T Qx$ plus a penalization term to encourage randomization, which can then be optimized via an SDP or the adapted Gram-Schmidt Walk vector balancing approach.

Theorem 5.C.1 (Non-symmetric MSE bound)**.** *Let $p := \mathbb{P}(z_i = 1)$ for all $i \in [n]$, $\beta_1 := 1/(2np(1-p))$, $\beta_2 := 1/(2np)$, $a = \langle \alpha, \mathbf{1}/n \rangle$, $b = \langle \phi, \mathbf{1}/n \rangle$, and $\langle s(z), \mathbf{1}/n \rangle \leq \sqrt{\delta}$ for all z. Let also $x \in \{-1, 1\}^n$ by defining $x := 2z - 1$. Define $X := \mathrm{Cov}(x)$. Then, for $q \geq 1$,*

$$\begin{aligned}\mathbb{E}[(\hat{\tau} - \tau)^2] \leq 7\{\mathrm{Tr}((\gamma\beta_1^2 L + \eta[\beta_1^2 + \beta_2^2]L^\dagger)X) + \kappa[\beta_1^2 + \beta_2^2]\|X\|_q \\ + ([a\beta_1 + b\beta_2]^2 + \beta_1^2\delta)\,\mathrm{Tr}(\mathbf{1}\mathbf{1}^T X)\}. \end{aligned} \tag{5.C.1}$$

We display the proof in Appendix 5.J.2.

Let $\Delta = (a\beta_1 + b\beta_2)^2 + \beta_1^2\delta$. Reparameterizing with $\tilde{\gamma}$,$\tilde{\eta}$, and $\tilde{\kappa}$ so that $\tilde{\gamma} := \frac{\gamma\beta_1^2}{\Delta}, \tilde{\eta} := \frac{\eta[\beta_1^2+\beta_2^2]}{\Delta}$, and $\tilde{\kappa} := \frac{\kappa[\beta_1^2+\beta_2^2]}{\Delta}$, we can write an SDP more explicitly in terms of the new trade-off parameters:

$$\begin{aligned} \min_{X \in \mathbb{R}^{n\times n}} \quad & \tilde{\eta}\mathrm{Tr}(L^\dagger X) + \tilde{\gamma}\mathrm{Tr}(LX) + \tilde{\kappa}\|X\|_q + \mathrm{Tr}(\mathbf{1}\mathbf{1}^T X) \\ \text{s.t.} \quad & X \succeq 0, \\ & \mathrm{diag}(X) = \mathbf{1}. \end{aligned} \tag{5.C.2}$$

Figure 5.E.2 presents the covariance matrices of the design across different treatment assignment probabilities p, generated from applying Algorithm 2 to the corresponding SDP solutions in Figure 5.E.1.

In the adapted Gram-Schmidt Walk, we can take $A := (\tilde{\eta}L^\dagger + \tilde{\gamma}L + \mathbf{1}\mathbf{1}^T)^{1/2}$. Then, we can take the robustness parameter to be $\lambda = \frac{\sqrt{\tilde{\kappa}n^{1/q}}}{\sqrt{\omega_A^2 n + \sqrt{\tilde{\kappa}n^{1/q}}}}$, where $\omega_A := \max_{i\in n}\|A_i\|$. Alternatively, one could append the vectors so $A_i^T = [\tilde{\eta}^{1/2}L_i^{\dagger/2} \quad \tilde{\gamma}^{1/2}L_i^{1/2} \quad \mathbf{1}\mathbf{1}^T{}_i^{1/2}]$.

5.D More examples

5.D.1 Homophily

Figure 5.D.1 depicts a concrete real example of homophily.

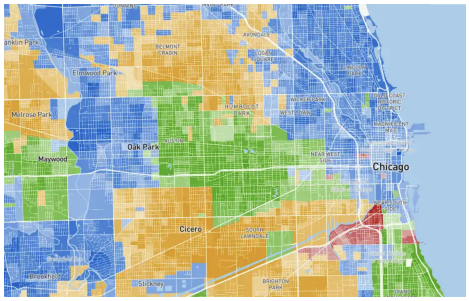


Figure 5.D.1: A real example of homophily. The map from Best Neighborhoods [n.d.] depicts majority race by geographic area in the Chicago region, based on self-identification on the US census. Darker shades indicate a larger racial majority. This spatial map can be modeled as a network, where nodes represent geographic areas and edge weights are inversely proportional to the distance between them. This network exhibits a high level of homophily, as racial composition varies smoothly across neighboring areas in the graph.

Example 5.D.1. One natural interpretation of homophily is via the latent space framework [Athreya et al., 2018, Hoff et al., 2002], where we identify a homophilous vector h with a smooth function of common latent variables that determine the graph weights W_{ij}. As a concrete example, each node may be associated to a latent vector u_i, and for $U := [u_1, \ldots, u_n]^T$, there is some $\beta \in \mathbb{R}^d$ such that $h = U\beta$, and $W_{ij} \propto \exp(-\|u_i - u_j\|^2)$.

Remark 5.D.2. The quadratic form $g^T Lg = \langle g, Lg \rangle_{\ell_2}$ directly quantifies the smoothness of the function g with respect to the graph structure. Its continuous analogue, replacing the discrete graph with a manifold or a domain equipped with a measure μ, is the quadratic form $\langle g, \Delta_\mu g \rangle_{L^2(\mu)} = \int g \Delta_\mu g \, d\mu = \int \|\nabla g\|^2 \, d\mu$, where Δ_μ is the Laplacian operator with respect to μ, and ∇g is the gradient of a sufficiently smooth function g supported on μ.

Remark 5.D.3. To define homophily with respect to similarity in node degrees d_i, a natural approach is to use the symmetric normalized Laplacian $L_{\text{sym}} := I - D^{-1/2} W D^{-1/2}$ and say that the vector h is (η, L_{sym})-homophilous if $h^T L_{\text{sym}} h \leq \eta$. This formulation incorporates an appropriate normalization of the edge weights by the square roots of the degrees of the corresponding nodes. Explicitly, $h^T L_{\text{sym}} h = \sum_{(i,j) \in E} W_{ij} \left(\frac{h_i}{\sqrt{d_i}} - \frac{h_j}{\sqrt{d_j}} \right)^2 \leq \eta$ captures the smoothness of h when degree similarity is taken into account.

Remark 5.D.4. Our notion of homophily is consistent with the classical view that homophily is associated with cluster formation in networks [McPherson et al., 2001, Currarini et al., 2009, Jackson

et al., 2023, Jackson, 2025]. In our setting, this is captured by the Rayleigh quotient of the graph Laplacian $h^T Lh/\|h\|^2$, the ratio η'/κ' of the homophily and heterogeneity of a centered vector h. When this ratio is small, the potential outcomes h have little discrepancy between neighbors in the graph compared to the variation of the vector. Equivalently, the variations in h are primarily inter-cluster rather than intra-cluster, and this can be characterized spectrally as h lying predominantly in the span of the eigenvectors associated with the smallest eigenvalues of the graph Laplacian. This is most insightful in our decomposition of α and ϕ into homophilous and heterogeneous components h_f and ε_f, where $f \in \{\alpha, \phi\}$, respectively. As will become apparent in bounds such as Theorem 5.4.2, this decomposition will be particularly effective numerically when h_f is selected so that the corresponding ratio η'/κ' is minimal, and vice versa for ε_f, $f \in \{\alpha, \phi\}$, thus aligning these two notions of homophily. In other words, if the outcomes associated with all the units in the population are highly heterogeneous and not driven by the underlying network structure, then a larger share of the variation in the outcomes can be attributed to ε (and hence κ) rather than h (and hence η).

5.D.2 Interference

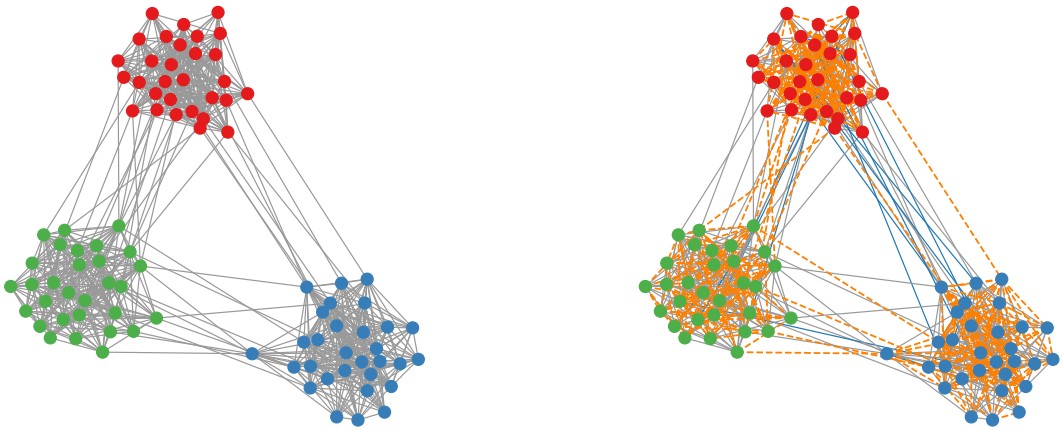


Figure 5.D.2: Example of noisy network observation. Left panel: True network, G. Right panel: Noisy observation G_{obs} of the network with grey edges preserved and newly added dark blue edges. The orange edges depict missing edges.

Example 5.D.5 (Neighborhood Interference)**.** A simple outcome model with a linear neighborhood interference term with a $(\gamma, \tilde{L})$-interferent s component paired with a $\sqrt{n}$-scaling reminiscent of the local asymptotic framework [Viviano et al., 2023, Hirano and Porter, 2009] is $Y_i = \alpha_i + \beta_i Z_i + \frac{\gamma}{\sqrt{n}} \sum_j \frac{W_{ij}}{\sum_j W_{ij}} z_j$. Indeed, the equation above can be rewritten as $Y_i = \alpha_i + \phi_i z_i - \frac{\gamma}{\sqrt{n}} e_i^T (I - D^{-1}W) z$, where $\phi_i := \beta_i + \gamma/\sqrt{n}$, $D = \text{diag}(d)$ for the vector $d \in \mathbb{R}^n$ of degrees of the nodes $d_i = \sum_{j \in N(i)} W_{ij}$, $W \in \mathbb{R}^{n \times n}$ is the matrix of the edge weights $\{W_{ij}\}_{i,j}$, and α, ϕ are vectors with entries α_i, ϕ_i, respectively, for $i = 1, 2, ..., n$. Let the normalized graph Laplacian $\tilde{L} := I - D^{-1}W$, i.e., $\tilde{L} = D^{-1}L$, and $\tilde{L}^\dagger$ to be the pseudoinverse of the normalized Laplacian. In matrix-vector form, $Y = \alpha + \text{Diag}(\phi) z - \frac{\gamma}{\sqrt{n}} \tilde{L} z$. Here, $s(z) = -\frac{\gamma}{\sqrt{n}} \tilde{L} z$, i.e., s is $(\gamma, \tilde{L})$-interferent. The normalized Laplacian term appropriately vanishes when the full population is treated or assigned to control.

Example 5.D.6 (Full Network Interference)**.** Take for example an instance of the weak-dependence model [Leung, 2022a] adapting α-mixing spatial processes (see, for example, Jenish and Prucha [2009]) to path lengths ρ_n in the network, where $\sum_{m=2}^n M_m^\partial \tilde{\theta}_m = o(\psi_1)$, for some $\psi_1 \ll 1$, $M_m^\partial := \frac{1}{n} \sum_{i=1}^n |\mathcal{N}^\partial(i, m)|$, $\mathcal{N}^\partial(i, m) := \{j \in [n] : \rho_n(i, j) = m\}$, and correlation between outcomes $\tilde{\theta}_m \to 0$, as $m \to \infty$. That is, the correlation $\tilde{\theta}$ between outcomes of units decay over (path length) distances ρ_n. Here $\tilde{L}$ can be taken to be the graph Laplacian of a complete graph, modeling a full interference structure, with edge weights $\tilde{\theta}$ corresponding to the correlation under path lengths in the original path length weak-dependence structure (see, for example, Figure 5.D.3). Under the setting with such a $\tilde{\theta}$ decaying quickly, we have that the full degree matrix $\tilde{D} = D + \text{diag}(\delta)$, and full adjacency matrix $\tilde{W} = W + \epsilon$, where $\delta_i = \sum_{j \in [n]} \epsilon_{ij}$, for all $i = 1, 2, \ldots, n$. Then, writing $L_{\text{res}} := \text{diag}(\delta) - \epsilon$, $\tilde{L} = L + L_{\text{res}}$. The spectrum of L_{res} are upper-bounded by some $\psi_2 \ll 1$, such that $\|\tilde{L}^\dagger L\|_{\text{op}} \leq 1 + O(\psi)$, for $\psi \ll 1$. Then, our worst-case MSE bound would consist of the perturbed bound, $\gamma(1 + O(\psi))$ instead of just γ.

Example 5.D.7. We note the distinction between our model of interference and a generalized version of the previous example when the interference term $s(z)$ is $\frac{\boldsymbol{\gamma}}{\sqrt{n}} \tilde{L} z$, where $\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j$ are not necessarily equal for any (i, j) pair. A somewhat similar interference structure encapsulated by our model is one where $s(z) = L^{1/2} \text{Diag}(\boldsymbol{\gamma}/\sqrt{2 d_{\max} n}) L^{1/2} z$, where $d_{\max}$ is the maximum degree of the nodes in the graph. Then, if $\|\boldsymbol{\gamma}\|_\infty^2 \leq \gamma$, indeed we have

$$s(z)^T L^\dagger s(z) = z^T L^{1/2} \text{Diag}(\boldsymbol{\gamma}/\sqrt{n}) L^{1/2} L^\dagger L^{1/2} \text{Diag}(\boldsymbol{\gamma}/\sqrt{n}) L^{1/2} z$$

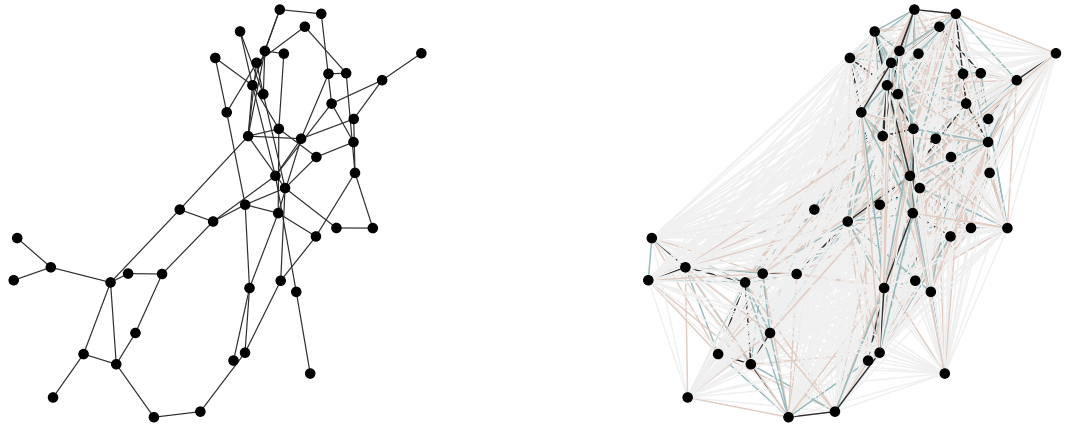


Figure 5.D.3: Full interference example. Left panel: First-order (neighborhood) interference, where outcomes depend only on treatments of immediate neighbors. Right panel: Full network interference induced by the same underlying graph, allowing interference to propagate along paths of increasing length, with strength decaying in path length. Darker edges indicate stronger (shorter-path) interference effects.

$$
\begin{aligned}
&= z^T L^{1/2} \operatorname{Diag}(\boldsymbol{\gamma}/\sqrt{2d_{\max}n})\mathcal{P} \operatorname{Diag}(\boldsymbol{\gamma}/\sqrt{2d_{\max}n})L^{1/2} z \\
&\leq \frac{z^T L z}{2d_{\max}n} \|\boldsymbol{\gamma}\|_\infty^2 \leq \gamma,
\end{aligned}
$$

where $\mathcal{P}$ is the orthogonal projection matrix onto the column space of L. Thus, s is (γ, L)-interferent.

5.E More on the experimental designs

5.E.1 Gaussian Rounding on the SDP solutions

5.E.1.1 Randomized hyperplane rounding

Algorithm 3 converts the SDP solution X_{SDP} into a treatment assignment by first sampling a Gaussian vector with covariance X_{SDP}. The entries of this vector are then rounded coordinatewise to obtain a binary assignment vector. In the symmetric case $p = 1/2$, this reduces to sign rounding; more generally, the rounding step is calibrated so that each unit has marginal treatment probability p. Thus, Gaussian rounding provides a practical way to pass from the continuous SDP relaxation to an implementable randomized binary design while retaining much of the covariance structure selected

Algorithm 3 GaussianRounding with treatment probability $p \leq 1/2$

Require: SDP solution $X_{\text{SDP}} \in \mathcal{E}_n$, treatment probability p, rounding rule $\mathsf{R} \in \{\mathsf{HR}, \mathsf{TR}\}$

Sample $\xi \sim \mathcal{N}(0, X_{\text{SDP}})$.
if $\mathsf{R} = \mathsf{HR}$ **then** ▷ Randomized hyperplane rounding
  for $i = 1, \ldots, n$ **do**
    $x'_i \leftarrow \mathbf{1}\{\xi_i \geq 0\}$
    if $x'_i = 1$ **then**
      Draw $\zeta_i \sim \text{Rademacher}(2p)$ ▷ $\mathbb{P}(\zeta_i = 1 \mid x'_i = 1) = 2p$
    else
      $\zeta_i \leftarrow -1$
    end if
  end for
else if $\mathsf{R} = \mathsf{TR}$ **then** ▷ Thresholding
  $t_p \leftarrow \Phi^{-1}(1 - p)$
  for $i = 1, \ldots, n$ **do**
    $\zeta_i \leftarrow \mathbf{2}1\{\xi_i \geq t_p\} - 1$
  end for
end if
return ζ

by the SDP.

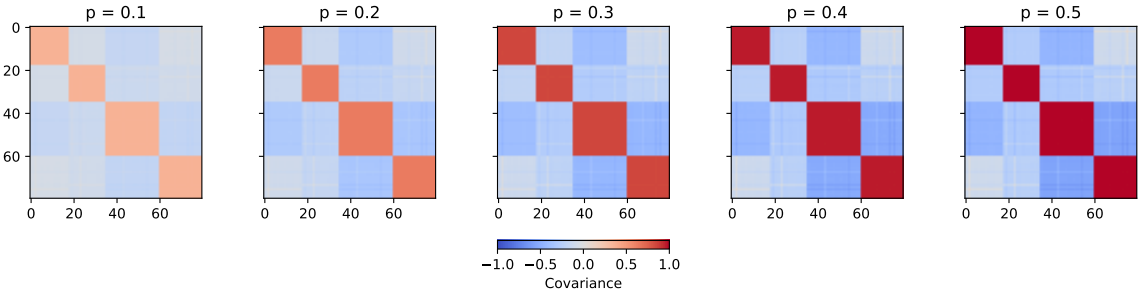


Figure 5.E.1: SDP solutions on the graph from Figure 5.B.1, with trade-off parameters $\gamma = 10, \eta = 1, \kappa = 1$, and $a = b = 1$, across varying treatment assignment probabilities p. Here p ranges between $(0.1, 0.2, 0.3, 0.4, 0.5)$ from left to right. The probabilities p scale each trade-off component of the SDP via β_1 and β_2, and thus the SDP solutions exhibit only slight variation across values of p.

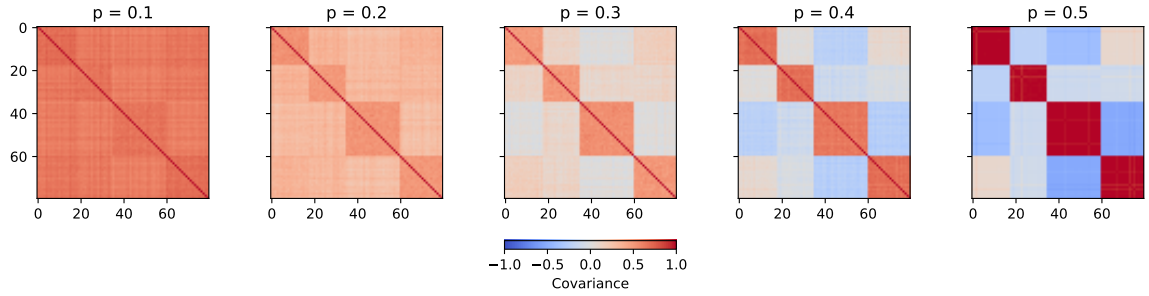


Figure 5.E.2: Covariance matrices of our designs by applying Algorithm 2 to the SDP solutions in Figure 5.E.1, across treatment assignment probabilities p ranging between $(0.1, 0.2, 0.3, 0.4, 0.5)$ from left to right.

5.E.1.2 Thresholding

One can also consider what we call "Thresholding", which is equivalent to Gaussian rounding when $p = 1/2$. Generally, thresholding for $p \leq 1/2$ is simply the procedure of selecting the threshold $t_p(i)$ for the rounding of $\xi \sim \mathcal{N}(0, X_{\text{SDP}})$, $\xi \in \mathbb{R}^n$ to $\zeta \in \{-1, 1\}^n$ based on the quantiles of ξ_i. That is, we take $t_p(i) = \sqrt{X_{\text{SDP}ii}} \cdot \Phi^{-1}(1-p) = \Phi^{-1}(1-p)$, and let $\zeta_i = 1$ if $\xi_i \geq t_p(i)$, and $\zeta_i = -1$ otherwise.

5.E.2 Adapted Gram-Schmidt Walk algorithm

We include the Gram-Schmidt Walk algorithm adapted to our trade-off problem in Algorithm 4. Algorithm 4 applies the Gram-Schmidt Walk to the augmented covariate matrix B, constructed from a factor A satisfying

$$A^\top A = \widetilde{\eta} L^\dagger + \widetilde{\gamma} L + \mathbf{1}\mathbf{1}^\top,$$

together with scaled standard basis vectors used to incorporate the robustness term. The algorithm is initialized at $z = 2p - 1$, ensuring the target marginal treatment probabilities p are satisfied. At each iteration, it computes an update direction that seeks to preserve balance in the span of the active columns of B, and then moves along this direction until one or more active coordinates hit ± 1. These coordinates are subsequently fixed, and the procedure repeats on the remaining active coordinates. When the algorithm terminates, all entries lie in $\{\pm 1\}$, yielding a treatment assignment that balances the induced covariates while preserving the desired level of randomization.

Algorithm 4 (Adapted) Gram-Schmidt Walk (adaptation of Harshaw et al. [2024]; see also Section 5.5.2)

Require: $q, \eta, \gamma, \kappa, \mathbf{p}$

Define A_i as the i-th column of $\Lambda^{1/2}V^\top$, where $V\Lambda V^\top = \tilde{\eta}L^\dagger + \tilde{\gamma}L + \mathbf{1}\mathbf{1}^\top$
(Alternatively, use the appended matrix A; see discussion above)

Define $\omega_A \leftarrow \max_{i\in[n]} \|A_i\|$

Set $\lambda \leftarrow \frac{\kappa n^{1/q}}{\eta\omega_A^2 n + \kappa n^{1/q}}$

Define $\left[B\right]_i \leftarrow \vec{b}_i \leftarrow \begin{bmatrix} \sqrt{\lambda}\, e_i \\ \omega_A^{-1}\sqrt{1-\lambda}\, A_i \end{bmatrix}$

Initialize iteration counter $t \leftarrow 1$

Initialize assignment vector $\vec{z}_1 \leftarrow 2\mathbf{p} - 1$

Select pivot index $i_{\text{piv}} \sim \text{Uniform}([n])$

while $\vec{z}_t \notin \{\pm 1\}^n$ **do**

  Define active set $\mathcal{A} \leftarrow \{i \in [n] : |\vec{z}_t(i)| < 1\}$

  if $i_{\text{piv}} \notin \mathcal{A}$ **then**

    Resample pivot $i_{\text{piv}} \sim \text{Uniform}(\mathcal{A})$

  end if

  Compute step direction:

$$\vec{u}_t \leftarrow \underset{\vec{u}\in\mathbb{R}^n}{\text{argmin}} \| \left[B\right] \vec{u}\|^2 \quad \text{s.t.} \quad u(i_{\text{piv}}) = 1,\ u(i) = 0\ \forall i \notin \mathcal{A}$$

  Define feasible step sizes:

$$\Delta \leftarrow \{\delta \in \mathbb{R} : \vec{z}_t + \delta\vec{u}_t \in [-1, 1]^n\}$$

  Set $\delta^+ \leftarrow |\max \Delta|$, $\delta^- \leftarrow |\min \Delta|$

  Randomly sample step size:

$$\delta_t \leftarrow \begin{cases} \delta^+ & \text{with probability } \frac{\delta^-}{\delta^+ + \delta^-} \\ -\delta^- & \text{with probability } \frac{\delta^+}{\delta^+ + \delta^-} \end{cases}$$

  Update assignment: $\vec{z}_{t+1} \leftarrow \vec{z}_t + \delta_t\vec{u}_t$

  Increment $t \leftarrow t + 1$

end while

return Final assignment vector $\vec{z}_t \in \{\pm 1\}^n$

5.F Generalizations and a special case

5.F.1 General Similarity Kernels and Stochastic Misspecification

We use the term "homophily" in a broad sense. Although it technically refers to similarity in social ties, we adopt it here to encompass similarities that may arise in contexts beyond social networks. If the practitioner, for instance, were to include some other likeness measure such as spatial positioning or geography, then a different similarity kernel, rather than or in addition to the Laplacian L, can be used. For example, under a similarity kernel $\mathcal{K}$, we can write our homophily formulation as $g^T\mathcal{K}g \leq \eta$. That is, g is $(\eta, \mathcal{K})$-homophilous.

This leads to a more general potential outcomes model,

$$Y = \sum_{i=1}^{N_\alpha} h_\alpha^i + \sum_{i=1}^{N'_\alpha} \varepsilon_\alpha^i + z \cdot \left(\sum_{i=1}^{N_\phi} \mathrm{diag}(h_\phi^i) + \sum_{i=1}^{N'_\phi} \mathrm{diag}(\varepsilon_\phi^i) \right) \tag{5.F.1}$$

where h_f^i, $f \in \{\alpha, \phi\}$, satisfy quadratic constraints ${h_f^i}^T \mathcal{K}_f^i h_f^i \leq \eta_f^i$, $\mathcal{K}_f^i$ a kernel/psd matrix, and the misspecifications ε_f^i, $f \in \{\alpha, \phi\}$, satisfy Schatten-norm constraints $\|\Sigma_f^i\|_{q_f^{i\star}} := \|\mathbb{E}[\varepsilon_f^i {\varepsilon_f^i}^T\|_{q_f^{i\star}} \leq \kappa_f^i$. Defining $\boldsymbol{\eta} := [\eta_\alpha, \eta_\phi]$, $\boldsymbol{\kappa} := [\kappa_\alpha, \kappa_\phi]$ for $\eta_f := (\eta_f^i)_{i=1}^{N_f}$, $\kappa_f := (\kappa_f^i)_{i=1}^{N'_f}$, where $f \in \{\alpha, \phi\}$, and $N := N_\alpha + N_\phi + N'_\alpha + N'_\phi$, this is summarized by the constraint set

$$\chi(\boldsymbol{\eta}, \boldsymbol{\kappa}) := \{h_f^i, \varepsilon_f^i : \ {h_f^i}^T \mathcal{K}_f^i h_f^i \leq \eta_f^i, \ \|\Sigma_f^i\|_{q_f^{i\star}} \leq \kappa_f^i\}.$$

Our framework extends to the following bound.

Theorem 5.F.1. *(General MSE) Let $p := \mathbb{P}(z_i = 1)$ for all $i \in [n]$, $a = \langle \alpha, \mathbf{1}/n \rangle$, $b = \langle \phi, \mathbf{1}/n \rangle$, and $\beta_1 := 1/(2np(1-p))$, $\beta_2 := 1/(2np)$. Let also $x \in \{-1, 1\}^n$ by defining $x := 2z - 1$, $X := \mathrm{Cov}(x)$. Then, for $q_f^i \geq 1$, $i \in [N'_f]$,*

$$\sup_{\chi(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\kappa})} \mathbb{E}[(\hat{\tau} - \tau)^2] \leq (N+1)\{(a\beta_1 + b\beta_2)^2 \,\mathrm{Tr}(\mathbf{1}\mathbf{1}^T X) + \sum_{f \in \{\alpha, \phi\}} \beta_f^2 \sum_{i=1}^{N_f} \eta_f^i \,\mathrm{Tr}((\mathcal{K}_f^i)^\dagger X) + \sum_{i=1}^{N'_f} \kappa_f^i \|X\|_{q_f^i}\}.$$

Before exploring the utility of this framework, let's first see how this corresponds to the basic structure established previously in the paper.

Example 5.F.2. If we set $\mathcal{K}_\alpha^1 = \mathcal{K}_\phi^1 = L^\dagger$, $\mathcal{K}_\alpha^2 = L$, $\eta_\alpha = [\eta, \gamma]$, $\eta_\phi = \eta$, $q_\alpha = q_\phi = q$, and $\kappa_\alpha = \kappa_\phi = \kappa$, then we arrive at the constraint set $\chi(\eta, \gamma, \kappa)$. Indeed, this gives us a potential outcome model

$$Y = h_\alpha^1 + h_\alpha^2 + \varepsilon_\alpha^1 + z \cdot \left(h_\phi^1 + \varepsilon_\phi^1\right) = (h_\alpha^1 + \varepsilon_\alpha^1) + z \cdot \left(h_\phi^1 + \varepsilon_\phi^1\right) + h_\alpha^2 =: \alpha + z \cdot \phi + s(z).$$

We note that $s(z) = h_\alpha^2$ may or may not explicitly depend on the assignment z in this framework, as we only specify that it satisfies a quadratic constraint. Adversarially selecting this term to maximize the MSE bound results in a typical graph interference setting.

Our extended framework provides a convenient means to incorporate information from other sources besides the graph and correlation motivated objectives discussed earlier.

Example 5.F.3. Of key importance in practice is the incorporation of covariates into the potential outcome model beyond the graph adjacency. In its most basic form, this is a covariate balancing problem, where one would like to penalize the squared discrepancy $x^T C C^T x$, for C the design matrix/matrix of additional covariates. In our framework, this corresponds to a potential outcome component satisfying a quadratic constraint with respect to $(CC^T)^{-1}$, with the typical example being a linear model $C\beta$. More general potential outcomes can be captured by kernel gram matrices $\mathcal{K} := [k(c_i, c_j)]_{i,j=1}^n$, c_i, c_j the covariates of units i and j. This provides particular utility for designs on disconnected graphs, which we discuss in more detail in Appendix 5.F.2.

Another point of theoretical importance is the disaggregation of the heterogeneous variation components in the model.

Example 5.F.4. For the majority of the paper, we streamlined our presentation by presenting different heterogeneous variation, such as deterministic residuals and stochastic network effects, as mutually exclusive modeling choices made by the practitioner. This is of course unnecessary, as our general MSE demonstrates, as both of these terms can be separately included. This is particularly important for large graphs, as the Schatten-q, $q < \infty$, norm of X may grow rapidly with the sample size. Thus, it may be desirable to model these components separately: deterministic residuals with bounded average squared magnitude, $\|\varepsilon_n^{\text{det}}\|_2^2 = O(n)$, whose worst-case contribution is governed by the operator norm of X, and stochastic network misspecification with bounded covariance operator norm, $\|\Sigma_n^{\text{stoch}}\|_{\text{op}} = O(1)$, whose worst-case contribution is proportional to $Tr(X)$. In other words,

as the graph grows, we expect the homophily and interference terms to capture the majority of the correlation structure.

5.F.2 Disconnected Components in the Graph

We have assumed, so far, that the graph is connected. In the following, we consider the worst-case bounds if the graph had T connected components instead.

Let Π be the projection to the null-space of the Laplacian matrix.

Theorem 5.F.5 (Disconnected MSE bound)**.** *Let $p := \mathbb{P}(z_i = 1)$ for all $i \in [n]$, $\beta_1 := 1/(2np(1-p))$, $\beta_2 := 1/(2np)$. Let also $x \in \{-1, 1\}^n$ by defining $x := 2z - 1$, and $X := \mathrm{Cov}(X)$. Then,*

$$\begin{aligned}\mathbb{E}[(\hat{\tau} - \tau)^2] \leq 6\{ &\operatorname{Tr}((\gamma\beta_1^2 L + \eta[\beta_1^2 + \beta_2^2]L^\dagger)X) \\ &+ \kappa[\beta_1^2 + \beta_2^2]\|X\|_q + \|\Pi(\beta_1\alpha + \beta_2\phi)\|^2 \operatorname{Tr}(\Pi X)\}.\end{aligned}$$

Proof. We decompose $\alpha = \Pi\alpha + (I - \Pi)\alpha,\ \phi = \Pi\phi + (I - \Pi)\phi$. For the latter terms orthogonal to the kernel of L, the same analysis as applied to the mean 0 residuals in Theorem 5.C.1 is applicable, hence it suffices to bound the projected terms. Following the proof of Theorem 5.C.1, we compute

$$\begin{aligned}&\mathbb{E}[(\langle \Pi\alpha, z/(np) - (1-z)/(n(1-p))\rangle + \langle \Pi\phi, z/(np) - \mathbf{1}/n\rangle)^2] \\ &= \operatorname{Var}(\langle \Pi\alpha, z/(np) - (1-z)/(n(1-p))\rangle + \langle \Pi\phi, z/(np) - \mathbf{1}/n\rangle) \\ &= \operatorname{Var}(\langle \Pi\alpha, \beta_1(x-\mu)\rangle + \langle \Pi\phi, \beta_2(x-\mu)\rangle) \\ &= \operatorname{Tr}(\Pi(\beta_1\alpha + \beta_2\phi)(\beta_1\alpha + \beta_2\phi)^T \Pi X) \\ &\leq \|\Pi(\beta_1\alpha + \beta_2\phi)\|^2 \operatorname{Tr}(\Pi X).\end{aligned}$$

■

5.F.3 Quadratic Penalization of X

So far in this paper, we have restricted our MSE optimization to first order terms linear in the design covariance matrix X, either via a trace inner-product $\mathrm{Tr}(AX)$ or norm penalties. However, there are many natural examples where higher-order dependence is a more appropriate model.

Example 5.F.6. A simple and practical model for the interference is $s(z) = Lz$. We can easily compute

$$(Lz)^T L^\dagger (Lz) = z^T Lz,$$

thus Lz is $z^T Lz$-interferent. In the MSE, this results in an objective $\propto (x^T Lx)^2$ for the Rademacher assignment variables x. This quadratic dependence on the cut is exactly because both the level of interference and the extent to which interference impedes the quality of our estimator are both linear $x^T Lx$.

Thus, assuming $x^T Lx$ interference requires optimization of $\mathbb{E}[(x^T Lx)^2]$, and this may similarly be of interest for other kernels $\mathcal{K}$. One strategy to deal with this is to uniformly upper-bound $x^T Lx$, reducing this to a linear optimization. However, this is quite loose, and fails to capture the desired behavior of the assumed outcome model.

Within the framework of Gaussian Rounding, Algorithm 2, we relax the Rademacher assignment vectors x to a multivariate Gaussian distribution. This allows us to apply Isserlis's theorem (also commonly known as Wick's probability theorem) [Isserlis, 1918], to yield

$$\mathbb{E}[(x^T Lx)^2] = \operatorname{Tr}(LX)^2 + 2\operatorname{Tr}([LX]^2) = \operatorname{Tr}(LX)^2 + 2\|L^{1/2} X L^{1/2}\|_F^2.$$

This is convex in X, and thus amenable to our optimization routine.

Example 5.F.7. Quadratic dependencies also naturally appear when considering alternative models for the interference. As a simple example, we may consider a typical setting where $s_i(z)$ is linear in the exposure fraction of treated neighbors adjacent to the vertex i (see, for example, Thiyageswaran et al. [2024]; see also Example 5.3.3). In other words, $s(\mathbf{1}) - s(z) = LD{-}1W[\mathbf{1} - z] = L/2W[x] + L/2\mathbf{1}$ where D is the degree matrix and W is the graph adjacency. When considering the treatment effect, we may decompose

$$\begin{aligned}\langle \mathbf{1}/n, s(\mathbf{1})\rangle - \langle z/np - (\mathbf{1}-z)/np, s(z)\rangle = \\ \langle \mathbf{1}/n - [z/np - (\mathbf{1}-z)/np], s(\mathbf{1})\rangle - \langle z/np - (\mathbf{1}-z)/np, s(z) - s(\mathbf{1})\rangle,\end{aligned}$$

so that the first term can be bounded by our typical error analysis, with some moment restrictions

placed on the vector $s(\mathbf{1})$. For the remaining term, its contribution can be bounded by

$$\langle z/np - (\mathbf{1}-z)/np, s(z) - s(\mathbf{1})\rangle = L/2\beta_1 x^T D^{-1} W x + L/2 x^T \mathbf{1} + L/2 c_1 n.$$

In other words, in general settings, we can evaluate the quantity

$$\begin{aligned}
&\langle z/np - (\mathbf{1}-z)/np, s(z) - s(\mathbf{1})\rangle \\
&= \langle \beta_1 x + c_1 \mathbf{1}, s(z) - s(\mathbf{1})\rangle \\
&= 2\beta_1 \langle z - \mathbf{1}, s(z) - s(\mathbf{1})\rangle + (c_1 - \beta_1)\langle \mathbf{1}, s(z) - s(\mathbf{1})\rangle,
\end{aligned}$$

and similar control of the MSE is achieved if we assume a quadratic domination $|\langle z-\mathbf{1}, s(z)-s(\mathbf{1})\rangle| \leq (z-\mathbf{1})^T A (z-\mathbf{1})$, as well as a linear bound $\langle \mathbf{1}, s(z) - s(\mathbf{1})\rangle \leq |v^T(\mathbf{1}-z)|$, for some matrix A and vector v. A similar analysis reduces the baseline $s(\mathbf{0})$ into suitable terms for our SDP, falling within the framework of Theorem 5.F.1.

5.G An alternative approach to the interference term

Instead of using the bound $\|\langle s(z), \mathbf{1}\rangle\| \leq \sqrt{\delta}$, one could also treat the interference term in the following manner. Suppose that s is $(\gamma, L + \delta \mathbf{1}\mathbf{1}^T)$-interferent, where $\delta > 0$. Then,

$$\begin{aligned}
&\sup_{\langle s(z), (L+\delta \mathbf{1}\mathbf{1}^T)^\dagger s(z)\rangle \leq \gamma} \langle x, s(z)\rangle \\
&= \sup_{\langle s(z), (L+\delta \mathbf{1}\mathbf{1}^T)^\dagger s(z)\rangle \leq \gamma} \left\langle x, (L+\delta \mathbf{1}\mathbf{1}^T)^{1/2} (L+\delta \mathbf{1}\mathbf{1}^T)^{\dagger/2} s(z)\right\rangle \\
&= \sup_{\langle s(z), (L+\delta \mathbf{1}\mathbf{1}^T)^\dagger s(z)\rangle \leq \gamma} \left\langle (L+\delta \mathbf{1}\mathbf{1}^T)^{1/2} x, (L+\delta \mathbf{1}\mathbf{1}^T)^{\dagger/2} s(z)\right\rangle.
\end{aligned}$$

Therefore, by Cauchy-Schwarz,

$$\sup_{\langle s(z), (L+\delta \mathbf{1}\mathbf{1}^T)^\dagger s(z)\rangle \leq \gamma} \mathbb{E}[\langle x, s(z)\rangle^2] = \gamma \operatorname{Tr}((L+\delta \mathbf{1}\mathbf{1}^T) X),$$

and the results follow.

5.H A practical consideration on trade-off parameters

Our framework allows practitioners to use historical experiments, pilot studies, pre-treatment measurements, and domain knowledge to calibrate the trade-off parameters η, γ, κ, and δ. These quantities define the model class used in our worst-case MSE bounds. In practice, a convenient way to calibrate them is to fit a working potential-outcome model and then evaluate the same quadratic and norm quantities that appear in the definitions of the uncertainty class.

A working model. Suppose that R historical experiments have been conducted on the same population and network. Let $z^{(r)} \in \{0,1\}^n$ and $Y^{(r)} \in \mathbb{R}^n$ denote the treatment assignment and observed outcome vector in experiment r, for $r = 1, \ldots, R$. Let $V \in \mathbb{R}^{n \times d}$ denote a matrix of pre-treatment covariates. Define the random-walk Laplacian

$$\widetilde{L} := I - D^{-1}W.$$

As a simple working interference model, take

$$s(z) = \rho \widetilde{L} z.$$

Since $\widetilde{L}\mathbf{1} = 0$, this specification satisfies

$$s(\mathbf{0}) = s(\mathbf{1}) = 0.$$

We fit the working model

$$Y^{(r)} = \alpha + \operatorname{diag}\Big(z^{(r)}\Big)\,\tau + \rho \widetilde{L} z^{(r)} + e^{(r)}, \qquad r = 1, \ldots, R,$$

where $\alpha \in \mathbb{R}^n$ and $\tau \in \mathbb{R}^n$ denote the unit-specific baseline and treatment-effect vectors. For example, these parameters may be estimated by least squares:

$$(\widehat{\alpha}, \widehat{\tau}, \widehat{\rho}) \in \operatorname*{arg\,min}_{\alpha,\tau,\rho} \sum_{r=1}^{R} \Big\| Y^{(r)} - \alpha - \operatorname{diag}\Big(z^{(r)}\Big)\,\tau - \rho \widetilde{L} z^{(r)} \Big\|_2^2 .$$

This estimation requires sufficient variation across historical treatment assignments to distinguish

direct treatment effects from neighborhood exposure effects. In particular, historical experiments with different treatment saturation levels can provide useful variation for estimating ρ. If the number of historical experiments is small, additional parametric structure or regularization can instead be imposed on α and τ.

Decomposing homophilous and heterogeneous variation. We next use the observed covariates to decompose the estimated baseline and treatment-effect vectors into graph-structured and residual components. Let

$$M := I - \frac{1}{n}\mathbf{1}\mathbf{1}^T, \qquad V_c := MV,$$

and define

$$P_V := V_c \left(V_c^T V_c\right)^\dagger V_c^T,$$

the orthogonal projection onto the column space of the centered covariates. Let

$$\widehat{a} := \frac{1}{n}\mathbf{1}^T\widehat{\alpha}, \qquad \widehat{b} := \frac{1}{n}\mathbf{1}^T\widehat{\tau}.$$

We define the fitted homophilous components by

$$\widehat{h}_\alpha := P_V M\widehat{\alpha}, \qquad \widehat{h}_\tau := P_V M\widehat{\tau},$$

and the remaining heterogeneous components by

$$\widehat{\varepsilon}_\alpha := M\widehat{\alpha} - \widehat{h}_\alpha, \qquad \widehat{\varepsilon}_\tau := M\widehat{\tau} - \widehat{h}_\tau.$$

Calibrating η. A natural plug-in calibration of the homophily parameter is

$$\widehat{\eta} := \max\left\{\widehat{h}_\alpha^T L\widehat{h}_\alpha, \widehat{h}_\tau^T L\widehat{h}_\tau\right\}.$$

Thus, $\widehat{\eta}$ is the largest graph-smoothness quadratic form among the fitted baseline and treatment-effect components.

Calibrating γ. Under the fitted interference model,

$$\widehat{s}(z) = \widehat{\rho}\,\widetilde{L}z.$$

For a collection of assignments $\mathcal{Z}$, define

$$\widehat{\gamma}(\mathcal{Z}) := \sup_{z\in\{0,1\}^n} \widehat{s}(z)^T L^\dagger \widehat{s}(z).$$

Equivalently,

$$\widehat{\gamma}(\mathcal{Z}) = \widehat{\rho}^2 \sup_{z\in\{0,1\}^n} z^T \widetilde{L}^T L^\dagger \widetilde{L} z.$$

A simple conservative upper bound is

$$\widehat{\gamma} \le \widehat{\rho}^2 n\, \lambda_{\max}\left(\widetilde{L}^T L^\dagger \widetilde{L}\right),$$

since $\|z\|_2^2 \le n$ for every $z \in \{0,1\}^n$.

Calibrating δ. The parameter δ controls the mean component of the interference through the condition

$$\left|\frac{1}{n}\mathbf{1}^T s(z)\right| \le \sqrt{\delta}.$$

Accordingly, a fitted value is

$$\widehat{\delta}(\mathcal{Z}) := \sup_{z\in\mathcal{Z}} \left(\frac{1}{n}\mathbf{1}^T \widehat{s}(z)\right)^2.$$

For the working model above,

$$\widehat{\delta} = \frac{\widehat{\rho}^2}{4n^2}\left\|\widetilde{L}^T\mathbf{1}\right\|_1^2.$$

To see this, let

$$c := \widetilde{L}^T\mathbf{1}.$$

Since $\widetilde{L}\mathbf{1} = 0$, we have $\mathbf{1}^T c = 0$, and hence

$$\sup_{z\in\{0,1\}^n} |c^T z| = \frac{1}{2}\|c\|_1.$$

If instead the working interference model is $s(z) = \rho L z$, then $\mathbf{1}^T s(z) = 0$ for every z, and therefore $\delta = 0$.

Calibrating κ. If the heterogeneous components are treated as deterministic residual variation, a direct calibration is

$$\widehat{\kappa} := \max\left\{\|\widehat{\varepsilon}_\alpha\|_2^2, \|\widehat{\varepsilon}_\tau\|_2^2\right\}.$$

This corresponds to the deterministic heterogeneous-variation formulation and induces the operator-norm robustness penalty on the design covariance matrix.

Alternatively, the heterogeneous components may be modeled as random effects. Suppose that

$$\Omega_\alpha := \mathbb{E}\big[\varepsilon_\alpha \varepsilon_\alpha^T\big], \qquad \Omega_\tau := \mathbb{E}\big[\varepsilon_\tau \varepsilon_\tau^T\big]$$

denote the second-moment matrices of the residual baseline and treatment-effect components, respectively. The stochastic robustness condition in our framework takes the form

$$\|\Omega_\alpha\|_{q^\star} \le \kappa, \qquad \|\Omega_\tau\|_{q^\star} \le \kappa.$$

Thus, if these second-moment matrices can be estimated from historical data, a natural plug-in calibration is

$$\widehat{\kappa} := \max\left\{\|\widehat{\Omega}_\alpha\|_{q^\star}, \|\widehat{\Omega}_\tau\|_{q^\star}\right\}.$$

One setting in which this is possible is when several comparable historical realizations of the residual components are available. Suppose, for example, that the fitting procedure yields residual baseline vectors

$$\widehat{\varepsilon}_\alpha^{(1)}, \ldots, \widehat{\varepsilon}_\alpha^{(R)}$$

from R comparable historical populations or experimental settings. Assuming that these residual vectors are mean zero and may be regarded as comparable realizations of the same stochastic component, one may estimate

$$\widehat{\Omega}_\alpha := \frac{1}{R}\sum_{r=1}^{R} \widehat{\varepsilon}_\alpha^{(r)} \widehat{\varepsilon}_\alpha^{(r)T}.$$

Similarly, if residual treatment-effect vectors

$$\widehat{\varepsilon}_\tau^{(1)}, \ldots, \widehat{\varepsilon}_\tau^{(R)}$$

are available, one may estimate

$$\widehat{\Omega}_\tau := \frac{1}{R} \sum_{r=1}^{R} \widehat{\varepsilon}_\tau^{(r)} \widehat{\varepsilon}_\tau^{(r)T}.$$

The Schatten norms of these estimated matrices can then be used to calibrate κ. This approach requires historical data that contains enough variation to separately recover the residual baseline and treatment-effect components, and enough comparable realizations to learn their dependence structure.

When repeated realizations are not available, a more parsimonious alternative is to impose a structured model on the residual second-moment matrix. For example, one may posit

$$\Omega_\alpha = \sigma_\alpha^2 I + \rho_\alpha K,$$

where K is a prespecified positive semidefinite kernel encoding a plausible source of residual dependence, such as spatial proximity, network similarity, or shared observed characteristics. With $\sigma_\alpha^2 \geq 0$ and $\rho_\alpha \geq 0$, this specification is positive semidefinite. The finite-dimensional parameters σ_α^2 and ρ_α can then be estimated from historical residuals using, for example, likelihood-based or moment-based procedures. For example, under the working model

$$\varepsilon_f \sim \mathcal{N}\left(0,\, \sigma_f^2 I + \rho_f K\right),$$

Gaussian maximum likelihood based on a single residual realization $\widehat{\varepsilon}_f$ gives

$$(\widehat{\sigma}_f^2, \widehat{\rho}_f) \in \underset{\sigma_f^2 > 0,\, \rho_f \geq 0}{\arg\min} \left\{ \log\det\left(\sigma_f^2 I + \rho_f K\right) + \widehat{\varepsilon}_f^T \left(\sigma_f^2 I + \rho_f K\right)^{-1} \widehat{\varepsilon}_f \right\}.$$

If R independent or approximately independent historical residual realizations $\widehat{\varepsilon}_f^{(1)}, \ldots, \widehat{\varepsilon}_f^{(R)}$ are available, the corresponding criterion becomes

$$R \log\det\left(\sigma_f^2 I + \rho_f K\right) + \sum_{r=1}^{R} \widehat{\varepsilon}_f^{(r)T} \left(\sigma_f^2 I + \rho_f K\right)^{-1} \widehat{\varepsilon}_f^{(r)}.$$

Restricted maximum likelihood may similarly be used when the mean and covariance components are estimated jointly.

A simpler method-of-moments calibration follows from

$$\mathbb{E}\left[\varepsilon_f^T \varepsilon_f\right] = n\sigma_f^2 + \rho_f \operatorname{Tr}(K),$$

and

$$\mathbb{E}\left[\varepsilon_f^T K \varepsilon_f\right] = \sigma_f^2 \operatorname{Tr}(K) + \rho_f \operatorname{Tr}(K^2).$$

For

$$m_{1,f} := \widehat{\varepsilon}_f^T \widehat{\varepsilon}_f, \qquad m_{2,f} := \widehat{\varepsilon}_f^T K \widehat{\varepsilon}_f,$$

matching these empirical and theoretical moments yields, provided K is not proportional to the identity,

$$\widehat{\sigma}_f^2 = \frac{m_{1,f} \operatorname{Tr}(K^2) - m_{2,f} \operatorname{Tr}(K)}{n \operatorname{Tr}(K^2) - \operatorname{Tr}(K)^2},$$

and

$$\widehat{\rho}_f = \frac{n m_{2,f} - \operatorname{Tr}(K) m_{1,f}}{n \operatorname{Tr}(K^2) - \operatorname{Tr}(K)^2}.$$

With several historical residual vectors, the moments $m_{1,f}$ and $m_{2,f}$ can be averaged across realizations. In finite samples, one may instead impose the constraints $\sigma_f^2, \rho_f \geq 0$ directly when fitting the moment equations.

Applying either procedure to $f = \alpha$ and $f = \tau$ gives estimated second-moment matrices

$$\widehat{\Omega}_f = \widehat{\sigma}_f^2 I + \widehat{\rho}_f K, \qquad f \in \{\alpha, \tau\}.$$

The stochastic robustness radius can then be calibrated as

$$\widehat{\kappa} := \max\left\{ \|\widehat{\Omega}_\alpha\|_{q^\star}, \|\widehat{\Omega}_\tau\|_{q^\star} \right\}.$$

This calibration remains model dependent. In particular, a single historical realization cannot nonparametrically identify an unrestricted residual covariance matrix. The choice of kernel K and covariance family therefore represents a substantive modeling assumption, and sensitivity analysis over plausible kernels or variance-component values provides a natural way to assess the robustness

of the resulting choice of κ.

Calibrating Δ. Finally, in the symmetric setting considered in the main text, the coefficient multiplying the constant-direction term is

$$\Delta = \left(a + \frac{b}{2}\right)^2 + \delta.$$

Thus, after calibrating the preceding quantities, we use

$$\widehat{\Delta} := \left(\widehat{a} + \frac{\widehat{b}}{2}\right)^2 + \widehat{\delta}.$$

This working model therefore yields a concrete calibration of the quantities entering our design objective:

$$\widehat{\eta}, \qquad \widehat{\gamma}, \qquad \widehat{\kappa}, \qquad \widehat{\delta}, \qquad \widehat{\Delta}.$$

The particular linear specification above is intended only as an illustrative working model. Other covariate models or exposure-response specifications may be used, after which the trade-off parameters can be obtained by evaluating the same quadratic and norm constraints defining the model class.

Additionally, rather than relying on a single estimate $(\widehat{\eta}, \widehat{\gamma}, \widehat{\kappa}, \widehat{\Delta})$, the practitioner may specify a plausible region

$$\mathcal{P} = [\eta_-, \eta_+] \times [\gamma_-, \gamma_+] \times [\kappa_-, \kappa_+], \times [\Delta_-, \Delta_+],$$

solve the design problem across this region, and examine the resulting covariance matrices and worst-case MSE bounds. If the same or similar designs remain optimal throughout a large portion of $\mathcal{P}$, then precise calibration of the parameters is relatively unimportant. Conversely, if small changes to the parameters lead to qualitatively different designs, this indicates that the associated modeling assumptions are consequential and deserve additional empirical or substantive investigation. One can also choose a single design robustly by solving

$$\min_X \sup_{(\eta,\gamma,\kappa,\Delta)\in\mathcal{P}} \mathcal{J}(X; \eta, \gamma, \kappa, \Delta),$$

where $\mathcal{J}$ denotes the corresponding worst-case MSE upper bound.

Ultimately, the strength of our framework lies in its adaptability. Much like the Gram-Schmidt

Walk algorithm in Harshaw et al. [2024], our approach does not prescribe a rigid estimation protocol. Instead, it provides a structured yet flexible methodology that practitioners can tailor to their specific experimental context, whether by refining parameter estimates through machine learning, incorporating alternative homophily metrics, or integrating auxiliary data sources. This adaptability ensures that domain experts can efficiently explore trade-offs and arrive at designs that balance interference, homophily, and robustness in a principled manner.

5.I Difference-in-Means estimator with Fixed Treatment Groups Sizes

In this section, we consider our proposed framework applied to the difference-in-means estimator. Let $n_1 := \sum_{i=1}^{n} z_i$, and $n_0 := n - n_1$, i.e., the number of treated and control units respectively. Then,

$$\begin{aligned}\hat{\tau}_{\mathrm{DiM}} &= \frac{1}{n_1}\langle z, Y\rangle - \frac{1}{n_0}\langle 1-z, Y\rangle \\ &= \frac{1}{n_1}\langle z, \alpha + \mathrm{Diag}(\phi)z - s(z)\rangle - \frac{1}{n_0}\langle 1-z, \alpha + \mathrm{Diag}(\phi)z - s(z)\rangle \\ &= \left\langle \frac{z}{n_1} - \frac{1-z}{n_0}, \alpha - s(z)\right\rangle + \left\langle \frac{z}{n_1}, \phi\right\rangle.\end{aligned}$$

Therefore,

$$\hat{\tau}_{\mathrm{DiM}} - \tau = \left\langle \frac{z}{n_1} - \frac{1-z}{n_0}, \alpha - s(z)\right\rangle + \left\langle \frac{z}{n_1} - \frac{\mathbf{1}}{n}, \phi\right\rangle.$$

Theorem 5.I.1 (MSE bound)**.** *Let $x \in \{-1,1\}^n$ by defining $x := 2z - 1$. Define $X := \mathbb{E}[xx^T]$. Let $\mathbf{1}^T X \mathbf{1} - \mathbf{1}^T \mu\mu^T \mathbf{1} = 0$, where $\mu = \frac{n_1 - n_0}{n}\mathbf{1}$, and let $\beta_1 := n/(2n_1n_0)$, and $\beta_2 := 1/(2n_1)$. Then, for $q \geq 1$,*

$$\mathbf{E}[(\hat{\tau}_{DiM} - \tau)^2] \leq 7\{\mathrm{Tr}((\gamma\beta_1^2 L + \eta[\beta_1^2 + \beta_2^2]L^\dagger)X) + \kappa[\beta_1^2 + \beta_2^2]\|X\|_q\}.$$

The proof is in Appendix 5.J.4.

Corollary 5.I.2 (MSE bound (Symmetric case))**.** *For the symmetric case, i.e., when $n_1 = n_0 = n/2$, let $x \in \{-1,1\}^n$ by defining $x := 2z - 1$. Let $\mathbf{1}^T X \mathbf{1} = 0$. Define $X := \mathbb{E}[xx^T]$. Then, for $q \geq 1$,*

$$\mathbf{E}[(\hat{\tau}_{DiM} - \tau)^2] \leq \frac{5}{n^2}\{\mathrm{Tr}((4\gamma L + 5\eta L^\dagger)X) + 5\kappa\|X\|_q\}.$$

Let $\mu = \frac{n_1 - n_0}{n}\mathbf{1}$. Reparameterizing with $\tilde{\gamma}$ and $\tilde{\eta}$ so that

$$\tilde{\gamma} := \frac{\gamma\beta_1^2}{\kappa[\beta_1^2 + \beta_2^2]} \text{ and } \tilde{\eta} := \frac{\eta}{\kappa},$$

we can write an SDP more explicitly in terms of the new trade-off parameters:

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times n}} \quad & \tilde{\eta}\mathrm{Tr}(L^\dagger X) + \tilde{\gamma}\mathrm{Tr}(LX) + \|X\|_q \\ \text{s.t.} \quad & X \succeq 0 \\ & \mathrm{diag}(X) = 1 \\ & \mathbf{1}X\mathbf{1} = \mathbf{1}^T\mu\mu^T\mathbf{1} \end{aligned} \tag{5.I.1}$$

Let X^* be the solution to the SDP above. Finally, we generate our treatment assignment vector x^* by running Algorithm 2 on X^*.

Remark 5.I.3. When $\kappa = 0$, we can proceed with a matrix lifting and a SDP relaxation as previously demonstrated, or solve the problem more exactly with a mixed-integer program as described below. We note however that solving an SDP is polynomial-time with efficient solvers that can be sped up by exploiting problem structure. This is not true for the mixed-integer programs. From the calculations in the proof of Theorem 5.I.1, when $\kappa = 0$, and $\tilde{\gamma} := \frac{\gamma\beta_1^2}{[\beta_1^2+\beta_2^2]}$ and $\tilde{\eta} := \eta[\beta_1^2 + \beta_2^2]$, we have the following MIP problem:

$$\min_{x \in \mathbb{R}^n} \quad \tilde{\eta}x^T L^\dagger x + \tilde{\gamma}x^T Lx \tag{5.I.2}$$

$$\text{s.t.} \quad x_i = \pm 1 \quad \forall i \tag{5.I.3}$$

$$x^T\mathbf{1} = \mathbf{1}^T\mu \tag{5.I.4}$$

5.I.1 Gaussian Rounding

One approach to proving theoretical guarantees of this approximation procedure is by leveraging and adapting results from [Raghavendra and Tan, 2012] who provide approximation guarantees to the min-bisection problem. For convenience, we restate the min-bisection problem followed by the main result of [Raghavendra and Tan, 2012].

Recall that the min-bisection problem is:

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{2}\sum_{i<j} w_{ij}(1 - x_i \cdot x_j) \\ \text{s.t.} \quad & x_i \in \{-1, 1\} \quad \forall i \\ & \sum_i x_i = 0 \end{aligned}$$

The SDP relaxation is given by,

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\big[\sum_{i<j} w_{ij}(v_i - v_j)^2\big] \\ \text{s.t.} \quad & \|v_i\|^2 = 1 \quad \forall i \in V \\ & \sum_i v_i = 0 \end{aligned}$$

or

$$\begin{aligned} \underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad & \text{Tr}(LX) \\ \text{s.t.} \quad & X_{ii} = 1 \quad \forall i \in V \\ & \mathbf{1}^T X \mathbf{1} = 0 \\ & X \succeq 0 \end{aligned}$$

For this min-bisection problem, Raghavendra and Tan [2012] propose an algorithm and provide a corresponding approximation error bound.

Theorem 5.I.4 (Raghavendra and Tan, 2012, Theorem 1.2)**.** *For every $\delta > 0$, there exists an algorithm, which given a graph with a bisection cutting ϵ-fraction of the edges, finds a bisection cutting at most $\mathcal{O}(\sqrt{\epsilon}) + \delta$-fraction of edges with high probability.*

Their algorithm comprises of two parts. First, they lift the problem through the Lasserre hierarchy and iteratively solve by conditioning on assignments of a subset of nodes, until a pre-specified α-level of independence is achieved between the solutions across the nodes. The second part of their approach, given this α-independence solution, is a rounding procedure reminiscent of [Goemans and Williamson, 1995]. Finally, to meet the cardinality constraints, they flip the assignments of the nodes with the

smallest degrees on the larger side of the cut.

In proving their guarantees from this rounding, they leverage the α-independence conditioning. While their α-conditioning in the first-part of their procedure does not apply to our setting, one could adapt their ideas to proving guarantees in our setting under a α_κ-independence level, dependent on the randomization tuning parameter κ as per Proposition 5.4.7. To adapt their final "flipping" procedure to achieve cardinality constraints to our settings, we propose targeting the nodes with the smallest a_i with a_i corresponding to the columns vectors of the matrix A, instead of the nodes of the smallest degrees. Then, one could prove analogous results for this difference-in-means setting.

5.I.2 Adapted Gram Schmidt Walk algorithm

As a heuristic, one could use the adapted Gram Schmidt Walk algorithm as in the Horvitz-Thompson section, followed by the assignment "flipping" procedure targeting the nodes, on the larger side of the cut, with the smallest Q_{ii} corresponding to columns vectors in the matrix Q. We also refer to [Harshaw et al., 2024, S8.4]. In particular, the authors describe how appending a constant vector to the covariate matrix helps make the treatment group sizes, n_1, n_0 closer to equal. A variant proposed by Harshaw et al. [2024, S8.4] is by adding a cardinality constraint within the Gram-Schmidt Walk algorithm. This variant, however, does not inherit the theoretical results from the Bernoulli rounding procedures in [Harshaw et al., 2024] due to violation of orthogonality of updates between pivot phases in the algorithm.

5.J Proofs to trade-off results

5.J.1 Proof to Proposition 5.4.7

We restate Proposition 5.4.7 and prove it below.

Proposition 5.4.7. *Let $r \in \{-1, 1\}^n$ denote the treatment assignment under a Bernoulli randomization design. Then, for fixed $p = \mathbb{P}(r_i = 1)$ for all $i \in [n]$, and $q > 1$*

$$\arg \min_{\substack{X \succeq 0; \\ X_{ii}=1}} \|X\|_q = \frac{1}{4p(1-p)} \operatorname{Cov}(r).$$

Proof. Let $X^* := I$. Then, X^* has n non-zero eigenvalues which are all equal to 1. We first show that X^* is the unique minimizer of $\|X\|_q^q$ over $X \in \mathcal{C} := \{X : X \succeq 0, X_{ii} = 1\}$, and then we show that

$X^* = \frac{1}{4p(1-p)}\,\mathrm{Cov}(r)$. It is easy to see that $X^* \in \mathcal{C}$. For $q > 1$, and any $X \in \mathcal{C}$, with eigenvalues λ_i,

$$\|X\|_q^q = (n) \times \frac{1}{n}\sum_{i=1}^{n} \lambda_i^q \geq (n) \times (\frac{1}{n}\sum_{i=1}^{n} \lambda_i)^q = (n) \times 1 = \|X^*\|_q^q,$$

as $f : x \to x^q$ is convex on $x \geq 0$, and $\sum_i \lambda_i = \sum_i X_{ii} = n$. That X^* is the unique minimizer follows from the strict convexity of $\|\cdot\|_q^q$ on $\mathcal{C}$ for $q > 1$. Indeed, if we express $\|X\|_q^q = g(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^n \lambda_i^q$, then

$$\nabla^2 g = q(q-1)\begin{bmatrix} \lambda_1^{q-2} & 0 & \ldots & 0 \\ 0 & \lambda_2^{q-2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n^{q-2} \end{bmatrix},$$

which is positive definite at X^* with its n eigenvalues $\lambda_i = 1$ for $i = 1, 2, ..., n$.

It is now sufficient to show that this matrix X^* is equivalent to the $\frac{1}{4p(1-p)}\,\mathrm{Cov}(r)$ under Bernoulli randomization with uniform treatment assignment probability. We begin by observing that $\mathrm{Cov}(r) = \mathbb{E}[rr^T] - \mathbb{E}[r]\mathbb{E}[r]^T$. Therefore, since $r \in \{-1, 1\}$ with $\mathbb{P}(r_i = 1) = p$ for all i, we have that $\mathrm{Cov}(r)_{ii} = 4p(1-p)$. Recall that unit-level Bernoulli randomization is a randomization scheme where each unit is independently assigned to treatment or control. Therefore, $\mathrm{Cov}(r)_{ij} = 0$ for all $j \neq i$. ■

We state below the analogous result for complete randomization.

Proposition 5.J.1. *Let $r \in \{-1, 1\}^n$ denote the treatment assignment under a complete randomization design. Then, for fixed $n_1 = n_0$, i.e., $r^T\mathbf{1} = 0$, and $q > 1$*

$$\arg\min_{\substack{X\mathbf{1}=\mathbf{0};\\ X\succeq 0;\\ X_{ii}=1}} \|X\|_q = \mathrm{Cov}(r).$$

Proof. Let $X^* := (1 + \frac{1}{n-1})I - \frac{1}{n-1}\mathbf{1}\mathbf{1}^T = \frac{n}{n-1}I - \frac{n}{n-1}(\mathbf{1}/\sqrt{n})(\mathbf{1}/\sqrt{n})^T$. Then, X^* has $n-1$ non-zero eigenvalues which are all equal to $\frac{n}{n-1}$, and one zero eigenvalue. Indeed, the first term has n eigenvalues that are equal to $\frac{n}{n-1}$, while the second term has one eigenvalue that is equal to $\frac{n}{n-1}$, and $n-1$ eigenvalues that are equal to 0, and as they share the same eigenvectors, their difference has the desired property. We first show that X^* is the unique minimizer of $\|X\|_q^q$ over $X \in \mathcal{C} := \{X : X\mathbf{1} = \mathbf{0}, X \succeq 0, X_{ii} = 1\}$, and then we show that $X^* = \mathrm{Cov}(r)$. It is easy to see that

$X^* \in \mathcal{C}$. For $q > 1$, and any $X \in \mathcal{C}$, with eigenvalues λ_i,

$$\begin{aligned}\|X\|_q^q = (n-1) \times \frac{1}{n-1} \sum_{i=1}^n \lambda_i^q &\geq (n-1) \times (\frac{1}{n-1} \sum_{i=1}^n \lambda_i)^q \\ &= (n-1) \times (\frac{n}{n-1})^q = \|X^*\|_q^q,\end{aligned}$$

as $f : x \to x^q$ is convex on $x \geq 0$, and $\sum_i \lambda_i = \sum_i X_{ii} = n$. That X^* is the unique minimizer follows from the strict convexity of $\|\cdot\|_q^q$ on $\mathcal{C}$ for $q > 1$. Indeed, if we express $\|X\|_q^q = g(\lambda_1, \ldots, \lambda_{n-1}) = \sum_{i=1}^{n-1} \lambda_i^q$, then

$$\nabla^2 g = q(q-1) \begin{bmatrix} \lambda_1^{q-2} & 0 & \ldots & 0 \\ 0 & \lambda_2^{q-2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_{n-1}^{q-2} \end{bmatrix},$$

which is positive definite at X^* with its first $n-1$ eigenvalues $\lambda_i = \frac{n}{n-1}$ for $i = 1, 2, ..., n-1$.

It is now sufficient to show that this matrix X^* is equivalent to the $\mathrm{Cov}(r)$ under complete randomization, with equal number of treated and control units, i.e., $r^T \mathbf{1} = 0$. We begin by observing that $\mathrm{Cov}(r) = \mathbb{E}[rr^T]$, and that $\mathrm{Cov}(r)\mathbf{1} = \mathbb{E}[rr^T\mathbf{1}] = 0$. Therefore, since $r \in \{-1, 1\}$ with equal probability, we have that $\mathrm{Cov}(r)_{ii} = 1$, and that the row sums of $\mathrm{Cov}(r) = 0$. This must imply that $\sum_{j \neq i} \mathrm{Cov}(r)_{ij} = -1$, for all $i \in [n]$. Recall that unit-level complete randomization is a uniform randomization process. Therefore, $\mathrm{Cov}(r)_{ij} = \mathrm{Cov}(r)_{ik}$ for all $j \neq k \neq i$. Therefore, this together with $\sum_{j \neq i} \mathrm{Cov}(r)_{ij} = -1$ imply that $\mathrm{Cov}(r)_{ij} = -\frac{1}{n-1}$ for all $j \neq i$, for every i. ■

5.J.2 Proof to Theorem 5.C.1

We first state and prove lemmas we use to prove Theorem 5.C.1.

Lemma 5.J.2. *Suppose that h is η-homophilous, $\langle h, \mathbf{1} \rangle = 0$, and $X := \mathbb{E}[xx^T]$, then*

$$\mathbb{E}[\langle x, h \rangle^2] \leq \eta \, \mathrm{Tr}(L^\dagger X).$$

Proof. Observe, as h is orthogonal to the kernel of L, and L is PSD,

$$\sup_{\langle h, Lh \rangle \leq \eta} \langle x, h \rangle = \sup_{\langle h, Lh \rangle \leq \eta} \left\langle x, L^{\dagger/2} L^{1/2} h \right\rangle = \sup_{\langle h, Lh \rangle \leq \eta} \left\langle L^{\dagger/2} x, L^{1/2} h \right\rangle.$$

Therefore, by Cauchy-Schwarz,

$$\sup_{\langle h, Lh\rangle \le \eta} \mathbb{E}[\langle x, h\rangle^2] = \eta \operatorname{Tr}(L^\dagger X).$$

■

Lemma 5.J.3. *Suppose that s is γ-interferent, $\langle s(z), \mathbf{1}/n\rangle = 0$ for all z, and $X := \mathbb{E}[xx^T]$, then*

$$\mathbb{E}[\langle x, s(z)\rangle^2] \le \gamma \operatorname{Tr}(LX).$$

Proof. Observe, as $s(z)$ is orthogonal to the kernel of L, and L is PSD,

$$\begin{aligned}\sup_{\langle s(z), L^\dagger s(z)\rangle \le \gamma} \langle x, s(z)\rangle &= \sup_{\langle s(z), L^\dagger s(z)\rangle \le \gamma} \left\langle x, L^{1/2} L^{\dagger/2} s(z)\right\rangle \\ &= \sup_{\langle s(z), L^\dagger s(z)\rangle \le \gamma} \left\langle L^{1/2} x, L^{\dagger/2} s(z)\right\rangle.\end{aligned}$$

Therefore, by Cauchy-Schwarz,

$$\sup_{\langle s(z), L^\dagger s(z)\rangle \le \gamma} \mathbb{E}[\langle x, s(z)\rangle^2] = \gamma \operatorname{Tr}(LX).$$

■

Lemma 5.J.4. *Suppose that ε is independent of x and $\Sigma := \mathbb{E}[\varepsilon\varepsilon^T]$ is such that $\|\Sigma\|_{q^*} \le \kappa$, $1/q^* + 1/q = 1$. Let $X := \mathbb{E}[xx^T]$. Then,*

$$\mathbb{E}[\langle x, \varepsilon\rangle^2] \le \kappa \|X\|_q.$$

Proof. We begin by writing

$$\mathbb{E}[\langle x, \varepsilon\rangle^2] = \operatorname{Tr}(\mathbb{E}[\varepsilon\varepsilon^T]\mathbb{E}[xx^T]) = \operatorname{Tr}(\Sigma X).$$

Then, the proof follows from Proposition 5.4.4. ■

In fact, the results in Lemma 5.J.2 can be tightened. We write this formally in the following lemma.

Lemma 5.J.5. *Suppose that h is η-homophilous, $\langle h, \mathbf{1}\rangle = 0$, and $X := \mathbb{E}[xx^T]$, then*

$$\mathbb{E}[\langle x, h\rangle^2] \le \eta\, \lambda_{\max}(L^{\dagger/2} X L^{\dagger/2}).$$

Proof. Observe, as h is orthogonal to the kernel of L, and L is PSD,

$$\begin{aligned}\sup_{\langle h, Lh\rangle \le \eta} \mathbb{E}[\langle x, h\rangle^2] = \sup_{\langle h, Lh\rangle \le \eta} h^T X h &= \sup_{\|L^{1/2}h\| \le \eta^{1/2}} h^T L^{1/2} L^{\dagger/2} X L^{\dagger/2} L^{1/2} h \\ &= \sup_{\|w\| \le 1} \eta w^T L^{\dagger/2} X L^{\dagger/2} w \\ &= \eta \|L^{\dagger/2} X L^{\dagger/2}\|_{\mathrm{op}}.\end{aligned}$$

■

We restate Theorem 5.C.1 below.

Theorem 5.C.1 ((General) MSE bound)**.** *Let $p := \mathbb{P}(z_i = 1)$ for all $i \in [n]$, $\beta_1 := 1/(2np(1-p))$, $\beta_2 := 1/(2np)$, $a = \langle \alpha, \mathbf{1}/n\rangle$, $b = \langle \phi, \mathbf{1}/n\rangle$, and $\langle s(z), \mathbf{1}/n\rangle \le \sqrt{\delta}$ for all z. Let also $x \in \{-1, 1\}^n$ by defining $x := 2z - 1$. Define $X := \mathbb{E}[xx^T] - \mu\mu^T$, $\mu := \mathbb{E}[x]$. Then,*

$$\begin{aligned}\mathbb{E}[(\hat{\tau} - \tau)^2] \le 7\{ &\operatorname{Tr}((\gamma\beta_1^2 L + \eta[\beta_1^2 + \beta_2^2] L^\dagger) X) \\ &+ \kappa[\beta_1^2 + \beta_2^2] \|X\|_q + ([a\beta_1 + b\beta_2]^2 + \beta_1^2 \delta) \operatorname{Tr}(\mathbf{1}\mathbf{1}^T X)\}.\end{aligned}$$

Proof. The relevant quantities of interest are

$$\hat{\tau} = \frac{1}{n} \langle Y, z/p - (1-z)/(1-p)\rangle = \frac{1}{n} \langle \alpha + s(z), z/p - (1-z)/(1-p)\rangle \tag{5.J.1}$$

$$+ \frac{1}{n} \langle \phi, z/p\rangle, \tag{5.J.2}$$

$$\tau = \langle \phi, \mathbf{1}/n\rangle, \tag{5.J.3}$$

$$\hat{\tau} - \tau = \langle \alpha + s(z), z/(np) - (1-z)/(n(1-p))\rangle + \langle \phi, z/(np) - \mathbf{1}/n\rangle. \tag{5.J.4}$$

Notice that we can re-write

$$\begin{aligned}&z/(np) - (1-z)/(n(1-p)) \\ &= \left[z/(np) - (1-z)/(n(1-p)) - \frac{1}{2}(1/(np) - 1/(n(1-p)))\mathbf{1}\right]\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{2}(1/(np) - 1/(n(1-p)))\mathbf{1} \\
&=: \beta_1 x + c_1 \mathbf{1}
\end{aligned}
$$

where $c_1 := \frac{1}{2}(1/(np) - 1/(n(1-p)))$. Similarly,

$$
\begin{aligned}
z/(np) - \mathbf{1}/n &= [z/(np) - \mathbf{1}/n - (1/(2np) - 1/n)\mathbf{1}] + (1/(2np) - 1/n)\mathbf{1} \\
&=: \beta_2 x + c_2 \mathbf{1}
\end{aligned}
$$

where $c_2 := (1/(2np) - 1/n)$. Before proceeding, we first center α, ϕ, decomposing them as $\alpha = \bar{\alpha} + \alpha'$, $\phi = \bar{\phi} + \phi'$, $\bar{\alpha}$, $\bar{\phi}$ denoting the projection onto the constant vector for each term respectively. Similarly, we decompose $s(z) = \bar{s}(z) + \tilde{s}(z)$. Then, we can write,

$$
\begin{aligned}
&\mathbb{E}[(\hat{\tau} - \tau)^2] \\
&= \mathbb{E}[(\langle \alpha + s(z), z/(np) - (1-z)/(n(1-p))\rangle + \langle \phi, z/(np) - \mathbf{1}/n\rangle)^2] \\
&= 7\mathbb{E}[(\langle \bar{\alpha}, z/(np) - (1-z)/(n(1-p))\rangle + \langle \bar{\phi}, z/(np) - \mathbf{1}/n\rangle)^2 \\
&\quad + \langle \bar{s}(z), z/(np) - (1-z)/(n(1-p))\rangle^2 \\
&\quad + \langle h_\alpha, \beta_1(x-\mu)\rangle^2 + \langle \tilde{s}(z), \beta_1(x-\mu)\rangle^2 + \langle \varepsilon_\alpha, \beta_1(x-\mu)\rangle^2 \\
&\quad + \langle h_\phi, \beta_2(x-\mu)\rangle^2 + \langle \varepsilon_\phi, \beta_2(x-\mu)\rangle^2],
\end{aligned}
$$

where the last two lines on the right-hand-side come from the mean-zero components of α, $s(z)$, and ϕ. Indeed, as these are orthogonal to $\mathbf{1}$ we can drop the $c\mathbf{1}$ terms from the expression, and add the μ terms. Thus, we can write

$$
\begin{aligned}
&\langle \alpha' + \tilde{s}(z), z/(np) - (1-z)/(n(1-p))\rangle + \langle \phi', z/(np) - \mathbf{1}/n\rangle \\
&= \langle h_\alpha, \beta_1(x-\mu)\rangle + \langle \tilde{s}(z), \beta_1(x-\mu)\rangle + \langle \varepsilon_\alpha, \beta_1(x-\mu)\rangle \\
&+ \langle h_\phi, \beta_2(x-\mu)\rangle + \langle \varepsilon_\phi, \beta_2(x-\mu)\rangle.
\end{aligned}
$$

We can upper-bound the contribution of the terms in the first two lines, as we notice that by the definition of our marginal probabilities,

$$
\mathbb{E}[\langle \bar{\alpha}, z/(np) - (1-z)/(n(1-p))\rangle] = 0 = \mathbb{E}[\langle \bar{\phi}, z/(np) - \mathbf{1}/n\rangle],
$$

hence,

$$\begin{aligned}
&\mathbb{E}[(\langle\bar{\alpha}, z/(np) - (1-z)/(n(1-p))\rangle + \langle\bar{\phi}, z/(np) - \mathbf{1}/n\rangle)^2] \\
&= \operatorname{Var}(\langle\bar{\alpha}, z/(np) - (1-z)/(n(1-p))\rangle + \langle\bar{\phi}, z/(np) - \mathbf{1}/n\rangle) \\
&= \operatorname{Var}(\langle\bar{\alpha}, \beta_1(x-\mu)\rangle + \langle\bar{\phi}, \beta_2(x-\mu)\rangle) \\
&= \operatorname{Var}([\beta_1 a + \beta_2 b]\langle\mathbf{1}, (x-\mu)\rangle) + \operatorname{Var}(\langle\bar{s}(z), (x-\mu)\rangle) \\
&= (\beta_1 a + \beta_2 b)^2 \operatorname{Tr}(\mathbf{1}\mathbf{1}^T X),
\end{aligned}$$

for $X := \mathbb{E}[xx^T] - \mu\mu^T$, $\mu := \mathbb{E}[x]$, $a = \langle\alpha, \mathbf{1}/n\rangle$, $b = \langle\phi, \mathbf{1}/n\rangle$.

Similarly,

$$\begin{aligned}
&\mathbb{E}[\langle\bar{s}(z), z/(np) - (1-z)/(n(1-p))\rangle^2] \\
&\leq \delta\mathbb{E}[\langle\mathbf{1}, z/(np) - (1-z)/(n(1-p))\rangle^2] \\
&= \delta \operatorname{Var}(\langle\mathbf{1}, z/(np) - (1-z)/(n(1-p))\rangle) \\
&= \delta \operatorname{Var}(\langle\mathbf{1}, \beta_1(x-\mu)\rangle) \\
&\leq \beta_1^2 \delta \operatorname{Tr}(\mathbf{1}\mathbf{1}^T X).
\end{aligned}$$

Thus, by Lemmas 5.J.2 to 5.J.4, we get

$$\begin{aligned}
\mathbb{E}[(\hat{\tau} - \tau)^2] \leq \operatorname{Tr}((\gamma\beta_1^2 7L + \eta[\beta_1^2 + \beta_2^2]7L^\dagger)X) \\
+ \kappa[\beta_1^2 + \beta_2^2]7\|X\|_q + ([a\beta_1 + b\beta_2]^2 + \beta_1^2\delta)7 \operatorname{Tr}(\mathbf{1}\mathbf{1}^T X).
\end{aligned}$$

■

5.J.3 Tighter bounds

In Theorem 5.C.1, we relied on the inequality from Lemma 5.J.2. If we, instead, used the tighter inequality from Lemma 5.J.5, we obtain the following worst-case bound.

Theorem 5.J.6 ((Tighter) MSE bound)**.** *Let $p := \mathbb{P}(z_i = 1)$ for all $i \in [n]$, $\beta_1 := 1/(2np(1-p))$, $\beta_2 := 1/(2np)$, $a = \langle\alpha, \mathbf{1}/n\rangle$, $b = \langle\phi, \mathbf{1}/n\rangle$, and $\langle s(z), \mathbf{1}/n\rangle \leq \sqrt{\delta}$ for all z. Let also $x \in \{-1, 1\}^n$ by*

defining $x := 2z - 1$. *Define* $X := \mathbb{E}[xx^T] - \mu\mu^T$, $\mu := \mathbb{E}[x]$. *Then, for* $q \geq 1$,

$$\mathbb{E}[(\hat{\tau} - \tau)^2] \leq 7\{\gamma\beta_1^2 \operatorname{Tr}(LX) + \eta[\beta_1^2 + \beta_2^2]\lambda_{\max}(L^{\dagger/2} X L^{\dagger/2}) \\ + \kappa[\beta_1^2 + \beta_2^2]\|X\|_q + ([a\beta_1 + b\beta_2]^2 + \beta_1^2\delta)\operatorname{Tr}(\mathbf{1}\mathbf{1}^T X)\}.$$

However, writing this as an SDP introduces an additional PSD constraint, and the resulting optimization problem is much more computationally intensive.

5.J.4 Proof to Theorem 5.I.1

We restate Theorem 5.I.1 below.

Theorem 5.I.1 (MSE bound)**.** *Let* $x \in \{-1, 1\}^n$ *by defining* $x := 2z - 1$. *Define* $X := \mathbb{E}[xx^T]$. *Let* $\mathbf{1}^T X \mathbf{1} - \mathbf{1}^T \mu\mu^T \mathbf{1} = 0$, *where* $\mu = \frac{n_1 - n_0}{n}\mathbf{1}$, *and let* $\beta_1 := n/(2n_1 n_0)$, *and* $\beta_2 := 1/(2n_1)$. *Then, for* $q \geq 1$,

$$\mathbf{E}[(\hat{\tau}_{DiM} - \tau)^2] \leq 7\{\operatorname{Tr}((\gamma\beta_1^2 L + \eta[\beta_1^2 + \beta_2^2]L^\dagger)X) + \kappa[\beta_1^2 + \beta_2^2]\|X\|_q\}.$$

Proof. The relevant quantities of interest are

$$\hat{\tau}_{\mathrm{DiM}} = \langle Y, z/n_1 - (1-z)/n_0\rangle = \langle \alpha + s(z), z/n_1 - (1-z)/n_0\rangle + \langle \phi, z/n_1\rangle\,, \tag{5.J.5}$$

$$\tau = \langle \phi, \mathbf{1}/n\rangle\,, \tag{5.J.6}$$

$$\hat{\tau}_{\mathrm{DiM}} - \tau = \langle \alpha + s(z), z/n_1 - (1-z)/n_0\rangle + \langle \phi, z/n_1 - \mathbf{1}/n\rangle\,. \tag{5.J.7}$$

Notice that we can re-write

$$\begin{aligned} &z/n_1 - (1-z)/n_0 \\ &= \left[z/n_1 - (1-z)/n_0 - \frac{1}{2}(1/n_1 - 1/n_0)\mathbf{1}\right] + \frac{1}{2}(1/n_1 - 1/n_0)\mathbf{1} \\ &=: \beta_1 x + c_1\mathbf{1} \end{aligned}$$

where $c_1 := \frac{1}{2}(1/n_1 - 1/n_0)$. Similarly,

$$z/n_1 - \mathbf{1}/n = [z/n_1 - \mathbf{1}/n - (1/(2n_1) - 1/n)\mathbf{1}] + (1/(2n_1) - 1/n)\mathbf{1}$$

$$=: \beta_2 x + c_2 \mathbf{1}$$

where $c_2 := (1/(2n_1) - 1/n)$. Before proceeding, we first center α, ϕ, decomposing them as $\alpha = \bar{\alpha} + \alpha'$, $\phi = \bar{\phi} + \phi'$, $\bar{\alpha}$, $\bar{\phi}$ denoting the projection onto the constant vector for each term respectively. Similarly, we decompose $s(z) = \bar{s}(z) + \tilde{s}(z)$. Then, we can write

$$\begin{aligned}
&\mathbb{E}[(\tau - \hat{\tau}_{\mathrm{DiM}})^2] \\
&= 7\mathbb{E}[(\langle \bar{\alpha}, z/n_1 - (1-z)/n_0 \rangle + \langle \bar{\phi}, z/n_1 - \mathbf{1}/n \rangle)^2 \\
&\quad + \langle \bar{s}(z), z/n_1 - (1-z)/n_0 \rangle^2 \\
&\quad + \langle h_\alpha, \beta_1 x \rangle^2 + \langle \tilde{s}(z), \beta_1 x \rangle^2 + \langle \varepsilon_\alpha, \beta_1 x \rangle^2 + \langle h_\phi, \beta_2 x \rangle^2 + \langle \varepsilon_\phi, \beta_2 x \rangle^2],
\end{aligned}$$

where the terms in the second line of the right-hand-side come from the mean-zero components of α, $s(z)$, and ϕ. Indeed, as these are orthogonal to $\mathbf{1}$ we can drop the $c\mathbf{1}$ terms from the expression. Thus, we can write

$$\begin{aligned}
&\langle \alpha' + \tilde{s}(z), z/n_1 - (1-z)/n_0 \rangle + \langle \phi', z/n_1 - \mathbf{1}/n \rangle \\
&= \langle h_\alpha, \beta_1 x \rangle + \langle \tilde{s}(z), \beta_1 x \rangle + \langle \varepsilon_\alpha, \beta_1 x \rangle \\
&\qquad + \langle h_\phi, \beta_2 x \rangle + \langle \varepsilon_\phi, \beta_2 x \rangle.
\end{aligned}$$

We can upper-bound the contribution of the terms in the first line, as we notice that by the definition of our marginal probabilities,

$$\mathbb{E}[\langle \bar{\alpha}, z/n_1 - (1-z)/n_0) \rangle] = 0 = \mathbb{E}[\langle \bar{\phi}, z/n_1 - \mathbf{1}/n \rangle],$$

hence,

$$\begin{aligned}
&\mathbb{E}[(\langle \bar{\alpha}, z/n_1 - (1-z)/n_0 \rangle + \langle \bar{\phi}, z/n_1 - \mathbf{1}/n \rangle)^2] \\
&= \mathrm{Var}(\langle \bar{\alpha}, z/n_1 - (1-z)/n_0) \rangle + \langle \bar{\phi}, z/n_1 - \mathbf{1}/n \rangle) \\
&= \mathrm{Var}(\langle \bar{\alpha}, \beta_1 x \rangle + \langle \bar{\phi}, z/n_1 - \mathbf{1}/n \rangle) \\
&= (\beta_1 a + \beta_2 b)^2 \,\mathrm{Tr}(\mathbf{1}\mathbf{1}^T [X - \mu\mu^T])
\end{aligned}$$

for $X := \mathbb{E}[xx^T]$, $\mu := \mathbb{E}[x]$, $a = \langle \alpha, \mathbf{1}/n \rangle$, $b = \langle \phi, \mathbf{1}/n \rangle$.

By our constraints $\mathbf{1}^T X \mathbf{1} - \mathbf{1}^T \mu\mu^T \mathbf{1} = 0$,

$$\mathbb{E}[(\langle \bar{\alpha}, z/n_1 - (1-z)/n_0 \rangle + \langle \bar{\phi}, z/n_1 - \mathbf{1}/n \rangle)^2] = 0.$$

Similarly,

$$\begin{aligned}
\mathbb{E}[\langle \bar{s}(z), z/n_1 - (1-z)/n_0 \rangle^2] &\leq \delta \mathbb{E}[\langle \mathbf{1}, z/n_1 - (1-z)/n_0 \rangle^2] \\
&= \delta \operatorname{Var}(\langle \mathbf{1}, z/n_1 - (1-z)/n_0 \rangle) \\
&= \delta \operatorname{Var}(\langle \mathbf{1}, \beta_1 x \rangle) \\
&= \beta_1^2 \delta \operatorname{Tr}(\mathbf{1}\mathbf{1}^T [X - \mu\mu^T]),
\end{aligned}$$

which is constrained to be zero by our constraints $\mathbf{1}^T X \mathbf{1} - \mathbf{1}^T \mu\mu^T \mathbf{1} = 0$.

Thus, by Lemmas 5.J.2 to 5.J.4, we get

$$\mathbb{E}[(\tau - \hat{\tau}_{\mathrm{DiM}})^2] \leq \operatorname{Tr}((\gamma \beta_1^2 7L + \eta[\beta_1^2 + \beta_2^2] 7L^\dagger) X) + \kappa[\beta_1^2 + \beta_2^2] 7 \|X\|_q.$$

■

5.K *Proofs to Theoretical Guarantees*

Our proof of Theorem 5.K.4 utilizes the seminal ideas underlying the approximation bound for MAXCUT, as introduced in Goemans and Williamson [1995]. For convenience, we reiterate the MAXCUT problem in the following subsection.

5.K.1 The MAXCUT problem

$$\begin{aligned}
\underset{x \in \mathbb{R}^n}{\text{maximize}} \quad & \frac{1}{2} \sum_{i<j} w_{ij} (1 - x_i \cdot x_j) \\
\text{s.t.:} \quad & x_i \in \{-1, 1\} \quad \forall i \in V
\end{aligned}$$

The SDP relaxation is given by

$$\begin{aligned} \underset{X \in \mathbb{R}^{n \times n}}{\text{maximize}} \quad & \operatorname{Tr}(LX) \\ \text{s.t.:} \quad & X_{ii} = 1 \quad \forall i \in V \\ & X \succeq 0 \end{aligned}$$

Let $\zeta = \operatorname{sign}(\xi)$, $\xi \sim \mathcal{N}(0, X^*)$, where X^* is the solution to the SDP relaxation of MAXCUT above. Goemans and Williamson [1995] proved that

$$\mathbb{E}\zeta^T L \zeta \geq 0.878 \max_{x \in \{-1,1\}^n} x^T L x,$$

by leveraging Grothendieck's identity

$$\frac{\pi}{2}\mathbb{E}[\zeta_i \zeta_j] = \arcsin(X^*_{i,j}).$$

We now prove our results generalizing Grothendieck's identity when $p \neq 1/2$, in Lemma 5.K.1. When $p = 1/2$, the special case of Grothendieck's identity is recovered.

Lemma 5.K.1. *Let $p := \mathbb{P}(x_i = 1)$, and X be the SDP solution to 5.C.2. Let also v_i, $i = 1, 2, \ldots, n$ be the column vectors of $X^{1/2}$. Denote the vector returned by Algorithm 3 by ζ. Then,*

$$\operatorname{Cov}(\zeta_i, \zeta_j) = \frac{8p^2}{\pi} \arcsin \langle v_i, v_j \rangle,$$

for $i \neq j$.

Proof. Write $p^i_\xi := \mathbb{P}(\zeta_i = 1 | \langle v_i, \xi \rangle > 0)$. Since $\mathbb{P}(\langle v_i, \xi \rangle > 0) = 0.5$, and since $\mathbb{P}(\zeta_i = 1 | \langle v_i, \xi \rangle \leq 0) = 0$, we know that

$$\begin{aligned} p := \mathbb{P}(\zeta_i = 1) &= \mathbb{P}(\zeta_i = 1 | \langle v_i, \xi \rangle > 0)\mathbb{P}(\langle v_i, \xi \rangle > 0) \\ &\quad + \mathbb{P}(\zeta_i = 1 | \langle v_i, \xi \rangle \leq 0)\mathbb{P}(\langle v_i, \xi \rangle \leq 0) \\ &= p^i_\xi / 2. \end{aligned}$$

Therefore, we have that $p^i_\xi = 2p$ for all i. Define θ to be the angle corresponding to the region of vectors

$\{v_k, v_l\}$ such that $\langle v_k, \xi\rangle > 0$, and $\langle v_l, \xi\rangle > 0$, i.e., $\theta := \{\max_{u,v} \arccos\langle u, v\rangle : \langle u, \xi\rangle > 0, \langle v, \xi\rangle > 0, \xi \in \text{Unif}(S^{n-1})\}$, where S^{n-1} is the $(n-1)$-dimensional sphere. Then, $\theta = \pi - \arccos\langle v_i, v_j\rangle$, and

$$\begin{aligned}\mathbb{P}(\zeta_i \neq \zeta_j) &= \frac{4p(\pi-\theta) + 4p(1-2p)\theta}{2\pi} \\ &= \frac{2p}{\pi}\left(\pi - \theta + (1-2p)\theta\right) \\ &= \frac{2p}{\pi}\left[\arccos\langle v_i, v_j\rangle + (1-2p)(\pi - \arccos\langle v_i, v_j\rangle)\right].\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[\zeta_i\zeta_j] &= 1 - 2\mathbb{P}(\zeta_i \neq \zeta_j) \\ &= 1 - \frac{4p}{\pi}\left[\arccos\langle v_i, v_j\rangle + (1-2p)(\pi - \arccos\langle v_i, v_j\rangle)\right] \\ &= \frac{\pi(1-4p(1-2p)) - 8p^2\arccos\langle v_i, v_j\rangle}{\pi}.\end{aligned}$$

$$\begin{aligned}\text{Cov}(\zeta_i, \zeta_j) &= \frac{\pi(1-4p(1-2p)) - 8p^2\arccos\langle v_i, v_j\rangle}{\pi} - (2p-1)^2 \\ &= \frac{8p^2}{\pi}\arcsin\langle v_i, v_j\rangle.\end{aligned}$$

■

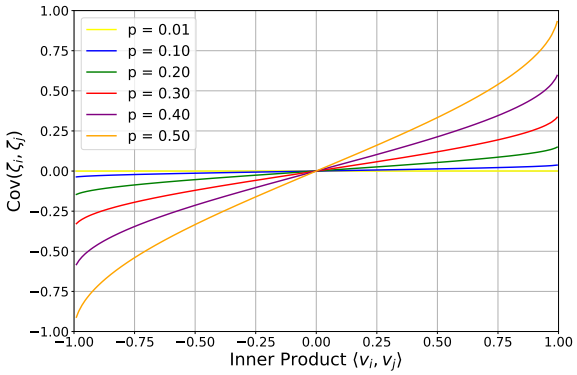


Figure 5.K.1: The relationship between $\text{Cov}(\zeta_i, \zeta_j)$ and vector inner-products v_i, v_j in the SDP relaxation for various treatment assignment probabilities p.

Figure 5.K.1 displays this relationship for various treatment assignment probabilities p.

We first prove Theorem 5.6.1.

Lemma 5.K.2 (Schatten-norm contraction under rounding)**.** *Let*

$$\mathcal{E}_n := \left\{X \in \mathbb{R}^{n\times n} : X \succeq 0,\ \mathrm{diag}(X) = \mathbf{1}\right\},$$

and let $X \in \mathcal{E}_n$. Fix $p \in (0, 1/2]$ and $q \in [1, \infty]$. Let $\zeta \in \{-1, 1\}^n$ be generated from X using either

1. *randomized hyperplane rounding, or*
2. *thresholding,*

with marginal treatment probability

$$\mathbb{P}(\zeta_i = 1) = p, \qquad i \in [n].$$

Let

$$\Xi := \mathrm{Cov}(\zeta)$$

denote the covariance matrix of the rounded assignment, and define its normalized covariance by

$$\overline{\Xi} := \frac{1}{4p(1-p)}\Xi.$$

Then

$$\overline{\Xi} \in \mathcal{E}_n$$

and, for every Schatten q-norm,

$$\|\overline{\Xi}\|_q \le \|X\|_q. \tag{5.K.1}$$

Proof. We first consider the symmetric setting, where $p = 1/2$.

5.K.1.0.1 Symmetric setting Let $C_k := X^{\circ(2k)}$. Then, $C_k \succeq 0$, and $(C_k)_{ii} = X_{ii}^{2k} = 1$, for all i. Therefore, for $q \in [1, \infty]$,

$$\|C_k \circ X\|_q \le \|X\|_q,$$

We can write

$$\|X^{\circ(2k+1)}\|_q = \|C_k \circ X\|_q \leq \|X\|_q.$$

From Lemma 5.K.1, we know that

$$\Xi = \frac{2}{\pi}\arcsin(X) = \frac{2}{\pi}\left(\sum_{k=0}^{\infty}\frac{\binom{2k}{k}}{4^k(2k+1)}X^{\circ(2k+1)}\right).$$

Then, by the triangle inequality, we get

$$\|\Xi\|_q = \left\|\frac{2}{\pi}\left(\sum_{k=0}^{\infty}\frac{\binom{2k}{k}}{4^k(2k+1)}X^{\circ(2k+1)}\right)\right\|_q \leq \frac{2}{\pi}\sum_{k=0}^{\infty}\frac{\binom{2k}{k}}{4^k(2k+1)}\|X^{\circ(2k+1)}\|_q \leq \frac{2}{\pi}\sum_{k=0}^{\infty}\frac{\binom{2k}{k}}{4^k(2k+1)}\|X\|_q = \|X\|_q.$$

For $p < 1/2$, we consider two approaches. The first being the randomized hyperplane rounding, and the second being thresholding-based Gaussian rounding.

5.K.1.0.2 Randomized hyperplane rounding Now suppose that $p < 1/2$. Define, as above,

$$G(X) := \frac{2}{\pi}\arcsin[X].$$

From Lemma 5.K.1, for $i \neq j$, the covariance identity for the randomized hyperplane procedure gives

$$\mathrm{Cov}(\zeta_i, \zeta_j) = 4p^2 G(X)_{ij} = \frac{8p^2}{\pi}\arcsin(X_{ij}).$$

On the diagonal,

$$\mathrm{Var}(\zeta_i) = 4p(1-p).$$

Therefore, at the matrix level,

$$\Xi = 4p^2 G(X) + 4p(1-2p)I. \tag{5.K.2}$$

After normalizing,

$$\overline{\Xi} = \frac{p}{1-p}G(X) + \frac{1-2p}{1-p}I. \tag{5.K.3}$$

The coefficients in (5.K.3) are nonnegative and sum to one. It remains to note that

$$\|I\|_q \le \|X\|_q. \tag{5.K.4}$$

Indeed, if $\lambda_1, \dots, \lambda_n$ are the eigenvalues of X, then $\lambda_i \ge 0$ and

$$\sum_{i=1}^{n} \lambda_i = \mathrm{Tr}(X) = n.$$

For $q < \infty$, Jensen's inequality gives

$$\|X\|_q^q = \sum_{i=1}^{n} \lambda_i^q \ge n \left(\frac{1}{n} \sum_{i=1}^{n} \lambda_i \right)^q = n = \|I\|_q^q.$$

For $q = \infty$,

$$\|X\|_\infty = \max_i \lambda_i \ge \frac{1}{n} \sum_{i=1}^{n} \lambda_i = 1 = \|I\|_\infty.$$

Combining the symmetric-case contraction $\|G(X)\|_q \le \|X\|_q$ with (5.K.3) and (5.K.4), we obtain

$$\begin{aligned} \|\overline{\Xi}\|_q &\le \frac{p}{1-p} \|G(X)\|_q + \frac{1-2p}{1-p} \|I\|_q \\ &\le \left(\frac{p}{1-p} + \frac{1-2p}{1-p} \right) \|X\|_q \\ &= \|X\|_q. \end{aligned}$$

5.K.1.0.3 Gaussian threshold rounding.

Let

$$t_p := \Phi^{-1}(1-p)$$

and define the centered, standardized threshold function

$$f_p(u) := \frac{\mathbf{1}\{u \ge t_p\} - p}{\sqrt{p(1-p)}}.$$

If $\{h_k\}_{k \ge 0}$ denotes the orthonormal Hermite basis of $L^2(\mathcal{N}(0,1))$, then

$$f_p(u) = \sum_{k=1}^{\infty} b_{k,p} h_k(u),$$

where the constant coefficient, i.e., when $k = 0$, vanishes because $\mathbb{E}[f_p(g)] = 0$ for $g \sim \mathcal{N}(0,1)$. See, for example, O'Donnell [2014][Definition 11.34]. Since $\mathbb{E}[f_p(g)^2] = 1$, Plancherel's/Parseval's identity (see, for example, O'Donnell [2014][Proposition 11.36]) yields

$$\sum_{k=1}^{\infty} b_{k,p}^2 = 1. \tag{5.K.5}$$

For jointly standard Gaussian random variables (g_1, g_2) with correlation ρ, orthogonality (see, for example, O'Donnell [2014][Proposition 11.31]) gives

$$\mathbb{E}[h_k(g_1)h_\ell(g_2)] = \mathbf{1}\{k = \ell\}\rho^k.$$

Consequently,

$$\begin{aligned} H_p(\rho) &:= \mathbb{E}[f_p(g_1)f_p(g_2)] \\ &= \sum_{k=1}^{\infty} b_{k,p}^2 \rho^k. \end{aligned} \tag{5.K.6}$$

Equivalently,

$$H_p(\rho) = \frac{\overline{\Phi}_2(t_p, t_p; \rho) - p^2}{p(1-p)},$$

where

$$\overline{\Phi}_2(t_p, t_p; \rho) := \mathbb{P}(g_1 \geq t_p, g_2 \geq t_p).$$

Under threshold rounding,

$$\frac{\zeta_i - \mathbb{E}[\zeta_i]}{2\sqrt{p(1-p)}} = f_p(\xi_i).$$

It follows that

$$\overline{\Xi}_{ij} = H_p(X_{ij}),$$

and hence, by (5.K.6),

$$\overline{\Xi} = H_p[X] = \sum_{k=1}^{\infty} b_{k,p}^2 X^{\circ k}. \tag{5.K.7}$$

For every $k \geq 1$, define

$$D_{k-1} := X^{\circ(k-1)},$$

with the convention $D_0 = \mathbf{1}\mathbf{1}^\top$. By the Schur product theorem,

$$D_{k-1} \in \mathcal{E}_n,$$

and

$$X^{\circ k} = D_{k-1} \circ X.$$

Thus, by Lemma 5.K.2,

$$\|X^{\circ k}\|_q \le \|X\|_q.$$

Using (5.K.5) and (5.K.7),

$$\begin{aligned}\|\overline{\Xi}\|_q &\le \sum_{k=1}^{\infty} b_{k,p}^2 \|X^{\circ k}\|_q \\ &\le \sum_{k=1}^{\infty} b_{k,p}^2 \|X\|_q \\ &= \|X\|_q.\end{aligned}$$

The Hermite expansion is first obtained for correlations $\rho \in (-1,1)$; the cases $\rho = \pm 1$ follow by continuity. This completes the proof for both rounding procedures. ■

Theorem 5.K.3 (Approximation bound)**.** *Fix $p \in (0, 1/2]$ and define*

$$\mathcal{C}_{n,p} := \left\{ \frac{\mathrm{Cov}(\zeta)}{4p(1-p)} : \zeta \in \{-1,1\}^n,\ \mathbb{P}(\zeta_i = 1) = p \text{ for every } i \right\}.$$

Define

$$\mathcal{E}_n := \left\{ X \in \mathbb{S}_+^n : X_{ii} = 1 \text{ for all } i \in [n] \right\} = \left\{ X \in \mathbb{R}^{n\times n} : X = X^\top,\ X \succeq 0,\ \mathrm{diag}(X) = \mathbf{1}_n \right\}, \tag{5.K.8}$$

Then

$$\mathcal{C}_{n,p} \subseteq \mathcal{E}_n.$$

Let $A_n \succeq 0$ and $\widetilde{\kappa}_n > 0$, and define

$$J_n(X) := \mathrm{Tr}(A_n X) + \widetilde{\kappa}_n \|X\|_q,$$

where $q^\star$ denotes the Schatten dual exponent:

$$\frac{1}{q} + \frac{1}{q^\star} = 1.$$

In the design problem considered here, one may take, for example,

$$A_n = \widetilde{\eta}_n L^\dagger + \widetilde{\gamma}_n L + \mathbf{1}\mathbf{1}^\top.$$

Let

$$X_{\mathrm{SDP}} \in \arg\min_{X \in \mathcal{E}_n} J_n(X),$$

and let

$$X^\star \in \arg\min_{X \in \mathcal{C}_{n,p}} J_n(X)$$

be an optimal achievable normalized design covariance. Apply either randomized hyperplane rounding with marginal adjustment or Gaussian threshold rounding to X_{SDP}, and let

$$\Xi_{\mathsf{R}} := \mathrm{Cov}(\zeta_{\mathsf{R}}), \qquad \overline{\Xi}_{\mathsf{R}} := \frac{\Xi_{\mathsf{R}}}{4p(1-p)}.$$

Define the approximation factor

$$c_{\mathsf{R}} := \frac{J_n(\overline{\Xi}_{\mathsf{R}})}{J_n(X_{\mathrm{SDP}})}.$$

Then

$$J_n(\overline{\Xi}_{\mathsf{R}}) \le c_{\mathsf{R}} J_n(X^\star), \tag{5.K.9}$$

and

$$1 \le c_{\mathsf{R}} \le 1 + \frac{\|A_n\|_{q^\star}}{\widetilde{\kappa}_n}. \tag{5.K.10}$$

If, in addition, $A_n \succ 0$, then

$$c_{\mathsf{R}} \le \frac{n\lambda_{\max}(A_n) + \widetilde{\kappa}_n \|X_{\mathrm{SDP}}\|_q}{n\lambda_{\min}(A_n) + \widetilde{\kappa}_n \|X_{\mathrm{SDP}}\|_q} \le \frac{\lambda_{\max}(A_n)}{\lambda_{\min}(A_n)}. \tag{5.K.11}$$

Consequently,

$$c_{\mathsf{R}} \le \min\left\{1 + \frac{\|A_n\|_{q^\star}}{\widetilde{\kappa}_n}, \frac{\lambda_{\max}(A_n)}{\lambda_{\min}(A_n)}\right\},$$

whenever $A_n \succ 0$. *Equivalently, if*

$$\Sigma^\star := 4p(1-p)X^\star$$

denotes the corresponding unnormalized optimal covariance, then

$$J_n(\Xi_{\mathsf{R}}) \le c_{\mathsf{R}} J_n(\Sigma^\star).$$

Proof. Because every normalized covariance matrix with marginal treatment probability p is positive semidefinite and has unit diagonal,

$$\mathcal{C}_{n,p} \subseteq \mathcal{E}_n.$$

Moreover, by construction,

$$\overline{\Xi}_{\mathsf{R}} \in \mathcal{C}_{n,p}.$$

Optimality of X_{SDP} over $\mathcal{E}_n$ therefore gives

$$J_n(X_{\mathrm{SDP}}) \le J_n(\overline{\Xi}_{\mathsf{R}}),$$

and hence $c_{\mathsf{R}} \ge 1$. Since $X^\star \in \mathcal{C}_{n,p} \subseteq \mathcal{E}_n$, SDP optimality also gives

$$J_n(X_{\mathrm{SDP}}) \le J_n(X^\star).$$

It follows that

$$\begin{aligned} J_n(\overline{\Xi}_{\mathsf{R}}) &= c_{\mathsf{R}} J_n(X_{\mathrm{SDP}}) \\ &\le c_{\mathsf{R}} J_n(X^\star), \end{aligned}$$

which proves (5.K.9).

Next, Hölder inequality gives

$$\mathrm{Tr}\left(A_n \overline{\Xi}_{\mathsf{R}}\right) \le \|A_n\|_{q^\star} \|\overline{\Xi}_{\mathsf{R}}\|_q.$$

By Lemma 5.K.2,

$$\|\overline{\Xi}_{\mathsf{R}}\|_q \le \|X_{\mathrm{SDP}}\|_q.$$

Therefore,

$$\begin{aligned} J_n(\overline{\Xi}_{\mathsf{R}}) &= \operatorname{Tr}\left(A_n\overline{\Xi}_{\mathsf{R}}\right) + \widetilde{\kappa}_n\|\overline{\Xi}_{\mathsf{R}}\|_q \\ &\leq \left(\|A_n\|_{q^\star} + \widetilde{\kappa}_n\right)\|X_{\mathrm{SDP}}\|_q. \end{aligned}$$

On the other hand,

$$J_n(X_{\mathrm{SDP}}) \geq \widetilde{\kappa}_n\|X_{\mathrm{SDP}}\|_q.$$

Consequently,

$$\begin{aligned} c_{\mathsf{R}} &= \frac{J_n(\overline{\Xi}_{\mathsf{R}})}{J_n(X_{\mathrm{SDP}})} \\ &\leq \frac{\left(\|A_n\|_{q^\star} + \widetilde{\kappa}_n\right)\|X_{\mathrm{SDP}}\|_q}{\widetilde{\kappa}_n\|X_{\mathrm{SDP}}\|_q} \\ &= 1 + \frac{\|A_n\|_{q^\star}}{\widetilde{\kappa}_n}. \end{aligned}$$

This proves (5.K.10). For the second bound, suppose that $A_n \succ 0$. Since both $\overline{\Xi}_{\mathsf{R}}$ and X_{SDP} belong to $\mathcal{E}_n$,

$$\operatorname{Tr}(\overline{\Xi}_{\mathsf{R}}) = \operatorname{Tr}(X_{\mathrm{SDP}}) = n.$$

Therefore,

$$\operatorname{Tr}\left(A_n\overline{\Xi}_{\mathsf{R}}\right) \leq n\lambda_{\max}(A_n),$$

whereas

$$\operatorname{Tr}\left(A_nX_{\mathrm{SDP}}\right) \geq n\lambda_{\min}(A_n).$$

Using again

$$\|\overline{\Xi}_{\mathsf{R}}\|_q \leq \|X_{\mathrm{SDP}}\|_q,$$

we obtain

$$J_n(\overline{\Xi}_{\mathsf{R}}) \leq n\lambda_{\max}(A_n) + \widetilde{\kappa}_n\|X_{\mathrm{SDP}}\|_q$$

and

$$J_n(X_{\mathrm{SDP}}) \geq n\lambda_{\min}(A_n) + \widetilde{\kappa}_n\|X_{\mathrm{SDP}}\|_q.$$

Hence

$$c_{\mathsf{R}} \leq \frac{n\lambda_{\max}(A_n) + \widetilde{\kappa}_n\|X_{\mathrm{SDP}}\|_q}{n\lambda_{\min}(A_n) + \widetilde{\kappa}_n\|X_{\mathrm{SDP}}\|_q} \leq \frac{n\lambda_{\max}(A_n)}{n\lambda_{\min}(A_n)}.$$

Finally,

$$J_n(\Xi_{\mathsf{R}}) = 4p(1-p)J_n(\overline{\Xi}_{\mathsf{R}})$$

and

$$J_n(\Sigma^\star) = 4p(1-p)J_n(X^\star).$$

Multiplying (5.K.9) by $4p(1-p)$ proves the corresponding unnormalized statement.

■

We now prove Theorem 5.K.4.

Theorem 5.K.4 (Approximation bound)**.** *Let $\Xi = \mathrm{Cov}(\zeta)$ be the covariance matrix of the associated Rademacher random variables. Let X be the SDP solution. We assume that X is non-degenerate, i.e., $X \succeq \delta I$, $\delta > 0$. For $c_\kappa = \left(\frac{8p^2}{\pi} + \frac{8p^2\psi_\kappa}{\pi\delta} + \frac{4p(1-p)-8p^2/\pi}{\delta}\right)$ for $\psi_\kappa = n\alpha_\kappa^3$, where α_κ is the maximum off-diagonal entry of X,*

$$\Xi \preceq c_\kappa X.$$

Proof. Recall that for Rademacher random variables $\zeta_i \in \{-1, 1\}$ with $\mathbb{P}(\zeta_i = 1)$, $\mathrm{Var}(\zeta_i) = 4p(1-p)$. Therefore, we know that $\Xi_{ii} = 4p(1-p)$, and what is left is to consider the off-diagonals. From our derivation above for (i, j) pairs, we know that for the off-diagonals,

$$\Xi = \mathbb{E}[\zeta\zeta^T] - \mu\mu^T = \frac{8p^2}{\pi}\arcsin(X),$$

where $\arcsin(X)$ is written to be the entry-wise function on the matrix X, with Taylor expansion,

$$\arcsin(X) = X + \sum_{k=1}^{\infty} \frac{\binom{2k}{k}}{4^k(2k+1)} X^{\circ(2k+1)}.$$

Let α_κ be the maximum entry in the off-diagonals $X - I$. Here we note the dependence on κ; by Proposition 5.4.7, as κ increases in the objective function relative to the other tradeoff parameters γ, η, α_κ decreases. Let also $f(X) := \sum_{k=1}^{\infty} \frac{\binom{2k}{k}}{4^k(2k+1)} X^{\circ(2k+1)}$, where $X^{\circ\beta}$ denotes all entries of X being raised to the power of β.

Denote the off-diagonal matrix of X by $\tilde{R}$. Therefore,

$$\begin{aligned}\Xi &= \frac{8p^2}{\pi}(\arcsin(X)) + 4p(1-2p)I \\ &= 4p^2((\frac{2}{\pi}\arcsin(X) - I) + I) + 4p(1-2p)I\end{aligned}$$

$$= \frac{8p^2}{\pi}(\tilde{R} + \mathcal{O}(\tilde{R}^{\circ 3})) + 4p(1-p)I,$$
$$= \frac{8p^2}{\pi}X + \frac{8p^2}{\pi}\mathcal{O}(\tilde{R}^{\circ 3}) + (4p(1-p) - \frac{8p^2}{\pi})I.$$

Let $\psi_\kappa = n\alpha_\kappa^3$, then $\|\mathcal{O}(\tilde{R}^{\circ 3})\|_{\text{op}} \leq \psi_\kappa$, and $\tilde{R}^{\circ 3} \preceq \frac{\psi_\kappa}{\delta}X$. In particular, for κ large, if $\alpha_\kappa = \mathcal{O}(n^{-1/3})$, then $\psi_\kappa = \mathcal{O}(1)$. It is easy to then see that

$$\left(\frac{8p^2}{\pi} + \frac{8p^2\psi_\kappa}{\pi\delta} + \frac{4p(1-p) - 8p^2/\pi}{\delta}\right)X \succeq \Xi.$$

■

We now prove Theorem 5.6.2 as a corollary. First, we restate it below for general p.

Theorem 5.6.2 (Gaussian hyperplane rounding MSE approximation bound)**.** *Let $\Xi := \text{Cov}(\zeta)$ be the covariance matrix of the rounded assignment returned by Algorithm 3 with $R = HR$, and let X_{SDP} be the solution to the SDP in 5.C.2. Assume that X_{SDP} is non-degenerate, i.e., $X_{SDP} \succeq \delta_\kappa I$, $\delta_\kappa > 0$. Let $c_\kappa = \frac{1}{4p(1-p)}\left(\frac{8p^2}{\pi} + \frac{8p^2\psi_\kappa}{\pi\delta_\kappa} + \frac{4p(1-p)-8p^2/\pi}{\delta_\kappa}\right)$, $\psi_\kappa = n\alpha_\kappa^3$, where α_κ is the maximum off-diagonal entry of X_{SDP} in absolute value. Define*

$$J(X) := 7\{\text{Tr}((\gamma\beta_1^2 L + \eta[\beta_1^2 + \beta_2^2]L^\dagger)X) + \kappa[\beta_1^2 + \beta_2^2]\|X\|_q$$
$$+ ([a\beta_1 + b\beta_2]^2 + \beta_1^2\delta)\,\text{Tr}(\mathbf{1}\mathbf{1}^T X)\},$$

i.e., the right-hand side of Equation (5.C.1). Let X^* denote the optimal achievable covariance matrix, i.e., one minimizing $J(X)$ over all covariance matrices achievable by a design. Then, for the estimator $\hat{\tau}_\zeta$ under the design associated with ζ, $\sup_{\chi(\eta,\gamma,\kappa)} \mathbb{E}[(\hat{\tau}_\zeta - \tau)^2] \leq J(\Xi) \leq c_\kappa J(X^*)$.

Proof. We see that $L, L^\dagger, \mathbf{1}\mathbf{1}^T$ are PSD. It is enough to notice that for PSD matrices A, B, C, with $B \succeq C$, $\text{Tr}(A(B-C)) \geq 0$. Finally, notice that by Theorem 5.K.4, for $X = 4p(1-p)X_{\text{SDP}}$, we have $J(\Xi) \leq c_\kappa J(X^*)$. ■

Proposition 5.K.5 (Harshaw et al., 2024, Theorem 6.4)**.** *The Gram-Schmidt Walk design with parameter $\lambda \in [0,1]$ provides balance-robustness guarantee $\|\text{Cov}(Ax)\|_\infty \leq \frac{\max_{i\in[n]}\|x_i\|_2^2}{1-\lambda}$ and $\|\text{Cov}(x)\|_\infty \leq \frac{1}{\lambda}$.*

We now state and prove the general version of Theorem 5.6.3, allowing for $p \neq 1/2$.

Theorem 5.K.6. *Let* $A := (\tilde{\eta}L^\dagger + \tilde{\gamma}L + \Delta\mathbf{1}\mathbf{1}^T)^{1/2}$, *where* $\tilde{\eta} = \eta[\beta_1^2 + \beta_2^2]$, $\tilde{\gamma} = \gamma\beta_1^2$, $\tilde{\kappa} = \kappa[\beta_1^2 + \beta_2^2]$, *and* $\Delta = (a\beta_1 + b\beta_2)^2$. *The Gram-Schmidt Walk algorithm, with randomization parameter* λ, *returns an assignment* x *such that* $\|\operatorname{Cov}(Ax)\|_{p'} \le \omega_A^2/(1-\lambda)\|I\|_{p'}$, *and* $\|\operatorname{Cov}(x)\|_q \le \|I\|_q/\lambda = n^{1/q}/\lambda$, *for* $\omega_A := \max_{i\in n}\|A_i\|$, A_i *the column* i *of the matrix* A. *Thus, for* $q \ge 1$,

$$\sup_{\chi(\eta,\gamma,\kappa)} \mathbb{E}[(\tau - \hat{\tau})^2] \le 7\left[\mathbb{E}[\|A(x-\mu)\|^2] + \tilde{\kappa}\|X\|_q\right] \le 7\left[\frac{\omega_A^2 n}{1-\lambda} + \tilde{\kappa}\frac{n^{1/q}}{\lambda}\right],$$

where $\mu_i = \mathbb{E}[x_i]$, *for all* $i \in [n]$. *This is minimized for* $\lambda = \frac{\sqrt{\tilde{\kappa}n^{1/q}}}{\sqrt{\omega_A^2 n} + \sqrt{\tilde{\kappa}n^{1/q}}}$.

Proof. From Theorem 5.C.1, we have

$$\sup_{\chi(\eta,\gamma,\kappa)} \mathbb{E}[(\tau - \hat{\tau})^2] \le 7\left[\operatorname{Tr}(XAA^T) + \tilde{\kappa}\|X\|_q\right] = 7\left[\mathbb{E}[\|A(x-\mu)\|^2 + \tilde{\kappa}\|X\|_q\right]$$

From the proof of Harshaw et al. [2024, Theorem 6.4], $\operatorname{Cov}(x) \preceq \lambda^{-1}I$, and (2) $\operatorname{Cov}(Ax) \preceq \omega_A^2(1-\lambda)^{-1}I$. Thus, from (1) we have $\|X\|_q \le \lambda^{-1}\|I\|_q = \lambda^{-1}n^{1/q}$. From (2), $\operatorname{Tr}(XAA^T) \le \omega_A^2(1-\lambda)^{-1}\,\mathrm{I} = \omega_A^2(1-\lambda)^{-1}n$. Therefore,

$$7\left[\operatorname{Tr}(XAA^T) + \tilde{\kappa}\|X\|_q\right] \le 7\left[\frac{\omega_A^2 n}{(1-\lambda)} + \tilde{\kappa}\frac{n^{1/q}}{\lambda}\right].$$

Next, to minimize the right-hand side above, we take the derivative with respect to λ and set it equal to zero. Let $m = \omega_A^2 n(1-\lambda)^{-1} + \tilde{\kappa}n^{1/q}\lambda^{-1}$. Then,

$$\frac{dm}{d\lambda} = \frac{\omega_A^2 n}{(1-\lambda)^2} - \frac{\tilde{\kappa}n^{1/q}}{\lambda^2} = 0$$
$$\lambda = \frac{\sqrt{\tilde{\kappa}n^{1/q}}}{\sqrt{\tilde{\kappa}n^{1/q}} \pm \sqrt{\omega_A^2 n}}.$$

Since $\lambda \in [0,1]$, $\lambda = \frac{\sqrt{\tilde{\kappa}n^{1/q}}}{\sqrt{\tilde{\kappa}n^{1/q}} + \sqrt{\omega_A^2 n}}$. Additionally, since $\omega_A^2 n(1-\lambda)^{-1} + \tilde{\kappa}n^{1/q}\lambda^{-1}$ is convex in λ, $\lambda = \frac{\sqrt{\tilde{\kappa}n^{1/q}}}{\sqrt{\tilde{\kappa}n^{1/q}} + \sqrt{\omega_A^2 n}}$ is its minimizer, yielding the claim. ∎

The result above also holds when $A^T := [\tilde{\eta}^{1/2}L^{\dagger/2} \quad \tilde{\gamma}^{1/2}L^{1/2} \quad \Delta^{1/2}\mathbf{1}\mathbf{1}^{T\,1/2}]$.

Table 5.2: Key of designs compared in this paper.

Design	**Description**
SDP	SDP solution X_{SDP}
SDP + GaussRound	SDP followed by Gaussian rounding
SDP + QuantRound	SDP followed by quantile rounding (see Section 5.E.1.2)
GSW	Adapted Gram–Schmidt Walk design [Harshaw et al., 2024]
GSW + covariates	Adapted Gram–Schmidt Walk design [Harshaw et al., 2024] using covariate encoding directly instead of $L^\dagger$
Louvain	Cluster-randomization with default Python implementation of the Louvain clustering algorithm
ϵ-net	Cluster-randomization under (ϵ=1)-net clustering [Ugander et al., 2013]
Spectral	Cluster-randomization under spectral clustering
Causal	Cluster-randomization under Causal clustering [Viviano et al., 2023] with a grid-search for optimal hyperparameters
Bernoulli	Uniform Bernoulli randomization

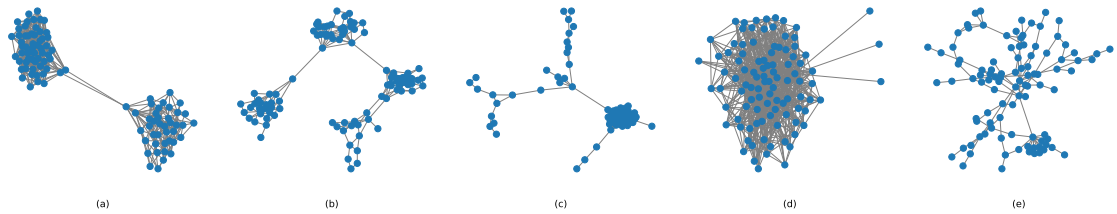


Figure 5.L.1: From left to right: Graphs generated from (a): SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$), with two clusters of equal membership assignment probability; (b): SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$) with four clusters of equal membership assignment probability; (c): SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$) with four clusters of membership assignment probabilities 0.7, 0.1, 0.1, 0.1; (d): Barabasi-Albert (preferential-attachment) with initial $n/10$ (=10) nodes connected according to the Erdös-Rényi model with connection probability $10/n$; Graph 5: Erdös-Rényi with connection probability $2/n$.

5.L Simulations

5.L.1 Real Data

Figure 5.2.1 illustrates a single village network in rural Karnataka, South India [Banerjee et al., 2013a,b]. For context, Banerjee et al. [2013a] collected network data from 75 villages before the introduction of financial services in these villages by a microfinance institute. In particular, in 2006, six months before the microfinance institute began their work, Banerjee et al. [2013a] conducted surveys on subsamples of villagers gathering demographic information. The survey included a portion on social network data along 12 dimensions. In Figure 5.2.1, we display Village No.6 which we had randomly selected. In this network, the nodes represent village members, and an edge between any pair of nodes is present if the corresponding villagers visit each other's homes. In this figure, the nodes in the network are colored according to the castes of the corresponding villagers, illustrating homophily in the network, as members of the same caste are more connected.

In Figure 5.L.2, we illustrate treatment assignments under various cluster-randomized designs, depicting almost complete overlap with the underlying homophily structure. In the worst case where the distinct homophilous groups have heterogeneous treatment effects, such cluster-randomized designs can lead to imprecise estimates of the GATE. Generally, this imprecision worsens as the number of distinct homophilous groups increases for a fixed population size.

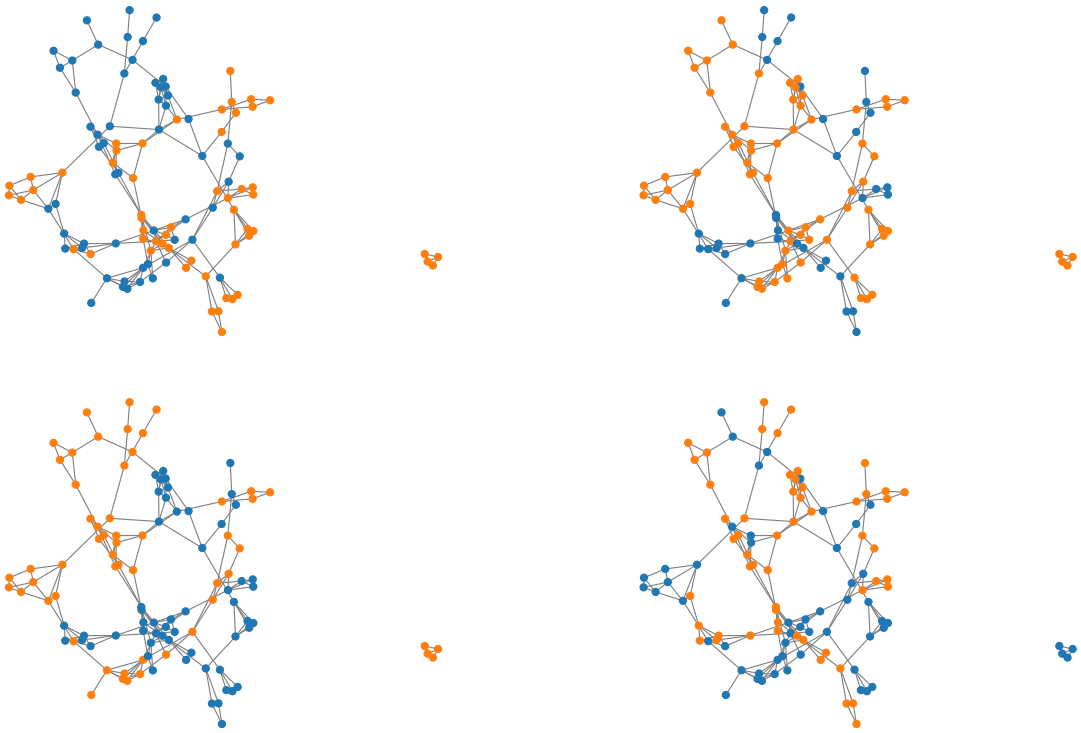


Figure 5.L.2: Assignments under various cluster randomization designs. Top (left to right): $\epsilon = 1$-net, causal cluster randomizations. Bottom (left to right): Louvain, and spectral cluster randomizations.

In the comparisons of the worst-case MSE bounds on this network data in Figures 5.L.3 and 5.L.5, we compare the bounds derived in Theorem 5.C.1 using covariance matrices from various designs. We describe the different designs we compare in Table 5.2. We note that the different cluster-randomized designs were also considered in [Viviano et al., 2023].

We also compare the MSEs by combining synthetic worst-case potential outcomes with the real network data together with the caste information. We generate the homophily components in the potential outcomes, h^α and h^ϕ from a random effects model across the castes. That is, we generate potential outcomes associated with each caste j

$$\nu_{jf} \sim \mathcal{N}(\mu_f, \sigma_f^2),$$

for some μ_f. Then, for each node i that belongs to caste j,

$$h_{if} = \nu_{jf},$$

for all j, and $f = \alpha, \phi$. In our simulations, we take $\mu_\alpha = 2$, $\mu_\phi = 3$, and $\sigma_f^2 = 3$, $f = \alpha, \phi$. Then, we take $\eta = {h_f}^T L h_f$.

To generate the heterogeneous variation components ϵ_f, for $f = \alpha, \phi$, for $q^* = 2$, we take

$$\begin{aligned}
\Sigma &\sim \mathcal{N}(0,1)^{n\times n},\\
X &= \Sigma/\|\Sigma\|_2,\\
z &\sim \mathcal{N}(0, I_n),\\
\epsilon_f &= \sqrt{\kappa} X^{1/2} z.
\end{aligned}$$

If $q^* = \infty$ instead, we replace X with $X = vv^T$, where v is the leading eigenvector of Σ.

To generate the interference component for each treatment assignment vector $x = 2z - 1$ generated from X from the corresponding covariance matrix, we take

$$\begin{aligned}
s(z) &= \frac{\sqrt{\gamma} L x}{\|x^T L^{1/2}\|_2},\\
s(z) &= s(z) - \left\langle s(z), \frac{\mathbf{1}}{n} \right\rangle \mathbf{1}.
\end{aligned}$$

We detail in Appendix 5.L.3 why these generative models appropriately capture worst-case potential outcomes.

In Figure 5.L.4, we compare the MSEs across various designs, under the generated worst-case potential outcomes, for various levels of γ and κ.

In Figure 5.L.5, in addition to the comparisons with the adapted Gram-Schmidt Walk design using the worst-case bounds with $\eta = {h_f}^T L h_f$, we also compare with the adapted Gram-Schmidt Walk design directly using a one-hot encoding of caste information (see the plotted line labeled "GSW + covariates"), with potential outcomes generated according to the worst-case models described above. Specifically, let J be the number of different covariate groups. In our village data example, J is the number of castes. Given the random effects model above we can write the homophily component of the MSE as

$$\begin{aligned}
\mathbb{E}[(h^T x)^2] &= \mathbb{E}\left[\left(\sum_{j=1}^{J}\left(\sum_{i:x_i=1}\mathbf{1}\{i\in\text{Caste}_j\}-\sum_{i:x_i=-1}\mathbf{1}\{i\in\text{Caste}_j\}\right)\nu_j\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{J}\left(n_{1j}-n_{0j}\right)\nu_j\right)^2\right] \\
&= \sum_{j=1}^{J}(n_{1j}-n_{0j})^2\,\text{Var}(\nu_j) = \sigma^2\sum_{j=1}^{J}(n_{1j}-n_{0j})^2.
\end{aligned}$$

Then, in this setting, we define $\eta = \sigma^2$, and take our pre-augmented covariate matrix $A \in \mathbb{R}^{(J+2n)\times n}$ to be such that its i-th column,

$$A_i^T = [\tilde{\eta}^{1/2}\xi_i^j \quad \tilde{\gamma}^{1/2}L_i^{1/2} \quad (\mathbf{1}\mathbf{1}^T)_i^{1/2}]^T,$$

where ξ_i^j is column vector with entry 1 in row j if unit i belongs to caste j and zero otherwise, and the parameters $\tilde{\eta}, \tilde{\gamma}, \tilde{\kappa}$ defined as before. Then, the worst case MSE bound can be written in the previous form

$$\mathbb{E}[(\hat{\tau}-\tau)^2] \le 7\Delta[\mathbb{E}[\|Ax\|_2^2] + \tilde{\kappa}\|X\|_q].$$

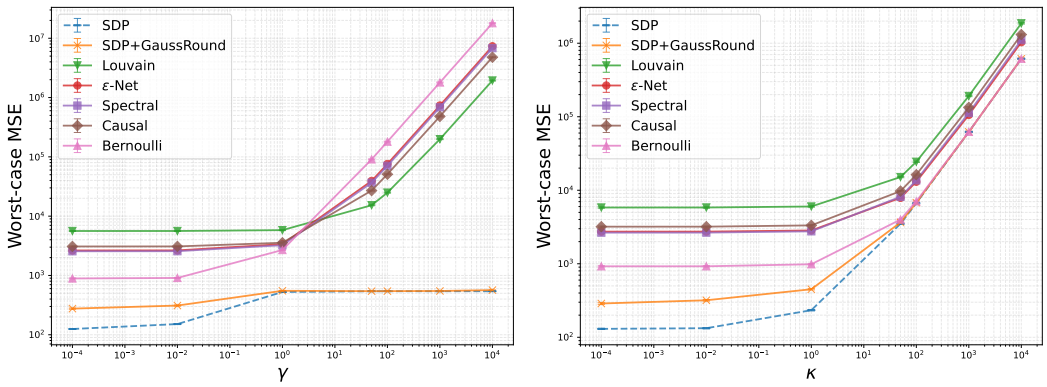


Figure 5.L.3: Comparisons of the worst-case MSE bounds across various designs, for different levels of γ and κ, with the average (over 1000 trials) $\eta \approx 196$. Error bars represent the standard error.

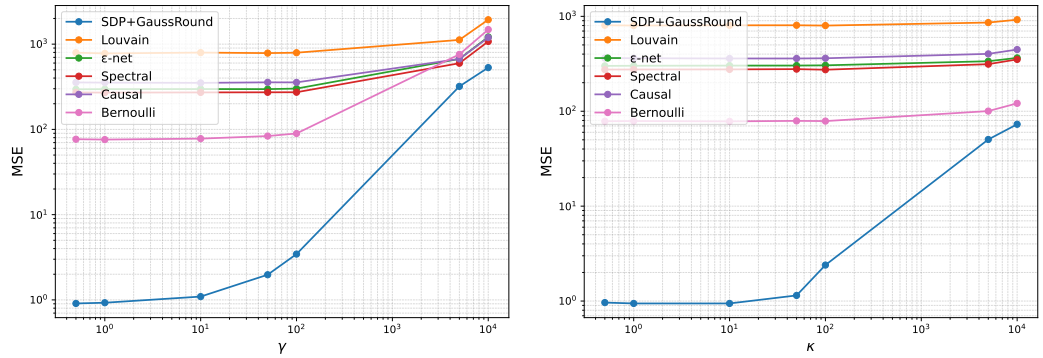


Figure 5.L.4: Comparisons of the MSEs across various designs, for different levels of γ and κ, with the average (over 1000 trials) $\eta \approx 196$. Error bars represent the standard error.

5.L.2 Worst-case MSE bounds comparisons on simulated networks

Throughout this subsection and the next, we compare the performance of our designs against various designs from the literature. We describe the different designs we compare in Table 5.2. We note that the different cluster-randomized designs were also considered in [Viviano et al., 2023].

In Figures 5.L.6 to 5.L.10, we compare the worst-case bounds by plugging into the bound in Theorem 5.C.1 the covariance matrices from various designs. We compare the worst-case MSE bounds for five different graph generation models, as described in Figure 5.L.1, averaged across 1000 graph instances of each graph model. For procedures where we cannot directly compute the design

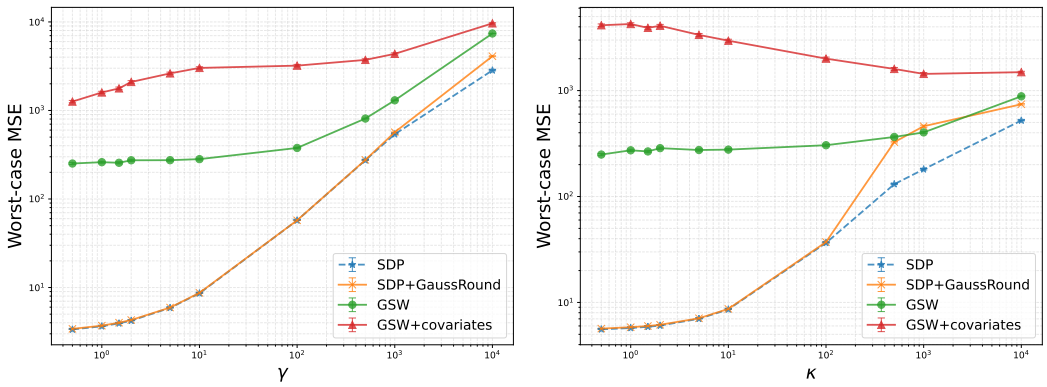


Figure 5.L.5: Comparisons of the worst-case MSE bounds across various designs, with the average (over 1000 trials) $\eta \approx 196$. Error bars represent the standard error.

covariance, we generate the sample covariance matrices of the designs over 1000 instances of the design vectors for each graph instance. We note that the gap between the optimal solution from solving the SDP and the other designs are due to unattainability of the optimal SDP solution given that the SDP formulation is a relaxation of the problem. In these settings, we set the fixed tradeoff parameters to be 1.

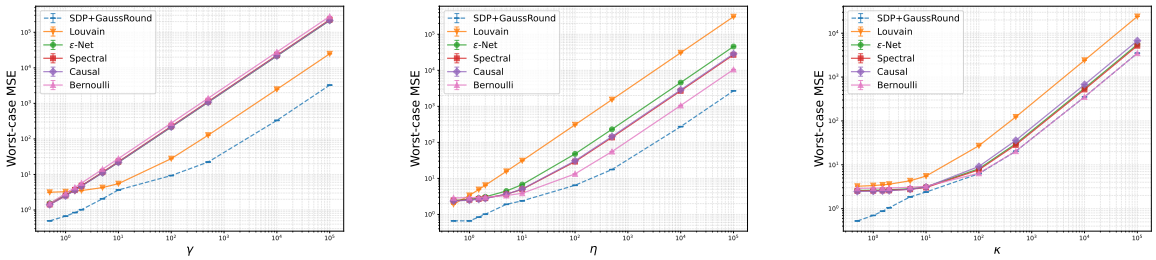


Figure 5.L.6: Comparison of worst-case MSE bounds between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$), with two clusters of equal membership assignment probability (see Graph 1 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

Figures 5.L.11 to 5.L.15 compare worst-case MSE bounds using our SDP approach followed by Gaussian rounding, and our adapted Gram-Schmidt Walk approach, across various n. Here, we set $\kappa = 100, \eta = 250, \gamma = 1000, q = 2$, and $\Delta = 0.01\gamma$. We observe that the adapted Gram-Schmidt

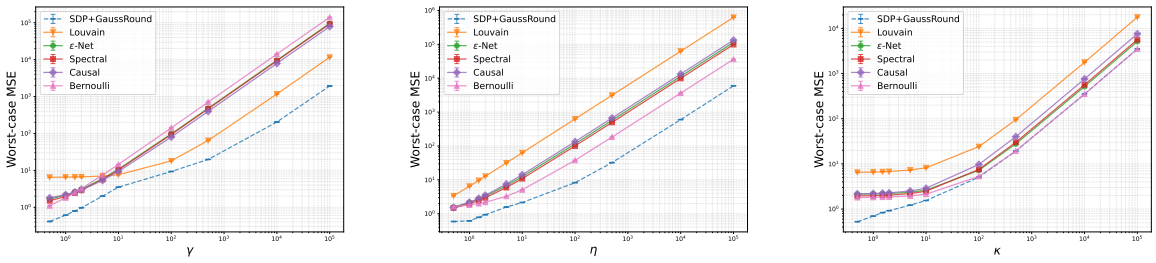


Figure 5.L.7: Comparison of worst-case MSE bounds between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$), with four clusters of equal membership assignment probability (see Graph 2 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

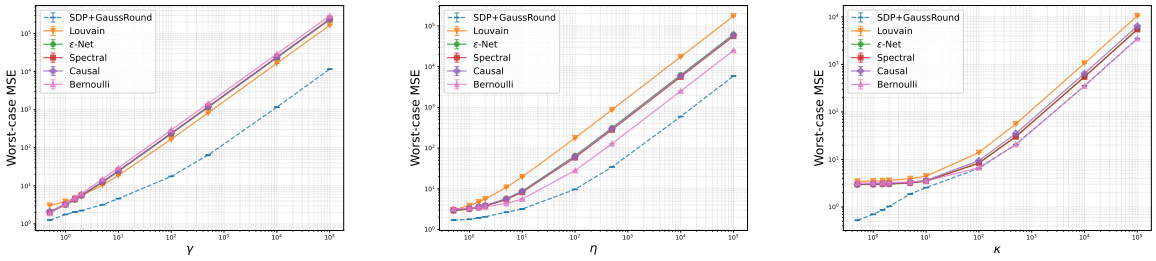


Figure 5.L.8: Comparison of worst-case MSE bounds between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$) with four clusters of membership assignment probability (0.70, 0.1, 0.1, 0.1) (see Graph 3 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

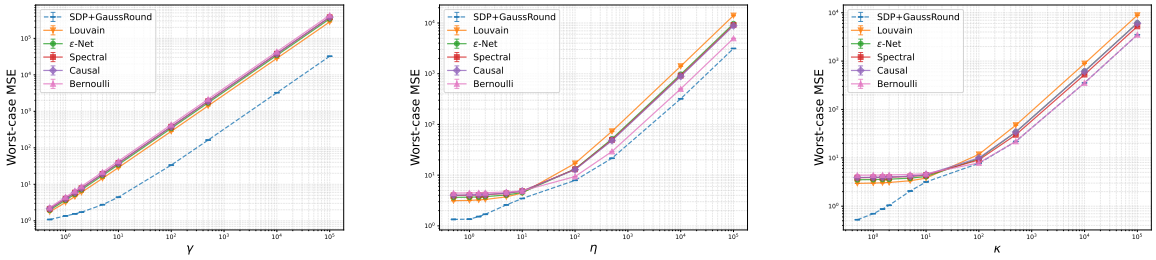


Figure 5.L.9: Comparison of worst-case MSE bounds between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from a Barabasi-Albert (preferential-attachment) model with initial $n/10$ (=10) nodes connected according to the Erdös-Rényi model with connection probability $10/n$ (see Graph 4 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

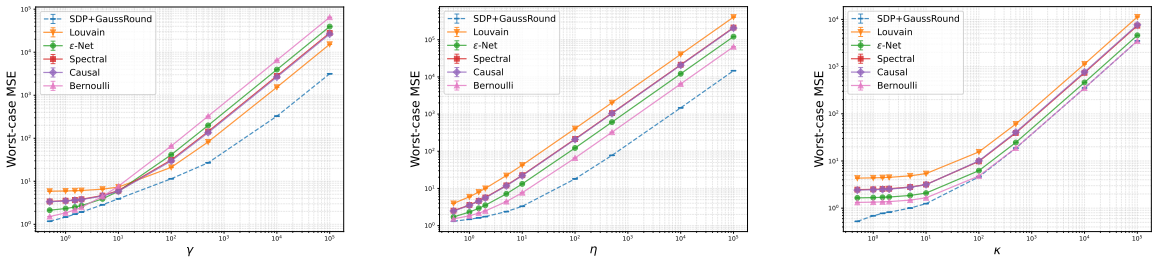


Figure 5.L.10: Comparison of worst-case MSE bounds between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from an Erdös-Rényi model with connection probability $2/n$ (see Graph 5 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

Walk designs have comparable worst-case error bounds to the SDP solutions, for n ranging between 200 and 400. The SDP solutions followed by Gaussian rounding perform worse than the adapted Gram-Schmidt Walk designs for settings of types Graphs 2 – 5 in Figure 5.L.1, while in settings of type Graph 1 in Figure 5.L.1, the Gaussian rounding procedure almost perfectly approximates the SDP solution.

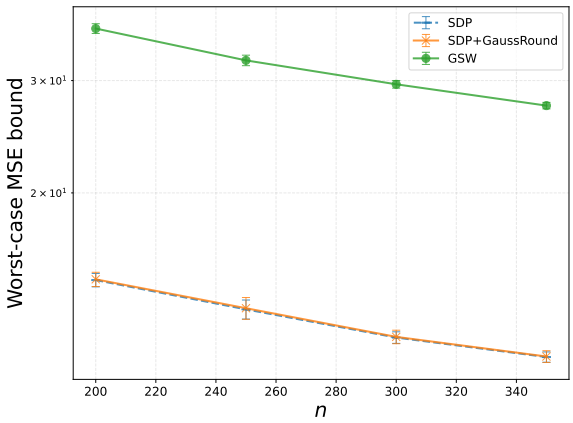


Figure 5.L.11: Comparisons with adapted Gram-Schmidt algorithm averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$), with two clusters of equal membership assignment probability (see Graph 1 in Figure 5.L.1), for different n. Error bars represent the standard error.

For completeness, we also compare the MSEs under simulated worst-case potential outcomes on various graphs. We present the results of these experiments in the next section.

5.L.3 MSE comparisons under synthetic worst-case potential outcomes

In Figures 5.L.16 to 5.L.20, we display the MSE using simulated potential outcomes across various designs, namely our design, cluster-randomization under the default Python implementation of the Louvain clustering, ($\varepsilon = 1$)-net clustering [Ugander et al., 2013], Spectral clustering, Causal Clustering [Viviano et al., 2023], and Bernoulli Randomization. We compare the MSEs for five different graph generation models, averaged across 1000 graph instances of each graph model, as described in Figure 5.L.1. In these settings, we set the fixed tradeoff parameters to be 1, and $a, b = 0.01\gamma$. We display the results in Figures 5.L.16 to 5.L.20.

The worst-case potential outcomes were generated according to the following model. For given

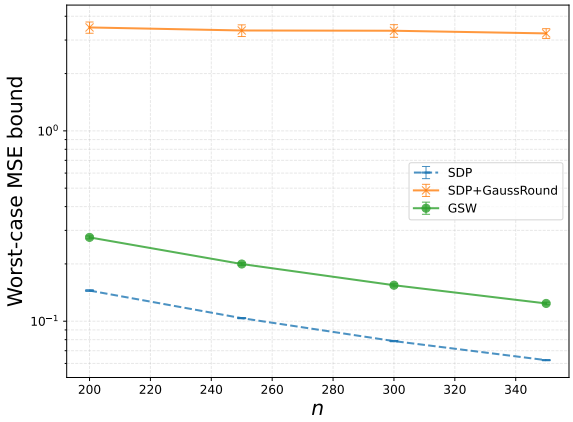


Figure 5.L.12: Comparisons with adapted Gram-Schmidt algorithm averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$), with four clusters of equal membership assignment probability (see Graph 2 in Figure 5.L.1), for different n. Error bars represent the standard error.

$L, L^\dagger, \eta, \gamma, \kappa$, and randomly generated baselines α_0, ϕ_0 (and, thus, Δ), we compute the covariance matrices X corresponding to the different designs described above. Then, adversarially, for the worst-case, we compute the potential outcome components. We describe this in detail in the following: To generate the homophily components h_f, $f \in \{\alpha, \phi\}$,

$$\begin{aligned} v = &\ \max \text{ eigenvector}(L^{\dagger/2} X L^{\dagger/2}), \\ h_f = &\ \sqrt{\eta} L^{\dagger/2} v, \\ h_f = &\ h_f - \left\langle h_f, \frac{\mathbf{1}}{n} \right\rangle \mathbf{1} \end{aligned}$$

In the above, we leverage the proof to the bounds in Lemma 5.J.5 to generate the worst-case homophily components. Indeed, let $w = L^{1/2} h_f / \sqrt{\eta}$, then, the h_f that achieves $\sup_{\|L^{1/2} h_f\|_2^2 \leq \eta} {h_f}^T L^{1/2} L^{\dagger/2} X L^{\dagger/2} L^{1/2} h_f$ is the same as the $\sqrt{\eta} L^{\dagger/2} w$ for the w that achieves $\sup_{\|w\|_2^2 \leq 1} w^T L^{\dagger/2} X L^{\dagger/2} w$, which is simply the maximum eigenvector of $L^{\dagger/2} X L^{\dagger/2}$.

To generate the heterogeneous variation components ϵ_f, $f \in \{\alpha, \phi\}$ $q^* = 2$, we take

$$\begin{aligned} \Sigma &= \kappa X / \|X\|_F, \\ z &\sim \mathcal{N}(0, I_n), \end{aligned}$$

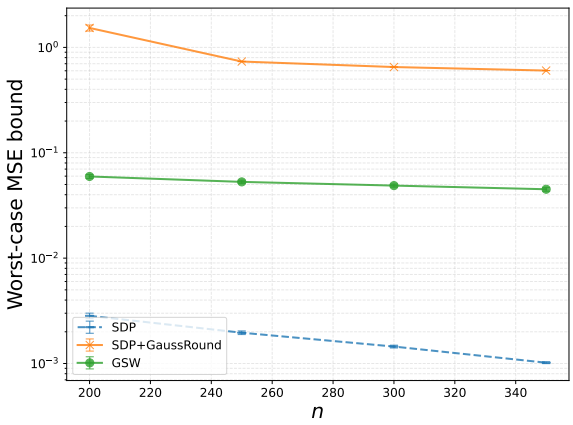


Figure 5.L.13: Comparisons with adapted Gram-Schmidt algorithm averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$) with four clusters of membership assignment probability (0.70, 0.1, 0.1, 0.1) (see Graph 3 in Figure 5.L.1), for different n. Error bars represent the standard error.

$$\epsilon_f = \Sigma^{1/2} z.$$

If $q^* = 1$ instead, we replace Σ above with $\Sigma = \kappa v v^T$, where v is the leading eigenvector of X. Indeed, it is not difficult to see that these are the ϵ_f that achieve the bounds in Lemma 5.J.4 and Proposition 5.4.4.

Finally, let $\tilde{\alpha} = h_\alpha + \epsilon_\alpha$, and $\tilde{\phi} = h_\phi + \epsilon_\phi$. Recall that here $a = \left\langle \alpha_0, \frac{\mathbf{1}}{n} \right\rangle$, and $b = \left\langle \phi_0, \frac{\mathbf{1}}{n} \right\rangle$. We take $\alpha = a\mathbf{1} + (\tilde{\alpha} - \left\langle \tilde{\alpha}, \frac{\mathbf{1}}{n} \right\rangle \mathbf{1})$ and $\phi = b\mathbf{1} + (\tilde{\phi} - \left\langle \tilde{\phi}, \frac{\mathbf{1}}{n} \right\rangle \mathbf{1})$.

To generate the interference component for each treatment assignment vector $x = 2z - 1$ generated from X from the corresponding covariance matrix, we take

$$s(z) = \frac{\sqrt{\gamma} L x}{\|x^T L^{1/2}\|_2},$$
$$s(z) = s(z) - \left\langle s(z), \frac{\mathbf{1}}{n} \right\rangle \mathbf{1}.$$

Indeed, through the proof the Cauchy-Schwarz inequality, it is not difficult to see that the $s(z)$ that achieves $\sup_{\|L^{\dagger/2} s(z)\|_2^2 \leq \gamma} \mathbb{E}[\langle x, s(z) \rangle^2]$, must satisfy $L^{\dagger/2} s(z) = \gamma^{1/2} L^{1/2} x / \|L^{1/2} x\|_2$.

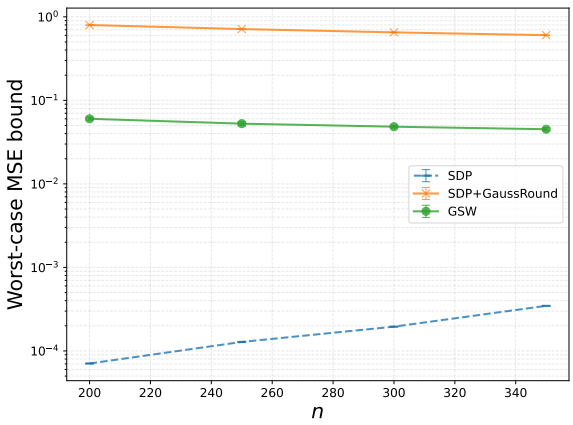


Figure 5.L.14: Comparisons with adapted Gram-Schmidt algorithm averaged over graphs generated from a Barabasi-Albert (preferential-attachment) model with initial $n/10$ (=10) nodes connected according to the Erdös-Rényi model with connection probability $10/n$ (see Graph 4 in Figure 5.L.1), for different n . Error bars represent the standard error.

5.L.4 Priors and Worst-Case

We take this opportunity to illustrate how our framework allows for the incorporation of priors in the bounds to improve the performance of the design otherwise tailored to the worst-case setting. Indeed, if the practitioner strongly believed that the homophily components of the potential outcomes followed a certain type of distribution, then we can refine the trade-off parameter bounds. In particular, if the practitioner believed that the homophily components followed a uniform distribution on a sphere, much like in the data generation procedure in the previous paragraph, we can replace the η parameter by η/n. Indeed, if $h \sim {L^\dagger}^{1/2}\mathrm{Unif}(S_{n-1}^{\eta})$ where S_n^{η} is the $(n-1)$-sphere of radius $\sqrt{\eta}$, i.e., the scaled normalized gaussians in the previous paragraph, then,

$$\begin{aligned}\mathbb{E}[(h^T x)^2] &= \mathrm{Tr}(X\mathbb{E}[{L^\dagger}^{1/2} v v^T {L^\dagger}^{1/2}]) \\ &= \mathrm{Tr}(X {L^\dagger}^{1/2}\mathbb{E}[vv^T]{L^\dagger}^{1/2}) \\ &= \frac{\eta}{n}\mathrm{Tr}(XL^\dagger).\end{aligned}$$

Similarly, if the practitioner believed that the interference $s(z)$ followed a uniform distribution on the sphere, we can replace γ by γ/n. Indeed, if $s(z) \sim L^{1/2}\mathrm{Unif}(S_{n-1}^{\gamma})$, where S_{n-1}^{γ} is the

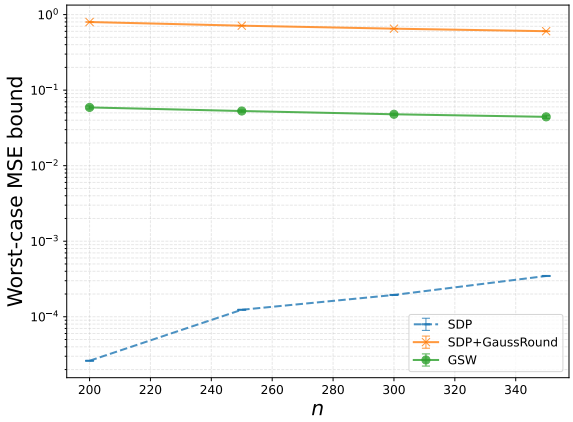


Figure 5.L.15: Comparisons with adapted Gram-Schmidt algorithm averaged over graphs generated from an Erdös-Rényi model with connection probability $2/n$ (see Graph 5 in Figure 5.L.1), for different n. Error bars represent the standard error.

$(n-1)-$sphere of radius $\sqrt{\gamma}$, then,

$$\begin{aligned}\mathbb{E}[(s(z)^T x)^2] &= \operatorname{Tr}(X\mathbb{E}[L^{1/2}vv^T L^{1/2}]) \\ &= \operatorname{Tr}(XL^{1/2}\mathbb{E}[vv^T]L^{1/2}) \\ &= \frac{\gamma}{n}\operatorname{Tr}(XL).\end{aligned}$$

5.L.5 Comparisons between SDP and adapted Gram-Schmidt Walk designs

In Figures 5.L.11 to 5.L.15, we compare the worst-case bounds of the designs produced by the SDP and adapted Gram-Schmidt Walk algorithms for $n \in \{200, 250, 300, 350\}$. As expected, the SDP solutions yield uniformly smaller worst-case bounds, since the SDP directly optimizes the worst-case objective. However, over this range of n, we do not observe a clear trend in the gap between the SDP- and Gram–Schmidt-Walk-based bounds that would reflect the predicted looseness of the adapted Gram–Schmidt Walk bound at large n. This is likely because we only consider a limited range of n, as solving the SDP becomes computationally prohibitive for larger n. Finally, for all the simulated networks besides the SBM graphs with two equal-membership clusters, the adapted Gram–Schmidt Walk algorithm yields smaller bounds than Gaussian rounding applied to the SDP solutions.

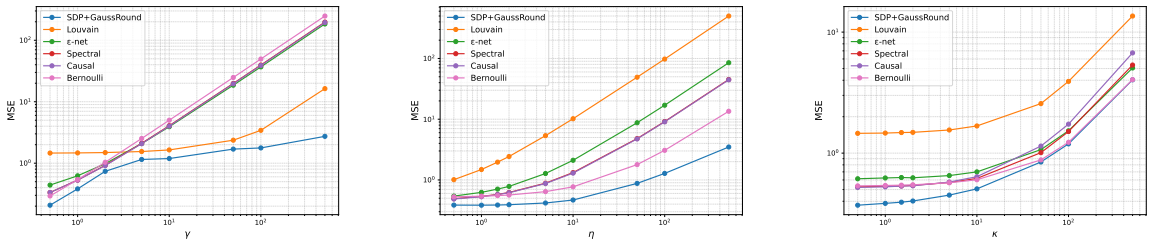


Figure 5.L.16: Comparison of MSEs between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$), with two clusters of equal membership assignment probability (see Graph 1 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

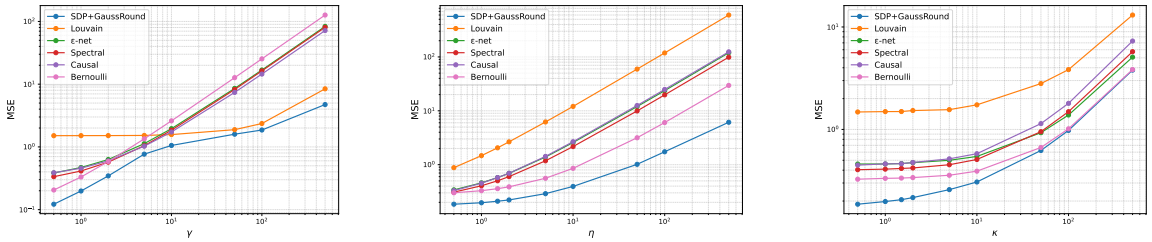


Figure 5.L.17: Comparison of MSEs between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$), with four clusters of equal membership assignment probability (see Graph 2 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

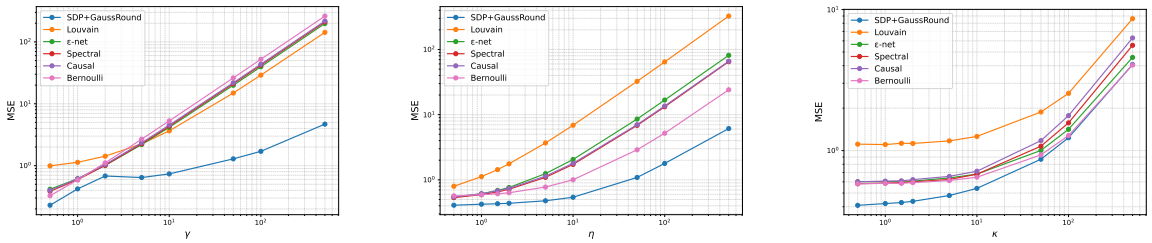


Figure 5.L.18: Comparison of MSEs between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from an SBM (within cluster connection probability $20/n$, across connection probability $\frac{0.1}{n}$) with four clusters of membership assignment probability (0.7, 0.1, 0.1, 0.1) (see Graph 3 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

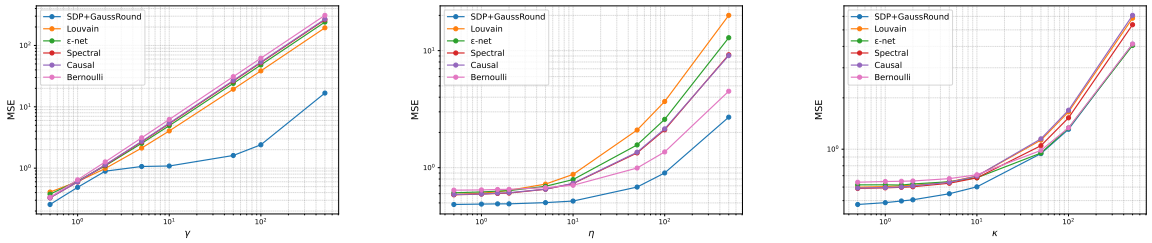


Figure 5.L.19: Comparison of MSEs between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from a Barabasi-Albert (preferential-attachment) model with initial $n/10$ (=10) nodes connected according to the Erdös-Rényi model with connection probability $10/n$ (see Graph 4 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

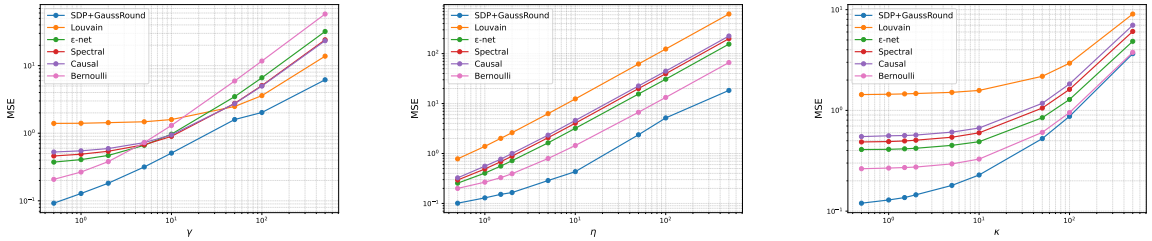


Figure 5.L.20: Comparison of MSEs between our designs, cluster randomization designs (with various clusterings) and Bernoulli randomization averaged over graphs generated from an Erdös-Rényi model with connection probability $2/n$ (see Graph 5 in Figure 5.L.1). The other two trade-off parameters are set to one. Error bars represent the standard error.

Chapter 6

DISCUSSION

This dissertation studies statistical problems that arise when the units in a population are dependent, with a particular focus on settings where this dependence is represented by a network. The chapters consider different parts of the statistical problem: obtaining asymptotic inference under dependent observations, detecting whether interference is present, estimating treatment effects when the interference structure is not completely known, and designing experiments to account for interference and other network structure before the outcomes are observed.

Dependence is both a complication and a source of information. If the dependence is ignored, procedures developed for independent observations may no longer provide valid inference or accurate treatment effect estimates. At the same time, when something is known about the structure generating this dependence, this information can be used to improve the statistical procedure. The results in this dissertation explore different angles of this idea.

Chapter 2 takes the most general and fundamental perspective. Rather than requiring dependence to vanish outside of a fixed neighborhood, the affinity-set framework separates relatively strong and weak dependence and places conditions directly on the resulting covariance quantities. Chapters 3–5 use more problem-specific network information. Chapter 3 uses the graph to select focal units that are more informative for testing interference. Chapter 4 uses network exposures together with the observed outcomes to adaptively select an exposure threshold. Finally, Chapter 5 uses the graph before the experiment to construct treatment assignment distributions that account for interference, homophily, and heterogeneous variation.

There are several broader lessons and open questions that come from considering these problems together.

6.1 Using Network Structure without Assuming it is Complete

Throughout the dissertation, the network is useful because it provides information about how units are related. At the same time, in most applications it would be unrealistic to assume that the

observed graph completely characterizes the dependence among the units.

This distinction appears in different forms across the chapters. In Chapter 2, observations outside of an affinity set are not required to be independent. Thus, an affinity set is not intended to identify all dependence associated with a particular observation. Instead, it should capture enough of the relatively strong dependence that what remains outside of the set is small in aggregate. This makes it possible, for example, to consider diffusion processes where every pair of infection indicators may be correlated at finite n.

A similar issue appears more directly in the interference setting. Exposure mappings simplify the dependence of a potential outcome on the full treatment assignment by mapping the assignment to a lower-dimensional exposure. Such a mapping is necessarily an approximation unless the interference mechanism is known exactly. Chapter 4 addresses one part of this uncertainty by allowing the exposure threshold itself to be selected from the data. Rather than specifying in advance exactly how much neighboring treatment is required before a unit should be regarded as treated-exposed, the procedure estimates the bias–variance trade-off associated with different choices.

Chapter 5 takes a different approach. Instead of trying to learn the full potential outcome model, the unknown structure is divided into components that can be associated with homophily, interference, and heterogeneous variation. The heterogeneous variation component is especially important in this interpretation. It represents the part of the potential outcomes that is not explained by the graph-based structure used to construct the design. Increasing the corresponding robustness parameter encourages additional randomization, moving the design away from one that relies strongly on the observed network.

There are two extremes that are generally unattractive. On one end, ignoring the network throws away information that can substantially improve statistical performance. On the other end, assuming that the observed network and the chosen exposure model exactly describe the underlying dependence can lead to procedures that are very sensitive to misspecification. The methods in this dissertation instead use the observed structure while allowing, in different ways, for some dependence or variation to remain outside of it.

An important future direction is to make this treatment of network uncertainty more explicit. For example, if several candidate networks are available, it would be useful to develop procedures that adapt across them rather than committing to one network before analysis. A similar problem arises when edge weights are estimated from historical interactions or when the observed network

is a noisy version of a latent interference network. In such settings, the uncertainty in the network could be incorporated directly into the inferential, estimation, or design objective.

6.2 Optimization across the Experimental Pipeline

Another recurring perspective in this dissertation is that optimization provides a natural framework for thinking about statistical decision-making under network dependence and interference. Once a statistical goal, such as maximizing power or minimizing mean squared error for a particular problem setting, is formalized, choices about which observations to use, how to construct an estimator, or how to assign treatments can be viewed as optimization problems.

In Chapter 3, the experiment has already been run and the goal is testing. The choice is which units should be focal units in a conditional randomization test. The power analysis shows that this selection should account not only for the number of focal units, but also for the amount of residualized exposure variation induced by rerandomizing the auxiliary units. This essentially turns focal selection into an effective sample size maximization problem.

In Chapter 4, both the assignments and outcomes have already been observed. Here, the statistical decision is the exposure threshold used by the estimator. Increasing the threshold reduces the bias induced by interference but also reduces exposure probabilities and increases variance. Thus, the appropriate optimization objective is the estimated MSE over the candidate thresholds.

Chapter 5 moves this decision before the experiment. The treatment assignment mechanism itself is chosen to minimize a worst-case MSE bound. In this setting, the optimization is over the covariance matrix of the design. The different components of the objective encode different statistical considerations: interference favors more correlated treatment assignments among nearby units, homophily favors assignments that are spread across the graph, and heterogeneous variation favors additional randomization.

Viewed together, these chapters suggest a broader point: design and analysis under interference should not necessarily be treated as completely separate problems. A design determines the exposure distribution observed by the estimator, and the estimator determines which features of the exposure distribution are useful. Similarly, a design that makes interference easier to estimate may also make it easier to detect. For example, Chapter 4 can be loosely viewed as adapting the analysis to the exposure variation produced by a fixed design, while Chapter 5 adapts the design so that the resulting outcomes are easier to estimate with a fixed estimator. A natural extension would be to

optimize these two objects jointly. One could, for example, choose a treatment design together with an exposure mapping or threshold to minimize the resulting worst-case MSE.

A related direction is to connect the testing problem in Chapter 3 to the design problem in Chapter 5. If the primary goal of an initial experiment is to determine whether meaningful interference exists, the optimal design for that experiment may differ substantially from one optimized for GATE estimation. One could first design an experiment to maximize power for detecting interference and then, depending on the result, use a second experiment designed for treatment effect estimation. More generally, this suggests sequential designs in which information about the interference structure is learned and then used to update future randomizations.

6.3 Robustness and the Amount of Structure to Use

The gains from using network structure depend on how accurately that structure represents the data-generating process. This creates a recurring tension between efficiency and robustness.

Chapter 3 gives one relatively clean separation between these two goals. The linear interference model motivates the focal-set criterion and therefore affects the power of the procedure. It is not, however, needed for the finite-sample validity of the conditional randomization test. Thus, a misspecified model may result in a focal set that is not power-optimal, while leaving the test valid. This is an attractive form of model-assisted inference: the model is used to improve performance without being required for the basic guarantee.

In Chapter 4, the role of the working model is somewhat stronger. The exposure regression is used to estimate how the bias changes as the threshold changes. Nevertheless, the linear model is used as a first-order approximation rather than assumed to be the exact potential outcome model. The simulation results with nonlinear exposure functions, together with the local linear extensions, illustrate how the same basic procedure can be useful outside of the exact linear setting.

In Chapter 5, this uncertainty is incorporated directly into the optimization problem. If the experimenter has strong information about the homophily and interference structure, a more structured and potentially nearly deterministic design can perform well. If there is substantial heterogeneous variation not explained by this structure, randomization becomes more important. Thus, randomization can be interpreted as the price paid for robustness to what is not known.

This raises the practical question of how much confidence should be placed in the available structural information. In Chapter 5 this is summarized through the parameters η, γ, and κ. Similar

unknown quantities appear implicitly in the other chapters: the appropriate size of affinity sets in Chapter 2, the strength and functional form of interference used for power analysis in Chapter 3, and the accuracy of the exposure-response approximation in Chapter 4.

In practice, the amount of robustness and structural information that should be incorporated into these procedures is generally not known a priori. Chapter 5 gives several practical approaches for calibrating these quantities from historical experiments, pilot studies, and pre-experiment data. For example, previous experiments at different treatment saturation levels can be used to estimate the effective spillover strength γ. The heterogeneous variation parameter κ can be informed by the covariance structure of residuals from historical outcomes, while the homophily parameter η can be estimated by measuring the graph smoothness of baseline outcomes or of outcome components predicted from observed covariates. More generally, Chapter 5 proposes fitting historical outcomes using network exposures and covariates, and using the fitted and residual components to calibrate the corresponding interference, homophily and heterogeneous-variation parameters.

These approaches provide practical methods for choosing the trade-off parameters, but are still largely heuristic. An important direction is therefore to develop a more systematic statistical framework for learning these quantities from previous data. For example, one could study estimators $\widehat{\eta}$, $\widehat{\gamma}$, and $\widehat{\kappa}$ obtained from one or several historical experiments, characterize their estimation error, and determine how this error propagates through the optimization problem to the performance of the resulting design.

The same idea applies more broadly across the experimentation pipeline. Historical experiments could be used to learn plausible interference strengths or exposure-response relationships before selecting focal units in Chapter 3. For the estimation problem in Chapter 4, previous experiments could provide information about the exposure-response function and the size of the approximation error used in estimating the bias, and therefore a more appropriate exposure threshold as it interpolates between structure and uncertainty. Repeated experiments could also provide information about how stable these quantities are across networks or populations, allowing the procedure to distinguish between uncertainty arising from limited data within one experiment and genuine variation across experimental settings.

Thus, a useful extension of the methods in this dissertation would be to move from user-specified or heuristically calibrated uncertainty levels toward principled and data-adaptive ones. Historical experiments, pilot studies, and repeated observations of related networks could be used to learn

how much structural information should be trusted, and therefore how much robustness should be incorporated at each stage of testing, estimation, and experimental design.

6.4 Connecting Covariate Balancing and Interference-Aware Design

Chapter 5 develops a connection between the covariate-balancing and network interference literatures. A standard covariate-balancing objective can often be written as

$$\|C^\top x\|_2^2 = x^\top C C^\top x,$$

where C is a matrix of unit-level covariates and x denotes the treatment assignment. More generally, whenever a positive semidefinite component of a statistical objective can be written as

$$x^\top Q x = \|Ax\|_2^2, \qquad Q = A^\top A,$$

the columns of A can be interpreted as transformed unit-level covariates to be balanced by the treatment assignment. This gives a useful way to translate certain network-design objectives into covariate-balancing problems.

The graph-based quantities arising in Chapter 5 have exactly this structure. In particular,

$$x^\top L x = \|L^{1/2} x\|_2^2, \qquad x^\top L^\dagger x = \|L^{\dagger/2} x\|_2^2.$$

The two choices induce different design incentives. Balancing the covariates associated with $L^{1/2}$ is equivalent to reducing treatment disagreements across edges and therefore encourages clustered assignments, which helps control interference in GATE estimation. Balancing those associated with $L^{\dagger/2}$ instead encourages treatment to be spread across the large-scale graph structure relevant to homophily. Chapter 5 combines these graph-induced covariates, together with robustness to heterogeneous variation, within the Gram–Schmidt Walk framework.

This representation suggests that other methods developed for covariate balancing could potentially be extended to network interference settings when the relevant statistical objective admits a similar form. For example, rerandomization procedures could use acceptance criteria involving both observed baseline covariates and graph-induced covariates. Matching or stratification procedures could similarly use both observed characteristics and appropriate graph representations. Conversely, interference-

aware objectives that can be written, or tightly approximated, as quadratic forms in the treatment assignment could potentially be combined with ordinary covariate-balancing objectives within a common framework. This could provide a way to construct designs that account jointly for network interference and covariate imbalance, rather than addressing the two considerations separately.

At the level of a randomized design, quadratic objectives satisfy

$$\mathbb{E}[x^\top Q x] = \mathrm{Tr}(QX), \qquad X = \mathbb{E}[xx^\top],$$

and therefore depend only on the second moment of the treatment assignment. More broadly, it would be interesting to characterize which interference and network-based objectives admit useful covariate-balancing representations, and which do not.

6.5 Sequential Network Observation and Adaptive Methods

Throughout this dissertation, the network structure is generally treated as observed before the statistical procedure is carried out. In many applications, however, this information may instead become available sequentially. New units may enter the population over time, new edges may be observed as interactions occur, or an initially partially observed network may become increasingly complete. An interesting extension of the methods in this dissertation is therefore to adaptive settings where statistical decisions are updated as the network is observed.

One way to formalize such a setting is to let $G_t = (V_t, E_t)$ denote the network observed at time t, with

$$G_1 \subseteq G_2 \subseteq \cdots \subseteq G_T,$$

where either new nodes, new edges, or both are revealed over time. At each time t, the statistician makes a decision using only the currently observed graph and the data available up to that point. This is different from simply applying the procedures in this dissertation repeatedly: earlier decisions cannot generally be changed once additional network information is observed, and those decisions may themselves affect the outcomes used to make later decisions.

There are natural versions of this problem for each of the settings studied in the dissertation. For the central limit theorem in Chapter 2, the sequential setting raises the question of whether asymptotic normality continues to hold when the dependence structure evolves with the observed network. Let G_t denote the network available at time t, and let $\mathcal{A}_i^{(t)}$ denote affinity sets appropriate for

the corresponding dependent observations. One could ask for conditions under which the covariance bounds of Chapter 2 hold uniformly or asymptotically along this evolving sequence, yielding a central limit theorem for estimators constructed from sequentially arriving observations. A further challenge arises when $\mathcal{A}_i^{(t)}$ is constructed from only the currently observed portion of the network: dependence induced by edges that have not yet been observed must then be controlled through the outside-affinity terms.

For interference detection in Chapter 3, focal-unit selection could similarly be carried out sequentially. As new nodes and edges are observed, the focal set could be updated to preserve or increase the amount of useful exposure variation. The optimization problem would then become an online version of the focal-selection problem: at each stage, one selects focal units based on the currently observed graph without knowing the future network. This raises a natural comparison between the power obtained by such a sequential procedure and the power of the offline procedure that observes the complete graph before selecting the focal set.

Chapter 4 is already adaptive in the sense that the exposure threshold is selected from the observed outcomes and exposures. A sequential version would allow this threshold to change as additional network information and outcomes become available. For example, one could maintain an estimate $\widehat{h}_t$ of the MSE-optimal threshold based on G_t and the observations up to time t, updating the estimated bias–variance trade-off as the distribution of exposures changes. This may be particularly useful when the initially observed network provides a poor approximation to the eventual exposure structure.

The design problem in Chapter 5 provides perhaps the most direct motivation for this extension. The current framework assumes that the network, together with bounds on homophily, interference, and heterogeneous variation, is available before the treatment assignment is constructed. In an adaptive version, these quantities could instead be learned sequentially. At time t, one could use the currently observed network G_t and estimates $\widehat{\eta}_t, \widehat{\gamma}_t, \widehat{\kappa}_t$ to determine the treatment assignment distribution for newly arriving units. As more of the graph and more outcomes are observed, these estimates could be updated and later assignments could place greater weight on the learned structure.

This leads naturally to an exploration–exploitation trade-off. Early in the experiment, when little is known about the interference or homophily structure, additional randomization may be valuable for robustness and for learning the relevant parameters. As information accumulates, future assignments could increasingly exploit the observed graph structure. This gives a sequential interpretation of

the robustness trade-off in Chapter 5: the amount of randomization need not be fixed before the experiment, but could decrease or otherwise adapt as uncertainty about the underlying structure is reduced. Indeed, Chapter 5 already suggests an online extension in which treatments are assigned sequentially while η, γ, and κ are learned from previous experiments or observations.

There are several additional difficulties that do not arise in the static setting. In particular, a unit may be assigned treatment before all of its eventual neighbors are known. Consequently, an assignment that is optimal with respect to G_t may perform poorly after additional edges are revealed. The mechanism through which nodes and edges are observed may itself also be informative; for example, highly active units may reveal more edges earlier. Thus, it may be necessary to model both uncertainty in the unobserved portion of the graph and uncertainty in the potential outcomes.

A useful theoretical objective in this setting would be to compare an adaptive procedure to an oracle that observes the complete graph in advance. For the testing problem this comparison could be made in terms of power, while for estimation and design it could be made in terms of MSE or worst-case MSE. One could then ask whether the additional cost of observing the network sequentially vanishes as more of the graph is revealed, or characterize the regret relative to the corresponding offline optimum.

More broadly, this direction would combine the main ideas developed throughout the dissertation: learning the structure of dependence, using that structure in an optimization problem, and retaining additional robustness when the available structural information is incomplete.

6.6 Scaling through Network Structure

The optimization frameworks developed in Chapters 3 and 5 are deliberately general, in the sense that they can be applied to broad classes of networks without requiring a particular graph structure. For sufficiently large networks, solving the corresponding mixed-integer or semidefinite optimization problems with generic solvers may no longer be practical. In these settings, an important direction is to tailor the optimization procedure to the underlying network and problem structure.

For focal-unit selection in Chapter 3, the exact combinatorial optimization problem becomes difficult as the number of nodes grows. However, the objective is itself induced by the network: the relevant information depends on the exposure variation generated through focal–auxiliary relationships and, under local interference, on relatively local portions of the graph. For sparse or highly structured networks, this suggests exploiting graph decompositions, separators, community structure, or restricted

candidate sets before applying the optimization procedure. For particular graph classes, such as graphs with bounded degree or other restricted structural complexity, more specialized algorithms may also be possible. The power characterization in Chapter 3 remains useful in this setting because it provides a statistical objective against which these more scalable procedures can be evaluated.

This issue is even more pronounced for the design problem in Chapter 5. The SDP optimizes over an $n \times n$ treatment-assignment covariance matrix, and therefore becomes computationally expensive when n is large. The adapted Gram–Schmidt Walk provides one more scalable alternative, but there remains substantial room for algorithms that exploit the particular structure of the graph-based objective. For example, the interference and homophily terms are determined by the graph Laplacian L and its pseudoinverse $L^\dagger$. Although $L^\dagger$ may be dense even when the original graph is sparse, these quantities have strong spectral structure. This suggests approaches based on low-rank spectral approximations, approximate Laplacian solves, graph partitioning, or decomposition across weakly connected portions of the network. One could also replace the full SDP with lower-dimensional or factorized optimization problems that retain the graph directions most relevant to the objective.

More generally, different network classes may call for different computational approaches. On graphs with clear community structure, one could first optimize at the level of communities and then refine assignments within them. On sparse graphs, one may be able to exploit locality in the interference structure. On graphs whose relevant Laplacian structure is approximately low-rank, the optimization may be carried out in a much lower-dimensional spectral representation. Thus, rather than seeking a single scalable algorithm for arbitrary graphs, it may be more useful to develop algorithms whose computational complexity adapts to the structural complexity of the network.

This also suggests theoretical questions that are different from generic worst-case approximation analysis. In addition to asking how the approximation error depends on n, one could characterize it in terms of graph quantities such as degree, spectral structure, effective rank, clusterability, or separator size. Such guarantees would make explicit when the general optimization formulations in Chapters 3 and 5 can be substantially simplified without losing much statistical efficiency.

From this perspective, the general optimization problems developed in this dissertation can be viewed as statistical targets rather than necessarily the final computational procedures. They identify the quantities that should be optimized for power or mean squared error. For large-scale networks, the next step is to use the specific structure of the application to approximate these targets efficiently.

6.7 Conclusion

The problems considered in this dissertation arise from a simple difficulty: once units interact, the statistical analysis can no longer be separated from the structure relating those units. Dependence affects asymptotic inference, interference affects which comparisons identify treatment effects, and the network affects the information generated by an experiment.

The main perspective developed here is to use this structure. Covariance structure can be used to establish central limit theorems even when dependence is not strictly local. Network structure can be used to select more informative focal units for detecting interference. Observed exposures can be used to adaptively trade off bias and variance when estimating treatment effects. Finally, the network can be incorporated before an experiment to choose treatment assignments that account jointly for interference, homophily, and uncertainty about the potential outcomes.

At the same time, the structure available to the researcher is rarely complete. This motivates the role of robustness throughout the dissertation: allowing weak dependence outside of affinity sets, using models to improve power without requiring them for test validity, treating exposure models as approximations, and retaining randomization when the observed network does not explain all heterogeneous variation.

There remains substantial room to connect these ideas more closely. In particular, jointly optimizing design and analysis, learning the relevant interference and robustness parameters from earlier experiments, developing inference after adaptive procedures, and constructing scalable optimization methods for large networks are natural next steps. More broadly, these problems suggest that statistical methods for interacting populations should account simultaneously for what is known about the dependence structure and for what remains unknown.

BIBLIOGRAPHY

Louis HY Chen and Qi-Man Shao. Normal approximation under local dependence. *The Annals of Probability*, 32(3):1985–2028, 2004.

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

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337, 2013.

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

Best Neighborhoods, n.d. https://bestneighborhood.org/race-in-chicago-il//, accessed August, 2025.

Christopher Harshaw, Fredrik Sävje, Daniel A Spielman, and Peng Zhang. Balancing covariates in randomized experiments with the Gram–Schmidt walk design. *Journal of the American Statistical Association*, 119(548):2934–2946, 2024.

D. R. Cox. *Planning of Experiments*. John Wiley & Sons, New York, 1958.

Donald B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978. doi: 10.1214/aos/1176344064.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, May 1990.

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

Peter M Aronow. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16, 2012.

Charles Stein. Approximate computation of expectations. *Lecture Notes-Monograph Series*, 7:i–164, 1986.

Arun G Chandrasekhar and Matthew O Jackson. A network formation model based on subgraphs. *Review of Economic Studies, forthcoming*, 2024.

Yosef Rinott and Vladimir Rotar. Normal approximations by stein's method. *Decisions in Economics and Finance*, 23:15–29, 2000.

Nathan Ross. Fundamentals of stein's method. *Probability Surveys*, 8:210–293, 2011.

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

Kenneth A Froot. Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in financial data. *Journal of Financial and Quantitative analysis*, 24(3):333–355, 1989.

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

John C Driscoll and Aart C Kraay. Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, 80(4):549–560, 1998.

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, 2002.

Shane Lubold, Arun G Chandrasekhar, and Tyler H McCormick. Identifying the latent space geometry of network models through analysis of curvature. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):240–292, 2023.

Pierre Baldi and Yosef Rinott. On normal approximations of distributions in terms of dependency graphs. *The Annals of Probability*, pages 1646–1650, 1989.

Larry Goldstein and Yosef Rinott. Multivariate normal approximations by stein's method and size bias couplings. *Journal of Applied Probability*, pages 1–17, 1996.

Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, pages 9–25, 1989.

Louis HY Chen. Poisson approximation for dependent trials. *The Annals of Probability*, 3(3):534–545, 1975.

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005. doi: 10.1214/154957805100000104.

Emmanuel Rio. Covariance inequalities for strongly mixing processes. *Annales de l'I.H.P. Probabilités et statistiques*, 29(4):587–597, 1993. doi: 10.1016/S0246-0203(93)80028-X.

Emmanuel Rio. *Weakly dependent sequences: theory and applications*. Springer, 2017. doi: 10.1007/978-3-319-54235-7.

Christophe Ange Napoléon Biscio, Arnaud Poinas, and Rasmus Waagepetersen. A note on gaps in proofs of central limit theorems. *Statistics & Probability Letters*, 135:7–10, 2018.

Joseph P Romano and Michael Wolf. A more general central limit theorem for m-dependent random variables with unbounded m. *Statistics & probability letters*, 47(2):115–124, 2000.

Richard C Bradley. *Introduction to strong mixing conditions*, volume 1. Kendrick Press, 2007.

Nazgul Jenish and Ingmar R Prucha. Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150(1):86–98, 2009.

Donald WK Andrews. Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21 (4):930–934, 1984.

Emmanuel Rio. Inequalities and limit theorems for weakly dependent sequences. 2013.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

C Alan Bester, Timothy G Conley, and Christian B Hansen. Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151, 2011.

Thomas Barrios, Rebecca Diamond, Guido W Imbens, and Michal Kolesár. Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*, 107 (498):578–591, 2012.

Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

Timothy G Conley and Giorgio Topa. Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics*, 17(4):303–327, 2002.

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

Patrick Billingsley. The lindeberg-levy theorem for martingales. *Proceedings of the American Mathematical Society*, 12(5):788–792, 1961.

IA Ibragimov. A central limit theorem for a class of dependent random variables. *Theory of Probability & Its Applications*, 8(1):83–89, 1963.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

Adrian Röllin. On quantitative bounds in the mean martingale central limit theorem. *Statistics & Probability Letters*, 138:171–176, 2018.

Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

David Puelz, Guillaume Basse, Avi Feller, and Panos Toulis. A graph-theoretic approach to randomization tests of causal effects under general interference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):174–204, 2022.

Eugene Katsevich and Aaditya Ramdas. On the power of conditional independence testing under model-x. *Electronic Journal of Statistics*, 16(2):6348–6394, 2022.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

Vydhourie Thiyageswaran, Alex Kokot, Jennifer Brennan, Marina Meila, Christina Lee Yu, and Maryam Fazel. Optimal design under interference, homophily, and robustness trade-offs. *arXiv preprint arXiv:2601.17145*, 2026a.

Wenshuo Wang and Lucas Janson. A high-dimensional power analysis of the conditional randomization test and knockoffs. *Biometrika*, 109(3):631–645, 2022.

Ronald Aylmer Fisher. Statistical methods for research workers. 1950.

Paul R Rosenbaum. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574, 1984.

Paul R Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the american statistical association*, 102(477):191–200, 2007.

Guillaume W Basse, Avi Feller, and Panos Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494, 2019.

Vydhourie Thiyageswaran, Tyler H McCormick, and Jennifer Brennan. Data-adaptive exposure thresholds under network interference. *Advances in Neural Information Processing Systems*, 38: 131360–131393, 2026b.

Mizuho Yanagi and Tomonari Sei. Improving randomization tests under interference based on power analysis. *arXiv preprint arXiv:2203.10469*, 2022.

Abba M Krieger, David Azriel, Michael Sklar, and Adam Kapelner. Improving the power of the randomization test. *arXiv preprint arXiv:2008.05980*, 2020.

Tadao Hoshino and Takahide Yanagi. Randomization test for the specification of interference structure. *arXiv preprint arXiv:2301.05580*, 2023.

Guillaume Basse, Peng Ding, Avi Feller, and Panos Toulis. Randomization tests for peer effects in group formation experiments. *Econometrica*, 92(2):567–590, 2024.

Wenxuan Guo, JungHo Lee, and Panos Toulis. Ml-assisted randomization tests for detecting treatment effects in a/b experiments. *arXiv preprint arXiv:2501.07722*, 2025.

Chao Gao, Christopher Harshaw, Fredrik Sävje, and Yitan Wang. On the impossibility of specification testing of interference models based on exposure mappings. *arXiv preprint arXiv:2605.09726*, 2026.

Jake Bowers, Mark M Fredrickson, and Costas Panagopoulos. Reasoning about interference between units: A general framework. *Political Analysis*, 21(1):97–124, 2013.

Jean Pouget-Abadie, Guillaume Saint-Jacques, Martin Saveski, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. Testing for arbitrary interference on experimentation platforms. *Biometrika*, 106(4):929–940, 2019.

Sarah Baird, J Aislinn Bohren, Craig McIntosh, and Berk Özler. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860, 2018.

Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497. PMLR, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Siddharth Joshi and Stephen Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2008.

Ragnar Frisch and Frederick V Waugh. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401, 1933.

Michael C Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

George Udny Yule. On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 79(529):182–193, 1907.

Jing Cai, Alain De Janvry, and Elisabeth Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, 2015.

Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

Daniel L Sussman and Edoardo M Airoldi. Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*, 2017.

Morgan Hardy, Rachel M Heath, Wesley Lee, and Tyler H McCormick. Estimating spillovers using imprecisely measured networks. *arXiv preprint arXiv:1904.00136*, 2019.

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

Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1):20150021, 2017.

Duncan J Watts. Small worlds: the dynamics of networks between order and randomness, 1999.

Damon Centola and Michael Macy. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, 113(3):702–734, 2007.

Guillaume W Basse and Edoardo M Airoldi. Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105(4):849–858, 2018.

Hongtao Zhu, Sizhe Zhang, Yang Su, Zhenyu Zhao, and Nan Chen. Integrating active learning in causal inference with interference: A novel approach in online experiments. *arXiv preprint arXiv:2402.12710*, 2024.

Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2):20180026, 2019.

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

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pages 9196–9205. PMLR, 2020.

Jianqing Fan and Irene Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, pages 2008–2036, 1992.

David Ruppert. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92(439):1049–1062, 1997.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR, 2018.

Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and James Robins. Lepski's method and adaptive estimation of nonlinear integral functionals of density. *arXiv preprint arXiv:1508.00249*, 2015.

OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

OV Lepskii. Asymptotically minimax adaptive estimation. ii. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3):433–448, 1993.

Oleg V Lepski and Vladimir G Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997.

Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

Alex Deng, Luke Hagar, Nathaniel T Stevens, Tatiana Xifara, and Amit Gandhi. Metric decomposition in a/b tests. 2024.

Jeffrey D Hart and Philippe Vieu. Data-driven bandwidth choice for density estimation based on dependent data. *The Annals of Statistics*, pages 873–890, 1990.

Aad W Van Der Vaart, Jon A Wellner, Aad W van der Vaart, and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Per-Gunnar Martinsson and Joel A Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

Alwyn Young. Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics*, 134(2):557–598, 2019.

Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

Arun G Chandrasekhar. Econometrics of network formation. 2016.

Arun G Chandrasekhar, Horacio Larreguy, and Juan Pablo Xandri. Testing models of social learning on networks: Evidence from two experiments. *Econometrica*, 88(1):1–32, 2020.

Noga Alon. Eigenvalues, geometric expanders, sorting in rounds, and ramsey theory. *Combinatorica*, 6(3):207–219, 1986.

Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):1–37, 2009.

Christos Gkantsidis, Milena Mihail, and Amin Saberi. Conductance and congestion in power law graphs. In *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and modeling of computer systems*, pages 148–159, 2003.

Fabian Kuhn, Thomas Moscibroda, and Rogert Wattenhofer. On the locality of bounded growth. In *Proceedings of the twenty-fourth annual ACM symposium on Principles of distributed computing*, pages 60–68, 2005.

Robert Krauthgamer and James R Lee. The intrinsic dimensionality of graphs. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 438–447, 2003.

Emmanuel Kowalski. *An introduction to expander graphs*. Société mathématique de France Paris, 2019.

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

Michael P Leung. Causal inference under approximate neighborhood interference. *Econometrica*, 90 (1):267–293, 2022a.

Ingvar Ziemann, Stephen Tu, George J Pappas, and Nikolai Matni. The noise level in linear regression with dependent data. *Advances in Neural Information Processing Systems*, 36, 2024.

Yandi Shen, Fang Han, and Daniela Witten. Exponential inequalities for dependent v-statistics via random fourier features. 2020.

Davide Viviano, Lihua Lei, Guido Imbens, Brian Karrer, Okke Schrijvers, and Liang Shi. Causal clustering: design of cluster experiments under network interference. *arXiv preprint arXiv:2310.14983*, 2023.

Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5–es, 2007.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation with high-dimensional data. Technical report, cemmap working paper, 2015.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Johan Ugander and Hao Yin. Randomized graph cluster randomization. *Journal of Causal Inference*, 11(1):20220014, 2023.

Jennifer Brennan, Vahab Mirrokni, and Jean Pouget-Abadie. Cluster randomized designs for one-sided bipartite experiments. *Advances in Neural Information Processing Systems*, 35:37962–37974, 2022.

Paul F Lazarsfeld, Robert K Merton, et al. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1):18–66, 1954.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

Matthew O Jackson, Stephen M Nei, Erik Snowberg, and Leeat Yariv. The dynamics of networks and homophily. Technical report, National Bureau of Economic Research, 2023.

Mayleen Cortez-Rodriguez, Matthew Eichhorn, and Christina Lee Yu. Analysis of two-stage roll-out designs with clustering for causal inference under network interference. *arXiv preprint arXiv:2405.05119*, 2024.

Zahra Fatemi and Elena Zheleva. Minimizing interference and selection bias in network experiment design. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 176–186, 2020.

Bradley Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. Causal inference for social network data. *Journal of the American Statistical Association*, 119(545):597–611, 2024.

Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6): 1115–1145, 1995.

Nikhil Bhat, Vivek F Farias, Ciamac C Moallemi, and Deeksha Sinha. Near-optimal AB testing. *Management Science*, 66(10):4477–4495, 2020.

Qianyi Chen, Bo Li, Lu Deng, and Yong Wang. Optimized covariance design for AB test on social network under interference. *Advances in Neural Information Processing Systems*, 36, 2024.

Kari Lock Morgan and Donald B Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

Xinran Li, Peng Ding, and Donald B Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Network exposure to multiple universes. In *NIPS Workshop: Social network and social media analysis: Methods, models and applications, Lake Tahoe, NV*, 2012.

Ravi Jagadeesan, Natesh S Pillai, and Alexander Volfovsky. Designs for estimating the treatment effect in networks with interference. *The Annals of Statistics*, 48(2):679–712, 2020.

Zahra Fatemi and Elena Zheleva. Network experiment designs for inferring causal effects under interference. *Frontiers in big Data*, 6:1128649, 2023.

Nick Doudchenko, Khashayar Khosravi, Jean Pouget-Abadie, Sebastien Lahaie, Miles Lubin, Vahab Mirrokni, Jann Spiess, et al. Synthetic design: An optimization approach to experimental design with synthetic controls. *Advances in Neural Information Processing Systems*, 34:8691–8701, 2021.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American mathematical monthly*, 69(1):9–15, 1962.

David Holtz, Felipe Lobel, Ruben Lobel, Inessa Liskovich, and Sinan Aral. Reducing interference bias in online marketplace experiments using cluster randomization: Evidence from a pricing meta-experiment on airbnb. *Management Science*, 71(1):390–406, 2025.

Jean Pouget-Abadie, Vahab Mirrokni, David C Parkes, and Edoardo M Airoldi. Optimizing cluster-based randomized experiments under monotonicity. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2090–2099, 2018.

Christopher Harshaw, Fredrik Sävje, David Eisenstat, Vahab Mirrokni, and Jean Pouget-Abadie. Design and analysis of bipartite experiments under a linear exposure-response model. *Electronic Journal of Statistics*, 17(1):464–518, 2023.

Fredrik Sävje. Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, 111(1):1–15, 2024.

Bar Weinstein and Daniel Nevo. Causal inference with misspecified network interference structure. *Biometrics*, page ujag023, 2026.

Wenrui Li, Daniel L Sussman, and Eric D Kolaczyk. Causal inference under network interference with noise. *arXiv preprint arXiv:2105.04518*, 2021.

Carlo Gaetan, Xavier Guyon, et al. *Spatial statistics and modeling*, volume 90. Springer, 2010.

Ronald Aylmer Fisher. Statistical methods for research workers. *Journal of the Ministry of Agriculture*, 1925.

Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

Chien-Fu Wu. On the robustness and efficiency of some randomized designs. *The Annals of Statistics*, pages 1168–1177, 1981.

Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):85–112, 2018.

Christina Lee Yu, Edoardo M Airoldi, Christian Borgs, and Jennifer T Chayes. Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119, 2022.

Keisuke Hirano and Jack R Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5): 1683–1701, 2009.

Michael P Leung. Rate-optimal cluster-randomized designs for spatial interference. *The Annals of Statistics*, 50(5):3064–3087, 2022b.

Vydhourie Thiyageswaran, Tyler McCormick, and Jennifer Brennan. Data-adaptive exposure thresholds for the horvitz-thompson estimator of the average treatment effect in experiments with network interference. *arXiv preprint arXiv:2405.15887*, 2024.

Matthew Eichhorn, Samir Khan, Johan Ugander, and Christina Lee Yu. Low-order outcomes and clustered designs: combining design and analysis for causal inference under network interference. *arXiv preprint arXiv:2405.07979*, 2024.

Angela Fontan and Claudio Altafini. On the properties of laplacian pseudoinverses. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 5538–5543. IEEE, 2021.

Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 563–568, 2008.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Yurii Nesterov. Quality of semidefinite relaxation for nonconvex quadratic optimization. Technical report, Université catholique de Louvain, Center for Operations Research and . . . , 1997.

Prasad Raghavendra and Ning Tan. Approximating csps with global cardinality constraints using SDP hierarchies. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 373–387. SIAM, 2012.

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(226):1–92, 2018.

Sergio Currarini, Matthew O Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

Matthew O Jackson. Inequality's economic and social roots: the role of social networks and homophily. *arXiv preprint arXiv:2506.13016*, 2025.

L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2331932.

Ryan O'Donnell. *Analysis of boolean functions*, volume 2. Cambridge University Press Cambridge, 2014.

Abhijit Banerjee, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. The Diffusion of Microfinance, 2013b. URL https://doi.org/10.7910/DVN/U3BIHX.